Université de Montréal

Application des librairies de codons dégénérés à l'étude du mécanisme de repliement et de la stabilisation de la structure du domaine liant *ras* de Raf

par
François-Xavier Campbell-Valois

Département de biologie moléculaire
Faculté des études supérieures

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Docteur
en biologie moléculaire

Décembre 2005

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée :

Application des librairies de codons dégénérés à l'étude du mécanisme de repliement et de la stabilisation de la structure du domaine liant *ras* de Raf

présentée par :

François-Xavier Campbell-Valois

a été évaluée par un jury composé des personnes suivantes :

Luc Desgroseillers, président-rapporteur
Stephen W. Michnick, directeur de recherche
Sergueï Chteinberg, membre du jury
Allan R. Davidson, examinateur externe
Luc Desgroseillers, représentant du doyen de la FES

Résumé

Les analogues structuraux sont des protéines à la structure tertiaire semblable, mais qui ne possèdent aucune homologie de séquence significative. Elles partagent donc la même topologie structurale tout en ne démontrant aucune autre caractéristique commune, tel que la fonction cellulaire ou l'origine évolutive. En ce sens, l'étude comparative de la réaction de repliement des protéines adoptant des topologies similaires représentent une stratégie intéressante afin d'améliorer notre compréhension de ce phénomène, bien qu'elle soit limité de façon intrinsèque aux topologies structurales fortement répandues.

La topologie typique de l'ubiquitine est rencontrée extrêmement souvent parmi les structures connues. Nous avons voulu vérifier expérimentalement les variations naturelles de séquence observées dans l'alignement de protéine adoptant cette topologie, particulièrement aux positions les plus conservées et déterminer le rôle de ces résidus dans la formation et la stabilisation de la structure. Le domaine liant *ras* (DLR) de Raf a été choisi comme modèle de la topologie d'ubiquitine, afin de le soumettre à une perturbation de séquence pratiquement exhaustive, qui a consisté à remplacer chaque codon original par un codon dégénéré permettant l'insertion des 20 types d'acides aminés et cela en 13 segments indépendants de séquence contigus. De façon remarquable, la fréquence des acides aminés observés à chaque position parmi ces variants du DLR de Raf qui ont conservé la capacité de former la structure native est très semblable à celle observée dans l'alignement des séquences des membres de la topologie d'ubiquitine. Les positions les mieux conservées correspondent principalement au cœur hydrophobe et les résultats suggèrent que la diversité de séquence obtenue par cette approche pourrait être valable pour étudier les déterminants des topologies protéiques peu répandues et faciliter leur synthèse *de novo*.

Ensuite, nous avons entrepris de déterminer par l'insertion de mutations ponctuelles le rôle de résidus choisis du DLR de Raf dans le repliement et la stabilisation de sa

structure. Dans un premier temps, une corrélation entre la déstabilisation induite par la mutation d'un résidu en alanine (ou en glycine pour les résidus alanine) et le niveau de conservation observé expérimentalement à un résidu donné a été démontré, indiquant un rôle prépondérant de la stabilité dans la pression sélective. Il a été aussi démontré que la stabilité du DLR de Raf n'est pas optimisée et que cela est en partie dû à sa fonction de liaison à *ras*. Finalement, une analyse des valeurs-Φ a permis de faire des prédictions concernant le mécanisme de repliement du DLR de Raf et de démontrer sa similarité avec celui d'ubiquitine, en accord en cela avec le principe que la topologie fixent des contraintes générales qui déterminent le mécanisme de repliement des protéines structuralement similaires. En résumé, les travaux présentés dans cette thèse ont permis d'approfondir les liens entre la séquence polypeptidique, les divers aspects de la structure soit sa formation et sa stabilisation ainsi que la fonction de liaison du DLR de Raf et la conservation de séquence dans les alignements de séquence de membres de la topologie d'ubiquitine.

Mots-clés : protéine, structure, repliement, topologie d'ubiquitine, interaction protéine-protéine, déterminants de séquence, librairies, DLR de Raf, stabilité, analyse des valeurs-Φ.

Abstract

Structural analogs are proteins displaying similar overall tertiary structure despite having no significant sequence homology. They are said to adopt a common topology, but rarely share other characteristics such as cellular function or evolutionary origin. In that sense, the study and the comparison of folding of proteins sharing similar topology has proven to be an interesting approach to enrich the comprehension of this phenomenon, although it is intrinsically limited to frequently occurring structural topologies.

The topology typical of ubiquitin is extremely frequent among known structures. We wanted to verify experimentally the natural sequence variation observed in the alignment of proteins adopting this topology, particularly at the most conserved positions and determine their roles in the formation and stabilization of the structure. We have selected the Raf binding domain of *ras* (RBD) as a model of the ubiquitin topology, and submitted it to quasi-exhaustive sequence perturbation by replacing every wild-type codon by a degenerate codon, which allowed for the insertion of the 20 amino acids in 13 independent segments contiguous in the sequence. Remarkably, the variations in occurrence of each amino acid observed in those Raf RBD mutants that retained the capacity to fold are very similar to those observed in sequence alignments of structural analogues classified in the ubiquitin topology. The better conserved residues correspond principally to the hydrophobic core of the protein and the results suggest that this approach could be suitable for studying the sequence determinants of poorly populated protein topologies and facilitate their *de novo* synthesis.

Next, we sought to determine through the insertion of point mutations the role of selected residues of Raf RBD in the folding and stabilization of its structure. First, a correlation between the destabilization induced by alanine mutation (or glycine in the case of alanine residues) and the level of conservation observed experimentally at a given residue was demonstrated, indicating that stability is a dominant factor in selective

pressure. It was also demonstrated that the Raf RBD stability is not optimized by evolution and that this could be partly attributed to its *ras* binding function. Finally, a Φ-value analysis has allowed for predicting the folding mechanism of Raf RBD and to reveal its similarity to ubiquitin's, in agreement with the principle that the topology fixes the general constraints that determine the folding mechanism of structurally similar proteins. In summary, this thesis has allowed to explore and expand the relationship between the polypeptide sequence, the diverse aspects of the structure e.g., its folding and stabilization, and the binding function of the Raf RBD, as well as the sequence conservation in alignments of proteins adopting the ubiquitin topology.

Keywords : protein, structure, folding, ubiquitin topology/fold, protein-protein interaction, sequence determinants, libraries, Raf RBD, stability, Φ-value analysis.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

Acides aminés (code 1 lettre/code 3 lettres)

A/Ala	Alanine
C/Cys	Cystéine
D/Asp	Aspartate
E/Glu	Glutamate
F/Phe	Phenylalanine
G/Gly	Glycine
H/His	Histidine
I/Ile	Isoleucine
K/Lys	Lysine
L/Leu	Leucine
M/Met	Methionine
N/Asn	Asparagine
P/Pro	Proline
Q/Gln	Glutamine
R/Arg	Arginine
S/Ser	Sérine
T/Thr	Thréonine
V/Val	Valine
W/Trp	Tryptophane
Y/Tyr	Tyrosine

AcP	Acylphosphatase
ACBP	Protéine liant l'acétyl-CoA
ADN	Acide désoxyribonucléique
AFM	Microscopie de force atomique

ANS	1-anilino-8-naphtalènesulfonater
ARN_m	Acide ribonucléique messager
ARN_t	Acide ribonucléique de transfert
ATP	Adénosine triphosphate
β_t	Beta-Tanford
C_α	Carbone alpha
CASP	« Critical Assesment of structure prediction »
CATH	« Class, architecture, topology and homologous superfamily » (banque de données)
CI2	Inhibiteur de la chymotrypsine 2
CKAAPs	« Conserved Key Amino Acids in Protein sequences » (banque de données)
CsPB	Protéine du choc au froid de *Bacillus Subtilis*
D	Dénaturant
[D]	Concentration de dénaturant
DC	Dichroïsme circulaire
D-C	Double-cortine
$\Delta\Delta G_{F-U}$	Variation de la différence d'énergie libre
$\Delta\Delta G_{U-‡}$	Variation de la différence d'énergie libre de repliement (U→‡)
$\Delta\Delta G_{‡-F}$	Variation de la différence d'énergie libre de dépliement (F→‡)
ΔG_0	Différence d'énergie libre en absence de dénaturant
ΔG_{F-U}	Différence d'énergie libre
$\Delta G_{U-‡}$	Différence d'énergie libre de repliement
$\Delta G_{‡-F}$	Différence d'énergie libre de dépliement
ΔS	Différence d'entropie
ΔH	Différence d'enthalpie
DHFR	Dihidrofolate réductase
DHM	Type de codon dégénéré

DHW	Type de codon dégénéré
DLR	Domaine liant *ras*
E. coli	*Escherichia coli*
EGF	Facteur de croissance épithélial
F	État natif
FRET	Transfert d'énergie de résonance de Förster
FSSP	« Families of structrally similar proteins » (banque de données)
ϕ	Angle fi
Φ	Valeur-Fi
G_F	Niveau d'énergie de Gibbs de l'état natif
GTP	Guanosine triphosphate
GTPase	Enzyme hydrolysant le GTP (dans cette thèse il s'agit spécifiquement des petites GTPases)
Gdm-HCl	Guanidine d'hydrochlorure
G_U	Niveau d'énergie de Gibbs de l'état dénaturé
H_2O	Eau
I	État intermédiaire
I→F	Transition de l'état intermédiaire vers l'état natif
Im7	Protéine de l'immunité liant E colicin type 7
Im9	Protéine de l'immunité liant E colicin type 9
K_{eq}	Constante d'équilibre
k_{obs}	Taux observé
k_f	Taux de repliement
k_u	Taux de dépliement
m	Variation de $\Delta G_{F\text{-}U}$ en fonction de la concentration de dénaturant
m_f	Variation de $\ln k_f$ en fonction de la concentration de dénaturant
m_u	Variation de $\ln k_u$ en fonction de la concentration de déna
NADPH	Forme réduite de la nicotinamide adénine dinucléotide phosphate

Na_2SO_4	Sodium sulfate
NIH	« National institute of Health » des USA
NNK	Type de codon dégénéré
NNS	Type de codon dégénéré
N-WASP	Protéine du syndrome de Wiscott-Aldrich neural
OC_r	Ordre de contact relatif
PCR	Réaction en chaîne de la polymérase
PDB	« Protein Data Bank » (banques de données)
PDZ	Domaine d'homologie à « PSD-95 large discs/ZO-1 »
PFAM	« Protein families » (banque de données)
PH	Pleckstrin Homology
Protéine-G	Domaine B1 liant les immunoglobulines de la protéine-G
Protéine-L	Domaine B1 liant les immunoglobulines de la protéine-L
PSI	« Protein Structure Initiative »
PTP1B	Phospatase des tyrosines protéiques 1B
R_G	Rayon de giration
Rap1A	Petite GTPase Rap1A
RMN	Résonance magnétique nucléaire
RT	Produit de la constante des gaz parfaits et de la température
SCOP	« Structural classification of proteins » (banque de données)
SH2	Src2 homology
SH3	Src3 homology
SMART	« Single module architecture reasearch tool » (banque de données)
ts	Type sauvage
U	État dénaturé
U→F	Transition état dénaturé vers état natif
U→‡	Transition état dénaturé vers état de transition
UV	Ultraviolet

‡	État de transition
ψ	Angle psi

A mes parents, Francene Campbell et Pierre Valois

"I can't get no satisfaction" – Mick Jagger et Keith Richards

Finalement au DLR de Raf,
"SNTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLL
HEHKGKKARLDWNTDAASLIGEELQVDFLD"

Utilisé avec l'autorisation de l'artiste (© et tous droits réservés à Catherine Campbell-Valois; cathcava@yahoo.ca)

Remerciements

Voilà, j'imagine que ça veut dire que c'est presque terminé! Je dois admettre que je n'en suis pas fâché! C'est beaucoup de travail que vous avez entre les mains et je vais commencer mes remerciements en utilisant un cliché, cette section est probablement la plus importante de ma thèse.

Incrédulité, c'est ce qui décrit le mieux a posteriori l'état d'esprit avec lequel j'ai entrepris mes études graduées. Quand j'étais enfant, j'avais appris qu'une maîtrise se terminait en deux ans et un doctorat en trois. Le pas entre ses fables et la réalité était tout autre! Je voudrais d'abord remercier mes collègues du département et du laboratoire qui au cours de l'été 1996 d'abord, durant mon stage, et ensuite pendant l'année 1997, qui a vu mon accession aux études graduées, m'ont permis de faire le saut de manière fluide. J'aimerais remercier en particulier Dr Joëlle Pelletier avec qui j'ai collaboré pour une de ses publications dans notre laboratoire, ce qui m'a permis d'obtenir des bourses d'études que je n'aurais sans doute pas obtenu sans cette note positive a mon curriculum vitæ.

Vivifiant, est le meilleur adjectif pour définir l'effet de l'enseignement qui m'a été transmis dans le cadre de mes études graduées a eu sur ma vision scientifique. Le cours donné à l'IRCM et étalé sur une année a été particulièrement intéressant et m'a ouvert l'horizon. Je remercie tous les conférenciers qui y ont participé. Je veux souligner le dynamisme du département de biologie moléculaire qui par le biais de sa journée porte ouverte contribue à assurer la cohésion et l'intérêt pour son programme d'études. J'en profite pour remercier la secrétaire des programmes Mme Viviane Jodoin dont l'aide a été fort apprécié et Dr Thrang Hoang, la directrice, qui a eu la gentillesse et le respect de s'intéresser à l'avancement de mon projet et de mon cheminement scientifique au fil de nos quelques rencontres. Aussi, je veux remercier les membres permanents de mon comité de thèse qui m'ont suivie depuis le début et qui m'ont bien conseillé : les Drs Franz Lang et Luc Desgroseillers.

Ensuite, je tiens à souligner la contribution de tous les employés du département au climat de travail qu'ils contribuent à instaurer. Je remercie personnellement André Racicot, Daniel Chevrier, Jacques Gauvreau, Ernest Lorange, Mme Ginette, Manon Moreau, Élaine Meunier, Sylvie Beauchemin et Denise Lessard qui m'ont aidé et facilité la vie à divers niveaux. Je tiens à remercier en particulier l'apport de Mme Mireille Fyfe à mes études dans le cadre desquelles elle a réalisé de nombreuses réactions de séquençage et la migration de nombreux échantillons. Pour les mêmes raisons et le support technique je remercie David Roquis et Pierre Lepage de la plateforme de séquençage de Génome Québec et de l'Université McGill. Je remercie les Dr Jeffrey Keillor Luc Desgroseillers ainsi que les membres de leur laboratoire, respectivement pour l'accès à l'appareil de mixage rapide/fluorimètre et à l'électroporateur.

La liberté et le respect des idées et de l'indépendance de la pensée d'autrui sont des beaux cadeaux à offrir à un étudiant qui tente de grandir. Enthousiasme, valorisation de l'imagination, toutes ces choses-là je les ai reçu durant ma formation de la part de mon directeur de thèse. Je remercie donc le Dr Stephen Michnick pour m'avoir fait bénéficier avant toute chose de ces grandes valeurs-là.

Quelque soit notre motivation, la poursuite d'études graduées en sciences fondamentales est à tout le moins faciliter par l'octroi de bourses qui nous permettent de vivre décemment et de nous concentrer sur notre travail d'étudiant. Pour cette raison, je veux souligner la contribution à mon degré de sérénité du FCAR, de l'IRSC, du programme de biologie moléculaire et de la FES par l'intermédiaire de la fondation J.A. de Sève, de même qu'à mon patron de thèse qui a su délier les cordons de sa bourse pour compléter mon salaire avant le début et après la fin de mes bourses. Je remercie aussi le département de biochimie et la famille Noël.

Une belle aventure tout de même. Je me rappellerai de mes amis collègues du département qui m'ont accompagné sur ce chemin et avec qui j'ai eu la chance de partager des bons moments, spécialement Sylvain Huard, Jean Buteau, Alexandre Benoît, Stéphane Angers et Ali Salahpour.

Élève, on le reste souvent en toute circonstance, alors même que c'est nous qui sommes dans la position de l'enseignant. En ce sens, Jérôme Dupras et Véronique Montplaisir m'ont beaucoup appris et je les en remercie.

Beaucoup de temps passé dans un laboratoire, ça signifie aussi des liens qui se tissent. J'ai dans mes souvenirs des dizaines de moments et de discussions mémorables que j'ai eu avec des collègues dont certains sont devenus par la force des choses des amis très chers. Je pense entre autre aux discussions à bâtons rompus, et un peu poudreuses parfois, de la belle époque. Donc à Alexis Vallée-Belisle, Dimitri Sans, Martin Primeau, J-F Turcotte, Geoff Denis, André Galarneau, Galia Ghaddar ainsi qu'à Guy Tremblay, chapeau bas. Pour d'autres types de discussion dont la teneur ne saurait être révélée ici, merci à mes amis Luciano Vidali et Hugo Lavoie.

Expérimentalement parlant, le collègue le plus important est le plus proche. Je remercie mes partenaires de paillasse J-M Brondello, Stéphanie Aquin et Martin Primeau pour leur civilité et des discussions. Pareillement à Philippe Nissaire.

Comment obtenir des produits sans se fatiguer? C'est facile, demander à Ingrid Remy, Annie Montmarquette, Nathalie Bourassa et J-F Paradis. Je les remercie chaleureusement pour leur aide d'appoint précieuse, leur célérité à commander ma dernière extravagance en toute urgence et plus encore! Finalement, je voudrais souligner l'aide, le support et les échanges scientifiques ou non que j'ai pu avoir avec les membres passés et présents du laboratoire que je n'ai pas nommés ci-dessus. Plusieurs membres du laboratoire

de l'édition printemps 2006 se sont libérés pour ma soutenance et je les remercie de cette marque de respect.

L'écriture et la préparation d'un manuscrit scientifique demande beaucoup d'efforts. Je remercie mon co-auteur, collègue et ami Kirill Tarassov pour sa contribution et son efficacité. Ça aura été extrêmement agréable et facile de travailler avec lui.

Immanquablement, la communication de résultats et d'idées scientifiques pour chaque article que nous avons préparé a nécessité des discussions, des débats et du polissage patient pour rendre réellement intelligible ce que je voulais dire, en fait ce que nous voulions dire, bien que les deux n'aient pas toujours concordés. Je remercie mon directeur de thèse le Dr Stephen Michnick pour les nombreuses choses qu'il m'a apprises sur la communication scientifique. Je lui suis bien sûr reconnaissant et redevable pour le temps et la patience dont il a fait preuve à mon égard.

Bien plus qu'un cadre académique ma formation universitaire a façonné ma vie actuelle de manière profonde. D'abord et avant tout, il y a des liens d'amitié qui désormais transcendent les relations de travail. Je remercie ma bonne étoile d'avoir mis sur mon chemin Stéphanie Pontier, Dimitri Sans, Martin Primeau, Jérôme Dupras et J-F Turcotte. À travers ces années, il y a une relation d'amitié constante et forte qui a duré. Dès mon premier stage au laboratoire, il était là. Nous avons traversé ensembles beaucoup des épreuves et remises en question typiques du cheminement académique. On a aussi fait beaucoup d'erreurs, mais c'était super pour ma part de les faire avec lui. Merci à Alexis Vallée-Belisle pour son amitié et pour avoir agrémenté le tout de son humour et de sa vision de la vie si personnelle.

Finalement, je voudrais remercier mes proches. D'abord mes parents, Francene Campbell et Pierre Valois, pour leur support à tous les niveaux, leur attention, leur amour et pour l'éducation et les valeurs qu'ils m'ont transmises, en particulier la rigueur, la persévérance et l'ouverture d'esprit que je tente d'honorer du mieux que je le peux. Mon frère et ma sœur pour leur support. Catherine, en particulier pour avoir accepté de prêter ses talents à la conception de la peinture qui précède. Ma grand-mère Marie-Paule Poulin, Alberte et Léopold Gravel ainsi que Paul Faure ont aussi joué des rôles très importants dans ma vie d'enfant et de jeune adulte. Je leur en suis reconnaissant et les assure de mon affection. Je voudrais remercier tous les autres membres de ma famille qui m'ont aidé et appuyé au cours de ces années et aussi ceux qui ont pu assister a ma soutenance (Geoffroy Campbell-Valois et Mélissa, Robert Valois et Jocelyne, Julie Rouleau, Henriette Poulin, John Campbell Sr ainsi que mes parents bien sûr, et des amis de la famille Jacques Laberge et Michel Béchard). Enfin, à Arthur et Stéphanie pour notre vie ensemble!

Avant-propos et mise en contexte

Les protéines sont essentielles à la vie tel que nous la connaissons, car elles constituent une part essentielle d'abord à la machinerie nécessaire à la réplication du matériel génétique et à sa transmission, mais aussi dans à peu près toutes les fonctions cellulaires qu'ils s'agissent du métabolisme, de la réponse aux hormones ou du remaniement et du maintien de la morphologie cellulaire. Or, l'activité biologique d'une protéine dépend de l'obtention en un temps raisonnable de la structure native qui doit être débusquée parmi toutes les conformations possibles qu'une chaîne polypeptidique peut adopter. En outre, les liens entre les propriétés de la structure et la fonction biologique des protéines restent à intégrer à bien des égards à notre vision du fonctionnement de la cellule et des organismes.

La constatation qu'un nombre grandissant de pathologies est lié à des phénomènes de défauts du repliement tel que ceux causés par les prions (par exemple la maladie spongiforme bovine, couramment dite maladie de la vache folle), ceux causant la maladie d'Alzheimer ainsi qu'une variété de maladies provoquées par l'agrégation de protéines (extension poly-glutamine ou poly-alanine), mais aussi des formes oncogéniques de p53 ou défectueuses du récepteur à la vasopressine pouvant mener à des formes rares de diabète néphrogénique, ont étendu à des cercles plus larges de la communauté scientifique et du public l'intérêt pour la structure des protéines et pour le processus de repliement en tant que tel. Dans le même ordre d'idées, il y a de plus en plus d'exemples de l'implication des processus de repliement et de dépliement dans la régulation des voies de signalisation, des interactions protéine-protéine et le destin des molécules protéiques (*1*).

Les études de repliement de protéine sont basées sur des expériences réalisées *in vitro* dont la conformité par rapport aux processus physiologiques est validée par la constatation qu'aucun facteur cellulaire tel que le ribosome ou les chaperonnes n'est essentiel au repliement de la vaste majorité des protéines et que donc l'information structurale est une propriété intrinsèque de la séquence. Les grands objectifs du champ d'études du repliement peuvent être énoncés brièvement de la manière suivante :

1. **La compréhension du mécanisme de repliement des protéines.** Quelles sont les étapes du processus de repliement? Quels en sont les éléments déterminants, en particulier au niveau de la séquence polypeptidique? En particulier, y a t il un mécanisme de repliement commun aux protéines adoptant la même topologie structurale, mais qui ne démontrent pas nécessairement d'homologie de séquence significative? Ultimement, il faudrait proposer une théorie unificatrice du mécanisme de repliement; c'est-à-dire un ensemble d'équations et de règles permettant d'expliquer le repliement de tous les polypeptides.

2. **La prédiction de la structure et de la fonction d'une protéine à partir uniquement de sa séquence d'acides aminés et partant de là, la capacité de générer à volonté des protéines à la structure et à la fonction déterminée.** Des outils informatiques capables de prévoir avec précision la structure, la fonction cellulaire et l'effet d'une mutation donnée dans un gène seraient des apports puissants à la recherche en sciences fondamentales et appliquées. Les applications industrielles en pharmaceutique, dans l'agro-alimentaire ou bien en environnement pourraient bénéficier de façon accrue de l'utilisation de protéines, en particulier d'enzymes, qui pourraient remplacer les procédés chimiques classiques de manière efficace et pour de moindres coûts financiers et écologiques.

L'obtention de la séquence de plusieurs organismes a permis dans les dernières années l'apparition de plusieurs projets d'envergure s'insérant dans un champ d'étude émergeant, la génomique structurale. Son objectif est d'identifier les protéines sans homologues et de tenter d'en déterminer la structure, dans l'optique de découvrir de nouvelles topologies structurales. Par ailleurs, l'efficacité d'une telle approche est limitée par la dégénérescence de la séquence ce qui se traduit par un haut niveau d'homologie structurale. Les outils permettant de trouver des homologues à un nouveau gène, dont le plus connu est « Psi-blast », permettent entre autre de choisir les cibles qui présentent les meilleures probabilités d'adopter une topologie inédite. Même en utilisant une telle procédure, la proportion de structures présentant des nouvelles topologies représente seulement environ 10% de celles qui ont été résolues par le consortium intégré de génomique structurale (voir leur site Internet à http://www.jcsg.org/). La multiplication des gènes à la fonction inconnue qui émergent des divers projets de séquençage de génome rend plus nécessaire que jamais auparavant le développement de méthodes bioinformatiques qui permettraient de prévoir la structure et la fonction d'une protéine à partir de sa structure primaire uniquement.

Durant mes études doctorales des progrès spectaculaires ont été réalisés en ce domaine, particulièrement en ce qui concerne la prédiction et le design de structure *de novo*. Remarquablement, l'algorithme ayant remporté le plus de succès repose sur la modélisation de la structure de courts segments de séquence du polypeptide d'intérêt à une banque de segments de taille similaire issus de structures de protéines résolues expérimentalement et à l'optimisation mathématique des contacts à longue distance entre ces segments (*2;3*). L'application de cette méthode a permis inversement le design de protéines à la structure extrêmement stable en se servant alternativement de protéines naturelles comme plan ou bien d'une topologie modèle non-répertoriée jusqu'à ce jour dans le répertoire naturel (*4;5*). Un grand défi à relever est de combiner ces nouveaux outils de design structural avec des méthodes permettant d'y intégrer des activités enzymatiques nouvelles ou connues pour les diverses applications souhaitées (*6*), car jusqu'à maintenant le design de structure *de novo* n'a pas été opéré en conjonction avec celui d'activité enzymatique.

Attardons-nous maintenant aux protéines qui ont strictement en commun des similitudes structurales. L'homologie de séquence entre de telles protéines est si faible qu'il est impossible de l'identifier sans le guidage de la structure, ce qui indique leur très grande distance évolutive et fonctionnelle. Cette observation est généralisable puisque pour quelques centaines de milliers de protéines, il y aurait seulement entre 1000 à 10 000 topologies structurales distinctes dont seulement quelques 400 représenteraient 80 % des domaines protéiques (*7;8*). Ces observations réitèrent la nature colossale du défi immense que constitue la compréhension du mécanisme de repliement, plus spécifiquement la découverte des éléments d'informations communs encodées dans la séquence de protéines à la topologie similaire et qui guident, par conséquent, la formation de leur structure native. C'est en particulier à cette question que ma thèse intitulée, « Application des librairies de codons dégénérés à l'étude du mécanisme de repliement et de la stabilisation de la structure du domaine liant *ras* de Raf », a tenté d'apporter une contribution. Dans un premier temps nous avons réalisé une perturbation massive de la séquence du DLR de Raf par l'utilisation et la sélection d'une collection aléatoire de mutants. L'approche expérimentale détaillée et

les résultats de cette étude sont exposés dans les deux premiers articles. Dans le cadre des 2 études subséquentes nous avons voulu vérifier le rôle dans la formation ou la stabilisation de la structure des résidus conservés dans l'expérience de perturbation de séquence (voir le **Chapitre 2 : Résultats**).

La section **Introduction** qui suit immédiatement est divisée en plusieurs sections suivant la nature des espèces et des changements encourus par les polypeptides lors de la réaction de repliement. Ainsi, certaines sections peuvent être ennuyeuses à lire pour le spécialiste et si, le lecteur n'a pas besoin de rappel sur ce sujet, je me permet de lui suggérer de passer outre, quitte à ce qu'il se serve de ces passages selon ses besoins. J'ai tenté d'intégrer le plus souvent possible et au mieux de mes connaissances des perspectives historiques remontant au tout début des champs d'études concernés et en axant la rédaction mon texte principalement sur les idées éclairantes qui y ont modelé la recherche. Finalement, je fais un rappel dans le **Chapitre 1** des équations et des principes fondamentaux qui ont été utilisés pour les analyses des expériences thermodynamiques et cinétiques discutées dans mes travaux.

Je conclus ma thèse en présentant des perspectives personnelles sur les résultats obtenus et sur les questions centrales d'avenir concernant plus généralement le repliement de protéine et la biologie structurale. Là-dessus, je vous souhaite une bonne lecture, en espérant que vous ne vous ennuyiez pas trop.

Introduction

Il est connu depuis le début du XX^{ème} siècle que les protéines sont composées d'acides aminés, mais c'est seulement en 1952 que la séquence d'une protéine sera déterminée pour la première fois (*9*). Ces études et d'autres sur la protéolyse de plusieurs protéines comme le lysozyme permirent d'établir la composition en acides aminés des peptides qui composent les protéines (réviser dans (*10*)). Les travaux d'Anfinsen et coll. sur la RNAse pancréatique bovine et sur la DNAse staphylococcique ont mené à la démonstration que la séquence primaire contient toute l'information nécessaire à la formation et au maintien de la structure de la protéine et qu'il n'y a pas de cofacteurs biologiques nécessaires à l'obtention de leur structure native. Ces travaux constituent la pierre d'assise de notre vision contemporaine du repliement des protéines. Ils énoncent plusieurs des questions scientifiques centrales de ce champ d'études en y établissant les deux pôles principaux de recherches (*11*):

1. L'étude de la réaction de repliement dans le but d'établir un mécanisme général.
2. La recherche des déterminants de séquence nécessaires à la formation et à la stabilisation des structures protéiques.
3. L'objectif étant à terme la prédiction de la structure d'une protéine à partir de sa structure primaire et le design de protéines à la structure et à la fonction prédéterminées.

Que l'on parle de la réaction, du mécanisme ou du processus de repliement, il est habituellement question du même phénomène, c'est-à-dire un ensemble d'événements (i.e. transitions, réactions etc.) menant aux remaniements des interactions chimiques – la plupart du temps non-covalentes, faibles et donc aisément réversibles – établies entre les acides aminés d'une protéine et qui lui permettent de passer d'une condition structurale dite dépliée à une dite native et vice-versa pour la réaction de dépliement. L'état natif est la conformation d'un polypeptide la plus stable en conditions quasi-physiologiques *in vitro* et sa structure est habituellement révélée par la cristallographie aux rayons X ou la résonance magnétique nucléaire (RMN). La structure native est habituellement compacte dans le cas

des protéines globulaires[1] et plusieurs des chaînes latérales des acides aminés particulièrement dans le cœur hydrophobe des protéines sont inaccessibles au solvant. Elle est habituellement considérée comme la forme biologiquement active des protéines. D'autre part, l'état déplié est pour la plupart des protéines globulaires impossible à observer en condition normale, car trop instable. Un état considéré comme équivalent est obtenu expérimentalement par des perturbations physiques ou chimiques précises comme l'ajout de dénaturant qui ont pour effet de stabiliser cette espèce structurale par rapport à l'état natif. L'état déplié obtenu de cette façon est nommé état dénaturé[2], et le processus en tant que tel, c'est la dénaturation. La dénaturation provoque un changement majeur de la conformation de la chaîne polypeptidique qui s'accompagne selon le modèle classique par la rupture de la plupart des interactions non-covalentes de la structure native et se traduit donc par le bouleversement de nombreuses propriétés physiques de la molécule. Par exemple, l'expansion de la chaîne polypeptidique résultant de la dénaturation est détectable par l'augmentation du rayon de giration (R_G)[3] et de la variabilité de ce dernier (**Figure 1**), ce qui indique la diversité structurale propre à cet état. Par ailleurs, la dénaturation de la structure d'une protéine ne doit pas être confondue à l'inactivation de son activité biologique (i.e. enzymatique, liaison etc.) qui peut habituellement être accomplie dans des conditions beaucoup moins rigoureuses comme une légère variation de pH ou de température. La capacité d'obtenir *in vitro* des conditions dénaturantes pour la plupart des protéines est bien entendu centrale à notre capacité d'étudier le processus de repliement.

---

[1] Le terme protéine globulaire dans le sens normal du terme, i.e. des structures qui comprennent un coeur hydrophobe enfoui. Par opposition aux protéines fibreuses comme le collagène par exemple (*13*).

[2] La distinction entre l'état déplié et dénaturé est nécessaire, parce qu'ils ne sont pas obtenus dans les mêmes conditions de solution. Cela veut dire que certaines de leurs charactéristiques pourraient varier et qu'en particulier, celles de l'état dénaturé sont dépendantes du mode de dénaturation choisi.

[3] Le R_G est la racine-moyenne-carré (« root-men-square ») des distances moyennes de toutes les parties d'une molécule par rapport à son centre de masse. Cette valeur peut être obtenue par mesure directe dans l'analyse de la dépendance angulaire de la dispersion de la lumière dans une solution de macromolécules. Ainsi le R_G est plus grand dans l'état dénaturé que dans l'état natif pour les protéines globulaires et est généralement beaucoup plus grand dans l'état natif des protéines fibreuses par rapport aux protéines globulaires. La comparaison entre le poids moléculaire et le coefficient de sédimentation donne une information comparable à R_G (*315*).

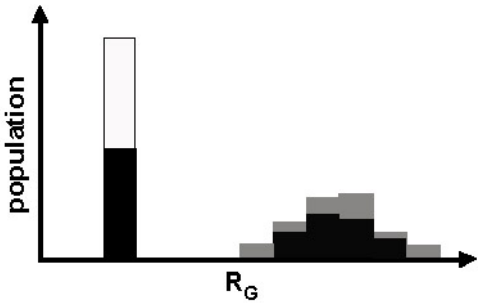


Figure 1. L'état natif et l'état dénaturé : une comparaison superficielle de leur conformation.
La variation théorique de la population d'un échantillon d'une protéine globulaire à l'équilibre évaluée en fonction du R_G obtenu à trois conditions expérimentales [(distribution de la population de l'échantillon de protéine en des conditions où la population est : 100% native (□); 50% native et 50% dénaturée (■); 100% dénaturée (■)]. Notez que l'état natif montre un R_G inférieur à l'état dénaturé et la diversité des conformations de l'état dénaturé.

Les termes de renaturation et de dénaturation font référence aux expériences utilisées par l'expérimentateur pour observer les phénomènes de repliement et de dépliement. Ils sont étudiés respectivement en suivant le plus souvent indirectement les changements structuraux qui interviennent dans un polypeptide transféré de conditions dénaturantes en conditions natives et vice versa. Ces phénomènes peuvent être étudiés dans des conditions dites à l'équilibre ou cinétiques. A l'instar des réactions chimiques plus simples, l'étude de la réaction de repliement se concentre sur trois états principaux et un état facultatif. Il s'agit respectivement de l'état dénaturé, l'état de transition et l'état natif ainsi que des états intermédiaires. Je vais procéder à rebours de la réaction de repliement en présentant dans la prochaine section les éléments de base essentiels à la compréhension de la structure des protéines, en particulier en ce qui a trait aux déterminants de la séquence dans le contexte des protéines adoptant la même topologie structurale. Par la suite, je discuterai de l'état de transition et des intermédiaires qui sont le sujet principal des études de repliement, car elles en déterminent la vitesse et donc le mécanisme. Finalement, je terminerai en présentant nos connaissances actuelles sur les structures résiduelles présentes dans l'état dénaturé de quelques protéines.

Structure native des protéines

Structure chimique des acides aminés naturels

Les acides aminés sont des molécules organiques simples structuralement arrangées autour d'un centre chiral le carbone-α (C_α). Celui-ci est coordonné à un groupement amine et à un groupement carboxylique ainsi qu'à un hydrogène et à un groupement de nature variable spécifique à chaque acide aminé et qui est nommé chaîne latérale. Il y a 20 types fondamentaux d'acides aminés qui sont directement encryptés par le code génétique et donc qui sont utilisés pour former les protéines chez l'immense majorité des êtres vivants (**Figure 2**). Ceux-ci sont tous de conformation L, excepté la glycine qui n'a pas de centre chiral, car deux hydrogènes sont coordonnées à son C_α. Les acides aminés peuvent aussi être groupés en diverses classes physico-chimiques (**Tableau AI**). Outre les 20 acides aminés fondamentaux, des dérivés de ceux-ci peuvent être obtenus via des modifications post-traductionnelles des protéines et d'autres types d'acides aminés sont utilisés comme intermédiaire dans certaines voies métaboliques (*12*).

Nature du lien peptidique

Les acides aminés successifs au sein d'une protéine sont unis par la formation de liens peptidiques entre le groupement carbonyle du premier acide aminé et le groupement amine du second acide aminé et ainsi de suite. La succession du groupement amine, du C_α et du groupement carbonyle de chaque acide aminé successif de la séquence forme la chaîne principale ou polypeptidique ou squelette carboné. Le lien peptidique est caractérisé par une rigidité semblable à un lien double dû à la résonance induite par la délocalisation partielle des électrons libres du groupement amine vers le lien carbonyle (**Figure 3**). Cette rigidité induit une conformation dans laquelle les groupements amine et carbonyle d'un lien peptidique se retrouvent dans le même plan, permettant en théorie l'adoption des conformations cis ou trans.

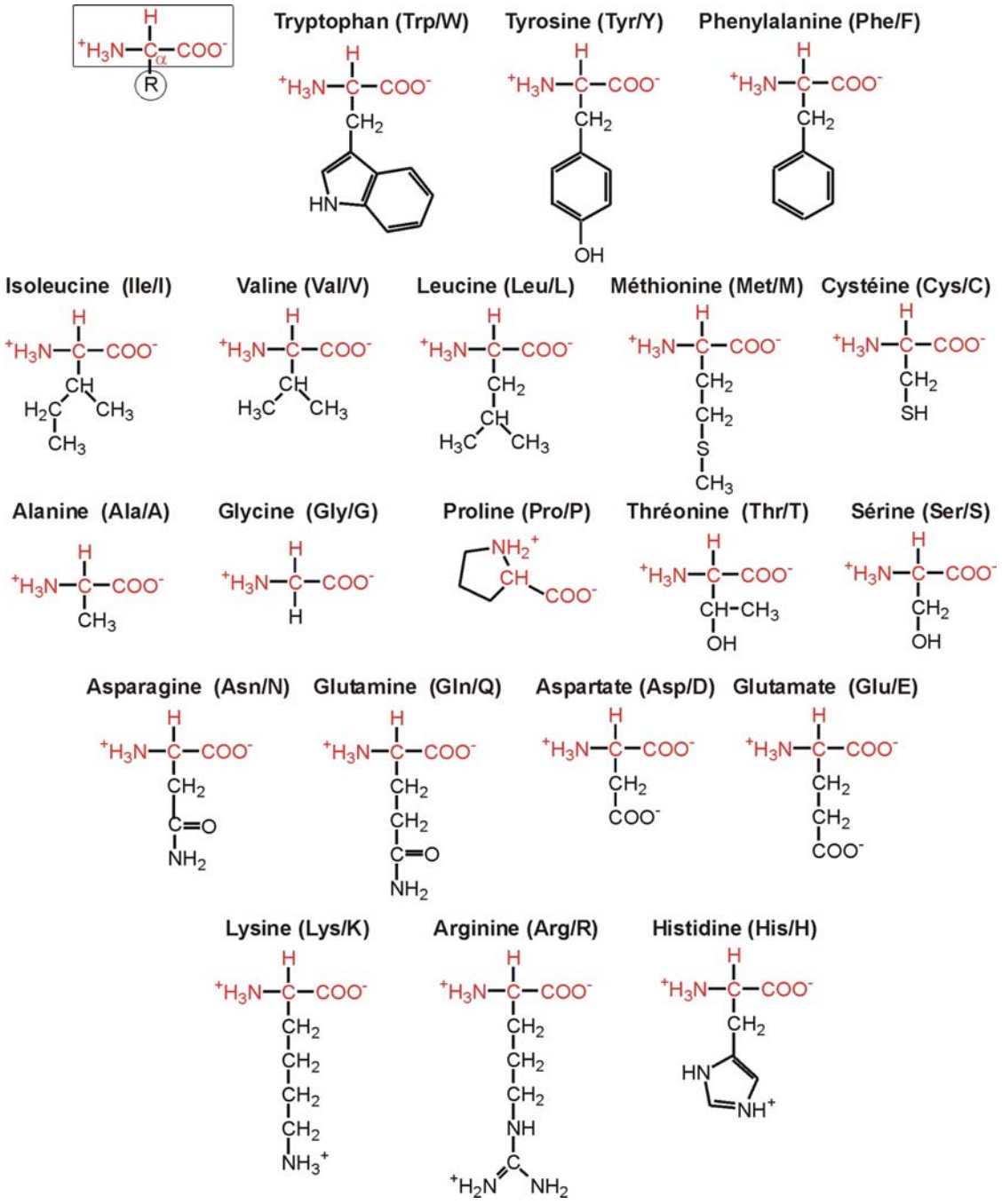


Figure 2. Structure chimique des 20 acides aminés naturels du code génétique.
Les acides aminés sont regroupés en classe (**Tableau AI**). Les atomes de la chaîne principale et de la chaîne latérale apparaissent respectivement en rouge et en noir.

Cependant, les groupes peptidiques adoptent dans la vaste majorité des cas la conformation trans, laquelle présente sur les côtés opposés d'un lien peptidique les deux C_α consécutifs, car celle-ci est plus favorable à l'optimisation des contraintes stériques. L'importance numérique des liens cis est limitée, et dans la structure native ils sont rencontrés à des liens peptidiques impliquant principalement la fonction imide des prolines. Je reviendrai plus tard à l'impact de l'isomérisation des prolines sur l'étude de la réaction de repliement. Aussi, les angles dihédraux, dits ϕ (phi) ψ (psi) désignent respectivement l'angle des liens N-C_α et C_α-C avec le plan formé par le lien peptidique.

Le diagramme de Ramachandran décrit l'ensemble des combinaisons d'angles ϕ et ψ permis par les contraintes stériques. Par ailleurs, certaines combinaisons d'angles sont considérées improbables ou impossibles, car elles mèneraient à des encombrements stériques entre les chaînes latérales et les groupements de la chaîne principale. Les combinaisons d'angles ϕ et ψ de chaque acide aminé d'un polypeptide sont suffisantes pour déterminer la conformation de son squelette carboné. Ces conformations particulières sont stabilisées par divers types d'interactions non-covalentes, entre autre de type ponts hydrogène. En effet, les atomes d'oxygène de groupements carbonyle et hydrogène de groupements amine de la chaîne principale sont respectivement accepteur et donneur dans la formation d'interactions régulières de type ponts hydrogène et en outre, la régularité avec laquelle celles-ci sont formées le long de la chaîne polypeptidique définit la structure secondaire des protéines (*13*) (**Figure 3**).

Connaissances de base sur la structure native des protéines : structure primaire, secondaire, tertiaire et quaternaire

Selon que les ponts hydrogène formés impliquent des groupes donneurs et accepteurs rapprochés ou distants, ils forment deux types de structures secondaires qui sont respectivement dénommées hélice-α et brin-β. Dans le premier cas, les ponts hydrogène se forment entre les groupements carbonyle en n et amine en n+4, forçant ainsi la chaîne à

tourner sur elle-même. Les hélices-α composées d'acides aminés naturels ont un enroulement de type droit[4] (**Figure 3**). Cette version canonique est parfois altérée par la présence de segment comportant un réseau de ponts hydrogène atypique en particulier à l'extrémité carboxylique. Il s'agit souvent de segment d'hélice 3_{10} qui comporte trois résidus par tour d'hélice au lieu des 3,6 de l'hélice-α. De courtes hélices 3_{10} sont aussi rencontrées hors du contexte de l'hélice-α. Les autres types d'hélices sont très rares au sein des structures protéiques. D'autre part, les résidus dans un brin-β doivent nécessairement former des ponts hydrogène avec ceux d'un autre brin-β, de sorte qu'ils apparaissent couramment sous la forme d'un feuillet plissé et tordu. L'aspect plissé provient de la périodicité d'ordre deux du brin-β qui fait en sorte que les C_α alternent entre les deux faces, soit tour à tour au-dessus et au-dessous du plan du feuillet. Les feuillets sont habituellement tordus afin d'optimiser l'empaquetage des résidus hydrophobes. Tout comme pour l'hélice-α, cette torsion est droite et résulterait de la conformation L des acides aminés. Par ailleurs, l'arrangement des ponts hydrogène peut prendre deux formes à l'intérieur d'un feuillet selon que le déroulement de la chaîne polypeptidique des brins qui le composent soit parallèle ou anti-parallèle, des épithètes qui servent donc à qualifier les feuillets rencontrés. Quelques 20 % des brins sont dits mixtes, parce qu'ils forment à la fois des interactions parallèles et anti-parallèles. De façon général, il est admis que la conformation des protéines tend à maximiser le nombre de ponts hydrogène afin de favoriser la création d'un environnement favorable à l'enfouissement des chaînes latérales hydrophobes (*13*) (voir la section **Stabilisation de la structure native : un rôle prépondérant pour l'effet hydrophobe**). L'inversion de la direction de la chaîne polypeptidique s'effectue par l'entremise de tours-β. Ce type de motif structural est très court et stabilisé d'ordinaire par au moins un pont hydrogène (**Figure 3**). Il en existe plusieurs types (i.e. nommément les types I-VIII) ayant des particularités de séquence et de conformations distinctes (*14*). Des acides aminés spécifiques se retrouvent plus fréquemment dans des segments de séquence

[4] Cela signifie que la chaîne s'enroule dans le sens donné par la rotation des doigts de la main droite autour de l'axe de l'hélice lorsque le bout du pouce est orienté de manière à indiquer le déroulement de la chaîne de l'amino vers le carboxy-terminal.

qui adopte l'un ou l'autre de ces types de tour-β. En effet, on retrouve fréquemment dans un tour-β une glycine, une proline ou même les deux à la fois. D'autres résidus flanquant les positions où apparaissent les glycines ou les prolines sont importantes pour la stabilisation des tours-β. Les tables de propension des tours-β permettent de prédire la localisation de ces éléments et indiquent les combinaisons d'acides aminés favorisant leur formation (*14;15*). Les tours-β se retrouvent habituellement en surface des protéines, par exemple entre deux brins-β antiparallèles dans un motif qui est nommé épingle à cheveux (« hairpin ») ou à la fin ou au début d'une hélice-α. En surface, d'autres segments à la séquence hydrophile et à la structure irrégulière que l'on nomme simplement boucles sont aussi rencontrés. Une partie très importante de la séquence des protéines globulaires se retrouve dans ce genre d'élément. Les boucles ne contiennent habituellement pas un réseau de ponts hydrogène régulier, ce qui les rend plus variables structuralement et dynamiquement que les autres éléments structuraux.

La séquence des acides aminés d'une protéine est aussi désignée par le terme "structure primaire". La disposition des structures secondaires les unes par rapport aux autres pour un polypeptide donné dans l'espace tridimensionnelle est dénommée structure tertiaire (**Figure 4**).

Figure 3. Géométrie de la chaîne polypeptidique, diagramme de Ramachandran et présentation des diverses structures secondaires en prenant comme exemple le DLR de Raf.
La figure se trouve à la page suivante. **A**, La géométrie du squelette carboné. Le lien peptidique a un caractère de double liaison partielle, ce qui introduit de la rigidité dans le plan amide. Les chiffres indiquent la longueur moyenne de chaque lien. Les angles ϕ et ψ sont aussi indiqués. **B**, Présentation de la structure secondaire de type hélicoïdal : structures schématisées de l'hélice-3_{10}, -α et -π avec la représentation du plan amide et la périodicité des ponts hydrogène. **C**, Diagramme de Ramachandran. Les combinaisons d'angles ϕ et ψ en ordre croissant de permissivité sont colorées en jaune, orange et rouge. Les régions du diagramme correspondant aux divers types de structures secondaires sont identifiées en fonction du code suivant : brin (β,b), hélice droite (α,a) et hélice gauche (L,l). Les points à l'intérieur du diagramme indique la valeur des angles ϕ et ψ pour les résidus 55-132 du DLR de Raf (selon le code PDB 1RFA). **D**, Présentation de la structure secondaire de type étendu (brin-β) : brin anti-parallèle, brin parallèle, motif épingle à cheveux entre deux brins antiparallèles et arrangement des ponts hydrogène pour deux types de tour-β fréquents (I et II). **E**, La structure tertiaire du DLR de Raf (code PDB 1RFA) ainsi que les divers éléments de structures secondaires qui y sont observables : i) hélice-α (α1), ii) épingle à cheveux, iii) tour-β1, iv) brins parallèles (β1 et β2) et anti-parallèles (β1 et β5). Notez que les atomes sont identifiables par le code de couleur suivant : C (noir), N (bleu), O (rouge) et la chaîne latérale (mauve). Les C_α sont numérotés et les ponts hydrogène sont indiqués en lignes pointillées (simple ou double).

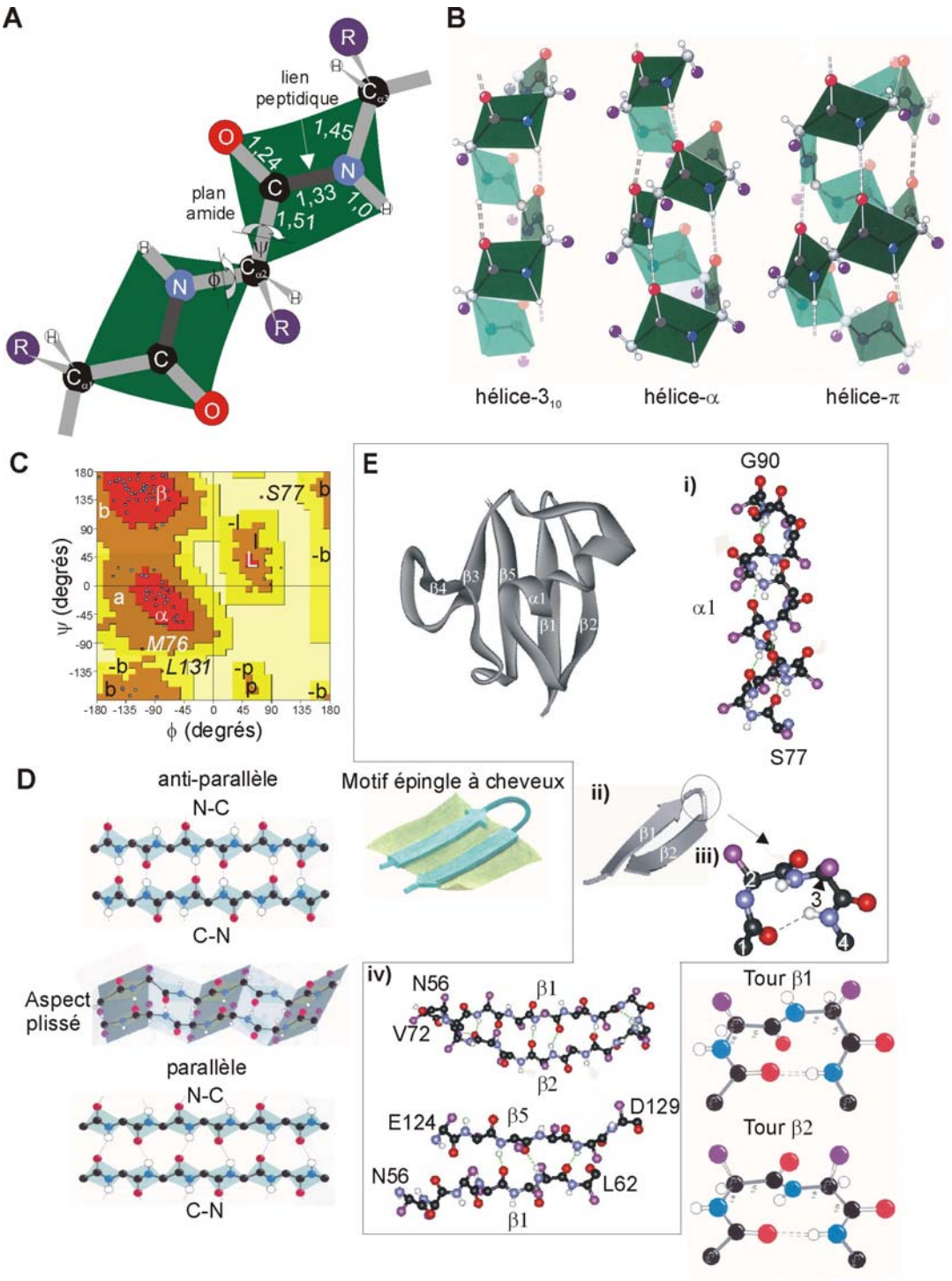


Figure 3. La légende est à la page précédente.

Certaines protéines ou sous-unités peuvent s'associer de façon non-covalente ou alternativement par la formation de ponts dissulfures pour former des oligomères obligatoire ou transitoire. Ce type de structure est qualifié de quaternaire. Dans sa première déclinaison, la formation de la structure native (i.e. sa réaction de repliement en tant que tel) d'une protéine, comme dans l'exemple classique de l'enzyme β-galactosidase, est couplée à l'association de ses sous-unités identiques. Cette définition de la structure quaternaire est la plus largement acceptée. D'autre part les complexes protéiques tels que ceux impliqués dans les voies de signalisation ou métaboliques sont transitoires et résultent la plupart du temps de l'association de polypeptides dont la structure est obtenue indépendamment de l'oligomérisation et donc à ce titre représentent la seconde déclinaison de complexe quaternaire observable. L'interaction du DLR de Raf avec la petite enzyme hydrolysant le GTP (GTPase) *ras* est un exemple pertinent de cette dernière catégorie (**Figure 4**). En ce qui concerne les protéines dont le repliement a été étudié, la structure tertiaire constitue le plus souvent la structure native. Les schémas de la **Figure 4** sont représentatifs de la représentation la plus usitée de la structure des protéines, en ce sens qu'elle se limite à la conformation du squelette carboné principal et qu'elle exclut donc les détails de la disposition des chaînes latérales pour des raisons principalement de simplicité. Ainsi, lorsque l'on dit des protéines à la séquence très divergente qu'elles adoptent des structures similaires, il est question de ressemblance dans la disposition générale de la chaîne polypeptidique. Par contre, dans le cas où les conformations des chaînes latérales seraient explicitement prisent en compte, de nombreuses distinctions dans le détail des conformations structurales apparaîtraient suivant la faiblesse du taux d'identité de séquence entre les analogues structuraux comparés.

Contrairement, à ce que pourrait laisser supposer la hiérarchie apparente entre la structure secondaire, tertiaire et quaternaire, il est incorrect de dire que celle-là se forme avant celle-ci. En effet, il n'y a pas de consensus à ce sujet et comme la littérature scientifique le révèle, l'ordre de la formation des différents niveaux d'organisation structurale pourrait varier en fonction des protéines ou bien dépendre de la méthode

d'observation utilisée. Je reviendrai sur ces questions dans les prochaines sections lorsqu'il sera question de la réaction de repliement comme telle.

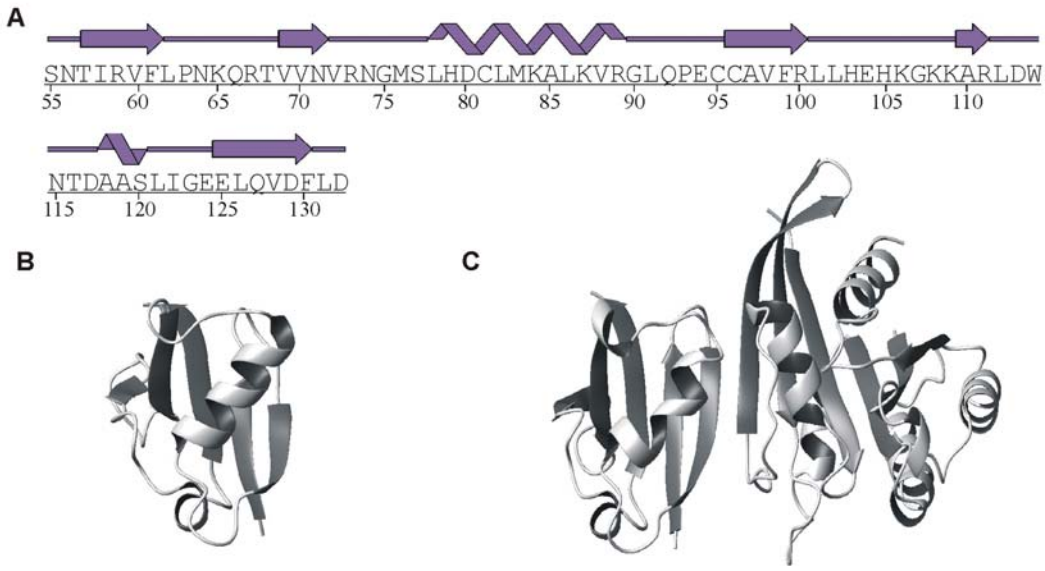


Figure 4. Présentation de la structure primaire, secondaire, tertiaire et quaternaire des protéines en prenant comme exemple le DLR de Raf.
A, La structure primaire et la structure secondaire du DLR de Raf. **B**, La structure tertiaire du DLR de Raf. **C**, La structure quaternaire du DLR de Raf : complexe formé par le DLR de Raf avec un double mutant de Rap1A, qui mime ainsi *ras* (code PDB, 1GUA). Notez que la GTPase est conjuguée à un analogue non-hydrolysable du GTP.

L'état natif est l'espèce structurale impliquée dans la réaction de repliement sur laquelle nous possédons le plus d'informations. Par le biais d'approches expérimentales qui fournissent des données atomiques précises tel que la RMN et la cristallographie par rayons X, les structures de 32 369[5] protéines, peptides ou complexes protéiques ont été résolues. A cause de leur instabilité et de leur plus grande variabilité, des structures représentatives des autres états (i.e., état dénaturé, intermédiaire et de transition) n'ont pu être déterminées sauf en de très rares exceptions. Par conséquent, l'état natif constitue le point d'ancrage structural le plus sûr et est donc essentiel à la description de la réaction de repliement.

[5] Ce nombre est celui annoncé sur le site de la PDB au 20 mars 2006 et inclue des doublons et des mutants d'une même protéine. A l'heure actuelle la progression dans la résolution de nouvelles structures est extremement rapide (plus de 5000 structures par an en 2004 et 2005) grâce à l'apport des efforts de la recherche en génomique structurale.

Peptides, protéines, domaines protéiques et « foldon » : distinction entre les termes

Le terme protéine (du grec *protos* « premier ») est le plus usité et répandu dans le langage populaire pour désigner une chaîne polypeptidique de grande taille. Il a d'abord fait son apparition dans le langage scientifique au début du XIX^{ème} siècle :

> « La matière organique, étant un principe général de toutes les parties constituantes du corps animal […] pourrait se nommer protéine » *Berzélius, lettre à Mulder, 1838, (Bulletin de sc. physiques en Néerlande, cité dans le Grand Robert de la langue française)*

Le sens donné au mot protéine a évolué et a été précisé par l'avancement de la connaissance scientifique qui permit la discrimination entre les divers matériaux fondamentaux du vivant. Ainsi, à notre époque le mot protéine est généralement dans le sens suivant :

> « Les protéines sont de très grosses molécules, de poids moléculaire variant de 10 000 à 1 000 000 ou plus. Ces macromolécules sont constituées par la polymérisation séquentielle de composés de poids moléculaire environ 100, appartenant à la classe des acides aminés.» *Jacques Monod (le Hasard et la nécessité, p. 69, cité dans le Grand Robert de la langue française)*

Dans son usage scientifique, ce terme en est venu à s'appliquer à toutes les chaînes polypeptidiques qui sont trop grandes pour être désignées sous le terme de peptide, cette taille limite étant sujet à controverse selon les sources et les points de vue. Pour cette raison entre autre, il n'y a pas de consensus quant à l'utilisation des termes protéine, polypeptide et peptide et leur usage se recouvre partiellement. Ils doivent donc être remplacés par des dénominations plus précises dès que le contexte le permet. Ainsi, un domaine est un élément modulaire d'une protéine qui a une structure et souvent une fonction définie. Le même domaine peut donc être retrouvé en tant qu'entité autonome dans plusieurs protéines. En effet, il y a plusieurs exemples de domaines qui sont extrêmement fréquents chez les protéines impliquées dans les voies de signalisation tels que les domaines SRC-Homology 3 (SH3), SRC-Homology 2 (SH2) et Pleckstrin-Homology (PH). La grande variété de

domaines rencontrés dans la nature autorise un éventail de combinaisons extrêmement diversifiées qui se reflète dans la riche diversité de l'arrangement des protéines impliquées dans les voies de signalisation cellulaires observées dans la nature. Par ailleurs, un domaine peut aussi être discontinu en séquence et n'a pas nécessairement la capacité de former sa structure native de façon indépendante, i.e. sans l'intervention du reste de la séquence ou d'un autre domaine. Des exemples correspondant à ce dernier type de domaines sont ceux liant l'adénosine triphosphate (ATP) des kinases à protéines et la forme réduite de la nicotinamide adénine dinucléotide phosphate (NADPH) chez la dihydrofolate réductase (DHFR).

Un « foldon » correspond à la séquence polypeptidique minimale et continue qui constitue une unité de repliement autonome *in vitro* (*8*). La première classe de domaine susmentionnée répond à ce critère. La vaste majorité de la littérature en biologie du repliement repose sur l'étude de domaines ou de protéines composées d'un seul « foldon ». Par ailleurs, le mot « foldon » a été utilisé plus récemment dans un sens différent par un autre groupe de recherche travaillant sur le cytochrome-c (*16*). Par conséquent, la prudence est de mise dans son utilisation et son interprétation tant que son usage ne sera pas plus répandu et consensuel.

Stabilisation de la structure native : un rôle prépondérant pour l'effet hydrophobe

La structure tridimensionnelle d'une protéine est stabilisée principalement par la formation de contacts non-covalents. Il y a des exceptions notables entre autre dans les protéines où l'on retrouve des ions ou des cofacteurs coordonnés et dans les protéines sécrétées ou exposées à la surface des cellules qui arborent des modifications post-traductionnelles qui impliquent la formation de nouvelles interactions covalentes menant à des changements importants au niveau structural telles que des glycosylations et la formation de ponts dissulfures. En excluant ces cas particuliers et les interactions de type

ponts hydrogène déjà décrites ci haut dans le cas des contacts intra chaîne principale, il faut souligner l'importance des interactions entre les chaînes latérales impliquant les forces électrostatiques et particulièrement l'effet hydrophobe dans la stabilisation de la structure native.

Les interactions ioniques ou les ponts salins établis entre des chaînes latérales de signes opposés tel que le glutamate (acide) et la lysine (base) sont fortes, mais étant donné leur rareté et les interactions avec les molécules de solvant qui nuisent à l'établissement de ce type d'interaction à la surface des molécules, leur importance dans la stabilisation de la structure des protéines est marginale. Plusieurs types d'interactions, dites de Van der Waals, peuvent se former entre groupements d'acides aminés formant des dipôles tel que le groupement carbonyle ou entre des dipôles induits qui sont formés entre autre sur des groupements aliphatiques, tel que des groupements méthyle et éthylène. On suppose que deux résidus forment un contact de Van der Waals lorsque les nuages électroniques d'un ou plusieurs de leurs atomes se retrouve à proximité. Malgré la faiblesse individuelle de ce type d'interactions, leur nombre considérable en fait une des forces majeures de la stabilisation des structures natives des protéines globulaires (*13*).

L'effet hydrophobe est un modèle qui a contribué à façonner notre vision contemporaine de l'organisation de la matière chez les êtres vivants (pour un point de vu intéressant et toujours pertinent le lecteur peut consulter (*17*)). Il permet entre autre de lier l'organisation des membranes lipidiques, des micelles de détergent et le phénomène d'enfouissement des chaînes latérales hydrophobes à l'intérieur de la structure des protéines globulaires à la tendance des substances hydrophobes de minimiser leurs contacts avec l'eau, qui est un solvant très médiocre de ces dernières. En effet, suivant un principe de physico-chimie trivial les molécules hydrophobes ont une meilleure solubilité dans des solvants non-polaires, et par conséquent leur transfert d'un milieu hydrophile à un milieu hydrophobe - de l'eau à l'intérieur d'une protéine, par exemple - est favorisé. Or, la force des interactions de type Van der Waals créées par une molécule d'H_2O avec un groupement méthyle ou éthylène d'une chaîne latérale hydrophobe ou bien de ces groupements entre

eux est comparable. Par contre, la présence de substance hydrophobe tel que les chaînes latérales d'acides aminés aliphatiques en surface, tel que cela pourrait survenir dans l'état dénaturé, force les molécules d'H_2O à s'ordonner autour de ces dernières afin de permettre leur solvatation. Cela aurait pour effet de diminuer l'entropie et donc d'augmenter l'énergie du système en réduisant le nombre d'arrangements potentiels du réseau de ponts hydrogène de l'eau (voir le **Chapitre 1 : Bases Théoriques**) d'où les arrangements spécifiques adoptés par les détergents et les protéines qui permettent de compenser cette pénalité thermodynamique. Les échelles d'hydrophobicité[6] permettent de classer les acides aminés en fonction de leur propension à former ou à éviter des contacts avec l'eau (**Tableau AI**) (*13;18*).

Voyons ce que dit W. Kauzman, en 1959, dans le cadre d'un article de synthèse des données de la littérature, qui est considéré comme un classique de la littérature scientifique, à propos de l'importance relative des divers types d'interactions non-covalentes qui peuvent être établies entre les acides aminés (*19*):

> « The fact that electrolytes generally fail to act as denaturing agents is probably an indication that salt linkages are not prominent contributors to the stability of proteins. The denaturing tendencies of detergents, of interfaces and of organic solvents, such as acetone, […] indicate the wide importance of hydrophobic bonds, since they should be weakened by these reagents. »

Les interactions de type ponts hydrogène et salins joueraient donc un rôle négligeable dans la stabilisation de la structure par rapport aux interactions de Van der Waals établies par les résidus non-polaires. Cela est probablement dû au fait qu'il y a peu de variations dans la contribution enthalpique et surtout entropique de l'un ou l'autre de ces types d'interactions entre l'état natif et dénaturé contrairement aux interactions hydrophobes. En résumé, selon

---

[6] Il y a plusieurs types d'échelles qui mesurent le niveau d'hydrophobicité. Certaines comme celle de Kyte et Doolittle utilisé dans le **Tableau A2** combine les tendances hydrophile et hydrophobes des acides aminés alors que d'autre sont basées sur la variation d'énergie du transfert d'un acide aminé d'un solvant hydrophile à un solvant hydrophobe.

ce modèle le phénomène de repliement des chaînes polypeptidiques découlerait de l'effet hydrophobe (i.e. de la propension des chaînes latérales hydrophobes à se regrouper afin de minimiser leurs contacts avec le solvant aqueux et donc de maximiser l'entropie du système solvant-protéine) et la structure serait stabilisée principalement par les interactions de type Van der Waals, lesquelles sont formées très majoritairement par ces mêmes résidus, mais dont leur formation en solution est favorisée principalement à cause de la composante entropique favorable décrite ci-dessus. L'utilisation du terme de lien hydrophobe doit être utilisé avec circonspection, car il diverge fondamentalement de la définition acceptée d'un lien chimique. Contrairement aux autres types d'interactions, les liens hydrophobes sont peu directionnels. Par conséquent, la notion de lien hydrophobe fait classiquement référence à la composante entropique qui favorise la réaction de repliement.

Dans la même veine, la notion de cœur hydrophobe (« hydrophobic core ») fait référence à l'intérieur de la structure des protéines globulaires qui regroupe la vaste majorité des résidus dont les chaînes latérales hydrophobes sont enfouies afin de minimiser le contact avec le solvant. De nombreuses interactions de Van der Waals sont formées entre les chaînes latérales de ces résidus. Il est généralement admis que seul l'arrangement optimal des chaînes latérales hydrophobes est compatible avec une structure tertiaire donnée et vice versa. Un seul cœur hydrophobe est habituellement présent chez les petites protéines globulaires.

Par ailleurs, quelques études intéressantes ont été publiées sur l'organisation du cœur hydrophobe. Tout d'abord, l'intérieur des protéines est empaqueté de manière extrêmement dense, proche du niveau observé dans les solides cristallins, en particulier pour les petites protéines (< 200 résidus) (*20*). Cela suggère que les résidus du cœur hydrophobe ont un faible niveau de tolérance à la mutation et à la variation de volume. Une étude a aussi constaté que le cœur hydrophobe de certaines protéines pouvait être organisé sur deux niveaux concentriques avec une couche interne presque totalement inaccessible au solvant et une couche externe, située à la limite du cœur interne et de l'interface de la protéine en contact avec le solvant (*21*). Nous avons confirmé une telle organisation du

DLR de Raf que nous avons confortée ultérieurement par divers résultats obtenus lors des études que nous avons menées (*22*) (voir le **Chapitre 2 : Résultats**).

Topologie structurale

La topologie (« fold » ou « topology ») est définie par l'enchaînement du squelette carboné de la chaîne polypeptidique, i.e. plus précisément l'arrangement approximatif des éléments de structures secondaires les uns par rapport aux autres dans l'espace tridimensionnelle. Donc, la topologie et la structure tertiaire sont des termes qui indiquent des réalités similaires en particulier pour les petites protéines monomériques. Cependant, il y a dans l'utilisation du mot « topologie », particulièrement lorsque deux protéines sont considérées adopter la même topologie, une plus grande tolérance à des différences structurales entre celles-ci. Par conséquent, la topologie fait référence à une certaine interprétation humaine ou informatique de la structure brute pour discriminer les éléments structuraux importants ou communs des variations structurales moins importantes. Les protéines ayant en commun uniquement leur topologie structurale seront désignées en tant qu'analogues structuraux (voir la **Figure 5** pour quelques exemples). Ces notions sont extrêmement importantes dans la mesure où les topologies sont beaucoup mieux conservées que la structure primaire des protéines. Par ailleurs, j'ai trouvé intéressant dans un premier temps de remettre ce thème dans son cadre historique.

La première structure cristallographique d'une protéine, c'est-à-dire celle de la myoglobine isolée du sperme de baleine, avait surpris par son irrégularité en comparaison à l'ADN double brin dont la structure avait été résolue quelques années plus tôt (*23*). Avec l'avancement des connaissances, la comparaison de la structure de la myoglobine et de l'hémoglobine de quelques espèces devient possible et révèle une remarquable similarité dans l'arrangement générale de leur chaîne polypeptidique, compatible avec une conservation générale de la topologie. Par ailleurs, l'alignement des séquences (9) de ces protéines met en lumière la conservation du type d'acides aminés à uniquement 9 des 140

résidus qui peuvent être alignés (*24*). Des études ultérieures reposant sur des regroupements comportant un nombre plus élevé de structures et de séquences renforcent et généralisent à d'autres classes de protéines les conclusions initiales de Perutz et coll. (*25-28*). Il apparaît donc d'ores et déjà que des séquences très diverses peuvent encoder la même topologie. Avec l'augmentation du nombre de séquences et de structures et des outils expérimentaux, plusieurs familles de protéines ou de topologies ont pu être soumises à ce genre d'études comparatives (*29-36*). Il transparaît de ces études que certaines positions du cœur hydrophobes sont conservées, mais pris isolément ces résultats ne suffisent pas à la détermination de règles générales qui pourraient expliquer le rôle spécifique des résidus conservés dans la formation et la stabilisation de la structure native. La question reste entière, quels sont les déterminants de la séquence qui permettent à des chaînes polypeptidiques aussi diverses d'adopter une topologie structurale commune. Parmi les millions de gènes protéiques potentielles, qui sont réévalués à environ 150 000 en tenant compte des redondances dues aux séquences très similaires entre les organismes (*37*), il y aurait seulement de 1000 à 10 000 domaines aux topologies distinctes, du moins, il s'agit des estimations les plus largement acceptées en ce moment (*7;8*). Ainsi, lorsque l'on classe les protéines et les domaines en fonction de leur topologie, on observe que quelques-unes de celles-ci sont très répandues. En effet, les 400 topologies les plus fréquentes ou méso topologies (« mesofold ») correspondent à la très vaste majorité des structures distinctes observées jusqu'à ce jour (*7*). Les 10 méso topologies les plus fréquentes sont dénommées supertopologies (« superfold ») (**Tableau I**). Le DLR de Raf appartient à l'une de ces supertopologies, celle dite similaire à ubiquitine (i.e., « ubiquitin-like roll », aussi connu sous le nom de « β-grasp ubiquitin-like » ou « ubiquitin-superfold »), qui se caractérise par l'apposition d'une helice-α sur un feuillet-β mixte de topologie 2-1-5-3-(4)[7]. Notez les similitudes dans l'organisation structurale de plusieurs membres de cette topologie représentés à la **Figure 5**. La diversité de séquences des protéines classifiées dans la topologie d'ubiquitine est très grande, et à cause de sa haute occurrence et versatilité

[7] Le quatrième brin est très court et n'apparaît pas chez tous les membres de la topologie.

fonctionnelle qui a suscité de l'intérêt, quelques études s'attachant à comparer les séquences de certains analogues structuraux, en particulier pour les membres de la superfamille d'ubiquitine (voir la prochaine section pour une description de l'organisation de la topologie d'ubiquitine), ont été rapportées (*29;30;36*).

Un des éléments clé dans les études que j'ai réalisées au cours de ma thèse repose justement sur la comparaison des séquences adoptant la même structure ou la même topologie. Dans ce cas, il est souhaitable de mesurer le niveau de conservation des résidus dans un alignement. Plusieurs méthodes existent pour ce faire et elles sont comparées en détail dans un article de synthèse de la littérature fort à propos (*38*). L'efficacité de ces méthodes afin d'obtenir des prédictions sur les déterminants structuraux d'une topologie est

Tableau I. Classement des 10 super topologies les plus fréquentes (données tirées de (*7*)).

Supertopologies[a]	# superfamilles[b]	% d'apparition dans les domaines à la structure connue
Doublement-blessé	122	8.8
Trèfle-β	2	0.1
Simili-ferrodoxine	65	4.7
Simili-imunoglobuline	55	4
Baril-TIM	28	2
Faisceau haut bas	17	1.2
Gâteau-roulé	17	1.2
Rouleau-ubiquitine	16	1.1
Topologie OB	16	1.1
Simili-globine	4	0.3
Supertopologies	342	24.7
Toutes les topologies	1386	100

[a] Traduction libre de l'anglais, respectivement : « doubly-wound », « β-trefoil », « ferredoxin-like », « immunoglobulin-like », TIM-barrel, « updown bundle », « ubiquitin-roll », « OB fold » and «globin-like».

[b] Selon la banque de données CATH.

fortement tributaire de la quantité et de la diversité des séquences disponibles. En effet, l'information disponible dans des alignements de séquences très diversifiées telle que l'on peut l'observer chez certaines topologies permet de révéler les propriétés et résidus conservés et peut donc être utile pour le design de protéines adoptant ces structures. Par ailleurs, d'autres topologies sont extrêmement rares et souvent limitées à un seul type de fonction cellulaire. L'information de séquence disponible pour ces topologies est donc très faible. D'autres topologies n'ont tout simplement jamais été observées dans la nature. Ainsi, en utilisant des simulations informatiques, une séquence a été optimisée afin d'adopter une telle topologie. Il est intéressant de noter que malgré la rareté supposée de ce type de topologie, cette protéine est extrêmement stable (*5*). Il est probable que les topologies structurales les plus flexibles aient été retenues au cours de l'évolution, car elles possèdent la capacité de réaliser diverses fonctions biologiques. D'autre part, la haute occurrence de certaines topologies pourraient être un indicateur de leur apparition hâtive au cours de l'évolution.

Les principales banques de données de classification et de comparaison structurale

Tout d'abord, les structures de toutes les biomolécules publiées peuvent être retrouvées sur le site de la « Protein Data Bank » (PDB). Plusieurs banques de données utilisent le répertoire structural complet de cette banque de données pour assembler les protéines et domaines arborant des similarités topologiques. Par exemple, la banque de données « Class, Architecture, Topology and Homologous superfamily » (CATH) classifie les structures de façon hiérarchique en groupes réunissant des structures de plus en plus semblables et ce, via un processus partiellement automatisé (*39;39;40*). La banque de données « Family of Structurally Similar Proteins » (FSSP) utilise un algorithme de comparaison de structure, le « Distance mAtrix aLIgnment » (DALI), pour classifier les structures en fonction de leur topologie (*35;41;42*). Elle n'est pas construite de façon hiérarchique ce qui diminue son efficacité à établir des liens entre les diverses topologies. En contrepartie, elle fournit un alignement de séquences des protéines sélectionnées en utilisant les recoupements entre les divers segments de structures secondaires et des

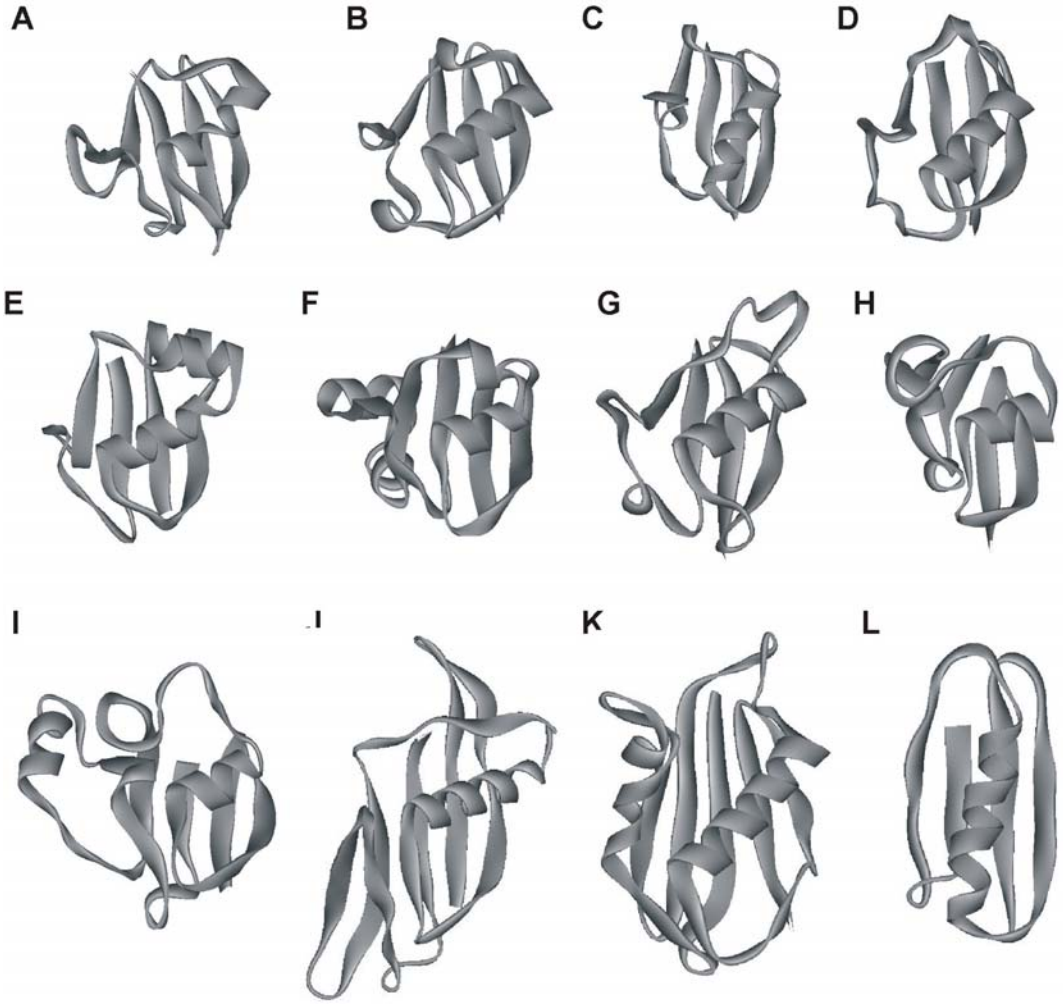


Figure 5. Représentation structurale de membres de certaines superfamilles de la topologie d'ubiquitine.
A, Le DLR de Raf (1RFA) et, **B**, Ubiquitine (1UBI) : superfamille simili-ubiquitine (s1). **C**, Le domaine de dimérisation de CAD (1C9Fa) et **D**, Le domaine de dimérisation de ICAD (1F2Ri) : superfamille de CAD et PB1 (s2). **E**, La sous-unité MoaD de la synthétase de la molybdoptérine (1FMAd) : superfamille MoaD/ThiS (s3). **F**, Domaine TGS en carboxy-terminal de la protéine YchF (1JALa) : superfamille TGS (s4). **G**, Domaine double-cortine (D-C) de la D-C (1MG4) : superfamille D-C (s7). **H**, Déhydrogénase au monoxyde de carbone (1FFVa) et, **I**, Ferredoxine (1L5Pa) : superfamille 2Fe-2S similaire à la ferredoxine (s5). **J**, Staphylokinase (1C78a) : superfamille staphylokinase/streptokinase (s6). **K**, Exotoxine C (1AN8) : superfamille des super-antigènes/toxines (s8). **L**, Protéine-L (1HZ6) : superfamille des domaines de liaison aux immunoglobulines (s9). Les 7 premiers panneaux (**A-G**) sont les représentants de la superfamille d'ubiquitine et de 4 superfamilles probablement reliées évolutivement avec cette dernière alors que les autres représentent des superfamilles divergentes (**H-L**). Le code attribué à chaque structure dans la banque de données PDB est indiqué entre parenthèses. Deux superfamilles ne sont pas représentées dans cette figure, parce qu'elles ont des caractéristiques distinctes (voir section « Supporting Information » de l'Article 2 (*22*)), soit celles du facteur d'initiation de la traduction IF3 (s10) et du domaine amino-terminal de la glutamine synthétase (s11). Une troisième superfamille liée aussi à la superfamille d'ubiquitine, soit celle de la TmoB-like (s12) n'est pas représentée, car elle a été annotée tout récemment (superfamille-1, s1 ; superfamille-2, s2 ; etc.).

statistiques pour comparer chaque séquence retrouvée par rapport à la séquence de la structure utilisée pour la recherche. La banque de données « Conserved Key Amino Acids

in Protein sequences » (CKAAPs) utilise les alignements de FSSP ou ceux obtenus par d'autres méthodes afin d'aider à l'identification statistique des résidus conservés dans les protéines structuralement similaires (*34;43;44*). La banque de données « Structural Classification of Proteins » (SCOP) regroupe des protéines et des domaines à la structure semblable dans des topologies (« fold ») réparties selon la classe structurale à laquelle elles appartiennent (i.e., α, β, α+β et α/β, où α = hélice-α et β= brin-β) (*45-49*). Les diverses topologies dans SCOP sont subdivisées en 1 ou plusieurs superfamilles (« superfamilies ») et celles-ci en sous-groupes dits familles (« famillies ») dans lesquelles le lien évolutif et fonctionnel, la similarité de séquence et de structure vont en ordre croissant. Afin de décrire schématiquement l'organisation hiérarchique dans SCOP, je mets ci-dessous la classification propre à la topologie d'ubiquitine telle qu'elle est y retrouvée, en prenant plus particulièrement comme exemple spécifique le DLR de Raf (**Figure 6**). La classification hiérarchique utilisée par SCOP est distincte de celle utilisée dans CATH, dans laquelle les superfamilles ont un lien fonctionnel commun plus grand. La banque de données SCOP est basée et mise à jour régulièrement par le truchement d'humains experts et l'utilisation d'outils informatiques. J'ai trouvé son organisation plus conviviale que toutes les banques de données mentionnées ci-dessus. Finalement, la banque de données « Superfamily » permet d'attribuer à des gènes protéiques dont la structure est inconnue, et qui sont issus de divers organismes pour lesquels le génome a été caractérisé, une structure hypothétique fondée sur la similarité de séquence avec des superfamilles de SCOP (*50*). À l'aide de cette banque, il est donc possible d'évaluer la fréquence d'utilisation d'une topologie donnée.

Maintenant que la notion de topologie et superfamilles a été introduite, je vais profiter de cette occasion pour discuter de l'organisation de la topologie d'ubiquitine. La classification dans la banque de données SCOP de la topologie d'ubiquitine et schématisée ci-dessus indique que cinq des superfamilles semblent avoir un lien évolutif avec la superfamille d'ubiquitine alors que les sept autres en seraient plus éloignées. A cet égard, il est intéressant de remarquer comment les liens évolutifs présumés entre les diverses protéines classifiées dans la supertopologie d'ubiquitine sont reliés à leur superfamille

d'origine et sont même visibles empiriquement dans les variations de la structure tertiaire de divers membres choisis et représentés à la **Figure 5**. Les liens entre le classement de SCOP et les variations de la structure tertiaire se répercutent sur l'organisation structurale des résidus du cœur hydrophobe interne de ces protéines (**Figure 7**), tel que perçu du point de vue du réseau de contacts établis entre les chaînes latérales de ces résidus.

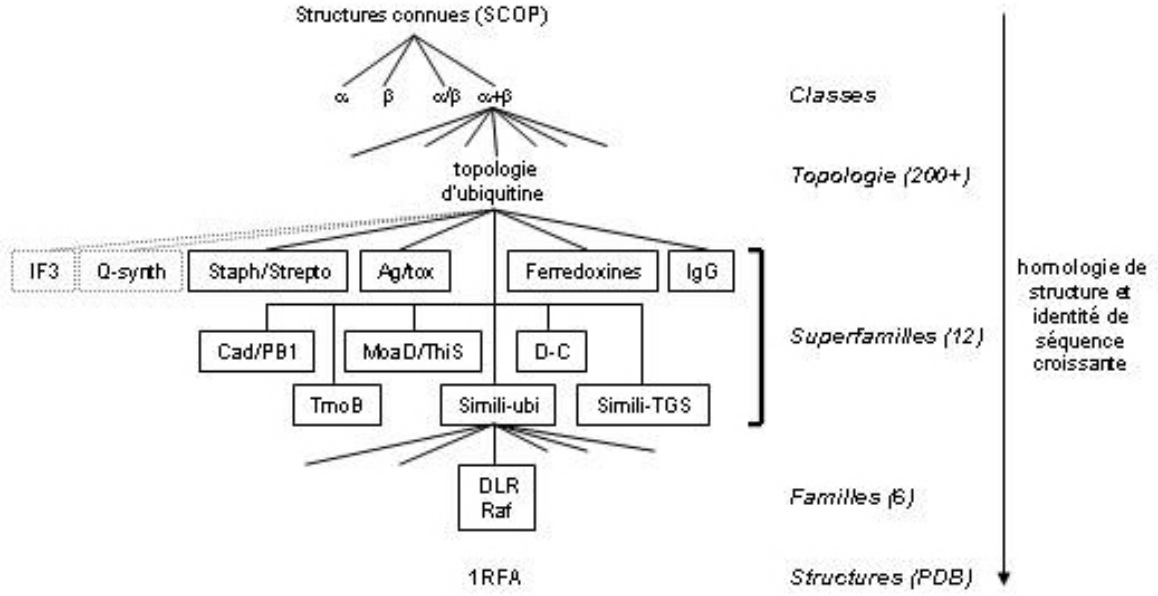


Figure 6. Organisation hiérarchique de la banque de données SCOP : exemple du DLR de Raf.
Les chiffres entre parenthèses font référence au nombre d'entités distinctes à chaque niveau de l'organigramme. Des liens évolutifs présumés entre les superfamilles sont dénotés par la position de leur embranchement au tronc commun. Deux superfamilles représentant des distinctions structurales très importantes sont dénotées par des lignes pointillées. Les superfamilles staphylokinase/streptokinase (s6), super-antigènes/toxines (s8), les domaines liant les immunoglobulines (s9) et dans une moindre mesure celle des ferredoxines (s5) sont aussi clairement différentes des superfamilles reliées à l'ubiquitine.

Ainsi, l'organisation structurale du cœur interne des protéines ou domaines classifiés dans les superfamilles reliées à celle d'ubiquitine est semblable et comprend des variations mineures principalement limitées à la disposition de l'hélice-α principale contre le feuillet-β. Par contre, les autres superfamilles ont des arrangements distincts entre eux suggérant que leurs relations évolutives les unes par rapport aux autres et par rapport aux

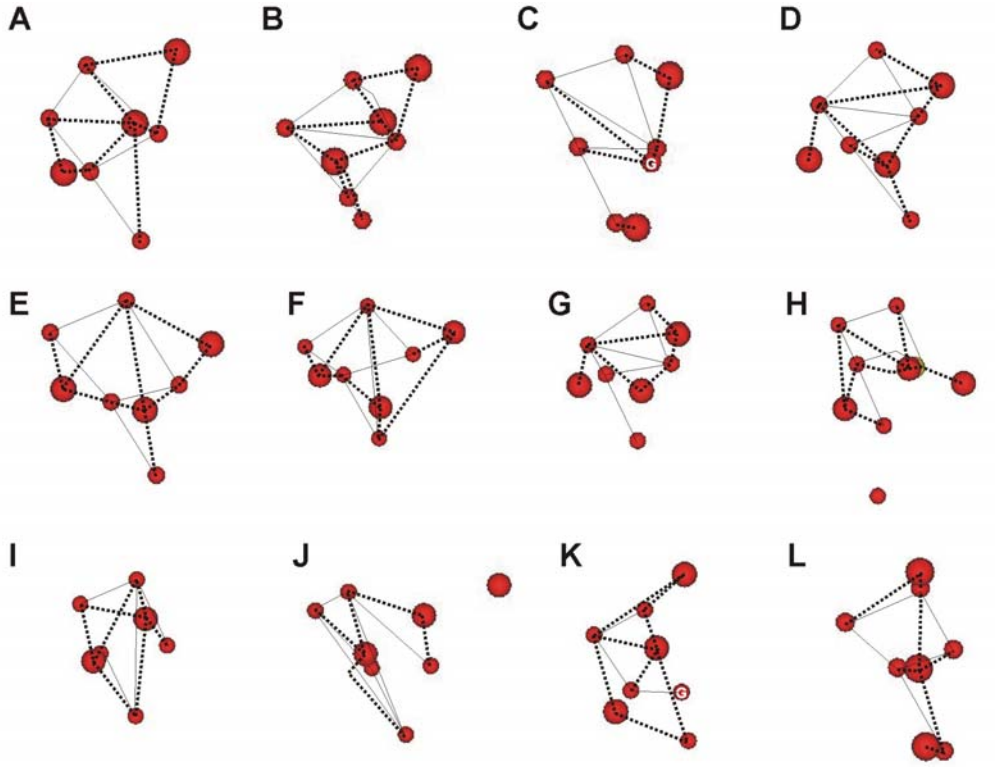


Figure 7. Représentation structurale simplifiée des membres de la topologie d'ubiquitine.
Les résidus du cœur hydrophobe interne (voir l'**Article 2** (*22*) et la section « Supporting Information » l'accompagnant) sont représentés par une sphère rouge correspondant à leur C_β ou au C_α dans le cas de la glycine. Deux résidus ayant au moins un atome non-hydrogène de leur chaîne latérale à moins de 6 Å de distance sont considérés comme étant en contact à condition qu'il n'y ait aucun atome d'un autre résidu entre ceux-ci. Les contacts répondant à ce critère sont reliés par un trait, soit plein pour des résidus dans la même surface ou pointillé pour des résidus localisés dans des couches distinctes (dans ce cas-ci, il s'agit de l'hélice-α principale et du feuillet-β). Les structures sont placées de manière à placer la couche de l'hélice-α en premier dans le plan (grosse sphère) suivi du feuillet- β (petite sphère). Les résidus situés dans un même élément de structure secondaire ne sont pas reliés dans ce schéma. Les motifs formés par ces contacts peuvent être classés en deux groupes : (**A-C**) semblable au DLR de Raf; (**D-G**) semblable à l'ubiquitine alors que les arrangements des autres structures sont diversifiés (**H-L**). **A**, Le DLR de Raf (1RFA). **B**, Le domaine amino-terminal de dimérisation de ICAD (1F2Ri). **C**, Le domaine amino-terminal de dimérisation de CAD (1C9Fa). **D**, Ubiquitine (1UBI). **E**, La sous-unité MoaD de la synthétase de la molybdoptérine (1FMAd). **F**, Domaine TGS en carboxy-terminal de la protéine YchF (1JALa). **G**, Domaine D-C de la D-C (1MG4). **H**, Déshydrogénase au monoxyde de carbone (1FFVa). Un résidu supplémentaire (résidu 10 coloré en jaune) dans le cœur hydrophobe de 1FFVa a été ajouté pour tenir compte de l'empaquetage distinct de cette protéine. **I**, Ferredoxine (1L5Pa). **J**, Staphylokinase (1C78a). **K**, Exotoxine C (1AN8). **L**, Protéine-L (1HZ6). Comparez les structures homologues de la **Figure 5** aux schémas ci-dessus, et observez leur similarité au niveau de la structure tertiaire.

superfamilles similaires à celle d'ubiquitine sont soit nulles ou à tout le moins beaucoup plus distantes. En résumé, il semble donc clair que la topologie contraint l'organisation structurale du cœur hydrophobe et vice-versa (voir le **Chapitre 2 : Résultats**; **Article 2** et **3**).

Les polypeptides ayant en commun uniquement la même topologie générale sont des analogues structuraux. Les protéines et domaines regroupés dans une famille peuvent donc avoir une fonction commune ce qui en fait plus que de simples analogues structuraux. Il faut alors parler d'orthologues, de paralogues ou d'homologues fonctionnels en fonction de la similarité de leurs rôles fonctionnels chez différents organismes ou à l'intérieur de la même espèce. Des banques de données comme SMART et PFAM classent des domaines ou des protéines en groupements fonctionnels sur la base de leur similarité de séquence et de leur rôle cellulaire (*51;52*).

Une théorie importante en biologie du repliement suggère que les protéines ayant une topologie similaire se replieraient en utilisant un processus semblable. Comme nous le verrons dans la section sur l'état de transition, des résultats contradictoires concernant les réactions de repliement et de dépliement de protéines à la topologie similaire suggèrent un modèle plus complexe, sans remettre en cause entièrement cette prédiction triviale.

Les banques de données fournissant des informations structurales peuvent être fort utiles à l'étude du repliement de protéine. Par exemple, les banques de données CATH, FSSP et surtout SCOP permettraient de choisir de manière plus rigoureuse les protéines modèles dont le repliement devrait être étudié afin d'élargir notre compréhension de ce phénomène biophysique. En effet, leur utilisation systématique faciliterait le choix rationnel des modèles expérimentaux représentatifs d'une topologie donnée ou les topologies à étudier et à comparer. En théorie, la capacité de perturber expérimentalement la séquence d'un polypeptide permet d'augmenter artificiellement la variabilité de séquences observée dans une protéine naturelle, une topologie ou une classe fonctionnelle donnée. Dans le cas d'une topologie peu répandue, cela pourrait permettre de déterminer les résidus importants à sa formation et sa stabilisation et ainsi aider au design de cette structure. Dans ce cas, les banques de données structurales et fonctionnelles mentionnées ci-dessus pourraient servir de base de comparaison à des résultats expérimentaux et

permettraient de discriminer entre des caractéristiques générales ou non du repliement de protéines adoptant une topologie donnée.

La perturbation de la structure primaire[8] afin d'étudier les déterminants de séquence de la structure des protéines

Afin de réaliser ce type d'expérience, il faut donc appliquer ou développer des approches permettant la perturbation de la structure primaire et la sélection des mutants qui permettraient la formation et la stabilisation de la structure native modèle. Par conséquent, la réussite de ce type d'expérience est premièrement tributaire de la capacité technique à créer expérimentalement des librairies dégénérées aléatoires[9] de la structure primaire. Celle-ci est habituellement insérée dans la structure primaire par des outils de biologie moléculaire remplaçant ainsi les codons natifs par des codons dégénérés appropriés, qui permettent l'insertion au hasard d'acides aminés. Par exemple, la dégénérescence concomitante de plus de 9 résidus pour les 20 acides aminés du code génétique (i.e. en utilisant des codons dégénérés, tel NNK; **Tableau AII**) est limitée par le nombre extrêmement élevé de clones indépendants à générer pour couvrir une partie significative des possibilités combinatoires ($9^{20} \approx 5x10^{11}$) et ainsi recouvrer un nombre suffisamment élevé de clones indépendants pouvant se replier à la structure originelle. Considérant cela et comme les domaines protéiques les plus courts sont composés d'une quarantaine de résidus, il est nécessaire de trouver des solutions alternatives à la dégénérescence concomitante de tous les résidus afin d'obtenir des informations sur les déterminants de la séquence pour un maximum de positions d'un polypeptide donné. De plus, le biais introduit par la méthode de perturbation de la séquence choisie sur l'occurrence théorique de chaque

---

[8] La perturbation de la structure primaire désigne toute approche expérimentale qui vise à remplacer plusieurs codons dans la séquence de type sauvage (ts) par un codon dégénéré qui permet l'insertion de plusieurs types d'acides aminés. Les études citées ci-dessus présentent les résultats de la mutation d'un nombre significatif de résidus.

[9] Le code génétique ne permet pas d'inserer un ratio équivalent de chaque acide aminé dans les librairies dégénérées. Le qualificatif d'aléatoire indique que les librairies doivent être soustraites autant que possible à tout biais autre que structural dans la fréquence des acides aminés observés à chaque position variée. Le choix de la méthodologie est crucial à cet égard.

acide aminé, qui est inévitable ne serait-ce que par la nature et la distribution des 64 codons pour les 20 acides aminés, doit être pris en compte dans l'analyse des résultats. Ensuite, la capacité à sélectionner les variants de la séquence pour leur capacité à se replier à la structure native de la protéine d'intérêt est la seconde étape expérimentale clé. À ce jour, tous les moyens appropriés pour le faire reposent sur des méthodes qui tirent parti d'essais fonctionnels tel que l'activité enzymatique, la répression de la transcription et sur d'autres méthodes qui permettent par exemple la sélection de variants sur la base de la conservation de la capacité à former un complexe avec un ligand ou un partenaire protéique. La capacité des séquences à se replier est donc sélectionnée indirectement, rendant donc nécessaire l'évaluation de l'ampleur du biais fonctionnel potentiellement introduit par l'essai de sélection.

Le laboratoire de Robert Sauer a accompli la tâche du pionnier dans le développement et l'application des méthodes de perturbation de séquence à l'élucidation des déterminants de séquences encodant la structure et la fonction rudimentaire d'une petite protéine. Dans le cadre de ses premières études sur le répresseur-λ, Sauer et coll. s'intéressèrent à la dégénérescence de la séquence correspondant à l'hélice-α constituant l'interface d'homodimérisation, qui est en outre essentielle à l'activité de répression transcriptionnelle, et de 7 résidus qui font partie du coeur hydrophobe du domaine (**Figure 8**) (*53;54*). Dans l'approche expérimentale qu'ils ont privilégiée, des codons dégénérés permettant l'insertion des 20 types d'acides aminés étaient introduits par une variante ingénieuse de la méthode de mutagenèse par cassette dans laquelle les bases dégénérées sur un brin sont appariées à une base inosine, qui s'apparie similairement avec toutes les autres bases Watson-Crick. Un essai de survie axé sur la complémentation d'une souche altérée de phage λ n'exprimant pas le répresseur-λ par les variants du répresseur-λ de la librairie fut utilisé afin de sélectionner les séquences compétentes pour le repliement. Leurs résultats ont démontré une forte prédilection pour les acides aminés hydrophobes et une faible variation du volume de la chaîne latérale aux positions enfouies et à l'interface de dimérisation. De plus, lorsque plusieurs résidus du coeur hydrophobe étaient covariés, ils

ont observé une nette augmentation de la variabilité du type d'acides aminés hydrophobes tolérés. Le même laboratoire a aussi étudié des contraintes structurales et fonctionnelles d'Arc-répresseur en utilisant une variante de cette approche, alors que l'organisation structurale de cette protéine était inconnue à l'époque (*55*).

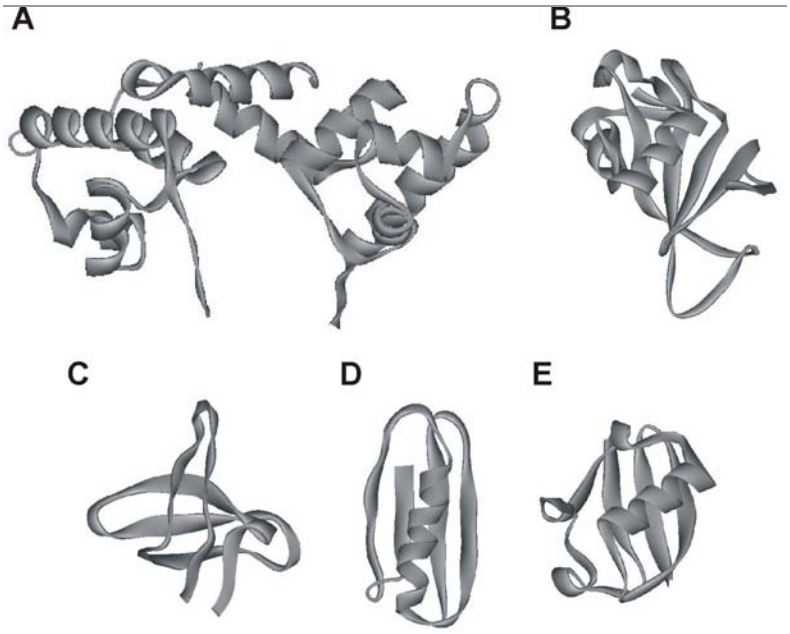


Figure 8. La structure de certaines protéines soumises à des expériences de dégénérescence de la séquence.
A, Le phage-λ (1LLI). **B**, Barnase (1A2P). **C**, Le domaine SH3 de SRC (1SRL) **D**, Protéine-L (1HZ6) **E**, Ubiquitine (1UBI). Le code de la structure dans la PDB est indiqué entre parenthèses.

Dans des études similaires à celles sur le répresseur-λ, des résidus du cœur hydrophobe de la protéine barnase et d'ubiquitine ont été dégénérés conjointement pour des acides aminés hydrophobes (V, I, L, F, M) (**Figure 8**) (*56;57*). Dans la première étude, les variant de barnase furent sélectionnés en utilisant un essai de transcription comme méthode de sélection Les auteurs ont noté une variation de l'ordre de +/- 5 % du volume des chaînes latérales des 13 résidus du coeur hydrophobe de barnase à l'étude, ce qui concordait avec les résultats obtenus sur le répresseur-λ. La conclusion principale de ces travaux est que l'hydrophobicité de la chaîne latérale des acides aminés qui est retrouvé aux résidus du coeur hydrophobe constituent un facteur prédominant dans la détermination de la capacité d'un variant à adopter la structure native. D'autre part, la dégénérescence simultanée de 8

acides aminés situés dans la moitié amino-terminale d'ubiquitine, soit les résidus 1, 3, 5, 13, 15, 17, 26 et 30 révèle une forte sélection pour les acides aminés de type-sauvage (ts), ce qui la distingue des études précédentes. Cela pourrait être dû à la méthode de sélection des séquences qui reposait sur un essai de résistance à la trypsine couplé à la méthode d'exposition sur phage (« phage display ») ou au confinement des résidus variés à une partie restreinte du cœur hydrophobe dans cette étude.

Le laboratoire de David Baker a publié des études de perturbation de séquence fort intéressantes par leur conjugaison avec des études de cinétique de repliement et de dépliement. Afin de parvenir à isoler les variants de séquences qui permettent un repliement correct, Gu et coll. ont développé des essais de liaison reposant sur la méthode d'exposition sur phage (*58*). La première étude consista en une tentative de simplification de la séquence du src SH3 qui limitait à 5 acides aminés, représentant les principales classes physico-chimiques des acides aminés du code génétique, la dégénérescence permise aux 46 résidus du domaine dont la séquence a été dégénérée dans le cadre de ce travail (**Figure 8**) (*59*). Les mutants obtenus après deux rondes de sélection ont démontré des caractéristiques de repliement similaires à la séquence de ts, ce qui confirmait la faisabilité d'obtenir des protéines fonctionnelles à partir d'un répertoire minimal d'acides aminés fondamentaux tel qu'il aurait pu l'être durant l'abiogenèse ou bien chez les premières protéines issues des protocellules. La seconde étude qui est plus pertinente est basée sur la dégénérescence de la séquence de la protéine-L (**Figure 8**), qui est classée dans la super topologie d'ubiquitine, mais qui se différencie de la structure canonique par la structure allongée de son feuillet-β et de son hélice-α. Son aspect général est beaucoup plus compact et elle est largement exempte de boucles, ce qui donne un arrangement relativement symétrique, composé successivement d'un premier motif épingle à cheveux, d'une hélice-α et d'un second motif épingle à cheveux. Quatre segments de la protéine-L comprenant en tout près de 50 % des résidus du domaine furent variés séparément par l'insertion de codons dégénérés dont la nature variait selon les régions, de telle sorte que pour les deux tour-β et l'hélice-α (codons NNS) ainsi que pour les brins-β 1 et 4 (i.e., respectivement les

codons DHM et DHW) (**Tableau AII**), il y avait respectivement 20 et 12 types d'acides aminés permis (*60;61*). Malgré ces lacunes expérimentales qui compliquent l'interprétation de données, il fut constaté que les tours-β des deux motifs épingle à cheveux se comportaient différemment, celui en amino-terminal démontrant une sensibilité accrue à la variation de sa séquence. Les conséquences sur le mécanisme et accessoirement sur l'optimisation de la cinétique de repliement, et la compréhension du rôle de la conservation de la structure primaire dans la détermination de ceux-ci seront revues plus en détail dans la section sur l'état de transition.

La lacune la plus apparente des études décrites ci-dessus est la faible couverture des expériences de perturbation de séquences qui y sont décrites. En d'autres mots, il y a trop peu de résidus dégénérés pour statuer sur le rôle et sur l'ensemble des déterminants de séquence. Il serait aussi souhaitable d'homogénéiser les codons dégénérés utilisés dans le cadre d'une même étude afin de réaliser une analyse comparative sérieuse. En second lieu, les variants ont été sélectionnés par des méthodes basées sur la conservation de la fonction et donc pour la structure native de manière indirecte à cause de l'absence de méthodes alternatives. Dans ce contexte, une analyse rigoureuse du biais expérimental devrait être effectuée en comparant les séquences obtenues expérimentalement et les données contenues dans les banques de données concernant les analogues structuraux et homologues fonctionnels des protéines ou domaines protéiques étudiés.

Retour sur des considérations techniques

Dans les études décrites dans la section précédente, trois méthodes ont été utilisées pour introduire les codons dégénérés dans un segment contigu ou à des positions disparates d'une séquence d'ADN. Tout d'abord, les codons dégénérés peuvent être introduits par PCR ou bien par mutagénèse de type cassette (*53; 61*) (voir aussi le **Chapitre2 : Résultats**; **Article 1** et **2**). Cette dernière approche a une limitation gênante, étant donné que la cassette doit être insérée par l'utilisation d'enzymes de restriction, ce qui signifie que des sites de reconnaissance appropriés pour ces dernières doivent être introduits. La méthode

présentée par Cocco et coll., qui est proche de la méthode par cassette, représente aussi un attrait certain, quoiqu'elle ait été peu utilisée au moment de rédiger cet ouvrage, si l'objectif recherché est de dégénérer des résidus qui sont dispersés sur toute la séquence (*62*). Les approches utilisant la PCR quant à elles risquent d'introduire un biais pour la séquence native lors de l'amplification, car les amorces plus similaires à la séquence originale ont une meilleure affinité pour le gabarit. A cause de ce potentiel écueil, la méthode Kunkel[10] représente une alternative avantageuse. Cette stratégie permet d'altérer la séquence de plusieurs segments à la fois. La méthode elle-même comporte plus d'étapes distinctes que la méthode par PCR, mais un protocole très détaillé et bien fait d'une variante améliorée de cette méthode est disponible (*63*). Par ailleurs, le biais de la méthode par PCR peut être minimisé en veillant à enlever la séquence du segment à dégénérer du gabarit de la réaction. C'est la solution que nous avons privilégiée pour nos propres expériences (*22;64*) (voir respectivement l'**Article 1** et **2**).

Comme je l'ai brièvement mentionné ci-dessus, la plupart des travaux présentant la sélection de séquence à partir de librairies dégénérées reposent sur l'utilisation de stratégie de sélection indirecte. La méthode la plus couramment utilisée est la méthode d'exposition sur phage qui permet de sélectionner les variants avec une certaine affinité pour leur ligand. Ce type de sélection a un attrait particulier, car il est plus facilement généralisable et applicable à nombre de protéines modèles. Par ailleurs, le « Protein-fragment Complementation Assay » (PCA) (le concept de cette méthode est révisée dans (*65;66*)) de la DHFR a été utilisé auparavant afin d'optimiser un hétérodimère de zipper de leucines à partir du criblage d'une librairie combinatoire de nucléotides dégénérés (*67*). Cela laissait présager que l'on pourrait tirer profit de cette technique dans le cadre d'une stratégie expérimentale de perturbation de la structure primaire.

D'autre part, il serait souhaitable de développer des approches qui permettraient de sélectionner les séquences uniquement en fonction de leur capacité à adopter la structure

[10] La méthode de Kunkel repose sur la production d'ADN simple brin qui sert gabarit à une polymérase dans une réaction simple (i.e., par opposition à l'amplification lors d'une PCR).

originale native ou à tout le moins une conformation soluble et repliée. Plusieurs approches ont été développées afin de sélectionner des polypeptides solubles d'une librairie de séquences en fusion à la GFP, grâce à l'intensité supérieure du signal de fluorescence produit dans ces cas-là (*68;69*). Une autre approche a été proposée avec l'utilisation du domaine SH2 dans une variante de l'exposition sur phage. Dans cette méthode, la librairie d'intérêt est insérée dans une boucle du domaine SH2 située en surface de sa structure ce qui mène à une forte altération de sa stabilité. L'hypothèse de travail de cette approche suggère que les séquences qui permettraient la formation d'une structure compacte mèneraient en retour à la stabilisation du domaine SH2 et donc à sa liaison à la colonne d'affinité (*70*). Cependant, les premiers résultats obtenus à partir d'une librairie aléatoire ont indiqué que des polypeptides solubles, mais avec des propriétés structurales similaires à des polypeptides non-structurés sont sélectionnées (*71*). Le concept de cette stratégie est trop élégant pour être abandonné et il serait tentant de l'appliquer à d'autres systèmes de criblage, en particulier *in vivo*. En conclusion, les quelques approches qui ont été proposées jusqu'à maintenant pour sélectionner directement la capacité d'une séquence à se replier ne permettraient pas d'atteindre les mêmes objectifs que les méthodes indirectes décrites précédemment. Il reste donc du travail à faire sur ce plan.

Dans les prochaines sections, je m'attarde principalement sur les espèces qui contribuent à définir le mécanisme de repliement. Tout au long de ces sections, je vais mettre en perspective le travail et la contribution des chercheurs qui ont été les pionniers de la recherche du mécanisme de repliement en insistant surtout sur les constatations et les questions centrales qu'ils ont formulées.

De la dénaturation des protéines : perspectives historiques

Les propriétés des protéines varient selon les conditions auxquelles elles sont exposées. Leur transfert dans des conditions non-physiologiques, que ce soit de température, de pH ou en présence de certaines substances organiques peut provoquer des changements détectables au niveau microscopique, voire même macroscopique.

Effectivement, le changement de la viscosité, de la sensibilité aux protéases, de la réactivité des chaînes latérales et du taux d'échange deutérium hydrogène des groupes peptidiques pour quelques protéines indiquaient déjà au tournant des années 1950 qu'au cours de ce processus il y a un remaniement de la conformation des protéines qui mène à l'exposition de certains groupements des acides aminés habituellement inaccessibles au solvant. La clémence des conditions utilisées pour dénaturer les protéines suggère qu'elle s'accompagne de la rupture d'interactions chimiques plutôt faibles. Très rapidement, le lien est fait entre l'état natif et l'activité biologique et l'abrogation de celle-ci dans des conditions favorisant l'état dénaturé. Ensuite, il fut constaté que plusieurs protéines, notamment l'hémoglobine et l'albumine, peuvent être alternativement dénaturées et renaturées. Cette transition est de type tout ou rien, signifiant que l'on retrouve toutes les molécules d'un échantillon dans l'un ou l'autre de l'état natif et de l'état dénaturé (*72*) (voir le **Chapitre 1 : Bases Théoriques**).

Les premières théories correctes de la dénaturation des protéines ont été énoncées dans les années 30 par Wu ainsi que par Mirsky et Pauling (*73;74*). Au cours des années subséquentes des expériences de dénaturation résumées dans un article de synthèse de la littérature de Tanford signale le caractère extrêmement coopératif du processus de dénaturation (*75*). De cet état de fait, il résulte que les courbes de dénaturation réalisées à l'état d'équilibre thermodynamique, adopte une forme sigmoïde dont la transition entre la forme native et dépliée est très aiguë. Tanford suggère l'explication suivante : l'empaquetage des chaînes latérales hydrophobes dans le centre d'une protéine est tel que le déplacement d'une seule d'entre elle de cet environnement aura un impact sur l'empaquetage des autres chaînes enfouies, et ainsi de suite menant inexorablement et rapidement à l'état dénaturé. Dans ces cas là, les méthodes de dénaturation et les propriétés physico-chimiques utilisées afin de suivre la réaction n'influe pas sur la nature de la courbe de dénaturation de sorte qu'elles sont ordinairement superposables (voir le **Chapitre 1 : Bases Théoriques**). Il appert selon les données de cette époque que les protéines complètement dénaturées perdent toute structure stable et adopte des propriétés semblables aux homopolymères classiques que l'on désigne par le vocable d'embobinage aléatoire

(« random coil »). Cet état est caractérisé par la fluctuation libre des angles ϕ et ψ et des autres angles des liens entre les atomes de la chaîne latérale. Il peut-être distingué de l'état natif par de nombreux changements des propriétés physiques et spectroscopiques comme le volume, le R_G, la viscosité, les propriétés spectrales et le spectre en dichroïsme cellulaire (DC). Tanford rapporte ainsi qu'en présence de hautes concentrations de guanidine d'hydrochlorure (Gdm-HCl) et d'urée, qui allaient devenir les dénaturants de choix pour étudier la réaction de repliement et de dépliement, les protéines testées démontrent des caractéristiques très proches de l'embobinage aléatoire. Aujourd'hui, grâce à des méthodes expérimentales beaucoup plus sensibles, il est possible d'affirmer que l'état dénaturé est probablement un peu plus structuré que le modèle de l'embobinage aléatoire strict (voir la section **L'état déplié et/ou dénaturé**).

Bien que la nature de l'état dénaturé puisse varier selon les méthodes employées pour l'obtenir, l'état natif démontre des caractéristiques très constantes quelque soit les conditions de dénaturation et de renaturation. Anfinsen proposa donc qu'il y ait une seule conformation native qui représente le minimum énergétique thermodynamique (*11*). En effet, si les protéines se repliaient vers un état métastable, la réaction de repliement devrait être beaucoup plus sensible aux conditions de dénaturation ou de la renaturation subséquente. Par conséquent, le mécanisme de repliement est robuste, accessible et conservé.

Le repliement des protéines peut être décrit par les mêmes théories, entre autre les lois de la thermodynamique, que les réactions chimiques simples. L'état dénaturé représente le réactif alors que l'état natif représente le produit. Dans le cas le plus simple, où seul ces deux états sont détectés :

$$D \Leftrightarrow N \quad \textbf{(1)}$$

l'état dénaturé est en équilibre avec l'état natif. Ces deux états sont aussi qualifiés de fondamentaux (« ground state »). L'état natif est favorisé en conditions physiologiques, parce qu'il est plus faible en énergie. La différence d'énergie entre les deux états

fondamentaux est appelée énergie libre ($\Delta G_{F\text{-}U}$). Ils sont séparés par une barrière d'activation[11] qui mène à la formation de l'état de transition, la hauteur de cette dernière déterminant le taux des réactions de repliement et de dépliement. Spécifiquement, les barrières d'activation entre l'état dénaturé ou l'état natif et l'état de transition sont appelées respectivement la différence d'énergie libre de repliement ($\Delta G_{U\text{-}‡}$) et la différence d'énergie libre de dépliement ($\Delta G_{‡\text{-}F}$) (**Figure 9**). En général, l'état natif des protéines est marginalement stable avec un $\Delta G_{F\text{-}U}$ habituellement entre 20-65 kJ.mol^{-1}, soit l'équivalent de trois ponts hydrogène et par conséquent, la transition comporte une large barrière relativement faible en énergie du moins dans des conditions physiologiques.

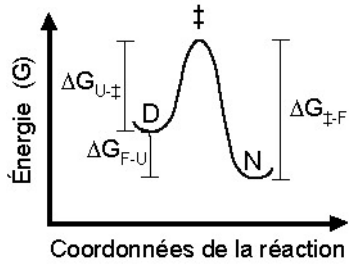


Figure 9. Diagramme d'énergie d'une réaction de repliement de type deux-états.
La stabilité d'une protéine est déterminée par la différence d'énergie entre l'état natif et l'état dénaturé, ($\Delta G_{F\text{-}U}$), ou simplement énergie libre. La différence d'énergie entre l'état dénaturé et l'état de transition ($\Delta G_{U\text{-}‡}$) et l'état natif et l'état de transition ($\Delta G_{‡\text{-}F}$) (i.e. respectivement énergie libre d'activation de la réaction de repliement et de dépliement) détermine respectivement le taux de la réaction de repliement et de dépliement, soit respectivement k_f et k_u.

Un autre type d'état, dit intermédiaire, peut intervenir dans les réactions de repliement plus complexes retrouvées chez certaines protéines, mais j'y reviendrai plus tard (voir section **Y-a-t-il des voies de repliement ou le mécanisme de repliement est-il séquentiel : introduction aux intermédiaires**). Par ailleurs, quelque soit la nature et le nombre d'états observables directement (i.e., deux-états, trois-états etc.; état dénaturé, intermédiaire et natif), le principe de la micro réversibilité stipule que les états de transition des réactions de repliement et de dépliement sont identiques et que les états rencontrés sur la voie de repliement sont les mêmes, ce qui veut dire que ces deux réactions se produisent

[11] Il y a des indices que certaines petites protéines/peptides (< 50 acides aminés) n'expérimenteraient pas de barrière d'activation significative lors de leur réaction de repliement (voir un exemple dans (*278*)).

par un processus exactement inverse l'un de l'autre. En pratique, cela implique que la dénaturation peut servir à obtenir de l'information sur la réaction de repliement.

L'état de transition

L'état de transition est le plus haut point en énergie sur un diagramme d'énergie d'une réaction chimique donnée qui montre la variation d'énergie libre de Gibbs en fonction des coordonnées de cette réaction (**Figure 9**). Il découle de cette description classique du diagramme d'énergie que la formation de l'état de transition est obligatoire à une réaction chimique, y compris celles menant à la formation et à la dénaturation de la structure native d'une protéine, et qu'elle en constitue l'étape limitante. En clair, les propriétés de l'état de transition, en particulier $\Delta G_{U-\ddagger}$ et $\Delta G_{\ddagger-F}$, respectivement dans le cas de la réaction de repliement et de dépliement, déterminent les taux respectifs de ces réactions (i.e. respectivement k_f et k_u). Le k_f correspond à l'inverse du temps de repliement moyen de la protéine. Il varie sur 6 ordres de grandeur pour les petites protéines (en général elles sont inférieures à 150 acides aminés), soit entre quelques microsecondes et plusieurs minutes.

Dans une réaction de repliement l'état de transition est relativement plus faible en énergie que lors d'une réaction chimique classique. Cela est principalement dû au fait qu'il n'y a pas rupture ni formation de liens covalents lors de la réaction de repliement, mais plutôt remodelage, établissement et renforcement d'interactions non-covalentes. Cette observation et les facteurs entropiques favorables discutés précédemment contribuent à faire de la réaction de repliement un processus spontané à des températures modérées (i.e. habituellement entre 5-45ºC). De plus, la nature de la réaction de repliement changerait les propriétés de la transition entre l'état dénaturé et l'état natif. Spécifiquement, cette dernière serait plus large que celles des réactions chimiques classiques, suggérant une nature plus dégénérée et moins rigide de la structure de l'état de transition. Dans une réaction de repliement de type deux-états, l'état de transition est la seule espèce dite de haute énergie (i.e., espèce dont l'énergie de Gibbs est supérieure à celle de l'état dénaturé) et la

connaissance de ces propriétés structurales permet de définir dans ses grandes lignes le processus/mécanisme de repliement. Par ailleurs, les propriétés énergétiques de l'état de transition le rendent instable et virtuellement impossible à observer directement. Par ailleurs, des méthodes indirectes faisant appel à des techniques de la biologie moléculaire et de l'étude du repliement peuvent être utilisées afin d'obtenir de l'information sur les propriétés de l'état de transition.

Ingénierie des protéines

L'amélioration des techniques de la biologie moléculaire va bouleverser l'étude du repliement des protéines en permettant de modifier de manière rationnelle leur structure primaire. La méthode d'ingénierie des protéines a été d'abord utilisée afin d'étudier le rôle de divers résidus dans l'activité enzymatique de la synthétase du tyrosyl-ARN_t (*76;77*). L'application de cette approche à l'étude du repliement permet de déterminer le rôle des interactions non-covalentes de la chaîne latérale d'un résidu donné dans la stabilisation de l'état de transition. Cette approche consiste à réduire progressivement la taille de la chaîne latérale d'un résidu donné en y introduisant des mutations non-disruptives telles que : I→V→A→G, D→E→A, T→S→A etc. Dans plusieurs études récentes, les mutations en alanine des résidus non-alanine et en glycine pour ces derniers sont utilisées afin de simplifier l'analyse comparative des résultats. Pour chacun des mutants générés, une valeur-Φ correspondant au ratio de la variation de l'énergie libre d'activation ($\Delta\Delta G_{U-‡}$ ou $\Delta\Delta G_{‡-F}$) sur la variation de l'énergie libre ($\Delta\Delta G_{F-U}$) est calculée à partir des données d'expériences thermodynamiques et cinétiques (voir le **Chapitre 1 : Bases Théoriques**). Ce paramètre mesurant la contribution des atomes qui ont été enlevés à la stabilisation de l'état de transition versus la stabilisation de la structure native, l'analyse globale des valeurs-Φ pour plusieurs résidus d'une protéine permet d'interroger la nature des contacts moléculaires, d'obtenir une estimation de leur importance relative et donc de déterminer le rôle de chaque région de la structure dans la stabilisation de l'état de transition. Plusieurs prémisses ont été énoncées dans le but de simplifier l'interprétation des

valeurs-Φ. Elles sont discutées en détail dans deux articles de Fersht et coll. (*78;79*). Mentionnons brièvement, la conservation des propriétés structurales de l'état natif et de l'état dénaturé et les caractéristiques énergétiques de ce dernier, de même que la nature native des contacts formés à l'état de transition. L'acceptation de la dernière prémisse revient à dire que les valeurs-Φ permettent de mesurer le ratio des contacts natifs formés à l'état de transition versus dans la structure native par les atomes en moins chez le mutant en comparaison avec le ts. Les états de transition tel qu'ils sont perçus par cette méthode arboreraient une structure plus ou moins dégénérée de l'état natif. Les deux premières applications de cette approche portèrent sur la protéine barnase (étude de l'intermédiaire (*80*) et de l'état de transition (*81;82*)) et sur CI2 (*83;84*) (voir respectivement la prochaine section et celle instituée

L'intermédiaire de repliement obligatoire de barnase; **Figure 10**). Encore aujourd'hui, ces études représentent le standard par excellence de cette approche tant par le nombre de mutants obtenus et la rigueur de l'analyse et de l'interprétation.

Prédiction de la structure de l'état de transition du repliement de CI2 et la théorie de nucléation-condensation

L'inhibiteur 2 de la chymotrypsine (CI2) est un domaine de 64 acides aminés, formant un cœur hydrophobe unitaire formé par la juxtaposition d'une hélice-α de 12 acides aminés sur un feuillet mixte de 3 brins-β (**Figure 10**), qui se replie via un mécanisme apparemment de type deux états (*85*), une primeur à l'époque. Sa renaturation/dénaturation a été suivie à l'aide de la fluorescence émise par le seul tryptophane, présent naturellement dans sa séquence. La photo instantanée de l'état de transition de CI2 révélé par l'analyse des valeurs-Φ obtenues des mutations de plusieurs résidus distribués à toute les régions de la structure indique que les contacts du cœur hydrophobe entre des résidus de l'hélice-α et du feuillet-β sont partiellement formés (*86*). Spécifiquement, Itzhaki et coll. proposèrent que les résidus A16 (hélice-α), I57 (brin-β) et L49 (brin-β) sont les résidus les plus importants à la stabilisation de l'état de transition,

bien que de toute évidence ils soient encore loin d'avoir établis tous leurs contacts natifs. D'autre part, l'hélice-α semble substantiellement structurée en amino-terminale alors que

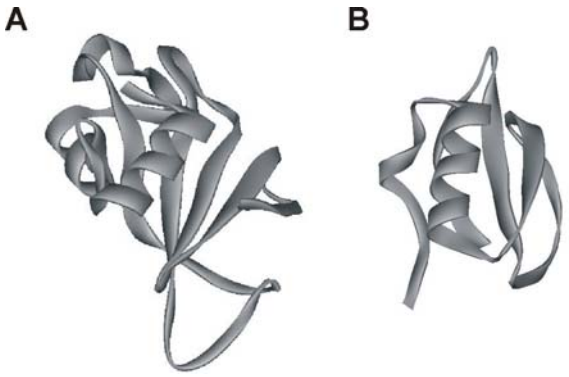


Figure 10. Représentation structurale schématique de deux protéines modèle utilisée pour l'élaboration de la méthode d'ingénierie des protéines.
A, Barnase. **B**, CI2.

la faible formation de contacts entre les résidus des brins-β indique que le feuillet-β n'a pas encore atteint un fort niveau de consolidation. Selon ce modèle, il semblerait donc que la structure secondaire et la structure tertiaire se forment de façon concomitante, donnant une apparence très concertée au mécanisme de repliement. Les trois résidus mentionnés ci-dessus font partie intégrante du cœur hydrophobe de CI2 et ils forment des contacts dans la structure native, des caractéristiques des résidus nucléus. Ce dernier terme identifie les résidus formant la fraction des interactions de type natif qui sont minimalement nécessaires à la stabilisation de l'état de transition (la définition originale du terme de nucléus est décrite à partir de simulations informatiques rudimentaires du repliement d'une protéine modélisée (*87*)). Dans le cas de CI2, le nucléus est donc constitué de contacts localisés dans l'hélice-α principale qui sont stabilisés seulement en présence des contacts de type natif et à longue distance établis entre A16 (résidu situé sur la face enfouie de cette hélice), d'une part et I57 et L49, d'autre part. Lorsque ces contacts sont formés, le reste de la chaîne polypeptidique dont la structure n'est pas encore formée peut se « condenser » autour de ce point focal pour atteindre finalement la structure native. Ce mécanisme de repliement correspond à la théorie de nucléation-condensation, qui est elle-même inspirée de la théorie

classique de nucléation-croissance (« nucleation-growth ») (*88*). Fersht précise cette théorie en énumérant entre autre ses implications (*86;89*) :

- Par convention, un résidu est considéré comme étant une partie intégrante du nucléus si sa mutation non-disruptive affecte le taux de repliement significativement et produit des valeurs-Φ élevées en plus de former un réseau de contact avec d'autres résidus nucléus potentiels dans la structure native de la protéine étudiée.

- Toutes les régions de la protéine participent à la stabilisation de l'état de transition, le nucléus rassemble seulement les éléments les mieux formés de la structure native.

- La présence d'intermédiaires et d'éléments de structure secondaire préformés à l'état dénaturé sont considérés comme étant un obstacle au repliement rapide.

- La présence de contacts non-natifs à l'état dénaturé est défavorisée évolutivement, car cela le stabiliserait.

- La théorie de nucléation-condensation rejette la présence des contacts formés par les résidus nucléus dans l'état dénaturé et suggère qu'ils sont formés uniquement à l'état de transition. Cela entre en contradiction avec le modèle séquentiel de repliement (voir la section **Y-a-t-il des voies de repliement ou le mécanisme de repliement est-il séquentiel : introduction aux intermédiaires**).

À cause de la découverte de la propriété deux-états du repliement de CI2, les études de repliement seront dorénavant axées principalement sur les protéines qui partagent cette caractéristique. De nombreux domaines et protéines ont été soumis à une analyse des valeurs-Φ d'un nombre substantiel de leurs résidus (**Tableau II**). La théorie de nucléation-condensation a été invoquée pour modéliser le repliement de plusieurs de ces protéines (voir ces références par exemple : (*90-95*)). Par ailleurs, cette théorie ne semble pas suffisamment générale pour être applicable à toutes les protéines, entre autre celles adoptant un état de transition polarisé, dans lequel une région, par exemple un élément de structure secondaire, est plus consolidée que le reste de la structure. Plusieurs des implications du modèle ont maintenant été remises en cause par des études portant sur d'autres protéines, mais même des études plus poussées sur CI2, en particulier dans des simulations informatiques et des études sur la structure résiduelle de l'état dénaturé, les ont ébranlées (*96*) (voir la section **L'état déplié et/ou dénaturé**).

Tableau II. Quelques exemples de domaines et protéines soumis à l'analyse des valeurs-Φ

Domaines modèle	Code PDB	Références
Ubiquitine†	1UBI	(*97*)
Protéine-G¶	2GB1	(*98*)
Protéine-L†	1HZ6	(*99*)
Domaine amino-terminal de L9	1DIV	(*100*)
Protéine du choc froid	1CSP	(*101*)
Domaine B/protéine-A	1SS1	(*102*)
Im7†	1AYI	(*103*)
Im9	1E0Ha	(*104*)
ACBP†	2ABD	(*105*)
WW de pin	1I8H	(*106*)
SH3 de src	1SRL	(*107*)
SH3 de la spectrine†	1SHG	(*108*)
SH3 de fyn¶	1FYN	(*109*)
Sso7d¶ (≈ SH3)	1SSO	(*110*)
Titine	1TIT	(*95*)
TNfn3	1TEN	(*93*)
FNfn10	1FNA	(*94*)
Domaine 1 de CD2	1HNF	(*111*)
CI2¶	2CI2	(*86*)
ADA2h¶	1AYE	(*91*)
Protéine ribosomale S6	1RIS	(*112*)
U1A (RNP-A)	2U1A	(*113*)
AcP de muscle¶	1APS	(*114*)
FKBP-12¶	1FKB	(*92*)
Répresseur λ	1LLI	(*115*)
Barnase†	1A2P	(*82*)
Barstar	1BTB	(*116*)
CheY†	1EHC	(*90*)
Cyt b562¶	1APC	(*117*)
Répresseur-arc	1ARR	(*118*)
Domaine tetramérique de p53	1AIE	(*119*)
Villine 14T	2VIK	(*120*)
CKS1	1QB3	(*121*)

¶ État de transition diffus *(124)*
† État de transition polarisé *(124)*

Il ne fait aucun doute que la méthode d'ingénierie des protéines a été d'un apport gigantesque à la compréhension du processus de repliement. Par exemple, elle a permis l'élaboration de nouvelles théories, tel que celle mettant l'emphase sur le rôle de la topologie native dans la définition du processus de repliement et elle a révélé la diversité des caractéristiques des états de transition, qu'ils soient localisés dans une région ou bien étendus à toute la structure. De la même manière, elle a permis la caractérisation d'intermédiaires obligatoires pour des protéines se repliant par une réaction trois-états. Finalement, cette méthode permet de comparer la structure approximative de l'état de

transition de nombreuses protéines partageant la même topologie et de comparer les topologies entre elles.

Graphe de Leffler

État de transition polarisé versus diffus

Le graphe de Leffler (alias graphe de Brønstead), qui nous vient d'application en chimie, permet de vérifier la variation du taux de la réaction (i.e., spécifiquement $\ln k_f$ ou $\ln k_u$) en fonction de la stabilité (en fait $\Delta\Delta G_{F\text{-}U}/RT$)[12] dans une protéine mutante versus sa contrepartie de ts (voir le **Chapitre 1 : Bases Théoriques** pour de plus amples détails). Habituellement, la corrélation entre ces deux paramètres est linéaire. La pente de cette corrélation permet de déterminer la valeur-Φ moyenne de toute la structure ou de sous-éléments en fonction de la dispersion de points obtenue. Cela permet de déterminer des caractéristiques brutes de l'état de transition : est-il polarisé ou diffus? Quelles régions participent à sa stabilisation? Quel est le niveau moyen d'implication dans la stabilisation de l'état de transition des résidus mutés? Etc. Les deux cas de figure principaux découlent encore cette fois des études originales sur les prototypes CI2 et barnase. Tout d'abord, dans le cas de CI2, l'effet de la mutation d'un résidu donné ne dépend pas de sa localisation et la relation entre tous les mutants obtenus est bien modélisée par une fonction linéaire (*86;122*). Il s'agit donc dans ce cas d'un état de transition dit diffus et la valeur-Φ moyenne obtenue est de 0,7 indiquant que l'état de transition est proche structuralement de l'état natif. Les observations qu'aucun élément de structure secondaire se démarque de la corrélation linéaire et que des fragments de CI2 ne forment pas de structure secondaires ont conduit à l'élaboration de la théorie de la nucléation condensation telle qu'elle a été stipulée plus haut (voir la section précédente pour les détails). Par ailleurs, les mutants de la protéine barnase se comportent différemment, en se ségrégant selon leur disposition dans la structure native, suivant leur position dans tel ou tels éléments de structures secondaires

[12] Inversement, de $\Delta G_{U\text{-}‡}$ ou $\Delta G_{‡\text{-}F}$ versus $\Delta\Delta G_{F\text{-}U}$.

(*122;123*). Ainsi les mutants de résidus localisés dans l'hélice-α principale démontrent une pente près de 1, suggérant sa formation quasi-complète à l'état de transition. Il est intéressant de noter que cette protéine se replie via un intermédiaire placé tout juste avant l'état de transition et que cette hélice-α y est déjà fortement structurée, mais nous y reviendront plus loin (voir la section **L'intermédiaire de repliement obligatoire de barnase**). Par ailleurs, les autres structures secondaires sont apparemment moins structurées dans cet état. Ce type d'état de transition est donc qualifié de « polarisé ». L'attribution du qualificatif de diffus et de polarisé n'est pas systématique et est difficile à définir dans les cas limites. Néanmoins, Sanchez et Kiefhaber ont classifié suivant cette simple nomenclature les états de transition de plusieurs protéines dont les structures ont été déterminées par une analyse des valeurs-Φ (*124*). Les états de transition des protéines classifiées selon cette nomenclature simple sont annotés dans le **Tableau II**.

Certaines mutations ou perturbations chimiques ou physiques peuvent mener à des déviations dans la linéarité de la corrélation entre le taux de la réaction de repliement ou de dépliement et $\Delta\Delta G_{F\text{-}U}/RT$ pour une protéine ou un segment adoptant une structure secondaire donnée. Ce type d'effet est ordinairement attribué à la présence de: voies de repliement parallèles ou alternatives (comportement d'anti-Hammond); déplacement de l'état de transition sur une large barrière d'énergie (comportement d'Hammond); changement d'état de transition dû à la présence d'intermédiaires sur une voie séquentielle de repliement (voir la section **Intermédiaires**); changement des propriétés structurales de l'état natif et dénaturé (tirer de (*125*)) (voir la section **L'état déplié et/ou dénaturé**).

Comportement d'Hammond et d'anti-Hammond

Selon le comportement d'Hammond, une perturbation déstabilisante provoque le mouvement de l'état de transition vers l'état fondamental le plus déstabilisé. La plupart du temps ce mouvement se fait vers l'état natif, car cet état fondamental est plus sensible aux perturbations que l'état dénaturé ou bien que d'éventuels états intermédiaires. Cet effet correspond à un déplacement parallèle aux coordonnées de la réaction de l'état de

transition. Cela indique que le déplacement se fait sur une large barrière en énergie. Originellement, un tel comportement fut décrit pour le repliement de barnase et de CI2 (*123;126*).

Le comportement d'anti-Hammond indique qu'une mutation déstabilisante déplace l'état de transition perpendiculairement aux coordonnées de la réaction et vers l'état dénaturé. Cela indiquerait la présence de voies parallèles de repliement. Un tel comportement a été décrit pour l'hélice-α principale de barnase (*122;123*) et il se traduit par une diminution de la pente de la corrélation de $\Delta G_{U-\ddagger}$ versus $\Delta\Delta G_{F-U}/RT$ pour les mutations dont $\Delta\Delta G_{F-U}/RT > 4$. Par ailleurs, le comportement d'anti-Hammond apparaît extrêmement rare et il n'aurait été observé que chez quelques autres protéines (voir la section **Entonnoir et multiplicité des voies de repliement**) (*125;127*).

Le mouvement de l'état de transition induit par une mutation individuelle est détecté par la variation du paramètre β-Tanford (β_t), qui mesure la position de l'état de transition par son degré d'exposition relativement au solvant en comparaison avec l'état natif et l'état dénaturé à partir d'un calcul impliquant des paramètres cinétiques et thermodynamiques (voir le **Chapitre 1 : Bases Théoriques**). Toutefois, tel que le rapporte une récente analyse des données de la littérature qui révèle la rareté de protéines démontrant un comportement d'Hammond malgré une variation appropriée de β_t, il importe de s'assurer que tout mouvement potentiel de l'état de transition n'est pas dû à un changement dans les propriétés structurales des états fondamentaux (*125*). Je rediscuterai de cela un peu plus loin (voir section **L'état déplié et/ou dénaturé**).

Les protéines partageant la même topologie se replient-elles par un mécanisme identique : oui et non

Étant donné que la structure d'un état de transition comporte plusieurs caractéristiques et contacts entre les chaînes latérales en commun avec la structure native, il en découle que la topologie générale de ce dernier état devrait déterminer celle du premier.

Un corollaire de ce raisonnement est donc que les protéines possédant une topologie commune devrait se replier via des états de transition semblables. Dans ce cas, jusqu'à quel point cela pourrait-il être vrai pour des protéines ne possédant qu'une très faible identité de séquence, par exemple? Deuxièmement, quel niveau de divergence structurale entre deux analogues structuraux peut être accepté sans remettre en cause la similitude de leur mécanisme de repliement? La comparaison des caractéristiques de l'état de transition de telles protéines pourrait permettre de définir les déterminants de la séquence essentiels à la définition du mécanisme de repliement.

Les taux de repliement d'orthologues appartenant à deux familles distinctes, respectivement celles de la protéine liant l'acyl-CoA (ACBP) et de la protéine du choc au froid de *Bacillus Subtilis* (CsPB) (i.e. approximativement 50 et 80% d'identité de séquence, respectivement), et donc ayant en commun des liens évolutifs et fonctionnels forts furent comparés dans des études distinctes (*128;129*). Celles-ci démontrèrent que plusieurs des paramètres cinétiques du repliement étaient conservés quoique la stabilité et les taux de repliement/dépliement puissent être très variables entre chaque orthologue. Ainsi, le caractère deux-états était généralement conservé ainsi que la position de l'état de transition par rapport à l'état natif et l'état dénaturé (voir le **Chapitre 1 : Bases Théoriques**). Par ailleurs, l'utilisation de modèles protéiques présentant une identité de séquences beaucoup plus faible est préférable afin d'attaquer correctement l'hypothèse de la contrainte topologique de l'état de transition, en ce sens que les positions qui sont conservées dans ce contexte joueraient un rôle potentiellement important dans l'encodage de la structure. L'étude de Sato et coll. est un bel exemple d'études comparatives du repliement de protéines à la structure analogue (*130*). Dans cet article, ils ont comparé le repliement de deux domaines amino-terminaux de protéines L9 issus de deux bactéries à un domaine semblable de la ribonucléase H1 et montré que la cinétique de leur repliement était liée à leur stabilité respective. Le DLR de Raf et ubiquitine ont une identité de séquence inférieure à 12%, tout en partageant des similitudes structurales évidentes. Les résultats obtenus dans notre laboratoire ont indiqué une sensibilité équivalente du k_f de ces protéines à la température, aux perturbations chimiques et à la mutation d'un résidu équivalent (*131*).

Dans le cadre de mes travaux de recherche, je compare plus en profondeur les mécanismes de repliement de ces deux protéines sur la base de l'analyse des valeurs-Φ que nous avons réalisée sur le DLR de Raf et celles qui ont été publiées récemment dans le cas d'ubiquitine (voir le **Chapitre 2 : Résultats**) (*97*).

Des études basées sur cette même approche ont démontré la similarité chez des protéines partageant la même topologie de la distribution des résidus impliqués dans la stabilisation de l'état de transition. En particulier, trois domaines SH3 soit ceux de la α-spectrine, de src et de fyn partagent un rôle dominant et remarquablement similaire de deux brins centraux de la structure (alias le « distal β-hairpin ») (*107-109;132;133*) (**Figure 11**). Bien qu'appartenant à la même famille structurale, les domaines SH3 de src et α-spectrine, par exemple, ne démontrent que 33% d'identité de séquence, cependant que leur similarité topologique et leur lien fonctionnel souligne leur relation évolutive proche.

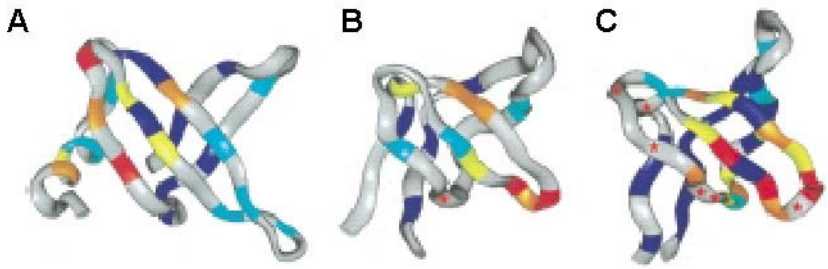


Figure 11. Comparaison des valeurs-Φ observées chez 3 membres de la topologie structurale des domaines SH3.
A, Sso7d. **B**, Domaine SH3 de l'α-spectrine. **C**, Domaine SH3 de src. Code de couleurs : bleu (0-0,2) ; cyan (0,2-0,4) ; jaune (0,4-0,6) ; orange (0,6-0,8) ; rouge (0,8-1,0). Les étoiles rouges indiquent les sites où des contacts non-natifs pourraient se former à l'état de transition. Figure adaptée de Guerois et coll. (*110*), avec la permission des auteurs responsables de la correspondance et des autorités compétentes du journal scientifique en question.

Les protéines-L et G adoptent la topologie d'ubiquitine et appartiennent plus particulièrement à la superfamille des protéines liant les immunoglobulines. Cette superfamille telle que discutée dans l'introduction et dans mes articles est atypique et ne semble pas avoir de liens évolutifs avec les superfamilles similaires à celle d'ubiquitine. Une de leur particularité qui est au centre de l'argumentaire de ces études est que la structure de la protéine-L et G adopte une topologie symétrique (*60;61;98;99;134*). Dans le

cas de la protéine-L le tour-β en amino-terminal joue le rôle le plus important dans la stabilisation de l'état de transition tel qu'il est prédit par le résultat de l'analyse des valeurs-Φ, alors que c'est le tour-β en carboxy-terminal qui joue se rôle prédominant pour la protéine-G. De plus, il a été démontré que les rôles respectifs des deux tours-β au cours du repliement de la protéine-G peuvent être inter changés par l'introduction de 11 mutations sélectionnées grâce à une procédure informatique (*135*). En clair, cela signifie que la structure de l'état de transition de ce variant de la protéine-G adopte une conformation plus similaire à celle de la protéine-L. Cela pourrait signifier que nonobstant la position divergente des éléments topologiques cruciaux au mécanisme de repliement de ces protéines, la symétrie de leur topologie les rend ardu à discriminer uniquement de ce point de vue. Ce dernier argument des auteurs pourrait être sujet à discussion. D'ailleurs, les valeurs-Φ des autres résidus de ces deux protéines les distinguent tel que cela a été remarqué plus récemment (*124*). Des différences notables dans la structure de l'état de transition des protéines et domaine U1A, S6 et le domaine pro-carboxypeptidase (ADA2H), qui adoptent la topologie de l'acylphosphatase (AcP), ont aussi été notées, alors que l'ADA2H et l'AcP musculaire (13 % d'identité de séquence) adoptent des mécanismes de repliement semblables (**Tableau II** pour les références appropriées). Finalement, un autre cas très intéressant de dissemblance significative dans la structure de l'état de transition porte sur la protéine sso7d, un analogue structural des domaines SH3 et démontrant une homologie de séquence non-significative (identité de séquence< 7%), dont le nucléus de l'état de transition est déplacé du troisième vers le deuxième motif épingle à cheveux (*110*) (**Figure 11**). Des variations structurales évidentes ont été évoquées pour expliquer ces différences (*136*). Quelques autres comparaisons intéressantes et une discussion sur ce sujet peuvent être retrouvées dans ce dernier article qui regroupe une synthèse de la littérature sur ce sujet. En tout et pour tout, il y a à ce moment-ci, 7 paires de protéines dont les caractéristiques structurales de l'état de transition ont pu être comparées de manière détaillée en utilisant l'analyse de leurs valeurs-Φ. Je reviendrai plus tard dans la section sur les **Intermédiaires** sur le cas spécifique de deux protéines homologues

démontrant le même état de transition, malgré des divergences quant au nombre d'états détectables sur leur voie de repliement.

En conclusion, ces études ont démontré des contraintes structurales importantes à l'état de transition de ces protéines, et qui sont centrées sur des éléments cruciaux et particulièrement simples dans la définition de la topologie d'une structure : les tours-β. Comme les résidus situés dans ces éléments de structure secondaire forment des contacts strictement locaux, leur rôle dans la stabilisation de l'état de transition ne peut être réconcilié à la définition classique de la théorie de nucléation-condensation. Ces études réhabilitent une vision plus séquentielle du processus de repliement pour les protéines se repliant en deux-états. Par ailleurs, elles indiquent que les caractéristiques de l'état de transition ne sont pas toujours conservées chez les analogues structuraux et qu'elles ont tendance à varier à mesure que l'identité de séquence et de petites différences structurales s'accumulent. Il est raisonnable de supposer à la lumière des connaissances acquises jusqu'à maintenant que chaque topologie a quelques sites potentiels permettant de consolider l'état de transition et donc de former la structure native. Il m'apparaît raisonable de prédire que les topologies moins fréquentes seraient moins flexibles à ce niveau là.

Topologie et ordre de contact

Les travaux de Plaxco et coll. ont mis au jour une corrélation significative entre le taux de repliement de la plupart des protéines[13] se repliant en deux états et l'ordre de contact relatif (OC_r), une mesure indirecte de la topologie de la structure native :

$$OC_r = \frac{1}{L \cdot N} \sum^{N} \Delta S_{i,j} \qquad \textbf{(2)}$$

où L est la longueur de la chaîne polypeptidique, N est le nombre de contacts et $\Delta S_{i,j}$ représente le nombre de résidus dans la chaîne polypeptidique séparant les résidus i et j qui

[13] A l'époque, 24 protéines se repliant en deux-états ont été utilisées pour démontrer la bonne corrélation de $\ln k_f$ avec l'OC_r, par rapport à d'autres paramètres caractéristiques de la structure (*138*)

forment un contact à l'état natif[14] (**Figure 12**) (*137-142*). En simple, ce que cette corrélation implique c'est que plus les résidus formant des contacts natifs dans la structure d'une protéine sont éloignés l'un de l'autre, plus l'OC_r est élevé et le taux de repliement serait lent. Dans ces propres études, le laboratoire de Dill a proposé un mécanisme trivial par lequel l'OC_r pouvait expliquer la coopérativité et déterminer le mécanisme de repliement des protéines (*143;144*).

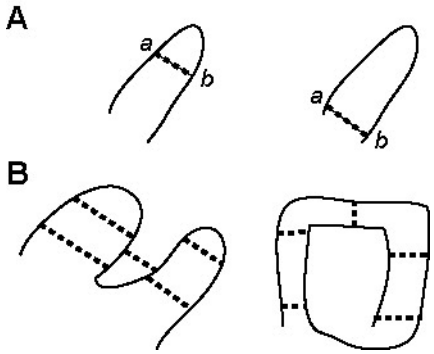


Figure 12. Démonstration empirique de la notion de l'ordre de contact.
A, Le contact (lien pointillé entre le point *a* et *b*) est à moins longue distance dans le panneau de gauche que dans le panneau de droite, car la boucle les séparant est plus courte dans ce cas. **B**, Protéines schématiques de taille égale, mais les contacts établis par celle dans le panneau de gauche sont plus locaux, faisant en sorte que son OC_r est plus petit que celui du panneau de droite.

La superfamille de l'AcP comprend des protéines et des domaines extrêmement divergents tel que l'AcP, la S6 et l'ADA2H dont les taux de repliement démontrent une corrélation particulièrement bonne avec ceux extrapolés de l'OC_r (*114*). Des résultats obtenus sur des permutants circulaires de la protéine S6 indique que le taux de repliement pourrait être déterminé par la stabilité de la structure native (*145*), une suggestion qui est soutenue par d'autres données (*146;147*) (voir la prochaine section). Il y a des exceptions par ailleurs, des domaines dont le taux de repliement n'est pas prévu adéquatement par l'OC_r. Il est possible que dans ces cas, il y ait des détails ou des propriétés supplémentaires spécifiques à la structure des ces protéines qui contribueraient à frustrer la formation de la structure native. Par exemple, des auteurs ont souligné que le contenu relatif en acides aminés

[14] La définition même d'un contact entre deux résidus pourrait être débattue. Dans cette étude les auteurs ont fixé un seuil de 6 Å de distance entre les atomes, en ne considérant pas les H de deux résidus non consécutifs. Ils n'ont pas noté de différence significative entre 3,5 et 8 Å.

hydrophobes pouvait aussi affecter le taux de repliement (*148*). Le DLR de Raf et ubiquitine sont aussi des protéines pour lesquelles le taux de repliement prédit à partir de l'OC_r est inférieur d'un ordre de grandeur à la valeur déterminée expérimentalement (*131*). La qualité prédictive de l'OC_r pour déterminer le taux de repliement souffre peut-être aussi de la diversité des méthodes et des conditions utilisées pour le mesurer expérimentalement.

En résumé, ces travaux ont démontré que la topologie de la structure native d'une protéine, par opposition aux détails de la séquence et des interactions au niveau atomique, détermine les grandes lignes de son mécanisme de repliement puisqu'il suffit pour obtenir un estimé correct du k_f d'une protéine de calculer son OC_r.

Les séquences des protéines ne sont pas optimisées pour la vitesse de repliement et pour la stabilité de la structure native

Plusieurs données récentes obtenues entre autres par l'utilisation de librairies dégénérées, de la méthode d'ingénierie des protéines et de modélisation informatique sur la protéine-L ou des domaines SH3 suggèrent que la séquence native n'est pas optimisée pour la vitesse de repliement (*59;61;109;133;147;149*). Par exemple, dans le cas de la protéine-L environ 50 % des clones obtenus par l'insertion aléatoire de codons dégénérés à des segments de résidus contigus (entre 5 et 11 acides aminés) ont un taux de repliement plus élevé que la séquence native. Par contre, aucun des mutants n'est plus stable que le ts. Ces observations se sont avérées pour les domaines SH3 étudiés. En bref, ces travaux sont en accord avec un modèle où la stabilité de la structure native et la fonction sont mieux optimisées que le taux de repliement.

Cette conclusion est en contradiction avec des travaux théoriques, entre autre de Shakhnovich et coll. ayant porté sur la conservation privilégiée au cours de l'évolution des résidus nucléus comparativement aux autres résidus du cœur hydrophobe, et en corollaire de cela, de l'optimisation du taux de repliement (*150;151*). Bien qu'une grande part de la controverse entre ces groupes viennent de différence entre leur définition respective de la

notion de conservation, leur perspective sur le mécanisme de repliement demeure très différente (*152-154*). Il semble assez clair maintenant qu'il n'y ait pas de corrélation entre le niveau de conservation d'un résidu et sa valeur-Φ. Une étude expérimentale et une étude informatique sur des domaines SH3 suggèrent que la stabilité est le facteur prédominant de la pression évolutive, particulièrement dans des conditions expérimentales où les contraintes nécessaires à la conservation de la fonction sont absentes (*146;147*). Par ailleurs, les données sur le domaine SH3 de Fyn indiquent une optimisation plus large de la séquence native afin de préserver la fonction de liaison à son substrat peptidique, incluant aussi des résidus important pour la formation et la stabilisation de la structure, en particulier les résidus du cœur hydrophobe faisant partie du nucléus (*147*). Il semblerait donc qu'un niveau de stabilité 'plancher' est maintenu au cours de l'évolution de la séquence des protéines globulaires et que cela serait suffisant pour déterminer leur taux de repliement.

Dans la nature, un compromis entre la stabilité et la fonction est atteint de telle sorte que la plupart des protéines sont probablement loin d'être aussi stable qu'elles le pourraient en théorie (*4*). En effet, des expériences récentes de design *de novo* de dix protéines globulaires dont la structure était calquée sur leur contrepartie naturelle dans un contexte où la fonction n'est pas conservée ont démontrées dans quatre des modèles une augmentation claire de la stabilité atteignant jusqu'à 7 $kCal.mol^{-1}$, soit une augmentation de l'ordre de 110% à 290% suivant les cas (*4*). À l'opposé, des expériences de design de domaines WW qui reposaient sur des alignements de séquences naturelles ont révélé des niveaux de stabilité de la structure native des variants artificiels comparables à leur équivalent naturel dans un contexte où la fonction de liaison à leur ligand peptidique était conservée (*155;156*).

Y-a-t-il des voies de repliement ou le mécanisme de repliement est-il séquentiel : introduction aux intermédiaires

L'observation des premières structures révèle que la conformation des acides aminés est limitée par des contraintes stériques (**Figure 3**) (*157*). Ramachandran et coll. observent aussi que les propriétés de la chaîne latérale d'un acide aminé telles que sa ramification et la nature des atomes dont il est composé influent sur les combinaisons d'angles permises. Malgré la limitation du nombre de conformations, les possibilités combinatoires de celles-ci demeurent innombrables.

À partir d'une argumentation logique, Cyrus Levinthal énonce lors d'une conférence, ce qui est maintenant connu sous le nom du paradoxe de Levinthal (traduction libre de la transcription par A. Rawitch d'un séminaire présenté par Cyrus Levinthal) (*158*):

> « Si l'on prend une protéine de 100 résidus et que l'on suppose que chaque résidu peut adopter trois conformations (angle ϕ et ψ typique d'une hélice-α, d'un feuillet-β ou bien d'une région désordonnée), on peut donc dire qu'il y a 3^{100} ou $1x10^{48}$ états possibles. Si l'on suppose que la protéine peut explorer 1 état par vibration moléculaire, et que celles-ci se produisent à un taux d'une par femtoseconde, on en déduit que l'exploration de tous les états possibles prendra $1x10^{48}$ femtosecondes ou $1x10^{33}$ secondes. Comme, il y a $1x10^{8}$ secondes dans une année, l'exploration de tous ces états prendraient $1x10^{25}$ années, soit plus d'années qu'en compte l'univers ».

La conclusion triviale tirée de ce raisonnement logique est donc que les protéines se replient vers la structure native par un phénomène qui permet d'éviter une recherche exhaustive de toutes les conformations possibles. Cela concorde avec les résultats expérimentaux de l'époque qui suggèrent que les protéines puissent se replier en quelques minutes (*11;159;160*), une constatation qui ne s'est que renforcée avec la découverte des protéines ou des domaines se repliant en deux-états (voir section **L'état de transition**). Il faut alors déterminer les mécanismes de la coopérativité qui permettent de passer outre cette recherche exhaustive.

Ces arguments mènent logiquement à la naissance de ce concept nommé voies de repliement (« folding pathways »). Dans un article fréquemment cité, Levinthal expose une théorie du repliement qui impliquerait la formation progressive de la structure native par la consolidation de structures locales, qui pourraient débuter à se structurer dès l'émergence du ribosome de la chaîne polypeptidique naissante. Ainsi, la condensation ou l'initiation de la formation de structures dans un segment de séquence entraînerait à son tour des changements du même genre dans une autre région du polypeptide et ainsi de suite jusqu'à l'obtention de la structure native (*161*). Pousser à son extrême le principe de voie de repliement prédit que la structure native d'une protéine est formée par la formation d'une série d'états intermédiaires suivant un ordre déterminé.

Les premières études de dénaturation et de renaturation de protéines (i.e., ferrihémoglobine et lysozyme entre autre) ont été réalisées dans les années 50 et 60 en utilisant des variations du pH ou de température (réviser dans (*75*)). Les protéines dénaturées dans ces conditions subissent souvent une seconde transition lorsqu'elles sont transférées dans des dénaturants plus forts tels que le Gdm-HCl ou l'urée. Par conséquent, les espèces obtenues en conditions douces de dénaturation possèderaient des caractéristiques à la fois de l'état natif et de l'état dénaturé. L'hypothèse est donc qu'elles partageraient des similarités avec des états intermédiaires formés au cours de la réaction, i.e. sur la voie de repliement.

L'amélioration des appareils et des techniques de mixage rapide en permettant de réduire le laps de temps qui s'écoule entre le mixage et le début de la mesure du repliement ainsi que son couplage avec diverses méthodes spectroscopiques de détection telles que le DC dans l'ultraviolet (UV) et la fluorescence ont permis la découverte d'intermédiaires cinétiques. Durant les années 1970-80, une série d'observations sur la cinétique du repliement de la β-lactamase, de la ribonucléase-A et de l'anhydrase carbonique-B avait suggéré la formation rapide d'intermédiaires possédant des éléments de structures secondaires natifs, mais peu de contacts à longue distance qui sont plutôt le propres de la structure native des protéines globulaires.

Intermédiaires

Comme je l'ai souligné plus haut une conséquence de la théorie des voies de repliement est que la formation de l'état natif à partir de l'état dénaturé nécessiterait la formation d'intermédiaires à la structure de plus en plus semblable à l'état natif:

$$U \Leftrightarrow I_h \Leftrightarrow \ldots \Leftrightarrow I_t \Leftrightarrow F \quad \textbf{(3)}$$

Une conclusion triviale qui peut être tirée de l'équation **(3)** est que les intermédiaires devraient démontrer des caractéristiques à la fois des états dénaturé et natif, la similarité supérieure avec l'un ou l'autre de ceux-ci dépendant de la position de l'intermédiaire en question sur le profil de la réaction. Mais quelle pourrait bien être la nature de ces intermédiaires? La préséance de la formation des structures locales dans cette théorie prédit que l'intermédiaire hâtif (I_h) devrait former des éléments de structures secondaires dans certaines régions, mais il serait largement dépourvu de tout contact à longue distance. D'autre part, l'intermédiaire tardif (I_t) typique adopterait une version légèrement dégénérée de la structure native qui démontrerait comme distinctions principales des fluctuations supérieures des chaînes latérales des acides aminés et un plus grand accès du solvant au cœur hydrophobe. Ces descriptions très brèves permettent d'introduire les prochaines sections qui présenteront les preuves de la présence de divers types d'intermédiaires. Par ailleurs, cette vision de la voie de repliement est tombée en désuétude, car elle a peu ou pas du tout été confortée par des données expérimentales. En effet, un nombre élevé de protéines se replie en deux-états et de toute évidence ne rencontrent pas d'intermédiaires majeurs successifs sur leur voie de repliement. Généralement, la littérature est plus en accord avec la présence d'un nombre restreint d'intermédiaires, si ce n'est aucun, au cours du processus de repliement.

Par ailleurs, même dans le contexte où il y a présence d'intermédiaires, il reste à vérifier qu'ils se trouvent vraiment sur la voie de repliement afin de démontrer leur nécessité à la formation de la structure native. Pour accomplir cette démonstration, ce type d'intermédiaires doit être distingué de ceux qui sont contre-productifs à la réaction de

repliement et conséquemment ne se retrouvent pas sur la voie de repliement (i.e., entre l'état dénaturé et l'état natif) ou de formes distinctes de l'état dénaturé qui pourraient se replier à des taux différents :

$$I \Leftrightarrow U \Leftrightarrow F \qquad \textbf{(4)}$$

$$U_R \Leftrightarrow F \qquad \textbf{(5)}$$

$$U_L \Leftrightarrow F$$

$$U \Leftrightarrow I \Leftrightarrow F \qquad \textbf{(6)}$$

où U_R et U_L sont des conformations de l'état dénaturé se repliant rapidement et lentement, respectivement. Une cause courante de ce dernier phénomène (équation **(5)**) est la présence de l'isomère non-natif d'une ou plusieurs prolines. Le cas de figure décrit à l'équation **(4)** évoque la présence de structures non-natives dans les conditions initiales de l'expérience de repliement. La déstabilisation de ce type d'intermédiaires par l'ajout de dénaturant devrait provoquer une accélération du repliement. La démonstration de la présence d'intermé-

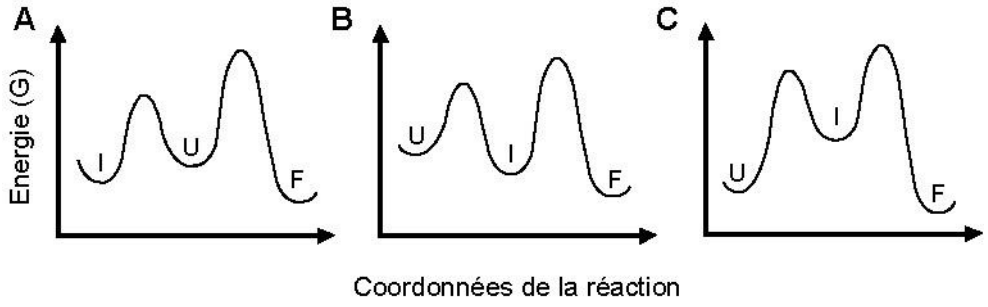


Figure 13. Modèle de diagramme d'énergie de réactions de repliement comportant un intermédiaire.
A, Diagramme d'énergie d'une réaction de repliement de type trois-états avec formation d'un intermédiaire cul-de-sac. **B**, Diagramme d'énergie d'une réaction de repliement de type trois-états avec formation d'un intermédiaire. **C**, Diagramme d'énergie d'une réaction de repliement avec formation d'un intermédiaire de haute énergie. Le dernier type d'intemédiaires **(C)** est moins stable que l'état dénaturé contrairement aux exemples présentés dans les deux panneaux précédents. Cela signifie que ces intermédiaires seront formés de manière plus transitoire et devraient être plus difficiles à observer. La stabilité de ce type d'intermédiaires dépendra aussi de la hauteur de la barrière d'activation de la transition I→F.

diaires authentiques (i.e. différent des cas de figure décris ci-dessus) sur la voie de repliement indique que la réaction de repliement analysée est de type trois-états ou plus en fonction du nombre d'intermédiaires détectés (équation **(3)** et **(6)**). Pour ces raisons, à chaque fois qu'un nouvel intermédiaire est découvert, il est de prime importance de

déterminer s'il se trouve sur (équation **(6)**) ou bien hors de la voie de repliement (équation **(4)**). Les intermédiaires sur la voie de repliement correspondant au modèle de l'équation **(6)** peuvent être plus ou moins stables, soit respectivement de faible ou de haute énergie relativement aux états fondamentaux (**Figure 13** et les sections ci-dessous). Par ailleurs, le rôle, la présence généralisée, voire la nécessité des intermédiaires aux mécanismes de repliement ou à leur processivité – à savoir si les intermédiaires sont généralement productifs, contre-productifs, un artéfact de certaines réactions de repliement ou bien s'ils se forment généralement avant ou après l'état de transition – sont toujours le sujet d'âpres débats.

Présences d'intermédiaires : preuves thermodynamiques et cinétiques

La présence d'intermédiaires peut être suspectée lorsque les courbes de dénaturation démontrent plus d'une transition coopérative ou qu'elles divergent en fonction des paramètres mesurés pour suivre la réaction de dénaturation d'une protéine donnée. Dans le premier cas, c'est un indice de la présence d'un intermédiaire stable à certaines conditions à l'équilibre. En second lieu, les courbes de dénaturation de la β-lactamase obtenues par la mesure de la fluorescence ou du DC dans l'UV de haute longueur d'onde montre une transition à plus faible concentration que celle réalisée par DC dans l'UV de basse longueur d'onde (*162*). Ces deux types de sondes permettant respectivement d'estimer le changement au niveau des structures tertiaires et secondaires d'une protéine dans un contexte donné, les résultats contradictoires obtenus dans le cas mentionné ci-dessus indiquent donc un découplage dans la formation de la structure native, qui suggère la présence d'un état intermédiaire comprenant des structures secondaires, mais au sein duquel il y aurait quasi absence des contacts à longue distance typiques de la structure tertiaire. Notez que la microcalorimétrie est une autre approche qui permet de confirmer à l'équilibre thermodynamique la validité du modèle de repliement deux-états pour une protéine donnée (*163;164*).

En second lieu, la présence d'un intermédiaire peut être suspectée dans le cas où les paramètres thermodynamiques obtenus à l'équilibre et calculés à partir des paramètres cinétiques diffèrent significativement. Advenant le cas que ce critère d'équivalence soit rempli cela ne signifie pas nécessairement qu'il y a absence d'intermédiaires, mais cela suggérerait à tout le moins que les intermédiaires potentiels présents dans ce contexte seraient instables, ce qui expliquerait la difficulté de les identifier.

La présence d'intermédiaires cinétiques peut être suspectée dès le moment que plusieurs phases exponentielles sont observées dans la réaction de repliement. Les intermédiaires peuvent fréquemment être reliés à la présence de ponts dissulfures ou de prolines dans les protéines. Les intermédiaires les plus intéressants sont ceux qui peuvent être reliés à d'autres phénomènes que ce soit la formation de structures locales ou des contraintes topologiques. Ceux-ci peuvent être révélés par des déviations dans la linéarité de la relation entre le taux de repliement ou de dépliement et la concentration de dénaturant, (voir section **Ingénierie des protéines** et **Chapitre 1 : bases théoriques**). Il pourrait s'agir d'intermédiaire de haute énergie, donc marginalement stables, dans le cas où les autres indicateurs de la présence d'intermédiaires mentionnés ci-dessus sont négatifs.

Intermédiaire hâtif

Il a été démontré que des segments de la structure primaire de certaines protéines qui correspondaient à des éléments de leur structure secondaire (i.e. hélice-α, épingle à cheveux) ainsi que des polypeptides modèle tel que des poly-alanine ou -valine pouvait se replier de manière autonome (réviser dans (*165;166*)). À cause de la nature strictement locale des ponts hydrogène impliqués dans la stabilisation des hélices-α, celles-ci devraient être particulièrement promptes à se replier indépendamment. Le peptide C de la ribonucléase-A a été le premier exemple d'une hélice-α se repliant isolément. Plusieurs autres cas, où il a été démontré que des hélices-α sont partiellement formées dans des conditions dénaturantes, ont été divulgués depuis ces premières études (par exemple (*96;167;168*)). Un autre motif, l'épingle à cheveux, peut se replier seul (*169*), j'y reviendrai

ci-dessus, dans un cas d'espèce chez ubiquitine. Ces observations cadrent bien avec l'hypothèse de la formation d'intermédiaire hâtif (« fast » ou « early intermediate ») comportant la formation de structures locales, en particulier de structures secondaires dès le tout début de la réaction. Les acides aminés affichent des propensions variables pour les divers types de structures secondaires, qui sont dues à des préférences dans l'adoption de leur conformation, soit plus précisément leurs angles ϕ et ψ (voir section **Nature du lien peptidique**). Par conséquent, les segments de la structure dépendamment de leur séquence ont des propensions diverses à former la structure secondaire native correspondante ou d'autres non-natives. Une méthode informatique prenant en compte ces diverses propensions a été développée et permet de calculer les éléments de structures secondaires les plus aptes à se replier de manière isolée (*170*). Par exemple, en ce qui concerne l'ubiquitine cette méthode prévoit que le motif épingle à cheveux en amino-terminal serait le segment possédant la tendance la plus forte à se replier de manière indépendante. Dans les faits, des peptides correspondant au motif épingle à cheveux en amino-terminal d'ubiquitine ou à des peptides plus longs, incluant aussi l'hélice-α, ont la capacité de se former de manière autonome (*171-175*). De plus, le motif épingle à cheveux en amino-terminal d'ubiquitine se replie à un taux supérieur à celui de la protéine complète, ce qui rend vraisemblable que le repliement de ce segment précède celui du reste du cœur hydrophobe (*166*) (voir la **Figure 14** pour un schéma de la structure du motif épingle à cheveux d'ubiquitine ou du dimère formé par les résidus 1-51 d'ubquitine).

Cependant, la plupart des éléments de structures secondaires isolés sont incapables de se replier. La véritable question est donc de savoir si la formation/déformation fluctuante de certaines structures secondaires, telles que des segments hélicoïdaux ou des tours-β, au tout début de la réaction de repliement pourrait limiter la recherche de la structure native parmi toutes les conformations accessibles (*176*). Il est difficile de vérifier expérimentalement une telle hypothèse, mais un détail important qui est toujours le sujet d'argumentations est le rôle joué respectivement par les ponts hydrogène et les interactions hydrophobes dans la formation de ces intermédiaires hâtifs.

Modèle de la charpente

Le modèle de la charpente (« framework ») est en accord avec une vision séquentielle du mécanisme de repliement. Il postule donc la formation de la structure native par la formation d'une série d'intermédiaires à la structure de plus en plus complexe. Par conséquent, ce modèle accorde beaucoup d'importance précisément à ces interactions locales, qui à partir de structures fluctuantes dans les stages initiaux de la voie de repliement, mèneraient à la formation des éléments de structures secondaires. L'établissement des contacts à longue distance serait ensuite facilité par l'arrimage d'éléments de structures secondaires préformés (*165;177*). Ce modèle s'appuie sur la formulation mathématique présenté par Karplus et coll. (*178*), sous le nom de la théorie de diffusion-collision. Celle-ci fait une description satisfaisante de la formation extrêmement rapide de la structure native de certaines protéines, pour lesquelles la limite théorique du taux de repliement est établie par la vitesse limite de la diffusion d'un polypeptide dans l'eau, et telle qu'elle a été observée chez des domaines strictement composés d'hélices-α. Par contre, il a été démontré qu'en dehors de ces cas limites, cette théorie demeure inapplicable.

Accrétion hydrophobe

L'accrétion hydrophobe (« hydrophobic collapse ») est le phénomène par lequel la chaîne polypeptidique se contracte par la formation d'interactions plus ou moins spécifiques impliquant des chaînes latérales hydrophobes. Ce processus se produirait pratiquement instantanément au moment du transfert de la chaîne polypeptidique d'un environnement permettant la solvatation des chaînes latérales hydrophobes à un autre la défavorisant, comme c'est le cas par exemple lors du transfert du ribosome au cytoplasme ou bien au moment de la dilution d'un échantillon de protéine dénaturé dans une solution renaturante. Cette désolvatation partielle des chaînes latérales hydrophobes pourrait favoriser la formation des ponts hydrogène essentiels à la formation des structures secondaires.

Originellement, l'accrétion hydrophobe est une théorie élaborée afin d'expliquer le mécanisme de repliement de la chaîne polypeptidique (*143;179;180*), qui a été échafaudée à partir de simulations informatiques soulignant l'importance du rôle des chaînes latérales hydrophobes (i.e. en particulier leur nombre relatif, leur fréquence et leur dispersion sur la chaîne d'acides aminés) à l'acquisition et à la conservation de la propriété de se replier, la caractéristique primordiale des protéines globulaires. Fondamentalement, ce mécanisme signifie que la réaction de repliement s'effectue à partir d'une conformation de la chaîne polypeptidique ramassée sur elle-même, donc dans un volume restreint. Cela pourrait mener à l'imposition de contraintes stériques qui peuvent limiter initialement le nombre de conformations accessible à une chaîne polypeptidique (d'après ces contraintes, $1x10^{-44}$ des conformations possibles sont réellement accessibles à une chaîne polypeptidique de 100 acides aminés (*179*); voir aussi **Y-a-t-il des voies de repliement ou le mécanisme de repliement est-il séquentiel : introduction aux intermédiaires**). Bien qu'elle ne fournisse pas un modèle détaillé du mécanisme de repliement, la théorie de l'accrétion hydrophobe contribue à aplanir le paradoxe de Levinthal. En effet, la formation de liens locaux d'origine native qui impliquent des chaînes latérales hydrophobes permettrait d'expliquer comment des contacts entre des résidus éloignés, mais à proximité des premiers peuvent alors être favorisés, et via ce phénomène introduirait une certaine forme de coopérativité au processus de repliement (*143*) (cette description est en quelques sorte précurseur de la vision dans laquelle la topologie de la structure native définit le mécanisme de repliement; voir la section **Les protéines partageant la même topologie se replient-elles par un mécanisme identique : oui et non**). Il demeure que plusieurs éléments reste à clarifier autour de cette théorie comme la préséance de l'accrétion sur la formation des structures secondaires ou vice-versa et de la spécificité des interactions hydrophobes formées initialement (*181;182*). Aussi, il n'est pas clair que le phénomène d'accrétion, particulièrement celui qui est relié à ce qui est appelé l'intermédiaire d'accrétion hydrophobe de la phase d'impulsion (« burst phase hydrophobic collapse intermediate »; voir le **Chapitre 1 : Bases Théoriques**) ne serait pas un artéfact des études de repliement *in vitro* qui procède des changements drastiques des propriétés du solvant.

Intermédiaire tardif : le globule fondu

Le terme de globule fondu s'applique à des intermédiaires plus structurés et en quelques sortes plus tardifs selon le schéma général de l'équation **(3)**. Plusieurs études expérimentales, portant entre autre sur l'α-lactalbumine, le cytochrome c et l'apomyoglobine, permettent de proposer un portrait-robot du globule fondu. Typiquement, ce type d'état intermédiaire démontre une rétention très forte des structures secondaires et du niveau de compaction de l'état natif. Par contre, l'empaquetage des chaînes latérales hydrophobes et aliphatiques en particulier ne serait pas terminé, suggérant que les contacts de nature tertiaire seraient non-formés ou à tout le moins qu'ils n'auraient pas encore acquis le niveau de rigidité cristalline attendu du cœur hydrophobe. Cela mènerait à l'exposition en surface de chaînes latérales hydrophobes ordinairement enfouies et à l'établissement de contacts potentiellement non-natifs (*183;184*). Ce type d'intermédiaire est l'espèce dominante chez plusieurs protéines dans des conditions de pH acide (ordinairement de pH 2 à 4 selon les protéines).

Est-ce que la détection d'états possédant les caractéristiques du globule fondu dans ces conditions expérimentales à l'équilibre thermodynamique est suffisante afin de proposer leur présence au cours du processus de repliement (i.e. dans les traces cinétiques du repliement)? Il est difficile de répondre succinctement à cette question. Par exemple, les traces de la cinétique de repliement de plusieurs protéines obtenues en présence de 1-anilino-8-naphtalènesulfonate (ANS)[15] révèlent deux phases exponentielles, une première ascendante et une seconde descendante, qui correspondraient respectivement à la transition de l'état dénaturé vers le globule fondu et de ce dernier à l'état natif. La présence généralisée de cet intermédiaire reste à démontrer en particulier pour les petites protéines dont le repliement est de type deux-états. En accord avec cela, des résultats concernant la protection à l'échange hydrogène/deutérium des groupements amines de plusieurs protéines

[15] L'ANS est une molécule qui interagit préférentiellement avec les chaînes latérales hydrophobes dont l'exposition au solvant est partielle. Il a donc une affinité très forte pour le globule fondu. Ce type d'interaction provoque une augmentation du rendement quantique de la fluorescence de l'ANS et un déplacement du maximum d'émission vers le bleu.

dont le cytochrome c et l'apomyogobine indiquent que les ponts hydrogène impliqués dans la formation de certaines structures secondaires se forment plus rapidement que les ponts hydrogène des contacts tertiaires. Par conséquent, ces résultats suggèrent que les espèces protéiques observées à l'équilibre thermodynamique et dans les cinétiques possèderaient des caractéristiques communes qui en outre, correspondent aux propriétés du globule fondu.

L'importance du globule fondu dans des phénomènes biologiques a été suggérée entre autre dans le repliement assisté par les chaperonnes, l'insertion membranaire ou la translocation à travers les membranes et la protéolyse par la voie d'ubiquitination (réviser dans (*184*)). Par exemple, il a été postulé que l'état intermédiaire correspondant au globule fondu chez la β-lactamase était nécessaire à la translocation dans le périplasme des bactéries gram négatives. Un état intermédiaire un peu plus dénaturé que le globule fondu[16] et qui est suspecté sur la voie de repliement de plusieurs protéines pourrait avoir un rôle pathogénique comme cela a été proposé dans le cas de la fibrillation en plaques de la protéine tau, un phénomène qui peut mener au déclenchement de la maladie d'Alzheimer (*185*).

Données à l'équilibre recueillies sur ubiquitine et suggérant la présence d'un intermédiaire tardif

Quelques données tirées de la littérature permettent de supposer la présence d'un état intermédiaire tardif sur la voie de repliement d'ubiquitine. Premièrement, il a été remarqué que le transfert d'ubiquitine dans une solution comportant 60 % de méthanol induit la formation d'une nouvelle espèce structurale majoritaire appelée l'état-A. Les résidus formant le motif épingle à cheveux amino-terminal ainsi que l'hélice-α principale adoptent une conformation stable dans laquelle jusqu'au pairage natif des résidus de ces deux brins-β est respecté, cependant que les résidus localisés en carboxy-terminal des régions susmentionnées semblent adoptés une conformation dénaturée (*186;187*). D'autre

[16] Espèce protéique, dite pré-globule fondu, qui démontre des propriétés intermédiaires entre le globule fondu et l'état dénaturé.

part, la variation du taux d'échange hydrogène/deutérium par RMN de l'ubiquitine en fonction de l'application d'une pression atmosphérique croissante (entre 30 et 3700 bars) dans la cellule contenant l'échantillon suggèrent la présence d'intermédiaires (*188;189*). Plus précisément, ces données indiquent que deux régions localisées autour des résidus 33 à 42 (brin-β 3) et 70 à 76 (fin du brin-β 5) démontrent des variations coopératives de leur taux d'échange en fonction du changement de pression (**Figure 14**). Ces données sont compatibles avec les cinétiques d'échange des protons d'ubiquitine à pression normale constante et sur le rôle potentiel de l'isomérisation de ces résidus prolines, soit P19 et surtout P37[17], dans ce phénomène (*190*). Or, l'interprétation des résultats obtenus d'autres expériences suggère que cet intermédiaire puisse avoir un rôle physiologique. En effet, la structure révélée par RMN d'un assemblage covalent de 2 à 4 ubiquitines successives, par la connexion de l'extrémité carboxy-terminale d'un monomère sur le K48 du suivant indique que des résidus hydrophobes situés en surface et à proximité des régions instables révélées par les expériences précédentes forment une structure isolée du solvant à l'interface de chaque monomère du dimère ou du tétramère. L'hypothèse proposée par les auteurs est que les variations structurales observées sur les monomères d'ubiquitine à l'intérieur des complexes versus le monomère libre pourraient être reconnus par la

Figure 14. Résultats de quelques-unes des études structurales portant sur ubiquitine et suggérant la présence de divers types d'intermédiaires lors de son repliement.
La figure se trouve à la page suivante. **A**, La structure (liens de la chaîne principale et atome d'oxygène du carbonyle et hydrogène de l'amide) du peptide correspondant aux résidus 1-17 d'ubiquitine (soit le motif épingle à cheveux : β1 et β2) obtenu par RMN (schéma de gauche ; code PDB 1E0Q) en comparaison avec la section correspondante de la séquence polypeptidique de l'ubiquitine (schéma du centre ; code PDB 1UBI). Notez la légère différence de conformation dans le tour-β et en carboxy-terminal. **B**, Le dimère formé par les résidus 1-51 d'ubiquitine soit par β1, β2, α1 et β3 (schéma de gauche ; code PDB 1GJZ) via l'échange de leur β3 respectif, soit 1-β3 et 2-β3, afin de former un feuillet-β mixte tel qu'observé par RMN (liens de la chaîne principale). **C**, Étude comparative par RMN de l'effet de la pression sur la structure d'ubiquitine. Les structures de gauche et du centre (le schéma représente les mêmes liens qu'en **(A)**) ont été obtenues respectivement à pression normale (code PDB 1V80) et à haute pression (code PDB 1V81). Les changements principaux se produisent entre β3 et β5 (respectivement en jaune et en rouge), alors que plusieurs ponts hydrogène (traits bleus) entre ces deux brins disparaissent à haute pression. Afin de servir de guide pour l'observation des structures schématiques, la structure de l'ubiquitine de ts obtenu par rayons-X (code PDB 1UBI) a été placée à l'extrême droite de chaque panneau en conservant la même orientation générale que les schémas d'intérêt. Les structures secondaires importantes à la description des topologies représentées sont annotées selon la même nomenclature que celles décrites précédemment pour le DLR de Raf (**Figure 3**).

[17] Notez que le résidu 93 du DLR de Raf qui correspond au résidu 37 d'ubiquitine est également une proline. Il n'y a pas d'équivalent à P19 chez le DLR, mais il a cependant une proline dans le premier tour-β (P63).

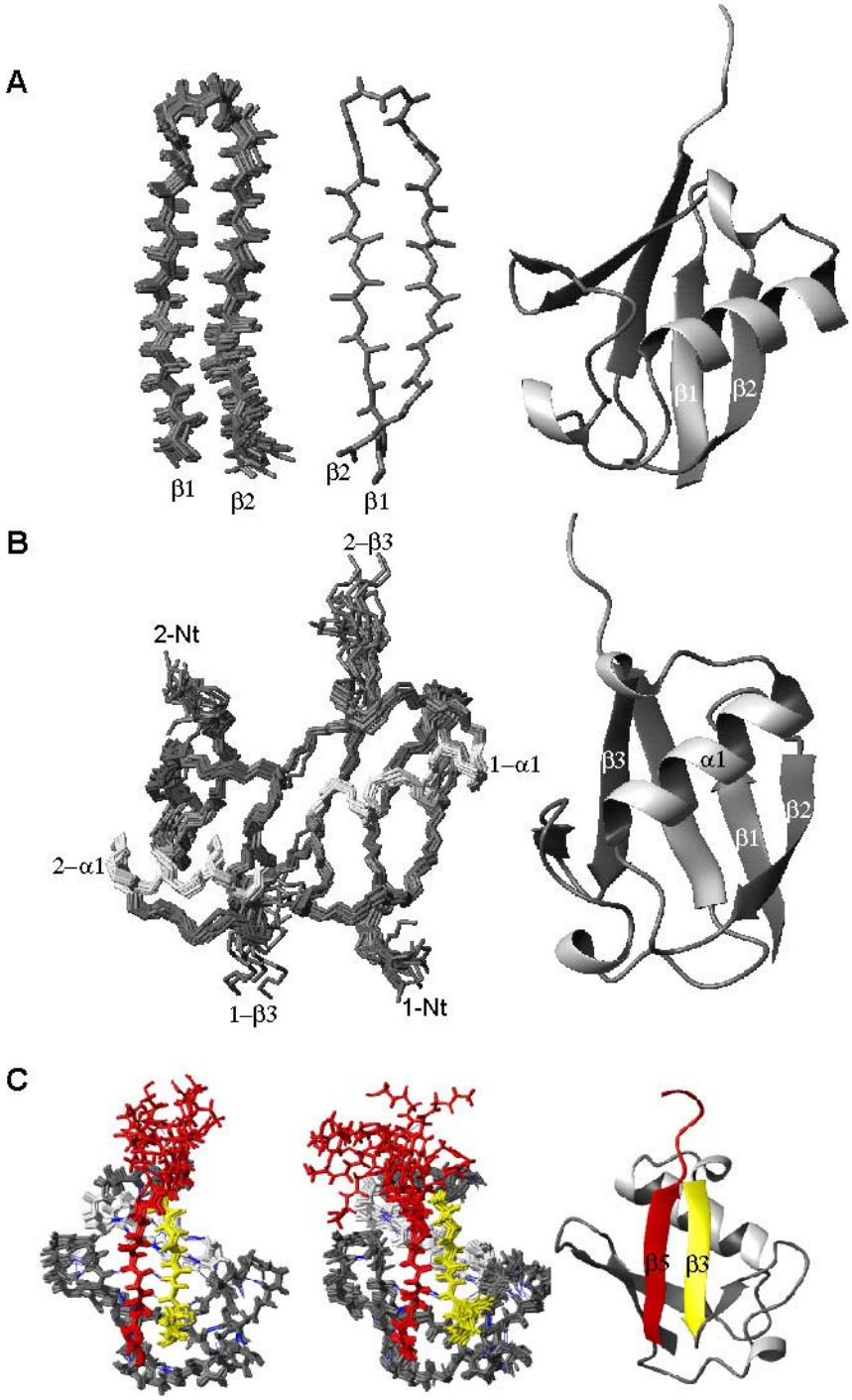


Figure 14. La légende est à la page précédente.

machinerie du protéasome[18] et d'autres complexes cellulaires, de telles sortes qu'elles agiraient ainsi comme un signal de ciblage permettant de déterminer les routages métaboliques et de la signalisation cellulaire divers des protéines marquées alternativement par la mono, di ou la tétra(poly)-ubiquitination (*191*). En accord avec cette théorie, les données expérimentales obtenues sur la dénaturation d'ubiquitine induite par la force mécanique et des simulations moléculaires ont démontré que les monomères d'un tétramère unis par K48 sont moins stables que ceux reliés par K63 ou via la formation d'un lien peptidique (*192*). D'autres travaux ont suggéré la présence d'un intermédiaire qui lie l'ANS dans le repliement cinétique d'ubiquitine (*193*) (voir aussi communications personnelles et thèse d'A. Vallée-Belisle), mais j'y reviendrai plus tard (voir section

Preuves de la présence d'intermédiaires cinétiques dans le repliement d'ubiquitine). Bien que ces expériences ne puissent confirmer que l'intermédiaire potentiel que je viens de décrire soit assurément compatible avec le modèle du globule fondu, ils sont en accord avec la présence d'un intermédiaire tardif aux propriétés proches de celles de l'état natif dans le cas d'ubiquitine.

Intermédiaires de haute énergie : études cinétiques

Le caractère deux-états apparent (tel que décrit ci-dessus) du repliement d'une protéine n'exclut pas la présence d'intermédiaires. Par exemple, si un intermédiaire est plus haut en énergie que l'état dénaturé et que l'état natif ou bien si la barrière d'activation pour la formation d'un intermédiaire est significativement plus haute que celle de sa conversion vers l'état natif, les traces de sa cinétique de repliement seront vraisemblablement modélisables par des fonctions exponentielles simples. En effet, ces d'intermédiaires sont extrêmement instables et par conséquent, ils seraient très peu peuplés (≈ 1%) au niveau

[18] L'ubiquitine est essentielle aux ciblages des protéines à dégrader vers un complexe protéique impliqué dans la protéolyse intracellulaire qui est connu sous le nom de protéasome. Pour ce faire l'ubiquitine est attachée par une cascade enzymatique à une lysine de la protéine à dégrader, et d'autres ubiquitines sont attachées à la première ubiquitine via une de ces nombreuses lysines, soit K48. La formation de la chaîne de polyubiquitine, habituellement une tétra-ubiquitine, sur une protéine entraîne son ciblage vers le protéasome.

cinétique et thermodynamique. Par ailleurs, les cinétiques de renaturation et de dénaturation (i.e. les expériences qui permettent de mesurer respectivement la variation du taux de la réaction de repliement et de dépliement en fonction de la concentration de dénaturant) peuvent fournir des preuves de la présence d'intermédiaires. En effet, une déviation descendante dans la linéarité de la relation entre le taux de repliement ou de dépliement et la concentration de dénaturant dans une courbe de chevron (voir section **Chapitre 1 : Bases Théoriques**) est considérée comme une preuve de la présence d'un intermédiaire de haute énergie. La détection de ce type d'intermédiaires est facilitée par leur stabilisation induite par des variations du pH, et l'ajout de certains sels et solvants organiques au tampon de renaturation. Parfois, la présence d'intermédiaires de haute énergie (intermédiaires rapides hydrophobes ou non) est suspectée lorsqu'un déficit du signal de fluorescence apparaît dans le temps mort de l'appareil (voir section **Accrétion hydrophobe** et **Chapitre 1 : Bases Théoriques**).

Une étude récente concluait que les intermédiaires de haute énergie seraient très fréquents chez les protéines qui montrent des déviations dans la relation linéaire entre leur énergie libre d'activation et la concentration de dénaturant. Les variations dans la relation linéaire de cette relation ont été postulés de découler d'un changement dans la position de l'état de transition par rapport à l'état intermédiaire; dans ce cas de figure l'état de transition se rapproche de l'état natif en fonction de l'augmentation de la concentration de dénaturant, ce qui voudrait dire que l'état de transition se situe respectivement avant et après l'intermédiaire en fonction de la concentration de dénaturant (*127*). Ces auteurs ont aussi noté que la seconde transition aurait un placement relativement équivalent chez les protéines étudiées suggérant un mécanisme généralement similaire quelque soit la nature de l'état de transition formé. Par ailleurs, une courbe de chevron ne démontrant aucune déviation ne garantit pas l'absence d'intermédiaires cinétiques. La consultation des travaux des groupes de Sosnick et Kiefhaber permettent de faire un tour d'horizon rapide des points de vue contemporains sur les intermédiaires de haute énergie (réviser dans (*127;194*)). Ci-dessous, je discute l'un des cas les mieux documenté d'intermédiaires cinétiques et je

présente certaines des preuves qui font suspecter de la présence d'intermédiaires cinétiques dans le cas d'ubiquitine.

L'intermédiaire de repliement obligatoire de barnase : visualisation via l'analyse des valeurs-Φ

Le squelette carboné de barnase est composé de 110 acides aminés arrangés en 3 hélices-α successives suivies d'un feuillet-β composé de 5 brins-β arrangés de manière antiparallèle (**Figure 10**). Le cœur hydrophobe principal est situé entre le feuillet-β et la grande hélice-α. Il y a en plus deux autres cœurs hydrophobes mineurs situés en périphérie, soit autour des hélices-α 2 et 3 et sur l'autre face du feuillet. La courbe de chevron de barnase démontre un fléchissement du bras de repliement à faible concentration de dénaturant, qui suggérait la présence d'un intermédiaire. D'autre part, barnase est la première protéine qui ait été soumise à une analyse des valeurs-Φ (voir section **Ingénierie des protéines**) étendue à des résidus répartis sur l'ensemble de la structure pour déterminer la structure de l'état de transition (*82*). L'intermédiaire a été caractérisé par la même méthode afin de proposer un modèle de sa structure (*80*). Les résultats obtenus ont suggéré que l'intermédiaire est sur la voie de repliement menant à la formation de l'état de transition à partir de l'état dénaturé, car des éléments de structure formés à l'état de transition sont déjà présents chez ce premier état. En effet, le cœur hydrophobe principale apparaît déjà fortement structuré et se renforce particulièrement aux extrémités du feuillet-β et de l'hélice-α principale lors de la formation de l'état de transition. Par contre, les deux cœurs hydrophobes mineurs ainsi que les boucles adopteraient leur conformation native après l'état de transition.

Le mécanisme de repliement de barnase dans son assemblage progressif d'éléments de structures secondaires préformés est en accord avec le modèle de la charpente. Il y a peu d'autres exemples où la structure d'un intermédiaire a pu être obtenue avec autant de détails. Une exception notable est celle de la protéine Im7, dans laquelle l'intermédiaire a été caractérisé par la même approche (*104;195*). La particularité de ces études est qu'elle

démontre la présence d'un intermédiaire sur la voie de repliement d'Im7, qui comprendrait des interactions hydrophobes non-natives.

Preuves de la présence d'intermédiaires cinétiques dans le repliement d'ubiquitine

La cinétique du repliement d'ubiquitine a été étudiée par plusieurs groupes de recherche et est à la source de polémiques fécondes en publications. Le caractère deux-états de la réaction de repliement d'ubiquitine apparaît incontestable d'après la comparaison des données obtenues d'expériences cinétiques et thermodynamiques (*131;196;197*). Par ailleurs, le nombre de phases détecté dans les traces cinétiques de repliement peut varier entre 2 et 4 selon les auteurs et les conditions de l'expérience. Dans leurs travaux Korasanizadeh et coll. ont identifié 3 transitions modélisables par des fonctions exponentielles (voir la section **Études cinétiques**): une phase majeure rapide, une phase moyenne en amplitude et en taux de repliement ainsi qu'une phase mineure lente reliée à l'isomérisation des prolines (*196;198*). À faible concentration de dénaturant, ils ont rapporté une courbature dans la linéarité de la dépendance du taux de repliement en fonction de la concentration de dénaturant qui a été attribuée à l'accumulation d'un intermédiaire hydrophobe durant le temps mort de l'appareil de mixage rapide. L'ampleur de ce signal de type phase d'impulsion démontrait une dépendance coopérative à la concentration de dénaturant entre l'état dénaturé et un intermédiaire potentiel. Par ailleurs, la présence de cet intermédiaire est contestée dans trois publications qui suggèrent des causes alternatives à cette courbature dans le bras de l'expérience de repliement: 1) l'agrégation de la protéine à partir d'une certaine concentration (*193*); 2) un défaut de lissage dû à la perte d'une partie du signal dans le temps mort de l'appareil (*131;197*). En ce qui concerne cette dernière hypothèse, vu que considérant son taux seul la première phase du repliement est sujette à être absorbée en tout ou en partie dans le temps mort de l'appareil, il est aisé de mélanger la première et la seconde transition dans les traces cinétiques du repliement ce qui advenant cela mènerait à une sous-évaluation du taux de la phase majeure. A ce propos, l'absence de déviation dans la cinétique de repliement à basse température (8ºC), qui a été présentée comme étant une preuve que l'intermédiaire était

stabilisé par des contacts hydrophobes, pourrait aussi découler simplement d'une diminution du taux de la phase majeure et donc, d'un amoindrissement de l'effet du temps mort. Confortant aussi cette explication, l'utilisation d'un appareil de mixage rapide avec un temps mort plus faible que celui utilisé dans le cadre des études mentionnées au début de ce paragraphe a permis de résoudre la déviation dans la courbe de chevron (*194;197*).

Or, les mêmes difficultés sont rencontrées dans l'analyse des données de repliement du DLR de Raf dont les traces cinétiques sont modélisées en utilisant quatre fonctions exponentielles qui auraient de surcroît des propriétés similaires à celle d'ubiquitine (*131*). Des expériences de double saut de dilution (voir la section **Études cinétiques**) ont permis de démontrer que les deux phases les plus lentes du DLR de Raf et d'ubiquitine pouvaient être attribuées à l'isomérisation cis/trans du lien imide d'une proline ou d'autres acides aminés. Par ailleurs, les deuxièmes transitions observées dans la réaction de repliement du DLR de Raf et d'ubiquitine sont particulièrement intrigantes, car leur taux respectifs ne sont pas sensibles à la concentration de dénaturant et trop élevés pour être attribuables au phénomène d'isomérisation des prolines. D'autre part, il subsiste d'autres preuves qui trahissent la présence d'intermédiaires transitoires dans la réaction de repliement d'ubiquitine. Tout d'abord, la détection d'une espèce structurale liant l'ANS en présence de Na_2SO_4 dans une expérience de renaturation favorise l'hypothèse d'un état intermédiaire obligatoire, mais instable sur la voie de repliement d'ubiquitine (*193*). Finalement, la protection de l'échange des protons amines avec le deutérium durant le repliement indique que deux régions sont découplées du reste de la structure, soit le début du brin-β 3 et le brin-β 5 (*190*) (voir aussi la section **Données à l'équilibre recueillies sur ubiquitine et suggérant la présence d'un intermédiaire tardif**).

La capacité de postuler la présence d'un intermédiaire est limitée par la puissance des moyens techniques disponibles pour son observation. Par exemple, l'utilisation de certaines sondes fluorescentes intrinsèques tel que des tryptophanes ou tyrosines peuvent faire conclure à un observateur qu'il n'y a pas d'intermédiaire, car malgré le fait que l'on se serve de celles-ci pour obtenir de l'information sur le remaniement globale de la chaîne

polypeptidique au cours du processus de repliement, leurs propriétés de fluorescence dépendent d'abord de la variation de leur environnement local. En fait, lorsque plusieurs acides aminés de ce type sont présents dans la même protéine, il est probable que la variation de l'environnement local autour de chacune de ces sondes influent sur le signal de fluorescence (i.e. sur l'amplitude et l'intensité) détecté lors de la réaction de repliement. Dans un tel contexte, il est plus probable que des découplages de la formation de la structure dans les diverses sous-régions puissent alors être détectés, révélant ainsi des « intermédiaires » qui passeraient inaperçus dans un autre contexte. Ce pourrait être le cas de barnase par exemple (voir la section précédente), puisqu'elle contient trois tryptophanes, soit un dans chacun des coeurs hydrophobes. Inversement, il est possible d'imaginer qu'une réaction de repliement deux-états pourrait apparaître plus complexe en présence de plusieurs sondes. Par conséquent, des sondes fluorescentes uniques (i.e. tryptophanes) introduites dans des endroits dispersés de la chaîne polypeptidique grâce à des outils classiques de biologie moléculaire pourraient permettre de décortiquer le mécanisme de repliement par l'identification d'intermédiaires autrement difficiles à étudier. Un collègue étudiant au doctorat de notre laboratoire a proposé et développé cette approche afin d'étudier le repliement du DLR de Raf et d'ubiquitine (voir thèse d'A. Vallée-Belisle).

Les résultats qu'il a obtenus par cette méthode suggèrent que la seconde phase modélisable dans les traces de repliement du DLR de Raf et d'ubiquitine est due à la conversion d'un intermédiaire vers la structure native et que bien qu'il semble y avoir des différences importantes dans la structure précise de ces intermédiaires, ils impliquent tous les deux des réarrangements dans la moitié carboxy-terminale de ces protéines (communications personnelles d'A. Vallee-Belisle), respectivement autour du brin-β 3 et, de l'hélice 3_{10} ainsi que du brin-β 5. Il est à noter que les traces de repliement obtenues avec les tryptophanes insérés au sein des régions impliquées dans les réarrangements démontraient une forte augmentation dans l'amplitude de la seconde phase, de tel sorte qu'elle devenait fortement majoritaire. De plus certains mutants d'ubiquitine démontraient une première phase ascendante suivie d'une phase descendante, en accord avec la

formation d'une espèce possédant des caractéristiques structurales non-natives à la suite de la première transition. Les caractéristiques structurales grossières de cet intermédiaire potentiel obtenues par cette approche sont compatibles avec l'interprétation des données cinétiques de protection à l'échange des protons d'ubiquitine au cours du repliement (*190*) (voir ci-dessus) et de l'analyse des valeur-Φ (*97*). De manière analogue, la structure d'un état intermédiaire du DLR de Raf prédite également par cette approche novatrice est compatible avec son placement après l'état de transition et l'analyse de ces valeurs-Φ (**Article 4**). Il reste maintenant à déterminer hors de tout doute que ces intermédiaires sont bien situés sur les voies de repliement d'ubiquitine et du DLR de Raf (par opposition à des intermédiaires cul-de-sac, situés en dehors de la voie de repliement principale; voir section **Intermédiaires**) et surtout, à comprendre la raison de la quasi-indépendance vis-à-vis de la concentration de dénaturant du taux de la phase moyenne de repliement correspondant à la transition hypothétique I→F.

Conclusion sur les intermédiaires

L'acceptation généralisée de la théorie qui postule l'existence de voies de repliement présentant une série d'intermédiaires se complexifiant structuralement progressivement au cours du processus s'est butée principalement à l'observation qu'il n'y a pas de preuves tangibles de la présence d'intermédiaires dans la réaction de repliement de plusieurs petites protéines se repliant apparemment en deux-états. Certaines données tendent à indiquer que l'absence de détection d'intermédiaires cinétiques pourrait être due aux méthodes expérimentales et analytiques utilisées ainsi qu'aux propriétés en tant que tel des intermédiaires de hautes énergies (*127*). De plus, bien qu'il soit de plus en plus évident que la plupart des intermédiaires ne sont pas contre-productifs à la réaction de repliement, il est ardu de démontrer que leur formation facilite la réaction de repliement ou est une conséquence normale du repliement des polypeptides et non pas celle de contraintes fonctionnelles ou topologiques imposées au cours du processus d'évolution naturelle. D'autre part, il est généralement admis que les protéines de grande taille (>150 a.a.) et à multi domaines (*199;200*) se replient via la formation d'intermédiaires (*201*).

Entonnoir et multiplicité des voies de repliement

La théorie de l'entonnoir (« funnel »), qui s'appuie sur la chimie des polymères, des simulations informatiques du repliement de protéine modélisé sur treillis (« lattice ») et la mécanique statistique, décrit la réaction de repliement d'un polypeptide vers sa structure native d'une façon nouvelle (*202-205*). La réaction de repliement y est représentée par un entonnoir tridimensionnel à la surface rugueuse (**Figure 15**). Trois assomptions principales différencient ce modèle des visions plus classiques du processus de repliement qui ont été précédemment décrites dans l'**Introduction** :

1. Les protéines se replient à partir d'un état déplié composé de structures diverses et fluctuantes par contraction et reconfiguration de la chaîne polypeptidique.

2. Les reconfigurations se produisent par diffusion des conformations de hautes énergies vers celles de basses énergies.

3. Les reconfigurations locales sont privilégiées, celles globales étant prohibées suite à la contraction initiale.

La vitesse à laquelle le polypeptide atteint sa structure native est dépendante de la distance entre l'état de transition et la transition de verre, cette zone correspondant à la crise entropique de la réaction de repliement. La nature processive du repliement dépend aussi du degré de rugosité général de l'entonnoir, en particulier autour de l'état natif. D'autre part, les aspérités de leur entonnoir sont réputées peu nombreuses et faibles chez les protéines se repliant en deux-états, de sorte que dans ces cas c'est principalement l'état natif qui détermine les caractéristiques de la voie de repliement. L'état natif est le plus stable de point de vue thermodynamique, mais il peut-être métastable par rapport à la diversité de conformations adoptables par une chaîne polypeptidique donnée. Une nouvelle notion qui est cruciale à la compréhension de cette perspective alternative sur le processus de repliement est celle d'ensemble. Un « ensemble » groupe des structures diverses qui possèdent toutefois des caractéristiques communes. Ainsi, dans cette perspective la nature de l'état de transition, l'état central à la détermination de la cinétique de repliement faut-il le rappelé, est transformée. La réconciliation avec la vision classique de l'état de transition

par l'analyse des valeurs-Φ est accomplie de la manière suivante. Les résidus/régions de la chaîne polypeptidique où les valeurs-Φ sont élevés seraient plus rigides et plus structurés que ceux ne participant pas à la stabilisation de l'état de transition (i.e. où les valeurs-Φ sont faibles). Conséquemment, ces dernières seraient principalement responsables de la diversité de la structure des molécules constituant l'ensemble état de transition. Finalement, la diversité structurale d'un ensemble est reliée de manière proportionnelle à son niveau d'énergie, de sorte qu'on retrouve un ensemble de molécules à la structure plus dégénérée au début de la réaction de repliement, c'est-à-dire à haute énergie.

Par extension, ce modèle prédit la présence de voies de repliement parallèles qui pourrait mener alternativement un polypeptide vers sa structure native. Cela pourrait expliquer la résistance à la mutagenèse de la réaction de repliement, car une mutation pouvant affecter une voie de repliement, laisserait potentiellement intact des voies alternatives (*206*). Cependant, jusqu'à ce jour, seulement quelques protéines sont suspectées de se replier ou de se déplier par des voies parallèles (*122;125;207;208*). Il se pourrait que la multiplicité des voies de repliement survienne principalement au début de la réaction, plus précisément aux stades précédents l'état de transition. En rétrospective, à la lumière des évidences présentées ci-dessus la généralité du modèle de l'entonnoir est loin d'être acquise, malgré le fait que ce modèle apparaisse cohérent en théorie.

Figure 15. Le modèle de l'entonnoir (« funnel »), la réaction de repliement des protéines sur une surface tridimensionnelle.
La figure se trouve à la page suivante. (**A-C**) Représentation du diagramme d'énergie de la réaction de repliement vers la structure native (N) à partir de la structure dépliée (D) sous la forme d'un entonnoir tridimensionnel modèle (adapter de (*209*) avec la permission de l'auteur). **A**, Entonnoir modèle en accord avec la théorie de la voie de repliement unique. **B**, Entonnoir modèle permettant de prédire la présence d'un ensemble de voies de repliement lentes (L) et rapides (R). **C**, Entonnoir modèle rugueux pour une réaction de repliement. Dans les trois cas l'état natif est le plus stable en énergie. Les panneaux (**B**) et (**C**) indiquent bien que plusieurs voies sont susceptibles de permettre le repliement du polypeptide vers sa structure native. Par contre, en (**C**) la surface de l'entonnoir est rugueuse et plusieurs aspérités peuvent agir en tant que cul-de-sac, car elles sont suffisamment stables et requièrent le franchissement d'une haute barrière en énergie afin de rejoindre la structure native. Le modèle de l'entonnoir rugueux est celui favorisé par les tenants de la théorie de l'entonnoir et de la multiplicité des voies de repliement, quoique le nombre d'aspérités varie d'une protéine à l'autre. **D**, Représentation schématique de la réaction de repliement sous forme d'entonnoir en deux dimensions. La diversité des conformations accessible à une molécule de polypeptide qui se replie diminue en fonction de sa position dans la réaction de repliement et à son niveau d'énergie comparativement à l'état natif. Les positions approximatives de l'état du globule fondu et de l'état de transition par rapport à l'état natif sont aussi indiquées.

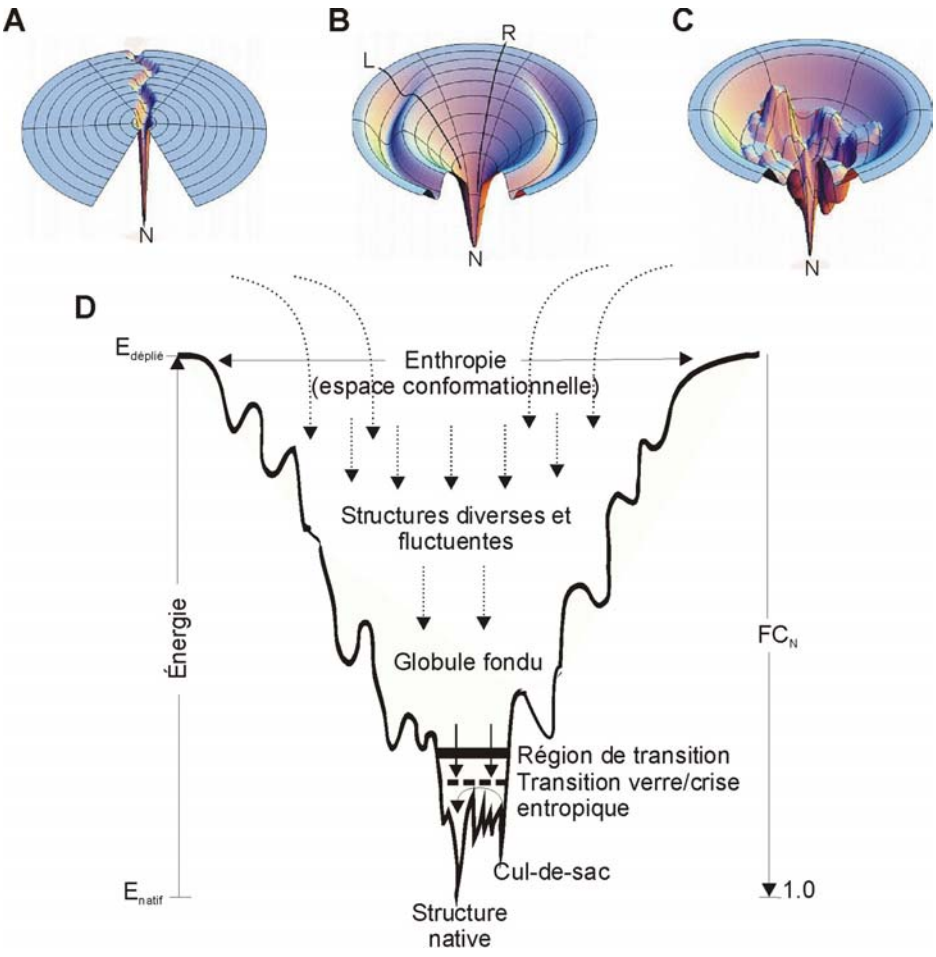


Figure 15. La légende est à la page précédente.

L'état déplié et/ou dénaturé

La conception de la structure de l'état dénaturé a évolué fortement au cours des dernières années, passant de la conception classique proche de l'embobinage aléatoire à des preuves claires directes et indirectes de la présence d'éléments de structures fluctuantes plus ou moins stables et variables selon les conditions dans lesquelles elle est étudiée.

Peu de méthodes expérimentales sont appropriées pour obtenir des informations structurales détaillées sur l'état dénaturé. Les techniques qui permettent d'obtenir de l'information précise sont toutes basées sur la RMN, mais même en utilisant ces approches

il est habituellement impossible d'en obtenir une structure précise, car l'état dénaturé est par définition composé de structures labiles et fluctuantes. Dans cette section très brève sur la nouvelle vision de l'état dénaturé et en particulier sur l'importance que celle-ci pourrait avoir sur notre compréhension du processus de repliement, je vais décrire brièvement les études les plus marquantes, en particulier celles qui ont porté sur des petites protéines.

La combinaison de simulations informatiques en dynamique moléculaire et de RMN sur l'état dénaturé de barnase obtenu à haute concentration d'urée permirent de démontrer que les éléments de la structure secondaire consolidés aux états intermédiaire et de transition, dont entre autre l'hélice-α majeure, étaient déjà partiellement renforcés en l'état dénaturé (résumer dans (*210*)). Ce résultat est donc quelque peu attendu et est en accord avec ce qui est connu sur le mécanisme de repliement de cette protéine. Il a été plus surprenant de découvrir par les mêmes approches que la protéine CI2 démontre elle aussi un état dénaturé, obtenu cette fois-ci à haute concentration de Gdm-HCl, démontrant des structures résiduelles, bien que plus restreintes que dans le cas précédent[19] (*96*). En fait, l'état dénaturé de CI2 apparaît fortement déplié, ne démontrant que deux régions significativement structurées soit l'hélice-α principale, en particulier en son centre, et des contacts non-natifs de type hydrophobe également au milieu du feuillet-β.

Dans une autre étude de RMN, cette fois sur l'état dénaturé du lysozyme obtenu à haute concentration d'urée ou bien à pH= 2, Klein-Seetharaman et coll. ont pu démontré la conservation de segments hélicoïdaux stabilisés par la formation de contacts hydrophobes non-natifs établis à longue distance et impliquant l'interface des deux domaines structuraux (*211*). Le remplacement d'un seul résidu tryptophane par une glycine a été suffisant pour détruire ces structures. Récemment, une étude remarquable de l'état dénaturé de l'homéodomaine d'engrailed[20] par RMN a été rapportée (*168*). Ces travaux ont ceci de

---

[19] La comparaison de l'état dénaturé de deux protéines obtenu avec des dénaturants différents doit être prise avec un grain de sel.

[20] L'homéodomaine est composé d'un faisceau de trois hélices-α. C'est l'un des domaines se replant le plus rapidement parmi ceux qui aient été documentés ($k_f \approx 39\,900\ s^{-1}$) (*274*).

particulier que les auteurs dans le cadre d'une étude d'ingénierie de protéine ont découvert par coïncidence un mutant aux propriétés remarquables. Effectivement, le mutant L16A se replie rapidement à haute concentration de sels par un procédé et un taux similaire à la protéine originale, mais à force ionique physiologique, il ne se replie pas et peut être dénaturé encore un peu plus par l'ajout d'urée via une transition faiblement coopérative. La structure RMN de ce mutant dans ces dernières conditions ioniques a révélé une structure largement dépourvue de contacts natifs à longue distance, mais possédant de nombreux éléments de structures secondaires natifs et un court segment hélicoïdal non-natif (**Figure 16**). L'obtention de

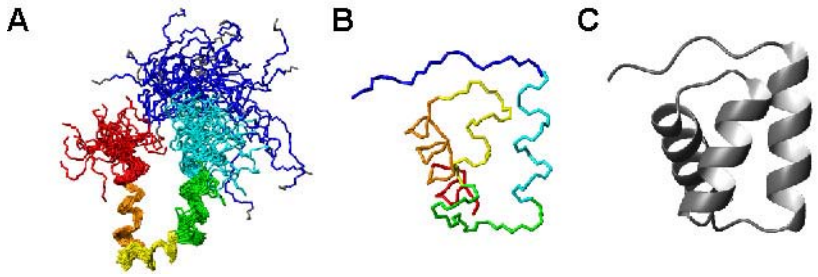


Figure 16. Comparaison entre la structure de l'intermédiaire et de l'état natif de l'homéodomaine d'engrailed. **A**, Structure (liens de la chaîne principale) du mutant L16A à faible force ionique obtenue en RMN (1HZR). **B** et **C**, Structure de l'homéodomaine engrailed de ts obtenue par cristallographie. L'extrémité amino-terminale est colorée en bleu. Présentation des structures dans le panneaux (**A**;**B**) inspirés de (*168*).

cette structure permet d'extrapoler avec un degré de confiance élevé une courbe de dénaturation et de calculer la stabilité de cet état. Les expériences cinétiques ont aussi montré que la structure native peut être obtenue à partir de l'état dénaturé formé par le mutant L16A, si la renaturation est monitorée à haute force ionique. Ce dernier apparaît similaire à la description d'un intermédiaire tardif tout en se distinguant par sa haute stabilité. Il s'agit donc d'un état dénaturé très structuré, qui pourrait s'avérer similaire à celui présent de façon transitoire en conditions physiologiques chez la protéine de type sauvage. Malgré le fait que les caractéristiques de cet intermédiaire soient particulières et qu'il est peu probable qu'une espèce protéique équivalente constitue une caractéristique générale des voies de repliement des chaînes polypeptidiques, ces travaux sont extrêmement instructifs et nous fournissent une première image détaillée d'un intermédiaire et par extrapolation d'un état dénaturé ainsi que de leur potentielle relation structurale et

thermodynamique. D'autres études ont suggéré la présence de structures résiduelles à l'état dénaturé d'autres petites protéines (*167;212;213*).

Par ailleurs, les corrélations entre la variation des paramètres thermodynamiques et cinétiques permettant de mesurer la sensibilité au dénaturant des transitions F→‡ et U→F (voir section **Ingénierie de protéines, analyse des valeurs-Φ**) ont démontré que les propriétés structurales de l'état dénaturé de la protéine ribosomale S6 varieraient en fonction des mutations introduites (*112*). Une étude portant sur 21 protéines dont les données ont été répertoriées dans la littérature a suggéré la présence d'un phénomène similaire chez sept d'entre elles, fournissant ainsi des preuves de la généralité de ce phénomène (*125*).

L'état dénaturé dans une expérience de repliement revêt une importance particulière. Il est évident que l'état dénaturé en présence d'agent de dénaturant puissant tel que l'urée et le Gdm-HCl est moins structuré que l'état déplié présent en conditions physiologiques. Or, lorsque l'on étudie le repliement *in vitro* à partir du premier de ces états, sa dilution rapide dans des conditions favorisant le repliement mène à la formation d'un état probablement plus près de l'état déplié biologiquement significatif. À savoir si l'espèce obtenue dans ce contexte devrait être aussi considérée comme le premier intermédiaire de la réaction de repliement, n'est pratiquement qu'une question de sémantique. L'occurence d'un état dénaturé très structuré pourrait théoriquement compliquer l'interprétation des données obtenues par l'ingénierie de protéine, particulièrement si celui-ci est stabilisé par des contacts de type non-natif (voir section **Ingénierie des protéines**). Cependant, la faible stabilité des interactions établies à l'état dénaturé versus à l'état de transition et à l'état natif permet de croire que la simplification de l'analyse des données est adéquate en pratique (voir section **Études cinétiques**). En résumé, ces résultats suggèrent que le paradoxe de Levinthal pourrait aussi être partiellement résolu (voir section **Accrétion hydrophobe**) par une restriction de l'exploration conformationnelle qui pourrait bien être une propriété intrinsèque de la chaîne polypeptidique sous sa forme dénaturée.

Présentation du DLR de Raf : structure et fonction

Plusieurs des aspects structuraux du DLR de Raf ont déjà été discutés dans les sections précédentes où il a servi de modèle à la présentation de notions fondamentales sur la structure des protéines (voir section **Structure native des protéines**). De plus, il en est abondamment question dans les **Articles 2-4** (voir le **Chapitre 2 : Résultats**). Je vais ici m'attarder sur quelques aspects structure fonction.

Le gène c-Raf qui est le centre d'intérêt principal de cette thèse a été originellement cloné à partir de la séquence d'un homologue rétroviral. Chez les vertébrés, outre c-Raf, il existe 2 autres gènes de la même famille soit a-Raf et b-Raf. L'interaction des protéines Raf avec la forme activée de la petite GTPase *ras* [i.e. lorsque conjuguée au guanosine triphosphate (GTP)], qui est induite entre autre par de nombreux facteurs de croissance reconnus par les RTK, déclenche leur phosphorylation à de nombreux sites et conséquemment l'activation de leur domaine kinase. La conséquence la mieux connue de l'induction de l'activité kinase des protéines Raf est la stimulation de la voie de signalisation dite des « kinases activées par les mitogènes aussi connu sous le nom des kinases régulées par les signaux extra-cellulaires » (MAPK/ERK). L'invalidation génétique des gènes Raf chez la souris provoque la mort utérine ou rapidement après la naissance. Cela serait dû respectivement pour les 3 gènes à des problèmes de développement neuronaux et gastro-intestinaux (a-Raf), des retards de croissance, neurologiques et vasculaires (b-Raf) et la mort massive des cellules hépatiques par apoptose (c-Raf). Contrairement à la conception classique qui identifiait c-Raf comme l'oncogène de la famille, il a été démontré récemment que b-Raf était le principal gène Raf muté dans les cancers humains. Je ne m'éterniserai pas sur les inconnus nombreux qui demeurent concernant le processus d'activation des protéines Raf et leurs diverses fonctions suspectées ou avérées, qui ont été discutées superbement ailleurs (*214*).

Les trois gènes Raf encodent des protéines de 75-100 kDa (648 acides aminés pour c-Raf) et comportent une organisation structurale commune arrangée autour de trois

régions conservées. A l'extrémité amino-terminale se trouve la région conservée numéro 1, qui est composée d'un domaine riche en cystéine et du DLR (celui de c-Raf est situé entre les résidus 55-132). Le DLR est suffisant pour l'interaction avec *ras in vitro* alors que l'interaction *in vivo* nécessiterait aussi le domaine riche en cystéine. L'identité de séquence des DLR des divers gènes de Raf chez *H. sapiens* est assez élevé, soit entre 52-56% L'importance physiologique et la modularité du DLR ont poussé la caractérisation rapide par RMN de la structure du domaine issu du gène c-Raf (*215-217*). Tel que cela l'a été présenté précédemment, le DLR forme une structure globulaire se repliant indépendamment et adoptant la topologie dite d'ubiquitine (voir section **Topologie structurale**). Les structures du DLR en complexe avec la GTPase Rap1A et un variant de cette dernière comportant une inversion de charge afin de mimer *ras* ont facilité la délimitation de l'interface de liaison et plus précisément l'identification des résidus impliqués dans la formation du complexe (*218;219*). Pour une raison inconnue, il a été impossible jusqu'à maintenant de déterminer la structure du complexe entre le DLR de Raf et *ras*. Par ailleurs, une autre étude de RMN a permis d'obtenir plus d'informations sur les résidus du DLR situés à longue distance de l'interface, mais qui pourraient être impliqués dans l'interaction en notant les résidus du DLR dont la variation de déplacement chimique était la plus grande en fonction de la concentration de *ras* ajoutée à l'échantillon (*220*). Plusieurs études ont aussi été conduites pour déterminer les résidus du DLR de Raf impliqués dans l'association avec *ras* ou alternativement qui pouvait affecter l'interaction en utilisant des essai de liaison *in vitro* basé sur des analogues fluorescent de GTP non-hydrolysable (*221-225*). En particulier, une étude a répertorié la plupart des résidus déstabilisant directement l'interaction discutée ci-dessus (*222*). D'autres études ont combiné des approches informatiques et expérimentales basées sur l'homologie de séquence entre des domaines potentiels d'interactions avec *ras*[21] afin de discriminer ceux réellement capables de former un complexe avec cette dernière protéine de la masse des

[21] Outre, les DLR à proprement parler (« *ras* binding domain » ou RBD ou RB), qui sont classés selon leur homologie fonctionnelle (i.e. de type Raf ou phosphoinositides-3 kinase), il y a les domaines associés à *ras* (« ras associated » ou RA) que l'on retrouve dans les banques de données tel PFAM et SMART.

protéines regroupées dans ces alignements, et d'identifier les résidus cruciaux à la liaison avec *ras* et Rap1A chez ceux-là (*226;227*). Dans le cas du DLR de c-Raf, l'hélice-α principale semble se comporter différemment, ce qui pourrait indiquer son implication dans une mécanisme distinct de reconnaissance et de stabilisation du complexe. En effet, elle forme une saillie plus prononcée et semble expérimentée un changement conformationel qui altère son arrangement canonique. L'interface de dimérisation du DLR de Raf est une surface basique alors que celle de *ras* est acide. Le résidu R89[22] du DLR est le seul résidu absolument essentiel à cette hétérodimérisation. L'interaction entre Raf et *ras* à cause de l'importance de cette dernière en tant qu'oncogène et du rôle clé de son interaction avec Raf dans la transmission de son signal a été la cible de nombreuses tentatives de synthèse d'inhibiteurs peptidiques et chimiques spécifiques (*228;229*) (réviser dans (*230*)). Un cas intéressant pour nous demeure l'inhibition faible, mais clair de l'interaction induite par un peptide de 7 résidus dont la séquence d'acides aminés correspond aux résidus 95-101 du DLR de c-Raf (*228*). Ce résultat a aussi été confirmé indépendamment par un autre groupe (*229*). Cette région, qui correspond grosso modo au brin-β 3, a démontré plusieurs caractéristiques intéressantes dans nos études et des résultats préliminaires indiquent la présence d'un réarrangement potentiel de la chaîne polypeptidique dans cette région qui prendrait place après la formation de l'état de transition (**Article 2-4** et communications personnelles d'A. Vallée-Belisle).

Le DLR de Raf est un nouveau modèle en repliement de protéines. En effet, outre les articles présentés dans la section **Résultats**, seulement deux autres publications issues de notre laboratoire ont rapporté des résultats concernant le repliement du DLR de c-Raf (*131;231*). Incidemment, ils seront bientôt réunis dans une thèse de Ph.D. en biochimie signée Alexis Vallée-Belisle. De plus, la famille fonctionnelle des DLR de Raf de type *ras* est relativement restreinte et les séquences retrouvées dans les banques de données tel que PFAM et SMART sont relativement peu dégénérées. D'autre part, l'appartenance à la topologie d'ubiquitine fait en sorte que de nombreux analogues structuraux ont été

[22] La mutation R89L abroge toute interaction détectable du DLR avec *ras*.

répertoriés. Tous ces critères font du DLR de Raf un modèle attrayant pour étudier les déterminants de séquence et établir ces liens avec la structure native, le mécanisme de repliement et la fonction de liaison *in vitro* ou intra-cellulaire.

Chapitre 1 : Bases Théoriques

Analyse de l'entropie positionnelle des séquences

En ce qui concerne le calcul de l'entropie de Shannon et des z-scores, le lecteur peut se référer à la section Matériels supplémentaires de l'**Article 2** et aux références ci-contre (*22;38;232;233*).

Analyse *in vitro* de la stabilité et de la cinétique de repliement et de dépliement

Il y a plusieurs prémisses nécessaires à l'interprétation et à la validation biologique des données de repliement obtenues *in vitro*. Le consensus central est que le mécanisme de repliement déterminé *in vitro* par l'utilisation de procédés de dénaturation chimiques ou physiques et celui prenant place *in vivo* pour lequel il est impossible d'obtenir des informations aussi détaillées sont similaires à tout le moins pour les petites protéines possédant la capacité intrinsèque de se replier. Compte tenu des conditions dans lesquelles sont réalisées les expériences qui permettent de mesurer les paramètres de la réaction de repliement, les protéines étudiées doivent démontrer une réversibilité rigoureuse de leur réaction de dénaturation. Deuxièmement, leurs voies de repliement et de dépliement suivant le principe de micro réversibilité sont purement l'inverse l'une de l'autre et par conséquent, elles procèdent via le même état de transition. Finalement, on considère qu'il n'y a qu'une seule espèce native et qu'elle est la plus stable dans les conditions de solution propices au repliement.

Comme cela a été mentionné brièvement ci-dessus, la fraction de la population totale en conformation native versus dépliée et la cinétique de la transition entre ces deux conformations pour une protéine donnée peut être suivie grâce à diverses méthodes spectroscopiques, dont entre autre le DC (longueurs d'onde de l'UV lointains en particulier, où les structures secondaires absorbent principalement), la RMN, la fluorescence intrinsèque de certains acides aminés (i.e. tryptophane et tyrosine), le changement de fluorescence de l'ANS induit par la liaison à des espèces repliées non-natives et le transfert

d'énergie de résonance de Förster (FRET) entre des acides aminés artificiels conjugués à des fluorophores ou entre ces derniers et des acides aminés naturels fluorescents. Bien qu'il faille souligner les progrès récents dans l'étude du repliement et de la stabilité de la structure des protéines au niveau unimoléculaire par FRET et microscopie de force atomique, l'écrasante majorité de nos connaissances actuelles reposent sur des études réalisées à l'aide de méthodes spectroscopiques permettant d'obtenir de l'information sur le repliement concomitant de millions de molécules[23]. Parmi les approches mentionnées ci-dessus, la fluorescence intrinsèque du tryptophane est la plus couramment utilisée pour suivre la renaturation et la dénaturation des protéines. La longueur d'onde d'émission ainsi que l'intensité et l'amplitude du signal de fluorescence produit varient en fonction de l'environnement structural dans lequel se retrouve le résidu utilisé en tant que sonde. À cet égard, les principaux facteurs qui entrent en ligne de compte dans la modulation du signal de fluorescence d'un tryptophane sont son exposition au solvant, plus précisément le contact avec des molécules d'H_2O dans l'état dénaturé et son environnement structural immédiat, c'est-à-dire la proximité de tels ou tels autres groupements d'acides aminés dans l'état natif. A ce propos, le cas du DLR de Raf est instructif. Ce dernier comporte un seul tryptophane (W114), et l'excitation à 280 nm[24] d'un échantillon maintenu en conditions natives, produit un spectre d'émission culminant à environ 335 nm (**Figure 17**). En comparaison, le spectre du DLR de Raf dénaturé est déplacé vers de plus hautes longueurs d'onde (≈ 350 nm) et est élargi, de tel sorte qu'il apparaît identique au tryptophane libre. Par conséquent, la fluorescence intrinsèque du W114 a été exploitée pour effectuer les expériences cinétiques et à l'équilibre (*131*) (voir **Article 2-4**). Finalement, notez que le spectre de fluorescence du DLR natif versus renaturé est similaire (**Figure 17**), ce qui est en accord avec une des prémisses énoncées au début de la présente section.

---

[23] Telle que la population totale d'un échantillon dilué de molécules identiques isolées par des procédures de purification.

[24] Les tryptophanes quels que soient leur environnement ont habituellement un pic dans leur spectre d'excitation autour de 280 nm.

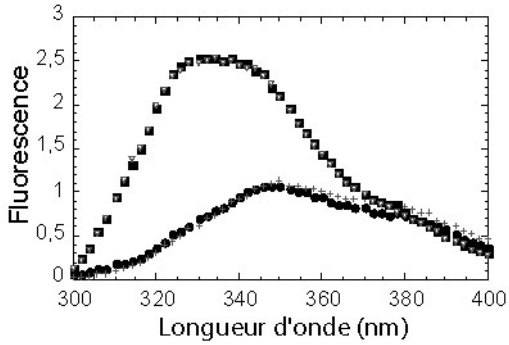


Figure 17. Le spectre de fluorescence du DLR de Raf : les états natif, dénaturé et renaturé.
Le spectre de fluorescence du DLR de Raf excité à 280 nm est principalement dû à W114 et il est équivalent chez les formes native (■) et renaturée (▽). Dans les mêmes conditions le spectre du DLR de Raf sous sa forme dénaturée (●) est quant à lui équivalent à celui du tryptophane en solution (+). Les données utilisées dans ce graphe sont une gracieuseté d'Alexis Vallée-Belisle.

Étude à l'équilibre thermodynamique

La modélisation mathématique des réactions de repliement *in vitro* est fondée sur le modèle de la thermodynamique des réactions chimiques simples (**Figure 18**), donc les équations classiques et les principes de la thermodynamique s'appliquent :

$$\Delta G = \Delta H - T\Delta S \qquad \textbf{(7)}$$

où ΔG est la différence d'énergie de Gibbs, ΔH et ΔS, respectivement la différence d'enthalpie et d'entropie entre le système à l'état de réactif et à celui de produit et T la température. Des études très poussées sur les déterminants enthalpiques et enthropiques du repliement des protéines ont été effectuées par calorimétrie (réviser dans (*163;164;234-236*)). Pour la réaction de repliement, le ΔG correspond à la stabilité de la protéine, soit la différence d'énergie entre l'état déplié versus l'état natif d'une protéine en solution, d'où son annotation sous la forme $\Delta G_{F\text{-}U}$. Pour la réaction de repliement, le ΔG est corrélée à K_{eq} par l'équation suivante :

$$\Delta G_{F-U} = -RT \ln K_{eq} \qquad \textbf{(8)}$$

$$K_{eq} = \frac{[P]}{[R]} \qquad \textbf{(9)}$$

où P et R sont respectivement le produit et les réactifs de la réaction, et RT le produit de la constante des gaz parfaits et de la température.

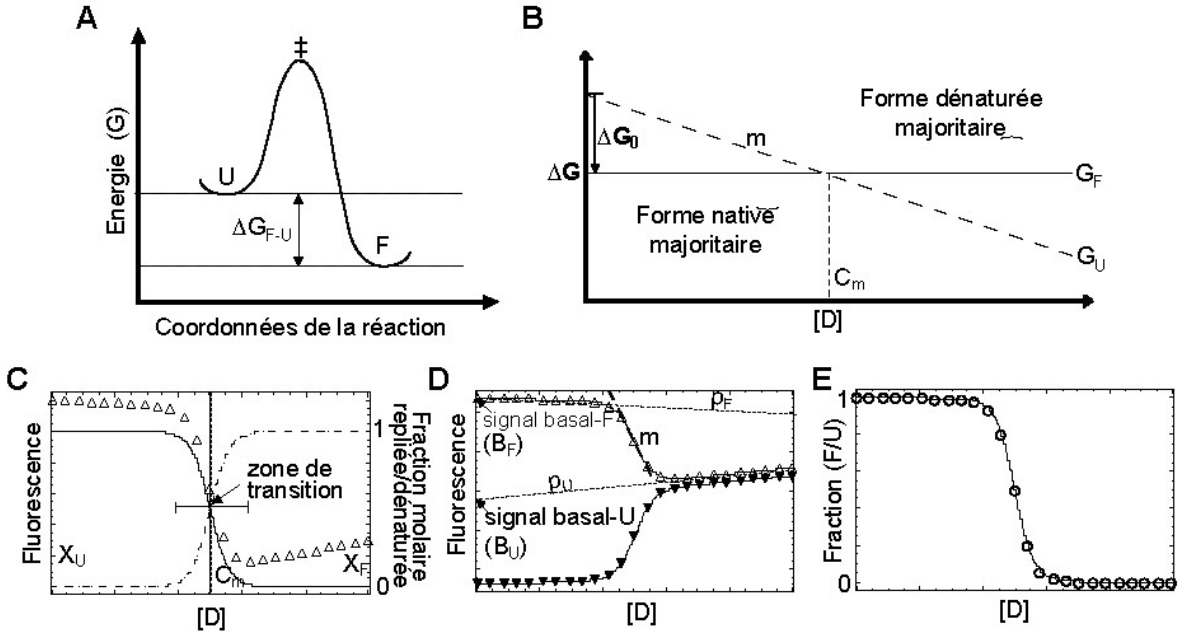


Figure 18. Les notions de base pour la compréhension des expériences réalisées à l'équilibre et des données qui sont extraites de la courbe de dénaturation.
A, Un diagramme d'énergie de Gibbs pour une réaction de repliement de type deux-états, i.e. un état dénaturé (U) se repliant vers l'état natif (F) via un état de transition (‡) en absence de tout autre état intermédiaire. Le ΔG_{F-U} correspond à la différence d'énergie du système lorsque la protéine est en son état natif versus dénaturé. Cela correspond aussi à la stabilité de F. **B**, Variation de l'énergie de F et U et par conséquent du ΔG_{F-U} en fonction de la concentration de dénaturant ([D]). Dans ce cas de figure, la pente de la droite indiquant la variation de G_U en fonction de la [D] correspond à la valeur m. La concentration de dénaturant à laquelle $\Delta G = 0$ est nommé C_m et elle correspond au point milieu de la zone de transition de la courbe de dénaturation. **C**, Courbe de dénaturation pour une protéine modèle se repliant en deux-états suivit par la variation de fluorescence. Les courbes X_U et X_F correspondent respectivement à la fraction F/U et U/F de la protéine modèle en fonction de la [D]. La zone de transition correspond à la région qui voit les ratios U/F et F/U s'inversés. **D**, La fluorescence peut diminuer ou augmenter lors de la dénaturation selon la sonde la sonde employée. Le lissage de la courbe de dénaturation (équation **(12)**) permet d'extrapoler la valeur de m, ΔG_0, P_F (pente plateau état natif), B_F (extrapolation de la fluorescence basale à partir du plateau de l'état natif), P_U (pente plateau état dénaturé) et B_U (extrapolation de la fluorescence basale à partir du plateau de l'état dénaturé). **E**, À partir des données issues de la courbe de dénaturation brute, il est possible de tracer une courbe de dénaturation de la fraction F/U de la protéine suivant la [D] (équations **(13;14)**). Cela à l'avantage de permettre la comparaison entre des expériences distinctes ou bien entre des protéines ayant des propriétés de fluorescence différentes.

L'équilibre entre la forme native et dépliée peut-être altéré par l'utilisation de dénaturant (**Figure 18**). L'addition progressive de dénaturant stabilise la conformation dénaturée de la protéine (augmente le ΔG_{F-U} en diminuant G_U) sans affecter l'énergie de la forme native (G_F), de sorte que sa population passe progressivement d'une région où la

structure native prédomine à une région ou c'est plutôt la structure dénaturée qui est majoritaire. La variation linéaire de ΔG_{F-U} en fonction de la concentration de dénaturant est exprimée par l'équation suivante :

$$\Delta G_{F-U} = \Delta G_0 + m \times [D] \qquad \textbf{(10)}$$

$$C_m = \frac{\Delta G_0}{m} \qquad \textbf{(11)}$$

où ΔG_0 représente l'énergie libre entre l'état dénaturé et natif de la protéine (i.e. la stabilité) en absence de dénaturant, m est le taux de variation de ΔG_{F-U} en fonction de la concentration de dénaturant et Cm est la concentration de dénaturant à $\Delta G_{F-U} = 0$, soit au point milieu de la zone de transition de la courbe de dénaturation. La valeur m correspond au taux d'ouverture de la protéine entre l'état dénaturé et l'état natif qui est induit par l'augmentation de la perturbation physique ou chimique pertinente. La valeur m augmente en fonction du pouvoir dénaturant de la procédure employée (les dénaturants chimiques utilisés les plus couramment sont en ordre croissant de pouvoir dénaturant [25]: urée, Gdm-HCl et le thiocyanate de guanidine). Les valeurs m et ΔG_0 pour une protéine donnée sont obtenues à partir de sa courbe de dénaturation à l'équilibre (**Figure 18**). La variation de fluorescence en fonction de la concentration de dénaturant suit une courbe sigmoïde dans laquelle les plateaux à basse et à haute concentration représentent respectivement l'intervalle de concentration où les conformations native et dénaturées de la protéine prédominent. Les variations de la population de molécule de polypeptide dans les états natif et dénaturé sont également représentées dans le même panneau. L'intervalle de concentration de dénaturant à laquelle les populations de ces deux états s'éloignent abruptement de leur population relative tel qu'observée en conditions normales correspond

[25] L'urée et le Gdm-HCl sont les dénaturants utilisés le plus fréquemment en repliement de protéine. L'hypothèse de la linéarité de la dépendance de ΔG_{F-U} en fonction de la concentration d'urée serait correcte alors qu'il y a des indications d'une légère déviation concave avec le Gdm-HCl, menant donc à une sous évaluation de ΔG_0 (*180*). Les résultats obtenus avec le DLR de Raf sont en accord avec cela (comparez par exemple, les données retrouvées dans (*131;231*) et l'**Article 3**).

à la zone de transition située entre les deux plateaux de la courbe de dénaturation. A l'aide de l'équation suivante, une courbe de ce type peut être lissée afin de déterminer la valeur de ΔG_0 et m pour une protéine dans des conditions données de température et de dénaturant:

$$I_f = B_F + P_F \times [D] + \left((B_U + P_U \times [D]) \times \frac{e^{((\Delta G_0 + m \times [D])/-RT)}}{\left(1 + e^{((\Delta G_0 + m \times [D])/-RT)}\right)} \right) \quad \textbf{(12)}$$

où I_f correspond à l'intensité du signal de fluorescence, P_F et P_U correspondent respectivement à la pente du plateau du signal de fluorescence à l'état natif et dénaturé et B_F et B_U correspondent respectivement au signal basal de fluorescence en absence de dénaturant extrapolée à partir de la pente du plateau de l'état natif et dénaturé.

Cette équation est générale et permet de calculer la valeur de ces paramètres quels que soit la méthode utilisée pour suivre la réaction de dénaturation (**Figure 18**). Par ailleurs, il est intéressant de comparer les courbes de dénaturation obtenues dans diverses conditions expérimentales ou bien simplement lors d'expériences distinctes. Pour ce faire, chaque mesure de l'intensité de signal de fluorescence doit être transformée en la fraction de la population adoptant la conformation native à une concentration de dénaturant donnée, soit F_F, afin d'obtenir des courbes de dénaturation comparables :

$$F_F = \frac{I_f - (B_U + P_U \times [D]) / \left(1 + e^{((-\Delta G_0 + m \times [D])/-RT)}\right)}{B_F + P_F \times [D]} \quad \textbf{(13)}$$

Une fois les données converties de cette manière, elles peuvent être lissées grâce à l'équation suivante :

$$F_F = \frac{e^{((\Delta G_0 + m \times [D])/-RT)}}{\left(1 + e^{((\Delta G_0 + m \times [D])/-RT)}\right)} \quad \textbf{(14)}$$

Cela permet de comparer visuellement toutes les courbes de dénaturation, car elles sont alors limitées à des valeurs entre 0 et 1sur l'axe des ordonnées. De cette manière, les

données récoltées lors d'expériences de renaturation et de dénaturation peuvent être combinées afin de tracer une courbe d'équilibre comportant plus de points expérimentaux et donc d'obtenir plus de précisions dans l'évaluation des paramètres thermodynamiques ΔG_0 et m (*237*).

Études cinétiques

La cinétique du repliement ou du dépliement d'une protéine est couramment suivie grâce à un fluorimètre couplé à un appareil de mixage rapide de type flux interrompu (« stopped-flow »), car ces réactions sont habituellement trop rapides chez les protéines se repliant en deux-états pour être suivies sans l'aide d'un tel dispositif (*238*) (**Figure 19**).

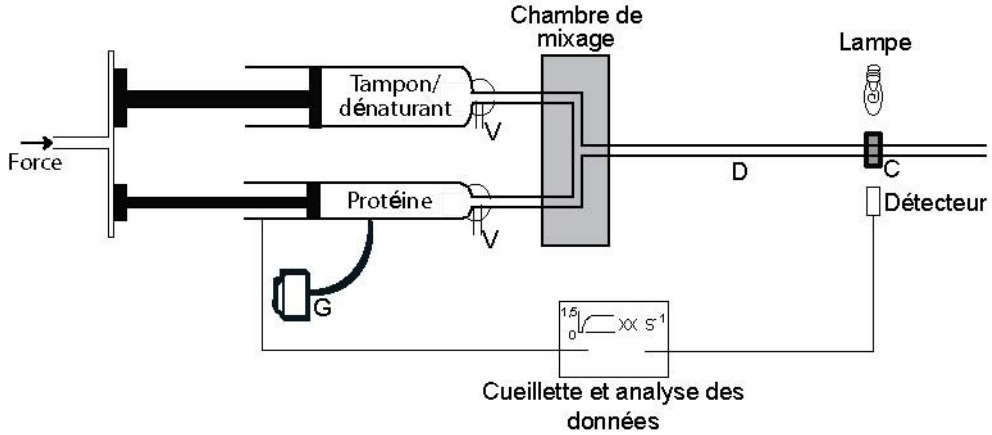


Figure 19. Schéma expliquant le fonctionnement d'un appareil de mixage rapide de type flux interrompu. L'appareil de mixage rapide que nous avons utilisé est de style flux interrompu (« stopped-flow »). Ce genre d'appareil est utilisé couramment pour étudier la cinétique de repliement/dépliement (ce schéma est inspiré de (*238*)) : gâchette (G), valve (V), boucle de délai (D), cellule d'observation (C). En bref, à un moment choisi par l'expérimentateur, une force est appliquée conjointement sur les pistons des deux seringues ce qui permet l'injection d'un certain volume (réglable) de la solution de dénaturant et de la solution de protéine qui sont mélangées dans la chambre de mixage. L'initiation de l'injection enclenche le début de la récolte des données par l'ordinateur. La réaction est suivie uniquement dans la cellule d'observation après le passage dans une boucle de délai favorisant le mixage des deux solutions.

Ce type d'appareil permet la dilution et le mixage d'un échantillon de protéine dans une solution de dénaturant, moyennant un court délai[26], que l'on nomme temps mort, entre le début du mixage et par conséquent, le commencement de la réaction et l'initiation de la

[26] Il est de l'ordre de quelques millisecondes pour les appareils contemporains, spécifiquement de 2,8 ms pour le SX 18MV d'Applied Biophysics Inc., soit l'appareil que j'ai employé.

récolte des données de fluorescence par le détecteur. Pour les expériences de renaturation un échantillon de protéine en condition dénaturée est dilué dans une solution contenant des concentrations variables de dénaturant, qui sont suffisamment faibles pour permettre la renaturation et, vice-versa pour les expériences de dénaturation (**Figure 20**).

Le taux de la réaction peut être déterminé dans le cas des traces cinétiques comportant une seule phase grâce à une fonction exponentielle simple :

$$I_{(t)} = A \times e^{(-k \times t)} + B_f \qquad \textbf{(15)}$$

où $I_{(t)}$ est l'intensité de fluorescence en fonction du temps t depuis l'initiation de la réaction, A est l'amplitude du changement dans le signal qui sert à suivre la réaction, k est le taux de la réaction (k_f et k_u sont respectivement les taux de repliement et de dépliement) et B_f correspond à l'extrapolation du signal de fluorescence à partir du plateau de la trace expérimentale. La détermination de k_f et de k_u, respectivement à partir de protéines dénaturées et repliées et à diverses concentrations de dénaturant par la dilution dans des solutions comportant des concentrations séquentiellement croissantes ou décroissantes de dénaturant est utilisée afin de construire ce que l'on nomme couramment une courbe de chevron (**Figure 20**) :

$$\ln k_{obs} = \ln\left(k_f^{H2O} \times e^{(-m_f \times [D])} + k_u^{H2O} \times e^{(m_u \times [D])}\right) \qquad \textbf{(16)}$$

où m_f et m_u sont respectivement le taux de variation de $\ln k_f$ et $\ln k_u$ en fonction de la concentration de dénaturant. Ces paramètres sont indicatifs respectivement du degré de variation de l'accès du solvant à la chaîne polypeptidique entre l'état dénaturé et l'état de transition et entre ce dernier et l'état natif. La variation de $\ln k_f$ et $\ln k_u$ en fonction de la concentration de dénaturant est linéaire, ce qui produit une courbe en forme de « v », d'où le qualificatif de chevron. La base de ces courbes est arrondie, car elle correspond à la région où les réactions de repliement et de dépliement peuvent être toutes deux observées, alors qu'elles dominent respectivement aux concentrations de dénaturant inférieures et

supérieures à la région de transition. La linéarité de la relation entre les taux permet de déterminer par extrapolation la valeur de k_f et k_u en absence de dénaturant (respectivement k_f^{H2O} et k_u^{H2O}).

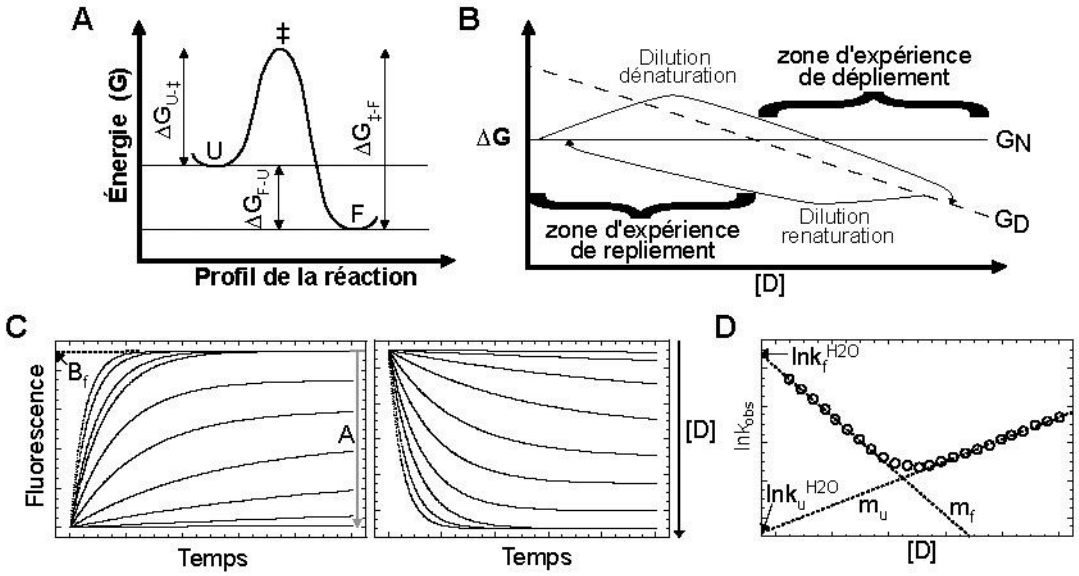


Figure 20. Fondements théoriques et expérimentaux nécessaires à la compréhension et à l'analyse des expériences de cinétiques.
A, Diagramme d'énergie de Gibbs présentant les variables, $\Delta G_{U-‡}$ et $\Delta G_{‡-F}$, qui sont importantes à la définition des propriétés cinétiques, i.e. des taux de la réaction de repliement et de dépliement (équation **(23;24)**). **B**, Les expériences permettant d'étudier la cinétique de repliement sont effectuées en mélangeant un échantillon de protéine initialement en condition dénaturante avec une solution à faible [D], qui permet la formation de la structure native (dilution renaturation). L'inverse est vrai pour les expériences de dépliement (dilution dénaturation). **C**, Traces obtenues pour les réactions de repliement et de dépliement en fonction de [D] croissante d'une protéine hypothétique dont la fluorescence de l'état natif est supérieure à celle de l'état dénaturé (où A, correspond à l'amplitude et B_f, au signal de fluorescence extrapolé à partir du plateau). Le taux observé (k_{obs}) à chaque condition de [D] est déterminé par le lissage à l'aide d'une fonction exponentielle simple (dans ce cas-ci) ou plus complexe (équations **(15;17;18)**). **D**, Lorsqu'une protéine démontre un repliement de type deux-états la combinaison dans un graphe des k_{obs} (i.e. k_f ou k_u en fonction de [D]) extrapolé à diverses [D] séquentielles forme une courbe de chevron. Le lissage de cette courbe (équation **(15)**) permet de déterminer par extrapolation k_f^{H2O} et k_u^{H2O}, c'est-à-dire k_f et k_u pour la [D]= 0.

Cependant, des déviations vers le bas sur l'un, l'autre ou les deux bras de la courbe de chevron surviennent respectivement à faible et/ou à haute concentration de dénaturant dans la réaction de repliement et/ou de dépliement chez environ 50% des protéines se repliant par un mécanisme deux-états apparent (un exemple est présenté à la **Figure 21**). Ces déviations peuvent être l'indication d'un mouvement de l'état de transition sur une large barrière d'énergie ou bien d'un changement de l'état de transition sur une voie de repliement comportant plusieurs transitions et des intermédiaires de haute énergie (*127*).

Tel que discuté précédemment dans le cas d'ubiquitine (voir section **Preuves de la présence d'intermédiaires cinétiques dans le repliement d'ubiquitine**), des erreurs inhérentes au temps mort dans le lissage des traces de repliement peuvent aussi potentiellement mener à ce genre de déviations, particulièrement dans le cas des protéines possédant un k_f près de la limite de résolution de l'appareil utilisé (*131;197*). L'agrégation ou l'homodimérisation spécifique de la protéine à faible concentration de dénaturant doivent aussi être considérées comme des facteurs pouvant induire la déviation du bras de repliement de la courbe de chevron[27]. Au contraire, des déviations dans la linéarité qui mèneraient à de plus hautes valeurs de m_f et m_u, respectivement à faible et à haute concentration de dénaturant seraient indicatives de voies de repliement parallèles (voir les sections **Intermédiaires de haute énergie : études cinétiques**, **Comportement d'Hammond et d'anti-Hammond** et **Entonnoir et multiplicité des voies de repliement**). Par ailleurs, il arrive assez souvent que les traces des expériences de renaturation ne puissent être lissées correctement à l'aide d'une fonction exponentielle simple, alors que les traces cinétiques de dépliement démontrent très rarement de telles complications (**Figure 21**).

Dans tous les cas où de telles complications surviennent, il faut tenter de lisser les traces obtenues avec une fonction bi exponentielle, tri exponentielle, etc., en fonction du niveau de complexité apparent :

$$I_f = A_1 \times e^{(-k_1 \times t)} + A_2 \times e^{(-k_2 \times t)} + B_f \qquad \textbf{(17)}$$

$$I_f = A_1 \times e^{(-k_1 \times t)} + A_2 \times e^{(-k_2 \times t)} + A_3 \times e^{(-k_3 \times t)} + B_f \qquad \textbf{(18)}$$

Afin de sélectionner l'équation permettant le lissage le plus adéquat des données expérimentales, il est utile de calculer la fonction résiduelle entre les traces et la fonction mathématique selectionnée (**Figure 21**). Dans le cas de réaction multi phasique, la phase correspondant à la transition deux-états est choisie en accord avec les données de

[27] Dans ce cas, la déviation devrait être corrélée à la concentration de la protéine dans l'échantillon.

expériences réalisées à l'équilibre thermodynamique (voir la section précédente et la suivante).

L'occurrence d'une ou plusieurs prolines dans la séquence d'une protéine est la cause la plus fréquente de l'apparition de phases supplémentaires dans les traces de repliement. En effet, la différence d'énergie entre la conformation trans et cis du lien imide de la proline, à 85 kJ/mol, est plus faible que pour les autres acides aminés. Cela fait en sorte que pour un polypeptide dénaturé contenant une proline, entre 10-30%[28] des molécules arborent la conformation cis à ce lien. Or, environ 7% des prolines adoptent la conformation cis dans la structure native des protéines (*239*). Dans ces cas, il est fort possible que l'isomérisation de ce lien constitue l'étape limitante de la réaction de repliement. Cette complication peut être habituellement évitée en choisissant un modèle d'études ne comportant pas de prolines en conformation cis dans sa forme repliée. Même si c'est le cas, la présence d'une ou des prolines dans la séquence polypeptidique peut provoquer l'apparition de phases plus lentes plus ou moins importantes en amplitude (0,1-0,001 s^{-1}) qui sont dues à la fraction de lien(s) imide(s) proline en conformation cis dans l'état dénaturé qui doivent s'isomériser avant de trouver leur environnement natif. Il y a des moyens expérimentaux, tel que les expériences de double dilution[29] (« double-jump ») ou alors l'ajout de protéine catalysant l'isomérisation du lien imide des prolines (peptide-prolyl isomérase) à l'échantillon lors de la réaction de repliement, qui permettent de vérifier que les phases lentes du repliement sont bien dues à ce phénomène (*240*). La première de ces approches a été utilisée dans le cas du repliement du DLR de Raf afin de déterminer si la présence de deux prolines dans sa structure primaire pouvait expliquer la présence des 3

---

[28] Cela varie en fonction de l'identité de l'acide amine qui est impliqué dans la formation du lien imide et de son environnement électrostatique.

[29] Ce genre d'expérience est réalisé à partir d'un échantillon de protéines en conditions natives qu'il faut traiter de la façon suivante : d'abord transfert en conditions dénaturantes, suivie après un court laps de temps par une seconde dilution, cette fois en conditions natives. Les propriétés de chacune des phases dans les traces de repliement obtenues par ce protocole sont comparées à celles obtenues par le protocole de dilution directe (i.e. échantillon de protéine en condition dénaturante diluer en condition native). Voir les référence ci-jointes pour un exemple de ce genre d'analyse avec le DLR de Raf (*131*) (voir la section **Preuves de la présence d'intermédiaires cinétiques dans le repliement** d'ubiquitine).

phases lentes observées dans les traces de repliement obtenues à faible concentration de dénaturant (*131*).

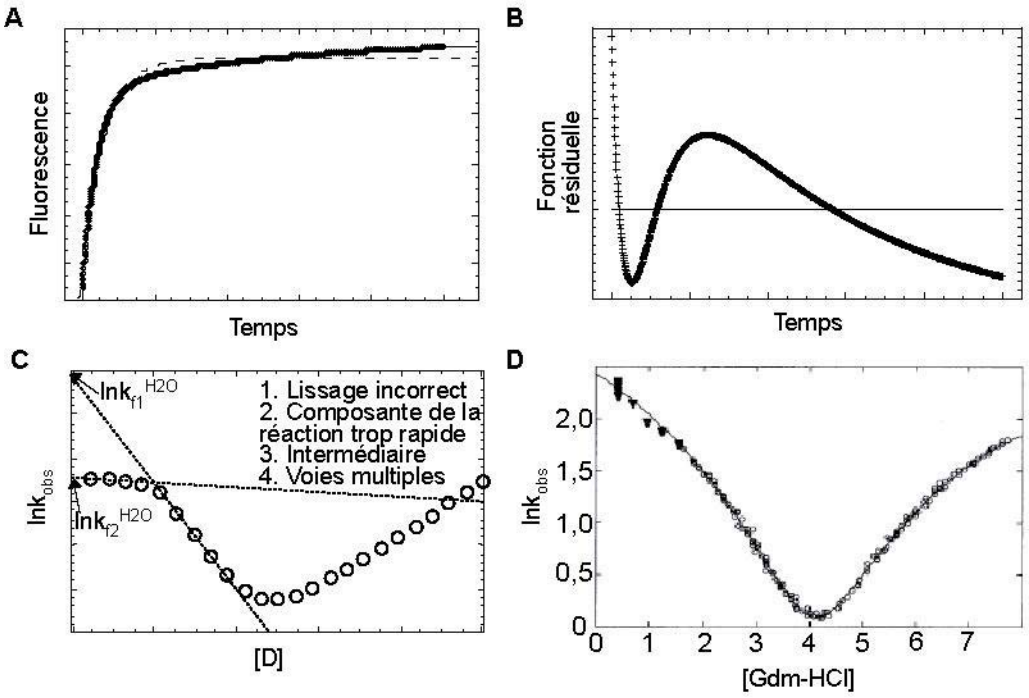


Figure 21. Réaction de repliement multi exponentielle et déviation de la linéarité des courbes de chevron. **A**, Modèle d'une trace hypothétique d'une réaction de repliement pour laquelle le lissage par une fonction simple exponentielle (---) est inadéquat, alors que le modèle bi exponentielle (—) est plus satisfaisant. **B**, Cela peut-être vérifié par le calcul de la fonction résiduelle qui détermine l'écart entre la fonction de lissage et les données du modèle expérimental : fonction résiduelle pour la fonction simple exponentielle (+) et pour la fonction bi exponentielle (---). **C**, Un lissage incorrect peut mener à une déviation dans la courbe de chevron, tel que celle représentée ci-dessus. Les facteurs annotés dans ce panneau peuvent induire ce genre de déviation dans la linéarité de k_f ou k_u en fonction de la [D], soit à faible et à haute [D], respectivement. D'autre part, des voies de repliement et de dépliement parallèles mènent à des déviation vers le haut de la relation entre le k_{obs} et la [D]. Ce phénomène a rarement été rapporté (voir section **Comportement d'Hammond et d'anti-Hammond**). **D**, La courbe de chevron de la protéine U1A montre des déviations sur le bras de repliement et de dépliement (tirer de Ternstrom et coll. (*113*) avec la permission de l'auteur responsable). Ce phénomène a été attribué à un mouvement de l'état de transition vers l'état natif.

Comparaison des données des expériences cinétiques et à l'équilibre thermodynamique

Afin de vérifier la validité de la modélisation mathématique des traces cinétiques et de s'assurer que le repliement suit un mécanisme deux-états, les principaux paramètres thermodynamiques extraits des expériences réalisées à l'équilibre sont comparés à ceux obtenus à partir des données cinétiques et des équations suivantes :

$$K_{eq} = \frac{k_f^{H2O}}{k_u^{H2O}} \qquad \textbf{(19)}$$

En combinant les équations **(8)** et **(19)**, on obtient :

$$\Delta G_0 = -RT \ln \frac{k_f^{H2O}}{k_u^{H2O}} \qquad \textbf{(20)}$$

De la même manière, la valeur de m peut aussi être dérivée :

$$m = -RT \times (m_f + m_u) \qquad \textbf{(21)}$$

En effet, la concordance entre les valeurs de ΔG_0 (ou plus généralement $\Delta G_{F\text{-}U}$) et de m extrapolée à partir des données d'expériences à l'équilibre et cinétiques est un critère essentiel pour confirmer que la réaction de repliement d'une protéine est de type deux-états. Lorsque plusieurs mutants sont caractérisés comme dans le cadre d'une étude classique d'ingénierie des protéines, l'équivalence et la corrélation des $\Delta\Delta G_{F\text{-}U}$ calculées à partir des données d'équilibre et des cinétiques sont habituellement présentées afin de démontrer le caractère deux-états de la transition, car ce paramètre est moins sensible aux erreurs d'extrapolation (équation **(22)** et (*98*)). Des déviations dans la similitude des estimés peuvent indiquer des mécanismes de repliement plus complexes ou si elles sont limitées à un nombre restreint de mutants, cela peut signifier que des changements dans la voie de repliement ont pris place. De plus, dans le cas d'une réaction de repliement dont les traces sont multiphasiques, cela permet aussi de déterminer la phase qui correspond à la transition deux-états. Il est courant de mesurer $\Delta G_{F\text{-}U}$ à partir des données cinétiques et à l'équilibre pour des concentrations plus élevées de dénaturant afin de minimiser l'erreur due à l'extrapolation.

Ingénierie de protéines, analyse des valeurs-Φ et position de l'état de transition

La méthode d'ingénierie des protéines est basée sur l'introduction de manière ponctuelle ou en combinaison de mutations non-disruptives tel que : I→V→A→G, Y→F→A, E→D→A, Q→N→A, T→S→A et A→G ou X→A, où X représente tous les (autres) acides aminés (*78*), à des résidus qui sont typiquement choisis par l'inspection de la structure ou grâce à l'alignement d'analogues structuraux. Dans un premier temps, les effets de chaque mutation ou combinaison(s) de mutations (mut) introduites sur la stabilité de la protéine native ou sur les taux des réactions de repliement et de dépliement sont mesurés en les comparant aux propriétés de la la protéine native de ts:

$$\Delta\Delta G_{F-U} = -RT \ln \frac{\Delta G^{mut}_{F-U}}{\Delta G^{ts}_{F-U}} \qquad \textbf{(22)}$$

$$\Delta G_{U-\ddagger} = -RT \ln k_f \qquad \textbf{(23)}$$

$$\Delta G_{\ddagger-F} = -RT \ln k_u \qquad \textbf{(24)}$$

$$\Delta\Delta G_{U-\ddagger} = -RT \ln \frac{k^{mut}_f}{k^{ts}_f} \qquad \textbf{(25)}$$

$$\Delta\Delta G_{\ddagger-F} = -RT \ln \frac{k^{mut}_u}{k^{ts}_u} \qquad \textbf{(26)}$$

Dans la pratique, divers estimés de $\Delta\Delta G_{F-U}$ peuvent être calculés à partir des données recueillies d'expériences cinétiques et à l'équilibre (en suivant les grands principes de la section précédente). La comparaison de ces paramètres équivalents est importante pour la validation de la qualité des données et du lissage des courbes (pour une description des divers types d'estimés voir (*98*) et l'**Article 4**). Similairement, les paramètres $\Delta\Delta G_{U-\ddagger}$, $\Delta\Delta G_{\ddagger-F}$, $\Delta G_{\ddagger-F}$ et $\Delta G_{U-\ddagger}$ utilisés dans le calcul des valeurs-Φ sont souvent déterminés à faible concentration de dénaturant plutôt qu'en absence de dénaturant afin de réduire les erreurs inhérentes à leur extrapolation à [D]= 0. A partir de ces paramètres, il est possible de mesurer le rôle des atomes enlevés et par extension du résidu muté en calculant sa

valeur-Φ (*78;79;238;241*). Ce paramètre correspond au ratio des variations d'énergie de l'état de transition versus l'état natif pour les mutants en comparaison avec le ts :

$$\Phi_F = \Delta\Delta G_{U-\ddagger}/\Delta\Delta G \qquad \textbf{(27)}$$

Alternativement, la valeur-Φ_U, qui correspond à l'inverse de Φ_F, est calculée à partir du taux de dépliement :

$$\Phi_U = \Delta\Delta G_{\ddagger-F}/\Delta\Delta G \qquad \textbf{(28)}$$

$$\Phi_F = 1-\Phi_U \qquad \textbf{(29)}$$

Suivant le principe de la micro réversibilité, les valeurs-Φ_F obtenues à partir de l'expérience de repliement et de dépliement doivent être équivalentes[30]. Si la valeur-Φ_F est de 1, cela signifie que la totalité des interactions natives brisées par la mutation sont déjà formées à l'état de transition et vice versa si le ratio est de 0. Les mutations produisant des valeurs-Φ entre 0 et 1 sont plus complexes à interpréter, mais l'interprétation la plus largement acceptée est que les interactions formées par une chaîne latérale ou un atome avec une valeur-Φ de ce type sont partiellement consolidées à l'état de transition (*122*). Ces ratios permettent de mesurer la contribution d'un résidu à la stabilisation de l'état de transition versus l'état natif (**Figure 22**). L'analyse des valeurs-Φ pour des mutations réparties sur l'ensemble de la structure permet de déterminer une structure approximative de l'état de transition (voir la section **Ingénierie des protéines**).

Un certain nombre de règles sont posé afin de permettre l'interprétation la moins ambiguë possible des valeurs-Φ obtenues chez une protéine donnée:

1. La mutation ne modifie pas la voie de repliement

[30] Des différences peuvent apparaître si m_f et/ou m_u du mutant et du ts sont significativement différents.

2. La mutation ne change pas significativement la structure de l'état natif et de l'état dénaturé.

3. La mutation n'introduit pas de nouvelles interactions durant le processus de repliement.

Bien sûr, ces prémisses sont autant de vœux pieux. En fait, elles constituent les conditions optimales pour que l'interprétation des résultats soit significative. Or, il est impossible de vérifier de façon directe l'applicabilité de ces règles dans le cadre de chaque analyse. Par ailleurs, la comparaison des courbes de chevron, en particulier des variations prononcées dans les paramètres m_f et m_u permet superficiellement de vérifier la prémisse 1. De plus, afin d'assurer le respect optimal de ces prémisses, des mutations non-disruptives (i.e., la taille de la chaîne latérale est réduite) sont privilégiées. Malgré ces précautions, des cas où ces limitations à l'interprétation des expériences d'ingénierie des protéines ne sont pas respectées surviennent fréquemment, tel qu'indiquer par l'observation de valeurs-Φ inférieures à 0 ou supérieures à 1. Des discussions intéressantes concernant des interprétations possibles pour ces valeurs-Φ atypiques sont discutées dans les références ci-jointes (*124;206;242-245*).

Mouvement de l'état de transition

La position relative de l'état de transition sur le profil de la réaction entre l'état dénaturé et l'état natif peut être évaluée à partir des données cinétiques en calculant le facteur β-Tanford (β_T) :

$$\beta_t = \frac{m_f}{(m_u + m_f)} \qquad \textbf{(30)}$$

Ce paramètre fournit une indication de la similarité structurale entre l'état de transition et les états observables à l'équilibre, une valeur près de 1 indiquant une proximité structurale avec la structure native et vice-versa pour 0. Par « proximité structurale » j'entends en fait une mesure relative de l'accessibilité au solvant de l'état de transition par rapport aux deux états fondamentaux. Habituellement, ce facteur est propre à une protéine donnée et il est

relativement résistant à l'introduction de mutations ponctuelles dans la séquence. Lorsque ce paramètre varie, cela peut indiquer la présence de mouvements de l'état de transition suivant les postulats d'Hammond et d'anti-Hammond. Cependant, ils peuvent aussi être dus

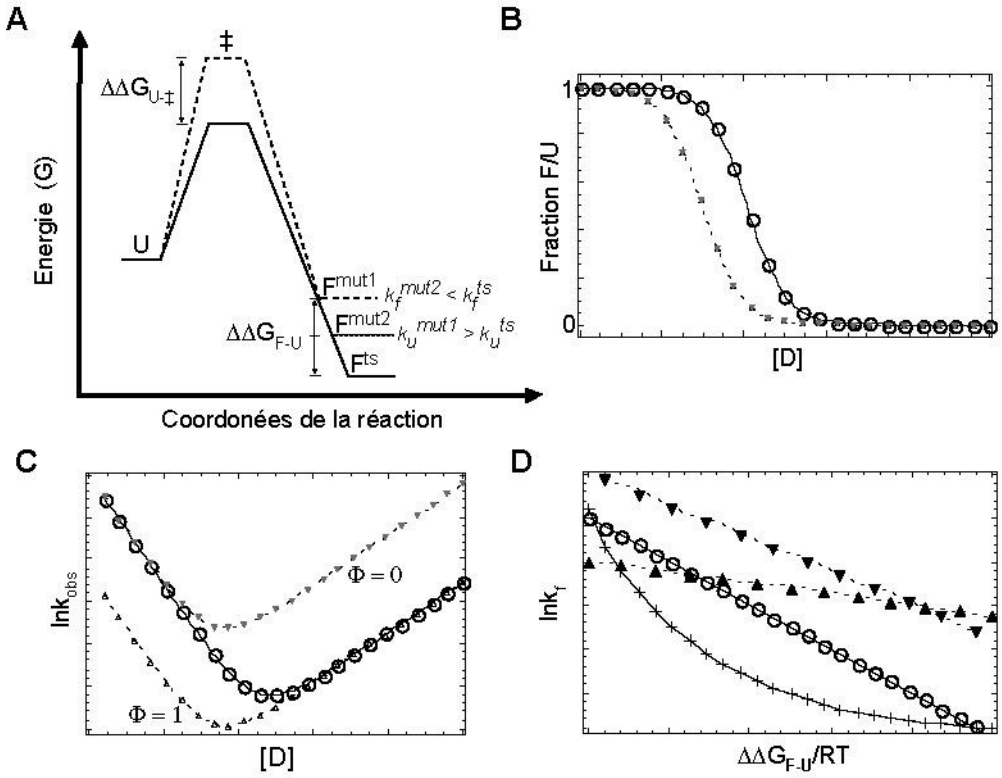


Figure 22. Bases théoriques et expérimentales pour l'analyse des valeurs-Φ.
A, Diagramme d'énergie pour une protéine modèle et deux de ces mutants qui en sont issus. Ces mutants (mut1 et mut2) sont déstabilisés dans la même mesure, mais à cause de composantes différentes des paramètres biophysiques de leur structure. En effet, le mutant1 (mut1) a un ‡ plus haut en énergie que le ts. Le mutant 2 (mut2) a un ‡ équivalent au ts. Dans le cas du premier mutant, seul k_f varie alors que dans le second cas, c'est k_u. L'analyse des valeurs-Φ repose sur la comparaison de $\Delta\Delta G_{U-‡}$ versus $\Delta\Delta G_{F-U}$ (équations **(22-29)**; **Figure 18** et **Figure 20**). **B**, Pour cette protéine hypothétique, les mutants sont tous deux déstabilisés dans la même mesure par rapport au ts (ts : ○; mut1 : ∆; mut2 : ▲). **C**, Les courbes de chevron correspondantes pour ces mêmes mutants. En accord avec ce qui a été dit dans les panneaux précédents mut1 et mut2 ont une valeur-Φ de 1 et 0 respectivement. **D**, Le graphe de Leffler (alias Brønstead) permet de visualiser la relation entre $\ln k_f$ et $\Delta\Delta G_{F-U}/RT$. Pour certaines protéines qui démontrent un état de transition diffus tous les mutants se regroupent sur une droite (○). Alternativement, pour d'autres protéines démontrant un état de transition polarisé les mutants sont regroupés sur la base de leur localisation dans des régions diverses de la structure (▲et ▼). Dans le cas hypothétique de voies de repliement parallèles (+), on observe en lieu et place de corrélations linéaires une courbe. En effet, au-delà d'un certain niveau de déstabilisation, il y aurait transfert de la voie de repliement la plus rapide vers une plus lente (voir les sections **Comportement d'Hammond et d'anti-Hammond** et **Entonnoir et multiplicité des voies de repliement**) (*241*).

à des changements dans les propriétés des états fondamentaux, et cette possibilité doit être écartée afin de confirmer le mouvement de l'état de transition (voir les sections **Comportement d'Hammond et d'anti-Hammond** et **L'état déplié et/ou dénaturé**) (*125*).

Pour une collection de mutants, Fersht et coll. ont proposé de comparer la dépendance de $\ln k_f$ ou $\ln k_u$ en fonction de la variation de stabilité ($\Delta\Delta G_0/RT$) à l'intérieur d'un graphe dit de Leffler (alias Brønstead) suivant les équations suivantes :

$$\ln k_f = \ln k_f^{wt} + \alpha_f x(\Delta\Delta G_{F\text{-}U}/RT) \qquad \textbf{(31)}$$

$$\ln k_u = \ln k_u^{wt} + (1-\alpha_f) x(\Delta\Delta G_{F\text{-}U}/RT) \qquad \textbf{(32)}$$

où α_f correspond à une valeur-Φ moyenne de tous les mutants considérés dans la corrélation linéaire. De la façon dont se répartisse les mutants, il est possible de poser les hypothèses concernant le processus de repliement à savoir, cette protéine se replie-t-elle via une voie de repliement unique ou multiple et surtout de déterminer la structure générale de l'état de transition, c'est-à-dire polarisée versus diffuse (*122;125;241*) (voir les exemples de la **Figure 22** et la section Graphe de Leffler).

Chapitre 2 : Résultats

Objectifs généraux de la thèse

L'objectif de cette thèse est de faire le lien d'une part, entre les déterminants de la séquence polypeptidique qui permettent l'encodage de la topologie structurale et d'autre part, le processus de repliement et la stabilisation de la structure native. Nous avions décidé d'approcher ces questions en nous servant de la haute diversité de séquence rencontrée chez les protéines adoptant la topologie d'ubiquitine afin d'énoncer des hypothèses sur le repliement et la stabilisation de la structure modèle du DLR de Raf. Un objectif secondaire était de développer des approches expérimentales qui pourraient être généralisables à l'exploration des déterminants de séquence. La capacité de réaliser cela pour les protéines consolidant des topologies structurales rares permettrait l'expansion de la diversité des séquences de ces dernières et pourrait contribuer à plus long terme au développement d'outils et d'algorithmes informatiques pour le design et la prédiction de structures.

Article 1 : Développement de stratégies expérimentales nécessaires à la synthèse de librairies dégénérées et à leur sélection par le PCA de DHFR : applications au DLR de Raf.

Article accepté dans « Methods in Molecular biology » vol. 352, p. 249-74

Présentation de l'article 1 :

Cet article présente des protocoles expérimentaux détaillés permettant la synthèse de librairies dégénérées de la structure primaire d'une protéine et la sélection, par l'entremise de la conservation de leur fonction de liaison, des variants de séquence dont la capacité à se replier correctement est conservée. Ces méthodes ont été développées sur le DLR de Raf et ont servi à déterminer la variation de séquence tolérée à chaque position de ce dernier tel que cela est rapporté dans l'**Article 2**. Les avancements dignes de mentions rapportées dans cet article sont :

1. L'établissement d'un protocole basé sur la technique de PCR qui permet l'insertion de courts segments de codons dégénérés contigus, tout en introduisant un biais minimal de la diversité de séquences des librairies.

2. L'adaptation du PCA DHFR au criblage de librairies du DLR de Raf par la détection de l'interaction avec *ras* dans des cellules *Escherichia coli* (*E. coli*).

Les techniques que nous avons développées pour la synthèse des librairies dégénérées et pour leur sélection sont simples à appliquer et se comparent donc avantageusement aux méthodes couramment citées dans la litérature scientifique, respectivement la mutagenèse Kunkel et l'exposition sur phage.

Contribution des auteurs à la préparation de l'article 1:

F.-X.C.V. : conception et réalisation des techniques et rédaction de l'article.

S.W.M. : suggestion et supervision du projet ainsi que rédaction de l'article.

Article 1. «Synthesis of Libraries and Screening with the DHFR PCA»

Running Head: Synthesis of Degenerated Libraries of the RBD of *Raf* and Rapid Selection of fast-folding and stable clones with the DHFR Protein fragment Complementation Assay

Authors: François-Xavier Campbell-Valois* and Stephen W. Michnick*†

* Département de Biochimie, Université de Montréal, C.P. 6128, Succ. centre-ville, Montréal, Québec, Canada H3C 3J7

†Corresponding author: email: stephen.michnick@umontreal.ca
phone: (514) 343-5849
fax: (514) 343-2015

Abstract

The protein-engineering field is mainly concerned with the design of novel enzyme activities or folds and in understanding the fundamental sequence determinants of protein folding and stability. Much effort has been put into the design of methods to generate and screen libraries of polypeptides. Screening for the ability of proteins to bind with high affinity and/or specificity is most often approached with phage display technologies. Herein, we present an alternative to phage display, performed totally *in vivo*, based on the DHFR Protein-fragment Complementation Assay (PCA). We describe the application of the DHFR PCA to the selection of degenerated sequences of the *ras*-binding domain (RBD) of *raf* for correct folding and binding to *ras*. Our screening system allows for enrichment of the libraries for the best behaving sequences through iterative competition experiments without the discrete library screening and expansion steps that are necessary in *in vitro* approaches. Moreover, the selected clones can be processed rapidly to purification by Ni-NTA affinity chromatography in 96-well plates. Our methods are particularly suitable for designing and screening of libraries aimed at studying sequence folding and binding determinants. Finally, it can be adapted for library against library screening, thus allowing for co-evolution of interacting proteins simultaneously.

Key words: Protein-fragment Complementation Assays, PCA, dihydrofolate reductase DHFR, bacterial survival assay, phage display, protein-protein interactions, protein engineering, protein folding, degenerated libraries, PCR, binding assays, 6xHis tag affinity-purification.

1. Introduction

Since the development of recombinant DNA technologies in the 70's and 80's, numerous ingenious approaches have been exploited to synthesize and screen oligonucleotide libraries to discover those that code for novel protein sequences displaying a desired characteristic, be it enzyme activity, binding or stability of a protein under selected conditions **(1-8)**. In any given case, to tackle such protein engineering efforts one must have two methods in place: a strategy to generate a diverse library of sequences and an efficient way to screen for desired characteristics of the products of the library. The choice of library synthesis method is crucial to providing a sufficiently large sequence search space, such that a maximum number of choices are available from which sequences coding for desired characteristics can be found. It is not so surprising then, that the development of such strategies has been and still is a focus of research in the field. There are a few examples reported in the literature of studies in which a region of a protein is completely randomized or highly degenerated to answer questions about protein folding **(9, 10)**. Nevertheless, examples of truly and highly degenerated libraries to explore sequence space in search of a novel fold, binding capability or enzyme activity are rare **(6, 11)**. The inherent limitations of generating highly randomized libraries and subsequently searching for the few sequences that display the desired characteristics in huge sequence space have prompted efforts towards the design of methods that explore more limited library sets. These include DNA shuffling strategies or completely alternative approaches that allow for recombination between genes devoid of any sequence homology **(12-18)**.

Even the most cleverly designed libraries will not yield useful products without an adequate screening and selection strategy. For example, an ideal way to screen a library coding for an enzyme activity, stability or ability to bind to a target protein is to express the library in a cell or organisms in which expression of library members with the desired characteristics confers specific growth capabilities on selective medium, in harsh conditions or in a specific genetic background **(1, 19-22)**. Such examples are unfortunately

rare and thus, protein engineers have sought more general approaches to screen libraries **(23-26)**. More specifically, binding assays can often serve the general purpose of selecting expressed polypeptides from a DNA library that are properly folded, stable and whose binding to some molecule, whether it be another protein, nucleic acid, organic substrate or transition state analogue imply the specific function desired. The most well established method of choice to do this is the phage display strategy **(4, 5, 27)** (reviewed in **(28-30)**). In this strategy, the expansion and the screening of libraries are performed in discrete steps, taking place respectively *in vivo* and *in vitro*. The method is well suited to proteins that bind to small molecules or peptides that can be easily cross-linked to a solid phase support, but it is not straightforward to adapt for studies of protein-protein interactions or for library against library screening. More recently, the Protein-fragment Complementation Assay (PCA) has emerged as an alternative technology **(31, 32)** (reviewed in **(33, 34)**). The PCA strategy relies on the association and folding of a reporter protein or enzyme from fragments, driven by the interaction of two proteins to which the fragments are fused. The reconstitution of the reporter protein fold and thus detectable catalytic activity depends on the interaction of the fused proteins. In particular, a simple survival-selection assay has been developed for screening libraries in *E. coli*, based on the murine dihydrofolate reductase (mDHFR) as reporter PCA **(32)**. In *E. coli*, as in all prokaryotes and eukaryotes, the DHFR product tetrahydrofolate is necessary for the synthesis of thymine, glycine, serine and adenine, while in prokaryotes, it is also required for synthesis of panthotenate. DHFR activity is thus absolutely required for cell growth and division in the absence of a source of DHFR end products. *E. coli* can be made dependant on expression of recombinant mDHFR by treatment of the cells with trimethoprim, a folate analog that is 12,000 times more potent an inhibitor of *E. coli* over mammalian DHFRs **(35)**. The principle of the mDHFR PCA then, is that two proteins fused to complementary fragments of mDHFR must be coexpressed and interact together in *E. coli* grown in minimal (M9) medium supplemented with trimethoprim in order for cells to grow and divide **(31)**. In a first demonstration of a library against library screen, the DHFR PCA was used to identify optimally heterodimerizing pairs of leucine zipper-forming sequences from individual

libraries containing $6x10^{10}$ possible combinations of sequences. Competition experiments and "library shuffling" strategies were devised to improve library screening coverage, to further optimize dimerizing pairs and finally to identify a "winner pair" **(32, 36)**. More recently, the DHFR PCA was adapted for screening and selection of single-chain antibodies *in vivo* **(37)**. The all in one genetic screening approach of the DHFR PCA strategy is the key feature allowing for simple performance of library against library screening, because selective pressure is applied concomitantly on both library populations over several cycles, without the tedious alternation between discrete expansion and screening steps associated with phage display. Thus PCA truly allows for the study of sequences co-variation of oligomeric partners.

The results obtained with leucine zippers convinced us that the assay could be useful for tackling more challenging problems. In the zipper studies, only a handful of key amino acid positions were varied and only from two to four amino acid substitutions were allowed. Based on previous theoretical work **(38)**, we are currently attempting to rigorously and exhaustively determine the sequence determinants for folding of the *ras* Binding Domain (RBD) of the serine/threonine protein kinase *raf* (ndlr, voir **Article 2**). The premise of our approach is similar to a previously published strategy aimed at identifying sequences that support rapid folding and stability of proteins selected by phage display **(39)**. The principle as applied to the *raf* RBD is as follows: if the sequence of a given RBD variant folds rapidly to the correct structure and is sufficiently stable, it should interact with it's natural binding partner, the small GTPase *ras*. Fusing the RBD library to one complementary fragment of DHFR and *ras* to the other, and then co-expressing these in *E. coli*, grown under selective pressure as described above, it can be reasoned that fast-folding and stable members of the RBD library will interact with *ras* and allow for the reconstitution of DHFR activity and the rescue of cell growth. Next, we needed to choose an efficient and realistic way to design libraries that at the same time allow for exploring the maximum sequence diversity in a meaningful way, while creating libraries of reasonable size. Based on the questions we chose to address in these studies, a meaningful

way to explore sequence space is to generate libraries in which only a small stretch of contiguous residues are varied at a time **(10)**. Examination of the RBD structure allows one to dissect it into 13 regions corresponding to individual β-turns or loops, β-strands and one α-helix (2 libraries were generated for this region corresponding to amino and carboxyl termini of the helix) ranging in length from 4 to 8 amino acids **(40, 41)**. On this basis, we have created 13 degenerated libraries, in which each wild type codon is replaced by a NNK codon (where N is any nucleotide and K is G or T) that allows the insertion of the 20 amino acids at each varied position in the sequence. These were screened for binding to *ras* by the DHFR PCA in *E. coli*. In addition to the screening being done entirely *in vivo*, a key advantage of this approach is that expressed RBD library members that interact well with *ras* can be purified for physical analysis without having to switch to another expression system.

Herein, we present the protocols and proposed trouble-shooting strategies, based on the technical challenges that we have encountered in the design and synthesis of the degenerated libraries and in their screening with the DHFR PCA. Hopefully, these protocols are general enough to be useful not only to those interested in folding, but more generally to problems requiring the optimization of protein-protein interactions.

2. Materials

2.1. Library synthesis

1. Oligonucleotide primers (IDT). The primers with positions where multiple bases are allowed are hand mixed to assure that the adequate ratio of each base is respected. These are SDS-PAGE purified.
2. Taq polymerase (Fermentas).
3. Agarose gel: agarose (Bioshop) Dark Reader™ (Clare chemical research) and Gelstar™ (Biowittaker Molecular Applications).

4. Gel purification, QIAEX™ II or even better QIAquick™ gel extraction Kit (Qiagen).

2.2. Library cloning and recovery

1. Plasmid pQE-32 Δ F [1,2] (derived from plasmid pQE-32 distributed by Qiagen. F [1,2] stands for DHFR fragment 1).
2. Plasmid pREP4 (Harbors *lac* repressor and kanamycin as selectable marker. Cells in which protein is expressed off of the pQE-32 plasmid, such as used in these studies, must contain this vector in order to limit expression from the otherwise very leaky *tac* promoter contained in the pQE-32 plasmid. Distributed by Qiagen).
3. Ligation, T4 DNA ligase (Fermentas) and ATP (Pharmacia).
4. SS320 electrocompetent cells (see **section 2.7**).
5. Genepulser™ II electroporation apparatus (Biorad).
6. Electroporation cuvette with 2 mm width slot (Invitrogen).
7. SOC medium: LB supplemented with 0.4% glucose, 2.5 mM KCl and 10 mM $MgCl_2$.
8. LB-agar supplemented with 10 μg/mL tetracycline, 10 μg/mL spectinomycin and 100 μg/mL ampicillin in 100 mm petri dishes.
9. 100-250 mL LB medium per library. LB is supplemented with 0.2 % glucose, 0.25x M9 salts solution (see step **2.3.4** for recipe), 10 μg/mL tetracycline, 10 μg/mL spectinomycin and 100 μg/mL ampicillin (Bioshop) added.
10. Plasmid Midi Kit (12143) (Qiagen).

2.3. Library screening

1. BL21 electrocompetent cells transformed with pREP4 (see section **2.2.2**) and then transformed with pQE-32 *ras*-F [3] (F [3] stands for DHFR fragment 2).
2. Genepulser™ II electroporator system (Biorad) or Electroporator 2510 (Eppendorf).
3. Electroporation cuvette with 1 mm width slot (Invitrogen).
4. SOC medium (see **step 2.2.7**).

5. Phosphate-buffered saline (PBS). For 1 L final volume in water, combine: 8 g NaCl, 0.2 g KCl, 1.44 g Na_2HPO_4 and 0.24 g KH_2PO_4. The pH is adjusted to 7.4 with HCl and autoclaved.
6. M9 minimal medium supplemented with the appropriate antibiotics (hence dubbed "selective medium"). For 1 L complete medium, combine: 740 mL of 2.5 % noble agar (Difco), 200 mL 5xM9 salts (for 1L, 64 g of Na_2HPO_4, 15 g KH_2PO_4, 2.5 g NaCl and 5 g NH_4Cl. Composition in **(42)**), 2 mL 1 M $MgSO_4$, 1 mL 100 mM $CaCl_2$ and 20 mL of 20% glucose solution. All salts and glucose are cell culture grade, from any source such as Sigma, Fisher or ICN except: 100 μg/mL ampicillin, 25 μg/mL kanamycin and 1mM IPTG (isopropyl-β-D-thio-galacto-pyranoside) (Bioshop), 10 μg/mL trimethoprim (ICN) and 800 μg/mL casamino acids (Difco) and 10 μg/mL Thiamine (Fisher). All solutions must be prepared with deionized water and sterilized by filtration for antibiotics, casamino acids, IPTG and thiamine (store at -20 °C) and by autoclave for salts (store at RT). The reconstituted medium is poured into 100 mm or 150 mm petri dishes. The reconstituted medium can be kept at 4°C for up to 2 months.
7. Plasmid Midi Kit (Qiagen) or alkaline lysis maxiprep.
8. Restriction enzymes: *Hpa*I, *Xma*I, *Eco*NI and *Xba*I (NEB or Fermentas)
9. XL1 blue chemiocompetent cells (see **section 2.8**).

2.4. Clones competition experiment

1. Glass culture or 15 mL conical tubes (Corning).
2. Solid and liquid selective medium (same protocol as in **step 2.3.6**, except that agar is not added for liquid medium).
3. Plasmid Midi Kit (Qiagen) or alkaline lysis maxiprep.
4. LB medium supplemented with 100 μg/mL ampicillin and 25 μg/mL kanamycin.

2.5. Isolation of clones and Sequencing

1. Restriction enzymes: *Hpa*I, *Xma*I, *Eco*NI and *Xba*I (NEB or Fermentas)

2. XL1 blue competent cells.
3. 24-well plates (Corning), LB-agar with 100 μg/mL ampicillin.
4. 2 mL V shaped 96-well culture block (VWR).
5. Montage™ Plasmid Miniprep 96 kit (Millipore, LSKP 096 01) or smaller scale prep kit, like QIAprep™ Spin Miniprep Kit (27104) (Qiagen) depending on the number of samples to be processed.
6. Vacuum manifold Multiscreen Resist™ (Millipore, MAVM 096 OR).
7. Oligonucleotide primer for sequencing specific to the plasmid harboring the library (IDT).

2.6. Protein purification

1. Appropriate restriction endonucleases (*Sal*I and *Xho*I in this case) and reagents necessary for ligation (see **section 2.2**).
2. BL21 pREP4 competent cells and LB 100 μg/mL ampicillin, 25 μg/mL kanamycin petri dishes.
3. Terrific broth (TB) supplemented with 100 μg/mL ampicillin and 25 μg/mL kanamycin
4. 50 mL conical tubes (Corning).
5. A centrifuge and rotor that accommodate 96-well plate such as Eppendorf 5810 or 5810 R and A4-62 respectively.
6. Ni-NTA Spin Kit or Ni-NTA Superflow™ 96 Biorobot Kit (Qiagen) depending on the number of samples to be processed. An affordable alternative to the Superflow 96 Biorobot Kit is the following: we use Ni-NTA Superflow resin (Qiagen), 0.25mm glass fiber filter 96 well plates (3510), 0.2 μm PVDF membrane 96-well plates (3504), 96-well volume extender (3584) and fraction collector (3958) (Corning).
7. Vacuum manifold Multiscreen Resist™ and the large collection and sealing block (Millipore, respectively MAVM 096 OR and OT).

8. Buffer A: 6 M Guanidinum-HCl (Gnd-HCl, ICN), 0.1 M NaH_2PO_4 (Fisher), 0.01 M Tris-Cl (Tris base, Bioshop), pH 8.0 and supplemented with PMSF 10 μM PMSF (phenylmethyl sulfonyl flourid, ICN), 7.2 mM β-mercaptoethanol (Fisher), 5 mM imidazole (Fisher) and 300 mM NaCl (Fisher)
9. Buffer B: same as buffer A, but pH 6.3 and supplemented with 7.2 mM β-mercaptoethanol and sometimes 15 mM imidazole
10. Buffer E: 4 M Gdn-HCl, 0.025 M NaOAc, pH 4.5.
11. Dithiothreitol (DTT)
12. KOH 6 M (Fisher).

2.7. Preparation of SS320 and BL21 pREP4 pQE-32 *ras*-F [3] electrocompetent cells

1. Overnight (O/N) preculture of SS320 or BL21 pREP4 pQE-32 *ras*-F [3].
2. 500 mL SOB medium: 10 g tryptone, 2.5 g yeast extract, 1 ml 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, supplemented with 0.2% glucose for SS320 strain.
3. 500 mL LB medium supplemented with 0.2% glucose for BL21 pREP4 strain.
4. 2 L of ice-cold autoclaved deionized water.
5. Autoclaved 10 % glycerol solution (Bioshop).

2.8. Preparation of XL1-blue and BL21 pREP4 chemiocompetent cells

1. O/N preculture of XL1-blue or BL21 pREP4.
2. 500 mL SOB medium supplemented with 0.2% glucose for SS320 strain.
3. 500 mL LB medium supplemented with 0.2% glucose for BL21 pREP4 strain.
4. Transormation buffer: 10 mM Pipes, 15 mM $CaCl_2$ and 250 mM KCl (Fisher). pH 6.7 with KOH (Fisher). After the pH is set, $MnCl_2$ is added to a concentration of 55 mM.
5. dimethyl solfoxide (DMSO) (Fisher).

3. Methods

3.1. General Considerations

3.1.1. Steric requirements in PCA

The spatial orientation of the PCA fragments is crucial to whether the PCA reporter protein can fold from it's cognate fragments and is determined by the orientations of the amino or carboxyl termini of the interacting proteins in the complex formed (see **Figure 1** for schematization of spatial considerations encountered when designing linkers). In the design of the protein-PCA fragment fusions, it is therefore important to determine *a priori*, whether the fragments would be brought together by a given combination of fusion constructs in such a way that the topology of the native structure could be achieved from a given configuration of the fusions. The two main factors that will determine whether correct folding can be attained are the orientation of the fusion (carboxyl and/or amino terminal) and second, the length of polypeptide linkers between the individual fragments and the proteins to which they are fused. Our experience has shown that linkers constituted of repeats of GGGGS behave better in *E. coli*, yeast and mammalian cells based on tests of a number of different interacting proteins. We rationalize that this type of linker improves flexibility and solubility of the fusions, thus easing their reassembly. Moreover, they ensure metabolic stability because of their lack of susceptibility to naturally occurring proteolytic activities. Although, these types of linkers are preferable, they are not essential for productive fragment complementation. However, bulky hydrophobic and rigid amino acids, such as proline and β-branched amino acids should be avoided. In most protein engineering problems, the structure of a protein complex of interest is already known and the orientation of the complex and requirement of linkers of a given length can be deduced. For the DHFR PCA, the spatial requirements for fusion of proteins at C- and N-termini are clear **(31, 43)**. For example, if the proteins of interest are fused respectively to the carboxyl terminus of F [1,2] and the amino terminus of F [3], the inserted linker can be quite short or may even not need to be included in the constructs, because it respects the normal topology

of the enzyme. However, if the oligomers are fused to the amino termini of both fragments, the topology of the enzyme is permuted, thus requiring a minimum of 2 amino acids (each peptide bond is approximated to 3.75 Å) in each linker to permit productive fragment reassembly, since the distance between the two amino termini is approximately 10 Å. This orientation was chosen in the case of *ras* and the RBD; however examination of the complex between *raf* RBD and the highly similar *ras* homologue's, *rap1A* **(44)**, reveals that the carboxyl terminus of each monomer are located 40 Å apart, thus requiring that a minimum of 6 amino acids have to be added to the linker of each construct. For the library screening, we have fixed the length of each linker to 14 amino acids total, including the restriction sites, to make sure that sufficient flexibility is allowed.

3.1.2. Controls and Stringency

Before beginning any protein engineering study and library screening with the DHFR PCA, one should perform rigorous controls to assess the sensitivity and stringency of the assay for the specific test system. Ideally, this means that the investigator should know before beginning library screens, roughly what is the dissociation constant limit of detection of the PCA for a given interaction. The sensitivity limit (maximum dissociation constant for which a PCA response is detected) is going to vary among different interacting pairs of proteins but in addition to the dissociation constant, it is modified by factors such as the level of expression of each fusion, the amount of soluble *versus* the total expression of protein fusions and the intrinsic properties of the protein, such as their stability, solubility and their kinetics of folding and binding. If the PCA is very sensitive and can detect very weak interaction between two specific proteins, it could prove impossible to distinguish clones that have the best desired properties from any other; that is, the assay can be too sensitive resulting in a loss of stringency. Thus, it is important to maintain a balance between these two factors. To assess these issues, general controls should be done prior to PCA studies. However, as will be discussed later, not all of these are necessarily relevant to a specific protein engineering study. 1) *Spurious reassembly*: For the PCA to work, assembly of fragments from weak or non-specific interactions cannot be allowed. The

effective sensitivity is intrinsic to a given interacting protein pair test system and PCA and can be assessed by Controls *4)* and *6)*. Then, if the sensitivity is too high, for example, if growth occurs for proteins that should not interact together (see **Figure 2 A)**), sensitivity can be reduced by decreasing expression levels and/or by using stringency mutants as described in *2)* (see **Figure 2 B)**). 2) *Stringency mutants*: The effects on DHFR PCA of side-chain truncation mutant at fragment interface such as Ile114 of F [3] have been reported **(31, 32)**. The problem encountered in the latter study was that when clones expressing leucine zipper forming pairs that formed complexes with varying efficiencies were compared, it was impossible to distinguish them based on growth rates or numbers of colonies formed. In contrast, by inserting the mutation Ile114Ala on F [3] we changed the wild type sensitivity of the PCA to an appropriate value for the leucine zipper system. A measure of this change in sensitivity was the "selection factor" in single-step selection, defined as the number of cotransformed cells plated divided by the number of colonies surviving under selective conditions. The result was an increase in stringency such that it allowed us to distinguish the best from more poorly behaving heterodimerizing pairs of leucine zippers in a reasonable number of iterative competition steps **(32)**. 3) *Fragment swapping*: An observed interaction between binding proteins should occur regardless of which PCA fragments either of the proteins are attached to. Therefore, an interaction observed with one protein-fragment configuration should give comparable result if proteins and fragments are swapped. 4) *Non-interacting proteins*: A PCA response should not be observed if a protein that is not known to interact with either of two interacting proteins being tested is used as a PCA partner (see **Figure 2 A)**), nor should over-expression of this protein alone compete for the known interaction. 5) *Ability to titrate and to decrease the observed reporter activity by competition*: The observed reporter activity should vary with the relative expression ratio of each of the protein-fragment constructs. As well, the PCA response should be diminished by simultaneous over-expression of one or the other interacting partners alone. However, one should bears in mind that the respective solubility and stability of each construct, the affinity of the interacting proteins for each other versus their cellular concentration and the sensitivity of the PCA affect the ability to titrate the

reporter activity. Modulation of reporter activity could then be achieved only by lowering expression level of the fusion constructs or by combining this type of control with *2)* and/or *6)* to reduce complementation efficiency. 6) *Disrupting the interaction*: Insertion of specific point or deletion mutation in one of the complex monomer's that is known to disrupt or diminish the interaction should also affect the PCA response in a predictable way.

In a protein-engineering project in which the model system under study is very well characterized, only controls 1, 2, 4 and 6 are essential for establishing the specificity and stringency of the assay. In the case of the RBD-*ras* interaction, a comprehensive mutagenesis study and its effect on the dissociation constant (K_d) for binding of the RBD had already been published **(45)**. These data permitted us to engineer several mutants that reduce the K_d for association of RBD-*ras* over three orders of magnitude (see **Figure 3** for example). These mutants and others were tested in the DHFR PCA allowing us to establish that the assay is able to detect binding for the RBD to *ras* for mutants with a K_d on the order of 1 μM. Also, published mutants that destabilize the protein fold, like core hydrophobic residues (valine, leucine or isoleucine) side chain truncation to alanine, could be used as a stringency test.

3.2. Library synthesis

1. In order to have a non-biased library we first generated a template where the region to be varied was deleted and replaced by a stop codon, inserting also a frame shift and a unique restriction site allowing for it's unequivocal identification (see **Note 1**).
2. To generate each library we synthesized two PCR products that partially overlap (typically 18-20 base pair (bp) hybridization region). For example, for PCR 1 we used one primer hybridizing in the promoter region of our vector (120 bp upstream of the start codon) and one primer hybridizing in the region immediately in 5' of the section targeted for degeneracy (see **Figure 4**). For PCR 2 we used one two-arms oligonucleotide and a primer that hybridizes to the F[1,2] (120 bp downstream of

the 3' end of the ORF). Typically, the PCR program was set as following: 1 min hotstart at 94°C, 25 cycles of 20 s at 94°C, 30 s at 52°C, 30 s at 72°C (see **Note 2**). Finally, the reactions put for 10 min at 72 °C to ensure completion of the elongation.

3. The PCR products are analyzed on agarose gel. If the desired product is obtained, the remainder of the PCR product is loaded on gel. We have advantageously used Gelstar™ and the Dark Reader™ (see **Note 3**) to visualize PCR products on agarose gels. It permits observation of bands under blue light (400-500 nm), wavelengths that do not damage DNA, in contrast to UV light (this allows one to cut the bands out and to proceed easily, in parallel, to the generation of several libraries).
4. Next the bands were gel purified with Qiaex™ II (see **Note 4**).
5. Approximately 300 ng of the PCR product from PCR 1 and 2 are combined (see **Note 5**) with 0.2 μM of the terminal primers (hybridizing in the promoter and in F [1,2]) that anneal in 5' and 3' respectively, to the product of PCR 1 and 2. The PCR 3 program is the following: 1 min hot start at 94°C, 10 cycles of 20 s at 94°C, 30 s at 52°C, 30 s at 72°C. Finally 10 min at 72°C to ensure completion of the elongation (see **Note 6**).
6. The entry vector pQE-32 Δ F [1,2] (see **Note 7**) and the resultant PCR products are digested with the appropriate restriction enzyme (*Sph*I and *Xho*I in this case).
7. Bands are purified according to **step 3.2.4**.

3.3. Library cloning and recovery.

1. The ideal insert versus vector ratio for ligation is 2:1 to 3:1. We try to limit the concentration of DNA ligated to 10 ng/μl and we use 1 mM ATP. We allow the ligation to proceed overnight (O/N) at 16°C (see **Note 8**).
2. The enzyme is heat inactivated at 65°C, chloroform extracted and precipitated with ethanol. The DNA pellet is then air dried for several minutes and resuspended in 30 μL of deionized water prior to electroporation.

3. SS320 *E. coli* strain (see **Note 9**) electrocompetent cells are prepared the same day (see **section 3.8.**). The ligation reaction from **step 3.3.2** is mixed in 300 µL of resuspended SS320 cells. The mix is transferred to 2 mm width electroporating cuvettes and electroporated on the Genepulser™ II. The apparatus parameters are adjusted to the following settings for the pulse: 2.5 kV, 25 µF and 200-400 Ω. For optimal results, the time constant should be between 3.8-4.5 for 200 Ω and 7.6-9.0 for 400 Ω. Immediately after the pulse, 1 mL of ice-cold SOC medium is added to the cuvettes. Cells are transferred to a 15 mL conical tube, the cuvettes washed two times with SOC medium to recuperate maximally the electroporated cells, and then allowed to recover for 30 min in 5 mL SOC medium (see **step 2.2.7.** for description) at 37°C with moderate shaking.
4. The efficiency of the ligation and cloning is evaluated by counting the colonies formed for plating of $1x10^{-4}$ of the electroporated bacteria (see **Note 10**). The remainder is directly inoculated into 250 mL of properly supplemented LB medium in a 500 mL flask (**step 2.2.9.**).
5. The DNA is isolated with Qiagen Midiprep Kit or similar kits or alternatively by alkaline lysis maxiprep **(46)**.

3.4. Library screening

1. 100 ng of the pooled library clones recovered from **step 3.3.5.** is ethanol precipitated and electroporated in 65 µl of BL21 pREP4 cells already harboring a plasmid expressing pQE-32 *ras*-F [3] (see **Note 11**) with 1 mm width electroporating cuvettes. The apparatus parameters are adjusted to the following settings for the pulse: 1.25-1.6 kV, 25 µF and 200Ω. The time constant is varied from 3.7-4.2 on the Genepulser™ II or from 4.0-4.6 on Electroporator 2510. The cells are allowed to recover during 30 minutes in SOC medium at 37 °C with moderate shaking.
2. The cells are washed twice with cold PBS to remove traces of SOC rich medium.

3. The cells are plated on selective medium as described in **section 2.3** and allowed to grow for 24 to 72 hours at 30 °C (see **Note 12**). Again, in order to be able to assess the efficiency of transformation and the ratio of clones in the library that rescue cell growth, a fraction of the electroporated cells, on the order of $1x10^{-3}$, should be plated separately to allow colony counting and comparison with a positive control. For example, in our procedure, we transform the same mass of a vector expressing the *wt* RBD fusion to F [1,2] and we plate a dilution of 10^{-4}-10^{-5} of the electroporated cells. All measures should be taken to avoid cross-contamination of the library pool mix with *wt* positive controls at every step of these manipulations.

3.5. Clonal competition experiment

This procedure is adapted from **(32)**.

1. After the appropriate incubation period, the cells plated at **step 3.4.3** are harvested with a small volume of selective medium and incubated in 25 mL of selective medium at 30°C in a shaker at 250 rpm (see **Note 13**).
2. Then after 24 hours of incubation, an aliquot of 1 μL of the saturated culture is diluted in 2 mL of fresh selective medium.
3. **Step 3.5.2** can be repeated until the targeted enrichment of the library is reached. Normally, we observe that the pool is greatly enriched for one to a few clones after 12 passages (12 days). However, it could vary from system to system depending on various factors such as the level of degeneracy of the libraries and the use of stringency mutants (see **section 3.1.2.** on **General Controls and Stringency**).
4. At any step a 10^{-4}-10^{-5} diluted aliquot of the saturated culture can be plated on solid selective medium to qualitatively check the efficiency of the competition. The heterogeneity in colony size should decrease and the average size of colonies should increase with each successive competition step.
5. At any passage, the clones represented in a pool mix can be recovered by inoculating 2 mL LB (100 μg/mL ampicillin and 25 μg/mL kanamycin) with 10 μL

of the pool mixture and incubated O/N. Then DNA is prepared with QIAprep™ (see **Note 14**).

3.6. Isolation of clones and Sequencing

This step consists of the isolation of the library bearing plasmid from independent clones or from the mixed pools obtained by the manipulations described in **section 3.4.** and **3. 5**.

1. 300 ng of DNA from our pool of clones are digested with a mixture of restriction enzymes that recognize sites present in the pREP4 and pQE-32 *ras*-F [3] plasmids but absent in the library plasmid. For this purpose, we used: *Hpa*I, *Xma*I, *Eco*NI and *Xba*I (see **Note 15**).
2. One tenth of the digested DNA is transformed in XL1-blue chemiocompetent cells and 20 μL is plated on 24-well plates containing LB-agar with 100 μg/mL of ampicillin (see **Note 15**).
3. Colonies are picked and incubated in LB, supplemented with 100 μg/mL of ampicillin, in the appropriate culture vessels.
4. High quality DNA minipreps are prepared for sequencing. For processing 96 samples in plates, we use the Montage™ kit. We have used QIAprep™ column kit for smaller scale preps.
5. For sequencing, we have used a primer that anneals only to the library plasmid, i.e. inside F [1,2] (see **Note 16**).
6. Sequencing.

3.7. Protein purification and characterization

After analysis of the obtained sequences, clones of interest are selected and rearrayed on the appropriate number of 96-well plates. At this moment, clones can be retransformed in XL1-blue cells and frozen to serve as back-up stock.

1. The selected clones at this step are recloned to express them as fusions with the 6xHis tag only, i.e. without the DHFR fragment (see **Note 17**). The expression of the 6xHis-clones is verified by an induction test (see **Note 18**).
2. The preps of the clones that express correctly are done with Montage™ kit and rearrayed again.
3. These clones are then transformed into BL21 pREP4 cells and plated on LB-agar medium containing 100 μg/mL of ampicillin and 25 μg/mL kanamycin. The plates are incubated at 37°C O/N. When processing several clones in parallel, we use 24-well plates. In this case, no more than 20 μl of competent cells should be used per transformation and this should be the maximum volume to be plated per well. Plating more than the maximum volume will not allow the plated cells to absorb completely into the medium.
4. The following day, one colony for each selected clone is picked and incubated O/N in 2.5 mL of LB supplemented with the appropriate antibiotics at 37°C with moderate shaking (see **Note 19**).
5. The saturated cultures are diluted 1:10 in 25 mL of TB supplemented with the appropriate antibiotics (see **Note 20**).
6. The cultures are then incubated for 90 to 120 min at 37°C in the shaker and then IPTG is added at 1 mM. After 2-4 hours of induction, the cells are harvested and the protein can be purified immediately or stored at -80 °C (see **Note 21**).
7. The cell pellets are resuspended in 1 mL of Buffer A (see **step 2.6.8.**) by agitation at room temperature until the solution becomes translucent, and then arrayed in a 96 well block (see **Note 22**). Most of the insoluble material is then removed by centrifugation at 4,000 rpm for 40 min on Eppendorf A4-62 rotor.
8. The 0.2 μm PVDF 96-well plate is filled with 200 μL 50% Ni-NTA Superflow™ resin. A volume extender is assembled on top of the 0.25 mm glass fiber 96-well plate. This assembly is placed on the sealing block on top of the vacuum manifold (see **Note 23**). The samples (900 μL) are then applied into the first filter plate and

100 µL of ethanol is added to reduce the risk of cross-contamination. Approximately 500 mbar of pressure is applied until all the samples are completely drawn through the plate (see **Note 24**).

9. The PVDF plate assembly should now contains the resin and the samples filtrate. The pressure is interrupted, and the PVDF plate assembly is moved from the collection chamber to be fitted on the sealing block on top of the vacuum manifold. Then, approximately 100 mbar of pressure is applied until all the samples are completely drawn through the resin (see **Note 25**).
10. Wash twice with 800 µL Buffer B (see **step 2.6.9.**). For all wash steps the vacuum pressure is set at 500 mbar (see **Note 26**).
11. Place a fraction collector in the collection chamber. The samples are eluted four times with 100 µL of Buffer E (see step **2.6.10.** and **Note 26**) at 100 mbar of pressure. The eluate is supplemented with 1 mM DTT and the pH is adjusted to 5 (see **Note 27**). The samples are now ready for immediate characterization (see **Note 28**).

3.8. Preparation of SS320 and BL21 pREP4 pQE-32 *ras*-F [3] electrocompetent cells

Cells were prepared according to **(47).**

3.9. Preparation of XL1-blue and BL21 pREP4 chemiocompetent cells

Cells were prepared according to H. Inoue *et al.* **(48)** with one slight modification: The cells are washed only once after the first centrifugation step. The bottles containing the cell pellets are placed inverted on a piece of paper at 4°C to remove most traces of medium (see **Note 29**).

4. Notes

1. Prior to PCR, the primers are phosphorylated with T4 polynucleotide kinase to allow ligation. The primers are designed to anneal respectively to each strand,

adjacent to the region to be deleted. We used high proofreading polymerases such as Pfu or Pfu Turbo for this type of PCR. The number of PCR cycles is kept low (10-16 cycles). The product of the PCR is a linear version of the plasmid without the deleted region. After the PCR, the resultant reaction mixture is digested with DpnI to digest the template DNA. Following ethanol precipitation, approximately one tenth of the PCR product is ligated (25 ng PCR product/50 μl ligation reaction) and transformed. The positive clones are screened for the insertion of the appropriate restriction site. A frame shift is included to reduce the possibility of read-through the stop codon. The PCR protocol is derived from the ExSite™ PCR-based site-directed mutagenesis kit (Stratagene).

2. We have used Taq polymerase since the sequences we wanted to amplify were relatively short. For longer genes we suggest Pfu or Pfu Turbo polymerase that gives more reliable results than Vent polymerase in our hands with this protocol. When the difference in size between the product of PCR 1 and PCR 2 is too large, particularly for large genes, we would recommend a mega-primer protocol **(49, 50)**. Both approaches with slight modification could be useful to generate combinatorial libraries in which regions far apart in the sequence are varied simultaneously.
3. The Gelstar™ is diluted $1x10^{-4}$ to $5x10^{-5}$ in TAE agarose gel. Gelstar is much more labile than ethidium bromide and thus, the gels should be prepared on the day that they will be used. Also, it is difficult to easily quantify DNA with Gelstar™ since the signal becomes saturated at lower quantities of DNA, and thus it is difficult to distinguish different amounts of DNA above a certain threshold. Also, one must be careful not to load large amounts of DNA per well since this could dramatically affect the migration of the samples (Typically we do not load more than 300 ng of plasmid DNA per 25 μL wells). Gel staining with ethidium bromide also permits visualization of DNA with the blue light of the Dark Reader, although more DNA must be loaded per well (more than 1.5 μg for digested plasmid DNA) to allow for visualization of small fragments or PCR products. In the worst case, UV ethidium

bromide visualization can of course be used, but care should be taken to process the bands as quickly as possible.

4. The QIAquick™ gel purification protocol is preferable for generating several libraries at the same time, but is more expensive. As an alternative to the procedure described in **step 3.2.3**, the electrophoresis step can be replaced by a 2-hour incubation at 37 ˚C of the PCR product with the restriction enzyme *Dpn*I to remove the plasmid template DNA. Then, the PCR products are purified according to the QIAquick™ PCR purification protocol.
5. This amount could be increased in proportion to the size of the gene under study. The relative quantities of product from PCR 1 and PCR 2 added to PCR 3 are adjusted according to their relative molecular weights.
6. The number of amplification steps in PCR 3 has to be minimized since failure to do so could be detrimental to library representativity. For this reason, the number of cycles is minimized and the concentration of the two terminal primers is reduced (see **Figure 3**). It is worth optimizing this step in order to obtain the maximum quantity of product in a minimum number of cycles. We recommend using Taq polymerase as described in **Note 2**.
7. This vector was obtained by intramolecular religation of pQE-32 digested with *Nhe*I and *Xba*I (compatible cohesive restriction sites). This procedure removed approximately 850 bp between the terminator and the origin of replication. The stability and the expression level of the ORF harbored in the new vector was exactly the same as the original, as far as we could judge, while at the same time allowing to increase by an order of magnitude, the number of cells transformed per µg of ligated plasmid DNA. Furthermore, the *Xho*I site present in the original vector was removed to allow use of this restriction site for library cloning.
8. We obtain best results with these ratios and when we use vector that is not dephosphorylated.
9. The SS320 cells are obtained by mating MC1061 with XL1-blue **(30)**. An equal volume of both strains at OD_{600}= 0.6 are mixed and incubated at 37 °C for 1 hour

with smooth shaking (50 rpm). Then, the conjugation is stopped by increasing the shaking to 250 rpm for 5 minutes. The new strains are isolated by plating on LB petri dishes supplemented with 10 μg/mL tetracycline and spectinomycin. The strains obtained combine the elevated electrocompetence of MC1061 (up to $5x10^{10}$ colonies/μg of supercoiled plasmid in our laboratory) with the episome overexpressing $LacI_q$ of XL1-blue necessary for cells transformed with pQE vectors. This strain could be replaced by any appropriate one for a given expression system.

10. The colonies are counted and the number multiplied by the dilution factor. With 300 ng vectors resuspended in 30 μl water and electroporated in 300 μl SS320 we usually obtained between 10^6-10^7 colonies.
11. We originally chose to use pQE vector because we wanted to be able to do the screening and Ni-NTA affinity purification without having to change vector. The *ras*-F [3] protein fusion is expressed from a pQE-32 derived vector **(31)**. This plasmid contains the same origin of replication and antibiotic resistance as the plasmid harboring the RBD of *raf* libraries. The selective pressure of the trimethoprim forces the cells to keep both plasmids (i.e. reconstitution of mDHFR from complementary fragments requires that both fusions are expressed). We had engineered a pQE-32 derived plasmid with an alternative origin of replication and antibiotic resistance to express the *ras* fusion, but the stringency of the screening was greatly diminished when we used this vector. In the present case, we mean by a decrease of stringency that the DHFR PCA selects clones that, based on sequence data, should not have the ability to bind to *ras*. These conclusions were drawn from experiments in which sequences of clones of the RBD of *raf* selected by the screening assay revealed that they contained stop codons and aberrant sequences in a region important for binding to *ras,* suggesting that these interactions were nonspecific. The problems we have encountered in these conditions could probably be corrected by inserting the destabilizing mutants in F [3], Ile114Val or Ile114Ala. We have not tested this yet. We think the stringency is high in the configuration we

have chosen, because the BL21 pREP4 cells harboring pQE-32 *ras*-F [3] have the maximum number of copies for the ColE1 origin (harbored by pQE vectors) prior to the electoporation of the library, thus allowing conservation of only a minimum number of copies of the library plasmid, as is necessary for growth in the selective conditions. A good alternative for future screens could be to use vector expressing the bait fusion construct at much lower level or to use DHFR destabilizing mutants **(31, 32)**.

12. Alternatively, directly inoculate 25 mL of liquid M9 minimal medium with the electroporated cells. This is particularly useful if one wants to directly start a competition experiment.
13. We have observed that the stringency, particularly for the first passages, is improved if cells are plated on solid medium prior to the competition experiment, meaning that growth on solid phase allows for selection of the most efficient clones faster.
14. We recommend using QIAprep™ spin kit at this stage, because the DNA extracted by alkaline lysis methods from BL21 pREP4 is not of good quality. Also as described in the manufacturer protocol, the columns are washed once with PB buffer (Qiagen) when preparing plasmid DNA.
15. Steps 2 and 3 obliterate the necessity to check if the cells are transformed with the appropriate plasmid prior to sequencing. Of the 700 colonies treated according to these procedures, we have never encountered a single one harboring pQE-32 *ras*-F [3] or pREP4 plasmid.
16. We used the same primer that anneals to 3' of PCR 2 and 3 (see **step 3.2.5.**).
17. We recommend removing the DHFR fragment before further *in vitro* characterization, unless it is required for specific experiments, because it diminishes yields, complicates purification procedures and might modify protein characteristics. We recommend engineering the plasmid in such a way that the DHFR fragments could be removed and the plasmid religated intramolecularly. To

do so we have used the compatible restriction sites *Xho*I and *Sal*I to clone respectively the library and F [1,2] fragment.

18. O/N saturated cultures are diluted 1:10 in TB 100 μg/mL ampicillin and incubated at 37°C at 300 rpm in a 96 well culture block. After 90 minutes, the cultures are induced with 1 mM IPTG. After 4 hours, 60 μl aliquots of each clone are pipetted and 1 volume of 2x SDS-PAGE loading buffer is added. Sample, markers and protein induction test samples are loaded on 15 % acrylamide SDS-PAGE and the protein bands are visualized by Coomassie brilliant blue staining. Alternatively, expression could be check by western blot with an antibody directed against the 6xHis tag (Qiagen).
19. Before proceeding to the large-scale purification, we recommend making test purification of several isolated clones to check yields obtained with different culture volumes. We recommend denatured condition purification, because it is easier to perform for obvious reasons when doing several purifications in parallel. Nevertheless, if the protein is very soluble, one could design a simple native state purification protocol amenable to such a scale. In our case, 25 mL culture of BL21 pREP4 cells expressing *raf* RBD mutants, processed as described in **section 3.7.**, yielded 400 μl of protein solution at concentrations of 200 to 600 μg/mL.
20. Depending on the quantity of protein needed or the expression level, the volume of culture can vary from 5 to 50 mL. Read **Note 19** for more details.
21. Alternatively, protein expression could be induced O/N at 30 °C.
22. Otherwise, if the quantity of clones to analyze was lower, we used Ni-NTA spin columns. If more protein from less independent clones is needed, regular Ni-NTA agarose column purification is the method of choice. In this case, it may be useful and reasonable to do purification under native conditions.
23. The described system only works in this configuration with the large collector and sealing block. It is also possible to replace vacuum steps by centrifugation.
24. A gauge is included with the Millipore manifold.

25. If required, the flow-through and the filtrate from the diverse washing steps are collected by placing a fraction collector in the collection chamber at **steps 3.7.9.** and **3.7.10.**
26. In our case, the protein is eluted under denaturing conditions and the sample diluted sufficiently to reduce denaturant to a concentration that does not interfere with protein function. The concentration of denaturant used in the elution buffer should be fixed according to the stability of the protein studied. Nevertheless, if it is desired to elute protein under native condition, Qiagen recommends using a gradient of decreasing concentration of Gdn-HCl. The protein is then eluted with the appropriate buffer, without denaturant. One could also simply purify protein under native conditions (see Qiagen Expressionnist and Ni-NTA spin column handbook). We have improved elution by using diluted glacial acetic acid, in quantities sufficient to buffer 25 or 50 mM NaOAc to pH= 5.0 (see **Note 27** concerning pH adjustment after elution). Urea could also be used as denaturant provided it will denature the protein under study at reasonable concentrations. Also care must be taken as urea in solution equilibrates with cyanate. On the other hand, urea does not precipitate SDS even at high concentration, thus facilitating the characterization of the protein at individual purification steps by SDS-PAGE electrophoresis..
27. DTT is added to the sample after elution because it is not recommended for use with Ni-NTA. The pH of the eluate can be adjusted to a relevant value by the addition of NaOH or of an appropriate weak base such as Tris or NaOAc. The quantity of base to be added is determined empirically on larger volumes with a pH meter.
28. If the samples are not going to be characterized immediately, we recommend adding 40% glycerol v/v and 1mM NaN_3 for storing of samples at -20°C. The glycerol can then be removed by dialysis or ultrafiltration prior to characterization. Alternatively, samples without glycerol can be flash frozen in liquid nitrogen and conserved at -80°C.

29. BL21 pREP4 chemiocompetent cells prepared according to QIAEXpressionnist Handbook with Transformation buffer B (RbCl) are usually more competent, but since we needed to transform only super coiled DNA in our protocols, the Inoue method was satisfactory **(48)**.

Aknowledgements

We are grateful to Emil Pai, David Waugh and Shigekazu Nagata for the cDNA's of *h-ras*, *raf* RBD and the CAD domain's of *cad* and *icad* respectively. We also want to thank Dimitri Sans for carefully reading this manuscript, Jérôme Dupras for help with the CAD constructs as well as Joelle Pelletier and other members of the laboratory that had contributed to the development of the DHFR PCA. F.-X. C.-V. is a CIHR and FCAR scholar.

References

1. Chen, K. Q. and Arnold, F. H. (1991) Enzyme engineering for nonaqueous solvents: random mutagenesis to enhance activity of subtilisin E in polar organic media. *Biotechnology (N Y)* **9,** 1073-1077.
2. Chen, K. Q., Robinson, A. C., Van Dam, M. E., Martinez, P., Economou, C. and Arnold, F. H. (1991) Enzyme engineering for nonaqueous solvents. II. Additive effects of mutations on the stability and activity of subtilisin E in polar organic media. *Biotechnol Prog* **7,** 125-129.
3. Iffland, A., Tafelmeyer, P., Saudan, C. and Johnsson, K. (2000) Directed molecular evolution of cytochrome c peroxidase. *Biochemistry* **39,** 10790-10798.
4. Scott, J. K. and Smith, G. P. (1990) Searching for peptide ligands with an epitope library. *Science* **249,** 386-390.
5. Lowman, H. B., Bass, S. H., Simpson, N. and Wells, J. A. (1991) Selecting high-affinity binding proteins by monovalent phage display. *Biochemistry* **30,** 10832-10838.
6. Keefe, A. D. and Szostak, J. W. (2001) Functional proteins from a random-sequence library. *Nature* **410,** 715-718.
7. Imanaka, T., Shibazaki, M. and Takagi, M. (1986) A new way of enhancing the thermostability of proteases. *Nature* **324,** 695-697.
8. Shih, P. and Kirsch, J. F. (1995) Design and structural analysis of an engineered thermostable chicken lysozyme. *Protein. Sci.***4,** 2063-2072.
9. Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. and Baker, D. (1997) Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4,** 805-809.
10. Kim, D. E., Gu, H. and Baker, D. (1998) The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl. Acad. Sci. U S A* **95,** 4982-4986.
11. Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. and Hecht, M. H. (1993) Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262,** 1680-1685.
12. Stemmer, W. P. (1994) Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370,** 389-391.
13. Crameri, A., Raillard, S. A., Bermudez, E. and Stemmer, W. P. (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391,** 288-291.

14. Zhao, H., Giver, L., Shao, Z., Affholter, J. A. and Arnold, F. H. (1998) Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat. Biotechnol.* **16,** 258-261.
15. Ostermeier, M., Shim, J. H. and Benkovic, S. J. (1999) A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat Biotechnol* **17**, 1205-1209.
16. Lutz, S. and Benkovic, S. J. (2000) Homology-independent protein engineering. *Curr. Opin. Biotechnol.* **11,** 319-324.
17. Coco, W. M., Levinson, W. E., Crist, M. J., Hektor, H. J., Darzins, A., Pienkos, P. T., Squires, C. H. and Monticello, D. J. (2001) DNA shuffling method for generating highly recombined genes and evolved enzymes. *Nat. Biotechnol.* **19,** 354-359.
18. Murakami, H., Hohsaka, T. and Sisido, M. (2002) Random insertion and deletion of arbitrary number of bases for codon-based random mutation of DNAs. *Nat. Biotechnol.* **20,** 76-81.
19. Crameri, A., Dawes, G., Rodriguez, E., Jr., Silver, S. and Stemmer, W. P. (1997) Molecular evolution of an arsenate detoxification pathway by DNA shuffling. *Nat. Biotechnol.* **15,** 436-438.
20. Christians, F. C., Scapozza, L., Crameri, A., Folkers, G. and Stemmer, W. P. (1999) Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling. *Nat. Biotechnol.* **17,** 259-264.
21. MacBeath, G., Kast, P. and Hilvert, D. (1998) Redesigning enzyme topology by directed evolution. *Science* **279,** 1958-1961.
22. Ostermeier, M., Nixon, A. E., Shim, J. H. and Benkovic, S. J. (1999) Combinatorial protein engineering by incremental truncation. *Proc. Natl. Acad. Sci. U S A* **96,** 3562-3567.
23. O'Neil, K. T. and Hoess, R. H. (1995) Phage display: protein engineering by directed evolution. *Curr. Opin. Struct. Biol.* **5,** 443-449.
24. Roberts, R. W. and Szostak, J. W. (1997) RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci. U S A* **94,** 12297-12302.
25. Firestine, S. M., Salinas, F., Nixon, A. E., Baker, S. J. and Benkovic, S. J. (2000) Using an AraC-based three-hybrid system to detect biocatalysts in vivo. *Nat. Biotechnol.* **18,** 544-547.
26. Hanes, J., Jermutus, L. and Plückthun, A. (2000) Selecting and evolving functional proteins in vitro by ribosome display. *Methods Enzymol.* **328,** 404-430.
27. Smith, G. P. (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228,** 1315-1317.
28. Dunn, I. S. (1996) Phage display of proteins. *Curr. Opin. Biotechnol.* **7,** 547-553.
29. Forrer, P., Jung, S. and Plückthun, A. (1999) Beyond binding: using phage display to select for structure, folding and enzymatic activity in proteins. *Curr. Opin. Struct. Biol.* **9,** 514-520.
30. Sidhu, S. S., Lowman, H. B., Cunningham, B. C. and Wells, J. A. (2000) Phage display for selection of novel binding peptides. *Methods Enzymol.* **328,** 333-363.
31. Pelletier, J. N., Campbell-Valois, F.-X. and Michnick, S. W. (1998) Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally-design fragments. *Proc. Natl. Acad. Sci.* **95,** 12141-12146.
32. Pelletier, J. N., Arndt, K. M., Plückthun, A. and Michnick, S. W. (1999) An in vivo library-versus-library selection of optimized protein-protein interactions. *Nat. Biotechnol.* **17,** 683-690.
33. Michnick, S. W., Remy, I., Campbell-Valois, F. X., Vallee-Belisle, A. and Pelletier, J. N. (2000) Detection of protein-protein interactions by protein fragment complementation strategies. *Methods Enzymol.* **328,** 208-230.
34. Michnick, S. W. (2001) Exploring protein interactions by interaction-induced folding of proteins from complementary peptide fragments. *Curr. Opin. Struct. Biol.* **11,** 472-477.
35. Appleman, J. R., Prendergast, N., Delcamp, T. J., Freisheim, J. H. and Blakley, R. L. (1988) Kinetics of the formation and isomerization of methotrexate complexes of recombinant human dihydrofolate reductase. *J. Biol. Chem.* **263,** 10304-10313.
36. Arndt, K. M., Pelletier, J. N., Muller, K. M., Alber, T., Michnick, S. W. and Plückthun, A. (2000) A heterodimeric coiled-coil peptide pair selected in vivo from a designed library-versus-library ensemble. *J. Mol. Biol.* **295,** 627-639.

37. Mossner, E., Koch, H. and Plückthun, A. (2001) Fast selection of antibodies without antigen purification: adaptation of the protein fragment complementation assay to select antigen-antibody pairs. *J. Mol. Biol.* **308,** 115-122.
38. Michnick, S. W. and Shakhnovich, E. (1998) A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Fold. Des.* **3,** 239-251.
39. Gu, H., Yi, Q., Bray, S. T., Riddle, D. S., Shiau, A. K. and Baker, D. (1995) A phage display system for studying the sequence determinants of protein folding. *Protein Sci.* **4,** 1108-1117.
40. Emerson, S. D., Waugh, D. S., Scheffler, J. E., Tsao, K. L., Prinzo, K. M. and Fry, D. C. (1994) Chemical shift assignments and folding topology of the Ras-binding domain of human Raf-1 as determined by heteronuclear three-dimensional NMR spectroscopy. *Biochemistry* **33,** 7745-7752.
41. Emerson, S. D., Madison, V. S., Palermo, R. E., Waugh, D. S., Scheffler, J. E., Tsao, K. L., Kiefer, S. E., Liu, S. P. and Fry, D. C. (1995) Solution structure of the Ras-binding domain of c-Raf-1 and identification of its Ras interaction surface. *Biochemistry* **34,** 6911-6918.
42. Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989). Bacterial media, antibiotics and bacterial strains. In Molecular Cloning (Chris Nolan ed.), vol. **3** Cold Spring Harbor Laboratory Press, New York, pp. A.3.
43. Remy, I., Wilson, I. A. and Michnick, S. W. (1999) Erythropoietin receptor activation by a ligand-induced conformation change. *Science* **283,** 990-993.
44. Nassar, N., Horn, G., Herrmann, C., Scherer, A., McCormick, F. and Wittinghofer, A. (1995) The 2.2 A crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with Rap1A and a GTP analogue. *Nature* **375,** 554-560.
45. Block, C., Janknecht, R., Herrmann, C., Nassar, N. and Wittinghofer, A. (1996) Quantitative structure-activity analysis correlating Ras/Raf interaction in vitro to Raf activation in vivo. *Nat. Struct. Biol.* **3,** 244-251.
46. Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989). Alkalyne lysis and PEG preparation for large scale DNA preparation. In Molecular Cloning, vol. **1** (Chris Nolan ed.) Cold Spring Harbor Laboratory Press, New York, pp. 1.38-1.41.
47. Seidman, C. E., Struhl, K., Sheen, J. and Jessen, T. (1997). Introduction of plasmid DNA into cells, basic protocol 2. In Current Protocols in molecular biology, vol. **1** (supplement 37) (Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. and Struhl, K., eds.) John Wiley & Sons, Inc., New-York, NY, pp. 1.8.4-1.8.5.
48. Inoue, H., Nojima, H. and Okayama, H. (1990) High efficiency transformation of Escherichia coli with plasmids. *Gene* **96,** 23-28.
49. Sarkar, G. and Sommer, S. S. (1990) The "megaprimer" method of site-directed mutagenesis. *Biotechniques* **8,** 404-407.
50. Brons-Poulsen, J., Petersen, N. E., Horder, M. and Kristiansen, K. (1998) An improved PCR-based method for site directed mutagenesis using megaprimers. *Mol. Cell Probes* **12,** 345-348.

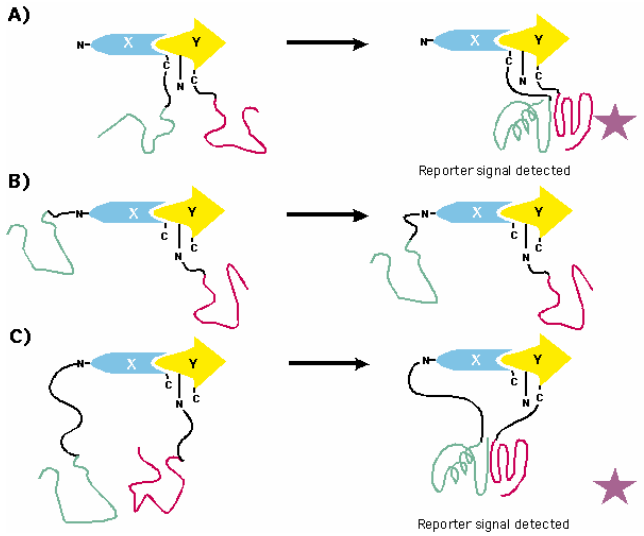


Figure 1 Schematic structure of a heterodimeric complex formed between proteins X and Y in which their respective amino termini (N) are far apart while their carboxyl termini (C) are proximal in space and do not directly participate in the binding interface. In this case the fragments should be attached as depicted in **A)**. The number of amino acids in each linker can be determined by assuming that a peptide bond is approximately 3.75 Å long and by taking into consideration the structural and spatial constraints of the X-Y complex of interest and of the protein used as PCA reporter. If the crystal structure of the X-Y complex is not available, linker length has to be determined empirically. In some cases it may not be possible to make the fusions in an optimum orientation. For example, in the (A) case, the carboxyl terminus of X has to be free for binding to Y. One could design constructs as pictured in **B)**. In this case, it is obvious that the short linkers designed for A) would not allow for reconstitution of native topology of the PCA reporter. Nevertheless, if longer linkers are used instead, as depicted in **C),** the folding of the reporter from it's cognate fragment is possible, thus making this protein fusion orientation adequate for PCA based protein engineering screening strategy.

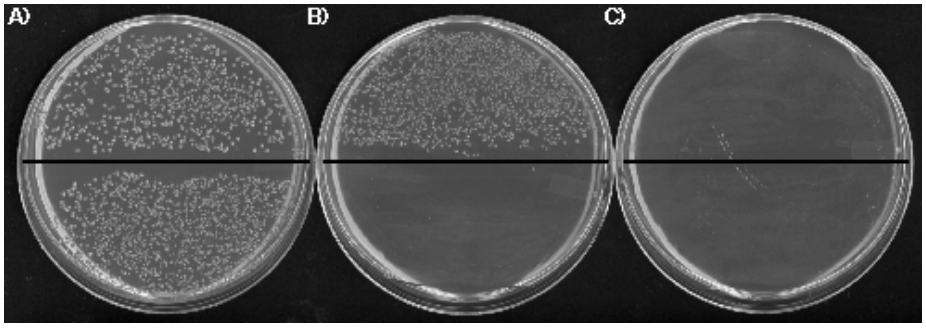


Figure 2 The caspase activated domains (CAD) of *icad* and *cad* were fused respectively to F [1,2] and F [3], F [3] Ile114Val or Ile114Ala. These *cad* constructs were cotransformed into BL21 pREP4 along with either pQE-32 Δ *icad*-F [1,2] or RBD of *raf*-F [1,2] and plated on selective medium respectively on the upper and lower part of the petri dishes, respectively and incubated for 24 hours at 30˚C: **A)** pQE-32 *cad*-F [3] I114 (wild-type DHFR fragment F[3]) **B)** pQE-32 *cad*-F [3] I114V **C)** pQE-32 *cad*-F [3] I114A. The cotransformation of *cad* fusions with the RBD of *raf* fusions served as an internal control as described in **section 3.1.2., General Controls and Stringency**, particularly as described under subheadings *1)* and *2)*. These tests allow for determination of the sensitivity limit of the PCA. Since there should not be any significant interaction between *cad* and the RBD of *raf*, no interaction should be detected by the PCA, and thus no colony formation observed on selective medium. The use of Ile114Val mutant is thus ideal for assuring sufficient stringency, because it allows for growth of cells cotransformed with the relevant *cad* and *icad* constructs, while in contrast to the wild-type F[3] fusion construct in **(A),** it does not lead to colony formation for cells cotransformed with the constructs of the RBD of *raf* and *cad*. The F[3] Ile114Ala mutant does not allow growth of neither the positive *icad-cad* nor the negative control *icad-raf* RBD pairs of fusions.

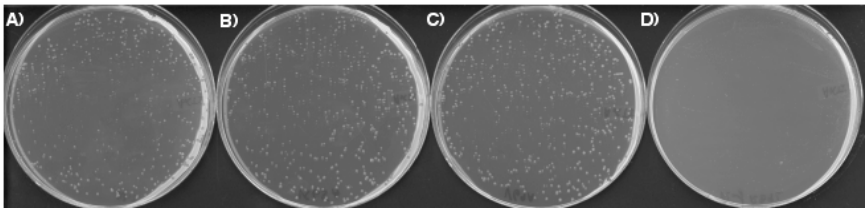


Figure 3 BL21 pREP4 cells were cotransformed with plasmids expressing *ras*-F [3] and either the RBD of *raf wt* or mutants fused to F [1,2]. The cells were allowed to grow for 48 hours on selective media and petris were scanned. RBD of *raf* **A)** *wt* (K_d=0.13 μM) **B)** K65M (K_d= 0.40 μM) **C)** V69A (K_d= 0.95 μM) **D)** R89L (K_d> 100 μM). These mutants and others not shown allowed us to situate the sensitivity limit of our detection assay in the 10-100 μM range.

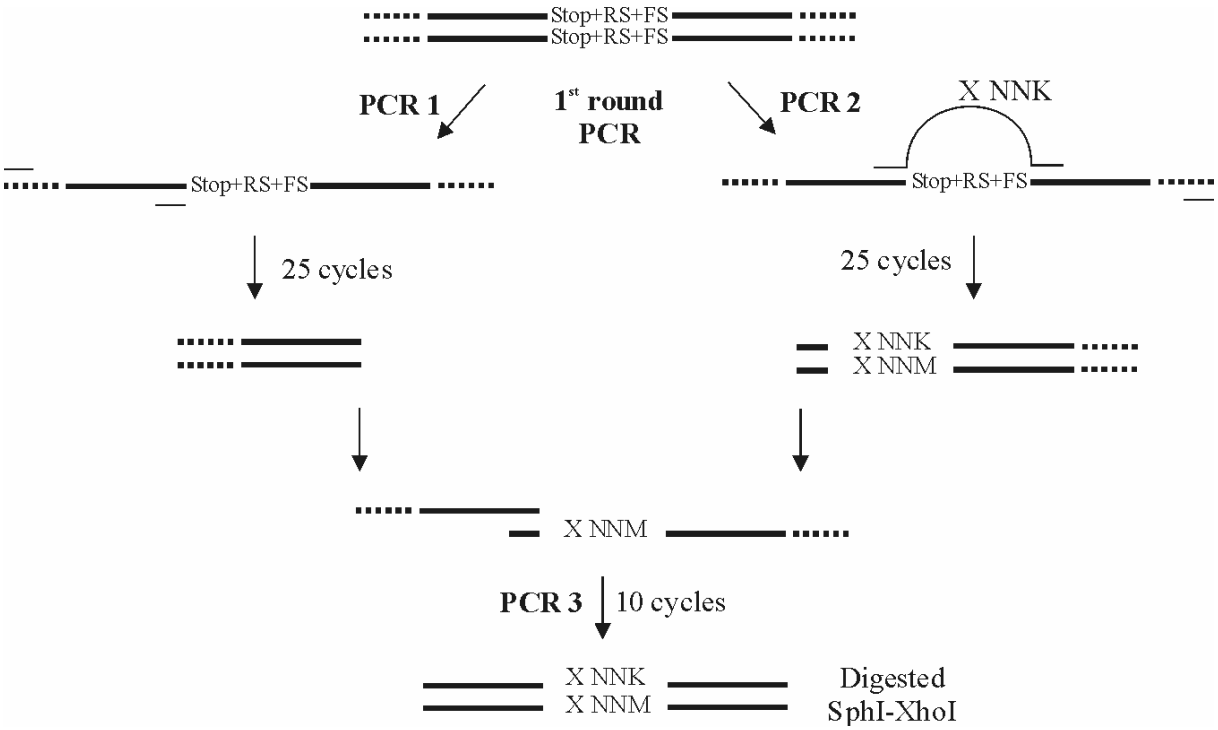


Figure 4 Schematic representation of the strategy for the synthesis of degenerated libraries (see **section 3.2. Library Synthesis** for detail). The strategy is divided into three distinct steps: First, the template is obtained by replacing the *wt* sequence of the region to be varied by an in-frame stop codon, a frame shift (FS) and a unique restriction site (RS) to allow for identification of the mutant (the full and dash line correspond respectively to DNA of the gene of interest and of the vector). Then, two PCR reactions are done on this template. PCR 1 product corresponds to the 5' end of the gene, while PCR 2 corresponds to the 3' end of the gene. In the latter case, the STOP-FS-RS sequence in the template is replaced by the appropriate number of NNK degenerated codons. Note that the products from round 1 PCRs partially hybridize through their 3' and 5' ends. . Thus, in PCR 3, the products from the 1st round PCR act as template and primers for the reaction. After a low number of cycles, usually 10, the full length degenerated gene is recovered by digestion with the restriction enzyme *Sph* I and *Xho*I. The library is then ready to be cloned.

Article 2 : Perturbation massive de la structure primaire du DLR de Raf

Article publié dans « Proceedings of the National Academy of Sciences U.S.A. »

(Octobre 2005), vol. 102, p. 14988-14993.

Présentation de l'article 2 :

Au moment où j'entrepris mes études graduées, S.W.M. terminait une collaboration avec E. Shakhnovich portant sur la conservation de séquence dans un alignement de protéines partageant la même topologie qu'ubiquitine. Ils avaient identifié une série de résidus conservés qui pourrait servir à la transmission évolutive du mécanisme de repliement (*30*). S.W.M me proposa alors d'entreprendre un projet où nous chercherions à déterminer expérimentalement la conservation de séquence du DLR de Raf, l'une des protéines considérée dans l'étude précédente. Nous pensions que cet objectif pouvait être atteint en synthétisant des séquences dégénérées du DLR de Raf et en sélectionnant celles qui pourraient former sa structure native. L'intérêt pour ce projet était triple. Premièrement, malgré sa simplicité, la topologie d'ubiquitine intègre plusieurs des divers motifs de structures secondaires et elle est retrouvée chez de nombreuses protéines qui remplissent des fonctions biologiques fort variées. Il n'est donc pas surprenant qu'elle soit parmi les 10 topologies les plus fréquemment observées (i.e. supertopologies) et que l'homologie de séquence parmi les domaines adoptant cette structure peut être très faible (< 10% dans plusieurs cas). L'identification des déterminants de séquence de la topologie d'ubiquitine aurait donc un grand intérêt du fait de son importance biologique et cela nous permettrait par la suite de poser des hypothèses quant aux causes et conséquences de l'utilisation répétitive de cette topologie au cours de l'évolution. En second lieu, S.W.M. voulait orienter les recherches dans son laboratoire vers l'étude des voies de signalisation cellulaires, en particulier sur l'organisation et la dynamique des réseaux d'interactions protéine-protéine. A cet égard, la protéine Raf constituait un modèle intéressant, car elle joue un rôle clé dans les voies de signalisation activées par les récepteurs tyrosine kinase, tel que le facteur de croissance épithélial (EGF). Le DLR de Raf est essentiel à la stimulation de l'activité kinase de Raf via son interaction avec les petites GTPases *ras* et Rap1A. A ce titre, l'interaction Raf et *ras* est importante à la transmission de signaux prolifératifs et/ou oncogéniques. Dans l'optique de développer un outil innovateur à l'étude des interactions protéine-protéine, Joëlle Pelletier et S.W.M était à mettre au point le PCA DHFR, une technique conceptuelle reposant sur la complémentation des fragments protéiques (*246*). Un objectif accessoire que nous avions donc aussi en tête à l'époque était de développer et de valider une méthode

basée sur le PCA qui pourrait se positionner comme une alternative à l'exposition sur phage par exemple, qui pourrait être applicable en particulier à la sélection de séquences se repliant vers une structure native donnée (voir l'**Article 1**).

Cet article présente donc la conclusion des 4 premières années d'études au doctorat en plus de mes premières années d'études graduées. Spécifiquement, nous y rapportons les approches expérimentales utilisées pour perturber 72 des 78 positions du DLR de Raf via l'insertion de librairies de codons dégénérés à des segments de résidus contigus ainsi que pour sélectionner les variants qui peuvent adopter la structure native (voir aussi l'**Article 1** pour de plus amples détails), mais surtout les résultats de ces expériences. Les découvertes principales publiées dans cet article sont :

1. Les séquences du groupe de variants du DLR de Raf obtenues expérimentalement a une diversité comparable à un alignement groupant des protéines naturelles très diverses évolutivement, dont l'unique caractéristique commune évidente est d'adopter la topologie d'ubiquitine.

2. Il est possible de prévoir une séquence consensus du DLR de Raf en discriminant les positions sur la base de leur importance structurale et fonctionnelle.

3. Le cœur hydrophobe du DLR de Raf est organisé en deux couches concentriques, le cœur interne et externe. L'appartenance des divers résidus hydrophobes à l'un ou l'autre de ces niveaux d'organisation peut être discriminée sur la base de leur niveau de conservation au cours de l'expérience de perturbation de la séquence.

Auparavant, il y avait eu peu d'études extensives de perturbation de la structure primaire d'une protéine ou d'un domaine protéique (*53-57;60;61*). Nos travaux suggèrent des solutions expérimentales simples et des méthodes d'analyses rigoureuses afin d'obtenir de l'information précise sur les déterminants de séquence globaux d'une petite protéine. Les résultats que nous avons obtenus permettent d'entrevoir l'utilité de notre approche dans la définition des déterminants de séquence de protéines adoptant des topologies rares, ce qui pourrait à terme faciliter la prédiction de leur structure et leur design par des programmes informatiques.

Contribution des auteurs à la préparation de l'article 2

F.-X.C.V. : conception et réalisation des expériences, analyse des données et rédaction de l'article.

K.T. : développement d'outils informatiques pour la présentation et l'analyse des données.

S.W.M. : suggestion et supervision du projet, ainsi que rédaction de l'article.

Article 2: «Massive sequence perturbation of a small protein»

Running-title: Massive sequence perturbation without long-range co-variation on Raf RBD approximates the sequence diversity explored by ubiquitin superfold in nature.

Authors: Campbell-Valois, F.-X.*†, Tarassov, K.*, Michnick, S.W.*‡

* Département de Biochimie and † Programme de Biologie Moléculaire , Université de Montréal, C.P. 6128, Succ. centre-ville, Montréal, Québec, Canada H3C 3J7

‡Corresponding author: email: stephen.michnick@umontreal.ca

phone: (514) 343-5849
fax: (514) 343-2015

Abstract

Most protein topologies rarely occur in nature, thus limiting our ability to extract sequence information that could be used to predict structure, function, and evolutionary constraints on protein folds. In principle, the sequence diversity explored by a given protein topology could be expanded by introducing sequence perturbations and selecting variant proteins that fold correctly. However, our capacity to explore sequence space is intrinsically limited by the enormous number of sequences generated from the 20 amino acids and the limited number of variants likely to fold. Here we sought to test whether the sequence space for naturally existing proteins can be explored by simple, sequential degeneration of a complete set of short sequence segments of a model protein, without long-range covariation. Using the Raf *ras* binding domain as a model of a small protein capable of autonomous folding, we degenerated 72 of 76 positions of the primary structure for the 20 amino acids in segments of four to seven residues defined by secondary structure and selected the folded species for interaction with *h-ras* by using an *in vivo* survival-selection assay. The methodology presented allowed for rigorous statistical analysis and comparison of sequence diversity. The ensemble of sequence variants of Raf *ras* binding domain obtained have recaptured the diversity observed for the ubiquitin-roll topology. A signature sequence for this fold and the implication of this strategy to protein design and structure prediction are discussed.

Keywords: massive mutagenesis | protein-fragment complementation assay | protein structure topology | ubiquitin superfold | Raf *ras* binding domain

Abbreviation footnote: RBD, ras binding domain; MSA, multiple sequence alignment; DHFR, dihydrofolate reductase; protein-fragment complementation assay, PCA; dissociation constant, K_d; wild-type, wt; signature sequence, SS; circular dichroism, CD; nuclear magnetic resonance, NMR.

Introduction

The ability of polypeptides to fold into a unique native structure is remarkably robust to mutations (1-3). Thus, polypeptides sharing structural topology, particularly if unrelated functionally, can display very low sequence identity. It follows that the comparison of diverse protein sequences adopting the same structure could be used to define the sequence determinants of a specific fold, because these residues are the most likely to be conserved across multiple sequence alignments (MSA). However, this approach is limited to a minority of folds for which a sufficient number of structures having divergent sequences are available. To expand the sequence space explored by a given topology, an interesting solution is to mimic nature by introducing massive degeneracy into the amino acid sequences and select variants for their folding capacity to identify the residues that are under selective pressure. Such information could build on achievements of protein design and structure prediction algorithms (4-7).

A mutagenesis strategy aimed at exploring the potential sequence diversity of an entire protein should allow for randomization of all residues for the 20 amino acids (aa). An algorithm for performing covariation of sequence segments has been proposed, but its experimental implementation would be difficult (8). Indeed, the ability to exhaustively explore sequence space is limited by the extraordinary number of sequences generated by the combination of "l" randomized residues (e.g., l is the number of residues in the polypeptide) that increases as 20^l and the limited number of sequences that can fold into the target structure. The obvious solution is to vary fewer residues at a time as has already been done to study compensation effects, principally in the hydrophobic core, by covarying residues dispersed over the primary sequence (9-12). This method is not suitable for large-scale sequence perturbation, because of the enormous number of covariation combinations to test and technical limitations in library synthesis. Alternatively, the primary structure can be degenerated in short contiguous segments (e.g., 4-10 residues). This approach has yielded interesting insights into protein folding (13, 14), but the residue degeneracy inserted was not constant across all segments, and, thus, it is difficult to interpret the significance to folding and stability of aa selection at specific positions. In principle, such a strategy would allow researchers to tackle the sequence perturbation of a protein in a simple and exhaustive

way. Strangely, no attempt has been made to entirely degenerate a protein segment-by-segment. Although it is clear that a fully exhaustive search of sequence space requires covariation of all residues, it would be possible to compare the sequence diversity obtained experimentally by segmental perturbation with those found in nature and ask whether the sequence space explored is similar. The latter exercise would require a model protein for which a large number of structures with highly diverse sequences are available.

The *ras* binding domain (RBD) of the Ser/Thr kinase Raf is composed of 78 aa and folds autonomously into a compact globular structure build by the packing of a single α-helix against a mixed β-sheet of connectivity 2-1-5-3-4 (Fig. 1*A*) (15, 16). Furthermore, the Raf RBD tertiary structure is characteristic of the ubiquitin superfold (also ubiquitin roll or β-grasp ubiquitin-like), which is one of the most common topologies in the protein universe (17). Therefore, sequences of several functional homologues (fh) and many structural analogues (sa) to Raf RBD can be retrieved from databases.

The strategy consists of creating discrete libraries of Raf RBD in which the codons of contiguous residues constituting an individual secondary structure element are randomized to allow insertion of the 20 aa and then selecting correctly folded clones by using an *in vivo* protein-fragment complementation assay (PCA) to detect protein-protein interaction (Fig. 1 *A*). The experimental design and the large data set described below allowed rigorous statistical analysis of sequence diversity by building positional entropy and aa selection profiles. Finally, the experimental data are compared with MSAs of fh and sa to validate the strategy. Strikingly, these analyses revealed that the sequence diversity observed experimentally in the Raf RBD sequences approximates the sequence space explored by a MSA of sa sharing the ubiquitin-roll topology.

Methods

Generation of Experimental Libraries. Each experimental library was prepared from the corresponding deletion tagged templates by two-round PCR (*Supporting Methods* in *Supporting Text*, which is published as supporting information on the PNAS web site) by

using Pfu turbo polymerase. The first round of PCRs yielded two products: one corresponds to the 5' region of the targeted segment and extending 120 bp upstream of the 5' end of the Raf RBD cDNA and a second one flanking the targeted segment in 3' and extending 120 bp downstream of the cDNA. Depending on the library position in the wild-type cDNA, one of the two products was generated with a loop out primer that allows reinsertion of the appropriate number of NNK codons (encoding the 20 types of aa and one terminator) into the targeted segment. All oligonucleotides were obtained from Integrated DNA Technologies (Coralville, IA) and were synthesized with hand-mixed nucleotides to insert an equal quantity of the appropriate bases at each position of the NNK codons. The PCR products generated in the first round had 18 complementary bp at their joining ends to enable the generation of the full-length degenerated cDNA through a second round PCR. The second round PCR is short (10 cycles) to maximize library representation (Fig. 6, which is published as supporting information on the PNAS web site). Detailed protocols for library recoveries can be found in *Supporting Methods* and ref. 18.

Screening Libraries with Dihydrofolate Reductase (DHFR) PCA. A fraction of pooled plasmid preparation for each library and the wt Raf RBD construct were precipitated with ethanol and dissolved in deionized water. Then, DNA concentration is estimated from OD_{260} and the volume adjusted to obtain similar concentration (e.g., 100 ng/μl). Routinely, 150 ng of the precipitated plasmid was electroporated into 60 μl of BL21 pREP4 strain carrying a vector that allows for expression of *h-ras*-DHFR [3] fusion under control of lacIq repressor. After incubation under vigorous shaking of electroporated cells in 2 ml of SOC medium during 30 min at 37°C, they were washed and plated on selective medium (as in ref. 19, except that thymine was replaced by 30 μM thiamine and 800 μg/ml casamino acids, and trimethoprim concentration was increased to 10 μg/ml). Petri dishes were incubated during 36 h at 30°C. For statistics, a dilution (wt plasmid: 1×10^{-4} and libraries: 1×10^{-3} are used) of the transformation reaction was plated on a separate Petri dish, and resulting colonies were counted (Table 1). The plasmids of selected clones were prepared in library pools by harvesting all colonies from Petri surface (18). Selected clones were sequenced, and nonredundant sequences were aligned with the wt Raf RBD (Table 2, which is published as supporting information on the PNAS web site).

Positional Entropy. Shannon entropy is calculated by using Eq. 1 (19):

$$S = -\sum p_i \ln p_i / \ln L \qquad [1]$$

For experimental entropy calculation, L was fixed to 20, because we considered every switch of aa at a given position as a mutation. Before calculation of the entropy, the frequency of each aa (p_i) was corrected according to the bias introduced by the NNK codon (*Supporting Methods*). A pseudo aa was added for the entropy calculations of natural sequence alignments to account for gap occurrences. Positions displaying >25% gaps in the natural sequences MSAs were not displayed in the graphs to avoid distorting the entropy profiles. The relative entropy scores range between 0 and 1, corresponding, respectively, to total conservation and maximal exploration of sequence space (20 aa occur at same rate).

We hypothesized that N experimental sequences equal to that sampled in the screen would represent sufficient sequence space coverage to assure that most tolerated substitutions at a residue have occurred. To verify this assumption, a simulation was performed by using the following algorithm programmed in C++: n sequences were randomly selected from a complete set of N sequences, where n = 20,..., N. For each n sequences, entropy was calculated according to Eq. **1**. Construction of a subset of n sequences and entropy calculation was repeated 1 x 10^6 times, and the average entropy was calculated.

Comparison of Entropy Profiles and aa Selection by Standard Error of Proportion. The entropy profile and aa selection comparisons were calculated according to the standard error of proportion formula:

$$Z = \frac{Fe_{posX} - Ft_{posX}}{\sqrt{Ft_{posX} - (1 - Ft_{posX})/N}} \qquad [2]$$

In which N is the number of sequences in the sample, Fe_{posX} the frequency of an aa observed experimentally at a given residue and Ft_{posX} its theoretical frequency. A positive z score

means that the entropy is higher in the experimental data versus either of the natural sequence MSAs or that an aa experimental occurrence is higher than expected by chance (Figs. 2*E* and 3). The opposite is true for negative values (*Supporting Methods*).

Results

Synthesis and Screening of Libraries. The libraries were synthesized by PCR. To avoid unwanted bias for wt codons that can be introduced by this technique, the sequence to be targeted for degeneracy was removed from the wt Raf RBD cDNA before an amplification reaction that allows for insertion of the appropriate number of NNK codons (*Methods*).

The simplest way to evaluate the capacity of a sequence to fold into a target structure is to screen for the folded protein species ability to bind a known protein partner or ligand. We previously reported a simple survival-selection assay to screen libraries for protein-protein interactions based on the DHFR PCA in *E. coli*, which can detect the interaction between *h-ras* and the c-Raf (thereof Raf) RBD (20, 21). The residues of the Raf RBD directly involved in the formation of the interaction with *h-ras* were identified meticulously in a mutagenesis study (22). Several of these mutants were used to assess the sensitivity of the DHFR PCA assay to detect formation of the complex with *h-ras*. Colony formation was observed for variants of the Raf RBD displaying dissociation constant (K_d) between 130 nM and 14 μM. However, the R89L mutation, known to disrupt binding to *h-ras* (22), does not allow growth in this assay, and this residue shows low tolerance to mutation (Fig. 1*B*; see also Fig. 7 and Table 3, which are published as supporting information on the PNAS web site). Thus, we concluded that the DHFR PCA is sensitive enough to detect clones that fold and bind to *h-ras* with biologically relevant affinities. Based on the experimental strategy described above, we synthesized and screened 13 independent libraries (Fig. 1 *A* and Tables 1 and 2).

Validation of Experimental Data Set Size. We first determined whether the interpretation of the experimental data could be biased because of the limited sampling of sequences in this study (Table 1). To test this assumption, we devised an algorithm to evaluate how

Shannon entropy changes as the number of randomly sampled sequences included in the calculation is increased (*Methods*). If the sequence coverage is reasonable, the rate of change in entropy should approach zero as sequences are added. Results suggest that for the number of sequences sampled in this study, entropy variation converged toward zero whether a residue had low (V60) or high overall entropy (G107) (Fig. 2 *A*).

Comparisons of Sequence Diversity in the Libraries Versus Natural Sequences. The key validity test of our experimental strategy is to show that individual positions have sequence diversity equivalent to those found in nature, despite the fact that the rest of the polypeptide sequence is held constant during the selection process. If this premise is true, we reasoned that the positional entropy profiles of the experimental data set should reflect what is observed in natural sequences. We retrieved sequences for fh and sa of Raf RBD from databases and generated three MSA. The SMART MSA includes strict fh, whereas the Structural Classification of Proteins (SCOP) and SCOP-Families of Structurally Similar Proteins MSA include sa (*Supporting Methods* and Table 2; see also Tables 4-6, which are published as supporting information on the PNAS web site).

The entropy profiles of the experimental data set and SMART MSA reveal little similarity, except in convergence of local minima of entropy (such as V60, L62, and P63) (Fig. 2*B*), because of the very high local sequence similarity of this database MSA. Overall, the higher entropy of the experimental data set reveals that the sequence diversity generated is well above what is observed in SMART MSA. The sa MSA entropy profiles are more similar to the experimental data set entropy profile (Fig. 2 *C* and *D*). Specifically, 11 residues have entropy scores below at least one standard deviation from the mean in the experiments (Table 7, which is published as supporting information on the PNAS web site): I58, V60, L62, P63, T68, L78, L82, L86, C96, L126, and V128. On the same basis, six positions (58, 60, 78, 82, 126, and 128) correspond also to local minima in the entropy profiles of the sa MSAs. The comparisons of the experiments versus the SMART and both sa MSA entropy profiles by z score analyses reveal more positions with significant differences for the former comparison (Fig. 2*E* and *Methods*). In retrospective, these results

suggest that the strategy used has succeeded in reproducing the sequence diversity observed in known natural structures sharing the ubiquitin-roll topology.

The experimental entropy profiles show that the main α-helix (spanning L78-R89) core residues (L78, L82, and L86) support less degeneracy than the core positions located in the β-sheet (Fig. 2 *B*-*E*). This result might arise from the importance of maintaining the α-helix core packing for binding to *h-ras* as R89, a critical residue for binding (Fig. 1*B*), is located in this region. It is also possible that the core residues in the helix are more important for folding or stability of the Raf RBD than those in the sheet. On the other hand, we cannot exclude the possibility that strong positional selection in the helix-coding sequence might be a consequence of the fact that it was varied in two discrete segments, thus restraining putative local compensations specifically crucial for helices. Nevertheless, these results do not change our general conclusions and subtle local sequence constraints could be tested by using a library in which the entire helix or various segments of it are varied (*Supporting Results* in *Supporting Text*; see also Table 6, which is published as supporting information on the PNAS web site).

Hierarchy in the Hydrophobic Core. To analyze more closely the positions under selective pressure in the experiments, the aa occurrences observed experimentally and in the natural sequence MSAs were examined residue by residue by using standard error of proportion (z score) to reveal all significant aa selection (*Methods* and Fig. 3; see also Tables 8 and 9, which are published as supporting information on the PNAS web site). The experimental data set reveals that 30 of 76 positions have a z score for at least one type of aa with a P value <0.01 (*Supporting Results* and Table 9). Of these residues, 26 show strong selection for wt aa. Further, obvious convergence between the experimental, SCOP, and SCOP-Families of Structurally Similar Proteins MSA revealed 18 of 30 residues, which reside in two structural regions. They consist of an inner core (I58, V60, L78, L82, L86, V98, L126, and V128) readily evident in the entropy profiles (Fig. 2 *B*-*D*) and an outer core (L62, T68, V70, V72, C81, A85, L91, C96, L112, A118, and L121), which surrounds the inner core and form contacts at the interface between the α-helix and β-sheet (Fig. 4). This

hierarchy in the core is not apparent in thermal b-factor or solvent accessibility data (Fig. 8 and Table 10, which are published as supporting information on the PNAS web site). For example, residues of the outer core such as V72, C81, A85, and A118 have solvent accessibility comparable to inner core positions. Based on these observations, a signature sequence in the form of a series of significant aa selections at key residues of Raf RBD is presented (Fig. 3). The convergence of the aa occurrences observed experimentally to either of the natural sequence MSAs is used to predict the type of constraint, either functional or structural, imposed by selection at each residue of the signature sequence.

Key Topological Constraints. To further validate our data, we searched for predictable aa selections that follow from accepted principles of protein structure. For example, aa such as Pro and Gly display negative z score value in helical and β-strand segments, whereas they occur frequently in clones selected from libraries corresponding to loop or β-turn elements (libraries S2, S4, S7, S9, and S12). Also, hydrophilic residues are largely absent from hydrophobic core positions. We also observed strong selection for residues constrained by topology of the ubiquitin superfold. For example, residues S77, D80, G90, and L91 are located at the extremities of the major α-helix and show aa selection typical of helix capping motifs. On the other hand, residues P63 and N64, which form the first β-turn, represent examples of Raf RBD specific constraints. Indeed, aa selections at these positions are not as strong in the sa MSAs, because of variations in conformation and length of the matching structural segment (Figs. 2 and 3, *Supporting Methods*, and Table 4).

A Conserved Raf RBD Binding Patch for *h-ras*. Other residues show strong comparable aa selection in the experimental and fh but not in the sa MSAs (Figs. 2*E* and 3). This observation could suggest that these residues play a role in Raf RBD binding to *h-ras*. For example, observed aa selection and spatial proximity of the side chains of residues Q66, T68, R89, and A85 in the Raf RBD structure support their role in forming a critical binding surface for *h-ras* (Fig. 3 and Table 3). As discussed previously, R89 is the single residue of the Raf RBD, which is critical for binding to *h-ras*. Interestingly, Q66A and T68A decrease the affinity for *h-ras* (22), whereas A85K increase the affinity putatively by allowing for

binding to a wider range of GTPase conformers (23, 24). Consistent with the latter observation, our data indicates a strong selection for lysine at position 85.

Another subset of residues, including R100, D117, E125, and D129, shows specific convergence of aa selections with the fh MSA. These residues are not known to be involved in binding to *h-ras* and are far from the binding interface. Their mutation to Ala have no or only marginal effect on the affinity for *h-ras* in an *in vitro* binding assay (Table 11, which is published as supporting information on the PNAS web site). They could possibly be involved in structure stabilization specific to Raf RBDs (SMART MSA).

Clones Display wt-Like Folding and *h-ras* Binding. The folding and *h-ras* binding properties of Raf RBD clones were compared with the wt Raf RBD. To do so, we attempted the purification of 96 variants selected from the 13 libraries; 64 clones were purified with reasonable yield (Table 2). Far UV circular dichroism and proton NMR data suggest that the variants have conserved Raf wt structural characteristics (Fig. 5*A*; see also Fig. 9, which is published as supporting information on the PNAS web site). Moreover, similarity in the folding kinetic parameters between the variants and wt Raf RBD suggest conservation of folding mechanism despite large variation in folding rate (k_f) (Fig. 5*B*; see also Table 12, which is published as supporting information on the PNAS web site) (16). Finally, the observation that the Raf RBD variants/*h-ras* complex can be competed by wt RBD in a pull-down experiment, along with the K_d determined for some of these complexes, are coherent with the estimated sensitivity of DHFR PCA and validate our experimental scheme (Table 11 and Fig. 1*B*; see also Fig. 10, which is published as supporting information on the PNAS web site).

Discussion

Above, we presented a general strategy to expand the sequence space explored by a fold. This approach was used to generate a massive (72 of 76 residues were perturbed) sequence perturbation of a small protein. Strikingly, this set of functionally related sequences derived from the Raf RBD closely approximates the sequence space observed in sa MSAs as shown by the entropy profile comparisons (Fig. 2*E*). Nevertheless, the inner

core residues, which are among the most constrained residues by the absence of long-range covariation, display, in most cases, at least a slightly predominant selection for wt aa (Fig. 3 and Table 9). However, comparison of our results versus studies reporting partial degeneracy of a small protein such as λ-repressor, barnase, protein-L, and ubiquitin (10, 11, 14) reveal similar behavior of core residues despite variation in the location and dispersion of degenerated positions (e.g., concomitantly on a series of core residues or on a short segment of residues), suggesting that it might be stemming more from the limited number of residues covaried in each of these studies than from the nature of sequence perturbation. The potential bias introduced by fixing the sequence around a degenerated segment could be compensated for by the lower overall destabilization penalty induced by the segmental sequence perturbation versus corandomization of core residues. Alternatively, it might indicate that the local nature of structure formation has been underestimated. Lau and Dill (3) had observed that proteins are remarkably robust to point mutation, which is the main motor of natural sequence space expansion in fold evolution, although this type of alteration occurs at a very low rate in normal conditions (10^{-10} to 10^{-11} $\cdot bp^{-1} \cdot replication^{-1}$) (25, 26). It is also noteworthy that the most successful algorithm for predicting structure and *de novo* design, Rosetta, is based on optimizing local elements of structure on successive stretches of residues (27). In conclusion, the induction of massive sequence perturbation through a segment-by-segment approach does not reveal the complete sequence space compatible to a fold, but as reported here, it is sufficient to outline the repertory of sequence variation and constraints imposed upon it.

At present, there is no reliable method to screen for sequences capable of forming a specific target fold (28). Previously, studies have used protein-protein or protein-peptide interactions to screen libraries to select clones capable of folding into a specific structure (13, 14, 29). The residues of c-Raf RBD directly involved in binding to *h-ras* were thoroughly identified in a preceding mutagenesis study (22). However, to detect any other deviation in the experimental aa selection introduced by the experimental method, we compared the experimental data with sa and fh MSAs recovered from databases. Based on these comparisons, we proposed a signature sequence for Raf RBD (Fig. 3). Most of these consensus positions are also conserved in the sa MSAs and, thus, are consistent with the

proposal that they are also key features of the ubiquitin superfold. The number of residues in the signature sequence versus Raf RBD length (30 of 76) is consistent with the average sequence identity observed between redesigned proteins and their natural counterparts as reported in ref. 5. These results suggest that the experimental strategy outlined here can produce a MSA containing significant structural information. Interestingly, Rosetta performance in structure predictions was improved by adding MSA information to its regular algorithm routine (4). As demonstrated by this example, the capacity to artificially extend sequence space explored by any (poorly populated) folds could be helpful in protein design and structure prediction.

Lockless *et al.* (30) have proposed a method to identify mechanistically coupled functionally important residue pairs. This approach necessitates large MSAs to test pairs of residues for covariation. One could easily design experiments based on our strategy to generate large RBD sets in place of two or more remote segments of the primary structure or in a mutant background either of the RBD or of its binding partner *h-ras*. Residues important for their dimerization and showing specific convergence in the experimental aa selection and SMART MSA (Fig. 1*B* and 3) could constitute a good starting point for these investigations.

Finally, the experimental data and the sa MSA reveal a two-layer assembly of the hydrophobic core of the Raf RBD and of the ubiquitin superfold, which is reminiscent of a measure of global hydrophobic core formation that is based on micelle-like models used, coincidently, to improve Rosetta performance (4). Furthermore, structural observations of proteins sharing the ubiquitin-roll topology suggest that the spatial dispositions of residues found in the inner and outer core are conserved. Interestingly, the volume of inner core residues side chain is fairly constant across this superfold, particularly in superfamilies grouped in the SCOP MSA, with an average volume of 594 ± 49 $Å^3$ (Table 6). Similar observations were described for data obtained by degenerating concomitantly core residues of λ-repressor and barnase (10, 11). Moreover, theoretical studies have suggested that conservation of core volume might be particularly relevant for domains smaller than 200 residues, because they usually have higher packing density than larger domains (31).

Therefore, combinations of volume density dispersion at key positions in MSAs with simple graphical representation of protein structure could represent a way of unraveling unsuspected architectural or evolutionary linkages between sa and structure topologies (32, 33). Such additional geometrical constraints could be useful to improve structure prediction and protein design algorithms.

Footnotes

‡Corresponding author: email: stephen.michnick@umontreal.ca

Acknowledgements

We thank J. Bonvin and J. Turnbull for circular dichroism; P. Lepage, D. Roquis, and M. Fyfe for sequencing; S. Bilodeau and T. Viet for NMR; and A. Vallée-Belisle, S. Pontier, and M. Lerch (discussions). Natural Sciences and Engineering Research Council and Canadian Institutes of Health Research (CIHR) funded this project. F.X.C.V. was a CIHR, Program de Biologie Moléculaire, and Faculté des Etudes Supérieures scholar. S.W.M. is the Canada Research Chair in Integrative Genomics.

References

1. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13,** 669–678.
2. Lesk, A. M. & Chothia, C. (1980) *J. Mol. Biol.* **136,** 225–270.
3. Lau, K. F. & Dill, K. A. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 638–642.
4. Bonneau, R., Strauss, C. E. & Baker, D. (2001) *Proteins* **43,** 1–11.
5. Dantas, G., Kuhlman, B., Callender, D., Wong, M. & Baker, D. (2003) *J. Mol. Biol.* **332,** 449–460.
6. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003) *Science* **302,** 1364–1368.
7. Scalley-Kim, M. & Baker, D. (2004) *J. Mol. Biol.* **338,** 573–583.
8. Arkin, A. P. & Youvan, D. C. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 7811–7815.
9. Reidhaar-Olson, J. F. & Sauer, R. T. (1988) *Science* **241,** 53–57.
10. Lim, W. A. & Sauer, R. T. (1989) *Nature* **339,** 31–36.
11. Axe, D. D., Foster, N. W. & Fersht, A. R. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 5590–5594.
12. Finucane, M. D. & Woolfson, D. N. (1999) *Biochemistry* **38,** 11613–11623.
13. Gu, H., Kim, D. & Baker, D. (1997) *J. Mol. Biol.* **274,** 588–596.
14. Kim, D. E., Gu, H. & Baker, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 4982–4986.
15. Emerson, S. D., Madison, V. S., Palermo, R. E., Waugh, D. S., Scheffler, J. E., Tsao, K. L., Kiefer, S. E., Liu, S. P. & Fry, D. C. (1995) *Biochemistry* **34,** 6911–6918.
16. Vallee-Belisle, A., Turcotte, J. F. & Michnick, S. W. (2004) *Biochemistry* **43,** 8447–8458.
17. Soding, J. & Lupas, A. N. (2003) *Bioessays* **25,** 837–846.
18. Campbell-Valois, F.-X. & Michnick, S. W. (2005) *Methods and Protocols in Molecular Biology*, in press.
19. Pelletier, J. N., Campbell-Valois, F.-X. & Michnick, S. W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 12141–12146.

20. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
21. Pelletier, J. N., Arndt, K. M., Pluckthun, A. & Michnick, S. W. (1999) *Nat. Biotechnol.* **17,** 683–690.
22. Block, C., Janknecht, R., Herrmann, C., Nassar, N. & Wittinghofer, A. (1996) *Nat. Struct. Biol.* **3,** 244–251.
23. Fridman, M., Maruta, H., Gonez, J., Walker, F., Treutlein, H., Zeng, J. & Burgess, A. (2000) *J. Biol. Chem.* **275,** 30363–30371.
24. Fridman, M., Walker, F., Catimel, B., Domagala, T., Nice, E. & Burgess, A. (2000) *Biochemistry* **39,** 15603–15611.
25. Grishin, N. V. (2001) *J. Struct. Biol.* **134,** 167–185.
26. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998) *Genetics* **148,** 1667–1686.
27. Simons, K. T., Strauss, C. & Baker, D. (2001) *J. Mol. Biol.* **306,** 1191–1199.
28. Waldo, G. S. (2003) *Curr. Opin. Chem. Biol.* **7,** 33–38.
29. Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. & Baker, D. (1997) *Nat. Struct. Biol.* **4,** 805–809.
30. Lockless, S. W. & Ranganathan, R. (1999) *Science* **286,** 295–299.
31. Liang, J. & Dill, K. A. (2001) *Biophys. J.* **81,** 751–766.
32. Kannan, N. & Vishveshwara, S. (1999) *J. Mol. Biol.* **292,** 441–464.
33. Lindorff-Larsen, K., Rogen, P., Paci, E., Vendruscolo, M. & Dobson, C. M. (2005) *Trends Biochem. Sci.* **30,** 13–19.

Figure legends, Figures and Tables:

Figure 1 Description of the sequence perturbation methodology. (*A*) Experimental strategy. (*1*) The Raf RBD is subdivided into 13 segments based on topological elements. Residues in parentheses were unvaried in the experiments. (*2*) Each segment is degenerated separately by PCR (Fig. 6). (*3*) Libraries are screened by using DHFR PCA. (*4*) Sequence diversity observed in experiments and database MSAs are compared. (*B*) Growth observed on selective media with *Escherichia coli* cells cotransformed with *h-ras* and a set of RBD mutants tethered to DHFR PCA fragments (K_d for each mutant is indicated in micromolars between parentheses below the Petri dish).

Figure 2. Validation of experimental data set and comparison of entropy profiles. (*A*) The variation in mean entropy calculated for two successive values of n (n and $n + 1$) is plotted for V60 and G107 representing, respectively, "low" and "high" entropy position. Next, the entropy profile obtained experimentally is plotted against the entropy profiles calculated for three MSA of natural sequences: Raf RBD fh (SMART) (*B*), close sa (SCOP) (*C*), and close and distant sa [SCOP-Families of Structurally Similar Proteins (FSSP)] (*D*). (*E*) The experimental entropy profiles and the three database MSAs are compared by z score after the color scale.

Figure 3. Amino acid selections. The z scores of aa selections for experimental libraries and the three natural sequence MSAs are represented in a color-coded matrix after the color scale shown. Under each residue, displayed left to right on top of the matrix, there is a column of 20 cells corresponding to the aa types. Each cell is divided into four squares (see scheme below the matrix) to facilitate comparison of the MSAs. The Raf RBD signature sequence is presented below the matrix by indicating significant amino acid selection, which are classified either as conserved broadly in the ubiquitin superfold (bold) and the ubiquitin-related superfamilies (black) or specifically in the Raf RBDs (blue) (Table 9).

Figure 4. Hierarchy in the hydrophobic core. Structural disposition of residues that constitute the inner (red) and outer core (green) of Raf RBD. Note that three residues for which insufficient experimental data are available likely belong to the outer core as we define it (Q66, R89, and W114).

Fig. 5. Clones selected display characteristics of folded proteins. (*A*) Circular dichroism spectra for five clones from five libraries interspersed in the Raf RBD sequence. (*B*) Chevron curves obtained from folding and unfolding experiments in GuHCl for the clones presented in *A*. *Inset* shows the distribution of folding rates in water extrapolated from the chevron curves for all clones with two-state folding behavior that were studied. Note that clones with folding rate (k_f) >450 s^{-1} are grouped in a single class.

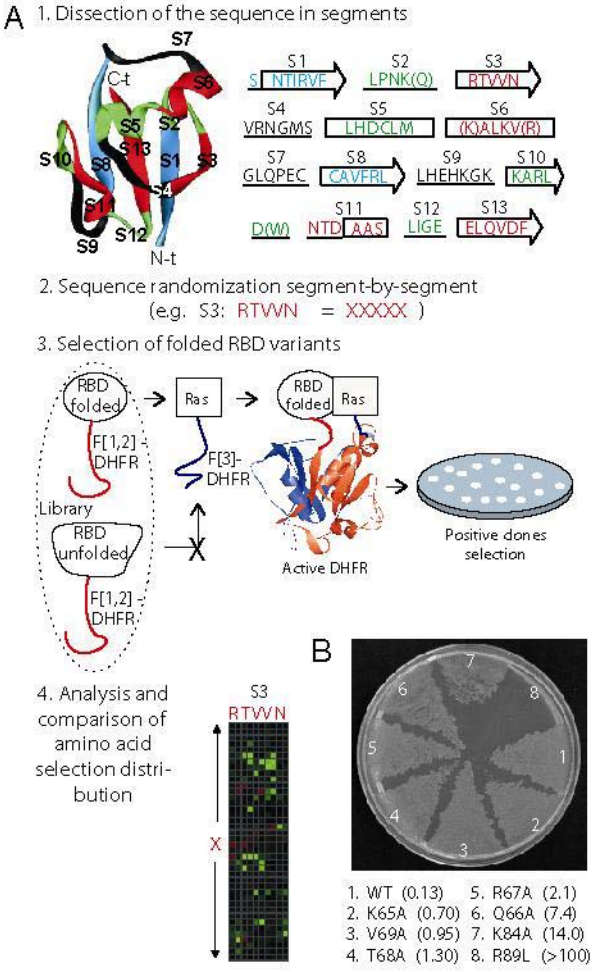

A
1. Dissection of the sequence in segments
S7
C-t
S6
S5
S2
S3
S10
S8
S1
S4
S11
S12
S9
N-t
S1 S2 S3
S NTIRVF LPNK(Q) RTVVN
S4 S5 S6
VRNGMS LHDCLM (K)ALKV(R)
S7 S8 S9 S10
GLQPEC CAVFRL LHEHKGK KARL
S11 S12 S13
D(W) NTD AAS LIGE ELQVDF
2. Sequence randomization segment-by-segment
(e.g. S3: RTVVN = XXXXX)
3. Selection of folded RBD variants
RBD folded
F[1,2] - DHFR
Library
RBD unfolded
F[1,2] - DHFR
Ras
F[3]- DHFR
X
RBD folded
Ras
Active DHFR
Positive clones selection
4. Analysis and comparison of amino acid selection distri-bution
S3
RTVVN
X
B
1. WT (0.13) 5. R67A (2.1)
2. K65A (0.70) 6. Q66A (7.4)
3. V69A (0.95) 7. K84A (14.0)
4. T68A (1.30) 8. R89L (>100)

Fig . 1

Table 1. Statistics concerning the synthesis and screening of the 13 degenerate libraries (see **Methods** and **Supporting Methods**).

Libraries	**Theoretical Number of sequences ($x10^6$)**	**Number of clones generated ($x10^6$)**	**% of clones selected**	**Number of clones sequenced**
S1	85.8	5.9	0.18	65
S2	0.19	2.2	0.28	61
S3	4.1	2.2	0.55	70
S4	85.8	2.4	0.25	118
S5	85.8	2.5	0.07	72
S6	0.19	0.9	0.7	81
S7	85.8	1.2	0.04	67
S8	85.8	2.3	0.3	91
S9	1800.1	3.1	1.29	74
S10	4.1	1.6	0.54	82
S11	85.8	3.2	0.39	72
S12	0.19	1.7	0.37	69
S13	85.8	1.9	0.16	64

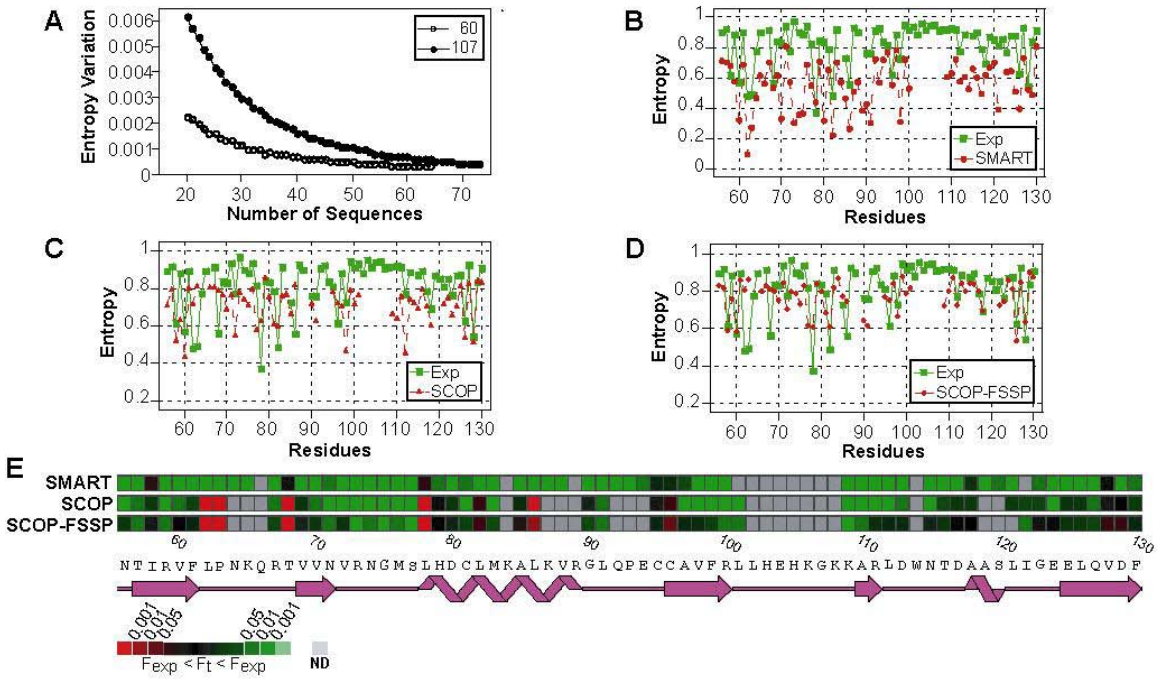

A
Entropy Variation
0.006
0.005
0.004
0.003
0.002
0.001
0
60
107
20 30 40 50 60 70
Number of Sequences
B
Entropy
1
0.8
0.6
0.4
0.2
0
Exp
SMART
60 70 80 90 100 110 120 130
Residues
C
Entropy
Exp
SCOP
60 70 80 90 100 110 120 130
Residues
D
Entropy
Exp
SCOP-FSSP
60 70 80 90 100 110 120 130
Residues
E
SMART
SCOP
SCOP-FSSP
60 70 80 90 100 110 120 130
NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLLHEHKGKKARLDWNTDAASLIGEELQVDF
0.001
0.01
0.05
0.05
0.01
0.001
Fexp < Ft < Fexp
ND

Fig. 2

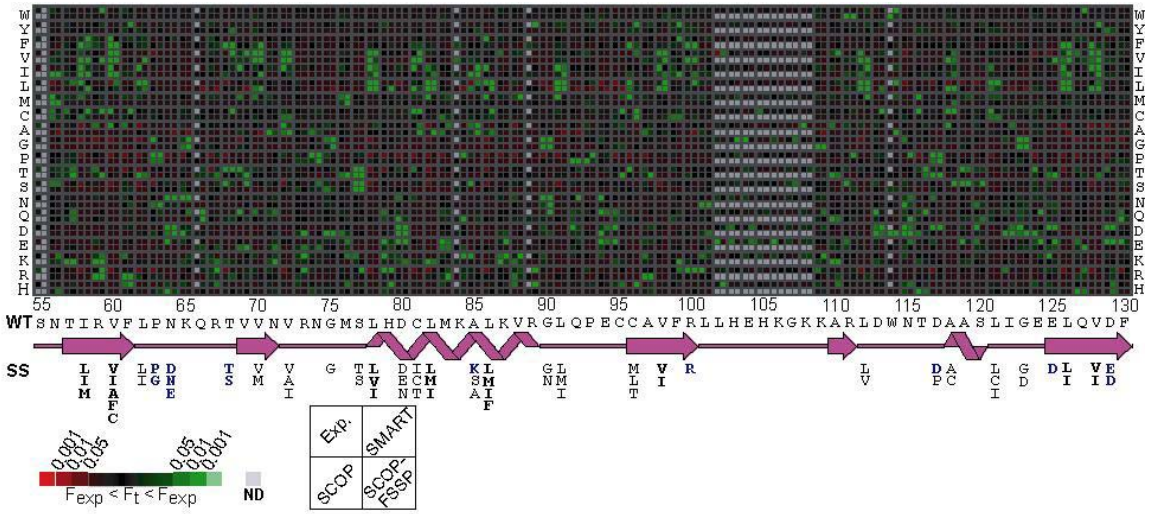

WT SNTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLLHEHKGKKARLDWNTDAASLIGEELQVDF
SS
Exp.
SMART
SCOP
SCOP-FSSP
ND

Fig. 3

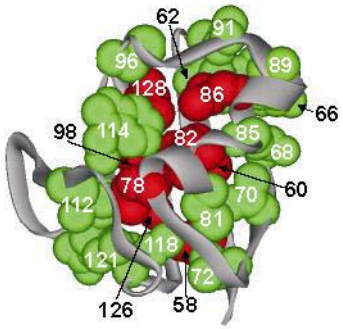

62
91
96
89
128
86
66
98
114
82
85
68
78
60
70
112
81
118
121
72
126
58

Fig. 4

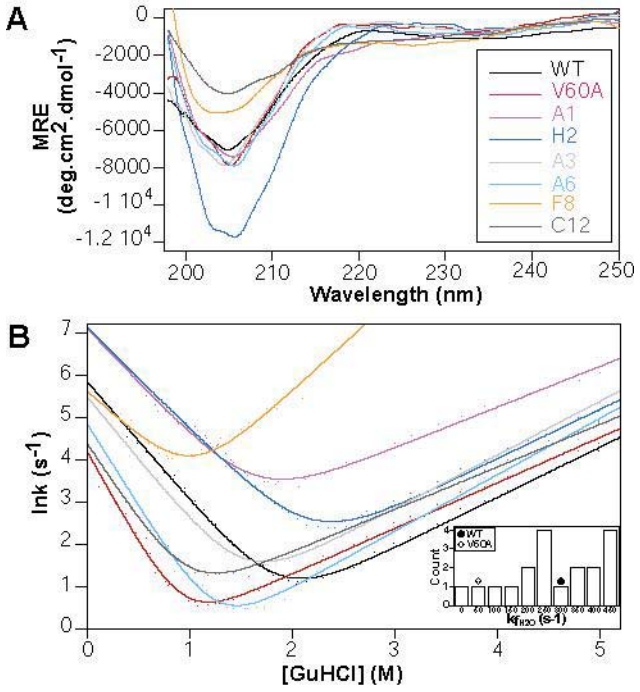

A
MRE (deg.cm2.dmol-1)
0
-2000
-4000
-6000
-8000
-1 10^4
-1.2 10^4
200
210
220
230
240
250
Wavelength (nm)
WT
V60A
A1
H2
A3
A6
F8
C12
B
lnk (s-1)
0
1
2
3
4
5
6
7
[GuHCl] (M)
Count
kH2O (s-1)

Fig. 5

Article 2 : Informations supplémentaires

Supporting Information for manuscript:

"Massive sequence perturbation of a small protein"

Running-title: Massive sequence perturbation without long-range co-variation on Raf RBD approximates the sequence diversity explored by ubiquitin superfold in nature.

Authors: Campbell-Valois, F.-X.*†, Tarassov, K.*, Michnick, S.W.*‡

* Département de Biochimie and † Programme de Biologie Moléculaire , Université de Montréal, C.P. 6128, Succ. centre-ville, Montréal, Québec, Canada H3C 3J7

‡Corresponding author: email: stephen.michnick@umontreal.ca

phone: (514) 343-5849

fax: (514) 343-2015

Supporting Text

Supporting Methods

Basic DNA Constructs. The construction of the two vectors (pQE32 Raf RBD-DHFR F[1,2] and pQE32 *h-ras*-DHFR[3]) necessary for the screening of the degenerated libraries (see below) has been described in ref. 1. The plasmid pQE32 Raf RBDDHFR F[1,2] have been modified as follows: NcoI and XhoI restriction sites were added respectively in 5' and 3' of the Raf RBD cDNA. The first site resulted in the addition of a methionine and a glycine codon between the hexahistidine tag and the beginning of the cDNA (thus there are three additional residues between the tag and cDNA). The second site was introduced in 3' of the cDNA by changing the codon of residue 131 and inserting the conservative mutation D132E. Finally, the vector size was reduced to improve transformation efficiency by removing ≈800 bp of DNA between the compatible cohesive restriction site XbaI and NheI located in 3' between the terminator and the origin of replication (2).

Design of Template for the Degenerated PCR. A deletion-tagged template for each of the 13 libraries was generated in which the region to be targeted for degeneracy (Fig. 1*A* and Table 1) is replaced with a unique recognition site for a restriction enzyme enabling unambiguous identification of the templates. In addition, a stop codon and a 1-bp frame shift were also inserted to eliminate a potential source of contamination with the wild-type (wt) cDNA in the screening process. The PCR protocol used to synthesize the templates is a variant of the ExSite protocol (Stratagene). The rest of the procedure described below is represented in Fig. 6.

Library Cloning and Recuperation. The product obtained after the second-round PCR was cloned in the entry vector, derived from pQE32 (Qiagen) (as described above), carrying DHFR F[1,2] by ligation through the restriction sites SphI and XhoI. The ligation product was transformed by electroporation in SS320 *Escherichia coli* strain (3). The number of resulting independent clones was estimated by counting the number of colonies formed after overnight incubation at 37°C on LB plates (supplemented with 25 μg/ml kanamycin/100

μg/ml ampicillin/10 μg/ml tetracycline) upon plating a fraction of the transformation reaction (1×10^{-4} dilution) (Table 1). The rest of the transformation mixture was used to inoculate 500 ml of LB supplemented with the antibiotics mentioned above. After 8-12 h of incubation, plasmidic DNA was prepared. Special care was taken at all steps to avoid contamination with the wt plasmid because it would confound the screening procedure.

Multiple Sequence Alignments (MSA) of Natural Sequences. An MSA of functional homologues was retrieved from data mining in the SMART database [Tables 4 (MSA S1) and 5] (4). Near homologues (>85% identity) of a sequence were removed from the alignment. In addition, we retrieved 54 structures [including 1RFA, the structure of the Raf ras binding domain (RBD)] sharing the ubiquitin fold, but displaying <35% sequence identity in pairwise alignment. Two alignments were constructed from these sequences. The Structural Classification of Proteins (SCOP) MSA included sequences classified strictly into five superfamilies belonging to the ubiquitin superfold according to the SCOP database (5): the ubiquitin-like superfamily and four superfamilies probably linked to it evolutionarily, the CAD and PB1 domains, MoaD/ThiS, TGS-like domain, and Double Cortin (DC). The SCOP-Families of Structurally Similar Proteins (FSSP) MSA included all sequences from the SCOP MSA, plus sequences retrieved through data mining in four of the remaining six superfamilies belonging to the ubiquitin superfold and in the FSSP database (6). Therefore, the SCOP-FSSP MSA includes structures more dissimilar to the Raf RBD than the SCOP MSA. These MSAs were constructed by aligning residues according to secondary structure or topological elements. Then, the side-chain gross orientation (e.g., facing core interior or solvent), the spatial position in the tertiary structure context, or similarities in inter-side-chain contacts were used in case the first criterion was insufficient. All insertions were deleted, thus keeping only the elements common with the Raf RBD across all sequences [Tables 4 (MSAs S2-S3) and 6]. Finally, note that for simplicity throughout the text, figures and tables that Raf RBD residue numbering is used to identify position in all MSAs.

Theoretical Amino Acid Frequency Used to Normalize Experimental Amino Acid Frequency for Shannon Entropy Calculation and for Natural Sequence MSA Ft_{posX}. The theoretical bias in occurrences imposed by the NNK codon before selection is as

follows: (W)1: (Y)1: (F)1: (V)2: (I)1: (L)3: (M)1: (C)1: (A)2: (G)2: (P)2: (T)2: (S)3: (N)1: (Q)1: (D)1: (E)1: (K)1: (R)3: (H)1: (stop)1. These biases in occurrence were used to normalize the experimental data amino acid occurrences for Shannon entropy calculation (Eq. **1**). The theoretical number of clones in each library varies after the function 21^{l} in which l is the number of different amino acids plus the stop codon allowed in the experiments (Table 1)

The following value of amino acid frequency according to swiss-prot (December 2003) were used to calculate standard error of proportion (Eq. **2**): (W)2.21: (Y)2.87: (F)3.11: (V)7.07: (I)8.40: (L)7.70: (M)7.87: (C)1:5.49 (A)5.25: (G)3.86: (P)3.41: (T)6.05: (S)4.68: (N)4.86: (Q)4.93: (D)4.51: (E)5.17: (K)5.23: (R)3.77: (H)3.56.

Entropy Score Amino Acid Selection Residue-by-Residue Representation with a Color-Coded Matrix. z scores were calculated for each residue to determine the difference in the entropy profile for the experimental data (Fe_{posX}) versus each of the three natural sequence MSAs (Ft_{posX}) (Fig. 2*E* and Table 7). For each comparison, *N* was fixed independently according to the *N* number of sequences in the alignment with the smallest number of sequences (SMART, SCOP, and SCOP-FSSP depending on the profiles compared; see Table 7). The color scale displayed in Fig. 2*E* was used for allocating color to each amino acid on a residue-by-residue basis (see below the description of amino acid distributions analysis for more details).

For calculation of the amino acid distributions and to identify the extent of selection for individual amino acids at individual sequence positions in the experimental data set, the theoretical frequency (Ft_{posX}) was fixed according to the bias introduced by using the NNK codon. Also, the number of clones (*N*) and the occurrence of each amino acid at a given residue were normalized to the smallest data set (Table 1) to avoid local bias due to differences in the number of clones isolated for each library (Eq. **2**). For the natural sequence MSAs, the amino acid frequencies in swiss-prot as it appeared in December 2003 were used as Ft_{posX} (for both of the aforementioned sets of Ft_{posX}, see previous section above).

The z scores are represented as a color-coded matrix (Fig. 3). Amino acids at a given residue with positive and negative z scores are displayed respectively in green and red after the scale for *P* value displayed in Fig. 3. The three brightest shades of green and red correspond from the darkest to the brightest, respectively, to *P* values of 0.05, 0.01, and 0.001 (see scale in Fig. 3). The z score corresponding to a given *P* value for each alignment (as *N* varies) was obtained from a *t* distribution critical *t* values limits table. Gray cells indicate sequence positions that were not degenerated in the libraries. The cells for residues 102-108, which correspond to a RBD specific loop, were also colored gray in the structural analogue MSAs. The z score value was calculated for all positions including those displaying >25% gaps. Sequences with gaps were counted in the *N* value, but gap amino acid distributions are not shown.

Construction of the Signature Sequence (SS). Thirty-eight positions have at least one significant ($P < 0.05$) amino acid selection (Table 9). In the manuscript, we discuss mainly the 33 positions showing even further significance ($P < 0.01$) in their amino acid selection. As discussed in Table 9, three positions were disregarded, because the significance of the amino acid selection was ambiguous. For these positions ($P < 0.01$), all amino acid with selection z score with $P > 0.05$ are displayed. This procedure reduces the SS presented in the manuscript to 30 positions. The amino acid selection distributions at these positions in the experiments were compared with those observed in the natural sequence MSAs to propose either functional homologue specific or structural analogue evolutionary pressure (Fig. 3 and Table 9).

Physical Characterization. The 96 clones selected for physical characterization were purified on Ni-affinity column (Qiagen). The circular dichroism experiments were performed on a JASCO-710 spectropolarimeter at 25°C in a quartz cuvette with a 1-mm path length. In these experiments, protein samples were diluted to 15 μM in 25 mM sodium acetate buffer (pH 4.9).

The kinetic and equilibrium folding experiments were performed with a stopped-flow apparatus model SX.18 MV under fluorescence mode (Applied Photophysics). The wild-type W114, unvaried in the experiment, is the intrinsic fluorescent probe used to monitor folding and unfolding reactions of the Raf RBD (7). The quantum yield of the tryptophan fluorescence decreased upon denaturation of the Raf RBD. GuHCl was used to denature protein in 25 mM NaoAc buffer at pH 4.9. Routinely, protein stock solutions at a concentration of 40 μM in 3.5 M and 0.3 M GuHCl, respectively, for folding and unfolding experiment, were diluted 1:10 in denaturant solution. The concentrations of GuHCl in all stock solutions and in the protein samples were estimated by refractometry.

The sequences of clones presented in Fig. 5 *A* and *B* are the following: a1 (library 1), "PWVDLDA"; h2 (library 2), "FTDG(Q)"; a3 (library 3), "GTRVT"; a6 (library 6), "(K)KLSE(R)"; f8 (library 8), "CKLMRR"; c12 (library 13), "TSCHDL," The other clones purified and characterized (Fig. 5 and Tables 11 and 12) are indicated below (Table 2).

Determination of the Dissociation Constant for the Raf RBDs. The dissociation constants (K_d) of several clones were determined according to a method described in ref. 8. Fluorescence was monitored in 96 or 384 well nonbinding surface (NBS) black plates (Corning no. 3650 and 3654) with a Gemini Xs plate reader (Molecular Devices). Typically, the K_d of Raf RBD variants for *h-ras* was calculated by fitting triplicate measurements at each Raf RBD concentration. The K_d determined experimentally are listed in Table 12.

NMR Experiments. Purified protein samples were transferred into deuterated 12 mM sodium acetate buffer, pH 5.3/0.1 mM DTT/1 mM NaN_3 on a PD10 desalting column (Amersham Pharmacia). Then, the samples were concentrated (between 0.5-1 mM) by ultrafiltration and ^2H_2O added to reach 94% H_2O/6% ^2H_2O. Proton NMR spectra were performed on a DMX-500 Bruker spectrometer at 25°C. Water suppression was accomplished by using a phase-sensitive pulsed field gradient approach, and NOESY spectra were collected with a mixing time of 125 ms (9, 10).

His-Tag Pull-Down and Competition. One hundred micrograms of 6×His-Raf (wt or clones) were incubated with moderate mixing by inversion at 4°C with 200 μl of Ni-NTA resin in 10 bed volumes (BV) of buffer C (20 mM Hepes/5 mM MgCl2/150 mM NaCl/25 mM imidazole, pH 7.4) for 20 min. Meanwhile 150 or 225 μg of GST-*ras* (depending on the affinity of the variant tested) bound to GMP-PNP (8) was preincubated either with four to six times more concentrated untagged Raf wt or an equal volume of a purification fraction from cells transformed with empty vector in 1 ml of buffer C. Next, the resin-bound Raf sample was split in two, the supernatant discarded, and the preincubated GST-*ras* added and incubated at 4°C for 30 min. The bound fraction was washed twice with 10 BVs buffer C. The volume of resin was estimated and adjusted with buffer C to ≈100 μl, and 3× SDS/PAGE loading buffer was directly added to the mix. The samples were boiled briefly, and 35 μl of the samples were loaded per lane on a 15% SDS/PAGE. The bands were revealed by Coomassie blue staining. Quantification of GST-*ras* pull-down in the uncompeted vs. untagged Raf wt treatment was done with quantity one software (BioRad) by subtracting the background and calculating the ratio between these two treatments for a Raf variants.

Software Used. The secondary structure representation for human Raf RBD (PDB ID code 1RFA) displayed in Fig. 2*E* and 3 was obtained on the PDBsum server (www.biochem.ucl.ac.uk/bsm/pdbsum). The graphs and curves were designed and fitted with prism (GraphPad) and kaleidagraph 3.52 (Synergy Software). The figures were assembled with corel draw 9.0 (Corel) and illustrator 9.0 (Adobe Systems). Structure representations were done with ds viewer pro (Accelrys). The NMR figure was prepared by using topspin 5.0.

Tous les tableaux du materiel supplémentaires sont disponibles sur Internet sur le site/All tables of Supporting information can be found on the internet at : www.pnas.org. Ils ont été annexées à la thèse suite aux commentaires du jury d'évaluation.

Supporting Results

Alignment of Sequences Obtained Through Screening of the 13 Degenerated Libraries. Each library was screened independently according to the protocol described above (Fig. 1*A*, *Methods*, and Fig. 6). The statistics presented in Figs. 2 and 3 were solely based on clones selected through a single screening step. The sequences for these clones are presented in Table 2. It also indicates the clones that were purified and further characterized (Fig. 5 and Table 11).

Each library was kept under selective pressure during several rounds of amplification in liquid media (according to ref. 16). The sequences of these clones are as follows (The indices used here have the same meaning as in Table 2): Library S1: "FKLILTY"†, "PDHLRFE"†; library S2: "LPDTQ*"†, "FTDGQ*"†‡§; library S4: "PDSSST"†, "AVPSLR"†, "VSLGHK"†; library S5: "LKKLLL"†; library S6: "K*RMTAR*"†‡§, "RSLHRR"†; library S7: "EINHLQ"†‡§, "EILPGQ"†‡§; library S8: "CKLMRR"†, "MVPMRE"†; library S9: "KTQCNG"†, "TSGRVLH"†; library S10: "KRTVSW*"†‡§; library S11: "TGEAIS"†, "SSGAHG"†; library S12: "VAGN"†, "LADC"†‡§; library S13: "QLLLEF"†‡§. From the limited sampling of clones (above and Table 2), no obvious bias for binding to GTP bound *h-ras*, stability, or kinetic of folding/unfolding is observed for those isolated at 1 versus 12 rounds of selection.

Alternative Experimental Libraries. We have synthesized three alternative degenerated libraries for the Raf RBD. The alternative libraries S2b, S6b, and S8b are homologous to libraries S2, S6, and S8 respectively. The libraries S2b and S6b were synthesized and screened to probe the impact of degenerating residues previously unperturbed (Fig. 1 and Table 1), known to be involved in binding to *h-ras* (e.g., Q66, K84 and R89) on the number

of colonies formed, and on the sequence space explored at neighboring residues (Fig. 7). Effectively, fewer colonies (on the order of 1×10^2) were formed in the survival-screening assay with libraries S2b and S6b than with the original S2 and S6 (Table 1). We have also made library S8b to probe the effect of keeping residue 100 constant for arginine, which is the amino acid occurring in the wt Raf RBD. The sequences selected from these three alternative libraries with the DHFR PCA survival-screening assay are listed in Table 3.

Simply by looking at the sequences (Table 3), there appears to be an interesting correlation between the nature of the amino acids observed at residues L62 and Q66. These two residues are located side by side near the turn of the first β-hairpin in the second and first β-strands, respectively. Small amino acids such as G and A at position 66 correlated with F at position 62. Residue 66 is also preferentially biased for Q. When this amino acid occurs at position 66, a leucine (the wt amino acid) is frequently observed at position 62. For the library 6b, the most important observation concerns the high conservation of residue 89 for arginine. As described in the article, mutation of this residue to leucine is sufficient to completely abrogate binding of the Raf RBD to *h-ras* in an *in vitro* binding assay (17) and in our survival selection assay (Fig. 1*B*). On the other hand K84, which was shown to increase the dissociation constant by two orders of magnitude ($K_d \approx 14$ μM), shows very low conservation. Overall, there is little entropy variation for residues in the two alternative libraries (sets 2b and 6b) versus the corresponding main libraries (Fig. 7 *A* and *B*). On the contrary, the effect of keeping R100 invariable on the amino acid diversity at residues V98 and L101 is evident (Fig 7*C*). R100 is a residue of the outer core located in the β-strand 3 (Figs. 3 and 4 and Table 9). This observation might explain why fixing the amino acid identity at position 100 reduces the sequence space explored by V98 the neighboring core residue of β-strand 3. Alternatively, the decrease in entropy for L101 in that context could be due to the fact that this residue has both of its flanking neighbors fixed. Indeed, a drastic decrease on the amino acid degeneracy tolerated at L101 upon fixation of the amino acid identities at both versus only one of its neighboring residues is clear. It is even more striking given that L101 displayed high entropy in the main experiment (Fig. 2*B* and 7). This observation indicates the strong potential of the polypeptide chain to adapt its conformation

to local sequence constraints. Alternatively, this effect could be due to Raf RBD specific factors, because R100 displays specific convergence in its amino acid selections with the functional homologues alignment (Fig. 3).

In conclusion, these results should be taken with care because of the small number of sequences analyzed. However, the results highlight the potential use of our approach to study covariation of residues in the same protein or between residues located across dimerization interfaces.

Description of the MSAs of Natural Sequences and of the Experimental Data Set.

Below are presented the alignments of natural sequences obtained through database mining. The sequences were edited as described above. Note that the Raf RBD residues numbering is used for all alignments throughout the manuscript and in this section.

The properties of the diverse sequences included in the functional homologues alignment (Table 4, MSA S1) were retrieved by searching on the blast search engine and are included in Table 5.

Properties of the different structures used in the structural analogue alignments were gathered from the SCOP database (Table 4, MSA S2 and S3). Other structural details gathered by manual evaluation of the structures and alignments are also listed in Table 6.

Variation in Position of the α-Helix over the β-Sheet and Superfamilies Absent in MSA S2 and S3. The arrangement of the α-helix over the β-sheet across the ubiquitin superfold is variable and can be approximated by the Ω-angle. In structures sharing this topology, a positive and negative Ω-angle would indicate, respectively, that the α-helix is packed over β-strand 1, 2, and 5 and 1, 3, and 5. Structures included in the alignment displayed a negative Ω-angle, with the vast majority of structures falling between -45° and -75°. One notable exception is the structure of the translation initiation factor IF3 (1TIF) classified in superfamily 10, which shows slightly positive angle (≈0-5°). Because the network of contacts in the core of this domain, particularly at the α-helix and β-sheet interface, is drastically different, it was not included in the SCOP-FSSP alignment. The

same argumentation was applied for rejecting structures (such as PDB ID code 1LGR) classified in superfamily 11.

Variation in α-Helix Length. The main α-helix (for the Raf RBD it comprises residues L78-R89) varies in length from 8 to 16 residues across superfold members included in the alignments. A majority of structures (18 of 27) in the SCOP alignment has a α-helix of 12 residues in length, similar to the Raf RBD. There is much more variation in the SCOP-FSSP alignment. It is mainly due to the large number of structures of the 2Fe-2S Ferredoxin-like superfamily (5), which display, in majority, a shorter α-helix. In most of these cases, the α-helix lack the last turn in carboxyl terminal (thus, their helix span eight residues in length). Thus, these domains do not possess a third inner core residue in the α-helix (inner and outer core residues are defined in Fig. 4). This structural variation is the most frequent case leading to discrepancy in the conservation of the inner core residues in members of the ubiquitin superfold (Figs. 3 and 4 and Table 6).

Effect of the Variation in the Register of the α-Helix on the Entropy Scores. An added feature of the analyses of structural differences among superfold members reveal cryptic sequence misalignments in the natural sequences. Discrepancies between experimental positional entropies in the α-helix (L78-R89) and structural analogues MSA were found to be due to changes of the α-helix position on the β-sheet. The main α-helix core residues facing the β-sheet in Raf RBD are *i* (L78), *i* + 4 (L82) and *i* + 8 (L86), whereas for ubiquitin these residues are *i* (I23), *i* + 3 (V26) and *i* + 7 (I30). The structures classified in the SCOP and SCOP-FSSP alignments adopt more frequently the ubiquitin arrangement (respectively, 21 of 27 and 32 of 54). Therefore, the entropy score for the second and third core residue of the α-helix were recalculated to take into account these observations. Straight forwardly, it was done by aligning position *i* + 3 and *i* + 7 with position *i* + 4 and *i* + 8 of structures displaying, respectively, ubiquitin and Raf-RBD type packing to calculate entropy. In this manner, the new positional entropy scores obtained for the SCOP-FSSP alignments are 0.51 and 0.61. These values correspond closely to those calculated from the experimental data of Raf RBD at residue L82 and L86, respectively, 0.49 and 0.59. These variations in the

packing of the α-helix create changes in the disposition of the hydrophobic core, including residues of the β-sheet. This latter phenomenon is illustrated by residue L62, which is intimately associated with the hydrophobic core in the Raf RBD, whereas in ubiquitin, the corresponding residue is largely exposed to solvent.

Inner Core Side Chain Volumes. The volume of the side chains of the inner core (Fig. 5 and Table 6) is fairly constant across the ubiquitin superfold with an average volume of 556 $Å^3$ and a standard deviation of 78 $Å^3$. For domains classified in the SCOP alignment (*Methods*), average side chain volumes were more similar (594 ± 49 $Å^3$). The variations in cumulative side-chain volume depending on the structural analogue alignments analyzed are mainly due to the slightly diverging α-helix arrangement described in the article and above (Table 6 and *Variation in α-Helix Length* above). The Raf RBD is close to that average with a volume of 601 $Å^3$ for the side chains of inner core residues. Interestingly, conservation of core residues volume was demonstrated in folded clones of the lambda repressor isolated from a library in which 7 core residues were concomitantly degenerated for the 20 types of amino acids (13). Similar conclusions can be drawn of a study on barnase (14). Based on an alignment of sequences retrieved in the SMART database, the volume of the side chain of inner core residues (based on the src-SH3 sequence, the positions in the alignments that correspond to the inner core are as follows: 12A, 26F, 32L, 34I, 45A, 54G, 56I, and 60V) in the large SH3 family seems also to be highly conserved, displaying an average of 443 ± 31 $Å^3$. The significance and generality of the conservation of the volume of inner core residue to structure or structure formation is unknown. However, recent theoretical work suggested that conservation of core volume might be particularly relevant for domain <200 residues because they have usually higher packing density than larger domain (15).

Evaluating Sequence Diversity in Alignments. The general properties of the experimental and natural sequence alignments can be summarized by the mean positional entropy and by the mean pairwise sequence identity. These values are listed in Table 7 to aid comparing the sequence diversity embedded in each alignment (Fig. 2*E*).

Raw z Score Value and Amino Acid Selections Residue by Residue. To simplify the analysis of amino acid selection distribution, the raw z scores at every residue as calculated for the experimental data are presented (Table 8). The calculation was done according to *Methods* (Eq. **2**), and these values were used for constructing Fig. 3. In Table 9, the z score of significant amino acid selections distribution observed in the experimental data set and in the three natural sequence alignments are presented.

As described in the article, the comparison of significant amino acid selections in the experimental data versus either the functional homologues alignment or the structural analogue alignments was used to establish a signature sequence for the Raf RBD (*Methods* and Fig. 3, see above). The residues with significant amino acid selections are listed in Table 9. These positions are classed into eight categories indicated at the bottom of Table 9. Outer core refers to residues that are located on an exterior layer around the inner core residues (Fig. 4). Some of the outer core residues, specifically R100, L112, and L121, are also part of a relatively independent mini core located in carboxyl terminus of the domain between β-strand 3 and 5. This mini core also includes in the Raf RBD residues E104, T116, and E124. Note that some residues mentioned in Table 8 were omitted in the article, because, for example, it was difficult to correlate the amino acid selection observed to any structural observation (e.g., residue M83) or because the structural database was probably an artifact (e.g., the methionine translation initiation codon aligned at residue N56).

Determination of the K_d of Raf RBD for *h-ras*. The *in vitro* binding of Raf RBD variants to *h-ras* was tested (*Methods*). The goal of these experiments was to show that the Raf RBD mutants isolated experimentally retain close to wild-type K_d (Table 11).

The K_d obtained by this procedure for the wt Raf RBD is consistent with ref. 17. Several clones were tested in regions known to be involved in binding to *h-ras*, e.g., libraries 2, 3, and 6. As one could predict, the clones from these regions formed among the most destabilized complexes.

Kinetic and Thermodynamic Parameters Obtained for Characterized Variants of the Raf RBD. The Raf RBD variant kinetic and thermodynamic parameters, which folding rate in water was displayed in Fig. 5*B Inset*, are listed in detail in Table 12.

Note that all termodynamic parameters were determined from kinetic experiments.

Hierarchy in the core is not fully apparent from the B factors and solvent accessibility values. The solvent accessibility and B factors were plotted for all residues of the Raf RBD based on one NMR and one crystallographic structure (1RFA and 1GUA, respectively), putting emphasis on the inner and outer core residues (Fig. 8). The average value and standard deviation are listed in Table 10.

Raf wt and Variants Display Similar 1D ^{1}H and NOESY Spectra

We performed ^{1}H and NOESY NMR experiments on Raf RBD wt and five mutants corresponding to positive clones selected from libraries of mutants to different regions of the structure. As probes of tertiary structure, we focused on the Trp-114 NεH proton (9.8 ppm for wt) and the two extreme upfield peaks correspond to the Cγ1H_3 (0.18 ppm) and Cγ2H_3 (-0.4 ppm) methyl protons of Val-98, which are ring-current shifted by interaction with the Trp-114 side chain (chemical shifts are from ref. 9). Fig. 9*A* shows spectra in amide region for the wt Raf RBD versus 5 clones (S1, A1; S2, H2; S3, A3; S7, C7; S12, H12). The sequence of clones and the corresponding segment of the wt are shown on the left of the spectra. Relative positions of the Trp-114 NεH (9.8 ppm for wt) are indicated with a broken red line. As can be seen, there are only moderate differences in the chemical shift of this proton for the clones, relative to that of the wt and consistent with maintenance of the integrity of the raf RBD structure. The amide envelopes themselves show minor variations,and one peak at 6.3 ppm is missing for clone H12. This peak could be that of one of the Gln-127 amine protons; this residue is substituted for a Leu in H12. Unfortunately these peaks have not been assigned. There were some interesting differences in the extreme upfield regions (Fig. 9*B*), including the appearance of a new peak, which overlaps the Cγ1H_3 (0.18 ppm) methyl protons of Val-98 in the spectrum of H2 and the appearance of two new peaks in the spectrum of C7. The increase in the three-bond coupling constants for Val-98 Cγ1H_3 and Cγ2H_3 could be due to reorientation of these methyl groups from cis (±60°) to trans (±120°) relative to the CβH proton from which this coupling arises. To confirm that the small changes in chemical shifts observed in the 1D spectra do not

represent substantial structural changes and to resolve any ambiguities about the chemical shifts in the upfield region, we performed proton NOESY experiments. These results confirm that NOESY crosspeaks for Trp-114 NεH are identical in each clone compared to the wt (Fig. 9*C*; broken red lines) and for the Cγ1H_3 and Cγ2H_3 methyl protons of Val-98 (Fig. 9*D*; broken red lines). This observation is true also of the spectrum for clone H2 in which another peak overlaps with the peak for Cγ1H_3 of Val-98. Details of crosspeak assignments are provided in Fig. 9 *C* and *D* legend.

GST-*ras* Pull-Down by Ni-NTA Bound 6×His-Raf wt or Clones Is Competed by Untagged Raf wt. We sought to demonstrate that the binding of clones isolated in our screening strategy had retained the same mechanism and binding surface to *h-ras* than the wt Raf RBD. To do so, we tested the ability of untagged Raf wt to compete the pull-down of GST-*ras* bound to GMP-PNP on Ni-NTA bound 6×His-Raf wt or clones (see above, Table 11 and Fig. 10). This experiment (Fig. 10) shows that the complex formed by the seven clones tested with GST-*ras* can be competed by untagged wt Raf, thus arguing in favor of conservation of the binding interface. Moreover, the lower level of GST-*ras* recovered in the pull-down for clones H2 and A3 versus Raf wt is coherent with their lower dissociation constants determined through a fluoresecent *in vitro* binding assay, as described above (Table 11).

References for Supporting Text

1. Pelletier, J. N., Campbell-Valois, F.-X. & Michnick, S. W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 12141–12146.
2. Campbell-Valois, F.-X. & Michnick, S. W. (2005) *Methods and Protocols in Molecular Biology*, in press.
3. Sidhu, S. S., Lowman, H. B., Cunningham, B. C. & Wells, J. A. (2000) *Methods Enzymol.* **328,** 333–363.
4. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5857–5864.
5. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
6. Holm, L. & Sander, C. (1994) *Nucleic Acids Res.* **22,** 3600–3609.
7. Vallee-Belisle, A., Turcotte, J. F. & Michnick, S. W. (2004) *Biochemistry* **43,** 8447–8458.
8. Manor, D. (2000) *Methods Enzymol.* **325,** 139–149.
9. Emerson, S. D., Waugh, D. S., Scheffler, J. E., Tsao, K. L., Prinzo, K. M. & Fry, D. C. (1994) *Biochemistry* **33,** 7745–7752.
10. Emerson, S. D., Madison, V. S., Palermo, R. E., Waugh, D. S., Scheffler, J. E., Tsao, K. L., Kiefer, S. E., Liu, S. P. & Fry, D. C. (1995) *Biochemistry* **34,** 6911–6918.
11. Janin, J. & Chothia, C. (1980) *J. Mol. Biol.* **143,** 95–128.
12. Richards, F. M. (1974) *J. Mol. Biol.* **82,** 1–14.

13. Lim, W. A. & Sauer, R. T. (1989) *Nature* **339,** 31–36.
14. Axe, D. D., Foster, N. W. & Fersht, A. R. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 5590–5594.
15. Liang, J. & Dill, K. A. (2001) *Biophys. J.* **81,** 751–766.
16. Pelletier, J. N., Arndt, K. M., Pluckthun, A. & Michnick, S. W. (1999) *Nat. Biotechnol.* **17,** 683–690.
17. Block, C., Janknecht, R., Herrmann, C., Nassar, N. & Wittinghofer, A. (1996) *Nat. Struct. Biol.* **3,** 244–251.
18. Fridman, M., Maruta, H., Gonez, J., Walker, F., Treutlein, H., Zeng, J. & Burgess, A. (2000) *J. Biol. Chem.* **275,** 30363–30371.

Tables for Supporting Information (starting on next page)

Table 2. Sequence of clones selected through our experimental strategy and used for statistics Eqs. 1 and 2

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
SNTIRVF	**LPNKQ***	**RTVVN**	**VRNGMS**	**LHDCLM**	**K*ALKVR***	**GLQPEC**	**CAVFRL**	**LHEHKGK**	**KARLDW***	**NTDAAS**	**LIGE**	**ELQVDF**
VTLVKVF	CPDGQ*	ASIKH	ARRGKS	VETHLH†	K*SMEHR*	SVPQNA	RLTVSR‡	SLDCFLT	VMNMPW*	GMCAAK	LCRE	DIQLEV
AEFFTVF	IIDGQ*†	RTMLM	IQWGTS‡	VENMLK	K*RMELR*	SFMQGH	TLTVWR	TLLCRVT	HDNLPW*	AMAAAF	STVE	HIHLEI
TNLMAVH	LQDGQ*	NSTMH	HKIGTG	VEKALR	K*SMVRR*	ALNQNE	KLGVRA	TVGCCSV	VAILDW*‡‖	SLDAGD	LTSE	HLFLEL‡
PMTMTVH	LADGQ*	WSIQR‡‖	PSKGAM	VERALQ	K*SMVLR*	EMLLND	KLVIAR‡§‖	DCIKRTN	MARLDW*	NRDAGK‡	LTGG	ELYTEI
MSWMPVY	IGQFQ*	SSYMR	PKKGLM	VKTTLF	K*SMGKR*	RVCLNE	MDRPWL	HTILRPT	TNRVDW*	FSPARD	LRGL	AIKIEF
RSLSCVH	IGLAQ*	QSTAV	PKCGLT	VSMILE	K*SMGLR*	KMCANR	MTRPVL	STHEGPV	RSALGW*	HSMAKD	LSGL	LIAVES
RGLMCIW	IGECQ*	RSTAS	VDRPRD	VRMILA	K*TMGHR*	KMDFVT	MKKPRL	ADTLLVY	MSALTW*	ECSARD	LRGK	KLVISP
RRLCFVM	IGSSQ*	KSTVV	MDRWAL	VHMVLL	K*TMYHR*	DCKAEQ	IDRLRL	TPTLHSD	NSRLLW*	ARVARG	LGSS	KLVIDY
ATCLFID	IGTSQ*	RSTVW	PSRPTS	ILDLVM‡	K*TMAMR*	NCPSEC	MSRIRL	SCTSKRD	GIRLQW*	HYSASV†	LPQS	GLVIEP
LMCMFVD	ILESQ*†	ISSVT	ANLEMT	ISDLMI	K*PMVSR*	QCDRET	LRGIRL	SHTYFRP	SRDLVW*	EMQASV	LRCS	SIVLQP
LLYLVVH‡	ISGDQ*	MSVVF	AMLGLT	IQQAVD	K*PMKQR*	SLSKEY	LICLRL	PDQSATG‡‖	LRLDVW*	NAFAST	LRHC	NLTIVS
LVLLEVL	ILADQ*	HSYCY	ARLSET	IRIHLN	K*PMDTR*	SYIPEY	GLTATL	EDQERIG	TRCLSW*	CVSNSA†	LSVN	NITYWQ
AMSLCCL	IQRTQ*	ISCCS	AGPVHT	IVDHLR	K*KMFSR*	SMSSES‡§¶	SLEARL	GPGQLTN	CRCLDW*	DVSCST	LADN	SLWCKQ
LTHLKCL	LGVHQ*	YALIA‡	CVLTDT	ISICLS	K*KMVTR*	GTSSQA	NMVIRI	TYGQGMG	PPLLMW*	QQDASA	LGVT	RIYVKE‡
LPKLNVK‡	LGAEQ*	YSRII	CGNSGT	LRWNLS	K*VMWRR*	GEDRQA	NMVLRK	MEPHIMT	PGWLEW*	NCDSSA	LGVL	RISVCT
TPRLKVG	LGNAQ*	LSTIA	GQCAMT	LKDNLS	K*AMCRR*	RVDSQQ	NMKVFM	GESHSTG	IPELYW*	PGCRYV	LNPQ	SIEVLE
CMRLYVK	LGHAQ*	MIALD	ACCCMT	LGVELS‡	K*KLSER*‡§¶	WLDRQF	MKVRRR	PEGSGSN	LDELHW*	PGWCME	ITPT	HIIVSH
SSRLLVL	VGHYQ*	HCRLS	SPRRET	LVKNLQ	K*KLTSR*	RISWRY	MKVQRS	PTKSKAE	LMMLNW*	GGIRSH	ISQT	HIVVHS
TVPLNVS	LGSSQ*	FSWLG	VPLPQT	LSKNLE	K*KLFMR*	RIGIGD	MKVLQM	PYRESSE	WNQLGW*	KLQRPA	IRGT‡	KICVTC
MMPLWVY	LGSTQ*	QCCAQ	VNEVVR	LYDCLM	K*KLEMR*	NIACVS	MVVERV	PYEVMSF	WQSLKW*	GQDDFK	KTTT	MVLVYE
SFNLAVK‡§¶‖	LPRLQ*‡§¶‖	QINPM	VNTPVG	LYICLI	K*KLCRR*	RIPRAS	MVVCRW	PQPNSVD	RNLLAW*	GSTPEK	SQTT	DGLVQE‡
TFMTRVW	LPHLQ*	DGNVY	VAKPLP	LQRCLQ	K*TLIRR*	DRMPSD	MAVWGG	GWLCMRR	RCALAW*	GMDGSR	LPTT	KVFIVE

Table 2 (continued).

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
SNTIRVF	**LPNKQ***	**RTVVN**	**VRNGMS**	**LHDCLM**	**K*ALKVR***	**GLQPEC**	**CAVFRL**	**LHEHKGK**	**KARLDW***	**NTDAAS**	**LIGE**	**ELQVDF**
RWALAVA	LPTMQ*	LTTVN	AYGKRP	LKRCLE	K*TLEKR*	TIIPTD	LRVNRT	GEGENRR	RVGLAW*	QMDVSQ	MPTG	TVDVVN
LWPLPVA	LPGHQ*‡§¶‖	GTRVT‡§‖	VYGKWP	LCSSLD	K*GLQKR*	DIRRKD	MGVNRT	ANAMPRE	RHGLTW*	GADGRE	RMLD	KVVVNT
PWVDLDA‡§¶	LPFVQ*	WTMVR	VKWKRE	LKQSLL‡	K*ALSKR*	VLTVQD†	ISVWRT	HNRRHEA	HVTRAW*	GHDVRT	MDLT†	EVAVDT
KSRVAVA	LPLVQ*	CTQSR	VPASSL	LYSTVL	K*RLLMR*	HLREQD	VEVFML	MNWESEG	HVKIAW*	HDDPND‡	NART	CITVDN
QSGIAVA	LPLTQ*	LTQMR	VFASQR	LYTTVY	K*RLNIR*	KLMGAD	TEVFRR	ARGRRLA	RVSVAW*	TFDTSD	LTRT	SLSVDN
GSTVIVA	LPMGQ*	LTSAL	VPAQIP	LMATLL	K*RLWNR*	MLAYAD	TSALQI	ASRARLQ	RIPTLW*	GSPCSQ	LSCT	RLMVSD‡§¶
TSCLRVA	LPSWQ*	LTSKC	VQAAES	LATTLH	K*CLFNR*	KLQRGD	TSITRI	RNGGRAD	RISTEW*	ASPMSN	VSRY	RLELLD
RDWFPNN	LPSAQ*	LTSSS	VACASS	LQTTLQ	K*RLASR*	SLQRAN	ASITLV	LNGLSKA	RQSTYW*	HSRCES	VGGS	SLAVHD
RYWVVAQ	LPSSQ*	LTIVS	VLHHPV	LRTLIG	K*RLALR*	KLSASH	RSDLIV	LIGSESA	RVRQYW*	ACEGSS	VGGG	SLKVVH
RMTVAIN	LPSNQ*	LTETI	VLGAIE	LRTLMK	K*VLAFR*	DLAVSH	MSDLRV	IDNFQSA	RMTATW*	NSEGYA	MGGS	NLWVRS
CRSLLAN‡	LPHNQ*†	WTQQL	VMRSIQ	LRTVLL	K*ALCYR*	SLSGST	IEQLTV	RSSVIPA	RMMAQW*	KVDSTR	ERGS†	YLRVDS
KMLLLAW	LPENQ*	ASTQL	TARAIA	LLGLLK	K*RLHYR*	TLKASD	LTTTRV	DFKVAFA	RMAALW*	DWDSVS	QDGF	QLYVSS
LPCLIAR	LTDNQ*	KTTQL	CGSNLR	LKRVCS	K*RLVHR*‡	ALRHSR	TTVYRV	NSTTNTA	EVTLPW*	MTPSTS	NVGW	RLCVKS
SRKLMAR	LTDHQ*	KQPKL‡	IGSRGM	LLDVSR	K*RLVSR*	DLGHDG	MTVKFK	EDGDGVP‡	MVTLPW*	DRESSE	VQDH	SLVVTS
RELLAAS	LPDHQ*	VTTKL†	VSSRAR	LNHILG	K*KLVWR*‡§¶	NLDKLS	MTIRSK	EILPGSI	IVTPGW*	CAPSWE	IQGH	TLTVTA
AAVVSAV	LPDNQ*	VTIGL	VGKAAR	INNILE	K*SLVTR*	NLDLKE	MFAIKA	IRWRGRV	EVLAEW*	DAPWSS	GQGC	SLTVRA
GCYVYAS	MPETQ*	VTTRI	TLHLRR	LQNIIV†	K*SLGSR*	NMFFKG	MEARMC	SRARGKA	LPATPW*	DQPIQF	HMGH	SLTVEV
HSATVAH	MPECQ*	STIMG	DRSLTR	LYEIIN	K*SLGNR*	NMKLKN	MVPRAA	NRKRPMP	LVAVTW*	SCPGQD	HLGA	TLSVLL
WSAVSFE	LPECQ*	KTIRM	LHGRKH	LRDELA	K*SLWSR*	NLQSTQ	MTELVA	RFKRGSK	ATQTDW*	QTDTEG	YLGL	KLVVCL
FAIIKFQ	LLETQ*	MTMFV	LGRHYH	VRDGLA	K*SLKNR*	NYRACH	MYIQRC	SIEGLNK	WTPADW*	NCDPEG	MRQR	KLSVEA
FAPITAI	LPESQ*	HTYRV	PAPGIV	LRKILA	K*SLKMR*	AASVEQ	MEIHRS	MSVRSNY	LTHHPW*	RHEMEG	MRGH	DVSVRM
FQALTAM	LPDRQ*	CTLRH	ASPGRV	LRDIMV‡	K*SCKIR*	AITVPQ	MYLTRR	SSVRINK	STHTGW*	YCNVDD	SRYI	DRMVRV

Table 2 (continued).

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
SNTIRVF	**LPNKQ***	**RTVVN**	**VRNGMS**	**LHDCLM**	**K*ALKVR***	**GLQPEC**	**CAVFRL**	**LHEHKGK**	**KARLDW***	**NTDAAS**	**LIGE**	**ELQVDF**
SCCFTAM	LPYRQ*	RTLQS	ARPRIL	LRDILV	K*SLMIR*	GWSAPQ	LGLPTR	TSRPSVS	MAAIKW*‡	DCNVDM	CLSS	DLKVQM
TRLIRIS	LPERQ*	YTLTC	AWSPIK	LREILR	K*SVMRR*	GVQGPD	LKLPFQ	HSRCTDK	IASIAW*	DSNSED	CLQS	CLEITM
YRPIVAW	LANNQ*	AVLNR	APILGS	LEYILR	K*SALRR*	GWRRGR	LGLHSV	RPRIKFV	DDAVAW*	RTNSEA	CQES	LCEVTK
RSVITVS	LCNGQ*	ATTLK	PMSSSS	LRASLR	K*SLSRR*	GSGVGR	LGLVVE	SPRSDSQ	WAAVDW*	TISCCT	CLTE	DMFVDA
GQVITCN	MPNSQ*	YTVLR	FPPESS	VRGCME	K*SLRHR*	GYPWGG	LVLTMG	NQYPRGV	IAGVSW*	RINCDT	CLGE	TLMLTR
AMNISVS	LPNSQ*	ITARQ	LLPGVT	LRKMME	K*SLRRR*	GTHVNK	LVLASM	EKRPKGA	QAVVNW*	RTDCIT	CCQL	TSCHDL‡§¶
DDMIACV	CPNMQ*	ITCTM	ILPSVS	LVQIME	K*SLHVR*	QRVVMV	LTPPEM	AGYNEPL	RAYKRW*	TTDCTP	CVSF	RSNPQL
MTALEFK	LPNDQ*	ATVTQ	PLANVP	LEVLMS	K*SLLVR*	LGVPGY	LAKPET	AETSETL	SATLRW*	SMSTRT	CAST	PSHNTA†
DTVLFFR	LPNFQ*	FSFRR	RSATAA	LSVSMG	K*HLMWR*	GYVLTA†	LGCYGP	GVNLESC	SISITW*	SMTCDD	AVNS	PVNMMS
RTNLTFQ	LPKEQ*	FTMRR	RWASMS	LGEAMC	K*GLRLR*	GFVSKS	LTPYIN	KVMLELL	SLEITW*	GLKGTS	IVNN	TFNGIP
GWMLMFD	LPWEQ*	FAVRM	EFVTLL	VAQLCQ	K*CCRLR*	GMDTNE	LEILPL‡	NKGNEFS	SFETEW*	TMKATP	CRNQ	SKFWMV
TGMLPFS	LPNEQ*	FMVTS	TTWTCY	VAESLS	K*PLGLR*	GLLTQE	LAIALL	GCYGTKS	SWLVGW*	STATTD	VNCA	IMQITV
HHQLEFA	MPDQQ*	VAQSI‡	KTTPPL†	AAEMLT	K*ACVQR*	GLALRE‡	LKINFR	YGTRVGS	SCMVRW*	GTAATT	TGAA	LLYIEI
NSQLEIF	IPFQQ*	VSQTF	FTENMK‡	LAETMT	K*ALQQR*‡‖	GLRLAS	LTIWFG	EGSKVDM	GPSASW*	KATATS	HRDA	PNRMIV
ANYLTTR	FNSSQ*	NAQVV	NIMSRR	LEAQLS	K*AVYRR*	GVESDL	LMATLQ	ERATYDS	IGSASW*	QNTATS	LVDA	MLDVRQ
AATLNIR	FNDGQ*	SSQVV	EIEDRA	VMGCLL	K*AVVSR*	GVLVDC	LSAAIK	EVDVLDE	GLSLSW*	KQCATL	LMDR	RLTCLS
PCTLSCS	FGDAQ*	RHVST	SWYQRS	LFNSIH	K*AFRSR*	GLERDS	LGITYK	ELEKHPQ	NLSLSW*	KTTAKE	QTDC	RLEITQ
NRTLICY		VHTTT	KPPNRC	LKDTVH	K*AMRVR*	QLDMHA	MVVYSR	EATKLML	GYTLSW*	FETVTG	VGDC	TEISVE
PPTLRIM		RSMLR	GRIARK†	LSRMVE	K*AAKLR*	GLTGHA	LVVFSR	NLSPLCR	LGRVSW*	SVTAME	RIPG	RLLVNN
VCRCRIK		RKMQA	QRWYAC	LQVLMC	K*ALKVR*	GLTAVT	LYVRSR	MLADCCS	LITVSW*	SVQAMS	ICGH	ACDVEQ
GCSLAIA		VTERT	HPLPSC	VSATLA	K*SFLTR*	GLQPEC	LRVRYF	QLLADIQ	KYPVSW*	SVEAEH	YAGT	
		MSAIP	ISTPCK	LINIMV	K*KFLLR*	ELNHRE‡	LRGVYM	QAWVDVS	VLMVSW*	SDGASH	QDGF	

Table 2 (continued).

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
		RTVVN	**VRNGMS**	**LHDCLM**	**K*ALKVR***	**GLQPEC**	**CAVFRL**	**LHEHKGK**	**KARLDW***	**NTDAAS**	**LIGE**	
		RHVMQ	CSPSGR	LEESIM	K*KFACR*	SCGIED	LRVARN	QSNKCET	HLVQSW*	AKGNMC	QRGW	
		STRLS	ALTEGN	LVALLR	K*VFSVR*		LRVKRK	QMNPGAL	TASMPW*	QLPAVR	LLNN	
		RSIMQ	ALLGHK	LEENLH	K*VFKHR*		LTVRNE	LYPIAES	TASPCW*	HQPANR	LLPA	
		RTLLP	AGPEMC	LQDFIG	K*QFVNR*		LTVRCS	WANQATF	TRSVQW*	EKPAER		
			FQPEQT	VINCVK	K*PFLNR*		LDVQME	GEQYPNL	TRSFSW*	RMPAGS		
			PGPRLS	VKISLM	K*PILVR*		LMVSKA	RAECKGA	NGESQW*[†]	SRNAAS		
			VSPRSS		K*PIRSR*		LAISRR	ELDIWEK	NEPMQW*			
			LKPYNT		K*PIRMR*		TAVSVR	IDKSEIV	KLSEQW*			
			IKPLNS		K*HINCR*		LSIGAR		KLEAQW*			
			IPPKSK		K*CINHR*		MSISAR		TLESGW*			
			ARRMQT		K*SINLR*		TSIRYR		TLLLGW*			
			VSVAQT		K*GIMLR*		LSLSYR		VESVQW*			
			VRKNER		K*KIFVR*		TCLNHI		HESVLW*			
			VRHPET		K*KIWGR*		TCIYVE[†]		EEKVVW*			
			ATPRST		K*RIGFR*		TWIHSM		MEKLHW*			
			ACVRST				TCVSLD		GESYRW*			
			CTCTST				TYVTLK					
			VAEYST				DKLGIW[‡§]					
			LLPAVR				ADLMIW					
			ALAALT				TRMGLW					
			VLSRVT				TRVGWR					

Table 2 (continued).

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
			VRNGMS				**CAVFRL**					
			GFSGVT				CKVGWG					
			PSARVT				GIMIWH[†]					
			IVQGGT				SRYTVS					
			IVKRYT				MLYGLY					
			IRPGLT									
			CHSKLS									
			CCGTLS									
			TTTALS									
			IHRAES									
			ARAPRS									
			ASGPGS									
			ATGVSS									
			PTRSSS									
			PASSYS									
			PDNQQS									
			ISSPVS									
			INSSMP									
			VCTETS									
			VTTGKK									
			VHMSIR									
			VHEATG									

Table 2 (continued).

S4
VRNGMS
VLLSRP
VLSSRV
VLARAP
VEPLKP
VMPLAS
VPRTCA
VPRPMN
PHRNCN
EHLTPK

*Residues not degenerated in the experiments.

†Clones that were not purified with sufficient yield.

‡Clones that were successfully purified.

§Clones for which CD spectra were done. Some of the spectra were presented (Fig. 5*A*).

¶Clones for which chevron curves are presented in Fig. 5*B*.

||Clone for which affinity curves were done (Table 11).

Table 3. Sequences of clones isolated through screening of three alternative libraries

S2b†	S6b‡	S8b§
FNGSG	DRLANR	LSIRR*R
FSGQG	KKLVVR	MSIRR*V
FSDLG	HSLRTR	MLISR*P
FYQLG	RPISMR	LVLKR*K
FPDYA	APMSHR	LVLFR*Y
MPHEQ	MPMTRR	LTHPR*Y
LPMEQ	RAMELR	LRVER*R
LPQHQ	STIKRR	TFVKR*K
LNGGQ		TWISR*R
LGSTG		TMLSR*R
LGSTM		TYIMR*T
LGTTQ		CIIHR*K
LADST		MWVTR*L
		KLQIR*L
		VMIYR*L

†S2b is identical to the main experimental library S2 except that Q66 is also degenerated (Table 1 and 2).

‡S6b is identical to the main experimental library S6 except that K84 and R89 are also degenerated (Table 1 and 2).

§S8b is identical to the main experimental library S8 except that R100 is kept constant (Table 1 and 2).

Table 4. Alignments of functional homologues (MSA S1) and structural analogues (MSA S2 and S3; note that wt Raf RBD sequence, 1RFA, is highlighted with bold lettering)

MSA S1. Functional homologues of Raf type RBD (SMART)

```
smart|RBD-sptrembl|Q8IN00|Q8IN      SLCRVIL-TDGATTIVQTRPGETVGELVERLLEKRNLVYPYYDIVF--QG---STK-SIDVQQPSQILAGKEVVIERR
smart|RBD-ENSANGP00000001206/3      TLCRVIL-SNGATTVVQTRSNETIKELVERLLEKRGIVYNAYEAFL--AG---STK-PLDLDGPSVSLAGKEVNIDQR
smart|RBD-ENSMUSP00000021945/2      KYCCVYL-PDGTASLALARPGLTIRDMLAGICEKRGLSLPDIKVYL--VG---NEQKALVLDQDCTVLADQEVRLENR
smart|RBD-ENSMUSP00000030984/3      KHCCVHL-PDGTSCVVAVKSGFSIKEILSGLCERHGINGAAVDLFL--VG---GDK-PLVLHQDSSILATRDLRLEKR
smart|RBD-SINFRUP00000158915/2      RQCRVML-PEG-SCSSILRPGSFIREVLQDLCQSIGVNIAAVDLFL--VG---GEK-PLVLDQDCMTLCSRDLRLEKR
smart|RBD-sptrembl|Q13878|Q138      PIVRVFL-PNKQRTVVPARCGVTVRDSLKKALMMRGLIPECCAVYR-I------------------------------
smart|RBD-sptremblnew|BAB32131      GTVKVYL-PNKQRTVVTVREGMSVYDSLDKALKVRGLNQDCCVVYR-LIK---GRKTVTAWDTAIAPLDGEELIVEVL
smart|RBD-sptrembl|Q98TC3|Q98T      STIRVYL-PNQQRTVVNVRPGMTLHNCLIKALKVRGLQPQCCAVFR-LHPGQRSKKLRMDWNTDSTSLIGQELLVEVL
smart|RBD-swissprot|P09560|KRA      STMRVYL-PNKQRTVVNVRSGMSLHDCLMKSLKVRGLQPECCAVFR-LIQDPKGK-LRLDWNTDAMSLVGAELQVDFL
smart|RBD-sptrembl|Q8I086|Q8I0      ILLRAHL-PNQQRTSVEVISGVRLCDALMKALKLRQLTPDMCEVST-THSG----RHIIPWHTDIGTLHVEEIFVRLL
smart|RBD-ENSANGP00000003633/1      MLLRAFL-PNQQRTSVQVIPGMRLKDALAKALKRRNLTCEFCEVTA-GNS-----NYPIPWETDVSALNCDEVFVRIL
smart|RBD-sptrembl|Q8MXT8|Q8MX      KMIMVHL-PFDQHSRVEVRPGETARDAISKLLKKRNITPQLCHVNASSDPKQESIELSLTMEEIASRLPGNELWVHSE
smart|RBD-sptrembl|Q9GT28|Q9GT      SLILLHL-PFNQHSKVEEKPGILARDAIAKILEKRAILPQMCRVCVGGDPSSPRTDLSMDLETLSGALEKKELWVHSA
smart|RBD-ENSANGP00000002581/1      KSYKVAL-PENTFATVYLREGMSVEEFLASACSRKNLNPMEHFVRVKKRRDM--EDHNYFVPHR------NDL-IETY
smart|RBD-ENSMUSP00000002588/7      TPSWFCL-PNNQPALTVVRPGDTARDTLELICKTHQLDHSAHYLRLKFLME---NRVQFYIPQP-----EEDI-YELL
smart|RBD-SINFRUP00000133952/7      TPSWVCL-PNDQPVLTIIKPGESALCVLESICKAHYLDPTRHYLRLKFLME---SQVKIYIPKP-----DEDV-CDLV
smart|RBD-ENSMUSP00000024562/8      VQTYVHF-QDNEGVTVTIKPEHRVEDVLALVCKMRQLEPTHYGLQLRKVVD---KSVEWCVPALYEYM-QEQASYDEI
smart|RBD-ENSP00000275245/186-      IQTYVHF-QDNHGVTVGIKPEHRVEDILTLACKMRQLEPSHYGLQLRKLVD---DNVEYCIPAPYEYM-QQQV-YDEI
smart|RBD-sptrembl|Q8IN00|Q8#I      VAFKLDL-PDPKVISVKSKPKKQLHEVIRPILSKYNYKMEQVQVIM--RDT----QVPIDLNQPVTMADGQRLRIVMV
smart|RBD-ENSANGP00000001206#/      VVFKLNL-PNRKMISVKSKAAKPLADVLRPILHKYNYELDEMKVVV--HSTV---DVCLDMTQPVTTVDGLCLYIRSA
smart|RBD-ENSMUSP00000021945#/      -TFQLELVGLERVVRISAKPTKRLQEALQPILAKHGLSLDQVVLHR--PGE----KQPMDLENPVSSVASQTLVLDTP
smart|RBD-ENSMUSP00000030984#/      -LFRLDLVPINRSVGLKAKPTKPVTEVLRPVVAKYGLDLGSLLVRL--SGE----KEPLDLGAPISSLDGQRVILEER
```

Table 4 (continued).

MSA S2. Structural analogues classified in the ubiquitin superfamily and 4 closely related super-families (SCOP)	
1RFA	**NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLKARLDWNTDAASLIGEELQVDF**
1UBI	MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYKESTLHLVL
1A5R	IKLKVIGQDSSEIHFKVK--MHLKKLKESYCQRQGVPMNSLRFLFEGQRIADNHT-PKEEEDVIEVYQ
1MG8a	MIVFVRFNSSYGFPVEVDSDTSILQLKEVVAKRQGVPADQLRVIFAGKELPNHLTVQNCQQSIVHIVQ
1VCBa	VFLMIR-RHKTTIFTDAKESSTVFELKRIVEGILKRPPDEQRLYKDDQLLDDGKTLGECAPATVGLAF
1M94a	IEVVVNDRLGKKVRVKCLAEDSVGDFKKVLSLQIGTQPNKIVLQKGGSVLKDHISLEDYDQTNLELYY
1J8Ca	IKVTVKTP-KEKEEFAVPENSSVQQFKEAISKRFKSQTDQLVLIFAGKILKDQDTLIQHDGLTVHLVI
1H8Ca	SKLRIRTPSGEFLERRFLASNKLQIVFDFVASK-GFPWDEYKLLSTRRDTDPNKSLLEVPQETLFLEA
1JRUa	TNIQIRLADGGRLVQKFNHSHRISDIRLFIVDA-AM--AATSFVLM-KELADQ-TLKEAAVIVQRLT-
1EO6a	VPVIVEKV-----KYLVPSDITVAQFMWIIRKRIQLPSEKIFLFVDKTVPQSSLTMGQLGFLYVAYSG
1EF1a	ISVRVTTD--AELEFAIQPNTTGKQLFDQVVKTIGLEVWFFGLQYQSTWLKLNKKVTAQSPLLFKFRA
1GG3a	MHCKVSLLDDTVYECVVEKHAKQDLLKRVCEHLLNLDYFGLAIWDNKTWLDSAKEIKKQ-PWNFTFNV
1LFDa	CIIRVSLDV--YKSILVTSQDKAPTVIRKAMDKHNLEPEDYELLQIKLKIPENANVFYA-NYDFILKK
1E8Xa	VFIVIHRS-TTSQTIKVSADDTPGTILQSFFTKMA---RDFVLRVRDEYLVGETPIKNF-EIHLVLDT
1K8Rb	CILRFIACNGQTRAVQSRG--DYQKTLAIALKKFSLDASKFIVCVSIKLITEE-------D-RLIIVP
1L7Ya	VTFKITLSD-PFKVLSVPESTPFTAVLKFAAEEFKVPAATSAIITNGVGVNPAQPAGNIFGSELRLIP
1D4Ba	RPFRVCDHKRIRKGLTAA---QLAKALETL-----L-NGVLTLVLE-TAVDSEDFFQLL-DTCLMVLQ
1C9Fa	KCVKLRALHSCKFGVAAR---SLLRKGCVRF---QLPMPGSRLCLDGTEVTDDFP---LNDAELLLLT
1F2Ri	KPCLLRRNHSDQHGVAAS---SLRSKACELL---AI----ITLVLAGTIVDDDY-FLCLSNTKFVALA
1IP9a	TKIKFYYK-DDIFALMLKGDTTLRSKIAPRI---DTD---FKLQTKSEEIKTDQ-VSNIIALKISVHD
1Q1Oa	ILFRISY---EIFTLLVEKVWNLIMAINSKISNT------IKIKYQFVVLGSDD-WNVAKFLNIRLY-
1FMAd	IMIKVLFF---ATEVAADFP-TVEALRQHMAAQSDLEDGKLLAAVNQTLVSFDHPL---DGDEVAFFP
1F0Za	MQILF--N-D--QAMQCAAGQTVHELLEQL----DQRQAGAALAINQQIVQWAHIV---DGDQILLFQ
1JSBa	MKFTVITD-DGKKILESGAPRRIKDVLGEL----EIPIETVVVKKNGQIVIDEEEI---DGDIIEVIR
1QF6a	PVITL--P-D-GSQRHYDHAVSPMDVALDI----PL-----AGRVNGELVDACDL---INDAQLSIIT
1JALa	QTYFTAGV-KEVRAWTVSVGATAPKAAAVIH---TF---I-RAEVIWRLEGKDYIV---DGDVMHFR-
1MG4	KKVRFYRNG-FGIVYAISPDFSFEALLADLTRTLNLP----TIYTILKKIS---SLDQLEGESYVCGS

Table 4 (continued).

MSA S3. Alignment of structural analogues (SCOP-FSSP)	
1RFA	**NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLKARLDWNTDAASLIGEELQVDF**
1UBI	MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYKESTLHLVL
1A5R	IKLKVIGQDSSEIHFKVK--MHLKKLKESYCQRQGVPMNSLRFLFEGQRIADNHT-PKEEEDVIEVYQ
1MG8a	MIVFVRFNSSYGFPVEVDSDTSILQLKEVVAKRQGVPADQLRVIFAGKELPNHLTVQNCQQSIVHIVQ
1VCBa	VFLMIR-RHKTTIFTDAKESSTVFELKRIVEGILKRPPDEQRLYKDDQLLDDGKTLGECAPATVGLAF
1M94a	IEVVVNDRLGKKVRVKCLAEDSVGDFKKVLSLQIGTQPNKIVLQKGGSVLKDHISLEDYDQTNLELYY
1J8Ca	IKVTVKTP-KEKEEFAVPENSSVQQFKEAISKRFKSQTDQLVLIFAGKILKDQDTLIQHDGLTVHLVI
1H8Ca	SKLRIRTPSGEFLERRFLASNKLQIVFDFVASK-GFPWDEYKLLSTRRDTDPNKSLLEVPQETLFLEA
1JRUa	TNIQIRLADGGRLVQKFNHSHRISDIRLFIVDA-AM--AATSFVLM-KELADQ-TLKEAAVIVQRLT-
1EO6a	VPVIVEKV-----KYLVPSDITVAQFMWIIRKRIQLPSEKIFLFVDKTVPQSSLTMGQLGFLYVAYSG
1EF1a	ISVRVTTD--AELEFAIQPNTTGKQLFDQVVKTIGLEVWFFGLQYQSTWLKLNKKVTAQSPLLFKFRA
1GG3a	MHCKVSLLDDTVYECVVEKHAKQDLLKRVCEHLLNLDYFGLAIWDNKTWLDSAKEIKKQ-PWNFTFNV
1LFDa	CIIRVSLDV--YKSILVTSQDKAPTVIRKAMDKHNLEPEDYELLQIKLKIPENANVFYA-NYDFILKK
1E8Xa	VFIVIHRS-TTSQTIKVSADDTPGTILQSFFTKMA---RDFVLRVRDEYLVGETPIKNF-EIHLVLDT
1K8Rb	CILRFIACNGQTRAVQSRG--DYQKTLAIALKKFSLDASKFIVCVSIKLITEE-------D-RLIIVP
1L7Ya	VTFKITLSD-PFKVLSVPESTPFTAVLKFAAEEFKVPAATSAIITNGVGVNPAQPAGNIFGSELRLIP
1D4Ba	RPFRVCDHKRIRKGLTAA---QLAKALETL-----L-NGVLTLVLE-TAVDSEDFFQLL-DTCLMVLQ
1C9Fa	KCVKLRALHSCKFGVAAR---SLLRKGCVRF---QLPMPGSRLCLDGTEVTDDFP---LNDAELLLLT
1F2Ri	KPCLLRRNHSDQHGVAAS---SLRSKACELL---AI----ITLVLAGTIVDDDY-FLCLSNTKFVALA
1IP9a	TKIKFYYK-DDIFALMLKGDTTLRSKIAPRI---DTD---FKLQTKSEEIKTDQ-VSNIIALKISVHD
1Q1Oa	ILFRISY---EIFTLLVEKVWNLIMAINSKISNT------IKIKYQFVVLGSDD-WNVAKFLNIRLY-
1FMAd	IMIKVLFF---ATEVAADFP-TVEALRQHMAAQSDLEDGKLLAAVNQTLVSFDHPL---DGDEVAFFP
1F0Za	MQILF--N-D--QAMQCAAGQTVHELLEQL----DQRQAGAALAINQQIVQWAHIV---DGDQILLFQ
1JSBa	MKFTVITD-DGKKILESGAPRRIKDVLGEL----EIPIETVVVKKNGQIVIDEEEI---DGDIIEVIR
1QF6a	PVITL--P-D-GSQRHYDHAVSPMDVALDI----PL-----AGRVNGELVDACDL---INDAQLSIIT
1JALa	QTYFTAGV-KEVRAWTVSVGATAPKAAAVIH---TF---I-RAEVIWRLEGKDYIV---DGDVMHFR-
1FRRa	YKTVLKTPSG-EFTLDVPEGTTILDAAEEA----GY---SCLGKVVEGFVLTIAIP---SDLVIETHK
1I7Ha	PKIVIL-P---GAVLEANSGETILDAALRN----GI---TCHCIVR-ESRLSQARV---EDLVVEIPR
1AYFa	ITVHFINRDGETLTTKGKIGDSLLDVVVQN----NLD--TCHLIFE-RSRLGQICLKAMDNMTVRVP-
1PUT	SKVVYVSHDGTRRQLDVADGVS----LMQAAVSNGI---TCHVYVN-NSRLCQIIMPELDGIVVDVPD
1L5Pa	GTITAVKG-GVKKQLKFEDDQTLFTVLTEA----GL---KCICKHV-NARLAAITLGENDGAVFEL--
1E0Za	PTVEYLNY-----TMEVAEGEYILEAAEAQ----GY---NCASIVKKDVRLTIGSP---DEVKIVYNA
1FEHa	KTIII--N-G--VQFNTDEDTTILKFARDN----NI---ICTVEVE--LVTADTLI---DGMIINTNK
1HLRa	IQKVITVN-GIEQNLFVDAEALLSDVLRQQL---GL---ACSVILDGKVVRAVTKMKRVDGAQITTIE

Table 4 (continued).

MSA S3 (continued)	
1RFA	**NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLKARLDWNTDAASLIGEELQVDF**
1FO4a	DELVFFVN-GKKVVEKADPETTLLAYLRRKL---GL---ACTVMLSHFSANALAPICTLHHVAVTTVE
1JROa	MEIAFLLN-GETRRVRIEDPQSLLELLRAE----GL---ACTVMIRSRAVNALMMLPQIAGKALRTIE
1FFVa	KIITVNVN-GKAQEKAVEPRTLLIHFLREEL---NL---ACTVDIDGRSVKSTHLAVQCDGSEVLTVE
2PIA	TPFTVRLRSGTSFEI--PANRS----ILEVLRDANV----TALC--TQIMVVSRA-----S-ELVLDL
1NENb	MRLEFSIY--QDYTLEADEGMMLLDALIQLKEKDSL---GLNM--NGKNGLAITPISALPGKKIVIRP
1KF6b	KNLKIEV---AFYEVPYDATTSLLDALGYIKDNLDL---GMMV--NNVPKLAKTFLRDYTD-GMKVEA
1QLAb	RMLTIRVF-FQEYKIEEAPSMTIFIVLNMIRTYDDL---GMMI--NGRPSLARTLTKDFEDGVITLLL
1JQ4a	HTITAVTEDGESLRFECRSDED----VITAALRQNI---TCKALCSLVLLCRTYP-----TDEIELPY
1KRHa	HQVALQFEDGVTRFICIAQGETLSDAAYRQ----QI---TCRAFCE-GYVLAQCRP----DAVFQIQA
1C78a	PYLMVNV--PHYVEFPIKPGLTIEYYVEWALDATAY-----SAKIETKSFPIGFVV---PGFNLITKV
1BMLc	SQLVVSV--LKFFEIDLT--LELLKAIQEQLIA--------DATITGKVYFATLPTQP--EFLLS-HV
1MG4	KKVRFYRNG-FGIVYAISPDFSFEALLADLTRTLNLP----TIYTILKKIS---SLDQLEGESYVCGS
1SE2	QNVLIRVN-KISFEVQTDKKVTLDIKARNFLINKNL---TGYIKFITFWY-----S---KSVKIEVHL
1AN8	HKLLGNLSG-QNLNIILEKDVTIDFKIRKYLMDNKI---SGRIEIGHEQI-----M---KNFHFDIYL
1AW7a	IELPLKVHG--KYWPKFDKKLALDFEIRHALTQIGL---GGYWKITTYQS-----I---DEITIEAEI
2IGD	TTYKLVIGK--ETTTKAV---DAEKAFKQYANDNGVD---GVWTYD---------------KTFTVTE
2PTL	VTIKANLANGQTAEFK-G---TATSEAYAYADTLDNG---YTVDVA----------------TLNIKF
1ACC	TTARIIF---VERRIAA---MTLKEALKIAF---GF-ITEFDFNFD----VLDIKL---AKMNILIRD
1G0Sa	HAAVLLP-----VAGMIEEGESVEDVARREAIEAGLI----RTKPV---------------SIMVGE-
1I9Aa	LAFSSWLF----VCGHPQLGESNEDAVIRRCRYEGV-----PPESI--------------CPVFAART

Table 5. Characteristics of the functional homologues included in MSA S1

Gene Id/smart	Organism	Accession no. in GenBank	Protein name	Residues range aligned
Raf-1h RBD*	*H. sapiens*	NP_002871	Raf-1	56-130
sptrembl|Q8IN00|Q8IN	*D. melanogaster*	NP_732773	Unknown/locomotion defect	360-430
ENSANGP00000001206/3	*A. gambiae*	XP_312542	RGS-like	362-432
ENSMUSP00000021945/2	*M. musculus*	NP_058038	RGS-14	303-374
ENSMUSP00000030984/3	*M. musculus*	NP_775578	RGS-12	962-1032
SINFRUP00000158915/2	*F. rubripes*	ND	Q8WX95/RGS-like	290-359
sptrembl|Q13878|Q138	*H. sapiens*	P15056	B-Raf/fragment	154-199
sptremblnew|BAB32131	*M. musculus*	AAH04757	A-Raf	19-79
sptrembl|Q98TC3|Q98T	*S. quinqueradiata*	BAB39747	Raf-1	54-129
swissprot|P09560|KRA	*X. laevi*	TVXLRF	Raf-1	56-130
sptrembl|Q8I086|Q8I0	*D. melanogaster*	AAN17541	Polehole	77-148
ENSANGP00000003633/1	*A. gambiae*	XP_318144	Raf-like	116-186
sptrembl|Q8MXT8|Q8MX	*C. elegans*	NP_741430	Lin-45/Raf-like	127-203
sptrembl|Q9GT28|Q9GT	*B. malayi*	AAG12472	Raf kinase/fragment	52-128
ENSANGP00000002581/1	*A. gambiae*	XP_316614	Tiam-1	1093-1160
ENSMUSP00000002588/7	*M. musculus*	NP_033410	Tiam-1	765-832
SINFRUP00000133952/7	*F. rubripes*	ND	Tiam-1	736-803
ENSMUSP00000024562/8	*M. musculus*	ND	NM_011878/tiam-2	831-903
ENSP00000275245/186-	*H. sapiens*	AAF05900	Tiam-2	186-257
sptrembl|Q8IN00|Q8#I	*D. melanogaster*	ND	CG5248-PC	431-501
ENSANGP00000001206#/	*A. gambiae*	XP_312542	RGS-like	433-504
ENSMUSP00000021945#/	*M. musculus*	NP_058038	RGS-14	376-446
ENSMUSP00000030984#/	*M. musculus*	NP_775578	RGS-12	1034-1104

*Raf RBD was not included in this alignment, because of high similarity with other sequences in the alignment. It is simply indicated to serve as a reference.

Table 6. Characteristics of the structural analogue included in MSA S2 and S3 (SCOP and SCOP-FSSP)

PDB code and chain aligned	Super-family* (SCOP)	Residues range correponding to ubi-like fold†	α-helix length‡	Ω-angle§	α-helix inner core residues¶	Volume of inner core residues sidechain‖	Amino acids at inner core Residues**
1RFA	**1**	**56-130**	**12**	**-50**	**I,I+4,I+8**	**601**	**I,V,L,L,L,V,L,V**
1UBI	1	1-71	12	-60	I,I+3,I+7	619	I,V,I,V,I,L,L,L
1A5R	1	22-92	12	-45	I,I+3,I+7	635	L,V,L,L,Y,F,I,V
1MG8	1	3-79	12	-60	I,I+3,I+7	601	L,I,V,L,V,L,V,L
1VCBa	1	3-79	12	-60	I,I+3,I+7	601	L,I,V,L,V,L,V,L
1M94a	1	2-72	12	-60	I,I+3,I+7	613	V,V,V,F,L,L,L,L
1J8Ca	1	33-102	12	-60	I,I+3,I+7	594	V,V,V,F,I,L,V,L
1H8Ca	1	4-80	12	-45	I,I+3,I+7	619	L,I,L,V,V,L,L,L
1JRUa	1	296-370	11	-45	I,I+3,I+7	646	I,I,I,I,I,L,Q,L
1EO6a	1	29-111	12	-45	I,I+3,I+7	599	V,V,V,F,I,L,V,Y
1EF1a	1	5-82	12	-60	I,I+3,I+7	542	V,V,G,L,V,L,F,F
1GG3a	1	1-77	13	-45	I,I+3,I+7	563	C,V,Q,L,C,I,F,F
1LFDa	1	17-99	12	-60	I,I+3,I+7	518	I,V,A,V,A,L,F,L
1E8Xa	1	219-309	12	-60	I,I+3,I+7	640	I,I,P,I,F,L,L,L
1K8Rb	1	71-147	12	-75	I,I+4,I+8	653	L,F,Y,L,L,V,L,I
1L7Ya	1	14-88	12	-45	I,I+3,I+7	604	F,I,F,V,A,I,L,L
1D4Ba	2	36-103	10	-45	I,I+4,I+8	575	F,V,L,A,L,L,L,V
1C9Fa	2	9-76	12	-15	I,I+4,I+8	567	V,L,L,G,F,L,L,L
1F2Ri	2	19-89	12	-60	I,I+4,I+8	511	C,L,L,A,L,L,F,A
1IP9a	2	13-84	12	-45	I,I+3,I+7	653	I,F,Y,L,I,L,I,V
1Q1Oa	2	762-854	15	-60	I,I+4,I+8	667	F,I,L,I,I,I,I,L
1FMAd	3	1-76	12	-60	I,I+3,I+7	557	I,V,V,L,M,A,V,F
1F0Za	3	1-61	8	-90	I,I+3,I+7	649	I,F,V,L,L,L,I,L
1JSBa	3	8-68	8	-60	I,I+3,I+7	594	F,V,I,V,L,V,I,V
1QF6a	4	2-61	8	-60	I,I+3,I+7	528	I,L,P,V,I,G,L,I
1JALa	4	280-361	8	-60	I,I+3,I+7	473	Y,T,A,A,I,A,M,F
1FRRa	5	2-90	8	-90	I,I+4	375	T,L,I,A,-,G,I,T
1I7Ha	5	2-102	8	-60	I,I+4	456	I,I,I,A,-,C,V,I
1AYFa	5	7-108	8	-90	I,I+4	531	V,F,L,V,-,L,V,V
1PUT	5	1-103	8	-60	I+4,I+8	449	V,Y,-,L,A,V,V,V
1L5Pa	5	1-93	8	-45	I,I+4	486	I,A,L,L,-,C,F,L
1E0Za	5	1-117	8	-60	I,I+4	458	V,Y,I,A,-,S,I,Y
1FEHa	5	2-76	8	-60	I,I+4	469	I,I,I,A,-,V,I,T
1HLRa	5	2-77	9	-60	I,I+4,I+8	619	K,I,L,L,L,V,I,T
1FO4a	5	6-89	9	-60	I,I+4,I+8	600	L,F,L,L,L,V,V,T
1JROa	5	1-79	8	-45	I,I+4	536	I,F,L,L,-,V,L,T
1FFVa	5	4-78	9	-60	I,I+4,I+8	571	I,V,L,L,L,V,V,T
2PIA	5	236-321	8	-45	I+4,I+8	536	F,V,-,I,L,T,L,L
1NENb	5	1-93	12	-90	I,I+3,I+7	612	L,F,V,A,L,M,I,I
1KF6b	5	4-92	12	-90	I,I+3,I+7	564	L,I,L,A,I,V,M,V

Table 6 (continued)

PDB code and chain aligned	Super-family* (SCOP)	Residues range correponding to ubi-like fold†	α-helix length‡	Ω-angle§	α-helix inner core residues¶	Volume of inner core residues sidechain‖	Amino acids at inner core Residues**
1QLAb	5	3-93	12	-90	I,I+3,I+7	638	L,I,I,V,I,I,I,L
1JQ4a	5	5-96	8	-90	I+4,I+8	388	I,A,-,V,A,A,I,L
1KRHa	5	4-97	8	-75	I,I+4	455	V,L,L,A,-,A,F,I
1C78a	6	23-131	13	-75	I,I+4,I+8	533	L,V,I,V,L,A,L,T
1BMLc	6	16-141	13	-75	I,I+4,I+8	500	L,V,L,I,L,A,L,-
1MG4	7	57-133	12	-60	I,I+3,I+7	621	V,F,F,L,L,I,Y,C
1SE2	8	129-234	16	-45	I,I+4,I+8	563	V,I,L,A,L,I,I,V
1AN8	8	100-206	16	-45	I,I+4,I+8	585	L,G,I,I,L,I,F,I
1AW7a	8	98-193	16	-45	I,I+4,I+8	639	L,L,L,I,L,W,I,A
2IGD	9	6-61	15	-30	I,I+4,I+8	592	Y,L,A,F,A,W,F,V
2PTL	9	18-76	16	-15	I,I+4,I+8	413	I,A,A,A,A,V,L,I
1ACC	other	487-593	9	-60	I,I+4,I+8	623	A,I,L,L,F,F,I,I
1G0Sa	other	58-148	13	-45	I,I+4,I+8	357	A,L,V,A,A,T,M,G
1I9Aa	other	33-125	13	-60	I,I+4,I+8	455	F,S,N,V,C,P,F,A

*SCOP database subdivides the β-grasp ubiquitin-like topology into 11 superfamilies: 1. ubiquitin-like 2. CAD/PB1 3. MoaD/ThiS 4. TGS-like 5. 2Fe-2S Ferredoxin-like 6. Staphylokinase/Strep-tokinase 7. Doublecortin (DC domain) 8. Superantigen toxins, C-terminal domain 9. Immunoglobulin-binding domains 10. Translation initiation factor IF3, N-terminal domain 11. glutamine synthetase N-terminal domain (this numbering is used in the table). The structures classified in "other" are categorized in other fold types, but display structure similar to the ubiquitin-roll topology (retrieved in the FSSP database). Note that a 12th superfamily was recently added to β-grasp ubiquitin-like, the TmoB-like. It is mostly similar to the 5 ubiquitin-like superfamilies and has only one member as it is now.

†Sequence range used in the alignment. Accounts for insertions or deletions inside the interval are mentioned.

‡The length of the α-helix was evaluated by inspection of the h-bond pattern of backbone atoms.

§Defined as the angle between the plane of the β-sheet (β-strand 5 is used as a guide) and the α-helix axis (11). The angle is approximated by manual observation of the structures.

¶The residues of the α-helix participating in the hydrophobic core are defined by visual examination of each structure. The numbering (e.g. I, I+3, I+4, I+7 and I+8) refers to positions 78, 81, 82, 85 and 86 of the alignments (for more information, see text under the subheading: Effect of the Variation in the Register of the α-Helix on the Entropy Scores).

‖Volume of inner core residues side chains in cubic angstrom (Å^3) were calculated according to the volume of amino acid chemical groups as defined by Richards *et al.* (12). The core residues were defined by visual evaluation of the structure and the entropy profile. They correspond to residues 58, 60, 78, 81 or 82, and 85 or 86, 98, 126 and 128 of the Raf RBD. Although there is diversity of the sequences and variations in fine details of the structure, the packing of the α-helix involved these residues most of the time. Nevertheless, it is observed quite frequently that residues 78 and 86 (core position 1 and 3 of the α-helix) are not true core residues.

**List of amino acid observed at inner core residue as described above in the two last footnotes and by taking into account the difference in α-helix packing. Note that structures with shorter α-helix have gaps at core position (78 or 86).

Table 7. Entropy profile characteristics and mean pairwise sequence identity for the experimental data and the 3 natural sequence alignments

Alignments	No. of sequences, *N*	Mean pairwise identity*, %	Highest pairwise identity†, %	Mean entropy‡	Standard deviation‡
Experimental	Table 1	ND	ND	0.82	0.14
SMART (MSA S1)	22	20.6	78.0	0.55	0.16
SCOP (MSA S2)	27	10.3	35.4	0.71	0.10
SCOP-FSSP (MSA S3)	54	9.4	35.4	0.77	0.09

*The pairwise amino acid identity is calculated for all pairs of sequences from the natural sequence alignments (Table 4). Positions displaying gaps in either of the two sequences are not considered in the calculation to avoid biasing the result.

†The highest pairwise identity observed within an alignment.

‡The entropy was calculated according to Eq. **1**. The mean entropy and its standard deviation were calculated by considering each position degenerated in the main experiment or each position in the natural sequence alignments plotted (Fig. 2 *B-D* and *Methods*).

Table 8. Raw z scores of occurrences for the 20 amino acids at all residues degenerated experimentally

	S	N	T	I	R	V	F	L	P	N	K	Q*	R	T	V	V	N	V	R	N	G	M	S
W	-0.7	1.4	0.7	-1.4	-0.7	-1.4	1.4	-1.4	-1.4	-0.7	-0.7		0.5	-1.4	-0.8	-1.4	-0.8	-1.4	-0.3	0.1	-1.0	-1.0	-1.4
Y	-0.7	-0.7	0.7	-1.4	0.0	-1.4	0.7	-1.4	-1.4	-0.7	-0.7		1.2	-1.4	0.5	-1.4	-0.1	-1.4	-0.6	-1.0	-0.3	-0.3	-1.0
F	0.7	0.0	-0.7	0.7	1.4	4.1	0.7	0.8	-1.4	0.1	0.1		1.8	-1.4	-0.8	-0.8	-0.1	-0.3	-0.3	-1.4	-1.4	-1.4	-1.4
V	-1.0	-1.0	0.5	2.0	0.0	10.9	-1.0	-1.5	-2.0	-1.5	-1.0		1.2	-1.6	1.2	3.1	0.7	7.3	-1.2	-1.2	-1.2	0.7	-0.9
I	-1.4	-1.4	-0.7	4.8	0.7	4.8	-0.7	6.7	-0.7	-1.4	-1.4		1.2	-0.1	3.1	1.2	1.2	3.2	-0.6	-0.3	-1.4	1.6	-1.4
L	0.4	-2.1	1.2	11.5	-0.9	-2.5	-0.9	15.1	-1.2	-1.2	-1.6		0.6	-2.5	-0.2	0.9	0.2	-1.1	0.7	-0.2	-1.1	0.2	-1.4
M	0.7	4.1	1.4	2.1	0.0	-1.4	1.4	1.5	-1.4	-0.7	0.1		1.2	-0.8	2.4	2.4	1.8	-1.0	0.1	-0.6	-1.0	2.0	-0.3
C	0.0	2.1	2.1	0.0	0.7	2.7	-1.4	0.1	-0.7	-1.4	0.8		-0.1	-0.1	0.5	-0.1	-0.1	1.3	0.1	0.5	-1.0	0.1	0.1
A	1.5	0.0	0.5	-2.0	2.5	4.4	2.5	-2.0	-1.0	-1.0	0.6		0.3	-0.2	-0.6	-0.2	-0.6	3.7	-0.4	1.0	1.5	0.2	-0.9
G	0.5	-1.0	-1.5	-2.0	-2.0	-2.0	-1.5	-2.0	4.9	-1.0	1.7		-1.6	-1.6	-2.0	-1.6	-1.1	-1.2	0.2	-0.1	2.4	-0.1	-1.2
P	0.0	0.0	0.5	-2.0	0.0	-2.0	-2.0	-2.0	16.0	-2.0	-2.0		-2.0	-2.0	-1.6	-1.6	-1.1	1.5	1.0	3.7	1.8	-1.2	0.4
T	1.5	1.0	1.5	-1.0	2.5	-1.5	-2.0	-2.0	-1.0	-1.0	0.6		-2.0	12.7	4.0	1.2	0.3	-0.9	0.4	-0.1	0.2	-0.4	5.6
S	-0.9	2.0	-1.3	-2.1	-0.9	-2.5	0.8	-2.5	-2.1	1.0	1.4		-1.0	5.9	-1.0	-1.0	0.6	-2.1	0.2	0.4	1.1	0.4	3.6
N	0.0	0.0	0.7	-1.4	0.7	-0.7	1.4	-1.4	0.1	5.2	3.0		-0.1	-1.4	-0.1	-0.8	-0.8	-1.0	0.1	-0.6	0.9	-0.6	-0.3
Q	-0.7	0.0	0.0	-1.4	-1.4	-1.4	0.7	-1.4	0.1	-0.7	0.1		0.5	-0.8	3.1	2.4	1.8	-1.0	0.1	-1.0	-0.3	0.9	-1.0
D	0.0	0.0	-1.4	-0.7	-1.4	-0.7	0.7	-1.4	-1.4	7.4	0.8		-0.8	-1.4	-1.4	-1.4	-0.8	-1.0	-0.3	-1.4	-1.0	-1.0	-1.0
E	-1.4	0.0	-1.4	-1.4	1.4	-1.4	-0.7	-1.4	-1.4	5.2	1.5		-1.4	-1.4	-0.1	-1.4	-1.4	-0.3	-1.0	0.5	0.9	0.9	-0.6
K	0.0	-1.4	0.0	-1.4	1.4	-1.4	2.1	-1.4	-1.4	-0.7	-1.4		1.2	-0.8	-1.4	1.2	-0.8	-0.6	1.3	0.9	0.5	0.1	2.0
R	1.6	0.0	-0.5	-2.5	-0.5	-2.5	-0.5	-2.5	-2.5	-1.6	-1.2		1.3	-2.5	-1.0	0.9	1.3	-2.1	0.0	0.7	0.4	0.4	0.2
H	0.0	-0.7	-0.7	-1.4	-1.4	-1.4	2.1	-1.4	-1.4	1.5	1.5		0.5	0.5	-1.4	-1.4	0.5	-0.6	1.3	-0.3	-0.6	-0.6	-0.6

Table 8 (continued).

	L	H	D	C	L	M	K*	A	L	K	V	R*	G	L	Q	P	E	C	C	A	V	F	R
W	-1.4	-1.4	-0.8	-1.4	-1.4	-1.4		-1.4	-1.4	0.8	-0.3		-0.7	-0.1	-1.4	-0.1	-1.4	-1.4	-1.4	-0.9	-1.4	0.1	1.1
Y	-1.4	1.7	-0.8	-1.4	-1.4	-0.8		-1.4	-1.4	-0.3	-0.3		-1.4	1.3	-1.4	-0.7	-1.4	1.3	-1.4	0.6	-0.4	1.1	1.1
F	-1.4	-0.8	-1.4	-0.8	-1.4	-0.8		-1.4	3.0	0.8	-0.3		-1.4	-0.1	-0.7	-0.1	-1.4	-0.7	-1.4	-0.9	-1.4	0.1	1.1
V	5.2	-0.2	-0.2	-0.2	1.1	-0.2		-0.4	-0.8	2.4	0.8		-1.5	0.9	-0.1	1.8	-0.6	-1.5	-1.7	0.5	9.0	0.1	0.1
I	3.0	-0.2	1.1	6.7	2.3	-0.2		-1.4	4.1	-0.8	0.3		-1.4	3.3	-0.1	-0.1	-1.4	-1.4	0.1	-0.4	6.5	1.6	1.1
L	15.4	-1.4	-2.5	0.8	13.9	-0.3		-2.5	10.1	-0.2	1.1		-2.1	8.3	-1.3	0.3	-2.1	-2.1	7.5	-0.5	0.7	0.4	-0.5
M	-1.4	-0.2	0.5	1.1	6.1	1.1		-1.4	8.0	0.8	1.9		-0.7	3.3	0.6	-0.7	-0.7	-1.4	10.4	1.1	-0.4	-0.9	0.6
C	-1.4	-0.8	-1.4	3.6	-0.2	-0.2		0.3	0.3	0.3	-0.3		-1.4	1.3	-0.1	-0.7	-0.7	0.6	-0.9	0.1	-0.4	-0.9	-0.9
A	-1.6	0.2	0.2	-0.2	-2.0	0.2		2.4	-1.2	0.0	-2.0		-0.1	-1.5	-0.1	1.4	0.4	0.9	-1.3	-0.2	-0.2	0.1	-0.6
G	-2.0	-1.1	-0.7	-1.6	-2.0	-0.2		-0.8	-2.0	0.8	-1.6		7.6	-1.5	-0.1	-0.1	1.4	-0.6	-1.3	0.1	-1.0	0.1	-1.3
P	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0		1.2	-2.0	-2.0	-2.0		-2.0	-2.0	-0.1	0.4	-0.6	-2.0	-2.0	-2.0	-1.0	0.5	-1.7
T	-2.0	-2.0	1.6	2.0	-2.0	-1.1		0.0	-2.0	-1.6	-0.4		-1.1	-1.1	-0.1	-1.1	-0.6	-0.1	2.9	1.9	-0.6	1.2	-1.0
S	-2.5	0.1	-1.8	0.5	-2.1	0.5		4.8	-2.5	-1.2	0.8		0.7	-2.1	0.7	0.3	-0.1	0.3	-1.9	1.0	-2.5	-0.7	-0.2
N	-1.4	-0.2	2.3	1.7	-1.4	-0.2		-1.4	-1.4	0.8	1.9		4.0	-1.4	-0.1	-1.4	3.3	-0.1	0.1	-1.4	-1.4	0.6	-0.9
Q	-1.4	2.3	1.1	-0.8	-1.4	1.7		-0.8	-1.4	-0.3	0.3		0.6	-1.4	1.9	0.6	3.3	2.6	-1.4	-1.4	-0.9	0.1	-0.4
D	-1.4	-1.4	6.1	-1.4	-1.4	-0.2		-1.4	-1.4	-0.8	-1.4		1.9	-1.4	4.6	-1.4	1.3	7.3	-0.9	0.6	-0.4	-1.4	-1.4
E	-1.4	4.2	3.6	-0.2	-1.4	3.6		-1.4	-1.4	0.8	-0.8		-0.1	-0.7	-0.1	-0.7	4.6	3.3	-1.4	1.6	-0.4	-0.9	-0.4
K	-1.4	3.0	1.7	-1.4	-1.4	1.1		5.2	-1.4	2.5	0.8		1.9	-1.4	0.6	-0.1	1.9	-0.7	-0.4	2.5	0.1	-0.4	-0.4
R	-2.5	3.1	-0.7	-2.5	-2.5	0.1		0.8	-2.5	0.1	0.8		-0.5	-1.7	-0.1	1.1	-1.3	-0.9	-1.9	0.1	-1.3	0.1	5.4
H	-1.4	-0.8	-0.8	0.5	-1.4	1.7		-0.3	-1.4	-0.3	2.5		-0.7	-1.4	-0.7	0.6	-0.1	1.3	-1.4	-1.4	-1.4	0.1	-0.9

Table 8 (continued).

	L	L	H	E	H	K	G	K	K	A	R	L	D	W*	N	T	D	A	A	S	L	I	G
W	0.6	-0.8	-0.8	0.4	-1.4	-0.8	-1.4	-1.4	0.8	-0.9	-0.9	-1.4	-1.4		-1.4	-0.8	-0.8	-0.8	-0.8	-1.4	-1.4	-1.4	-1.4
Y	-0.9	-0.8	1.0	0.4	-0.2	-0.8	-1.4	-0.2	-1.4	-0.3	-0.9	-0.9	0.2		-0.8	-0.8	-1.4	-1.4	-0.2	-1.4	-0.1	-1.4	-0.8
F	-0.9	-1.4	-0.2	-1.4	-0.8	-0.2	0.4	-0.2	-1.4	-0.9	-1.4	-0.9	-1.4		-0.2	-0.8	-0.8	-1.4	-0.8	-0.2	-1.4	-1.4	-1.4
V	0.8	-2.0	-0.3	-1.1	0.2	-1.1	0.6	0.6	-0.4	1.9	-1.2	4.7	-0.8		-2.0	0.7	-1.6	0.2	-1.1	-0.7	0.8	0.3	-0.1
I	0.6	0.4	0.4	-0.2	0.4	0.4	0.4	-0.8	1.3	1.3	-0.9	1.3	-1.4		-1.4	-0.2	-0.8	-0.8	-0.8	-1.4	2.5	-0.8	-1.4
L	1.0	-1.4	0.0	-1.1	-0.3	-0.3	-1.1	-0.3	0.1	0.4	-0.6	7.0	-1.2		-2.5	-1.0	-2.5	-2.5	-2.5	-2.1	6.0	0.6	-1.7
M	1.6	1.0	-0.8	-0.8	-0.8	-0.2	1.0	-0.8	1.3	1.3	0.8	0.2	-0.9		-0.8	4.2	-0.8	-0.2	1.1	-0.8	1.9	0.5	-1.4
C	-0.4	-1.4	0.4	-1.4	2.2	0.4	-0.2	-0.8	-0.9	-0.3	-0.3	-1.4	-0.9		-0.2	3.0	0.5	4.2	-0.8	-0.8	4.5	0.5	0.5
A	-0.2	0.6	-0.3	-0.3	-1.1	-0.3	-0.7	2.8	-1.6	2.3	1.5	1.1	1.1		0.2	0.2	-0.7	10.1	-0.7	0.7	-1.5	-0.1	-1.5
G	-0.6	1.0	-0.7	2.8	-0.7	1.9	-0.3	0.2	0.0	-0.4	-0.8	-2.0	0.7		2.5	-0.7	-1.1	0.7	-0.7	0.2	-1.5	1.7	8.3
P	-1.7	0.6	-0.3	-0.7	0.6	-0.7	0.2	-0.7	-1.2	-0.4	-0.4	-1.2	0.7		-1.1	-2.0	3.4	-0.7	-1.6	-1.1	-2.0	-0.6	-0.1
T	-0.6	0.2	-0.7	1.5	-1.1	-1.1	1.0	0.2	1.1	-0.4	1.1	0.7	0.3		-0.2	1.6	1.1	-0.2	2.9	1.6	-1.5	1.3	0.3
S	-1.3	0.4	0.4	-1.1	0.4	0.0	1.5	0.4	0.1	-1.5	3.4	-1.9	1.7		1.2	0.5	-0.3	0.5	3.8	1.6	-1.3	-0.6	-0.6
N	-0.4	1.6	1.6	1.6	0.4	-0.2	1.0	0.4	0.8	0.2	-0.3	-1.4	-0.3		1.7	-0.8	2.3	-0.2	-0.2	-0.8	-0.1	-0.1	1.2
Q	-0.4	1.0	-0.2	0.4	0.4	-0.8	-1.4	1.0	-0.9	-0.3	-0.3	-0.3	3.0		1.7	1.7	0.5	-1.4	-0.2	-0.2	1.2	1.9	1.9
D	-0.9	-0.2	2.2	0.4	-0.2	0.4	1.0	1.0	-0.9	0.2	-0.9	-0.9	2.4		3.0	-0.2	9.2	-0.8	1.1	5.5	-1.4	0.5	3.2
E	0.6	4.7	2.2	1.0	1.6	2.8	1.6	1.0	0.2	1.9	2.4	-0.9	0.8		0.5	-0.8	1.7	-1.4	4.2	2.3	-0.8	-1.4	-0.8
K	2.1	-0.8	-0.2	1.6	1.6	1.6	0.4	1.6	0.2	-1.4	0.2	-0.9	-0.3		1.7	-0.2	-0.2	-1.4	-0.2	1.1	-0.8	-1.4	-1.4
R	2.8	-0.7	-0.7	0.4	0.7	0.4	-0.3	-1.4	2.1	-0.6	-0.6	-2.2	-1.2		-0.7	-1.0	-2.1	-1.4	-0.3	-0.7	-1.7	2.1	-1.0
H	-0.9	0.4	-0.8	-0.8	-0.2	0.4	-1.4	-1.4	1.3	-0.9	-0.3	-0.9	-0.3		1.7	-0.2	-1.4	-1.4	-1.4	0.5	0.5	-1.4	-0.8

Table 8 (continued).

	E	E	L	Q	V	D	F
W	-0.1	-1.4	-1.4	0.0	-0.7	-0.7	-1.4
Y	-0.8	-0.7	-1.4	1.4	-0.7	-0.7	-0.7
F	0.5	-1.4	-0.7	1.4	-1.4	-1.4	-0.7
V	-2.0	-2.0	1.5	2.0	16.1	0.5	1.0
I	-0.8	-0.7	7.7	0.0	5.6	0.0	0.7
L	-0.6	-1.3	10.5	-1.3	0.0	-0.8	-0.4
M	-1.4	0.0	0.0	0.7	0.0	0.0	0.7
C	1.2	0.0	0.0	0.7	0.0	0.0	-0.7
A	0.8	-1.0	-2.0	-0.5	-2.0	-2.0	0.5
G	-0.1	-1.5	-1.5	-2.0	-1.5	-2.0	-2.0
P	-2.0	-0.5	-2.0	-2.0	-1.5	-2.0	0.0
T	4.1	1.5	-2.0	1.5	-1.5	2.5	-0.5
S	1.4	1.7	-1.3	-0.4	-2.1	-0.8	1.7
N	1.2	0.7	-0.7	0.7	-0.7	0.0	1.4
Q	-0.1	-0.7	-1.4	0.0	-1.4	1.4	2.1
D	-0.8	2.8	-1.4	0.7	-1.4	3.5	0.7
E	1.9	0.0	-0.7	2.1	-1.4	6.3	2.8
K	-0.8	3.5	-0.7	0.7	-1.4	0.7	-0.7
R	-1.7	1.3	-2.1	-1.7	-2.5	-0.4	-2.1
H	1.9	1.4	-1.4	0.0	-0.7	0.0	0.0

*Grey columns indicate unvaried residues

Table 9. Listing and classification of significant amino acid selections distribution ($p < 0.05$) observed experimentally and in the three natural sequence alignments

Residue*	Experimental†	SMART (MSA S1)†	SCOP (MSA S2)†	SCOP-FSSP (MSA S3)†	Role‡
56	M, C, S	K, S	M, I, C	M, I, H	G/oC
58	L, **I**, M	C, F	I, V, F, C	I, V, L, F	G/C
60	**V**, I, A, F, C	V, L	V, I, F	V, I, F	G/C
62	**L**, I	L	T	V, T, L	G/oC
63	**P**, G	P	N, P	N, P	S/t
64	D, **N**, E	N, D	-	D	S/t
66	ND (Table 3)	Q	-	Q	S/B
68	**T**, S	T, V, C	F	F, Y	S/B
70	**V**, M, Q	V	V, L	V, F, L, I	G/oC
72	**V**, A, I	V	V, C, A	V, A, C	G/oC
75	**G**	G	D, S	G, E	G/t
76	**M**	M, K, E, H	D, T	T, M	G/-
77	**T**, S, K	T, R, S	T, S	T, S	G/Nh
78	**L**, V, I	V, L, I	L, V	L, I	G/C
80	**D**, E, N	D, E	D, Q	D, K	G/Nh
81	I, **C**, T	V, C, A	L, V	A, V	G/oC
82§	**L**, M, I	L	L/**L,V**	K, L/**L, A, V, I**	G/C
83	E	E	E, C, A	R, E	G/-
84	ND (Table 3)	K, P	V	Q	S/B
85	K, S, **A**	I, A, L	I, L	A, I	S, G/Bo, C
86§	**L**, M, I, F	C, L	-/**I, L**	L, A/**L, I**	G/C
89	ND (Table 3)	R, Y	-	-	S/B
90	**G**, N	G, N, Q	G	G, N	G/Ch
91	**L**, M, I	L, I	L	L, I	G/Ch, oC
96	M, L, T	C, Y, H, V	F, L	C	G/oC
98	**V**, I	V, L	L	L, V, I	G/C
100	**R**	L, R	V, F	V, F, I	S/oC, mC, Sb

Table 9 (continued).

Residue*	Experimental†	SMART‡	SCOP‡	SCOP-FSSP‡	Role§
112	**L**, V	L, M, I	L, V, I	V, L	G/oC, mC
113	Q, **D**	D, C	D	D, L	G/t
114	ND	W, L	D, W	D, A	S/C
117	**D**, P, N	P, D	T, P	P, T	G/t
118	**A**, C	S, C, V	L, V	L, M, V, I	G/oC
121	**L**, C, I	L	C, L	C	G/oC, mC
123	**G**, D	G	G, D	G, D	G/t
125	K, D	E, D	T	V, T	S/mC,Sb
126	**L**, I	L, V	L, F, I, V	I, F, L, V	G/C
128	**V**, I	V, I, Y, L	L, F, V	L, V, I, T	G/C
129	E, **D**, T	E, D, H	Y	H	S/Sb

*Residues with amino acid selections distribution with $p < 0.05$ for at least one amino acid are listed in this table if the selection is also observed for the same amino acid type or class in either the structural analogue alignments or the functional homologue alignment.

†The amino acid type frequencies with $p < 0.05$ are presented from the most to the least significant deviation. In the experimental data set, the amino acids in bold are observed in the wild-type Raf RBD.

‡The residue with significant amino acid selections are classified on the basis of the convergence between the experimental data and either the functional homologues (S) or the structural analogues (G). By observation of the structure and consultation of the literature (10, 17, 18) residues are further described with one of the following acronyms: inner core (C), outer core (oC), mini core in carboxyl terminus (mC), turn (t), α-helix N-cap (Nh), α-helix C-cap (Ch), putative salt bridge (Sb), binding to *h-ras* (B). To simplify the descriptions in the article the outer core also included three residues classed in the mini core in this table: R100, L112, and L121. The residues that participate in the mini core and the organization of their side chains vary across structures included in the structural analogue alignments. Several structures do not even display this mini core (e.g. 2IGD and 2PTL).

§Two sets of amino acid (set1/**set2**) for which significant amino acid selections were observed are presented in this table and are based on: (*i*) The z score for position 82 and 86 of the SCOP and SCOP-FSSP alignments as displayed in Fig. 3 and (*ii*) a composite z score calculated by aligning inner core residues 2 and 3 of the α-helix, thus correcting for the two types of α-helix inner core residues register observed across the ubiquitin superfold (see text under the subheading: Effect of the Variation in the Register of the α-Helix on the Entropy Scores).

Table 10. Statistics concerning solvent accessibility and B factors distribution for inner versus outer core and all other residues

Structural characteristics and structure coordinates utilized	Types of residues	Average	Standard deviation
Solvent accessibility based on 1RFA	All other	78.8	44.7
	Inner	2.3	2.1
	Outer	19.2	19.7
Solvent accessibility based on 1GUA	All other	75.2	47.8
	Inner	1	1.3
	Outer	15.9	24.8
B factors for main chain atoms based on 1GUA	All other	17.2	8.7
	Inner	11.6	1.2
	Outer	13.3	3.0
B factors for lateral chain atoms based on 1GUA	All other	21.0	10.7
	Inner	12.1	2.3
	Outer	15.3	5.8

Table 11. Binding affinity of Raf RBD variants for *h-ras*

Libraries	Clone names	Sequences*	K_d†‡§, μM
wt	-	-	0.11 ± 0.029
S1	CF3	SFNLAVK	4.6
S2	A2	LPRLQ	3.6
S2	B2	LPGHQ	0.12
S2	H2	FTDGQ	8.6
S3	A3	GTRVT	7.4
S3	F3	WSIQR	0.87
S6	D6	QPLRLR	4.3
S6	E6	KALQQR	16
S6	G6	KRMTAR	5.4
S7	G3	EINHLQ	4.2
S8	H7	KLVIAR	0.42
S8	F8	CKLMRR	3.0
S9	A9	PDQSATG	0.72
S9	D9	KTQGCNG	0.11
S9	E9	TSGRVLH	0.07
S10	G9	VAILDW	3.7
C96M	-	-	0.38
R100A	-	-	0.64
D117A	-	-	0.13
E125A	-	-	0.76
D129A	-	-	0.74

*Sequences of the clones in the degenerated region.
†According to *Methods*.
‡The dissociation constant for the wt Raf RBD was determined independently from four experiments (each in triplicate). The standard deviation is indicated.
§The clone D6 was isolated from library S6 after one round of selection. An unexpected point mutation occurred: K84Q.

Table 12. Kinetic and thermodynamic parameters for characterized mutants of the Raf RBD

Libraries and clones	Sequence	k_f, s^{-1}	k_u, s^{-1}	m_f, kCal	m_u, kCal	ΔG, kCal. M^{-1}	M, kCal	C_m, M	β_t
wt	--	347	0.19	0.67	0.28	-4.44	2.37	1.87	0.70
V60A	--	66.0	0.43	1.05	0.26	-2.99	3.24	0.92	0.80
S1 (A1)	PWVDLDA	1220	4.01	0.61	0.23	-3.39	2.08	1.62	0.73
S1(CF3)	SFNLAVK	321	1.22	0.77	0.29	-3.30	2.62	1.26	0.73
S2 (A2)	LPRLQ*	179	0.40	0.86	0.27	-3.62	2.82	1.28	0.76
S2 (B2)	LPGHQ*	185	0.12	0.68	0.29	-4.34	2.41	1.80	0.70
S2 (H2)	FTDGQ*	1405	0.52	0.60	0.28	-4.69	2.18	2.15	0.68
S3 (A3)	GTRVT	242	0.39	0.71	0.30	-3.82	2.50	1.52	0.70
S6 (A6)	K*KLSER*	128	0.18	0.93	0.32	-3.87	3.09	1.25	0.74
S6 (C6)	K*RLVWR*	361	1.49	0.92	0.28	-3.25	2.96	1.10	0.77
S6 (G6)	K*RMTAR*	24.3	0.17	0.91	0.30	-2.93	2.99	0.98	0.75
S7 (B7)	SMSSES	183	0.72	0.93	0.25	-3.28	2.91	1.13	0.79
S7 (G7)	EILPGQ	379	0.77	0.93	0.25	-3.67	2.91	1.26	0.79
S7 (G3)	EINHLQ	249	1.85	0.79	0.29	-2.91	2.69	1.08	0.73
S8 (A8)	DKLGIW	230	0.88	0.84	0.27	-3.30	2.75	1.20	0.76
S8 (F8)	CKLMRR	453	11.4	0.88	0.36	-2.18	3.80	0.71	0.71
S10 (C10)	KRTVSW*	386	1.57	0.74	0.27	-3.26	2.50	1.31	0.73
S12 (H11)	LADC	541	1.82	0.79	0.33	-3.37	2.77	1.22	0.71
S13 (C12)	TSCHDL	74.1	0.76	0.84	0.25	-2.71	2.70	1.00	0.77
S13 (D12)	RLMVSD	95.3	0.81	0.93	0.24	-2.82	2.91	0.97	0.79
S13 (H12)	QLLLEF	232	0.16	0.60	0.28	-4.30	2.18	1.97	0.68

*Residues marked with an asterisk were not varied in the experiment.

Supporting Information Figure legends and Figures

Figure 6. Scheme of the PCR strategy used for generating the experimental libraries. Broad lines represent one DNA strand of the template DNA, whereas thinner lines represent primers. The codons located in the target section (dashed lines) are deleted and replaced by a deletion tag sequence composed of a restriction site, a stop codon, and an additional base pair to introduce a frame shift. The PCR reaction on the right is designed to replace the deletion tag by the number of NNK codons matching the number of codons in this segment of the wild-type cDNA (*Methods*).

Figure 7. The entropy scores of main libraries are compared with those obtained for alternative libraries (Fig. 1*A* and Tables 2 and 8). (*A*) Entropy scores determined for library S2 (residues L62 to K65 are degenerated) versus S2b (residues L62 to Q66 are degenerated). (*B*) Entropy scores determined for library S6 (A85 to V88 are degenerated) versus S6b (residues K84 to R89 are degenerated). (*C*) Entropy scores determined for library S8 (residues C96 to L101 are degenerated) versus library S8b (residues C96 to L101, except R100, are degenerated).

Fig. 8. Solvent accessibility values for each residue were extracted from the DSSP file of 1RFA (NMR data) (*A*) and of 1GUA (crystallographic data of the complex with Rap1a) (*B*) and plotted. (*C*) An average B factor for main chain atoms was calculated (values for Cα, C (carbonyl), and N (amide) were extracted from the PDB file of 1GUA) and plotted. (*D*) Similarly average B factors for side chain atoms (except Cβ) were calculated. Residues are subdivided as inner core and outer core, but all other residues are also shown for comparison. Insets show the distribution of solvent accessibility and B factors value for these three groups (Table 10).

Figure 9. ^{1}H and NOESY NMR spectra in amide and upfield aliphatic regions for Raf RBD wt and five clones (*Supporting Methods*). (*A*) 1D proton NMR spectra in the amide region for the wt Raf RBD versus five clones (S1, A1; S2, H2; S3, A3; S7, C7; S12, H12). The sequence of clones and the corresponding segment of the wt are shown on the left of the spectra. Relative positions of the Trp-114 NεH (9.8 ppm for wt) are indicated with a broken red line. (*B*) 1D proton NMR spectra in the upfield region for the wt versus five clones. The two extreme upfield peaks correspond to the Cγ1H_3 (0.18 ppm) and Cγ2H_3 (-0.4 ppm) methyl protons of Val-98 that are ring-current shifted by interaction with the Trp-114 side chain. Additional pair of peaks in the C7 spectrum could be due to a Pro replaced by a Leu at position 93, whose side chain could be in contact with the side chain of Trp 114. (*C*) Proton NOESY crosspeaks of Trp-114 NεH for wt versus five clones. Both intra- and interresidue crosspeaks are observed. Here we assign only the intraresidue crosspeaks because the interresidue correlations could be ambiguous. Intraresidue crosspeaks include CαH, 4.49; CβH, 2.9; Cδ1H, 7.08; Cζ2H, 7.4. (*D*) Proton NOESY crosspeaks for the Cγ1H_3 (0.18 ppm) and Cγ2H_3 (-0.4 ppm) methyl protons of Val-98 wt versus five clones. Common intrareside crosspeak correlations include NOEs from both Cγ1H_3 and Cγ2H_3 and include NH, 7.99; CαH, 4.9; and CβH, 1.03. These crosspeaks appear in spectra of wt and of the clones. Additional peaks in spectrum of H2 Cγ1H3 correspond to the overlapping peak (*B*) and are likely due to ring current shift of some aliphatic amino acid in proximity to the side chain of residue 62, which is a substitution of a Leu for Phe. The observed crosspeak correlated chemical shifts do not correspond to assignments of any aliphatic amino acids originally described.

Figure 10. Pull-down of GST-*ras* on Ni-NTA-bound Raf clones (expressed as fusions to a 6×His tag but without the DHFR F[1,2]) and binding competition with Raf RBD wt. GST-*ras* in complex with GMP-PNP preincubated with either untagged Raf RBD wt or a similar volume of a mock purification (equivalent purification fraction from cell containing an empty vector) was allowed to interact with Ni-NTA-bound 6×His-Raf RBD wt or seven clones isolated from the library screening (Fig. 5, *Supporting Methods*, and Table 11). The input of 6×His Raf loaded in each lane is also shown. Note residual untagged Raf in the appropriate lanes. The numbers at the bottom of the GST-*ras* lanes represent the ratio of the band intensities in the competition versus control treatment.

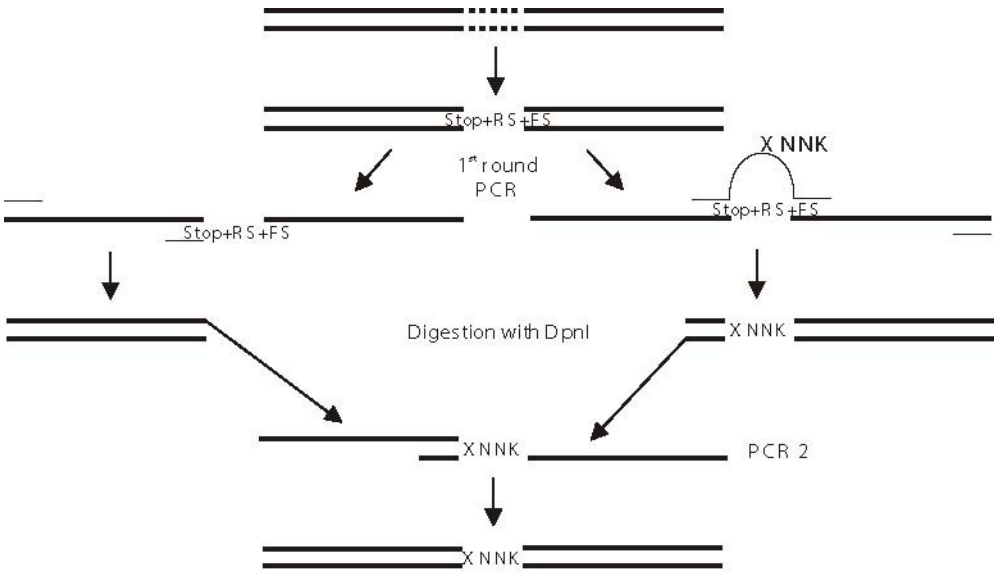

Stop+RS+FS
1st round
PCR
X NNK
Stop+RS+FS
Stop+RS+FS
Digestion with DpnI
X NNK
X NNK
PCR 2
X NNK

Fig. 6

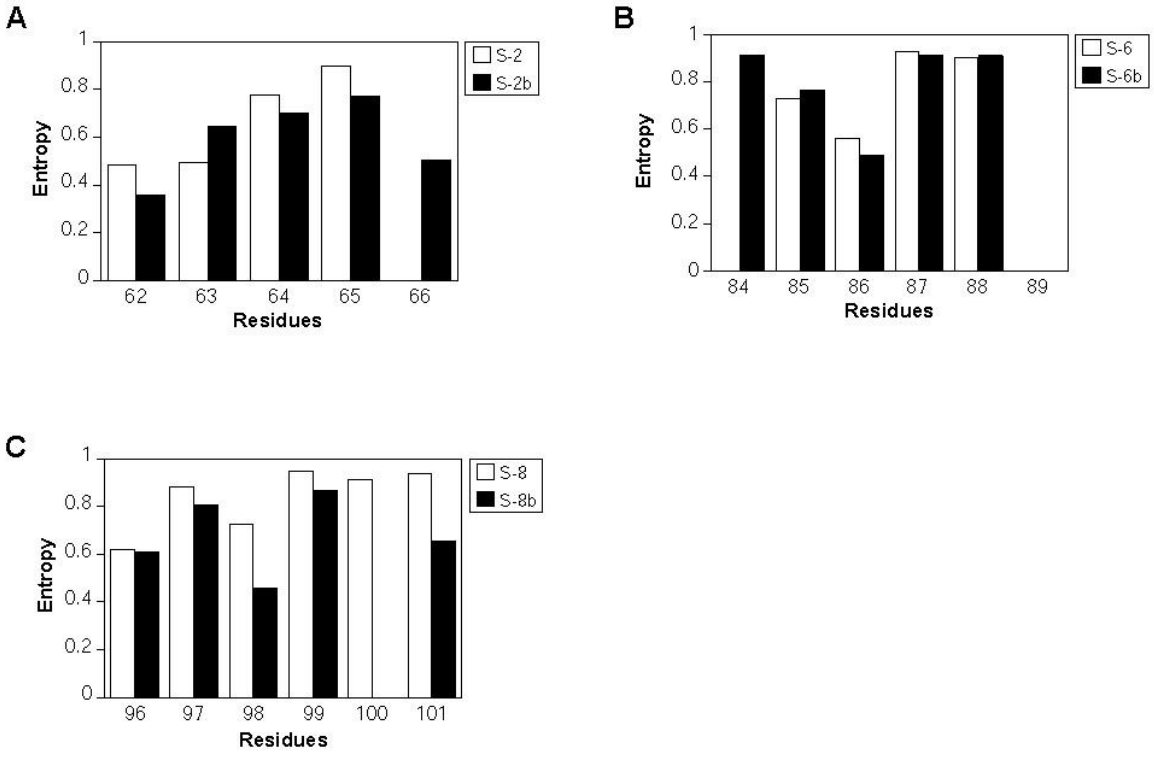

A
Entropy
1
0.8
0.6
0.4
0.2
0
62
63
64
65
66
Residues
S-2
S-2b
B
Entropy
1
0.8
0.6
0.4
0.2
0
84
85
86
87
88
89
Residues
S-6
S-6b
C
Entropy
1
0.8
0.6
0.4
0.2
0
96
97
98
99
100
101
Residues
S-8
S-8b

Fig. 7

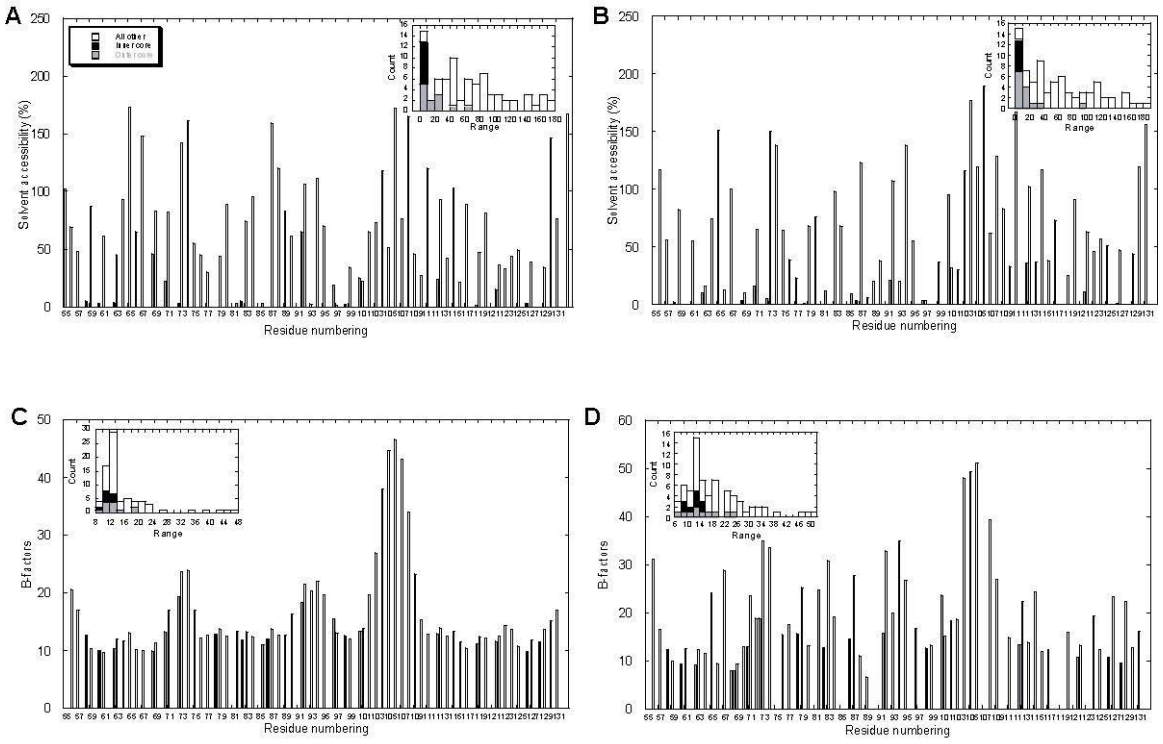

A
B
C
D
Solvent accessibility (%)
B-factors
Residue numbering
Count
Range

Fig. 8

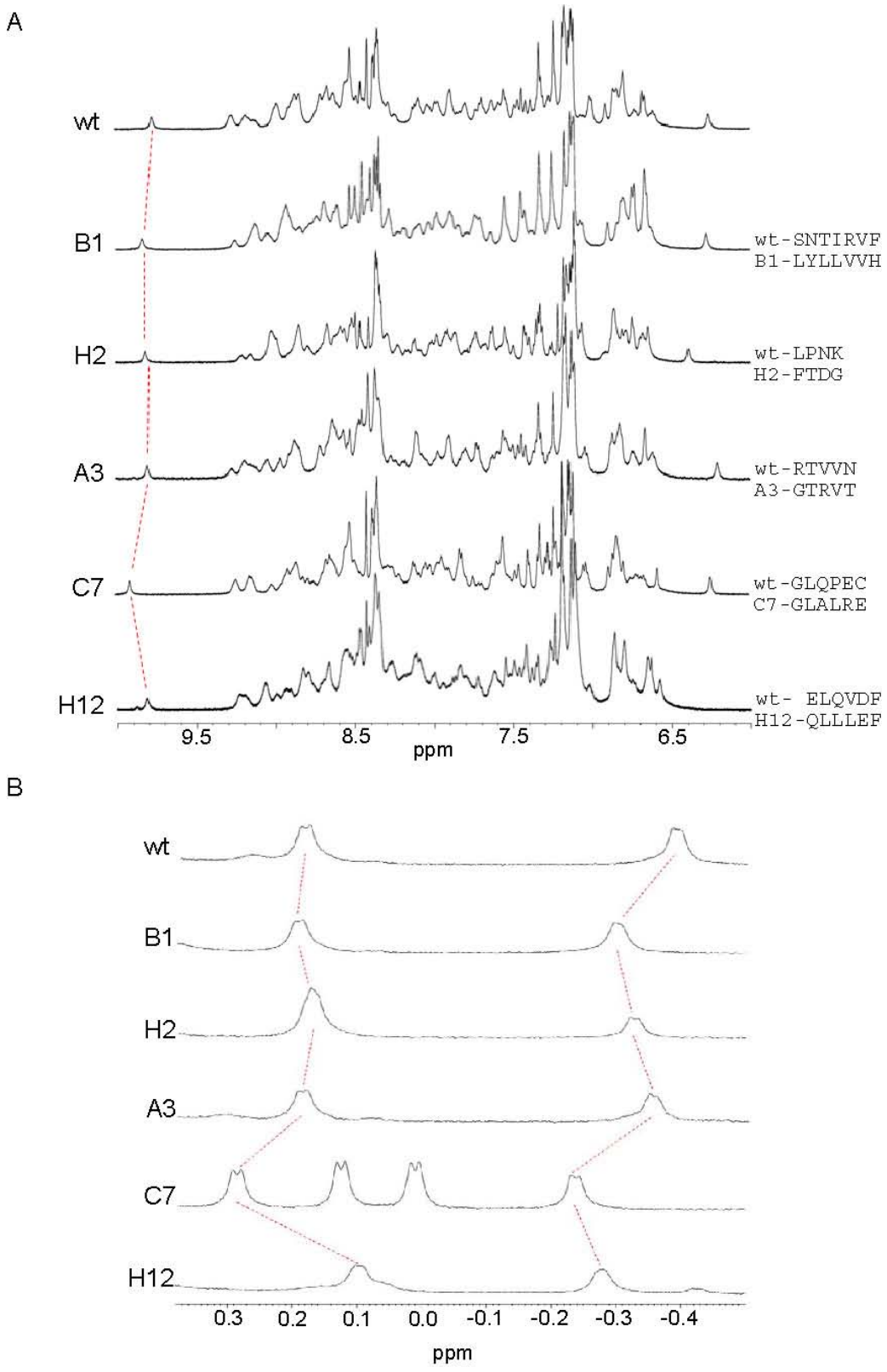

A
wt
B1
H2
A3
C7
H12
wt-SNTIRVF
B1-LYLLVVH
wt-LPNK
H2-FTDG
wt-RTVVN
A3-GTRVT
wt-GLQPEC
C7-GLALRE
wt- ELQVDF
H12-QLLLEF
9.5
8.5
7.5
6.5
ppm
B
wt
B1
H2
A3
C7
H12
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
ppm

Fig. 9

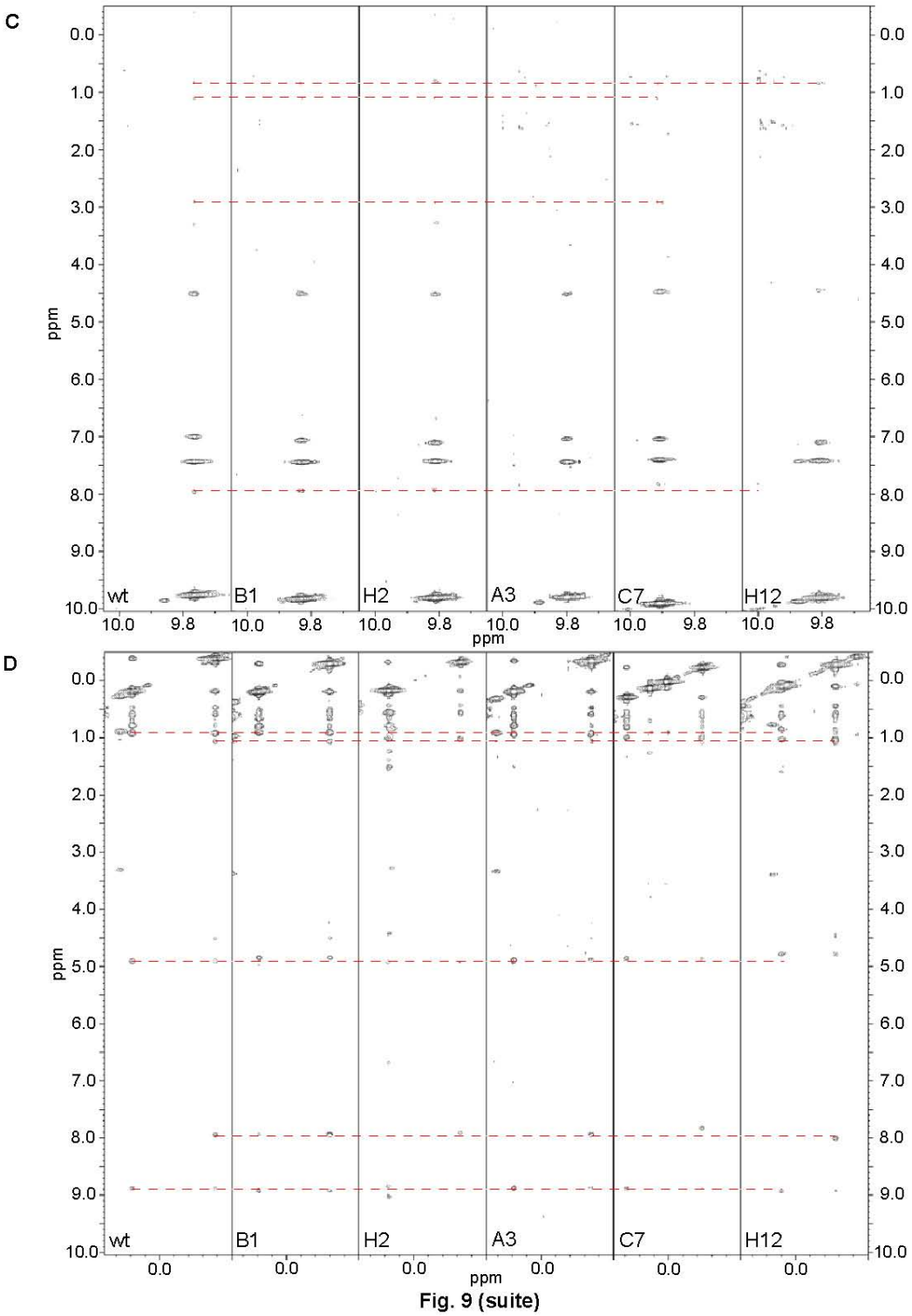


Fig. 9 (suite)

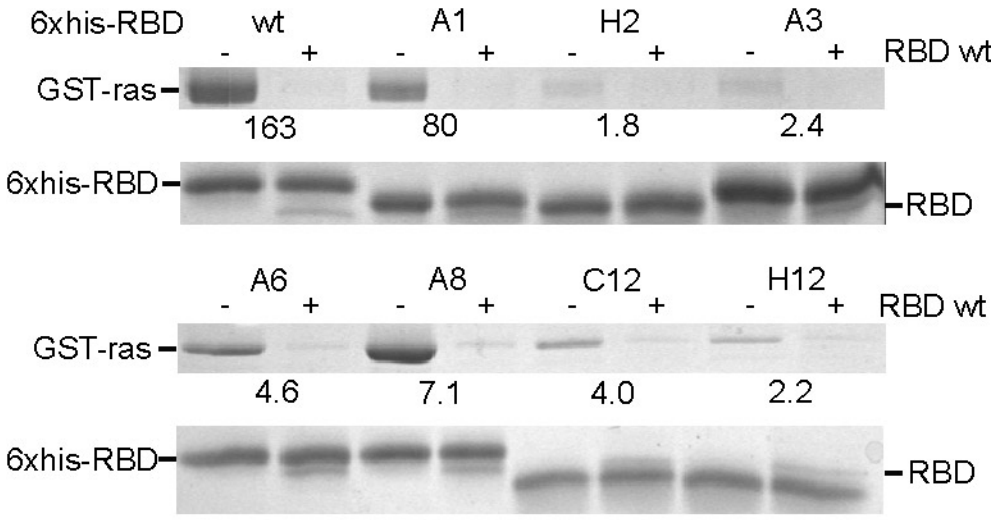

6xhis-RBD
wt
A1
H2
A3
- + - + - + - +
RBD wt
GST-ras
163
80
1.8
2.4
6xhis-RBD
RBD
A6
A8
C12
H12
- + - + - + - +
RBD wt
GST-ras
4.6
7.1
4.0
2.2
6xhis-RBD
RBD

Fig. 10

Article 3 : La perturbation massive de la séquence du DLR de Raf révèle des liens entre l'entropie de séquence, la propensité pour les structures secondaires, la conservation du volume dans le cœur hydrophobe, la stabilité et la fonction.

Article accepté sous conditions à « Journal of Molecular Biology »

Présentation de l'article 3 :

Cet article poursuit l'analyse des caractéristiques de la séquence des mutants du DLR de Raf isolés par la stratégie de perturbation de séquence (**Article 1** et **2**); spécifiquement, nous discutons des variations dans la conservation de la structure secondaire, du volume cumulatif et l'identité des acides aminés observés dans le coeur hydrophobe ainsi que de la polarisation de la distribution des charges à la surface du DLR de Raf. En second lieu, les variations dans l'organisation du cœur hydrophobe à l'intérieur de la topologie d'ubiquitine sont étudiées et mis en relations avec les analyses précédentes sur la perturbation de séquence du DLR de Raf. Sur cette base, des mutants ponctuels choisis du DLR de Raf ont été générés. Les mutations se regroupaient en deux classes générales : les troncations de la chaîne latérale (Ala ou Gly) et les atypiques qui ne respectent pas les limitations de la méthode d'ingénierie des protéines (voir section **Ingénierie de protéines, analyse des valeurs-□ et position de l'état de transition**). Les propriétés cinétiques de ces mutants seront décrites dans le quatrième article, alors que dans la présente étude, nous rapportons les paramètres thermodynamiques obtenus par des expériences réalisées à l'équilibre. Les découvertes principales de cet article sont :

1. La mesure de l'intolérance à la mutation dégénérative d'un résidu donné (i.e. entropie de séquence) du DLR de Raf telle que déterminée par la perturbation expérimentale de sa séquence est corrélée de manière prépondérante au degré de déstabilisation induit par la mutation de ce résidu en alanine ou, en glycine pour les résidus alanine. Cela suggère que la stabilité est un facteur prépondérant de la pression sélective.

2. Une cartographie du rôle des résidus du cœur hydrophobe : un rôle prépondérant pour le cœur interne, mais diversifié pour le cœur externe.

3. Le DLR de Raf n'a pas une stabilité optimale et cela est dû en partie à sa fonction de liaison à *ras*. Le mutant le plus stabilisant correspond à une délétion de trois acides aminés dans une boucle qui est raccourcie dans le cas d'autres gènes Raf connus.

4. Les structures secondaires natives du DLR de Raf sont très bien conservées dans le cadre des expériences de perturbation de la structure primaire.

5. L'organisation du réseau de contact dans le cœur hydrophobe est conservée dans les superfamilles similaires à celles d'ubiquitine. En outre, le volume cumulatif du cœur interne hydrophobe est aussi conservé.

L'apport le plus intéressant de ce manuscrit consiste à la présentation d'une preuve expérimentale solide de la prépondérance de la stabilité sur le taux de repliement ou le mécanisme de repliement dans le processus de sélection évolutive est sans doute son apport le plus intéressant. Bien que la conservation de la stabilité au cours de l'évolution ait été postulée dans plusieurs publications, auparavant des preuves expérimentales de sa prépondérance par rapport aux facteurs susmentionnés étaient peu nombreuses et indirectes (*32;109;146;147*). De plus, ce résultat valide l'approche expérimentale que nous avons utilisée et indique que le biais fonctionnel, qui peut être logiquement supputé de par la nature de notre essai de sélection des séquences, serait significatif uniquement à un nombre réduit de positions dégénérées. Il est aussi intéressant de souligner que ce résultat a été obtenu en dépit du fait que la stabilité du DLR de c-Raf n'est pas optimisée.

Contribution des auteurs à la préparation de l'article 3:

F.-X.C.V. : conception et réalisation des expériences, analyse des données et rédaction de l'article.

K.T. : développement d'outils informatiques pour la présentation et l'analyse des données.

S.W.M. : supervision du projet et rédaction de l'article.

Article 3 : «Massive sequence perturbation of the Raf *ras* binding domain reveals relationships between sequence positional entropy, structural propensity, volume conservation and stability»

Authors: Campbell-Valois, F.-X.*†, Kirill Tarassov*, Michnick, S.W.*‡

* Département de Biochimie and † Programme de Biologie Moléculaire, Université de Montréal, C.P. 6128, Succ. centre-ville, Montréal, Québec, Canada H3C 3J7

‡ Corresponding author: stephen.michnick@umontreal.ca

Summary

Evolutionary selection on natural proteins must achieve a delicate balance between function, stability and folding rates. It is widely recognized that contributions of residues to these evolutionary constraints can be discerned partly by comparing the sequences of structures having identical folds, but no discernable sequence homology. Recently, we have devised an experimental strategy to thoroughly explore residue substitutions consistent with a specific class of structure and function based on segmental perturbation of a protein sequence and applied this approach to the c-Raf/Raf-1 *ras* binding domain (thereof Raf RBD), an exemplar of the common β-grasp ubiquitin-like topology. Using this approach, a snapshot of the sequence diversity tolerated at virtually every residues of the Raf RBD was obtained, which allowed for defining the sequence determinants of this fold. Herein, we present analyses suggesting that more subtle sequence selection pressure including propensity for secondary structure, the hydrophobic core organization and charge distribution are imposed on the Raf RBD sequence. Secondly, using the Gibbs free energies ($\Delta G_{F\text{-}U}$) obtained for 54 mutants of Raf RBD, we demonstrate a strong correlation between amino acid conservation and the destabilization induced by truncating mutants. In addition, four mutants are shown to significantly stabilize Raf RBD native structure. Two of these mutations, including the well studied R89L mutant, are known to severely compromise binding affinity for *ras*. Another stabilized mutant consisted of a deletion of amino acids E104-K106 that occurs in the a-Raf and b-Raf variants. Finally, the combination of mutations affecting 5 of 78 residues of Raf RBD allows for stabilizing the structure by approximately 12 kJ.mol^{-1} ($\Delta G_{F\text{-}U}$ is -22 and -34 kJ.mol^{-1} for wt and mutant, respectively) due to a 10 fold improvement in folding rate and an equivalent reduction in unfolding rate. In conclusion, the results of the sequence perturbation approach and β-grasp structure sequence analysis have allowed us to predict sequence-specific requirements for function, stability and folding rate of the Raf RBD. The results are further discussed in the perspective of recent progress in the design of protein structure and function.

Keywords: Raf *ras* binding domain (RBD), β-grasp ubiquitin-like topology, sequence entropy, core volume, secondary structure propensity, stability, charges distribution.

Introduction

It became clear very early in the emerging structural biology field that proteins with diverging sequences can share the same overall arrangement of their backbone atoms (e.g., topology) ([1-5]). While this could appear at odds with Anfinsen's principle that the structural information is encoded fully and completely in the sequence ([6]) (reviewed in ([7])), it is rather an indication of the degenerated nature of the message encrypted in polypeptide sequences. To uncover this code, the most instructive structural comparisons are based on proteins displaying no apparent evolutionary links – specifically in the absence of significant sequence homology and common biological functions – because it could then be reasonably argued that the few commonly constrained positions in the sequences are important for defining the structural analogies. The comparisons of the primary structure of proteins adopting similar topology and of their folding reactions have been used to try to decipher the redundant messages embedded in polypeptide sequences ([8-23]) (**Article 4**).

In general, the analysis of sequence alignment of proteins adopting similar fold but low sequence identity identifies primarily hydrophobic core positions as conserved ([8, 9, 11-13]), but yields little information about the precise role and interplay between the residues in stabilization and formation of the structure. The comparison of the folding reaction of several homologous and more distantly related proteins have indicated that their folding mechanism might be comparable ([14-16, 18, 19, 21, 22]) (**Article 4**), but in some cases significantly different ([17, 20, 24, 25]) (these examples and other reviewed in ([26])). These results suggest that the folding mechanism may be encoded in the polypeptide sequence in a less constrained manner than classically thought and that clearly distinct pathways to the native state might be populated in proteins adopting the same topology. This hypothesis could provide an explanation for the diversity of fold occurrences in natural proteins, by which the most versatile topology would be favored at the structure-function level. Surprisingly, few studies have combined folding kinetics and sequence alignment analysis in an integrated manner that hold the promise of novel insights of the relationships between sequence conservation, folding mechanism, stability and function ([12, 13, 27]). Nevertheless, databases combining sequence and structural information such as "protein families" (PFAM), "simple

modular architecture research tool" (SMART), "structural classification of proteins" (SCOP), "class, architecture, topology and homologous superfamily" (CATH) and "families of structurally similar proteins" (FSSP) are useful to recover and analyze sequences and select experimental models of structurally close or distantly related proteins.

The protein universe is not homogenous in the sense that some protein topologies have been selected and spread much more than others throughout evolution. For example, an estimated 80% of known structures adopt one of the 400 most frequent topologies of the 10,000 predicted to exist in nature ([28]). This statement is apparently confirmed by the low number of novel topologies discovered by structural genomics research programs (consult http://www.jcsg.org/ and http://www.strgen.org/ for statistics on this matter and http://www.rcsb.org/pdb/strucgen.html for a list of other structural genomics centers) that established experimental methodologies specifically designed to do so ([29, 30]). A corollary of these observations is that most of known topologies are rare and therefore limited in sequence space. Rarely occurring folds pose a specific challenge to the study of structure stabilization and formation in conjunction with sequence determinants, precisely because of the paucity of sequence information available. The utilization of degenerated libraries to experimentally expand the sequence space covered by such poorly populated folds is in theory a simple means of circumventing this problem. In the past, several approaches to introduce targeted randomization into the primary structures of proteins and select the variants for their capacity to form the native structure based on various functional assays were reported ([31-37]). These types of sequence perturbation strategies are also attractive, because they allow, without the unwanted bias embedded in natural sequence evolution, for experimentally determining the relationships between sequence, structure and function. Building on these classic studies, we have recently reported a method to massively perturb the sequence of small proteins, and demonstrated its application to test the sequence variation tolerated at virtually every residue of c-Raf/Raf-1 *ras* binding domain (RBD) ([23]).

The most recognized biological function of the Ser/Thr kinase Raf is to activate the MAPK pathway, but it is also suspected to play key roles in other processes. The classic scheme by which the MAPK cascade is activated is through recruitment of Raf to the

membrane through binding of Raf RBD to GTP loaded *ras*, which then relieves the auto-inhibition of the kinase activity by phosphorylation at several sites on Raf proteins ([38]). So far, no structures of the heterodimeric complex between Raf RBD and *h-ras* have been reported. However, the complex between mutated Rap1A with charge reversal at position 31 to mimic *ras* and c-Raf/Raf-1 RBD (thereof Raf RBD) has been used to model the Raf-*h-ras* complex ([39, 40]). Further, residues located on a basic surface of the Raf RBD were shown to be implicated in formation of the complex in a mutagenesis study ([41]). The Raf RBD is constituted of 78 amino acids that form a globular structure, which is classified in the β-grasp ubiquitin-like topology according to the Structural Classification of Proteins (SCOP) database (http://scop.mrc-lmb.cam.ac.uk/scop) (Figure 1 (a)). This protein topology groups several superfamilies, some linked by putative common evolutionary origins, while others appear to result from convergent evolution, yielding very high discrepancy in sequence identity even within the same superfamily. For example, the Raf RBD and ubiquitin that are classified in the ubiquitin-like superfamily display below 12% sequence identity (based on alignment of sequences according to secondary structure). Comparable sensitivity of the folding rate of these proteins to mutations and chemical perturbation has been demonstrated ([22]). In an **Article 4**, we demonstrate that the structure of the Transition State (TS) of mammalian ubiquitin ([42]) and Raf RBD as determined by Φ-value analysis share common characteristics (**Article 4**).

Applied to the Raf RBD, our sequence perturbation strategy mentioned above, consisted in randomizing the polypeptide sequence in 13 discrete segments corresponding to secondary structure elements and selecting the variants able to fold into the native structure based on their capacity to interact *in vivo* with *h-ras* ([23]). In that study, the focus was on analyses of the tolerance to mutation of each position (e.g., sequence entropy) and the specific amino acid selection at each position (Figure 1). Specifically, we have discriminated between the functional and structural constraints at each conserved residues and shown that the conservations recapture the sequence variability observed in alignments of structural analogues recovered in SCOP β-grasp ubiquitin-like topologies. In addition, we proposed that the hydrophobic core is organized in a bi-level hierarchy and that other

topological constraints limit sequence space. Herein, we discuss more subtle aspects of selection-pressures, such as secondary structure propensity, hydrophobic core organization and charge distribution, which are imposed on the Raf RBD sequence as we observed experimentally and by evolution as deduced from alignments of natural protein sequences sharing the β-grasp ubiquitin-like topology or Raf RBD function.

An important debate about sequence evolution concerns the conservation of residues that determine the folding rate, specifically residues displaying high Φ-values, sometimes referred as the nucleus for folding ([43-46]). In one specific example, computer simulations and experiments suggested that native state stability and function are the major determinants of sequence conservation in the SH3 domains structural families {Di Nardo, Larson, et al. 2003 141 /id}([47]).ghg Here and in an **Article 4**, we report how the knowledge of sequence conservation that we obtained by sequence perturbation is combined with studies of the kinetic and thermodynamic properties of point mutants of Raf RBD, allowing for exploration of the relationship between sequence conservation, folding and stabilization of its native structure.

Recently, sequence conservation and co-varying positions in alignments were used to generate artificial WW domains with stability and binding affinities toward their natural substrate comparable to their natural counterparts ([48, 49]). In contrast, *de novo* design or redesign of natural proteins have tended to show that the absence of selective pressure for function in the computer simulation algorithms lead to hyper stabilized variants and conversely, that the natural proteins are slightly sub-optimal for native structure stability ([50, 51]). Accordingly, the results presented below suggest that the stability and folding rate of the Raf RBD is not optimized and that this could be linked, in part, to conservation of residues for binding to *h-ras*.

Results and discussion

The massive sequence perturbation experiment on Raf RBD allowed us to construct an experimental sequence-positional entropy profile so that we could establish the residue conservation at each position degenerated ([23]). Remarkably, this experimental entropy

profile is similar to that obtained for proteins sharing the ubiquitin-like topology aligned according to their secondary structure (Materials and Methods, Figure 1 (b) and 2). This is evident particularly from the agreement between positions that correspond to local minima in entropy. The experimental entropy profile is also in very good agreement with the theoretical prediction from the Conservatism of Conservatism database (CoC) of the Raf RBD structure (e.g. PDB file, 1C1Yb) (http://kulibin.mit.edu/coc/index.html) and a theoretical study based on sequence alignment and computer simulations on Raf RBD and ubiquitin ([9]). Furthermore, the analysis of the specific bias in occurrences of amino acids in the experimental *versus* alignments of structural analogues or functional homologues can be used to define the sequence space constraints and discriminate grossly between their structural or functional origins (Figure 1 (c)). It is clear from Figure 1 (b) and (c) that the regions with the lowest entropy and the strongest bias involved mostly hydrophobic positions conserved in the β-grasp ubiquitin-like topology that we have sub-grouped in the inner and outer hydrophobic core based on their decreasing level of sequence conservation and dispersion in the native structure. In addition, we identified a subgroup of residues (e.g., I58, S77T, C81 and C96) that displayed a predominant bias for non-wt amino acids. We will come back to these issues later. On the other hand, it appeared that important topology-defining residues were conserved, particularly obvious in some β-turns and in the α-helix. Following these observations, we examined sequence biases for more subtle selection, including propensities for wt secondary structure in the sub-segments of the primary structure.

Specific conservation of wt secondary structure propensity in segments of the β-grasp ubiquitin-like topology

An interesting question concerning amino acid bias observed in proteins sharing the same topology is the relative importance of local factors such as secondary structure propensity imposed on the sequence space, the folding mechanism, thermodynamic stability and structure prediction. Several lines of experimental evidence argue for the presence of residual fluctuating segments of secondary structure in the denatured state of proteins, even for apparently two-state folders ([52-56]), which could constrain the conformational search

early in the folding process, thus generally agreeing with the sequential model for protein folding (reviewed in ([57, 58])). Based on this theoretical background, Rose and co-workers have proposed that secondary structure content could be used to classify protein topologies, predict the secondary structure elements able to fold in isolation and the folding rate of specific polypeptides ([59-61]). Interestingly, Rosetta the most efficient *de novo* design and structure prediction algorithm uses a library of short structural segments to find local matches with the target sequence to initiate the simulations ([62, 63]).

To explore the issue of secondary structure propensity conservation, we used the scale of Koehl and Levitt ([64]) to compare the profiles of average propensity for α-helix and β-strands at all positions varied in the sequence perturbation of Raf RBD *versus* the proteins in the alignment of β-grasp ubiquitin-like topology (Figure 3 (a) and Materials and Methods). As would be predicted, the patterns of propensities for α-helix and β-strand share several similarities in the experiments *versus* ubiquitin-roll topology alignment. For example, the observed experimental propensities in the segments corresponding to the first β-strand (T57-F61) and the major α-helix (L78-R89) show strong preference for amino acids with high propensity for these secondary structures. In contrast, β2 and β5 sequences observed in the sequence perturbation studies showed low propensities for the appropriate secondary structure except at core positions. Some segments show low propensity for wild-type secondary elements, such as β4 (A110-R111) and α2 (A118-S120). Rose and colleagues have proposed that secondary structure elements with the highest propensity for the native structure could form early in the folding process ([60]). In Raf RBD (**Article 4**), as in ubiquitin ([42]), the amino-terminal β-hairpin is the most native-like region in the transition state according to Φ-value analysis. Moreover, this β-hairpin in ubiquitin was found to fold in isolation ([65, 66]). Other evidences obtained by NMR and single-force spectroscopy techniques argue in favor of formation by ubiquitin of non-native states that could be compatible with a sequential model for folding, but all are consistent with a well stabilized amino-terminal β-hairpin ([67-74]).

In order to confirm that the sequence variants isolated through the sequence perturbation strategy have conserved the wt secondary structure, we next sought to verify the secondary structure of Raf RBD variants isolated in the sequence perturbation experiments by utilizing three of the most accurate secondary structure prediction algorithms available (e.g. PhD, PROF and PSI-PRED) to calculate the average percentage of variants adopting the wt secondary structure at each position (Materials and Methods). The overview of Figure 3 (b) reveals globally that the wt secondary structures are largely dominant with variations appearing at the margins of secondary structure elements as in β1 and β5, but more profoundly at most residues of β3 (e.g., C96-L101) for which a small fraction of the variants obtained are predicted to switch to α-helical conformation (black bars in Figure 3 (b)). In fact, 6.2±2.8% of clones are predicted to have at least 4 consecutive residues in α-helical conformation between C95-L102 (data not shown). As apparent from the error bars, the algorithms are not perfectly consistent in this region, with PSI-PRED and particularly PROF predicting a higher frequency of α-helix. We also noticed that the identity of variants experiencing the putative secondary structure switch is far from being perfectly match among the three algorithms (e.g., PSIPRED versus PROF≈ 50% and PHD versus PROF or PSIPRED≈ 0). Therefore, the particularity of β3 in this prediction test could represent either common limitation of the algorithms, which have a success rate of approximately 72-78% for globular proteins. Alternatively it could suggest genuine and rare local changes of structure, but for now we have no direct evidence supporting this hypothesis and the low level of correlation between the algorithms is of no use to address this question. It is noteworthy that immediately before β3, there is a tight β-turn reminiscent of a single turn α-helix; particularly the P93 and E94 could act as efficient N-capping residue of a putative α-helix. We hypothesize that this element of secondary structure could favor in the context of the segmental sequence perturbation methodology the induction of a secondary structure switch or could confound the prediction algorithms.

In nature, the evolution of novel protein folds from a template protein gene is dependant on the accumulation of several single point mutations over a long time and on several mechanisms, including addition/deletion of structural elements, circular

permutation, strand invasion/withdrawal and β-hairpin flip/swap, that can induce much more dramatic effects on structure ([75]). In this perspective, some elements in the structure such as isolated β-turns could represent a potential nexus between known structure and the evolution of novel protein folds from the accumulation of a relatively small number of point mutations or insertion of few amino acids.

The hydrophobic core in the ubiquitin related superfamilies adopt two distinct patterns

Analysis of the sequence perturbation experiment showed that the hydrophobic core of Raf RBD has a two-layer concentric organization (Figure 4 (a)) and this organization could be extrapolated to ubiquitin (Figure 4 (b)) using the alignment of natural sequences displayed in Figure 2. The innermost layer (inner core: red) includes the residues I58, V60, L78, L82, L86, V98, L126 and V128 (residue numbering according to Raf RBD) and includes most of the positions with the lowest minima in entropy, apart from residues putatively affecting binding. The outermost layer (outer core: green) includes the residues L62, Q66, T68, V70, V72, C81, A85, R89L, L91, C96, R100, L112, W114, A118 and L121, located in the immediate periphery of the inner core. These residues showed generally below average entropy, although generally higher than the inner core ([23]). However, there are subtle distinctions in the arrangement of the inner core due to differences in register of the main α1 helix.

Indeed, the packing of α1 over the β-sheet is found to occur mainly following two arrangements across the β-grasp ubiquitin-like topology. For example, in Raf RBD, the inner core residues of α1 are at position i, i+4 and i+8, while in the case of ubiquitin the second and third residues are at position i+3 and i+7. Careful scrutiny of the entropy profile reveals a discrepancy at α1 hydrophobic core, which is annotated in the consensus sequence of the β-grasp ubiquitin-like topology (Figure 1 (b) and (c)) ([23]). The impact of this on the networks of contacts established by the hydrophobic core is schematized for Raf RBD and ubiquitin (Figure 4 (c) and (d)). This graph indicates that the contacts established in the hydrophobic core between residues located in the α-helix and principally L62, T68, V70,

V98 and L126 of the β-sheet vary across the ubiquitin-roll topology following the mode of packing of α1. For example, L62 is more intimately associated with the inner core in Raf RBD *versus* ubiquitin. The differences in arrangement are also confirmed by Φ-value analysis for Raf RBD and ubiquitin that reveal comparable involvement in TS stabilization of i+4 *versus* i+3 and i+8 *versus* i+7, respectively ([42])(**Article 4**). Therefore, the proposal of distinct α-helix over the β-sheet packing made initially on the alignment of β-grasp ubiquitin-like members are confirmed by the entropy profile, structural observation and folding kinetics.

The β-grasp ubiquitin-like topology is arranged into 12 superfamilies according to SCOP. The inner core arrangement of Raf RBD and more frequently ubiquitin occur most frequently in 5 superfamilies, which are said to be evolutionarily related to the ubiquitin superfamily. The packing adopted by ubiquitin is much more frequent in this group (Table 1). Structures classified in these 5 superfamilies represent half of the sequences aligned in Figure 2. Among the 6 other superfamilies, 4 adopt structures somewhat similar to ubiquitin-related superfamilies, but much more degenerate core packing arrangements (Materials and Methods). The evolutionary relationship between proteins displaying dissimilar or similar α1 packing is not clear and is not reflected in the SCOP classification. The comparison of contact maps for the hydrophobic core of Raf RBD and ubiquitin highlight the insights that this kind of scheme can bring to understanding the structural organization of proteins sharing similar topological structure and could be used as a method to establish evolutionary links between structural analogues and topologies ([76, 77]). In summary, the inner core includes the residues which are the most highly conserved both experimentally and in alignment of natural sequences.

The volume of the inner hydrophobic core is evolutionarily conserved

The observation of a common trend in the organization of the hydrophobic core contacts network spurred the analysis of the volume distribution in superfamilies of the β-grasp ubiquitin-like topology. To calculate the volume of side-chains, we used the volume of amino acids reported by F. M. Richards ([78]). Previously, we noticed that the inner core is

less tolerant than the outer core to volume variation. This was particularly true for 5 ubiquitin-related superfamilies (Materials and Methods), which showed an average cumulative volume of 594 ± 49 $Å^3$ in their inner core ([23]). Furthermore, if the volume observed in the diverse superfamilies integrated into the β-grasp ubiquitin-like alignment presented at Figure 2 is plotted, the aforementioned superfamilies constitute a discrete subgroup in which side chain volume follows a normal distribution (Figure 5). The inner core volume for these 5 superfamilies appears to be constrained and does not tolerate a variation in volume of greater than the equivalent of three methyl groups. Twelve of the structures in the other superfamilies, the most numerous examples being in the ferredoxin-like superfamily due to shorter major α-helix, lack one of the inner core residues. Nevertheless, even among the structures harboring 8 inner core residues, the volume requirements appear more diverse and particularly small amino acids, such as Ala or Thr occur more frequently. The significance of these observations in addressing the evolutionary relationships in the β-grasp ubiquitin-like would be worth exploring.

We next asked whether the conservation of cumulative volume is reflected in the amino acids requirements observed at the inner core residues in the more homogeneous ubiquitin-related superfamilies (Table 1). As could be predicted from the low sequence identity in the alignment, the sequence requirements are flexible, with most of the positions appearing equally constrained and only 4 of them (e.g., 60, 81/85, 98 and 128) displaying above 40% of selection for a given amino acid. It is noteworthy, that the amino acids present in the wt Raf RBD sequence are predominant at all positions, making it a good representative model of the average inner core composition. In summary, there appears to be a selection bias for specific cumulative volume in the inner hydrophobic core, although a variety of amino acids combinations are tolerated at every position of the inner core. Similarly, Gerstein *et al.* have observed in alignments of three protein families that the cumulative volume of the hydrophobic core is better conserved than the sequence identity or the volume accepted at specific residues ([79]). It was demonstrated that several positions in a very large sequence alignment of SH3 domains (i.e., 266 sequences ([12])) are highly constrained to amino acid type ([13]). This distinction might stem from the tighter

evolutionary relationships among proteins with similar biological function. Accordingly, the comparison of the amino acid selection obtained experimentally for Raf RBD (Figure 1 (c)) to those listed in Table 1 clearly demonstrate a more constrained variability of amino acid composition in the hydrophobic core. On the other hand, it is not clear that the β-grasp ubiquitin-like topology hydrophobic core as envisioned from conservation in the natural sequence alignment is particularly more flexible than the average natural topologies or if it is simply the result of subtle structure variation resulting from evolutionary drift. To start delineating the biophysical meaning of the hydrophobic core organization and the contribution of other conserved positions into the formation and stabilization of the β-grasp ubiquitin-like topology structure, we performed kinetic and thermodynamic studies on Raf RBD mutants.

Thermodynamic study on mutants of Raf RBD

On the basis of the sequence perturbation experiments and tertiary structure features of Raf RBD, 37 Ala/Gly mutations (e.g., residues are mutated to Ala, except Ala residues which are mutated to Gly) and 17 atypical mutations were introduced at selected position. Herein, we report the thermodynamic parameters for these 54 mutants (Table 2 and Materials and Methods). In Figure 6 (a), we show five representative urea melting curves obtained. Some of the findings that have been gathered using the thermodynamic data of this set are reported below. The kinetic parameters and a Φ-value analysis of the TS structural properties are reported in the companion manuscript (**Article 4**).

Two different estimates of $\Delta\Delta G_{F\text{-}U}$ were calculated (Material and Methods). The $\Delta\Delta G_{F\text{-}U}{}^{0M}$ and $\Delta\Delta G_{F\text{-}U}{}^{Cm}$ are generally comparable and indicate that the quality of the data is good. However, the $\Delta\Delta G_{F\text{-}U}{}^{Cm}$ is considered to be more accurate, because it necessitates less extrapolation for its calculation. Correlation between thermodynamic and kinetically derived $\Delta\Delta G_{F\text{-}U}$ and m suggest that the Raf RBD is a two-state folding protein and therefore that the two-state equation can be adequately applied (Table 2). Briefly, the most destabilizing mutations are concentrated in the hydrophobic core and the analysis of the

thermodynamic and kinetic parameters suggest that the native state structure of Raf RBD is unaltered by mutation despite strong destabilization (**Article 4**).

The relationship between sequence conservation and stability

The next question that we asked is whether the sequence conservation observed in the sequence perturbation experiment can be correlated either with the destabilization or reduction in folding rates induced by the Ala/Gly mutations. In order to do so, the positional entropy of Ala/Gly mutants was plotted against $\Delta\Delta G_{F\text{-}U}^{Cm}$ or $\ln k_f^{1.6M}$, respectively (Figure 6 (b) and (c)). Sequence entropy correlated best with stability, rather than folding rates as indicated by regression of the linear fits (r= 0.88 and r= 0.68, respectively). Only three of the 37 Ala/Gly mutants (e.g. P63A, T68A and C81A) were excluded from the graph to produce these correlations. In the case of T68A, the low entropy observed could have resulted from its involvement in binding to *ras*, the basis upon which sequences had been selected in the massive perturbation study. The strong correlation between entropy and stability is also consistent with the native structure being generally unaltered by mutations. It is noteworthy that the extrapolation of the linear fits to an entropy value of 1 (e.g., a theoretical case, in which a position would display absolutely no selective pressure, meaning in other words that all amino acids are equally well tolerated, including Ala) would yield a theoretical $\Delta\Delta G_{F\text{-}U}^{Cm}$ and $k_f^{1.6M}$ of -0.7 kJ/mole and 327 s^{-1}, respectively, therefore within measurement errors of wt Raf RBD values (e.g. 0 kJ/mole and 321 s^{-1}, respectively). The normalization of $\Delta\Delta G_{F\text{-}U}^{Cm}$ according to volume variation does not improve the correlation with sequence entropy. Moreover, even for the hydrophobic core residue mutations, stability and folding rate are better correlated with sequence entropy than with volume variation (data not shown). This is in agreement with results on the Fyn SH3 domain where the sensitivity of folding rate and stability to hydrophobic core mutations correlated with the conservation of the altered residues in a large sequence alignment of SH3 domains ([13, 27]), suggesting that the measure of sequence conservation accounts indirectly for sensitivity to side-chain volume variations induced upon mutation at a specific residue. Hence, sequence entropy is reliable, provided sufficient sequence diversity are reached in a dataset or databank, to define the importance of a given residue in the

stabilization of native structure, because it is a highly context specific parameter in comparison with volume variation, for example. Recently, a theoretical study using only thermodynamic stability as a selection constraint in the simulations was sufficient to generate artificial sequences similar to natural SH3 domains at 86% of positions ([47]). Interestingly, the correlation observed between the positional entropy and the level of destabilization induced by inserting Ala/Gly mutations demonstrate experimentally that a similar relationship exists for the Raf RBD, although a weaker correlation is also seen between entropy and folding rate. This is not very surprising given that most destabilizing mutants of Raf RBD also induced a significant reduction in folding rate (**Article 4**). In contrast, proteins with more polarized TS should display an equivalent correlation between entropy and stability, but reduced toward folding rate as the uncoupling between the thermodynamic and kinetic parameters would be higher in this case, hence confirming unambiguously the predominance of stability over folding rate in selection pressure. As previously reported, we observed no significant correlation between positional entropy and Φ-values, suggesting again that core residues are conserved to maintain stability of the native, rather than specifically the transition-state ensemble (data not shown) ([43, 45, 46]). In conclusion, our approach for obtaining sequence conservation constraints from segmental sequence perturbation is further validated by the correlation of sequence entropy with meaningful measures of biophysical characteristics for formation and stabilization of Raf RBD native structure. It has allowed for analysis of the sequence determinants for formation and stabilization of the native structure, which are not possible otherwise given the currently low number of known natural Raf RBD and the poor diversity of their sequences.

A map of the hydrophobic core role in the stabilization of the Raf RBD

Next, we compared the structural organization of the hydrophobic core *versus* the degree of destabilization induced by mutation of these residues to Ala/Gly (Figure 7). The most destabilizing mutations are principally located in the inner hydrophobic core, including those located in α1, β5, plus I58 and the outer core residue L62, both located in β1. Following this set of most important residues, there is a more disparate group that includes the inner core V60 and V98, A85 in α1 and a subset of residues of the outer core

(e.g. C96, L112 and L121) located mainly in the carboxy-terminal half of the domain. The β2 and remaining outer core residues dispersed on the structure play a more marginal role in the stabilization of the Raf RBD native structure. Overall, it is clear that the crucial determinant in stabilization of Raf RBD structure is located at the interface of the β-sheet and α-helix along an axis defined by the residues L78, L82 and L86, the surface of both of these topological elements forming the inner core. The similarity in entropy profile between the experimental data and the β-grasp ubiquitin-like topology (Figure 1 (b)) suggests that the role of residues, particularly of the hydrophobic core, in stabilization of this topology could be conserved. The structural distribution of stabilizing core residues in ubiquitin and its comparison to the results of Raf RBD presented herein is instructive. The distribution of stabilizing residues on ubiquitin reveals a more prominent role for β2 and β3 and a lesser role for β5 ([42]), which is broadly in agreement with the variation in the hydrophobic core organization (Figure 4).

The stability of Raf RBD is not optimal

The sequence perturbation experiment revealed a group of residues displaying high occurrence biases for non-wt amino acids. We sought to determine whether substitution of these residues into the wt sequence improved thermodynamic stability. In this study, 6 such variants were tested: N56M, I58L, S77T, C81I, C96L and C96M (Table 3). Only S77T showed clearly improved stability stemming mainly from lower k_u (Figure 8 (a) and Table 2). This mutation could possibly stabilize the structure by improving packing against the side-chain of close by N115. On the other hand, N56M had higher k_f and k_u and displayed only marginal stabilization. The I58L, C96M and C96L mutants showed similar stability to the wt or very minor destabilization. The fact that our strategy for screening libraries of degenerated sequence of Raf RBD *in vivo* with the DHFR PCA was sensitive to mutation disrupting the binding interface (see text above) ([23]), prompted us to evaluate whether some of these mutants could have improved binding affinity. Two mutants (e.g. C81I and C96M) close to the interface for binding to *ras* were selected for testing this hypothesis using an *in vitro* binding assay, but the results obtained did not confirm this hypothesis (data not shown). Alternatively, these 5 mutations might confer better behavior in *E. coli* cells during

selection. It is also foreseeable that these amino acids allowed more structural plasticity and can compensate better for destabilizing variation elsewhere in the perturbed segment.

However, three other mutations that stabilized Raf RBD were found. The mutant R89L that is known to disrupt the Raf RBD/*h-ras* complex (see text above) increases stability by approximately 3.8 $kJ.mol^{-1}$, mainly through improvement in folding rate. Mutant H2 recovered from the sequence perturbation experiment in the β-turn1 displayed a switch in turn type, which renders it more similar to the equivalent structure in ubiquitin (Figure 2 and Materials and Methods). This mutant showed a modest improvement in stability, mainly through an increase in the folding rate (**Article 4**). It is usually agreed that natural proteins are not optimized for stability, because this parameter is in competition with conservation of biological function. R89L, which is the most critical residue for binding to *h-ras* ([41]), is a paragon of this concept. *A posteriori,* the localization of this residue in the outer core, partly-buried and bridging α1 to β2 suggest that it could indeed be well accommodated by a hydrophobic amino acid. In addition, α1 appears to unwind partly in a central segment between C81 and A85 due to irregularities in the h-bond patterns in the crystal structure of the complex with Rap1A *versus* the monomeric RBD structure. It is noteworthy that A85 is the only α1 residue that tolerates Pro substitution in the sequence perturbation experiment ([23]). The unwinding of the α-helix as seen in the crystal would make R89 side-chain protrude further away from the protein interior. Therefore, we hypothesize that the suboptimal packing of the carboxy-terminal half of α1 might be necessary for formation of a stable complex with *ras*. More speculatively, data on the H2 mutant suggests similar sequence requirements for binding to *ras* in the β-turn1 and adjacent residues that compromise the stability of the Raf RBD. This region constitutes a second major linear epitope involved in the *h-ras* binding surface and the residues mutated in H2 and R89 are adjacent in Raf RBD structure (Figure 4).

The mutant Δ104-6 corresponds to deletion of residues E104-K106. This mutant was devised based on the observation that a-Raf and b-Raf lack these 3 amino acids present in the c-Raf RBD used in these studies. Strikingly, this alteration produced such improved

stability that the thermodynamic parameters could not be derived precisely from the urea melting curve due to the absence of a sufficiently long unfolded baseline. The Δ101-8 mutant was designed on the same premises based on comparison with ubiquitin sequences, which is shorter in this region, but was found to be destabilizing.

Next, starting from the Δ104-6 background, double- and triple-cycle mutants, integrating the other stabilizing mutations, were generated to determine whether they would improve thermodynamic stability of Raf RBD in an additive or non-additive way. Equilibrium and chevron curves were generated for the various mutants, using the strong denaturant Gdm-HCl (Figure 8 (b) and (c) and Table 3). Firstly, more precise thermodynamic and kinetic parameters for Δ104-6 were obtained using Gdm-HCl; this mutant is stabilized by approximately 6 $kJ.mol^{-1}$ compared to the wt RBD. The mechanism by which the Δ104-6 mutant could improve stability to such an extent is not obvious. The region comprising residues L102-K108 is relatively unstructured and is one of the most flexible region of the protein according to NMR data ([80]). This sequence bridges the β3 to the β4 and the deletion of residues E104-K106 could allow for the formation of a tighter turn in this region, resulting in a reduced entropic cost for loop closure and thus increased folding/decreased unfolding rates. It is nevertheless very surprising that this small deletion leads to an improvement in folding rate given the peripheral location of this segment relative to regions structured in the transition state (**Article 4**). The variation in loop size of L102-K108 suggests that the stability and kinetics for folding of c-Raf *versus* a-Raf/b-Raf might be significantly different. It would be interesting to explore how these differences could affect the normal and pathologic cellular functions that are fulfilled by the Raf genes ([38]). The double and triple mutants (e.g., Δ104-6/S77T, Δ104-6/S77T/H2 and Δ104-6/S77T/R89L) are even more stabilized and yielded maximum improvement in folding and unfolding rates of 12 fold and 16 fold, respectively. The mutant Δ104-6/S77T/R89L displays the most improved stability, with ΔG_{F-U}= -34.3 $kJ.mol^{-1}$, which is -12.5 $kJ.mol^{-1}$ lower than what is observed for the wt as determined in the same denaturant condition. This improvement in stability is quite remarkable given that only five residues are affected in this mutant. In addition, the effect of the mutations seems to be additive, suggesting that the

mutations are optimizing different details of the native structure. Finally, the capacity of Δ104-6, double and triple mutants to bind *ras* was tested in an *in vitro* pull-down assay and competition experiment (Figure 8 (d)). The Δ104-106 and Δ104-106/S77T retain a strong capacity to bind *ras*. As previously reported on wt Raf RBD, the insertion of H2 and R89L in the Δ104-106/S77T background reduces and abrogates binding to *ras*, respectively ([23, 41]). The specificity of the binding assay was confirmed by competition of the retention of *ras* on the resin bound Raf RBD mutants with untagged wt Raf RBD.

The level of stabilization obtained easily by combining data from the sequence perturbation experiment and literature knowledge indicates the non optimal character of Raf RBD sequence toward stability. Thus, there is little doubt that by using the data already accumulated in the sequence perturbation experiment and by other means, the Raf RBD stability could be even more optimized. In this regard, the other residues involved in the binding surface constitute probable sub-optimized positions. The degree of stabilization observed in *de novo* design experiments have indicated that the wt counterparts of designed proteins could be dramatically stabilized in the absence of selective pressure for function ([50]). For example, the most drastically stabilized protein in this study, the redesigned procarboxypeptidase domain showed 33 $kJ.mol^{-1}$ reduction in $\Delta G_{F-U,}$ which corresponds to a nearly threefold improvement in stability. The absence of functional constraints that could lead to sub-optimized structural arrangements was invoked as well to explain the very high thermal stability of *de novo* designed proteins ([51]).

Conservation of a polarized distribution of charged amino acids in Raf RBD

The Raf RBD is a very basic protein with an estimated pI of 8.9 [e.g., 12 basic (Lys and Arg) and 9 acidic (Asp and Glu) amino acids for a length of 78 residues]. A thorough mutagenesis study identified the most important residues for binding to *h-ras* on Raf RBD as R89, K84 and Q66, mostly in agreement with the crystal structure of the model based on Rap1A mutant ([40, 41]). In the sequence perturbation experiments, we retrieved a clear amino acid bias at expected residues including Q66, T68, K84, A85 and R89 ([23]). The structure of

the complex reveals also that the binding surface on Raf RBD includes several basic amino acids. In fact, the RBD structure displays basic and acidic patches that are segregated on opposite faces of the structure (Figure 10 (a) left and right panel, respectively), located roughly in the N- and C-terminal half of the domain, respectively. In addition, the Raf RBD displays several putative salt bridge pairs (e.g. R100-E124, K109-D129 and R59-E125) that might be important in the packing of the C-terminal half of the domain, encompassing β3-β5, and establishing long range contacts with the binding surface.

The distribution of charged residues on the surface of Raf RBD and the fact that several of these residues showed some level of conservation in our experiments prompted us to compare the distribution of charged amino acids in an alignment of Raf type RBD recovered from the (SMART) database and in the sequence perturbation experiment on c-Raf/Raf-1 (Figure 10 (b) and (c)). The comparison of both histograms reveals conservation of a polarized distribution of basic and acidic amino acids. Basic residues cluster in the binding surface located in the N-terminal half of the domain, e.g. in the stretch R67-S77, K84-R89, and around R100. On the other hand, acidic residues are favored in the β-turn1 (N64-K65), the amino terminal half of α1, in the loop constituted by the stretch G90-C95 immediately after α1 and generally in the C-terminal half of the domain. Based on studies that delineated the binding surface for *ras* on Raf RBD discussed above ([40, 41]), residues K84 (*) and R89 (**) were not varied in the main sequence perturbation experiment in order to maximize the number of clones recovered. These residues were degenerated in independent sequence perturbation experiments, which as expected yielded a very low number of binding competent clones. Nevertheless, this limited data indicated only average conservation of K84 for basic amino acids while position 89 was extremely intolerant to any other amino acid than arginine ([23]), in agreement with the mutagenesis data ([41]). In summary, charge polarization on the surface of Raf RBD was conserved in the sequence perturbation experiment. It is possible that this could be due to the segment-by-segment degeneracy approach that we have utilized. However, the fact that naturally evolved Raf-type RBDs display similar charged amino acid distributions suggests that these selections are not solely the result of the experimental methodology.

The diversity in the properties of protein-protein interfaces is very rich as observed in large sets of complexes ([81-83]). Generally, the side-chains that get buried upon complex formation are as highly packed as in the hydrophobic core. The involvement of charged amino acids at buried protein interfaces reflects the less penalizing entropic contribution to protein-protein interactions than in protein folding (the role of electrostatic forces in protein-protein interactions is reviewed in ([84])). It has been hypothesized that the higher polarity at the interfaces of regulated dimerizing proteins reflects the necessity of the protomers to be stable and soluble on their own under physiological conditions disfavoring complex formation ([82]). In fact, it is known that charged amino acids can act early in the association process (e.g. in the encounter complex) by grossly orienting and retaining together the colliding protein molecules through long range attracting and repulsing electrostatic interactions. In a key study on the importance of electrostatic interactions on the association of proteins Schreiber and Fersht suggested that increasing favorable electrostatic interactions between proteins-forming complex would accelerate association rates by favoring a less specific transition state ([85]). The moderate conservation of charges in the Raf RBD perturbation experiment might stem from this phenomenon and could be required to ensure fast association *in vivo*. In addition, the conserved segregation of basic and acidic amino acids on Raf RBD surface could allow for making the basic side-chains available for the intermolecular interaction with *ras* by avoiding intramolecular electrostatic interactions, while ensuring relatively neutral pI and repulsing forces in the case of non-productive encounters.

It is noteworthy that the CAD domains of CAD and ICAD, which are classified in the β-grasp ubiquitin-like (Figure 2; 1C9F and 1F2Ri, respectively), formed a heterodimer bearing a polarized charge distribution similar to Raf RBD on the structure surface, with the basic N-terminus of CAD and acidic C-terminus of ICAD forming the interface ([86]). The basic and acidic amino acids are also segregated in Ral GDS, but differently than in the previous cases, as can be observed from the crystallographic structure of the complex formed with *ras* ([87]). These observations suggest that the polarized charges distribution has

been retained by evolution in ubiquitin-like superfamilies members involved in protein-protein interactions.

The combinations of elegant theoretical work with experiment showed that some amino acids are coupled over a long range in protein structures ([88]). Piloted on the analysis of a large PDZ domain alignment, this method was used to determine the amount of evolutionary co-variation observed in an alignment by comparing the amino acid frequency variation at all positions in the presence or absence of amino acid constraint at a given residue (for example, residue "x" is coupled with "y", if when x is restricted to Leu, y is more frequently Tyr that when x is not constrained) relationship perform sequence coupling analysis ([88]). Specifically, it has been demonstrated that double mutant cycles of such coupled residues could synergistically affect the affinity of a PDZ domain for its substrate, despite the fact that some residues were located far from the ligand binding pocket. Other examples with more complex proteins, including membrane receptors and proteases were also reported ([89]). These data comforted the hypothesis that coupled residues define a path for energy distribution across protein structure, which could play a role not only in improving binding function, but also to transmit information intramolecularly, e.g. from one face of a protein structure to another, and as in the case of receptor from the cell surface to the cytoplasm. Putatively such mechanism could be broadly implicated in protein conformational changes that follow ligand binding or in post-translational modification, in which they could participate in fine tuning the stability of an oligomeric state and the efficiency of proteins as biological machines and switches in signaling cascade, respectively. The polarized charge distribution on Raf RBD provides a tantalizing means of transmitting energy across the protein surface upon binding of *ras* at the basic binding patch. Indeed, the change in the net charge or redeployment of electrostatic interaction of Raf RBD upon binding *ras* would provide a means to do so. However, as exemplified by the studies mentioned above, less obvious residue connections might be involved in that process.

Conclusions

The analysis of the sequence perturbation data described above have revealed that there are significant similarities in the local propensities for α-helix and β-strand between the mutated Raf RBD and an alignment of proteins sharing the ubiquitin-roll topology. Some of the discrepancies in the comparison can be attributed to variation in the packing of the hydrophobic core, specifically due to different α1 arrangement over the β-sheet. Next, the determination of the thermodynamic stability and folding rate of numerous variants of Raf RBD indicates a stronger relationship of the former with sequence entropy. The Raf RBD hydrophobic core was previously described to be composed of two concentric layers, the inner and outer core ([23]). The mutation of inner core residue was shown to have the most dramatic impact on thermodynamic stability and also transition state stabilization, while the mutation of outer core residues had less predictable effects on thermodynamic stability and folding kinetics (Figure 9 and **Article 4**). The correlation in the entropy profiles in the inner core residues (Figure 1) and the conservation of their structural organization (Figure 4) and of the cumulative volume of their side-chains in the ubiquitin-related superfamilies (Figure 5) argues for similar relationships, while the other superfamilies seem to have different properties.

We also present evidence suggesting that the polarization of charges at the surface of Raf RBD is conserved and that this may occur for functional reasons as it is conserved across natural Raf RBDs. So far, the determinants of Raf RBD structure and function have been tackled by a classical approach in which the residues are mostly treated as independent entities ([22, 23, 41, 90]). A new strategy called sequence coupling analysis has been devised to go further than the position-by-position assessment of sequence diversity across sequence alignments. As described above, this approach can allow for determining residues connected energetically across protein structures and that are impossible to detect otherwise. Recently, a protein design scheme building on the sequence coupling analysis approach and sequence information of a large WW domain alignment (> 100 residues) succeeded in producing unnatural WW domains sharing the characteristics of various natural counterparts, while the scheme that considered residues as independent entities failed ([48, 49]). The paucity of natural Raf RBD sequences has precluded the utilization of the sequence coupling analysis strategy. The sequence perturbation experiment could be used to perform

co-variation experiment, but it would be a daunting experimental challenge to do this without an a priori determination of potentially linked positions. In the case of Raf RBD, the already accumulated data might help to restrain the search space. Nonetheless, in a more general perspective, it would be worth designing new *in silico* or experimental strategies to complement sequence coupling analysis and allow for its application to more protein families.

Finally, the combination of a few stabilizing mutation in Raf RBD indicated that its stability could be dramatically improved by mutation at only 5 residues, leading approximately to a 150% increase in $\Delta G_{F\text{-}U}$. Also, the stabilization induced by mutation of R89, the major residue necessary for binding *ras*, is in agreement with the hypothesis of an evolutionary non-optimized stability due to the compromises necessary to binding function (e.g. more generally, any function). This is a very interesting observation as it highlights the insights that integrating the study of the biophysical characteristics of a protein or domain in the perspective of tackling its biologically meaningful roles could bring to understanding cell signaling and function. Indeed, the suboptimal thermodynamic stability in protein structures, from locally disordered regions up to fully disordered proteins, might represent a mean to expand functions and/or capacities to modulate it by creating alternative protein-protein interaction surface or by accommodating post-translational modification sites ([91]). Furthermore, noel type of allosteric sites in signaling proteins or enzymes, such as those described for the phospho-tyrosine phosphatase PTP1B and the Wiskott-Aldrich syndrome protein (N-WASP) ([92, 93]), might be potentially generated in destabilizing loops, unstructured or thermodynamically sub-optimized segments or take advantage of structural remodeling in these elements to perform its regulatory activity.

Materials and methods

Description of the β-grasp ubiquitin-like alignment

The alignment was constructed from sequences recovered in SCOP and FSSP database. At the time this alignment was constructed there was 11 superfamilies in the β−grasp ubiquitin-like according to SCOP. The members of 9 superfamilies were integrated

in the alignment. Raf RBD belongs to the ubiquitin-like superfamily, which is probably evolutionarily related to 4 other superfamilies, including CAD/PB1, MoaD/ThiS, TGS-like and Double-Cortin sequences. Recently a twelfth superfamily, TmoB-like, with only one member was added. The sequence was analyzed *a posteriori*, particularly at the level of hydrophobic core and was found to match with the observations described for the 5 ubiquitin related superfamilies in the SCOP database (http://scop.mrc-lmb.cam.ac.uk/scop). The 4 other superfamilies in the alignment are ferredoxin-like, staphylokinase/streptokinase, superantigen toxins and IgG binding domain. Prototypes of two other superfamilies are too degenerate to be integrated in the alignment. Other proteins in the alignment were recovered from the FSSP database, but classified in other topology by SCOP database. The highest sequence identity between any two sequences allowed is 35%. Any sequences that were above this threshold were not included in the alignment. Overall, the mean pair-wise identity across the alignment is 9.4% in the alignment of all superfamily sequences and 10.3% for ubiquitin-like and the 4 ubiquitin-related superfamilies sequences (more details concerning the alignment can be found in ([23])).

Entropy calculation

Sequence entropy was calculated following a modified version of Shannon entropy formula ([94]), using experimental data and the β-grasp ubiquitin-like alignment (Fig. 2) ([23]).

Interactions network in the hydrophobic core and inner core volume in the β-grasp ubiquitin-like

Residues of the hydrophobic core having at least one side-chain atom involved in direct Van der Waals contacts were determined by manual observation of ubiquitin (1UBI) and Raf RBD (1RFA) structure and were linked through their C_β atoms (Figure 3). The delineation of inner and outer core residues was done as described in the text. The cumulative volume of residue side-chains in the inner core was calculated using estimates of their volumes by Richards (Figure 4) ([78]).

Residue-by-residue variation of volume in the hydrophobic core, mean propensity and dispersion of acidic/basic amino acid occurrences

The normalized amino acid occurrences at each position varied in the experimental data and the ubiquitin-roll ([23])were used to average propensity and occurrence of charged (acidic or basic) amino acids residue-by-residue (Figure 5-7). The secondary structure propensity scale is taken from Koehl and Levitt ([64]). The average propensities are calculated straightforwardly from the normalized amino acid occurrences (Figure 6). The normalized occurrences for acidic/basic amino acids is directly plotted (Figure 7).

Comparing secondary structure prediction of clones with wt structure

Secondary structures of all sequence variants isolated through the sequence perturbation strategy were predicted using three prediction algorithms: PhD, PROF and PSI-PRED. Each of prediction algorithms was run with their default parameters. The wt secondary structure was also predicted using these three methods. Next, for each wt position we calculated the mean percentage of sequence variants for which secondary structure motif (helix, strand or loop) was predicted to be the same as predicted for wt sequence at a given position. We also calculated the mean percentage of sequence variants predicted to adopt α-helix conformation in the region C95-L102 and counted the ratio of these variants that displayed at least three consecutive residues in that type of secondary structure.

Mutants description, cloning and purification

Mutants of human Raf-1/c-Raf RBD were synthesized with a variant of the ExSite™ protocol (Stratagene) using the high-fidelity Pfu polymerase. Most variants synthesized carry a single point mutation. The mutant H2 was recovered from the sequence perturbation experiment. In this mutant, residues 62-65 (Leu-Pro-Asn-Lys) of Raf RBD are replaced with the amino acid Phe-Thr-Asp-Gly. Mutant H2_F62L revert residue 62 to the wt amino acid (Leu-Thr-Asp-Gly). Mutant Δ104-6 and Δ101-8+AG are deletion mutants. In the former case amino acid 104 to 106 are deleted (Glu-His-Lys). In the latter case, amino acid 101 to 108 are replaced by Ala-Gly, as in ubiquitin. The mutation insertion was confirmed by sequencing. The protein expressed included residue 55-132 of Raf-1 plus an amino-terminal located hexahistidine tag separated by a spacer of 3 amino acids (Ser-Met-Gly). Proteins were purified from bacterial cell lysate under denaturing conditions using urea, on a Ni-NTA column.

Stability

The thermodynamic parameters were calculated from denaturant induced melting curves obtain from the endpoint fluorescence (raw fluorescence at 10s) of the unfolding traces obtained on Applied Photophysics SX.18MV stopped-flow fluorimeter (**Article 4** and below). All experiments were done at 25°C, in urea and 50mM sodium phosphate buffer at pH=7.0. The data were converted to fraction of folded protein and fit to a two-state model. Most mutants displayed minor error between the kinetic and thermodynamic estimates of m (< 20%), stemming principally from error in the baseline. In cases with error > 20%, the thermodynamic parameters were recalculated from fluorescence melting curves perform on a Varian Eclipse spectrofluorimeter. In most cases, the discrepancies were resolved and were attributed to obvious deviations in the baselines. Note that for the mutants displaying high resistance to urea induced unfolding (Δ104-106 and double and triple-cycle mutants), the thermodynamic parameters were calculated from the endpoint fluorescence of the unfolding traces in Gdm-HCl 50 mM sodium phosphate buffer, pH=7.0.

Thermodynamic parameters

$\Delta\Delta G_{F\text{-}U}{}^{Cm}$ was calculated using a method described previously ([25]):

$$\Delta\Delta G_{F-U}^{Cm} = \langle m \rangle \left(Cm^{mut} - Cm^{wt} \right) \quad (1)$$

where <m> is the average m value for all the mutants (3.90 (±0.33) kj mol^{-1} M^{-1}), Cm^{wt} and Cm^{mut} are the concentration of urea at which 50% of wt and mutant proteins are folded.

Kinetics and chevron curves

The kinetics experiments were performed as described (**Article 4**). The chevron curves for stabilized mutants (Δ104-6 and double and triple cycle mutants that derived) were done under the same conditions, using Gdm-HCl as denaturant. The refolding reaction was initiated from proteins diluted in ~ 6.25 M Gdm-HCl. The unfolding reaction was initiated from proteins diluted in ~ 0.5 M Gdn-HCl.

Ni-NTA pull-down of *ras* bound to non-hydrolysable GTP analogs using Raf RBD

Ni-NTA pull-down of GST ras bound to GMP-PNP with tagged Raf RBD mutants and competition with wt untagged Raf RBD was done using the same protocol as previously reported ([23]).

Structure representation

Raf RBD and ubiquitin structural representation were rendered from the atomic coordinates 1RFA and 1GUA (for figure 10 (c)) and 1UBI, respectively, of the protein databank with the molecular graphics packages MolMol and Weblab viewer (Acelrys) software.

Acknowledgements

The authors thank: Dr Jeffrey W. Keillor for providing access to stopped-flow and fluorimeter apparatus as well as for collaboration. Dr Luc Desgroseillers and member of his lab for access to electroporating device. Alexis Vallée-Belisle for discussions and data exchange. The NSERC and CIHR funded this project. FXCV is a scholar of CIHR, le programme de biologie moléculaire and the FES. SWM is the Canada Research Chair in Integrative Genomics.

References

1. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). Structure and function of haemoglobin. *J. Mol. Biol.* **13**, 669-678.
2. Rossmann, M. G. & Argos, P. (1976). Exploring structural homology of proteins. *J. Mol. Biol.* **105**, 75-95.
3. Richardson, J. S. (1977). beta-Sheet topology and the relatedness of proteins. *Nature* **268**, 495-500.
4. Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34**, 167-339.
5. Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225-270.
6. Anfinsen, C. B., Redfield, R. R., Choate, W. L., Page, J. & Carroll, W. R. (1954). Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *J. Biol. Chem.* **207**, 201-210.
7. Anfinsen, C. B. & Scheraga, H. A. (1975). Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.* **29**, 205-300.
8. Mirny, L. A., Abkevich, V. I. & Shakhnovich, E. I. (1998). How evolution makes proteins fold quickly. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4976-4981.
9. Michnick, S. W. & Shakhnovich, E. (1998). A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Fold. Des* **3**, 239-251.
10. Chothia, C., Gelfand, I. & Kister, A. (1998). Structural determinants in the sequences of immunoglobulin variable domain. *J. Mol. Biol.* **278**, 457-479.

11. Hill, E. E., Morea, V. & Chothia, C. (2002). Sequence conservation in families whose members have little or no sequence similarity: the four-helical cytokines and cytochromes. *J. Mol. Biol.* **322**, 205-233.
12. Larson, S. M., Di Nardo, A. A. & Davidson, A. R. (2000). Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J. Mol. Biol.* **303**, 433-446.
13. Di Nardo, A. A., Larson, S. M. & Davidson, A. R. (2003). The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core. *J. Mol. Biol.* **333**, 641-655.
14. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.* **6**, 1005-1009.
15. Kragelund, B. B., Hojrup, P., Jensen, M. S., Schjerling, C. K., Juul, E., Knudsen, J. & Poulsen, F. M. (1996). Fast and one-step folding of closely and distantly related homologous proteins of a four-helix bundle family. *J. Mol. Biol.* **256**, 187-200.
16. Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M. A., Jaenicke, R. & Schmid, F. X. (1998). Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nat. Struct. Biol.* **5**, 229-235.
17. McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of beta-hairpin formation in protein G folding. *Nat. Struct. Biol.* **7**, 669-673.
18. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nat. Struct. Biol.* **6**, 1016-1024.
19. Martinez, J. C. & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat. Struct. Biol.* **6**, 1010-1016.
20. Guerois, R. & Serrano, L. (2000). The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304**, 967-982.
21. Friel, C. T., Capaldi, A. P. & Radford, S. E. (2003). Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: similarities and differences in the folding of homologous proteins. *J. Mol. Biol.* **326**, 293-305.
22. Vallee-Belisle, A., Turcotte, J. F. & Michnick, S. W. (2004). raf RBD and Ubiquitin Proteins Share Similar Folds, Folding Rates and Mechanisms Despite Having Unrelated Amino Acid Sequences. *Biochemistry* **43**, 8447-8458.
23. Campbell-Valois, F. X., Tarassov, K. & Michnick, S. W. (2005). Massive Sequence Perturbation of a Small Protein. *Proc. Natl Acad. Sci. U.S.A.* **102**, 14988-14993.
24. Ternstrom, T., Mayor, U., Akke, M. & Oliveberg, M. (1999). From snapshot to movie: phi analysis of protein folding transition states taken one step further. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 14854-14859.
25. Kim, D. E., Fisher, C. & Baker, D. (2000). A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298**, 971-984.
26. Zarrine-Afsar, A., Larson, S. M. & Davidson, A. R. (2005). The family feud: do proteins with similar structures fold via the same pathway? *Curr. Opin. Struct. Biol.* **15**, 42-49.
27. Northey, J. G., Di Nardo, A. A. & Davidson, A. R. (2002). Hydrophobic core packing in the SH3 domain folding transition state. *Nat. Struct. Biol.* **9**, 126-130.
28. Soding, J. & Lupas, A. N. (2003). More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* **25**, 837-846.
    Ref Type: Journal
29. Service, R. (2005). Structural biology. Structural genomics, round 2. *Science* **307**, 1554-1558.
30. Service, R. (2005). Structural biology. A dearth of new folds. *Science* **307**, 1555.
31. Reidhaar-Olson, J. F. & Sauer, R. T. (1988). Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. *Science* **241**, 53-57.
32. Lim, W. A. & Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature* **339**, 31-36.
33. Bowie, J. U. & Sauer, R. T. (1989). Identifying determinants of folding and activity for a protein of unknown structure. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2152-2156.

34. Axe, D. D., Foster, N. W. & Fersht, A. R. (1996). Active barnase variants with completely random hydrophobic cores. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 5590-5594.
35. Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. & Baker, D. (1997). Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4**, 805-809.
36. Kim, D. E., Gu, H. & Baker, D. (1998). The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4982-4986.
37. Finucane, M. D. & Woolfson, D. N. (1999). Core-directed protein design. II. Rescue of a multiply mutated and destabilized variant of ubiquitin. *Biochemistry* **38**, 11613-11623.
38. Wellbrock, C., Karasarides, M. & Marais, R. (2004). The RAF proteins take centre stage. *Nat. Rev. Mol. Cell Biol.* **5**, 875-885.
39. Nassar, N., Horn, G., Herrmann, C., Scherer, A., McCormick, F. & Wittinghofer, A. (1995). The 2.2 A crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with Rap1A and a GTP analogue. *Nature* **375**, 554-560.
40. Nassar, N., Horn, G., Herrmann, C., Block, C., Janknecht, R. & Wittinghofer, A. (1996). Ras/Rap effector specificity determined by charge reversal. *Nat. Struct. Biol.* **3**, 723-729.
41. Block, C., Janknecht, R., Herrmann, C., Nassar, N. & Wittinghofer, A. (1996). Quantitative structure-activity analysis correlating Ras/Raf interaction in vitro to Raf activation in vivo. *Nat. Struct. Biol.* **3**, 244-251.
42. Went, H. M. & Jackson, S. E. (2005). Ubiquitin folds through a highly polarized transition state. *Protein Eng Des Sel* **18**, 229-237.
43. Plaxco, K. W., Larson, S., Ruczinski, I., Riddle, D. S., Thayer, E. C., Buchwitz, B., Davidson, A. R. & Baker, D. (2000). Evolutionary conservation in protein folding kinetics. *J. Mol. Biol.* **298**, 303-312.
44. Mirny, L. & Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* **308**, 123-129.
45. Larson, S. M., Ruczinski, I., Davidson, A. R., Baker, D. & Plaxco, K. W. (2002). Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *J. Mol. Biol.* **316**, 225-233.
46. Tseng, Y. Y. & Liang, J. (2004). Are residues in a protein folding nucleus evolutionarily conserved? *J. Mol. Biol.* **335**, 869-880.
47. Larson, S. M. & Pande, V. S. (2003). Sequence optimization for native state stability determines the evolution and folding kinetics of a small protein. *J. Mol. Biol.* **332**, 275-286.
48. Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H. & Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. *Nature* **437**, 512-518.
49. Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B. & Ranganathan, R. (2005). Natural-like function in artificial WW domains. *Nature* **437**, 579-583.
50. Dantas, G., Kuhlman, B., Callender, D., Wong, M. & Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* **332**, 449-460.
51. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-1368.
52. Bond, C. J., Wong, K. B., Clarke, J., Fersht, A. R. & Daggett, V. (1997). Characterization of residual structure in the thermally denatured state of barnase by simulation and experiment: description of the folding pathway. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 13409-13413.
53. Wong, K. B., Clarke, J., Bond, C. J., Neira, J. L., Freund, S. M., Fersht, A. R. & Daggett, V. (2000). Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding. *J. Mol. Biol.* **296**, 1257-1282.
54. Kazmirski, S. L., Wong, K. B., Freund, S. M., Tan, Y. J., Fersht, A. R. & Daggett, V. (2001). Protein folding from a highly disordered denatured state: the folding pathway of chymotrypsin inhibitor 2 at atomic resolution. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4349-4354.
55. Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T., Imoto, T., Smith, L. J., Dobson, C. M. & Schwalbe, H. (2002). Long-range interactions within a nonnative protein. *Science* **295**, 1719-1722.

56. Religa, T. L., Markson, J. S., Mayor, U., Freund, S. M. & Fersht, A. R. (2005). Solution structure of a protein denatured state and folding intermediate. *Nature* **437**, 1053-1056.
57. Ptitsyn, O. B. (1987). Protein Folding: Hypothesis and Experiments. *J. Prot. Chem.* **6**, 273-293.
58. Baldwin, R. L. & Rose, G. D. (1999). Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* **24**, 26-33.
59. Przytycka, T., Aurora, R. & Rose, G. D. (1999). A protein taxonomy based on secondary structure. *Nat. Struct. Biol.* **6**, 672-682.
60. Srinivasan, R. & Rose, G. D. (1999). A physical basis for protein secondary structure. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 14258-14263.
61. Gong, H., Isom, D. G., Srinivasan, R. & Rose, G. D. (2003). Local secondary structure content predicts folding rates for simple, two-state proteins. *J. Mol. Biol.* **327**, 1149-1154.
62. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209-225.
63. Simons, K. T., Strauss, C. & Baker, D. (2001). Prospects for ab initio protein structural genomics. *J. Mol. Biol.* **306**, 1191-1199.
64. Koehl, P. & Levitt, M. (1999). Structure-based conformational preferences of amino acids. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 12524-12529.
65. Searle, M. S., Williams, D. H. & Packman, L. C. (1995). A short linear peptide derived from the N-terminal sequence of ubiquitin folds into a water-stable non-native beta-hairpin. *Nat. Struct. Biol.* **2**, 999-1006.
66. Zerella, R., Evans, P. A., Ionides, J. M., Packman, L. C., Trotter, B. W., Mackay, J. P. & Williams, D. H. (1999). Autonomous folding of a peptide corresponding to the N-terminal beta-hairpin from ubiquitin. *Protein Sci.* **8**, 1320-1331.
67. Harding, M. M., Williams, D. H. & Woolfson, D. N. (1991). Characterization of a partially denatured state of a protein by two-dimensional NMR: reduction of the hydrophobic interactions in ubiquitin. *Biochemistry* **30**, 3120-3128.
68. Briggs, M. S. & Roder, H. (1992). Early hydrogen-bonding events in the folding reaction of ubiquitin. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 2017-2021.
69. Cox, J. P., Evans, P. A., Packman, L. C., Williams, D. H. & Woolfson, D. N. (1993). Dissecting the structure of a partially folded protein. Circular dichroism and nuclear magnetic resonance studies of peptides from ubiquitin. *J. Mol. Biol.* **234**, 483-492.
70. Stockman, B. J., Euvrard, A. & Scahill, T. A. (1993). Heteronuclear three-dimensional NMR spectroscopy of a partially denatured protein: the A-state of human ubiquitin. *J. Biomol. NMR* **3**, 285-296.
71. Kitahara, R., Yamada, H. & Akasaka, K. (2001). Two folded conformers of ubiquitin revealed by high-pressure NMR. *Biochemistry* **40**, 13556-13563.
72. Kitahara, R. & Akasaka, K. (2003). Close identity of a pressure-stabilized intermediate with a kinetic intermediate in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3167-3172.
73. Varadan, R., Walker, O., Pickart, C. & Fushman, D. (2002). Structural properties of polyubiquitin chains in solution. *J. Mol. Biol.* **324**, 637-647.
74. Carrion-Vazquez, M., Li, H., Lu, H., Marszalek, P. E., Oberhauser, A. F. & Fernandez, J. M. (2003). The mechanical stability of ubiquitin is linkage dependent. *Nat. Struct. Biol.* **10**, 738-743.
75. Grishin, N. V. (2001). Fold change in evolution of protein structures. *J. Struct. Biol.* **134**, 167-185.
76. Kannan, N. & Vishveshwara, S. (1999). Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol.* **292**, 441-464.
77. Lindorff-Larsen, K., Rogen, P., Paci, E., Vendruscolo, M. & Dobson, C. M. (2005). Protein folding and the organization of the protein topology universe. *Trends Biochem. Sci.* **30**, 13-19.
78. Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1-14.
79. Gerstein, M., Sonnhammer, E. L. & Chothia, C. (1994). Volume changes in protein evolution. *J. Mol. Biol.* **236**, 1067-1078.
80. Emerson, S. D., Madison, V. S., Palermo, R. E., Waugh, D. S., Scheffler, J. E., Tsao, K. L., Kiefer, S. E., Liu, S. P. & Fry, D. C. (1995). Solution structure of the Ras-binding domain of c-Raf-1 and identification of its Ras interaction surface. *Biochemistry* **34**, 6911-6918.

81. Lo, C. L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177-2198.
82. Nooren, I. M. & Thornton, J. M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. *J. Mol. Biol.* **325**, 991-1018.
83. Shaul, Y. & Schreiber, G. (2005). Exploring the charge space of protein-protein association: a proteomic study. *Proteins* **60**, 341-352.
84. Sheinerman, F. B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153-159.
85. Schreiber, G. & Fersht, A. R. (1996). Rapid, electrostatically assisted association of proteins. *Nat. Struct. Biol.* **3**, 427-431.
86. Otomo, T., Sakahira, H., Uegaki, K., Nagata, S. & Yamazaki, T. (2000). Structure of the heterodimeric complex between CAD domains of CAD and ICAD. *Nat. Struct. Biol.* **7**, 658-662.
87. Huang, L., Hofer, F., Martin, G. S. & Kim, S. H. (1998). Structural basis for the interaction of Ras with RalGDS. *Nat. Struct. Biol.* **5**, 422-426.
88. Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295-299.
89. Suel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* **10**, 59-69.
90. Fridman, M., Maruta, H., Gonez, J., Walker, F., Treutlein, H., Zeng, J. & Burgess, A. (2000). Point mutants of c-raf-1 RBD with elevated binding to v-Ha-Ras. *J. Biol. Chem.* **275**, 30363-30371.
91. Dyson, H. J. & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197-208.
92. Wiesmann, C., Barr, K. J., Kung, J., Zhu, J., Erlanson, D. A., Shen, W., Fahr, B. J., Zhong, M., Taylor, L., Randal, M., McDowell, R. S. & Hansen, S. K. (2004). Allosteric inhibition of protein tyrosine phosphatase 1B. *Nat. Struct. Mol. Biol.* **11**, 730-737.
93. Peterson, J. R., Bickford, L. C., Morgan, D., Kim, A. S., Ouerfelli, O., Kirschner, M. W. & Rosen, M. K. (2004). Chemical inhibition of N-WASP by stabilization of a native autoinhibited conformation. *Nat. Struct. Mol. Biol.* **11**, 747-755.
94. Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**, 56-68.

Figure legends, Figures and Tables:

Figure 1. The Raf RBD model and insights obtained from sequence perturbation experiment. (a) Tertiary structure of Raf RBD. (b) Comparison of normalized entropy of the Raf RBD obtained experimentally *versus* 54 naturally occurring proteins displaying the ubiquitin-roll topology. (c) Primary and secondary structure of Raf RBD. Below the secondary structure is indicated the signature sequence of the Raf RBD, depicted as a serial group of amino acid bias derived from the sequence perturbation experiment (Exp). The positions in bold have very low entropy both in the experiment and sequence alignments of proteins with the ubiquitin-roll topology. The positions in regular and italic lettering indicate respectively average entropy or conservation specific to Raf RBD. The consensus position within the ubiquitin-roll topology (Ubi-topo) are also indicated (h: hydrophobic; c: helix capping; capital letters indicate higher conservation). Positions marked by stars in the major α-helix indicate discrepancy between the consensus Exp and Ubi-topo due to variation in its arrangement over the β-sheet.

Figure 2. Alignment of 54 proteins or domains adopting the ubiquitin-roll topology recovered from structural databases. The alignment was done manually, according to overlap of common secondary structural elements. The 27 first structures starting at the top of the alignment are classified in the ubiquitin-like and the ubiquitin-related superfamilies (Materials and Methods). Residue numbering follows the Raf RBD sequence. An insertion that occurs in the Raf RBD between positions 101-109 was removed from the alignment.

Figure 3. Segmental secondary structure conservation deduced from the sequence perturbation experiments: propensities and secondary structure predictions. (a) Mean propensities for α-helix and β-strand observed at all positions in the experiment and in the β-grasp ubiquitin-like alignment are shown in the top and bottom panel, respectively. Note that residues Q66, K84, R89 and W114 were not degenerated in the main experiment. Nevertheless, K84 (*) and R89 (**) showed very good conservation for amino acids with high propensity for α-helix in a separate perturbation experiment. In the case of position 89, conservation for Arg was observed in all mutants obtained. (b) The mean percentage of wt secondary structure that is conserved, according to secondary structure prediction algorithms, at each position in Raf RBD variants obtained experimentally (■). Among the core elements of the structure, only a segment corresponding to β3 (e.g., segmental library 8 ([23])) showed significant reduction in wt secondary structure prediction. In this region a low but significant percentage of positions are predicted to switch to α-helical conformation (■). Further analysis of the predictions showed that 6.2±2.8% of clones for the region C95-L102 had at least 4 consecutive residues in α-helical conformation (data not shown). Secondary structure predictions were performed using the sequence of each segmental variant of the 13 libraries in the context of the wt sequence of Raf RBD as queries for submission to three secondary structure prediction algorithms (e.g., PSIPRED, PHD and PROF) and the standard deviation to the mean percentage were plotted. The gapped positions indicate the unperturbed residues mentioned in the legend for panel (a).

Figure 4. Core structural organization and side-chain contact networks; ubiquitin and Raf RBD as prototype of differing structural arrangement in the ubiquitin-roll topology. The hydrophobic core of the Raf RBD is organized into two layers defined as inner and outer core, readily apparent in the sequence perturbation experiment. (a) These positions are shown on Raf RBD tertiary structure (1RFA), respectively in red and green. (b) The homologous positions in ubiquitin are displayed on ubiquitin tertiary structure (1UBI). The direct Van der Waals contacts between any two side-chains participating in the inner and outer core residues are shown by connecting C_β of residues involved for (c) Raf RBD and (d) Ubiquitin. These proteins are constituted into two surfaces, grossly defined as α1 and the β-sheet. The residues in loops are classified according to the surface in which they are integrated. The network contacts are represented in the same orientation than in the first panels. The C_β for residues located in the α1 layer are represented by bigger spheres than the β-sheet. Thick lines connect residues part of the same secondary element (black: β-strand; grey: α-helix). In the case of α1, the residues on the same side of the helix are connected by a thick line. The thinner lines connect residues whose side-chains are in contact, whether the contact is intra-surface (full lines) or inter-surface (dashed lines). The inner core network is shown in isolation in the bottom panel with the numbering identifying inner core residues 2 and 3 of α1. Note the variation in the arrangement of α1 over the β-sheet in Raf RBD *versus* ubiquitin.

Figure 5. Variation of inner core residue side-chain cumulative volume in the ubiquitin superfold according to superfamily classification. Inner core (I58, V60, L78, L82, L86, V98, L126 and V128) total volume for each structure (54) included in the β-grasp ubiquitin-like topology. The volume distribution is classified according to each superfamily included in the alignment. The ubiquitin-related superfamilies are annotated with black and white columns. Note that some superfamilies, particularly the highly populated ferredoxin-like lack residue L86, which could explain partly the reduction in volume. The variation in α1 packing was considered to decide which residues participate in the inner core (Figure 4 and ([23])).

Figure 6. Denaturant induced melting curves of Ala/Gly mutants of Raf RBD and the relationship between stability, folding rate and entropy. (a) Urea induced melting curves of selected mutants of Raf RBD: V60A (□), V72A (○), S77T (◊), V98A (Δ) and V128A (∇). Idealized Wt melting curve is shown for reference (grey line). (b) Plot of positional sequence entropy *versus* $\Delta\Delta G_{F\text{-}U}^{Cm}$ induced by Ala mutation of non-Ala residue and by Gly mutation for Ala residue. The linear regression is significant (r= 0.88). (c) Plot of positional sequence entropy *versus* $\ln k_f^{1.6M}$. In this case, the correlation is weaker (r= 0.68). Note that three Ala mutations that deviated significantly were not plotted in this graph (P63A, T68A and C81A).

Figure 7. Map of the stabilizing hydrophobic core residues in Raf RBD structure. Comparison of the hydrophobic core organization determined from the sequence perturbation experiment and sequence alignments of proteins sharing the ubiquitin-roll topology (see inner core: red; outer core: green; up panel) *versus* the destabilization induced by Ala/Gly mutation of these residues (all residues mutated to Ala, but Gly mutation for Ala residues; middle panel). The residues are colored following the ratio $\Delta\Delta G_{F\text{-}U}^{Cm}/\Delta G_{F\text{-}U}$ (0-0.25: grey; 0.25-0.35: blue; 0.35-0.45: green; 0.45-0.55: orange; 0.55-0.65: red). R89 and W114 are not shown because they were not mutated for Ala/Gly. A cartoon representation of the Raf RBD is presented for reference (bottom panel).

Figure 8. Characterization of stabilized mutants of Raf RBD. (a) Urea induced melting curves of stabilized mutants of Raf RBD: H2 (□), S77T (○), R89L (◊) and Δ104-6 (Δ). (b) Gdm-HCl induced melting curves of Δ104-6 and double and triple mutants generated from it: Δ104-6 (□), Δ104-6/S77T (○), Δ104-6/S77T/H2 (◊) and Δ104-6/S77T/R89L (Δ). (c) Chevron curves for Δ104-6, double and triple mutants (legend is the same than in the previous panel). Modeled wt melting and chevron curves are shown in the corresponding panels to serve as reference (grey line). (d) Ni-NTA pull-down of GST-ras bound to GMP-PNP using the stabilized variants of his-tag Raf RBD and competition with untagged wt Raf RBD. The proteins were revealed by coomassie blue staining. The picture also shows that the amount of loaded Raf RBD is similar in each lane.

Figure 9. Conservation of the polarized charges distribution at the surface of the Raf RBD. (a) Basic (K, R: blue) and acidic patch (D, E: red) on c-Raf/Raf-1 RBD structure in CPK representation are shown on the left and right panel, respectively. The surfaces correspond roughly to the plane of the page and are identical image obtained by a 180º rotation around the Y axis. The basic patch corresponds to the GTPase binding surface as indicated by the interaction of a section of Rap1A polypeptide sequence (e.g. amino acids I27-K42). The colored amino acids starting from the N-termini are E31, D33, E37, D38, R41 and K42. Distribution over all positions of charged amino acids, e.g. acidic or basic, observed (b) in an alignment of Raf type RBDs retrieved from the SMART database and (c) in the massive sequence perturbation experiment of Raf RBD. K84 (*) was not varied in the main experiment, but showed average conservation for basic residues when tested independently. On the other hand, R89 (**) when tested independently displayed complete conservation for arginine (see text).

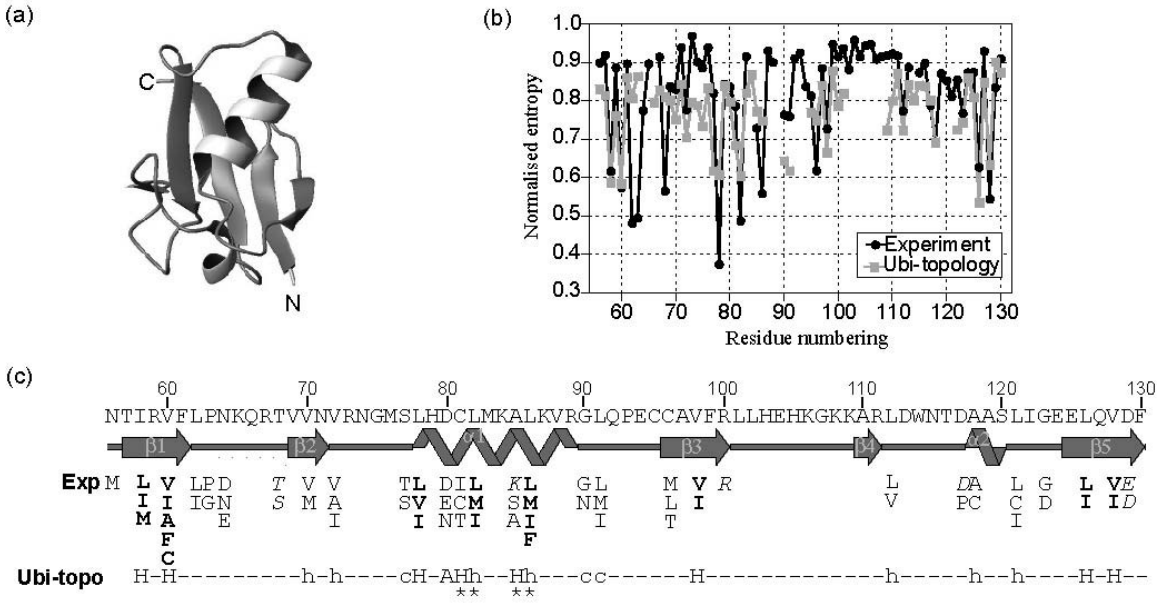

(a)
C
N
(b)
Normalised entropy
Residue numbering
Experiment
Ubi-topology
(c)
NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLLHEHKGKKARLDWNTDAASLIGEELQVDF
Exp
Ubi-topo

Fig. 1


60
70
80
90
100 110
120
130
1RFA NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLKARLDWNTDAASLIGEELQVDF
1UBI MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYKESTLHLVL
1A5R IKLKVIGQDSSEIHFKVK--MHLKKLKESYCQRQGVPMNSLRFLFEGQRIADNHT-PKEEEDVIEVYQ
1MG8a MIVFVRFNSSYGFPVEVDSDTSILQLKEVVAKRQGVPADQLRVIFAGKELPNHLTVQNCQQSIVHIVQ
1VCBa VFLMIR-RHKTTIFTDAKESSTVFELKRIVEGILKRPPDEQRLYKDDQLLDDGKTLGECAPATVGLAF
1M94a IEVVVNDRLGKKVRVKCLAEDSVGDFKKVLSLQIGTQPNKIVLQKGGSVLKDHISLEDYDQTNLELYY
1J8Ca IKVTVKTP-KEKEEFAVPENSSVQQFKEAISKRFKSQTDQLVLIFAGKILKDQDTLIQHDGLTVHLVI
1H8Ca SKLRIRTPSGEFLERRFLASNKLQIVFDFVASK-GFPWDEYKLLSTRRDTDPNKSLLEVPQETLFLEA
1JRUa TNIQIRLADGGRLVQKFNHSHRISDIRLFIVDA-AM--AATSFVLM-KELADQ-TLKEAAVIVQRLT-
1EO6a VPVIVEKV-----KYLVPSDITVAQFMWIIRKRIQLPSEKIFLFVDKTVPQSSLTMGQLGFLYVAYSG
1EF1a ISVRVTTD--AELEFAIQPNTTGKQLFDQVVKTIGLEVWFFGLQYQSTWLKLNKKVTAQSPLLFKFRA
1GG3a MHCKVSLLDDTVYECVVEKHAKQDLLKRVCEHLLNLDYFGLAIWDNKTWLDSAKEIKKQ-PWNFTFNV
1LFDa CIIRVSLDV--YKSILVTSQDKAPTVIRKAMDKHNLEPEDYELLQIKLKIPENANVFYA-NYDFILKK
1E8Xa VFIVIHRS-TTSQTIKVSADDTPGTILQSFFTKMA---RDFVLRVRDEYLVGETPIKNF-EIHLVLDT
1K8Rb CILRFIACNGQTRAVQSRG--DYQKTLAIALKKFSLDASKFIVCVSIKLITEE-------D-RLIIVP
1L7Ya VTFKITLSD-PFKVLSVPESTPFTAVLKFAAEEFKVPAATSAIITNGVGVNPAQPAGNIFGSELRLIP
1D4Ba RPFRVCDHKRIRKGLTAA---QLAKALETL-----L-NGVLTLVLE-TAVDSEDFFQLL-DTCLMVLQ
1C9Fa KCVKLRALHSCKFGVAAR---SLLRKGCVRF---QLPMPGSRLCLDGTEVTDDFP---LNDAELLLLT
1F2Ri KPCLLRRNHSDQHGVAAS---SLRSKACELL---AI----ITLVLAGTIVDDDY-FLCLSNTKFVALA
1IP9a TKIKFYYK-DDIFALMLKGDTTLRSKIAPRI---DTD---FKLQTKSEEIKTDQ-VSNIIALKISVHD
1Q1Oa ILFRISY---EIFTLLVEKVWNLIMAINSKISNT------IKIKYQFVVLGSDD-WNVAKFLNIRLY-
1FMAd IMIKVLFF---ATEVAADFP-TVEALRQHMAAQSDLEDGKLLAAVNQTLVSFDHPL---DGDEVAFFP
1F0Za MQILF--N-D--QAMQCAAGQTVHELLEQL----DQRQAGAALAINQQIVQWAHIV---DGDQILLFQ
1JSBa MKFTVITD-DGKKILESGAPRRIKDVLGEL----EIPIETVVVKKNGQIVIDEEEI---DGDIIEVIR
1QF6a PVITL--P-D-GSQRHYDHAVSPMDVALDI----PL-----AGRVNGELVDACDL---INDAQLSIIT
1JALa QTYFTAGV-KEVRAWTVSVGATAPKAAAVIH---TF---I-RAEVIWRLEGKDYIV---DGDVMHFR-
1MG4 KKVRFYRNG-FGIVYAISPDFSFEALLADLTRTLNLP----TIYTILKKIS---SLDQLEGESYVCGS
1FRRa YKTVLKTPSG-EFTLDVPEGTTILDAAEEA----GY---SCLGKVVEGFVLTIAIP---SDLVIETHK
1I7Ha PKIVIL-P---GAVLEANSGETILDAALRN----GI---TCHCIVR-ESRLSQARV---EDLVVEIPR
1AYFa ITVHFINRDGETLTTKGKIGDSLLDVVVQN----NLD--TCHLIFE-RSRLGQICLKAMDNMTVRVP-
1PUT SKVVYVSHDGTRRQLDVADGVS----LMQAAVSNGI---TCHVYVN-NSRLCQIIMPELDGIVVDVPD
1L5Pa GTITAVKG-GVKKQLKFEDDQTLFTVLTEA----GL---KCICKHV-NARLAAITLGENDGAVFEL--
1E0Za PTVEYLNY-----TMEVAEGEYILEAAEAQ----GY---NCASIVKKDVRLTIGSP---DEVKIVYNA
1FEHa KTIII--N-G--VQFNTDEDTTILKFARDN----NI---ICTVEVE--LVTADTLI---DGMIINTNK
1HLRa IQKVITVN-GIEQNLFVDAEALLSDVLRQQL---GL---ACSVILDGKVVRAVTKMKRVDGAQITTIE
1FO4a DELVFFVN-GKKVVEKADPETTLLAYLRRKL---GL---ACTVMLSHFSANALAPICTLHHVAVTTVE
1JROa MEIAFLLN-GETRRVRIEDPQSLLELLRAE----GL---ACTVMIRSRAVNALMMLPQIAGKALRTIE
1FFVa KIITVNVN-GKAQEKAVEPRTLLIHFLREEL---NL---ACTVDIDGRSVKSTHLAVQCDGSEVLTVE
2PIA TPFTVRLRSGTSFEI--PANRS----ILEVLRDANV----TALC--TQIMVVSRA-----S-ELVLDL
1NENb MRLEFSIY--QDYTLEADEGMMLLDALIQLKEKDSL---GLNM--NGKNGLAITPISALPGKKIVIRP
1KF6b KNLKIEV---AFYEVPYDATTSLLDALGYIKDNLDL---GMMV--NNVPKLAKTFLRDYTD-GMKVEA
1QLAb RMLTIRVF-FQEYKIEEAPSMTIFIVLNMIRTYDDL---GMMI--NGRPSLARTLTKDFEDGVITLLL
1JQ4a HTITAVTEDGESLRFECRSDED----VITAALRQNI---TCKALCSLVLLCRTYP-----TDEIELPY
1KRHa HQVALQFEDGVTRFICIAQGETLSDAAYRQ----QI---TCRAFCE-GYVLAQCRP----DAVFQIQA
1C78a PYLMVNV--PHYVEFPIKPGLTIEYYVEWALDATAY-----SAKIETKSFPIGFVV---PGFNLITKV
1BMLc SQLVVSV--LKFFEIDLT--LELLKAIQEQLIA--------DATITGKVYFATLPTQP--EFLLS-HV
1SE2 QNVLIRVN-KISFEVQTDKKVTLDIKARNFLINKNL---TGYIKFITFWY-----S---KSVKIEVHL
1AN8 HKLLGNLSG-QNLNIILEKDVTIDFKIRKYLMDNKI---SGRIEIGHEQI-----M---KNFHFDIYL
1AW7a IELPLKVHG--KYWPKFDKKLALDFEIRHALTQIGL---GGYWKITTYQS-----I---DEITIEAEI
2IGD TTYKLVIGK--ETTTKAV---DAEKAFKQYANDNGVD---GVWTYD----------------KTFTVTE
2PTL VTIKANLANGQTAEFK-G---TATSEAYAYADTLDNG---YTVDVA----------------TLNIKF
1ACC TTARIIF---VERRIAA---MTLKEALKIAF---GF-ITEFDFNFD----VLDIKL---AKMNILIRD
1G0Sa HAAVLLP-----VAGMIEEGESVEDVARREAIEAGLI----RTKPV----------------SIMVGE-
1I9Aa LAFSSWLF----VCGHPQLGESNEDAVIRRCRYEGV-----PPESI----------------CPVFAART

Fig. 2

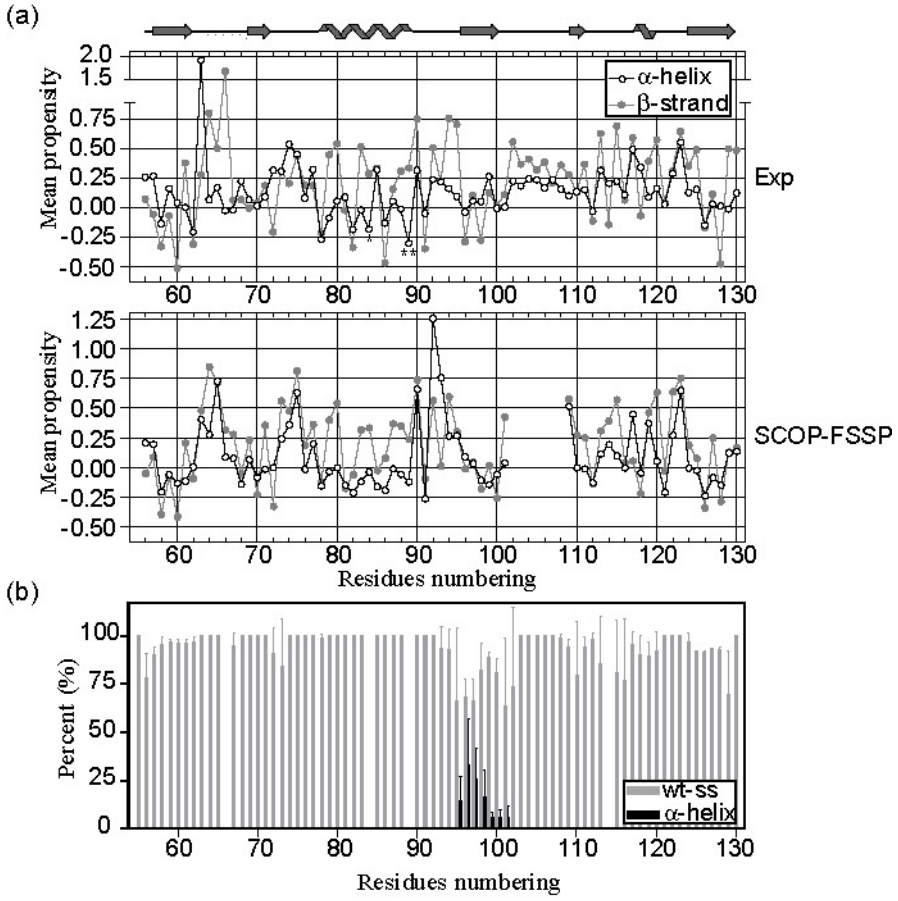

(a)
α-helix
β-strand
Mean propensity
Exp
SCOP-FSSP
Residues numbering
(b)
Percent (%)
wt-ss
α-helix
Residues numbering

Fig. 3

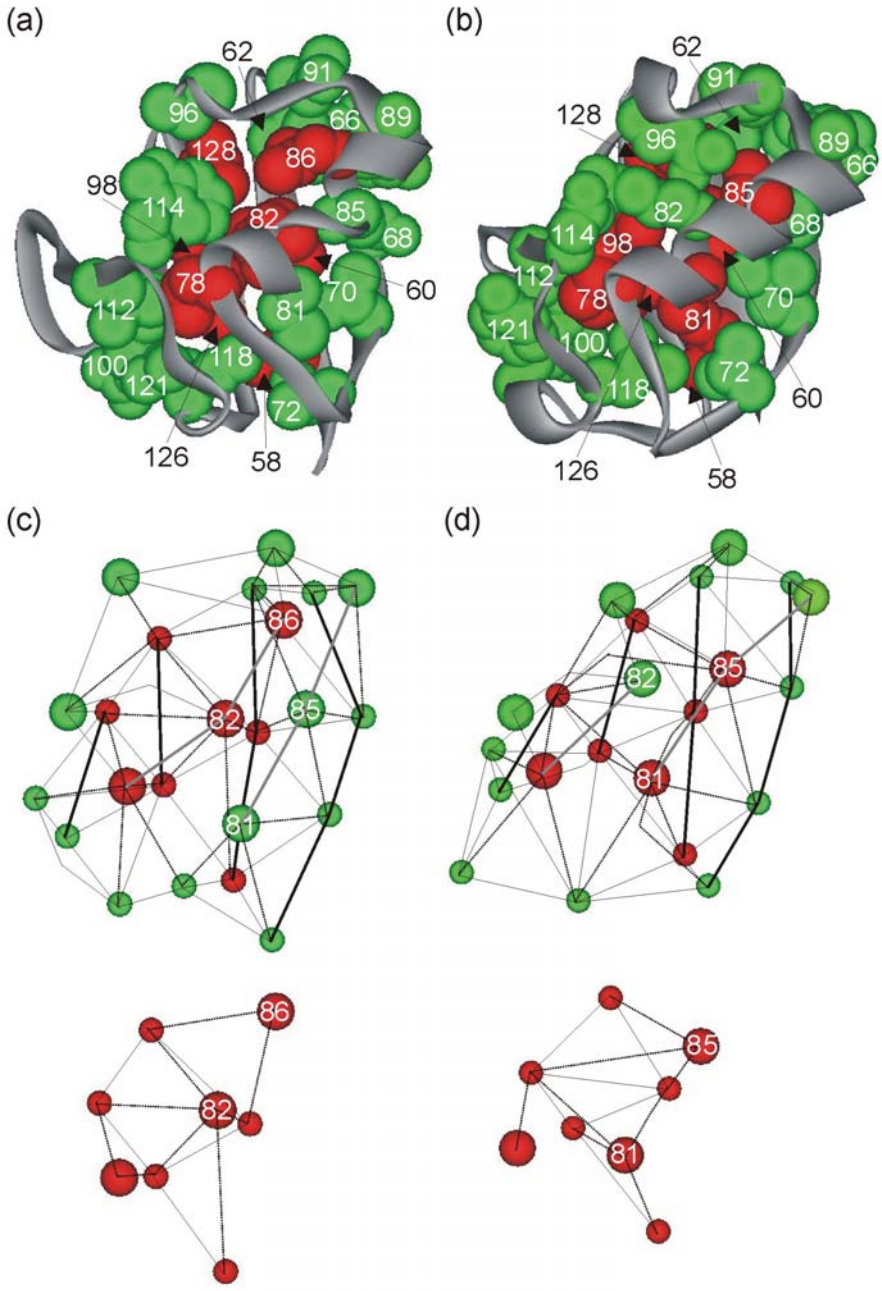

(a)
62
91
96
66
89
128
86
98
114
82
85
68
78
70
60
112
81
118
100
121
72
126
58
(b)
62
91
128
96
89
66
85
82
68
114
98
112
78
70
81
121
100
72
118
60
126
58
(c)
86
82
85
81
86
82
(d)
82
85
81
85
81

Fig. 4

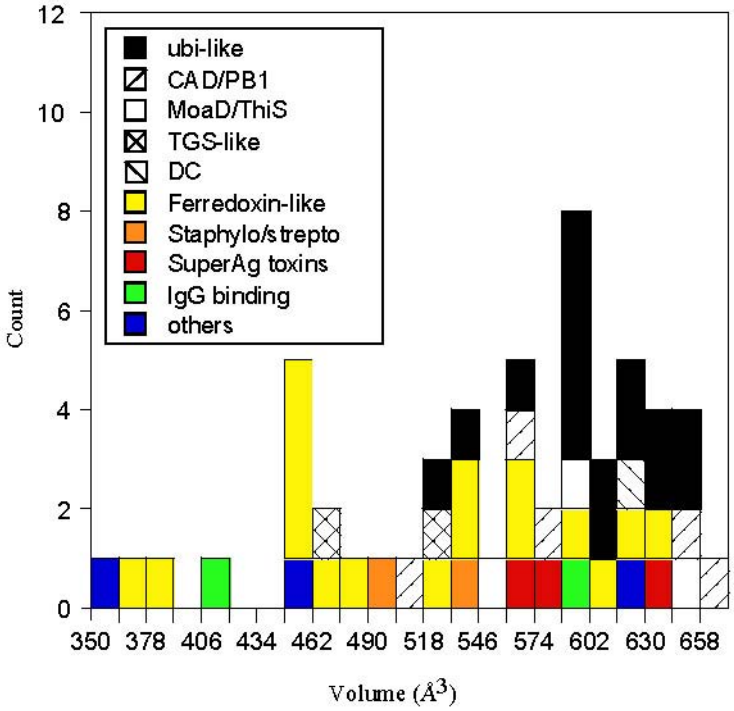

ubi-like
CAD/PB1
MoaD/ThiS
TGS-like
DC
Ferredoxin-like
Staphylo/strepto
SuperAg toxins
IgG binding
others
Count
0
2
4
6
8
10
12
350 378 406 434 462 490 518 546 574 602 630 658
Volume (Å3)

Fig. 5

Table 1. Amino acids observed at inner hydrophobic core residues in ubiquitin-related superfamilies.

Code PDB	58[a]	60[a]	78[a]	81/85[a]	82/86[a]	98[a]	126[a]	128[a]
1RFA*	I	V	L	L	L	V	L	V
1UBI	I	V	I	V	I	L	L	L
1A5R	L	V	L	L	Y	F	I	V
1MG8a	V	V	I	L	V	V	V	I
1VCBa	L	I	V	L	V	L	V	L
1M94a	V	V	V	F	L	L	L	L
1J8Ca	V	V	V	F	I	L	V	L
1H8Ca	L	I	L	V	V	L	L	L
1JRUa	I	I	I	I	I	F	Q	L
1EO6a	V	V	V	F	I	L	V	Y
1EF1a	V	V	G	L	V	L	F	F
1GG3a	C	V	Q	L	C	I	F	F
1LFDa	I	V	A	V	A	L	F	L
1E8Xa	I	I	P	I	F	L	L	L
1K8Rb*	L	F	Y	L	L	V	L	I
1L7Ya	F	I	F	V	A	I	L	L
1D4Ba*	F	V	L	A	L	L	L	V
1C9Fa*	V	L	L	G	F	L	L	L
1F2Ri*	C	L	L	A	L	L	F	A
1IP9a	I	F	L	L	I	L	I	V
1Q1Oa*	F	I	L	I	I	I	I	L
1FMAd	I	V	V	L	M	A	V	F
1F0Za	I	F	V	L	L	L	I	L
1JSBa	F	V	I	V	L	V	I	V
1QF6a	I	L	P	V	I	G	L	I
1JALa	Y	T	A	A	I	A	M	F
1MG4	V	F	F	L	L	I	Y	C
Frequency range of amino acids (%)								
40-55		V		L		L		L
20-40	I, V	I	L, V	V	I, L		L	
10-20	L, F	F, L	I	F, I, A	V	I, V	I, V, F	V, F, I
3-10	Y, C	T	F, P, A, G, Y, Q	G	F, A, M, C, Y	F, A, G	M, Y, Q	Y, A, C

* Structures with packing i (res. 78), i+4 (res. 82) and i+8 (res. 86) of α-helix inner core residues. All others adopt i, i+3 (res. 81) and i+7 (res. 85) packing.

[a] Inner core residues as described in ([23]) and Figure 4.

Table 2. Thermodynamic parameters.

	m^{b} (kJ mol^{-1} M^{-1})	$\Delta G_{F\text{-}U}^{eq}$ (kJ mol^{-1})	Cm (M)	$\Delta\Delta G_{F\text{-}U}^{0M}$ (kJ mol^{-1})	$\Delta\Delta G_{F\text{-}U}^{Cm}$ (kJ mol^{-1})	$m^{kin\,a,b}$ (kJ mol^{-1} M^{-1})
Wt	3.8	-24.1	6.3	nsap	nsap	4.0
N56M	3.8	-25.1	6.6	-1.0	-1.1	4.4
I58A	4.1	-11.8	2.8	12.3	13.5	3.8
I58L	4.3	-23.3	5.5	0.8	3.3	4.2
I58F	4.0	-19.4	4.9	4.7	5.6	4.3
R59A	3.6	-23.2	6.5	0.9	-0.6	4.0
V60A	3.6	-13.2	3.7	10.9	10.1	4.1
L62A	3.6	-8.8	2.4	15.3	15.1	3.9
P63A	3.6	-17.3	4.7	6.9	6.1	3.5
N64A	3.9	-20.0	5.2	4.1	4.5	4.1
H2	4.0	-27.4	6.8	-3.3	-1.8	3.8
H2_F62L	3.6	-18.5	5.1	5.6	4.6	3.9
Q66A	3.8	-21.6	5.7	2.5	2.3	4.2
T68A	3.8	-23.3	6.1	0.8	0.7	4.3
V69A	3.9	-20.0	5.1	4.1	4.8	4.1
V70A	4.0	-17.4	4.3	6.7	7.7	4.1
V72A	4.2	-21.1	5.1	3.0	4.8	4.3
V72I	3.9	-20.9	5.3	3.2	3.8	4.6
M76A	3.1	-16.5	5.3	7.6	4.1	4.3
S77A	4.0	-18.1	4.6	6.0	6.8	4.0
S77T	4.0	-27.6	7.0	-3.5	-2.6	4.2
L78A	4.0	-9.2	2.3	14.9	15.7	4.1
D80A	3.6	-20.0	5.6	4.1	2.6	4.0
C81A	3.6	-24.2	6.7	-0.1	-1.5	3.9
C81I	5.0	-24.9	5.0	-0.8	5.2	4.7
L82A	4.8	-13.7	2.8	10.4	13.5	5.7
A85G	4.4	-17.7	4.0	6.4	9.0	4.6
L86A	4.1	-11.6	2.8	12.5	13.5	3.9
R89L	4.1	-29.5	7.3	-5.4	-3.8	4.3
L91A	4.0	-18.4	4.6	5.7	6.7	4.6
P93A	4.0	-24.8	6.2	-0.7	0.5	4.2
C95A	4.1	-26.2	6.4	-2.1	-0.3	3.8
C96A	3.6	-14.7	4.0	9.5	8.8	4.0
C96L	3.6	-23.3	6.4	0.8	-0.4	3.7
C96M	4.4	-25.5	5.7	-1.4	2.2	4.3
A97G	4.0	-22.2	5.5	1.9	3.2	3.8
V98A	3.5	-13.7	4.0	10.4	9.1	3.2
R100A	3.8	-21.5	5.6	2.6	2.7	3.9
E104A	3.9	-24.7	6.3	-0.6	0.0	4.1
K109A	3.9	-23.4	6.0	0.7	1.2	4.1
L112A	3.8	-14.0	3.7	10.1	10.3	3.4
D117A	3.6	-19.8	5.5	4.3	3.1	4.2
A118G	3.8	-16.3	4.3	7.8	7.9	3.7
A118L	3.3	-14.1	4.2	10.0	8.1	3.6
L121A	4.0	-14.5	3.7	9.6	10.3	3.8
E124A	4.1	-25.8	6.3	-1.7	0.1	4.0
E125A	3.8	-21.8	5.7	2.3	2.3	4.1
L126A	4.1	-11.0	2.7	13.1	14.0	4.1
V128A	4.0	-11.0	2.7	13.1	13.9	3.8
D129A	4.0	-22.2	5.5	1.9	3.0	4.3
Δ104-6	3.5	-28.0	8.0	-3.9	-6.7	4.1
Δ101-8+AG	3.6	-16.1	4.5	8.0	7.2	3.6

See Materials and Methods for parameters description.

[a] Calculated from RT * ($-m_f + m_u$) (accompanying paper JMB).

[b] Average m and m^{kin} are 3.90±0.36 and 4.07±0.41, respectively. For Ala/Gly mutants only, m and m^{kin} are 3.88±0.29 and 4.05±0.39, respectively.

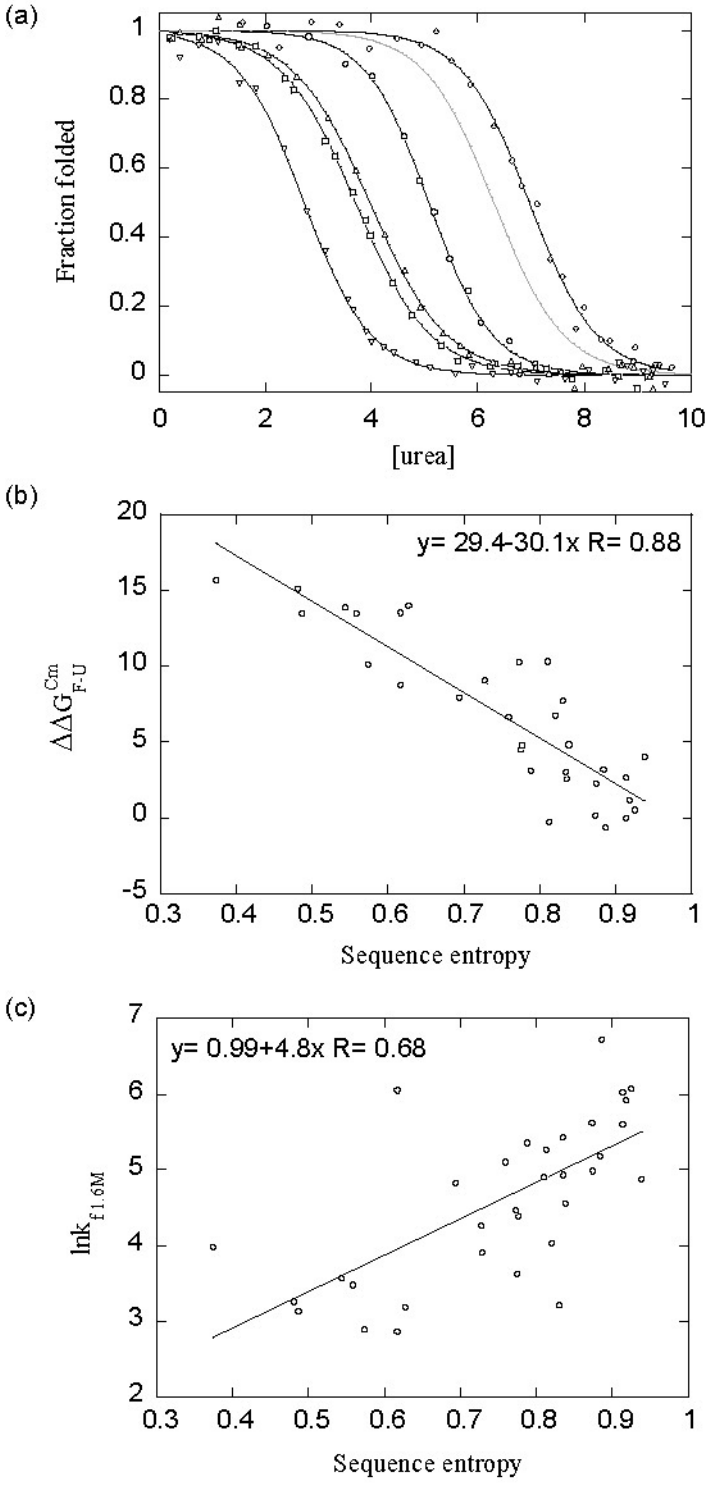
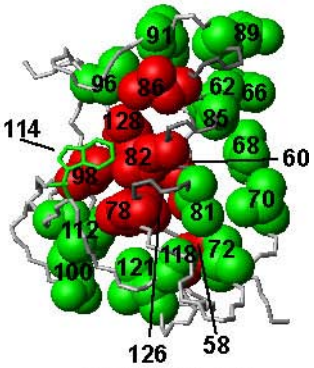
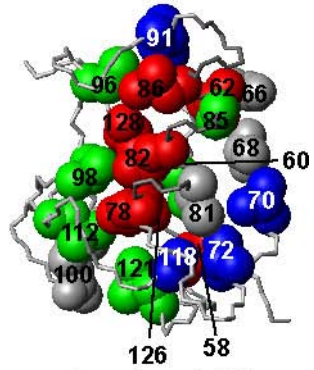

(a)
Fraction folded
[urea]
(b)
y= 29.4-30.1x R= 0.88
$\Delta\Delta G^{Cm}_{F-U}$
Sequence entropy
(c)
y= 0.99+4.8x R= 0.68
$\ln k_{f1.6M}$
Sequence entropy

Fig.6



Inner/outer core



Impact on stability

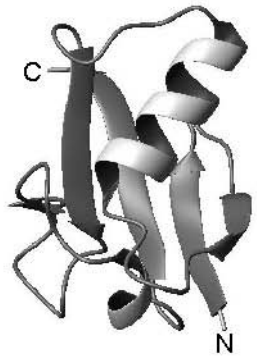


Fig. 7

Table 3. Thermodynamic and kinetic parameters for significantly stabilized mutant and double-, triple-cycle mutants.

	m^{eq} (kJ mol^{-1} M^{-1})	ΔG_{F-U}^{eq} (kJ mol^{-1})	$-m_f$ (M^{-1})	k_f (s^{-1})	m_u (M^{-1})	k_u (s^{-1})	βt[e]
wt[a]	3.8	-24.1	1.22	2263	0.41	0.05	0.75
Simple mutants[a]							
H2	4.0	-27.4	1.03	3874	0.49	0.10	0.68
S77T	4.0	-27.6	1.20	2566	0.52	0.01	0.70
R89L	4.1	-29.5	1.26	30681	0.46	0.04	0.73
Δ104-6[b]	3.5	-28.0	1.32	9661	0.32	0.02	0.81
wt[c]	10.9	-21.8	3.66	1100	0.98	0.31	0.79
Δ104-6: -double-triple[d]							
Δ104-6	10.1	-27.6	3.05	2747	1.01	0.05	0.75
Δ104-6/S77T	9.9	-29.0	3.25	4714	1.10	0.02	0.75
Δ104-6/S77T/H2	10.4	-32.9	2.68	7082	1.03	0.11	0.72
Δ104-6/S77T/R89L	10.4	-34.3	2.72	13162	1.22	0.04	0.69

See Materials and Methods for parameters description.

[a] Data taken from Table 1 and accompanying article (accompanying paper JMB).

[b] ΔG_{F-U}^{eq}, m_u, k_u and β_t are not reliably measured, because the protein is too resistant to urea induced unfolding.

[c] Data on wt Raf RBD obtained from experiments in Gdm-HCl and tris buffer, pH= 7.5 ([22]).

[d] Experiments performed in Gdm-HCl as described in Materials and Methods.

[e] Calculated from kinetic data using $\beta_t = m_f/(m_f+m_u)$

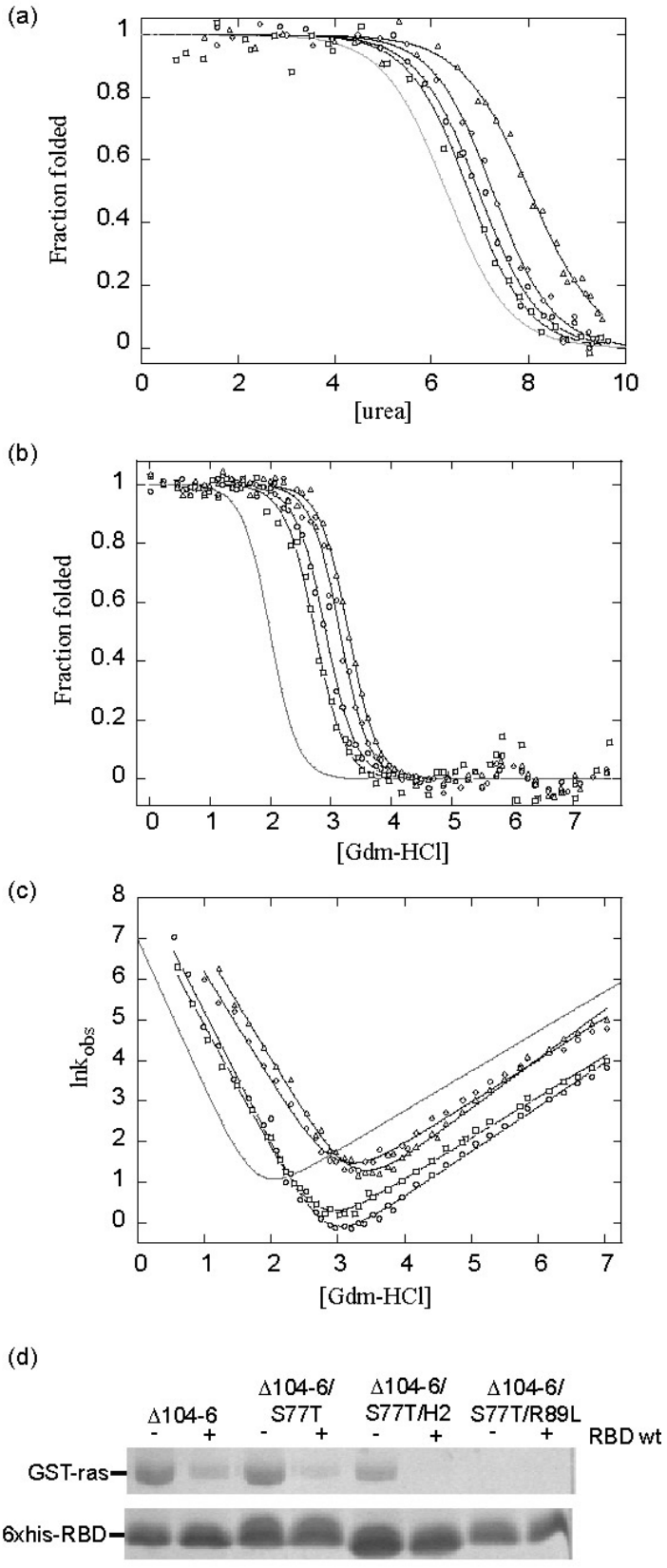

(a)
Fraction folded
[urea]
(b)
Fraction folded
[Gdm-HCl]
(c)
lnk$_{obs}$
[Gdm-HCl]
(d)
Δ104-6
Δ104-6/ S77T
Δ104-6/ S77T/H2
Δ104-6/ S77T/R89L
RBD wt
GST-ras
6xhis-RBD

Fig. 8

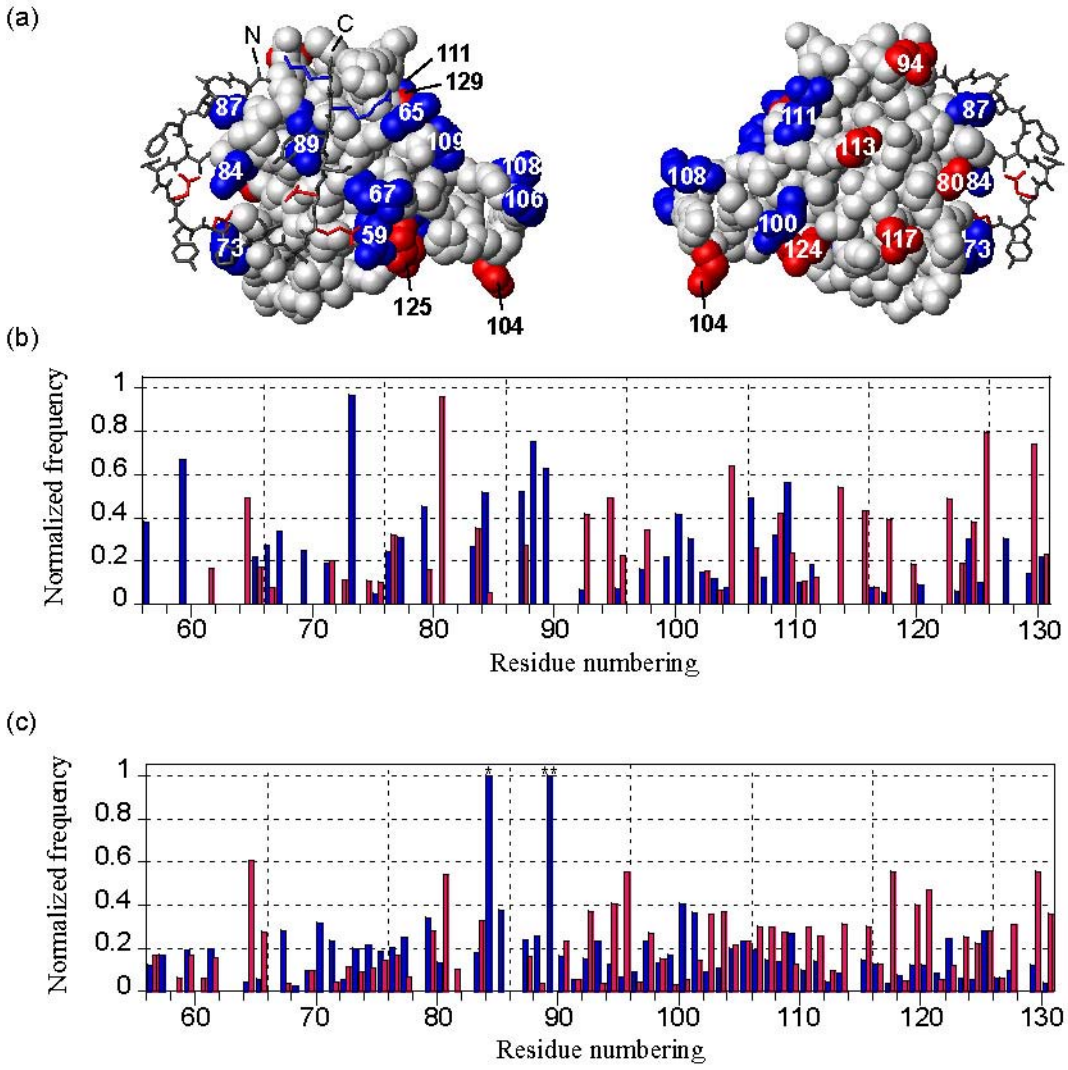

(a)
N
C
111
129
87
65
89
109
108
84
67
106
73
59
125
104
94
87
111
113
108
80
84
100
124
117
73
104
(b)
Normalized frequency
1
0.8
0.6
0.4
0.2
0
60
70
80
90
100
110
120
130
Residue numbering
(c)
Normalized frequency
1
0.8
0.6
0.4
0.2
0
60
70
80
90
100
110
120
130
Residue numbering

Fig. 9

Article 4 : Description de l'état de transition du DLR de Raf en utilisant la méthode d'analyse des valeurs-Φ : comparaison avec ubiquitine

Article soumis à « Journal of Molecular Biology »

Présentation de l'article 4 :

Cet article présente les expériences et les paramètres cinétiques des mutants du DLR de Raf présentés dans l'**Article 3** ainsi qu'une analyse complète des valeurs-Φ de ces mutants. La structure de l'état de transition du DLR de Raf qui est obtenue y est comparée a celle d'ubiquitine qui a été publiée récemment (*97*). Les découvertes principales de cet article sont :

1. L'état de transition du DLR de Raf est polarisé autour du motif épingle à cheveux localisé à l'extrémité amino-terminale, mais inclut aussi plusieurs des autres résidus du cœur hydrophobe interne.

2. L'état de transition apparaît plus diffus selon une mesure strictement énergétique de l'état de transition. Ce résultat souligne aussi l'importance de considérer l'effet absolu d'une mutation sur le taux de repliement et donc la stabilisation de l'état de transition en conjonction avec l'approche plus classique de l'analyse des valeurs-Φ.

3. L'état de transition et plus largement le mécanisme de repliement du DLR de Raf et d'ubiquitine sont similaires.

4. Certaines données indiquent que le mécanisme de repliement comporterait un resserrement de l'empaquetage en carboxy-terminal de l'hélice-α immédiatement à l'état de transition ou immédiatement avant ou après celui-ci. Cela pourrait être en accord avec la présence d'état(s) intermédiaire(s) de haute énergie sur la voie de repliement du DLR de Raf.

Avant le parachèvement de cette étude, il avait été démontré pour un petit nombre de topologies structurales que des analogues structuraux à l'identité de séquence relativement faible pouvaient posséder un état de transition ou à tout le moins une cinétique de repliement aux propriétés similaires (*104;107-109;114;128-130*) (ces exemples sont révisés dans (*136*)). Un collègue de notre laboratoire dans le cadre de ces travaux de doctorat avait commencé à définir des similarités dans le mécanisme de repliement du DLR de Raf et d'ubiquitine. Il avait démontré une sensibilité similaire de la réaction de repliement aux mutations, à un sel stabilisant la structure native (Na_2SO_4) et à la température (*131*). La constatation que la structure de l'état de transition d'ubiquitine et du DLR de Raf obtenue par la méthode d'ingénierie des protéines partage plusieurs similarités concorde avec ces résultats antérieurs. Par ailleurs, la très faible identité de séquence entre le DLR de Raf et

ubiquitine (< 12%) et leur lien évolutif distant permet d'ajouter une nouvelle topologie à la courte liste de celles pour lesquelles des analogues structuraux distants évolutivement ont vu leur état de transition caractérisé (voir section **Les protéines partageant la même topologie se replient-elles par un mécanisme identique : oui et non**).

Contribution des auteurs à la préparation de l'article 4:

F.-X.C.V. : conception et réalisation des expériences, analyse des données et rédaction de l'article.

S.W.M. : supervision du projet et rédaction de l'article.

Article 4: «Protein engineering of the *ras* binding domain of Raf reveals a polarized distribution of residues with high Φ-values, but an energetically diffuse transition state»

Authors: Campbell-Valois, F.-X.[1, 2], Michnick, S.W.[1*]

[1]Département de Biochimie and [2]Programme de Biologie Moléculaire, Université de Montréal, C.P. 6128, Succ. centre-ville, Montréal, Québec, Canada H3C 3J7

[*]Corresponding author

Summary

The *ras* binding domain (RBD) of the Ser/Thr kinase c-Raf/Raf-1 spans 78 residues and adopts a structure characteristic of the β-grasp ubiquitin-like topology. Remarkable similarities in folding kinetics between this RBD and ubiquitin have been reported, despite insignificantly low sequence similarity. Further, the primary sequence of RBD has been nearly exhaustively perturbed experimentally by insertion of stretches of degenerate codons, which revealed sequence conservation and hydrophobic core organization similar to that found in an alignment of β-grasp ubiquitin-like proteins. These results now allow us to examine the relationship between sequence conservation and the folding process, particularly viewed through the analysis of Transition State (TS) structure. Specifically, we present herein a protein engineering study combining classic truncation (Ala/Gly) and atypical mutants to predict folding TS ensemble properties. Based on classical Φ-value analysis, Raf RBD TS structure is particularly polarized around the N-terminal β-hairpin. However, all residues constituting the inner layer of the hydrophobic core are involved in TS stabilization, although they are clearly found in a less native-like environment. The TS structure can also be probed by a direct measure of its destabilization upon mutation, $\Delta\Delta G_{U-‡}$. Viewed with this parameter, Raf RBD TS is a more diffuse structure, in which all residues of the hydrophobic core including β-strands 1, 2, 3 and 5 and the major α-helix play similar roles in TS stabilization. Moreover, the comparison of TS structures obtained from Φ or $\Delta\Delta G_{U-‡}$ for Raf RBD *versus* ubiquitin reveals striking similarities between the two proteins, albeit ubiquitin TS appears slightly more denatured-like and polarized. In light of these results, we suggest that in the frame of a protein engineering based description of TS, Φ-values interpretation and discussion should also consider the direct impact of mutations on TS energetic ($\Delta\Delta G_{U-‡}$). Finally, the impact of these findings on the modeling of protein folding is discussed.

Keywords : protein folding; chevron curves; protein engineering; Φ-value; Raf *ras* binding domain (RBD).

Introduction

At the heart of the protein folding problem lays the search for unifying principals that adequately describe the processes by which a polypeptide chain spontaneously folds from the denatured state into a unique three-dimensional structure. The discovery of apparently two-state folding proteins has clarified our vision of the folding process by providing simpler theoretical and experimental models ([1]). The implication of two-state models is that through such a folding pathway, only one species determines the rate of the reaction, i.e. the transition state (TS). As a result much of the emphasis in the last decade has been centered on the study of this state. The capacity to alter the polypeptide sequence by directed mutagenesis and specifically, the application of protein engineering to folding through the development of the Φ-value analysis method has provided a framework for straightforward interpretations of the importance of a given amino acid residue in stabilizing the TS ([2, 3]). The importance of the protein engineering methods in the development of our contemporary view of protein folding is clear ([4-21]), despite the need to address mounting concerns regarding the interpretation of Φ-values and the comparison of TS ensemble properties ([22-27]).

The role of various residues in the folding process of a model protein is classically addressed by the insertion of non-disruptive point mutations (usually alanine) and the measurement of their impact on the folding rate and hence, TS stability ($\Delta\Delta G_{U-\ddagger}$). To make meaningful prediction of TS structural properties, $\Delta\Delta G_{U-\ddagger}$ is normalized by the change in native state stability ($\Delta\Delta G_{F-U}$) induced by a given mutation, yielding $\Phi_F = \Delta\Delta G_{u-\ddagger} / \Delta\Delta G_{F-U}$. The interpretation of Φ-value is based on a key set of assumptions discussed in classic works by Fersht and colleagues ([3, 28]). A collection of Φ-values obtained at a significant number of residues dispersed on a protein structure is used routinely to define the structural characteristic of the TS. Based on this type of study, the TS structure of protein are often described as polarized or diffuse depending on the topological distribution of residues having high versus low Φ values. A delocalized TS has been very well described for the important model CI2 and for other proteins ([5, 9, 18, 29]). In contrast, polarized TS structures have been described for barnase, two SH3 domains, protein-L and more recently a cold

shock protein and ubiquitin ([4, 7, 8, 17, 29-31]). The fundamental physical meaning behind this categorization of TS structure, if any, is not perfectly well established and might depend on the experiments and parameters utilized to described the TS.

Proteins sharing a common structural topology are hypothesized to adopt a similar folding mechanism (reviewed in ([32])). For example, the functionally homologous SH3 domains of src, spectrin and fyn share TS's with similar structural characteristics ([7, 8, 19, 33]). The combination of classic Φ-value analysis and molecular dynamics suggest high similarity of the TS properties among three members of the homeodomain superfamily. However, the nature of the pathways to the native states is divergent with domains folding seemingly through either of the framework, nucleation condensation or intermediate models ([21]). Two members of the immunoglobulin-like greek key fold displaying insignificant sequence identity were shown to display striking similarities in the dispersion of the most significant Φ-values ([14, 15]), although experiments and simulations revealed large differences in the level of compaction of their TS ([34]). The folding rates of proteins sharing the acylphosphatase (AcP) fold are extremely well correlated with relative contact order and AcP. Moreover, the activation domain of procarboxypeptidase A2 (10% sequence identity), two proteins that adopt this topology display similarities in their folding nucleus ([9]). This clear-cut effect of the native state topology on the definition of the folding process is being challenged by contrasting results as significant differences are observed between other members of the AcP fold ([10]). The symmetrically organized IgG binding domains of Protein-L and G display notable differences in the structure of their TS both at the topological level, as revealed by the contrasting roles of their N- and C-terminus, and in the organization of their hydrophobic core, which appears more diffuse in Protein-G ([17, 18]). These results suggest that some folds could form through distinct pathways and TS's. This hypothesis is attractive, because it could provide an explanation for the structural and functional versatility of the topologies that are recurrent in the natural protein universe. Clearly, more folds and diverse structural superfamilies have to be studied to clarify our vision of the folding process and the rules that govern it.

The *ras* binding domain of c-Raf/Raf-1 (Raf RBD) Ser/Thr kinase is an interesting model to study the relationships between folding kinetics, native state stability and binding. The interaction of this domain with GTP-loaded *ras* induces activation of Raf kinase activity, thereby triggering the activation of the mitogen activated pathway (MAP) kinase signal transduction cascade (reviewed in ([35])). The domain is composed of 78 amino acids and folds independently into a compact globular structure formed by the packing of an α-helix (α1) against a mixed β-sheet in which the β-strands are organized in the following order β2-β1-β5-β3-β4 ([36]) (Figure 1 (a) and (b)). The only other notable element of secondary structure is a short 3^{10} helix (α2) located before β5. Its structure is classified in the β-grasp ubiquitin-like topology (aka ubiquitin-roll topology) according to the Structural Classification of Proteins (SCOP) database (http://scop.mrc-lmb.cam.ac.uk/scop/). This topology frequently occurs in the protein universe and encompasses sequences with high sequence and functional diversity ([37-39]).

Ubiquitin and the Raf RBD share remarkable structural similarities, despite insignificant sequence identity (< 12%, based on alignment of secondary structure elements). Moreover, their folding pathways display similar sensitivities to mutation, temperature and stabilizing salts, despite significant differences in native state stability and TS placement on the reaction coordinates ([40]). The refolding traces of both proteins are complex and must be fitted with 4 exponentials. Mammalian ubiquitin TS appears to be highly polarized around the N-terminal β-hairpin ([31]), in a manner vaguely reminiscent of the TS formed by protein-L ([17]), a distant structural analogue. Recently, we reported an exhaustive sequence perturbation of Raf RBD. The results described a detailed sequence conservation profile and suggested a bi-layer organization of the hydrophobic core, which is also observed in an alignment of β-grasp ubiquitin-like proteins (Materials and Methods, Figure 1 (c) and (d)) ([38]). This study provided key information about the sequence determinants of Raf RBD structure and function and the framework for directed mutagenesis studies described below.

We have engineered 54 mutants of Raf RBD that can be separated into two categories: 1) side-chain truncation mutants (Ala/Gly) and 2) atypical mutants. Mutations were selected based on the perturbation experiments, literature information and alignments of sequences recovered in SCOP and SMART databases. In the accompanying manuscript, a thermodynamic study of these variants and new insights into the sequence perturbation experiments are presented. The combination of these two types of datasets allows for testing different hypotheses concerning the relationships between sequence, function, stability and folding/unfolding kinetics and their conservation throughout evolution ([39]). Herein, we present the kinetic parameters of these mutants derived from chevron curves and present the TS structure according to Φ-value and a pure measure of TS destabilization ($\Delta\Delta G_{U-\ddagger}$). Strikingly, the structure of Raf RBD TS appears polarized according to Φ, but diffuse based on $\Delta\Delta G_{U-\ddagger}$. These analyses also reveal that Raf RBD and ubiquitin fold through very similar TS.

Results

A subset of residues of the Raf RBD was chosen for Φ-value analysis based on their low tolerance to mutation in sequence perturbation experiments ([38]). These include all residues of the hydrophobic core except W114, which serves as the folding/unfolding fluorescent probe (Figure 1) and some residues involved in the topological arrangement of the domain and in the binding interface with *ras*. Additionally, other residues were selected to cover all regions of the structure, including surface exposed positions. In total 37 of the 78 residues of the Raf RBD were probed by mutations with the majority of untested residues located at the surface. These 37 residues were mutated to Ala (except Ala residues that were mutated to Gly) to facilitate experiments and analyses. In addition, more atypical mutations were designed based on information found in the literature (I58F, V72I and R89L), sequence alignments (A118L, Δ101-6 and Δ101-8+AG) or in the Raf RBD sequence perturbation experiments, particularly positions displaying significant bias toward non-wt amino acids (N56M, I58L, S77T, C81I, C96L, C96M, H2 and H2_F62L) (see Materials and Methods for description of the Δ101-6, Δ101-8+AG, H2 and H2_F62L mutants).

Fitting traces and chevron curves

The kinetics of folding/unfolding of Raf RBD in both Gdm-HCl and urea were previously reported ([40, 41]). The folding traces of wt Raf RBD are complex and must be fit with 4 exponential functions. The three slower phases are only resolved at concentration below 3 M urea and there rates are insensitive to denaturant concentration (Figure 2 (a)). The two slowest of them were attributed to prolyl and non-prolyl peptide bond isomeriation, whereas the remaining one could not, its origin remaining unknown as yet ([40, 41]). The two-state transition is embodied by the fastest phase in the refolding traces of Raf RBD and hence, variation in its rate upon mutation is used to delineate the TS structure. The dependence of folding/unfolding rates on urea concentration was followed for all Raf RBD mutants using stopped-flow fluorescence spectroscopy (Materials and Methods). The resulting chevron curves are grouped by secondary structure segments in Figure 2 (b-j).

Kinetic and thermodynamic parameters

The fast folding rate of Raf RBD and the small mixing volume ratio used in these experiments can increase errors due to long extrapolation necessary for the determination of k_f^{H2O} and k_u^{H2O} from chevron curves (Material and Methods). In order to minimize these effects, the rate of folding at 1.6 M ($k_f^{1.6M}$) and the unfolding rate at 5.8 M and 8 M ($k_u^{5.8M}$ and k_u^{8M}, respectively) are also reported (Table 1), and were used to derive Φ-value estimates as described in the next section.

Most of the Ala/Gly mutations induced a decrease in the folding rate, an increase in the unfolding rate or combination of both of these effects (Table 1). However, some mutants that increased amino acid side-chain length accelerated both the folding and the unfolding rate (N56M, V72I and C81I). In addition, four mutants (H2, S77T, R89L and Δ104-6) were found to increase the folding rate and/or decrease the unfolding rate and improve stability. Overall, the folding rates obtained span approximately 2 orders of magnitude according to $k_f^{1.6M}$, with the slowest and fastest mutants folding approximately 15 times slower (18 s^{-1}; I58A) and 15 times faster (4092 s^{-1}; R89L) than the wt Raf RBD (322 s^{-1}). A similar range of unfolding rates was observed according to $k_u^{5.8M}$, but the slowest and fastest reactions

proceeded 5 times slower (0.11 s^{-1}; Δ104-6) and 100 times faster (57.00 s^{-1}; L78A) than the wt (0.53 s^{-1}), respectively.

The variation in m_f and m_u observed across all Raf RBD variants was very close to a normal distribution and the averages calculated are 1.25 ± 0.12 and 0.39 ± 0.06, respectively. The mutants displaying the highest variation in m_f and m_u are: V98A< L112A< A118G< Δ101-8+AG< m_f^{wt}< A85G< L91A< V72I< C81I< L82A; L62A< P63A< H2_F62L< L86A< L126A< V128A< m_u^{wt}< C96M< H2< S77T (Table 1 and Figure S1 in Supplementary material). The variation of these parameters was similar to those found in other two-state folding proteins ([7, 8, 17-19, 31]). In a two-state model of protein folding, m_f and m_u can be used to calculate the β-Tanford value [$\beta_t = m_f/(m_f+m_u)$], which indicates the TS position relative to the ground states. Change in β_t upon mutation of a protein or the imposition of other chemical and physical perturbations can be interpreted using Hammond or the rarer anti-Hammond postulates ([42-44]). A perturbation (e.g., chemical, mutation, temperature etc.) that destabilizes the native state is said to trigger Hammond behavior if it concomitantly induces a shift of the TS toward the native state therefore, leading to an increase in β_t. The β_t factors calculated for Raf RBD variants are relatively consistent with a mean of 0.76 ± 0.04, spanning from 0.68 to 0.83 (Table 1). In this study, 12 mutants (β_t> 0.8) display potential Hammond behavior. β_t is a relative measure of TS position between the denatured and native state, thus variations in the properties of the ground states must be ruled out before concluding that a true shift in the TS position could be occurring. This can be tested indirectly by verifying that variations in m_f are compensated by an opposite change in m_u and *vice versa* in a collection of variants of a given protein. In contrast, as described in ([11, 29]), correlation of m_f or m_u with m_f+m_u would indicate modification in the properties of the denatured or native state, respectively. The data on Raf RBD reveals a very good correlation of m_u+m_f *versus* m_f (slope= 0.88; R= 0.90), but not m_u (slope= 0.40; R= 0.18) (Figure 3 (a)). Virtually identical correlations of m_u+m_f *versus* m_f are observed for strictly Ala/Gly mutants (data not shown). Thus, the access of solvent to the denatured state but not the native state is significantly affected upon mutation of Raf RBD.

Urea unfolding curves were obtained from the endpoint fluorescence signal of unfolding traces and were used to calculate the thermodynamic parameters grouped in Table 2. The m value estimates obtained from equilibrium and kinetic data were similar overall, yielding mean values of 3.90±0.36 and 4.07±0.41 ([39]), respectively. Measures were again taken to avoid errors due to extrapolation by calculating various $\Delta\Delta G_{F\text{-}U}$ estimates using both solely thermodynamic parameters ($\Delta\Delta G_{F\text{-}U}^{Cm}$ and $\Delta\Delta G_{F\text{-}U}^{5.8M}$) and kinetic parameters $\Delta\Delta G_{F\text{-}U}^{kin}$ (Material and Methods). The kinetic estimate of the change in free energy correlates well with the estimates from the equilibrium experiments (slope= 0.98, R=0.96; Figure 3 (b)), in agreement with a two-state model for folding in which $\Delta G_{F\text{-}U}$ is correlated with $\ln(k_f/k_u)$. The correlation between thermodynamic and kinetic estimates of $\Delta\Delta G_{F\text{-}U}$ diminishes the likelihood that the apparent variations in the denatured state properties discussed in the preceding paragraph are energetically significant. Therefore, it simplifies the Φ-value analysis, because the change in the denatured state can be considered to have only marginal effects on $\Delta\Delta G_{U\text{-}‡}$ and $\Delta\Delta G_{F\text{-}U}$ ([29]).

Φ-value analysis

As described previously, the calculation of independent Φ-value estimates obtained from different kinetic and thermodynamic parameters provides an indication of the magnitude of experimental error ([17]). Using the different methods of calculating $\Delta\Delta G_{F\text{-}U}$ described above, we have obtained Φ-value estimates similar to those described by Kim et al. (Table 3). Equations and parameters used for calculations of the Φ-value estimates are detailed in Materials and Methods. Briefly, Φ_F^{kin} was calculated solely from kinetic data using $k_f^{1.6M}$ and k_u^{8M} and the relationship between change of energy of the native state and folding/unfolding rate in a two-state model for folding, Φ_F^{Cm}, was calculated with k_f^{H2O} and a fixed m-value, using a minimal extrapolation of the thermodynamic data. Finally, $1\text{-}\Phi_U$ was obtained from $k_u^{5.8M}$ and another minimal extrapolation of the thermodynamic data. No atypical Φ-value (Φ< 0 ou Φ> 1) was consistently observed across the three estimates for a given non-disruptive mutation once those with $\Delta\Delta G < |2|$ kJ mol^{-1} were excluded (Table 3). The only notable exception to this statement is R89L, which could result from larger errors in estimation of $k_f^{1.6M}$ due to its extremely fast folding rate. Overall, we found a good

correlation between the Φ_F^{kin} and Φ_F^{Cm} (slope= 1.05; R= 0.90) or 1-Φ_U (slope= 1.03; R= 0.91) (Figure S2 of Supplementary Material). The major outlying mutants in the Φ-value correlation plot are P63A, V72A, C96A, A97G and R100A and Δ101-8+AG. The most significant discrepancies occur in Φ_F^{Cm}. Accordingly, we found that the values of Φ_F^{kin} are more similar to 1-Φ_U than to Φ_F^{Cm} (average standard deviation for all mutants of 0.07 and 0.09, respectively). The higher discrepancy of Φ_F^{Cm} is predictable given its calculation from k_f^{H2O}, a parameter with larger inherent error and sensitivity to variation in m_f, and from fixed m-value. The comparison with Φ_F^{exp}, calculated from fully extrapolated equilibrium and kinetic data (Table S1 of Supplementary material), suggest that the mechanistically unjustified assumption of fixed m-value is the main cause of deviation of Φ_F^{Cm} from the other estimates.

It is not a common practice in the protein folding field to present different estimates of Φ-values. In order to verify that the magnitude in errors observed in our study is not aberrant, we have compared the average standard deviation of the three Φ-value estimates collected for each Raf RBD mutant to those of protein-L as reported by Kim and colleagues, and found similar errors (e.g. averages of the three estimates for Raf RBD and protein-L are 0.41±0.11 and 0.26±0.10, respectively). Given the much faster folding rate of Raf RBD, we concluded that the quality of our dataset was adequate. Because of the smaller extrapolation used in their calculations, the Φ_F^{kin}-values (thereof Φ or Φ-value) are more precise, particularly for less destabilizing mutations ([27]). Therefore, they were used in the structural description of Raf RBD folding TS and in all figures that follow.

TS ensemble characteristics determined by Φ_F value (Ala/Gly mutations)

The hydrophobic core

The highest Φ-values at core residues occur mainly in the N-terminal β-hairpin, which spans residues T57-N71 (Figure 4). The mutant V60A (Φ= 0.82; Φ-value for a given residue appear in parenthesis in Results, unless mentioned otherwise), in the middle of β1, has a higher Φ-value than I58A (0.54) and L62A (0.44), located at the edges of β1. The mutants V70A (0.96) and V72A (0.98) that are located in β2 showed the highest Φ-values

at hydrophobic core positions. In contrast, the inner core residues of α1 have much lower Φ-values. L78A (0.28), the residue at the N-terminus of α1, displays a lower Φ-value than residues towards the C-terminus of α1, e.g. L82A (0.45) and L86A (0.44). The native-likeliness of the contact formed by the outer core residue A85G (0.57) at the TS is compatible with the Φ-value of the two latter inner core residues. It was proposed that the insertion of Gly mutations in an α-helical segment reduces the ratio of molecules in a given population that forms a stable α-helix and consequently, it could be used to monitor the impact of helix formation on folding rate ([45, 46]). Therefore, the mutant A85G probes secondary structure formation as well as the contacts formed by the β-methyl group of the side chain. Residues in the hydrophobic core located in the carboxy-terminal half of Raf RBD sequence play a more minor role in stabilization of the TS. The mutation yielding the most significant Φ-value in that region is V98A (0.50). Mutations probing residues of the outer core and proximal to V98, such as C96A (-0.14) and R100A (0.13) have among the lowest Φ-values in our dataset. Outer core residues, including L112A (0.26), A118G (0.26) and L121A (0.19), further away on the polypeptide chain yield similar Φ-values. The inner core mutants of β5, L126A and V128A displayed close to average Φ-values of 0.45 and 0.39, respectively. In summary, the clustering of mutation producing high Φ-values at residues located in the first β-harpin suggest that folding proceed through a polarized activated state, although the rest of hydrophobic core has also formed a certain fraction of native-like structure

First β-turn, binding patch to ras and capping of the major α-helix

The strongest amino acid selection observed in the sequence perturbation experiment at topologically constrained positions are found in β-turn1 and in residues involved in N- and C-capping of α1 (e.g., S77, G90 and L91). The two residues in the middle of the β-turn1 have apparently contrasting roles in the stabilization of the TS. In fact, P63A and N64A display low (0.32), and very high Φ-values (0.98), respectively. After β-turn1, and before β2, there is a region of undetermined structure (Q66-T68). This small stretch constitutes part of the binding surface for *ras* in which Q66 and T68 are most critical

for binding. In addition, these residues were classified as outer core residues, based on structural data and consensus among ubiquitin-roll topology members ([38]). This classification should be revised for Raf RBD as the mutation of these residues for Ala does not significantly destabilize the protein ([39]). Thus, the low entropy of these residues in the sequence perturbation experiment is probably mainly due to selective pressure for conservation of the binding function. The β-turn2 immediately before α1 was probed solely by M76A, which displays an average Φ-value (0.51).

N-capping of the α-helices usually consists of N-terminally located Ser or Thr whose side chain hydroxyl groups can contribute to satisfying the unpaired backbone hydrogen donor of the first residues adopting helical conformation. A negatively charged residue is also favored at the N-terminal end of the α-helix to neutralize its induced dipole (reviewed in ([47])). In Raf RBD, S77 and D80 could fulfill these roles; mutation of these residues to Ala yielded high Φ-values of 0.70 and 1.05, respectively. On the other hand, distant residues were also shown to be involved in N-capping of Raf RBD (N-capping in trans) ([36]), suggesting that the role of S77 in the stabilization of α1 might be reduced in comparison to a more common N-capping motif involving proximal hydroxyl bearing lateral chains. The N-capping in trans of α1 involves backbone carbonyls of residues W114 and T116, as indicated by the H-bond networks they formed with L78 and H79. T116 is connected through H-bonds to D117, an exposed residue which displayed significant amino acid selection in the sequence perturbation and a relatively low Φ-value upon mutation to Ala (0.38). Accordingly, W114 has a very minor impact on TS formation as indicated by kinetic data obtained in another study (Φ_F^{kin} for W114M is 0.07; the Met was the least destabilized mutation tested) (A. Vallée-Belisle and S.W. Michnick unpublished data). It is notable that this phenomenon of N-capping in trans of α1 is in agreement with the low Φ-value of L78A (0.28). Nevertheless, it must be noted that the discrepancies in the importance of the residues involved in N-cap formation for TS stabilization could stem from the obvious limitation of the protein engineering approach in probing interactions formed by backbone atoms. At the other extremity of α-helices, C-capping is often composed of the motif Gly/Asn-Aliphatic. The mutation of the matching aliphatic residue in

Raf RBD, L91, displayed a low Φ-value (0.25). The nearby β-turn3 was probed by the mutations P93A and C95A. They induced insufficient $\Delta\Delta G_{F-U}$ hence, precluding calculation of significant Φ-value (Table 2 and 3).

Solvent exposed positions and putative electrostatic interaction

Five mutations were designed to probe the exposed surface of the β-sheet. First, mutations R59A (β1) and A97G (β3) displayed, respectively, insignificant $\Delta\Delta G_{F-U}$ and low Φ-value (Table 2 and 3). V69A, which is located in β2, showed Φ-value well above average (0.69), in agreement with its contiguous position to the very important TS stabilizing residue, V70. E125 is the first residue of β5 and could be involved in a salt bridge with R59. In contrast to the latter, the E125A mutation showed a high Φ-value (0.67). In light of this contrasting result, we decided to probe the role in TS formation of a few other potential charge-charge interactions, which are concentrated in the C-terminal half of Raf RBD. At the other end of β5, the low Φ-value of D129A (0.24) suggested a modest role of this residue in TS formation. This residue could be involved in a salt bridge with K109A, but this latter residue showed an insignificant $\Delta\Delta G_{F-U}$, arguing against the importance of this putative interaction. Residues E104 and E124 are located close to the side-chain of R100 and may also be involved in charge–charge interactions, but similarly to K109A, mutations to Ala of both residues resulted in insignificant $\Delta\Delta G_{F-U}$ (Table 2). In summary, the results concerning these putative salt bridges suggest that they play a negligible role in stabilizing the TS and the native state.

Φ-value for disruptive mutants confirm that α2 is not involved in TS formation

The introduction of bulkier side-chains was recently proposed as a manner of obtaining more information on the properties of TS structure ([19, 48]). Several proteins sharing the ubiquitin-roll topology display a contact triad involving the side-chains of residues corresponding to I58 (β1), V72 (β2) and A118 (α2) of Raf RBD. Since the contacts established between the side-chains of these three residues are spatially constrained and α2 is packed against the rest of the hydrophobic core strictly through A118, we reasoned that introducing larger amino acids at any of the sites of the triad should disrupt

the proper packing of α2 and could be used to confirm its absence in the TS ensemble as predicted from the Φ-value of A118G. In agreement with this hypothesis, I58F and A118L displayed very low Φ-values, similar to A118G. The V72I mutation produced a negative Φ-value, suggesting that it induces formation of non-native interactions before the TS. Taken together, these results confirm that α2 is not properly packed at the TS.

Comparison of Φ and $\Delta\Delta G_{U-\ddagger}$ parameters and impact on predicted TS structure

The Φ-value analysis revealed a polarized TS with a nucleus located in the N-terminal β-hairpin. However, there are very few residues throughout the protein that have near-zero Φ-values, suggesting that some level of native structure is formed in almost all parts of the Raf RBD. In view of these results, we sought to compare the TS characteristics delineated from Φ-value analysis with those deduced from a direct energetic measure of TS destabilization. To do so, $\Delta\Delta G_{U-\ddagger}^{1.6M}$ was normalized against $\Delta\Delta G_{U-\ddagger}^{1.6M}$ max, which is the $\Delta\Delta G_{U-\ddagger}^{1.6M}$ of the mutant with the most destabilized TS (e.g. I58A) (Materials and Methods). This new normalized parameter, $\Delta\Delta G_{U-\ddagger}^{1.6M}$rel, ranges between 0-1 for destabilizing non-disruptive mutations, allowing for a straightforward comparison to Φ-values (Figure 4). What is immediately obvious from the comparison of Φ-value *versus* $\Delta\Delta G_{U-\ddagger}^{1.6M}$rel is a drift in the TS structural properties, from polarized to more diffuse, respectively. Indeed, according to the latter parameter, all regions of the inner core participate roughly equally to TS stabilization including residues located in α1 (particularly L82 and L86) and β5, which are not in such a highly native-like environment. In contrast, the role of core residues in the most native-like region, the N-terminal β-hairpin, does not change much, although the relative role of L62 increased while that of V72 diminished. To ensure that the non-thermodynamically normalized parameter $\Delta\Delta G_{U-\ddagger}^{1.6M}$rel does not introduce a bias for the most drastic mutation (e.g., I→A *versus* V→A), the effect of normalizing according to side-chain volume variation was tested. We found that this manipulation of the data did not significantly change the results for Raf RBD and ubiquitin (see Discussion section) or other proteins discussed below (data not shown).

Literature, sequence perturbation and alignment insights: folding/unfolding kinetics of atypical mutants

Results of the sequence perturbation experiment suggested that some residues of Raf RBD favored non-wt amino acids ([38]). We reasoned that some of these mutations might have been selected, because they improve the stability, folding rate or binding function to *ras*. Therefore, we tested the effect on kinetics and stability of mutations corresponding to the strongest amino acid selection observed in the perturbation experiment ([39]). The mutation of N56 to Met accelerated k_f ($\approx 4 \cdot k_f^{1.6M}$ wt), but also k_u ($\approx 2.5 \cdot k_u^{8M}$ wt), leading only to a marginal stabilization (Table 2). The mutant I58L folded slower than the wt and displayed a similar Φ-value to I58A, suggesting that this mutation is non-disruptive. The S77T unfolds slower ($\approx 5 \cdot k_u^{8M}$ wt) and is stabilized, displaying a low Φ-value (Table 1-3). Mutation C81I folds much faster ($\approx 6 \cdot k_f^{1.6M}$ wt), but is slightly destabilized. Mutations of residue C96 to Met and Leu were among the most highly favored mutations in the sequence perturbation experiment, but their mutation in the wt background produced no or marginal destabilizing effect. Intrigued by these results, mutants C81I and C96M were tested for improved *ras* binding *in vitro*. The K_d extrapolated from the binding curves were equivalent to the wt (data not shown) ([38]).

The mutant H2 that was recovered in the sequence perturbation experiment displayed a faster k_f and k_u and a slight improvement in stability (Table 1 and 2). This mutation changed the turn type at β-turn1 through replacement of amino acids 62-LPNK-65 for 62-FTDG-65. The β_t factor for this mutant (0.68) is the lowest among all mutants tested (Table 3) and is very similar to the β_t reported for ubiquitin (0.66) ([40]). It is noteworthy that H2 and the ubiquitins possess the same turn type in the β-turn1 (the corresponding residues in ubiquitin are: 7-TLTG-10). We hypothesized that the presence of a phenylalanine at residue 62 of H2 might destabilize the native state, leading to an increase of the k_u. However, mutant H2_F62L, which is obtained by reverting residue 62 to Leu in the H2 background was destabilized, displayed a low Φ-value (0.10) and a high β_t (0.83).

Another interesting mutation was suggested by a study, which evaluated the importance of most Raf RBD residues directly involved in binding to *ras* ([49]). The mutation R89L was found to completely disrupt formation of the complex. The non-obvious choice of Leu to replace R89 brought us to determine the folding/unfolding kinetics of this variant. Strikingly, R89L dramatically stabilizes the domain through an impressive improvement in k_f ($\approx 10 \cdot k_f^{1.6M}$ wt), without significantly affecting the k_u ([39]), yielding a Φ-value slightly greater than 1. This result indicates that the C-terminal part of the α-helix is not optimized and that R89L increase stability of the TS, hence suggesting the importance of this region in the folding mechanism.

c-Raf/Raf-1 has an insertion of three amino acids *versus* a-Raf or b-Raf RBD, in a loop connecting β3 and β4. Specifically, alignment of Raf type RBD obtained from the SMART database suggest that residues equivalent to E104-H105-K106 are absent from a-Raf and b-Raf. Therefore, a deletion mutant of c-Raf with the stretch E104-K106 removed (Δ104-6) was generated and its kinetics tested. This mutant folded faster ($\approx 3.5 \cdot k_f^{1.6M}$ wt) and unfolded slower ($\approx 5 \cdot k_u^{8M}$ wt), yielding a Φ-value of 0.4. Moreover, Δ104-6 is so stable that it cannot be completely unfolded by urea, which leads to imprecision in the unfolding parameters m_u and k_u. The kinetic and thermodynamic parameters of this mutant in Gdm-HCl and further experiments are reported in the companion article ([39]). Ubiquitin does not possess the region of undetermined structure located between β3 and β4 (L102-K108) of Raf RBD, but two flexible residues (Ala-Gly) forming a tight β-turn. To check if stability or kinetics could be improved by replacing the segment (L102-K108) with Ala-Gly, the mutant Δ101-8+AG was generated. This mutation failed to improve any of the parameters and yielded a close to zero Φ-value.

Overall behavior of mutations

The Leffler (aka Brønstead) plot is an established method in chemistry that has been adapted to protein folding ([28]). For a series of mutations engineered into a protein, $\ln k_f$ or $\ln k_u$ is plotted against $\Delta\Delta G_{F\text{-}U}/RT$ and the slope (e.g. β_F or β_U) of the best linear fit obtained corresponds to an average Φ_F. This fit can be determined for probed mutants as a whole or

in sub-groups corresponding to various secondary structure elements and regions of the tertiary structure (e.g. core or surface etc.). Sub-grouping allow for assessing the consolidation of the diverse structural regions at the TS. The Leffler plot regrouping all Raf RBD mutants described in this study display very good correlations (Figure 5 (a); β_F= 0.47 and R= 0.82). The observation that the correlation is significant for all mutations is in agreement with the results of Φ-value analysis at individual positions and the low number of residues displaying extreme values (e.g. 0 or 1). Nevertheless, a better correlation can be obtained if the N-terminal β-hairpin residues (N56-V72) the residues between α1 and β5, and the remaining mutants are fit separately, as shown by the best linear fit for $\ln k_f$ (Figure 5(b); β_F= 1.03 and R= 0.89, β_F= 0.25 and R= 0.84, and β_F= 0.50 and R=0.88, respectively). However, this very good correlation among β-hairpin residues is obtained if the major outliers (I58A, L62A and P63A), corresponding here to the most destabilizing mutations, are ignored (y= 6.20+0.65x; R= 0.85 for linear regression including all mutants of the β-hairpin subgroup). Thus, it is not clear that the linearity of the dependence of $\ln k_f^{1.6M}$ *versus* $\Delta\Delta G_{F\text{-}U}^{kin}/RT$ is true for mutants of the β-hairpin with $\Delta\Delta G_{F\text{-}U}^{kin}/RT > 4$. All the observations noted above with regard to the Leffler plot hold true if strictly Ala/Gly mutants are considered (data not shown) or for correlation of folding and unfolding rate with equilibrium data (Figure S3 of supplementary material). Overall, these results support the hypothesis that the residues located in the β-hairpin are in a more native-like environment, while all regions of the protein seem to participate to stabilization of the TS although the vast majority of residues located between α1 and β5 do to a lesser extent. This model suggests a very progressive extension of consolidated structure in the TS ensemble from a core nucleus formed in the amino-terminal β-hairpin. Specific curvature in Leffler plots for barnase were suggested to indicate the presence of parallel pathways in which the helix located at the N-terminus was either completely structured or partially disrupted ([28, 44]). Sanchez *et al.* reported that such a pattern in a Leffler plot was exceptional, as similar phenomenon was not observed in any other two-state folding proteins of the large set they analyzed ([29]). More mutants of Raf RBD should be obtained to clarify the discrepancy in the Leffler correlations. In addition, given the moderate scattering of the data, it is likely that more discrete sub-elements in the structure could be identified during this process.

Discussion

Indirect evidence of residual structure in the denatured state of Raf RBD and ubiquitin

Several small proteins were previously shown to display apparent variation in the denatured state structure upon mutation based on correlation of m_f *versus* m_f+m_u ([11, 29]), including ribosomal protein S6, CI2, Sso7D SH3, protein-L and G. In particular, the study by Sanchez and Kiefhaber suggested that this could be a rather common phenomenon as 7 of 21 proteins displaying apparent Hammond behavior showed significant structural changes in their denatured state. We have presented similar results for Raf RBD (Figure 3 (a)) and ubiquitin display a similar phenomenon (data not shown, based on ([31])). In both cases, the most significant changes in m_f occur upon mutation of hydrophobic core residues, which argues in favor of a denatured state structure stabilized by hydrophobic side-chain interactions. In Raf RBD the most drastic changes in m_f for non-disruptive mutations (e.g., Ala/Gly mutations) are observed for L82A, A85G, L91A, V98A and L112A (Table 1 and Figure S1 of Supplementary material). A corollary of these observations is that β_t is not a reliable measure of TS shift. On the other hand the analyses performed suggest that the native state is insensitive to mutation (Figure 3 (a)). Therefore, Raf RBD mutants displaying low m_u compared to the wt are interesting because they could indicate true TS shift. It is noteworthy that most of the mutations corresponding to this criterion appear in a structural region corresponding to the N-terminal β-hairpin and inner core residues of the α1 C-termini and β5 (Table 1; see section on the Description of the folding pathway for further discussion).

In theory, the observation of structure in the denatured state and its sensitivity to mutation causes a problem for the straightforward interpretation of Φ-values. Indeed, one assumption of protein engineering is precisely that the denatured state does not vary in energy upon mutation ([3]). The correlation described for m_f *versus* m_f+m_u solely indicates variation in exposure to solvent of the denatured state, not necessarily a significant variation in energy. In this sense, contacts formed in the denatured state could be so weak and

transiently formed that it is reasonable to state that the majority of $\Delta\Delta G_{U-\ddagger}$ occurs because of TS destabilization ([29]). This assumption is in agreement with the correlation observed among the Φ-value estimates determined from folding and unfolding data on Raf RBD mutants (Table 3 and Figure S2). Direct experimental evidence suggests that native and non-native contacts are involved in stabilizing fluctuating substructures in the denatured state of several proteins, including barnase, CI2 and engrailed homeodomain that were previously submitted to Φ-value analysis ([50-52]).

Description of Raf RBD TS characteristics: polarized (Φ) and diffuse ($\Delta\Delta G_{U-\ddagger}$)

The Φ-value analysis suggests that Raf RBD adopts a polarized TS structure characterized by a nucleus located in the amino-terminal β-hairpin. In fact, this segment contains 4 of the 5 hydrophobic core residues with the highest Φ-values (I58, V60, V70 and V72) (Table 3). The β-turn1 is well formed at the TS, although the residues tested showed some discrepancy. The L62A and P63A mutants had close to average Φ-values, while N64A displayed close to unity Φ values. The latter residue could be involved in forming side-chain backbone H-bonds that could stabilize the β-turn1 at the rate limiting step as was suggested for N14 in protein-L ([17]). From the β-hairpin, the nucleus extends, albeit with less native-like characteristics, to all residues located in the inner core including the α1 (specifically L82A and L86A), β3 (V98A) and β5 (L126A and V128A). All of these residues showed close to average Φ-values (Figure 4 and 5 (a)). Furthermore, high Φ-values obtained upon mutation of N-capping residues (S77A: 0.70; D80A: 1.05) and by introducing a Gly mutation in α1 (A85G: 0.57) suggest that it is well formed at least between residues D80 and L86. The low Φ-value observed for L78A (e.g., the first inner core residue of α1) agrees well with the N-capping in trans of this region of α1 by the segment D113-D117, which could occur later in folding according to Φ-values of residues mutated in this region. On the other hand, the C-cap (L91A) shows few native contacts in the TS, suggesting that the last turn of α1 is not very well structured. The observation that the mutation R89L displays a Φ-value close to unity and a 10 fold improvement in folding rate is in agreement with a non-optimal packing of this part of α1 in the wt TS. In summary,

these observations indicate that the most consolidated region of the hydrophobic core in the TS is located in the amino-terminal β-hairpin, but that all regions forming the inner core participate to a lesser degree in its stabilization. This extension of the nucleus from a polarized focal point is coherent with the Leffler plot obtained (Figure 5 (b)), which could be described to be somewhere between the proteins CI2 and protein-L, prototypes of diffuse and highly polarized TS, respectively ([5, 17]). The conclusions drawn from Φ_F^{kin} hold true for the other Φ-value estimates, including with Φ_F obtained from fully extrapolated data, particularly for the most highly destabilizing mutations (Table S1 of Supplementary material).

Starting from this first sketch of the Raf RBD TS structural properties, we reasoned that it would be interesting to confront our interpretation of Φ-values to $\Delta\Delta G_{U-\ddagger}$, a direct measure of TS energetic. To facilitate the comparison between the two parameters, we normalized $\Delta\Delta G_{U-\ddagger}^{1.6M}$ of all mutants against $\Delta\Delta G_{U-\ddagger}^{1.6M}$max, yielding $\Delta\Delta G_{U-\ddagger}^{1.6M}$rel. We found that using this direct measure of the stabilizing contribution of residues to the TS that the energy contribution in the structural ensemble is more diffuse than expected from the Φ-value analysis. According to $\Delta\Delta G_{U-\ddagger}^{1.6M}$rel obtained for I58A and V60A (inner core; β1), followed closely in decreasing order by L62A (β1 outer core), V70A (β2 outer core), L82A, L86A (α1 inner core residue 2 and 3, respectively), L126 and L128 (β5 inner core), these residues are the most important for TS stabilization. Next, a significant group of residues, as indicated by N64A, S77A, A85A and V98A, play lesser but nevertheless significant roles in consolidating the TS. This energetic perspective on the TS ensemble properties suggests that the inner core plays a central role, with some auxiliary contribution of the outer core particularly strong in the N-terminally located L62 and V70 (Figure 4 (c)). On the other hand, both Φ and $\Delta\Delta G_{U-\ddagger}^{1.6M}$rel showed that all residues probed between L91-L121, excluding V98, play a minor role in TS stabilization.

The divergence between the TS structure revealed by Φ-value and $\Delta\Delta G_{U-\ddagger}$rel highlights the fact that the former is not a pure measure of TS structure. Indeed, because of the introduction of $\Delta\Delta G_{F-U}$ as a denominator in Φ calculation, Φ-values represent a scale of

the native-likeliness of the environments and contacts formed by a residue involved in TS stabilization. Consequently, residues impacting equally on $\Delta\Delta G_{U-\ddagger}$, but differently on $\Delta\Delta G_{F-U}$ display different Φ-values, converging to one for mutants in which $\Delta\Delta G_{U-\ddagger}$ approximates $\Delta\Delta G_{F-U}$. Therefore, the impacts of mutations that induce the greatest destabilizing effect on TS, but that have not completely formed their native contacts at this stage are underestimated by Φ-value. As described above, several residues of the Raf RBD hydrophobic core show such behavior. Hence, proteins likely to behave similarly to Raf RBD should display a moderately polarized TS in combination with a rather homogenous contribution of hydrophobic core residue to native state stability. Nonetheless, two drawbacks of TS description using the normalized parameter $\Delta\Delta G_{U-\ddagger}{}^{1.6M}$rel should be kept in mind in order to interpret the results correctly. First, this parameter is dependant on normalization by the most destabilizing single point mutations, which could vary for the same protein depending on the set of mutations tested. Secondly, in contrast with the $\Delta\Delta G_{U-\ddagger}{}^{1.6M}$rel, the Φ parameter is indirectly corrected for the severity of mutations (e.g., the level of side chain volume variation due to mutation) through the $\Delta\Delta G_{F-U}$ denominator. As noted in the Results section the normalization of the $\Delta\Delta G_{U-\ddagger}{}^{1.6M}$ parameter by the variation in side chain volume upon mutation did not change dramatically the TS structure obtained. Moreover, the TS ensemble properties predicted by interpreting Φ-value in light of $\Delta\Delta G_{U-\ddagger}{}^{1.6M}$rel is in agreement with the near native placement of the TS along the folding reaction coordinate, suggested by high β_t ($\beta_t{}^{avge}$= 0.76 ± 0.04) and contribution of k_f to stability ($\Delta G_{\ddagger-U}{}^{wt}/\Delta G_{F-U}{}^{wt}$= 0.72) of Raf RBD. According to this scheme, most of the topology of Raf RBD is already determined at the TS, but with significant rearrangement in the hydrophobic core still required to reach the native state. Clearly, the interpretation of Φ-value analysis is improved by considering $\Delta\Delta G_{U-\ddagger}$rel, and we believe this contributes to building a more comprehensive view of Raf RBD TS.

Raf RBD and Ubiquitin share similar TS

Raf RBD and ubiquitin belong to the same topology and SCOP superfamily, but display insignificant sequence identity (< 12%). According to Φ-values, their TS structures are similarly consolidated around their N-terminal β-hairpin, but ubiquitin TS is less native-

like and more polarized ([31]) (Figure 6 (a) and 7 (a)). A notable differences is that β5 is not involved at all in the stabilization of ubiquitin TS. Nevertheless, we found a low but significant correlation between Raf RBD and ubiquitin Φ-value (slope= 0.51; R= 0.63, Figure S4 of Supplementary material), which can be significantly improved by ignoring the major outlier and correcting for difference in packing (R of 0.75 and 0.84, respectively). This is comparable to correlations published on 7 pairs of structurally similar proteins studied previously ([32]). Through the perspective of $\Delta\Delta G_{U-\ddagger}$ rel, ubiquitin TS extends to α1 as described for Raf RBD, but still excludes β5 and residues located in β2 contribute more (Figure 6 (b) and 7 (a)). These differences between Raf RBD and ubiquitin TS ensemble could stem from slight native structure dissimilarities due to variation in the packing of α1 over the β-sheet ([38, 39]). Indeed, the second and third inner core residue positions in α1 of Raf RBD and ubiquitin are misaligned as inner core residues L82 and L86 in the former are analogous to V26 and I30 in the latter, but are aligned with C81 and A85 according to secondary structure. This variation in α1 arrangement, produce local variations in packing of the protein, most obvious in the local contexts and role in stabilizing the native structure of residues L62 and, T68 and V70 of Raf RBD and of their matching residues in ubiquitin ([39]). In addition, it is noteworthy that the $\Delta\Delta G_{U-\ddagger}$ patterns observed for Raf RBD and ubiquitin are virtually identical to $\Delta\Delta G_{U-\ddagger}$rel (Figure S4 of supplementary material), suggesting that matching residues in the two proteins, excluding those in β5, induce comparable destabilization of the TS. We also provide the correction for volume of $\Delta\Delta G_{U-\ddagger}$rel for Raf RBD and ubiquitin to demonstrate that it does not affect drastically the properties of the TS's of these proteins (Figure S4 of supplementary material). Finally, the m_u variation upon mutation of both proteins are comparable and the most significant changes occur in a homologous structural region (Figure 6 (c)), as discussed later.

The model of the ubiquitin TS obtained using Φ-value and $\Delta\Delta G_{U-\ddagger}$rel is similar, but not identical to that predicted using engineered metal binding sites and Ψ-value analysis. This approach predicted that ubiquitin TS populates significant structure in α1 with an almost fully formed β-sheet, including β5 ([24, 53]). However, important concerns need to be

addressed concerning the validity of Ψ-value methods, before the discrepancy between diverging interpretations of TS structure can be explained ([54, 55]). Among key concerns is whether mutations of proteins alter the free-energy folding landscape by, for instance, altering TS stability and whether these changes can allow for easy interpretation of TS structure.

In summary, the comparison of TS structure of Raf RBD *versus* ubiquitin suggests that while these two proteins have very different sequences, they fold via TS's sharing several common properties whether they are viewed through Φ-value or $\Delta\Delta G_{U-\ddagger}$ analyses. However, the higher polarization of ubiquitin *versus* Raf RBD TS is in agreement with the lower β_t (0.66 *vs* 0.76) and $\Delta G_{\ddagger-U}^{wt}/\Delta G_{F-U}^{wt}$ (0.56 *vs* 0.72). The similarities between the two proteins at the kinetic level are matched by the data on stabilization of their native structure ([39]). The similarities in the entropy profile of the experimentally obtained Raf RBD sequences and natural proteins classified in 5 ubiquitin-related superfamilies suggest that the native state stabilization and formation determinants could be broadly conserved.

Comparison of Raf RBD TS with more distant structural analogues: protein-L and G

Protein-L and G, which have been subjected to Φ-value analysis ([17, 18]), are classified in the IgG binding domain superfamily of the β-grasp ubiquitin-like topology and thus are distant structural analogues of Raf RBD. The protein-G and L nucleus are located in one of two, symmetrically disposed C and N-terminal β-hairpins. In contrast to Raf RBD and ubiquitin, in protein-L and G the role of the hydrophobic core is reduced in favor of topology forming elements, specifically the β-turns. The dispersion in protein-L structure of residues with Φ-value distribution is slightly more polarized than that of Protein-G and corresponds more closely to ubiquitin and Raf RBD TS. However, the TS structures for these IgG binding domains, does not change dramatically by reinterpreting the TS in terms of $\Delta\Delta G_{U-\ddagger}$ rel (Figure 7 (b) and (c)), unlike what we described for Raf RBD and ubiquitin. Particularly, the role of the major α-helix in Protein-G and Protein-L, which is analogous to α1 although there are differences in length and packing, appears drastically diminished. Further evidence obtained from sequence perturbation and the introduction of several Gly

into this α-helix confirmed the disruption of this secondary structure element at the TS of protein-L ([46]). The IgG binding domain and ubiquitin-like superfamily members share the same succession of secondary elements, but the IgG binding domain β-sheet and α-helix are extended longitudinally and their axes are almost parallel. Taken together, these observations indicate the inherent complexities in relating structural analogues and superfamilies and highlight how fold classification may be challenged by folding kinetics. From this perspective, it would be interesting to examine whether the TS structures of Raf RBD and ubiquitin are conserved across the various superfamilies of the β-grasp ubiquitin-like topology.

Literature cases: reevaluation of TS structure using $\Delta\Delta G_{U\text{-}\ddagger}$rel

In light of results presented here, we reasoned that it would be interesting to reevaluate the properties of TS ensembles by comparing Φ-value and $\Delta\Delta G_{U\text{-}\ddagger}$rel, looking specifically for evidence of drift from polarized to diffuse TS in proteins adopting other types of fold. For most of the proteins scrutinized, the change in TS structure by using $\Delta\Delta G_{U\text{-}\ddagger}$rel was minor. However, we found one example in which the TS of a cold-shock protein was described to be polarized based on Φ parameter ([30]), but appear more diffuse according to $\Delta\Delta G_{U\text{-}\ddagger}$rel (Figure 7 (d)). Such a delocalized TS would correspond better to the extremely high β_t (0.9) measured for this cold-shock protein.

Description of the folding pathway

Based on our data, we can make a first attempt at describing the sequence of events along the Raf RBD folding pathway (Figure 8). The Φ-value analyses indicate that β2 and the central part of β1 (e.g. V60) adopt the most native conformation in the TS ensemble, whereas α1 adopts a partially native-like structure with the strongest contacts in the segment D80-L86. The relatively low Φ-value of α1 inner core residues (e.g., L78, L82 and L86) and of V98 (β3), L126, V128 (β5) and L62 (β1) suggests that the α-helix is not properly packed against the β-sheet. However, the observation that all these inner core elements contribute significantly to $\Delta G_{U\text{-}\ddagger}$ (β1< β5< β3) suggests that they have formed numerous stabilizing interactions at the TS. All probes between C96-L121 (excluding

V98A) suggest that β3, β4 and α2 are mostly in a denatured-like conformation in the activated state.

As described in the Results section, a change in m_u induced upon a few given Raf RBD mutations produced an apparent TS shift, in agreement with Hammond behavior (Figure 6 (c)). Significant decreases in m_u were noted for Ala mutations at L62, P63 (β-turn1) and L86 (α1) and also at L126, V128A (β5), which are all proximally located in Raf RBD tertiary structure. This small variation in m_u indicates a maximum reduction in solvent exposure of the TS of 33% in comparison with the native state (e.g., considering that $m_u^{L62,L86}$= 0.27, m_u^{wt}= 0.41) (Table 1). We hypothesize that this might indicate that following nucleus formation centered on the β-hairpin, the Raf RBD polypeptide chain proceeds to partial consolidation of native hydrophobic contacts between the C-termini of α1 and β1 and possibly also β5 (Figure 8). The high Φ-value (≈ 1) observed for the stabilized variant R89L indicates that better hydrophobic packing at this end of α1 improved folding efficiency. Interestingly, R89 is adjacent in the tertiary structure to L86 and the β-hairpin through direct contacts with L62. The chain of events in this proposed pathway is also in agreement with the lower Φ-value of L78A, located at the N-terminal extremity of α1.

As described above, Φ-value analyses and $\Delta\Delta G_{U-\ddagger}$ reveal similar polarization of native-like contacts in the N-terminal β-hairpin and energetically diffuse TS of ubiquitin and Raf RBD, although the former is more polarized by excluding β5 from the rate limiting step (Figure 6(a)-(c) and 7 (a)). The amino terminal β-hairpin of ubiquitin has been found to fold independently in the 20 μs range ([56, 57]), at least one order of magnitude faster than the full-length protein, suggesting that the consolidation of native structure in this region could well be the first significant step on the folding pathway. Moreover, the decrease in m_u observed for certain variants of Raf RBD are matched in ubiquitin although the latter also shows specific variation upon mutation in β3 and β2 (Figure 6 (c)). These resemblances of the m_u profiles suggest that α1 of both proteins pass through similar processes in consolidating their structure at the TS. A simulation using C_α Go-type models of ubiquitin suggested that

the C-terminus of α1 docks to the β-sheet, precisely at the level of the β-turn, β2 and β3 and that the β5 adopts native state contacts only late in folding ([58]).The results of Ψ-value analysis on ubiquitin are also in agreement with the C-terminal part of α1 being preferentially consolidated at the TS ([53]). Taken together these results suggest that ubiquitin and Raf RBD reach their native structure through highly similar folding pathways.

Atypical mutants and input of sequence perturbation experiments

A polemic has arisen concerning the level of conservation of residues most important to stabilization of the TS, often referred to as the nucleus ([59-61]). Our results have shown that positions displaying low sequence entropy in the sequence perturbation experiments roughly encompassed residues of Raf RBD displaying high $\Delta\Delta G_{U-\ddagger}$rel, but not all of those with high Φ-value (Figure 4 (c)). This is not likely to be a general principle as it is not true for ubiquitin, which folds through a more polarized TS. However, we have brought direct experimental evidence that stability is the predominant factor in evolutionary pressure, by demonstrating that positional entropy in Raf RBD is correlated with the level of destabilization of the native structure induced by the insertion of truncating mutations at the corresponding residues ([39]).

In addition, a series of residues showed very dominant occurrence bias for non-wt amino acids in the perturbation study. We reasoned that these might indicate mutations that could improve the folding or stability of Raf RBD. In this study, the folding/unfolding kinetics of N56M, I58L, S77T, C81I, C96L and C96M were determined. The most interesting results were obtained with S77T, which showed improved stability mainly through a decrease in k_u ([39]). The others had only marginal effects on stability and folding rate.

Finally, three other mutants (e.g. R89L, H2 and Δ104-6) showed a dramatic increase in folding rate accompanied by an increase in stability (Table 1 and 2). R89L abrogates binding to *ras* and is the fastest folding variant with over one order of magnitude improvement of k_f over the wt. Interestingly, α1 in Raf RBD is apparently remodeled upon

binding to Rap1A or a Rap1A variant mimicking *ras* ([62, 63]), as indicated by the non-canonical structure it adopts between C81 and A85. Therefore, the sub-optimal packing of the α-helix might be necessary for tight binding to *ras*. The mutant Δ104-106 is the most stabilized mutant and both k_f and k_u are improved over wt, which is striking given the deletion of 3 amino acids peripheral to the core and the putatively structured regions in the TS (Table 3 and Figure 4). It would be interesting to determine how the TS ensemble has changed by protein engineering in the background of stabilized mutants, particularly for Δ104-6. A more thorough discussion of thermodynamic effect of these mutations and thermodynamic/kinetic data in Gdm-HCl for protein variants combining these mutations are presented in the companion manuscript ([39]).

Conclusions and implications for models of protein folding

In summary, we have described the TS of c-Raf/Raf-1 RBD with a protein engineering approach. The TS structure is built using classic Φ-value interpretation and comparison with the newly introduced parameter $\Delta\Delta G_{U-\ddagger}$rel. The TS ensemble appears polarized around the β-hairpin according to the Φ-value, but delocalized to all inner hydrophobic core residues according to the $\Delta\Delta G_{U-\ddagger}$rel, indicating a dominant role for the hydrophobic core in the stabilization of the TS. Interestingly, TS's and proposed folding pathways for Raf RBD and ubiquitin are highly similar, despite insignificant sequence identity (< 12%). This observation is in agreement with the statement that the folding mechanism is defined by coarse amino acid composition and topological characteristics rather than by fine sequence details.

Φ-value analyses have been used successfully as constraints in folding simulations to build models of TS ([64-67]). However, our results suggest that for some proteins, such as Raf RBD, ubiquitin and the cold-shock protein, the addition of the $\Delta\Delta G_{U-\ddagger}$ parameter to Φ-value analysis in the modeling scheme could improve our capacity to represent the TS ensemble and the folding pathway. Also, in the perspective of TS structural description from an energy point of view, the protein engineering assumption concerning the dominance of native over non-native contacts in stabilizing TS becomes less necessary for

interpretation of the data, consequently emphasizing the putative role of non-native contacts in TS stabilization. Experimental data from various sources have suggested a significant role for non-native contacts in TS stabilization ([20, 68-70]). No Ala/Gly mutations introduced in Raf RBD produce coherent Φ-value estimates less than 0, which are thought to indicate the destruction of non-native contacts that stabilize the TS ([71, 72]). In theory, a residue forming TS stabilizing non-native contacts, but involved in native contacts equal to or superior in energy contribution to the native state would not display atypical Φ-values. In the case of Raf RBD, it is difficult to imagine a model in which V70 and V72 could be forming virtually all of their native contacts, while the inner core residues located in β1 (I58 and V60) and α1 (L78, L82, L86) and α2 (A118) have formed barely half of them. Interestingly, a molecular dynamics simulation has shown that the native-likeliness of TS was overestimated if non-native contacts were excluded ([26]). Currently, there is not enough experimental evidence reported in the literature to judge on the generality of this phenomenon ([25]).

Materials and Methods

Entropy calculation

Sequence entropy was calculated using a modified version of the Shannon entropy formula ([73]), with experimental data and the ubiquitin-roll topology alignment ([38, 39]).

Mutants cloning and description

Mutants of human Raf-1/c-Raf RBD were synthesized with a variation of the ExSite™ protocol (Stratagene) using the high-fidelity Pfu polymerase. Most variants synthesized carry a single point mutation. The mutant H2 was recovered from the sequence perturbation experiment. In this mutant, residues 62-65 (Leu-Pro-Asn-Lys) of Raf RBD are replaced with the sequence Phe-Thr-Asp-Gly. Mutant H2_F62L reverts residue 62 to the wt amino acid (Leu-Thr-Asp-Gly). Mutants Δ104-6 and Δ101-8+AG are deletion mutants. For Δ104-6 residues 104 to 106 (e.g., E104-H105-K106) are deleted. For Δ101-8+AG amino acids 101 to 108 are replaced by Ala-Gly, as in ubiquitin. The inserted mutation was

confirmed by sequencing. The protein expressed included residues 55-132 of Raf-1 plus an amino-terminal hexahistidine tag separated by a spacer of 3 amino acids (Ser-Met-Gly). Proteins were purified from bacterial cell lysate under denaturing conditions in 9M urea, on a Ni-NTA column.

Kinetics and chevron curves

The kinetic experiments were designed in agreement with the consensus experimental standards ([41]). The reactions were followed using Applied Photophysics SX.18MV stopped-flow fluorimeter. The experiments were performed at 25 (±0.1)°C in 50mM sodium phosphate buffer (pH=7.0) containing 1 mM DTT, using urea as denaturant. Refolding and unfolding traces were followed by excitation at 280 nm of the single tryptophan (W114) and detection of the emission signal using a 320 nm cut-off filter. We used the 20 μl mixing chamber and 1:10 mixing volume, resulting in a mixing dead-time of approximately 0.7 msec. Data between 3.5 msec and 10 sec was used for fitting traces. Protein concentration was set to be at 3-4 μM after dilution. The refolding reactions were initiated from proteins diluted in 9-9.75 M urea. The unfolding reactions were initiated from proteins diluted in 1-2 M urea (varying according to mutant stability) to avoid aggregation. Typically, 4 to 7 traces were averaged for each data point. Refolding traces at low denaturant concentration were fit to 4 exponentials as described previously for wt Raf RBD ([40, 41]). No clear deviation from linearity of denaturant dependence of $\ln k_f$ was observed at either end of the chevron plot for all mutants tested, except for V72A. This variant showed a moderate roll-over at low concentration of denaturant on the refolding arm (data not shown) and consequently, these data points were ignored in chevron curve fitting. The unfolding traces were fit to simple exponentials. The introduction of supplementary phases into the fitting of either reaction did not decrease the residual significantly, except in the case of the unfolding traces of C96M. For this mutant, the fitting of the unfolding traces with a double exponential resulted in improved modeling of the chevron curve (data not shown). All chevron curves were fit to two-state equations ([74]). Kinetically derived $\Delta G_{F\text{-}U}$ and m were calculated from k_f^{H2O} and k_u^{H2O} and m_f and m_u, respectively, using classical equations ([74]).

$\Delta\Delta G_{F\text{-}U}$ and Φ-value estimates

A study on protein-L presented a rigorous analysis of Φ-value estimates based on three independent calculations of free energy changes ([17]). These methods are used to minimize extrapolation errors in determination of the parameters obtained from the kinetic and thermodynamic experiments, which are used in Φ-value calculation ($\Phi = \Delta\Delta G_{U-\ddagger}/\Delta\Delta G_{F-U}$). Φ is particularly sensitive to errors in the $\Delta\Delta G_{F-U}$ parameter. Therefore, three equations were used to calculate free energy changes:

$$\Delta\Delta G_{F-U}^{kin} = RT\left(\ln\left(k_f^{mut(1.6M)}/k_u^{mut(8M)}\right) - \ln\left(k_f^{wt(1.6M)}/k_u^{wt(8M)}\right)\right) \quad (1)$$

$$\Delta\Delta G_{F-U}^{Cm} = \langle m\rangle\left(Cm^{mut} - Cm^{wt}\right) \quad (2)$$

$$\Delta\Delta G_{F-U}^{5.8M} = m^{mut}\left(Cm^{mut} - 5.8\right) - m^{wt}\left(Cm^{wt} - 5.8\right) \quad (3)$$

where <m> is the average m value for all the mutants (3.90 (±0.33) kj mol^{-1} M^{-1}), Cm^{wt} and Cm^{mut} are the concentrations of urea at which 50% of wt and mutant proteins are folded, $k_f^{wt\ (1.6M)}$ and $k_f^{mut\ (1.6M)}$ are the folding rates in 1.6 M urea of the wt and mutants, respectively. Similarly, $k_u^{wt\ (8M)}$ and $k_u^{mut\ (8M)}$ are the unfolding rates in 8 M urea for the wt and mutants, respectively. As previously suggested, mutants which produced insignificant change in $\Delta\Delta G_{F-U}$ (< |2| kJ mol^{-1}) were ignored in the Φ-value analyses ([23]).

Using these free energy changes, three different Φ-values estimates were calculated:

$$\Phi_F^{kin} = -RT\ln\left(k_f^{wt(1.6M)}/k_f^{mut(1.6M)}\right)/\Delta\Delta G_{F-U}^{kin} \quad (4)$$

$$\Phi_F^{Cm} = -RT\ln\left(k_f^{wt}/k_f^{mut}\right)/\Delta\Delta G_{F-U}^{Cm} \quad (5)$$

$$\Phi_U = -RT\ln\left(k_u^{wt(5.8M)}/k_u^{mut(5.8M)}\right)/\Delta\Delta G_{U-F}^{5.8M} \quad (6)$$

where k_f^{wt} and k_f^{mut} are the folding rates in water extrapolated from kinetic data for the wt and mutants respectively, $k_u^{wt(5.8M)}$ and $k_u^{mut(5.8M)}$ are the folding rates in 5.8 M urea for wt and mutants, respectively. $1-\Phi_U$ provides an estimate of Φ_F from the unfolding data following the micro-reversibility principle. The concentration of urea at which k_f and k_u were determined for utilization in the calculation of the various $\Delta\Delta G_{F-U}$ and Φ-value estimates were chosen based on the characteristic of the wt chevron curves. Concentration in urea of 1.6, 5.8 and 8 M, which corresponds to the lower limit of resolution by direct measurement of the folding rate, the lower limit of the transition region and roughly the middle of the unfolding arm, respectively, were chosen to balance the advantage of

minimizing extrapolation with the trade-off inherent to using non-extrapolated parameters in the Φ-value calculations. Φ_F^{ext} mentioned in the text correspond to Φ_F calculated from fully extrapolated thermodynamic and kinetic data, i.e. k_f^{H2O} and $\Delta\Delta G_{F\text{-}U}$ (Table S1 of Supplementary material).

$\Delta\Delta G_{U\text{-}‡}$ and $\Delta\Delta G_{U\text{-}‡}$ rel calculation

The $\Delta\Delta G_{U\text{-}‡}^{1.6M}$ is obtained from:

$$\Delta\Delta G_{U-‡}^{1.6M} = -RT \ln\left(k_f^{wt(1.6M)} / k_f^{mut(1.6M)}\right) \quad (7)$$

$\Delta\Delta G_{U\text{-}‡}$ rel is simply the normalization of $\Delta\Delta G_{U\text{-}‡}^{1.6M}$ obtained for a mutant against the most TS destabilized mutant ($\Delta\Delta G_{U\text{-}‡}^{1.6M}$ max) obtained in a collection of variants engineered for a given protein:

$$\Delta\Delta G_{U-‡}^{1.6M}\,\text{rel} = \Delta\Delta G_{U-‡}^{1.6M} \Big/ \Delta\Delta G_{U-‡}^{1.6M}\,\text{max} \quad (8)$$

In the case of Raf RBD and ubiquitin, the maximum destabilization of the transition state is observed with I58A and L15A (equivalent to V70 in Raf RBD), respectively. For Figure S3, the $\Delta\Delta G_{U\text{-}‡}$ of Raf RBD and ubiqutin mutants were normalized for side-chain volume variation according to the volume scale of Richards ([75]) and by arbitrarily giving to mutation A→G, a correction factor of 1. According to this mutation A→G reduced side-chain volume of 25.95 $Å^3$ and V→A, for example, of 37.8 $Å^3$. Therefore, the correction factor for the latter is (37.8/25.95)= 1.46.

Manipulation of the data of previous studies: Φ-values and $\Delta\Delta G_{U\text{-}‡}$ of other proteins discussed

Ubiquitin average Φ-value obtained from unfolding and folding experiments were used in the graph and structural scheme for comparison with Raf RBD. The $\Delta\Delta G_{U\text{-}‡}$ were calculated from k_f^{H2O} ([31]). For protein-L the equivalent of Φ_F^{kin} was utilized. $\Delta\Delta G_{U\text{-}‡}$ were calculated from $k_f^{0.4M}$, value used for the calculation of the previous parameter ([17]). For protein-G the equivalent of Φ_F^{kin} was utilized. $\Delta\Delta G_{U\text{-}‡}$ were calculated from $k_f^{0.5M}$, value used for the calculation of the previous parameter ([18]). For the cold shock protein mean Φ-value were obtained from averaging Φ-values obtained from thermodynamic and refolding

kinetics or from unfolding kinetics in some exceptional cases, and Φ-value determined solely from kinetic experiments, respectively. The $\Delta\Delta G_{U-\ddagger}$ were calculated from k_f^{H2O} ([30]).

Structure representation

Structural representations were created using MolMol software and the following molecular coordinates recovered from the protein databank: 1RFA (Raf RBD), 1UBI (ubiquitin), 1HZ6 (protein-L), 2GB1 (protein-G) and 1CSP (cold shock protein). Hydrophobic core residues, in which lateral chains are represented in the figures were identified using their regular residues numbering, except in the case of ubiquitin. To facilitate comparison, we decided to identify the residues of ubiquitin using the Raf RBD residues numbering and their sequence alignment guided by secondary structure. Details on the alignment and structural comparisons between Raf RBD and ubiquitin can be found elsewhere ([38, 39]).

Acknowledgements

The authors thank: Dr Jeffrey W. Keillor for collaboration and access to stopped-flow apparatus; Alexis Vallée-Belisle for discussions and data exchange; Emily Manderson for carefully reading this manuscript. The NSERC and CIHR funded this project. FXCV is a scholar of CIHR, le programme de biologie moléculaire and the FES. SWM is the Canada Research Chair in Integrative Genomics.

References

1. Jackson, S. E. & Fersht, A. R. (1991). Folding of chymotrypsin inhibitor 2. 1. Evidence for a two state transition. Biochemistry 30, 10428-10435.
2. Matouschek, A., Kellis, J. T., Jr., Serrano, L. & Fersht, A. R. (1989). Mapping the transition state and pathway of protein folding by protein engineering. Nature 340, 122-126.
3. Fersht, A. R., Matouschek, A. & Serrano, L. (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. J. Mol. Biol. 224, 771-782.
4. Serrano, L., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme. III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure. J. Mol. Biol. 224, 805-818.
5. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. J. Mol. Biol. 254, 260-288.
6. Kragelund, B. B., Osmark, P., Neergaard, T. B., Schiodt, J., Kristiansen, K., Knudsen, J. & Poulsen, F. M. (1999). The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. Nat. Struct. Biol. 6, 594-601.

7. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. Nat. Struct. Biol. 6, 1016-1024.
8. Martinez, J. C. & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. Nat. Struct. Biol. 6, 1010-1016.
9. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.* **6**, 1005-1009.
10. Ternstrom, T., Mayor, U., Akke, M. & Oliveberg, M. (1999). From snapshot to movie: phi analysis of protein folding transition states taken one step further. *Proc. Natl. Acad. Sci. U.S.A.*. **96**, 14854-14859.
11. Otzen, D. E. & Oliveberg, M. (2002). Conformational plasticity in folding of the split beta-alpha-beta protein S6: evidence for burst-phase disruption of the native state. *J. Mol. Biol.* **317**, 613-627.
12. Jager, M., Nguyen, H., Crane, J. C., Kelly, J. W. & Gruebele, M. (2001). The folding mechanism of a beta-sheet: the WW domain. *J. Mol. Biol.* **311**, 373-393.
13. Deechongkit, S., Nguyen, H., Powers, E. T., Dawson, P. E., Gruebele, M. & Kelly, J. W. (2004). Context-dependent contributions of backbone hydrogen bonding to beta-sheet folding energetics. *Nature* **430**, 101-105.
14. Hamill, S. J., Steward, A. & Clarke, J. (2000). The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* **297**, 165-178.
15. Fowler, S. B. & Clarke, J. (2001). Mapping the folding pathway of an immunoglobulin domain: structural detail from Phi value analysis and movement of the transition state. *Structure. (Camb.)* **9**, 355-366.
16. Wright, C. F., Lindorff-Larsen, K., Randles, L. G. & Clarke, J. (2003). Parallel protein-unfolding pathways revealed and mapped. *Nat. Struct. Biol.* **10**, 658-662.
17. Kim, D. E., Fisher, C. & Baker, D. (2000). A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298**, 971-984.
18. McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of beta-hairpin formation in protein G folding. *Nat. Struct. Biol.* **7**, 669-673.
19. Northey, J. G., Di Nardo, A. A. & Davidson, A. R. (2002). Hydrophobic core packing in the SH3 domain folding transition state. *Nat. Struct. Biol.* **9**, 126-130.
20. Capaldi, A. P., Kleanthous, C. & Radford, S. E. (2002). Im7 folding mechanism: misfolding on a path to the native state. *Nat. Struct. Biol.* **9**, 209-216.
21. Gianni, S., Guydosh, N. R., Khan, F., Caldas, T. D., Mayor, U., White, G. W., DeMarco, M. L., Daggett, V. & Fersht, A. R. (2003). Unifying features in protein-folding mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13286-13291.
22. Ozkan, S. B., Bahar, I. & Dill, K. A. (2001). Transition states and the meaning of Phi-values in protein folding kinetics. *Nat. Struct. Biol.* **8**, 765-769.
23. Sanchez, I. E. & Kiefhaber, T. (2003). Origin of unusual phi-values in protein folding: evidence against specific nucleation sites. *J. Mol. Biol.* **334**, 1077-1085.
24. Sosnick, T. R., Dothager, R. S. & Krantz, B. A. (2004). Differences in the folding transition state of ubiquitin indicated by phi and psi analyses. *Proc. Natl. Acad. Sci. U.S.A.*. **101**, 17377-17382.
25. Feng, H., Vu, N. D., Zhou, Z. & Bai, Y. (2004). Structural examination of phi-value analysis in protein folding. *Biochemistry* **43**, 14325-14331.
26. Settanni, G., Rao, F. & Caflisch, A. (2005). Phi-value analysis by molecular dynamics simulations of reversible folding. *Proc. Natl. Acad. Sci. U.S.A.*. **102**, 628-633.
27. Los Rios, M. A., Muralidhara, B. K., Wildes, D., Sosnick, T. R., Marqusee, S., Wittung-Stafshede, P., Plaxco, K. W. & Ruczinski, I. (2006). On the precision of experimentally determined protein folding rates and phi-values. *Protein Sci.* **15**, 553-563.
28. Fersht, A. R., Itzhaki, L. S., elMasry, N. F., Matthews, J. M. & Otzen, D. E. (1994). Single versus parallel pathways of protein folding and fractional formation of structure in the transition state. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 10426-10429.
29. Sanchez, I. E. & Kiefhaber, T. (2003). Hammond behavior versus ground state effects in protein folding: evidence for narrow free energy barriers and residual structure in unfolded states. *J. Mol. Biol.* **327**, 867-884.
30. Garcia-Mira, M. M., Boehringer, D. & Schmid, F. X. (2004). The folding transition state of the cold shock protein is strongly polarized. *J. Mol. Biol.* **339**, 555-569.

31. Went, H. M. & Jackson, S. E. (2005). Ubiquitin folds through a highly polarized transition state. *Protein Eng Des Sel* **18**, 229-237.
32. Zarrine-Afsar, A., Larson, S. M. & Davidson, A. R. (2005). The family feud: do proteins with similar structures fold via the same pathway? *Curr. Opin. Struct. Biol.* **15**, 42-49.
33. Larson, S. M. & Pande, V. S. (2003). Sequence optimization for native state stability determines the evolution and folding kinetics of a small protein. *J. Mol. Biol.* **332**, 275-286.
34. Geierhaas, C. D., Paci, E., Vendruscolo, M. & Clarke, J. (2004). Comparison of the transition states for folding of two Ig-like proteins from different superfamilies. *J. Mol. Biol.* **343**, 1111-1123.
35. Wellbrock, C., Karasarides, M. & Marais, R. (2004). The RAF proteins take centre stage. *Nat. Rev. Mol. Cell Biol.* **5**, 875-885.
36. Emerson, S. D., Madison, V. S., Palermo, R. E., Waugh, D. S., Scheffler, J. E., Tsao, K. L., Kiefer, S. E., Liu, S. P. & Fry, D. C. (1995). Solution structure of the Ras-binding domain of c-Raf-1 and identification of its Ras interaction surface. *Biochemistry* **34**, 6911-6918.
37. Soding, J. & Lupas, A. N. (2003). More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* **25**, 837-846.
38. Campbell-Valois, F. X., Tarassov, K. & Michnick, S. W. (2005). Massive Sequence Perturbation of a Small Protein. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 14988-14993.
39. Campbell-Valois, F. X., Tarassov, K. & Michnick, S. W. (2006). Massive sequence perturbation of the Raf *ras* binding domain reveals relationships between sequence positional entropy, secondary structure propensity, hydrophobic core volume conservation and stability, Submitted to JMB.
40. Vallee-Belisle, A., Turcotte, J. F. & Michnick, S. W. (2004). raf RBD and Ubiquitin Proteins Share Similar Folds, Folding Rates and Mechanisms Despite Having Unrelated Amino Acid Sequences. *Biochemistry* **43**, 8447-8458.
41. Maxwell, K. L., Wildes, D., Zarrine-Afsar, A., Los Rios, M. A., Brown, A. G., Friel, C. T., Hedberg, L., Horng, J. C., Bona, D., Miller, E. J., Vallee-Belisle, A., Main, E. R., Bemporad, F., Qiu, L., Teilum, K., Vu, N. D., Edwards, A. M., Ruczinski, I., Poulsen, F. M., Kragelund, B. B., Michnick, S. W., Chiti, F., Bai, Y., Hagen, S. J., Serrano, L., Oliveberg, M., Raleigh, D. P., Wittung-Stafshede, P., Radford, S. E., Jackson, S. E., Sosnick, T. R., Marqusee, S., Davidson, A. R. & Plaxco, K. W. (2005). Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci.* **14**, 602-616.
42. Matouschek, A. & Fersht, A. R. (1993). Application of physical organic chemistry to engineered mutants of proteins: Hammond postulate behavior in the transition state of protein folding. *Proc. Natl. Acad. Sci. U.S.A.*. **90**, 7814-7818.
43. Matouschek, A., Otzen, D. E., Itzhaki, L. S., Jackson, S. E. & Fersht, A. R. (1995). Movement of the position of the transition state in protein folding. *Biochemistry* **34**, 13656-13662.
44. Matthews, J. M. & Fersht, A. R. (1995). Exploring the energy surface of protein folding by structure-reactivity relationships and engineered proteins: observation of Hammond behavior for the gross structure of the transition state and anti-Hammond behavior for structural elements for unfolding/folding of barnase. *Biochemistry* **34**, 6805-6814.
45. Sosnick, T. R., Jackson, S., Wilk, R. R., Englander, S. W. & DeGrado, W. F. (1996). The role of helix formation in the folding of a fully alpha-helical coiled coil. *Proteins* **24**, 427-432.
46. Kim, D. E., Yi, Q., Gladwin, S. T., Goldberg, J. M. & Baker, D. (1998). The single helix in protein L is largely disrupted at the rate-limiting step in folding. *J. Mol. Biol.* **284**, 807-815.
47. Aurora, R. & Rose, G. D. (1998). Helix capping. *Protein Sci.* **7**, 21-38.
48. Northey, J. G., Maxwell, K. L. & Davidson, A. R. (2002). Protein folding kinetics beyond the phi value: using multiple amino acid substitutions to investigate the structure of the SH3 domain folding transition state. *J. Mol. Biol.* **320**, 389-402.
49. Block, C., Janknecht, R., Herrmann, C., Nassar, N. & Wittinghofer, A. (1996). Quantitative structure-activity analysis correlating Ras/Raf interaction in vitro to Raf activation in vivo. *Nat. Struct. Biol.* **3**, 244-251.
50. Wong, K. B., Clarke, J., Bond, C. J., Neira, J. L., Freund, S. M., Fersht, A. R. & Daggett, V. (2000). Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding. *J. Mol. Biol.* **296**, 1257-1282.

51. Kazmirski, S. L., Wong, K. B., Freund, S. M., Tan, Y. J., Fersht, A. R. & Daggett, V. (2001). Protein folding from a highly disordered denatured state: the folding pathway of chymotrypsin inhibitor 2 at atomic resolution. *Proc. Natl. Acad. Sci. U.S.A.*. **98**, 4349-4354.
52. Religa, T. L., Markson, J. S., Mayor, U., Freund, S. M. & Fersht, A. R. (2005). Solution structure of a protein denatured state and folding intermediate. *Nature* **437** , 1053-1056.
53. Krantz, B. A., Dothager, R. S. & Sosnick, T. R. (2004). Discerning the structure and energy of multiple transition states in protein folding using psi-analysis. *J. Mol. Biol.* **337**, 463-475.
54. Fersht, A. R. (2004). Phi value versus psi analysis. *Proc. Natl. Acad. Sci. U.S.A.*. **101**, 17327-17328.
55. Bodenreider, C. & Kiefhaber, T. (2005). Interpretation of protein folding psi values. *J. Mol. Biol.* **351**, 393-401.
56. Searle, M. S., Williams, D. H. & Packman, L. C. (1995). A short linear peptide derived from the N-terminal sequence of ubiquitin folds into a water-stable non-native beta-hairpin. *Nat. Struct. Biol.* **2**, 999-1006.
57. Zerella, R., Evans, P. A., Ionides, J. M., Packman, L. C., Trotter, B. W., Mackay, J. P. & Williams, D. H. (1999). Autonomous folding of a peptide corresponding to the N-terminal beta-hairpin from ubiquitin. *Protein Sci.* **8**, 1320-1331.
58. Zhang, J., Qin, M. & Wang, W. (2005). Multiple folding mechanisms of protein ubiquitin. *Proteins* **59**, 565-579.
59. Plaxco, K. W., Larson, S., Ruczinski, I., Riddle, D. S., Thayer, E. C., Buchwitz, B., Davidson, A. R. & Baker, D. (2000). Evolutionary conservation in protein folding kinetics. *J. Mol. Biol.* **298**, 303-312.
60. Mirny, L. & Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* **308**, 123-129.
61. Larson, S. M., Ruczinski, I., Davidson, A. R., Baker, D. & Plaxco, K. W. (2002). Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *J. Mol. Biol.* **316**, 225-233.
62. Nassar, N., Horn, G., Herrmann, C., Scherer, A., McCormick, F. & Wittinghofer, A. (1995). The 2.2 A crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with Rap1A and a GTP analogue. *Nature* **375**, 554-560.
63. Nassar, N., Horn, G., Herrmann, C., Block, C., Janknecht, R. & Wittinghofer, A. (1996). Ras/Rap effector specificity determined by charge reversal. *Nat. Struct. Biol.* **3**, 723-729.
64. Vendruscolo, M., Paci, E., Dobson, C. M. & Karplus, M. (2001). Three key residues form a critical contact network in a protein folding transition state. *Nature* **409**, 641-645.
65. Li, L. & Shakhnovich, E. I. (2001). Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations. *Proc. Natl. Acad. Sci. U.S.A.*. **98**, 13014-13018.
66. Paci, E., Clarke, J., Steward, A., Vendruscolo, M. & Karplus, M. (2003). Self-consistent determination of the transition state for protein folding: application to a fibronectin type III domain. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 394-399.
67. Hubner, I. A., Edmonds, K. A. & Shakhnovich, E. I. (2005). Nucleation and the transition state of the SH3 domain. *J. Mol. Biol.* **349**, 424-434.
68. Canet, D., Lyon, C. E., Scheek, R. M., Robillard, G. T., Dobson, C. M., Hore, P. J. & van Nuland, N. A. (2003). Rapid formation of non-native contacts during the folding of HPr revealed by real-time photo-CIDNP NMR and stopped-flow fluorescence experiments. *J. Mol. Biol.* **330**, 397-407.
69. Ventura, S., Vega, M. C., Lacroix, E., Angrand, I., Spagnolo, L. & Serrano, L. (2002). Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nat. Struct. Biol.* **9**, 485-493.
70. Friel, C. T., Beddard, G. S. & Radford, S. E. (2004). Switching two-state to three-state kinetics in the helical protein Im9 via the optimisation of stabilising non-native interactions by design. *J. Mol. Biol.* **342**, 261-273.
71. Martinez, J. C., Pisabarro, M. T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol.* **5**, 721-729.
72. Li, L., Mirny, L. A. & Shakhnovich, E. I. (2000). Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. *Nat. Struct. Biol.* **7**, 336-342.
73. Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**, 56-68.

74. Zarrine-Afsar, A. & Davidson, A. R. (2004). The analysis of protein folding kinetic data produced in protein engineering experiments. *Methods* **34**, 41-50.
75. Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1-14.

Abbreviations used: RBD, *ras* binding domain; TS, transition state; β1, β-strand 1; β2, β-strand 2; α1, α-helix; β3, β-strand 3; β4, β-strand 4; β5, β-strand 5; Gdm-HCl, guanidinium-hydrochloride; AcP, acylphosphatase; RCO, relative contact order.

Email address of the corresponding author: stephen.michnick@umontreal.ca

Figure legends, Figures and Tables:

Figure 1. (a) Primary, secondary structure and (b) tertiary structure of Raf RBD. (c) Comparison of normalized entropy of Raf RBD obtained experimentally *versus* an alignment of 54 protein structures sharing the ubiquitin topology. (d) Hierarchical arrangement of the hydrophobic core into an inner (red) and outer core (green) as revealed by the sequence perturbation experiment.

Figure 2. (a) Chevron curves for wt Raf RBD and dependence on urea concentration of the lnk_{obs} of additional phases modeled in the fitting of refolding traces. (b) β1: N56M (□), I58A (○), I58L (◊), I58F (Δ), R59A (∇) and V60A (x). (c) β-Turn1: L62A (□), P63A (○), N64A (◊), H2 (Δ) and H2_F62L (∇). (d) β2: Q66A (□), T68A (○), V69A (◊), V70A (Δ), V72A (∇), V72I (x). (e) N-t capping and N-terminus of α1: M76A (□), S77A (○), S77T (◊), L78A (Δ) and D80A (∇). (f) C-terminus of α1: C81A (□), C81I (○), L82A (◊), A85G (Δ), L86A (∇) and R89L (x). (g) Loop following the α-helix: L91A (□), P93A (○), C95A (◊), C96A (Δ), C96L (∇) and C96M (x). (h) β3 and the following loop: A97G (□), V98A (○), R100A (◊), E104A (Δ), Δ104-6 (∇) and Δ101-8 (x). (i) β4, loop and α2: K109A (□), L112A (○), D117A (◊), A118G (Δ) and A118L (∇). (j) β5: L121A (□), E124A (○), E125A (◊), L126A (Δ), V128A (∇) and D129A (x). In all panels, the modeled wt chevron curve is shown for comparison (grey line).

Figure 3. (a) Plot of m_f (○) or m_u (●) versus m_f+m_u. A very high correlation is seen with m_f (slope= 0.96; R= 0.93) indicating that denatured state is sensible to mutation. Only a poor correlation (slope= 0.40; R= 0.18) is seen for m_u indicating insignificant change in the native state. The outlier L82A was removed from both graphs. (b) Plot of the difference in the free energy of folding between mutant and Wt determined from equilibrium denaturation, $\Delta\Delta G_{F\text{-}U}^{Cm}$, *versus* the difference in energy calculated from kinetics, $\Delta\Delta G_{F\text{-}U}^{kin}$. The thermodynamic and kinetic data correspond well to the value expected for a two-state folding model (slope= 0.98, R= 0.96).

Figure 4. (a) Comparison between Φ-value (■) and $\Delta\Delta G_{U-\ddagger}$rel (□) observed at residue with significant Φ-values. Kinetic Φ-values and $\Delta\Delta G_{U-\ddagger}$ calculated for $k_f^{1.6M}$ are used in this histogram. (b) Comparison of the dispersion of residues with high Φ-value and high $\Delta\Delta G_{U-\ddagger}^{rel}$ within the secondary structure of Raf RBD. Residues in the sequence are classified as: positions mutated for Ala/Gly, but producing insignificant destabilization of the native state (black and bold lettering), those with Φ-value/ $\Delta\Delta G_{U-\ddagger}$rel between 0-0.25 (blue), 0.25-0.5 (green), 0.5-0.75 (orange) and 0.75-1 (red). (c) Dispersion in the tertiary structure of Raf RBD of residues in the hydrophobic core broke down from left to right panel according to: Φ-values, $\Delta\Delta G_{U-\ddagger}$rel and the inner/outer core hierarchy (reproduced from Figure 1, using the same color code). For the two former classifications, the color code used in panel (b) is utilized, except that mutations that produced insignificant destabilization (< |2| kJ mol^{-1}; see Table 3) of the native state are colored grey. A ribbon representation of Raf RBD is shown for reference.

Figure 5. (a) Leffler plot of all mutants tested using $\ln k_f^{1.6M}$ and $\Delta\Delta G_{F-U}^{kin}$ (○). (b) The same Leffler, but with separate correlation for residues in the amino-terminus β-hairpin (56-72) (●), residues located between α1 and β5 (e.g. L91-E124, but V98) (Δ), and the remaining residues (○). Outlier mutants in the β-hairpin (I58A, L62A and P63) are indicated (●), but neglected in the fitting presented on this panel. It these mutants are nevertheless included in the β-hairpin subgroup the correlation obtained is less good and the slope extrapolated is dramatically changed (y= 6.20+0.65x; R= 0.85).

Figure 6. Comparison of the residues involved in the TS stabilization and change of m_u in Raf RBD (■) *versus* ubiquitin (□) (data on ubiquitin taken from ([31])). The residue numbering of the Raf RBD is used throughout all panels. (a) Φ-value. Φ_F^{kin} and Φ_F are used for Raf RBD and ubiquitin, respectively. (b) $\Delta\Delta G_{U-\ddagger}$rel. (c) Comparisons of mu variation. The change of m_u is represented as the ratio= m_u^{mut}/m_u^{wt} for both proteins. See Materials and Methods for details on the calculation of Φ_F^{kin} and $\Delta\Delta G_{U-\ddagger}$ rel. Mutations marked with one star (*) did not yield significant Φ-value, because of too low $\Delta\Delta G_{F-U}$. The two stars (**) indicate that the mutation at residue of ubiquitin corresponding to L86 of Raf RBD was not reported. All other columns with zero values identify untested positions. Negative $\Delta\Delta G_{U-\ddagger}$ rel and occur mostly in mutants inducing insignificant k_f change. Only Ala mutants were considered, except at Ala residues at which Gly mutations were introduced. The corresponding residues in the two proteins are matched using the secondary structure alignment previously reported ([38]). Two of the inner core residues of the α-helix could not be straightforwardly aligned. Therefore, residues V26 and I30 of ubiquitin (corresponding to C81 and A85 of Raf RBD in the alignment) should be compared with their tertiary structure equivalent in Raf RBD: L82 and L86.

Figure 7. Structural dispersion and importance of residues involved in the stabilization of TS structure of Raf RBD *versus* ubiquitin and other proteins using Φ-value and $\Delta\Delta G_{U-\ddagger}$ rel. (a) Raf RBD *versus* ubiquitin (top and bottom panel, respectively). Only matching hydrophobic core residues for which data was available for Raf RBD and ubiquitin are shown here. The residue numbering of the Raf RBD is used throughout all panels. As discussed in the Results, two of the inner core residues of the α-helix are not aligned. Therefore, residues V26 and I30 of ubiquitin corresponding to C81 and A85 of Raf RBD should be compared with their tertiary structure equivalent in Raf RBD: L82 and L86. (b) Protein-L (1HZ6). (c) Protein-G (2GB1). (d) Cold shock protein (1CSP). The left and middle panel show the relative importance of residues in stabilizing TS structure according to Φ-value and $\Delta\Delta G_{U-\ddagger}$ rel, respectively. The color code is the same as that used previously (Figure 7). Ribbon representations of the protein structures are shown in the right panel as reference.

Figure 8. Schematic model of the folding pathway of Raf RBD based on Φ-value analysis and $\Delta\Delta G_{\ddagger-U}$ interpretation. The role of various structural regions in stabilization of the TS is colored using the following scale in increasing order of importance: grey< blue< green< orange< red. Other results suggest that α1 is in the process of consolidating its packing over the β-hairpin and β5 at the TS. This putative event is represented by grey double-head arrows in the scheme.

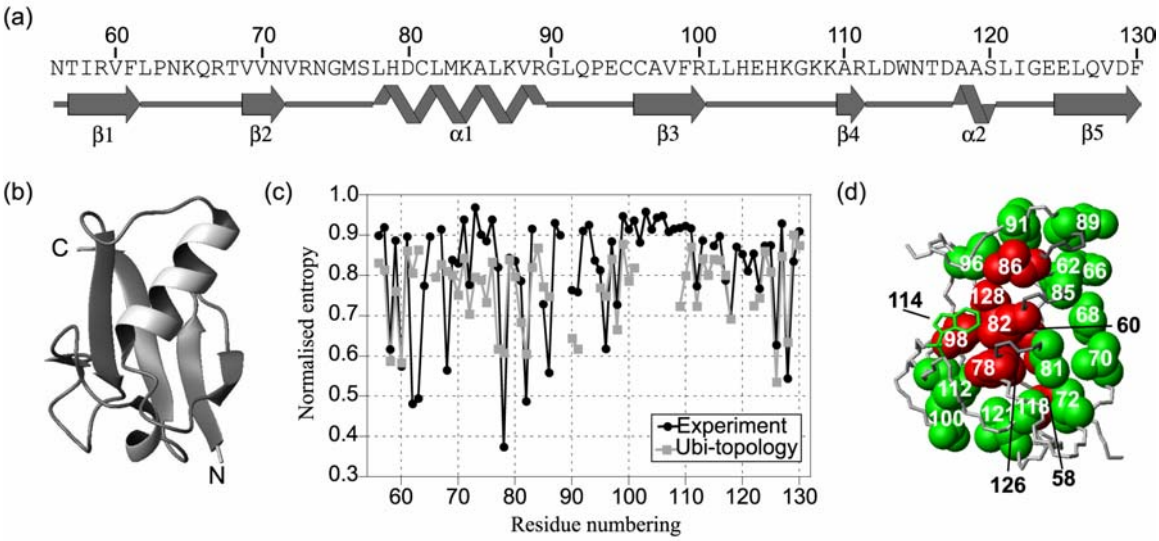

(a)
60 70 80 90 100 110 120 130
NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLLHEHKGKKARLDWNTDAASLIGEELQVDF
β1 β2 α1 β3 β4 α2 β5
(b)
C
N
(c)
Normalised entropy
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3
60 70 80 90 100 110 120 130
Residue numbering
Experiment
Ubi-topology
(d)
91 89 96 86 62 66 85 114 128 68 60 82 98 78 81 70 112 72 100 121 118 126 58

Fig. 1

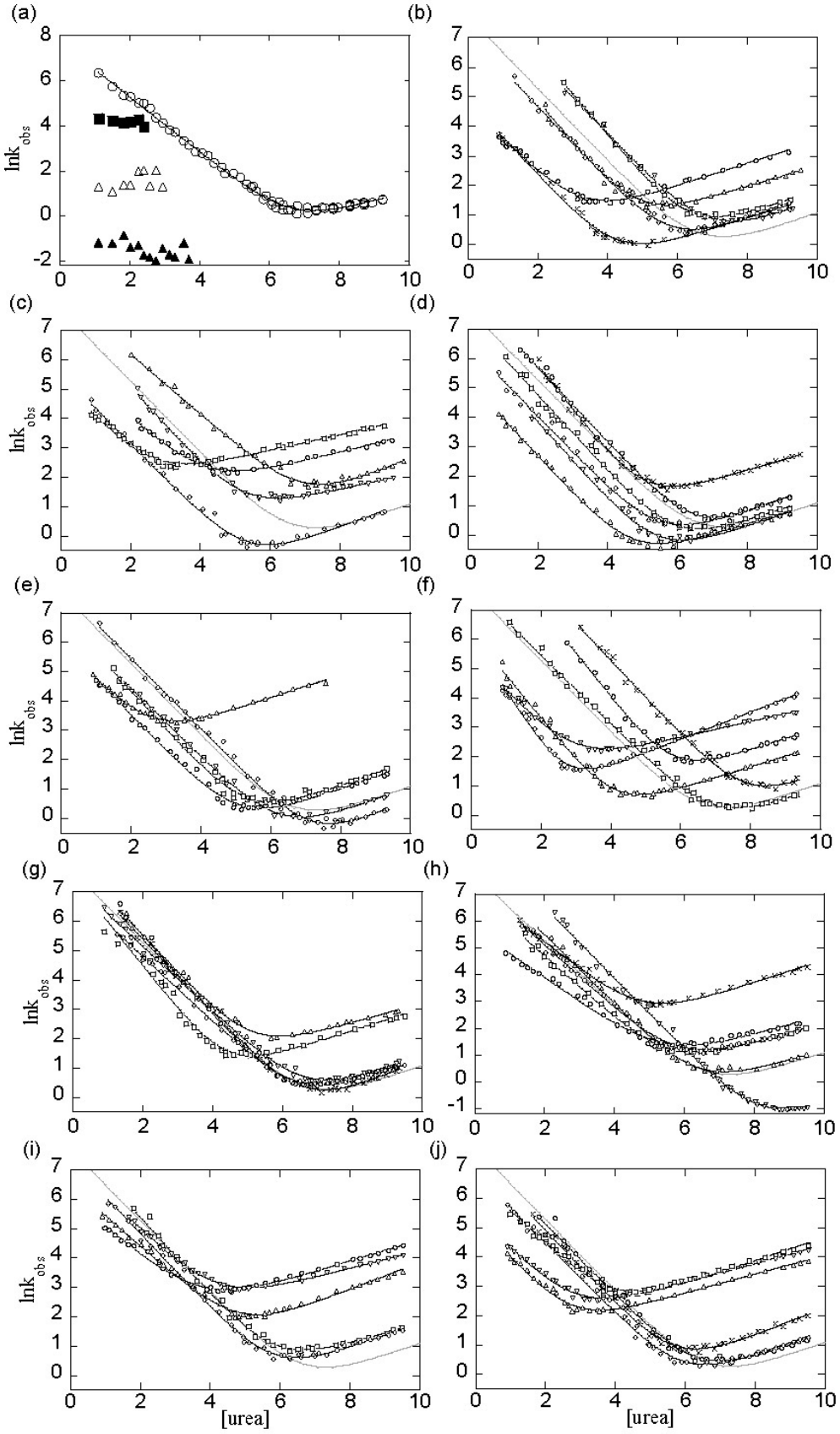


Fig. 2

Table 1. Kinetic parameters.

	$-m_f$ (M^{-1})	k_f^{0M} (s^{-1})	$k_f^{1.6M}$ (s^{-1})	m_u (M^{-1})	k_u^{0M} (s^{-1})	$k_u^{5.8M}$ (s^{-1})	k_u^{8M} (s^{-1})	β_t[a]
Wt	1.22	2263	322	0.41	0.05	0.53	1.30	0.75
N56M	1.40	10964	1174	0.38	0.14	1.26	2.92	0.79
I58A	1.16	112	18	0.38	0.73	6.78	15.77	0.75
I58L	1.24	1304	178	0.43	0.08	0.96	2.49	0.74
I58F	1.40	2146	228	0.35	0.42	3.30	7.20	0.80
R59A	1.23	5966	831	0.39	0.09	0.86	2.03	0.76
V60A	1.29	142	18	0.37	0.13	1.10	2.48	0.78
L62A	1.29	205	26	0.27	3.68	17.45	31.50	0.83
P63A	1.14	568	92	0.29	1.72	9.22	17.44	0.80
N64A	1.21	260	38	0.44	0.04	0.51	1.35	0.73
H2	1.03	3874	749	0.49	0.10	1.79	5.26	0.68
H2_F62L	1.33	2326	279	0.27	0.58	2.72	4.89	0.83
Q66A	1.30	1601	201	0.40	0.06	0.67	1.62	0.76
T68A	1.27	3854	504	0.47	0.05	0.74	2.10	0.73
V69A	1.23	680	95	0.41	0.09	0.92	2.27	0.75
V70A	1.29	197	25	0.36	0.08	0.65	1.45	0.78
V72A	1.33	677	81	0.40	0.06	0.56	1.34	0.77
V72I	1.51	6808	606	0.35	0.56	4.21	9.04	0.81
M76A	1.30	1061	131	0.42	0.11	1.23	3.08	0.76
S77A	1.22	401	57	0.41	0.10	1.11	2.72	0.75
S77T	1.20	2566	378	0.52	0.01	0.22	0.68	0.70
L78A	1.26	399	54	0.39	6.04	57.00	133.50	0.76
D80A	1.18	908	138	0.43	0.04	0.49	1.25	0.73
C81A	1.17	2479	381	0.42	0.04	0.45	1.13	0.74
C81I	1.52	24921	2183	0.38	0.46	4.10	9.36	0.80
L82A	1.86	453	23	0.46	0.85	12.25	33.71	0.80
A85G	1.48	530	50	0.37	0.27	2.31	5.20	0.80
L86A	1.28	251	32	0.27	2.77	13.54	24.71	0.82
R89L	1.26	30681	4092	0.46	0.04	0.59	1.63	0.73
L91A	1.50	1789	163	0.38	0.48	4.26	9.74	0.80
P93A	1.34	3714	433	0.37	0.10	0.82	1.84	0.78
C95A	1.11	1149	194	0.43	0.05	0.63	1.62	0.71
C96A	1.30	3453	434	0.36	0.90	6.17	12.82	0.78
C96L	1.11	1608	272	0.40	0.07	0.75	1.81	0.73
C96M	1.25	2693	367	0.47	0.03	0.49	1.38	0.73
A97G	1.17	1151	178	0.37	0.21	1.88	4.28	0.76
V98A	0.92	308	71	0.37	0.29	2.55	5.80	0.71
R100A	1.19	1812	271	0.39	0.20	1.84	4.30	0.76
E104A	1.29	3239	412	0.37	0.08	0.69	1.56	0.78
K109A	1.31	3016	372	0.34	0.19	1.40	2.99	0.79
L112A	1.00	432	87	0.36	2.74	22.25	49.26	0.73
D117A	1.26	1595	211	0.41	0.10	1.05	2.59	0.75
A118G	1.05	671	125	0.46	0.48	6.79	18.63	0.70
A118L	1.15	1454	233	0.31	3.22	19.37	38.24	0.79
L121A	1.14	833	135	0.40	1.96	19.97	48.18	0.74
E124A	1.21	1914	276	0.40	0.07	0.73	1.79	0.75
E125A	1.21	1002	146	0.43	0.06	0.74	1.91	0.74
L126A	1.34	208	24	0.32	2.30	14.96	30.43	0.81
V128A	1.21	244	35	0.33	3.20	21.21	43.46	0.79
D129A	1.25	1682	227	0.46	0.09	1.37	3.82	0.73
Δ104-6	1.32	9661	1166	0.32	0.02	0.11	0.22	0.81
Δ101-8+AG	1.04	1471	280	0.42	1.36	15.94	40.55	0.71

See Materials and Methods for parameters description.

[a] Calculated from kinetic data using $\beta_t = m_f/(m_f+m_u)$

Table 2. Thermodynamic parameters used for Φ-value calculation.

	m ($kJ\ mol^{-1}\ M^{-1}$)	Cm (M)	$\Delta\Delta G_{F\text{-}U}^{kin}$ ($kJ\ mol^{-1}$)	$\Delta\Delta G_{F\text{-}U}^{Cm}$ ($kJ\ mol^{-1}$)	$\Delta\Delta G_{F\text{-}U}^{5.8M}$ ($kJ\ mol^{-1}$)
Wt	3.8	6.3	nsap	nsap	nsap
N56M	3.8	6.6	-1.2	-1.1	-1.1
I58A	4.1	2.8	13.4	13.5	14.2
I58L	4.3	5.5	3.1	3.3	3.3
I58F	4.0	4.9	5.1	5.6	5.6
R59A	3.6	6.5	-1.2	-0.6	-0.4
V60A	3.6	3.7	8.8	10.1	9.4
L62A	3.6	2.4	14.1	15.1	14.2
P63A	3.6	4.7	9.5	6.1	5.8
N64A	3.9	5.2	5.4	4.5	4.4
H2	4.0	6.8	1.4	-1.8	-2
H2_F62L	3.6	5.1	3.6	4.6	4.4
Q66A	3.8	5.7	1.7	2.3	2.3
T68A	3.8	6.1	0.1	0.7	0.6
V69A	3.9	5.1	4.4	4.8	4.8
V70A	4.0	4.3	6.6	7.7	7.9
V72A	4.2	5.1	3.5	4.8	4.9
V72I	3.9	5.3	3.2	3.8	3.8
M76A	3.1	5.3	4.4	4.1	3.6
S77A	4.0	4.6	6.1	6.8	6.8
S77T	4.0	7.0	-2	-2.6	-2.7
L78A	4.0	2.3	15.9	15.7	16.1
D80A	3.6	5.6	2	2.6	2.5
C81A	3.6	6.7	-0.7	-1.5	-1.3
C81I	5.0	5.0	0.2	5.2	6.1
L82A	4.8	2.8	14.6	13.5	16.2
A85G	4.4	4.0	8.1	9.0	9.9
L86A	4.1	2.8	13	13.5	14
R89L	4.1	7.3	-5.7	-3.8	-4
L91A	4.0	4.6	6.7	6.7	6.8
P93A	4.0	6.2	0.1	0.5	0.4
C95A	4.1	6.4	1.8	-0.3	-0.5
C96A	3.6	4.0	5	8.8	8.3
C96L	3.6	6.4	1.2	-0.4	-0.3
C96M	4.4	5.7	-0.2	2.2	2.2
A97G	4.0	5.5	4.4	3.2	3.2
V98A	3.5	4.0	7.5	9.1	8.3
R100A	3.8	5.6	3.4	2.7	2.6
E104A	3.9	6.3	-0.2	0.0	-0.1
K109A	3.9	6.0	1.7	1.2	1.2
L112A	3.8	3.7	12.2	10.3	10
D117A	3.6	5.5	2.8	3.1	3
A118G	3.8	4.3	8.9	7.9	7.8
A118L	3.3	4.2	9.2	8.1	7.2
L121A	4.0	3.7	11.1	10.3	10.4
E124A	4.1	6.3	1.2	0.1	0
E125A	3.8	5.7	2.9	2.3	2.2
L126A	4.1	2.7	14.2	14.0	14.5
V128A	4.0	2.7	14.2	13.9	14.2
D129A	4.0	5.5	3.5	3.0	3
Δ104-6	3.5	8.0	-7.6	-6.7	-5.8
Δ101-8+AG	3.6	4.5	8.9	7.2	6.8

See Materials and Methods for parameters description.

Table 3. Φ-values[a], β_t and structural information.

	Φ_f^{kin} [a]	Φ_f [a]	1-Φ_u [a]	$\Delta\Delta G_{U-\ddagger}^{1.6M}$ (kJ mol^{-1})	Structure[b]	Solvent access[c]	Contacts[d]
Wt							
N56M*					und	69	N71,I122
I58A	0.54	0.55	0.55	7.21	β1/IC	5	V70,V72,L82,A118,L126
I58L	0.48	0.42	0.55	1.47	β1	5	V70,V72,L82,A118,L126
I58F	0.17	0.02	0.19	0.85	β1	5	V70,V72,L82,A118,L126
R59A*					β1	87	T57,F61,R67,V69,E125
V60A	0.82	0.68	0.80	7.15	β1/IC	3	T68,V70,L82,A85,L86,L126,V128
L62A	0.44	0.39	0.39	6.22	t1/OC	4	Q66,T68,A85,L86,R89,L91,V128
P63A	0.32	0.56	-0.22	3.10	t1	45	F61,L91,C96,Q127,V128,D129
N64A	0.98	1.19	1.02	5.32	t1	93	Q66,L91,D129
H2*					t1	nsap	nsap
H2_F62L	0.10	-0.01	0.07	0.36	t1	nsap	nsap
Q66A*					und/OC	65	L62,R89
T68A*					und/OC	46	V60,L62,V70,A85,R89
V69A	0.69	0.62	0.71	3.03	β2	83	T57,R59,N71
V70A	0.96	0.78	0.93	6.34	β2/OC	22	I58,V60,C81,A85
V72A	0.98	0.63	0.97	3.43	und/OC	3	I58,C81,A118
V72I	-0.48	-0.71	-0.36	-1.57	und/OC	3	I58,C81,A118
M76A	0.51	0.46	0.42	2.22	t2	45	V72,D80,C81,K84
S77A	0.70	0.63	0.73	4.31	Ch	30	D80,N115
S77T	0.20	0.12	0.20	-0.40	Ch	30	D80,N115
L78A	0.28	0.27	0.28	4.45	α1/IC	0	V98,L112,W114,A118,L121,L126
D80A	1.05	0.87	1.08	2.09	α1	89	M76,M83
C81A*					α1/OC	3	V70,V72,A118
C81I*					α1/OC	3	V70,V72,A118
L82A	0.45	0.30	0.52	6.52	α1/IC	5	I58,V60,V98,W114,L126,V128
A85G	0.57	0.40	0.63	4.63	α1/OC	3	V60,L62,T68,V70
L86A	0.44	0.40	0.42	5.69	α1/IC	0	V60,L62,L91,C96,V128
R89L	1.10	1.72	1.07	-6.30	α1/OC	83	L62,Q66,T68,L91
L91A	0.25	0.09	0.24	1.68	Ch/OC	65	L62,Q66,L86,R89,C96
P93A*					t3	2	L86,W114
C95A*					t3	70	Q92,L131
C96A	-0.14	-0.11	0.26	-0.69	T3/OC	19	L86,L91,V128
C96L*					T3/OC	19	L86,L91,V128
C96M*					T3/OC	19	L86,L91,V128
A97G	0.33	0.53	0.01	1.47	β3	1	F99,R111,D129,L131
V98A	0.50	0.54	0.53	3.75	β3/IC	2	L78,L82,L112,W114,L126,V128
R100A	0.13	0.21	-0.19	0.43	β3/OC	25	L112,L121,L126
E104A*					t4	114	R100,L101,K106
K109A*					und	46	N64,F99,L101,L102,Q127,D129
L112A	0.26	0.40	0.08	3.23	und/OC	24	L78,V98,L121,L126
D117A	0.38	0.28	0.43	1.04	und	89	S120
A118G	0.26	0.38	0.18	2.34	α2/OC	1	I58,V72,L78,C81,L121
A118L	0.09	0.14	-0.24	0.80	α2/OC	1	I58,V72,L78,C81,L121
L121A	0.19	0.24	0.13	2.15	t7/OC	15	L78,R100,L112,A118,L126
E124A*					t7	44	R59,R100,H103
E125A	0.67	0.89	0.61	1.96	β5	49	R59,F61,L101,H103
L126A	0.45	0.42	0.43	6.42	β5/IC	3	I58,V60,L78,L82,V98,L112,L121
V128A	0.39	0.40	0.35	5.47	β5/IC	0	V60,L62,L82,L86,C96,V98,W114
D129A	0.24	0.24	0.21	0.86	β5	34	P63,N64,A97,F99,K109,Q127,F130
Δ104-6	0.42	0.53	0.33	-3.19	t4	nsap	nsap
Δ101-8+AG	0.04	0.15	-0.25	0.35	t4	nsap	nsap

Table 3 (continued)

* Mutations with $\Delta\Delta G_{F-U} < |2|$ kJ mol^{-1} according to kinetic $\Delta\Delta G_{F-U}$.
[a] See Material and Methods for details of the calculation.
[b] S, t and h stand respectively for strand, helix and b-turns. β-strands are: β1 (57-61), β2 (69-71), β3 (97-101), β4 (110-111) and β5 (125-130); α-helix are : α1 (78-89), α2 (118-120) ; b-turns are: t1 (62-65), t2 (73-76), t3 (93-96), t4 (102-105), t5 (105-108), t6 (113-116) and t7 (121-124). Ch stands for residues involved in the capping of the major helix; und stand for undefined structure (loop). IC and OC stands for inner and outer core, respectively.
[c] Information extracted from the DSSP file of 1RFA.
[d] Determined by manual inspection of the structure. Only non-consecutive residues for which side-chains are in direct contacts are listed (distance cut-off of 7Å).

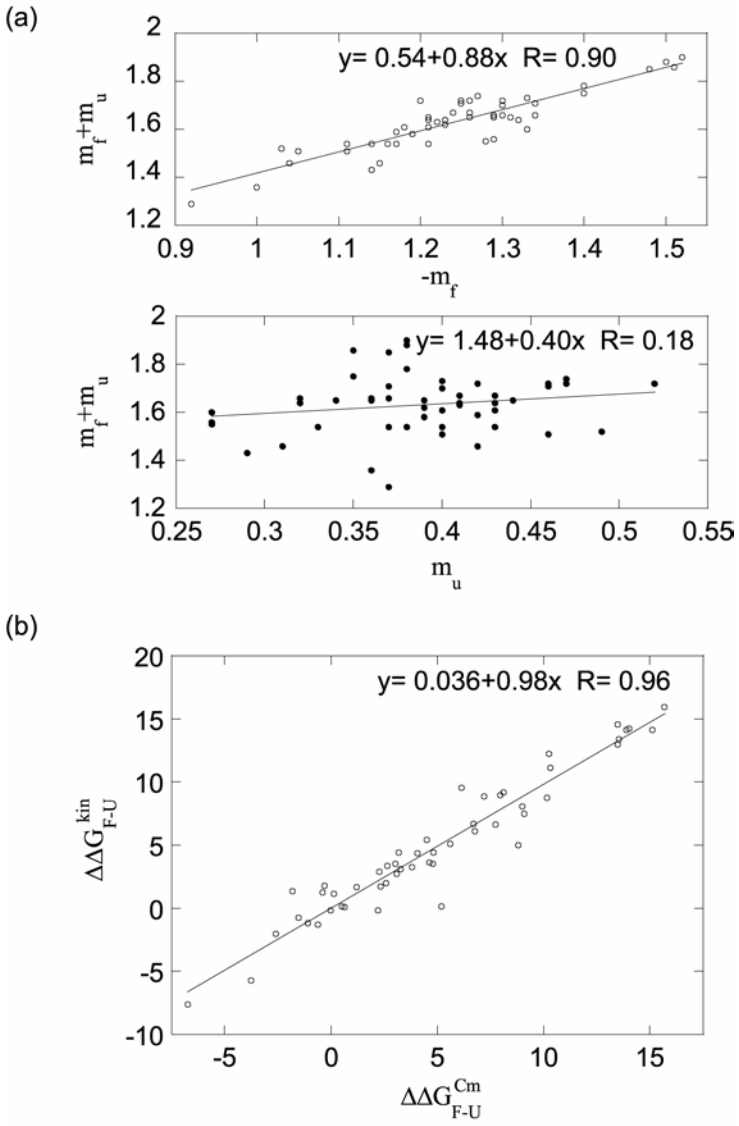

(a)
y= 0.54+0.88x R= 0.90
m_f+m_u
$-m_f$
y= 1.48+0.40x R= 0.18
m_f+m_u
m_u
(b)
y= 0.036+0.98x R= 0.96
$\Delta\Delta G^{kin}_{F-U}$
$\Delta\Delta G^{Cm}_{F-U}$

Fig. 3

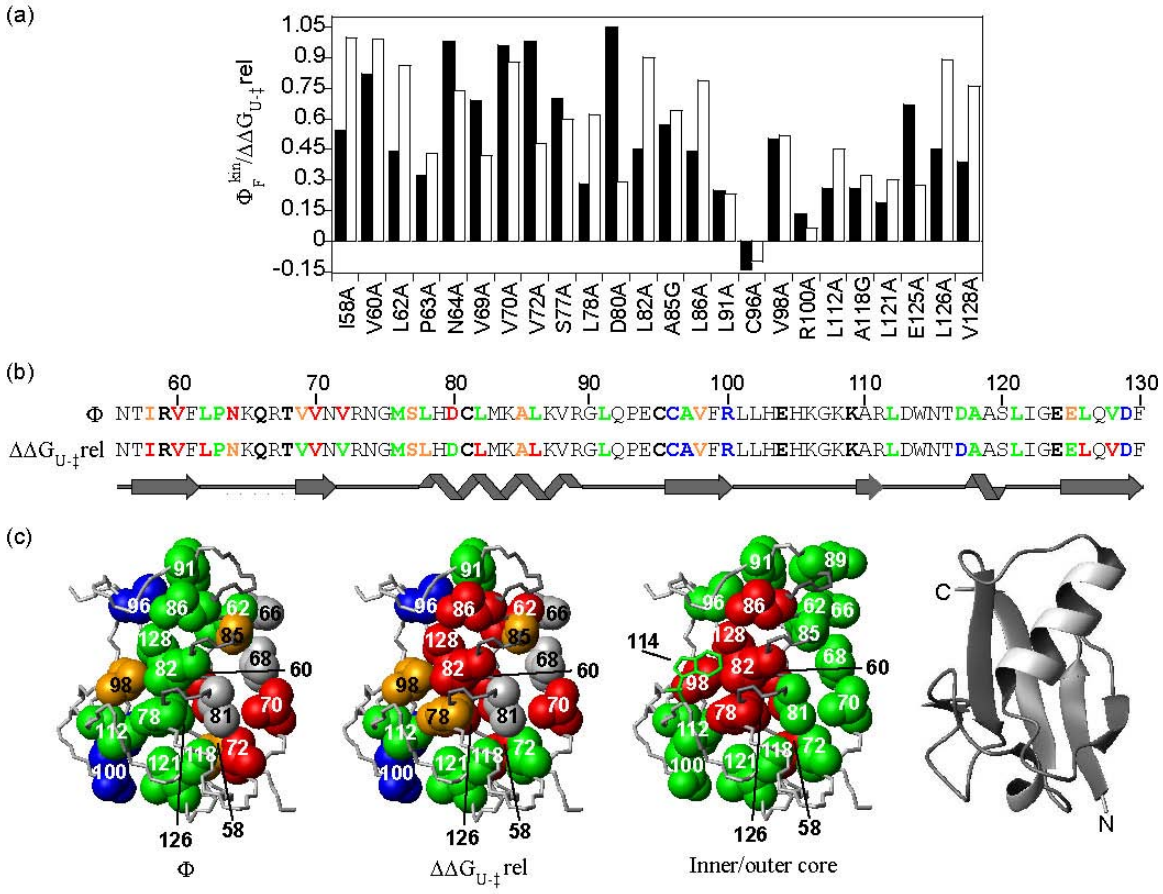

(a)
1.05
0.9
0.75
0.6
0.45
0.3
0.15
0
-0.15
$\Phi_F^{kin}/\Delta\Delta G_{U-\ddagger}^{rel}$
I58A V60A L62A P63A N64A V69A V70A V72A S77A L78A D80A L82A A85G L86A L91A C96A V98A R100A L112A A118G L121A E125A L126A V128A
(b)
60 70 80 90 100 110 120 130
Φ NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLLHEHKGKKARLDWNTDAASLIGEELQVDF
$\Delta\Delta G_{U-\ddagger}^{rel}$ NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLLHEHKGKKARLDWNTDAASLIGEELQVDF
(c)
Φ
$\Delta\Delta G_{U-\ddagger}^{rel}$
Inner/outer core
C
N

Fig. 4

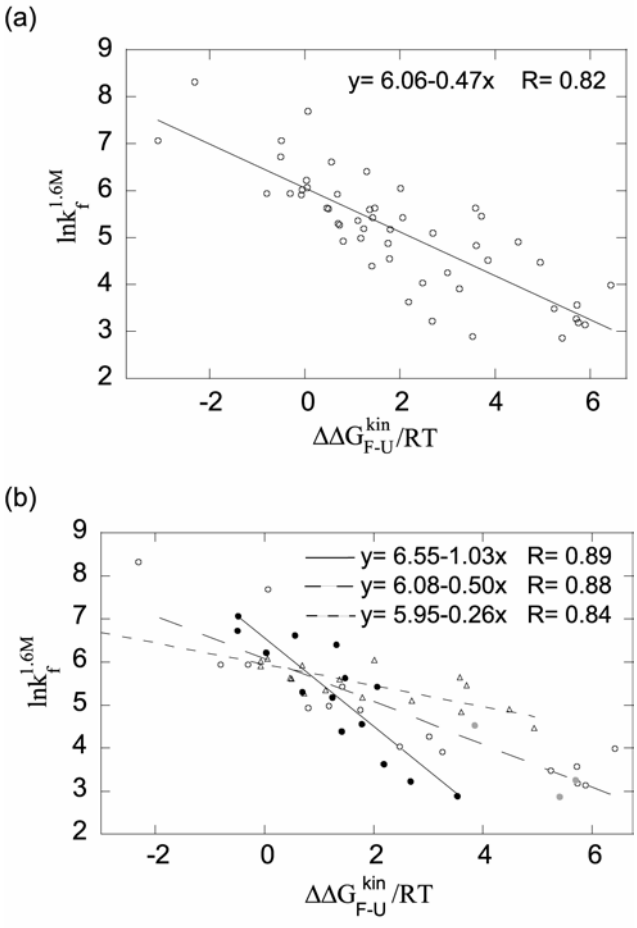

(a)
y= 6.06-0.47x R= 0.82
$\ln k_f^{1.6M}$
$\Delta\Delta G_{F-U}^{kin}/RT$
(b)
y= 6.55-1.03x R= 0.89
y= 6.08-0.50x R= 0.88
y= 5.95-0.26x R= 0.84
$\ln k_f^{1.6M}$
$\Delta\Delta G_{F-U}^{kin}/RT$

Fig. 5

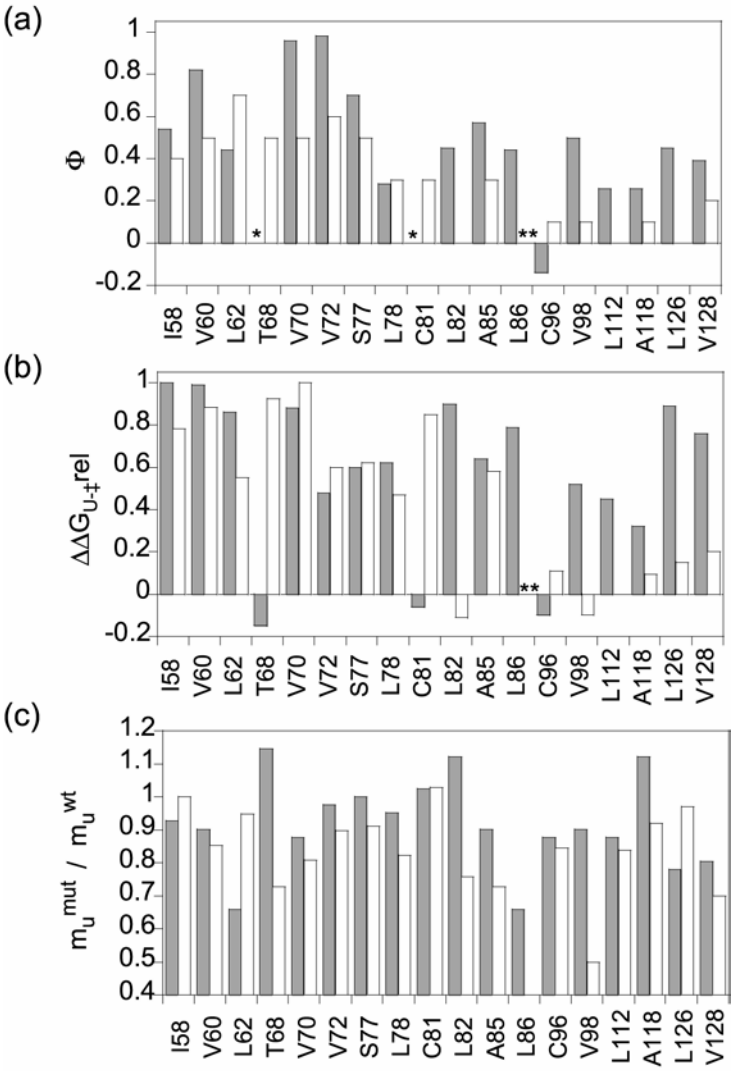

(a)
Φ
1
0.8
0.6
0.4
0.2
0
-0.2
*
*
**
I58 V60 L62 T68 V70 V72 S77 L78 C81 L82 A85 L86 C96 V98 L112 A118 L126 V128
(b)
ΔΔG‡U-rel
1
0.8
0.6
0.4
0.2
0
-0.2
**
I58 V60 L62 T68 V70 V72 S77 L78 C81 L82 A85 L86 C96 V98 L112 A118 L126 V128
(c)
mu mut / mu wt
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
I58 V60 L62 T68 V70 V72 S77 L78 C81 L82 A85 L86 C96 V98 L112 A118 L126 V128

Fig. 6

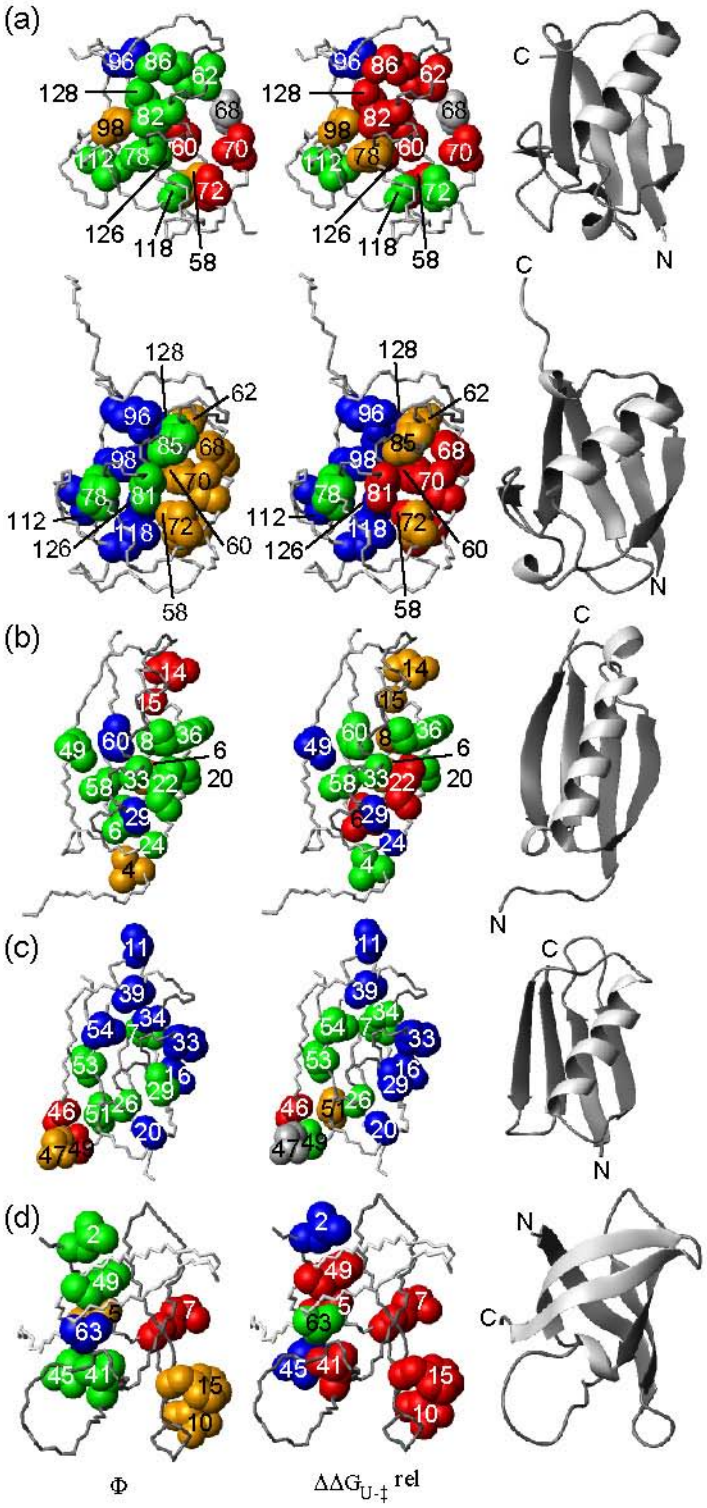


Fig. 7

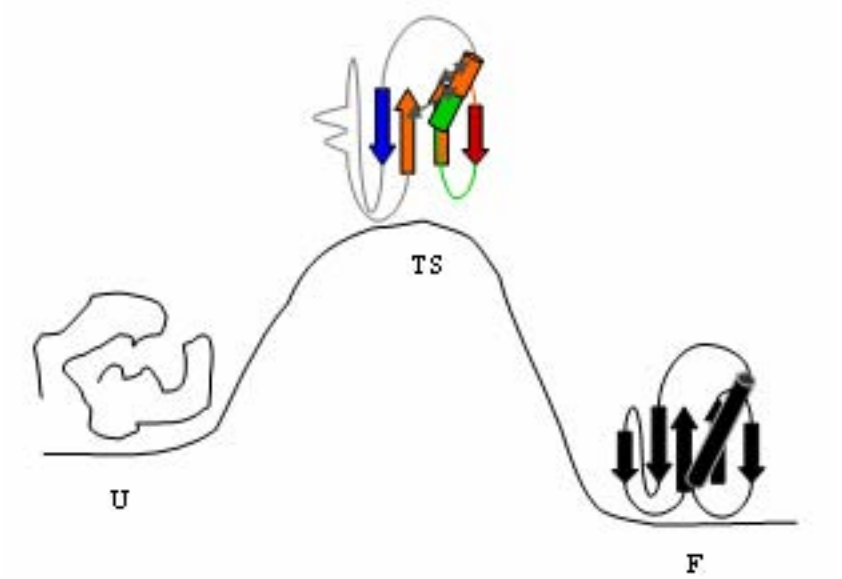


Fig. 8

Supplementary Material

Figure S1. (a) m_f and (b) m_u obtained for characterized mutants. Dashed lines indicate one standard deviation from the mean value of both parameters.

Figure S2. Plot of the correlation of Φ_F^{kin} versus Φ_F^{Cm} (□; filled line) and 1-Φ_U (○; dashed line) for mutants that display significant destabilization ($\Delta\Delta G_{F-U} < |2|$ kJ mol^{-1}). The slope is close to 1, the y intercept is close to 0 and R≈ 0.9 for both correlations. The plot of Φ_F^{kin} with Φ_F^{ext} (see Table S1) is not shown here, but displayed similar correlation.

Figure S3. (a) Leffler plot of all mutants tested using kinetic and thermodynamic data: $lnk_f^{1.6M}$ and $\Delta\Delta G_{F-U}{}^{1.6M}$ (○) and lnk_u^{8M} $\Delta\Delta G_{F-U}{}^{8M}$ (■). The values of β_F and β_U obtained are in agreement with the two-state model and the principle of micro-reversibility. (b) The same Leffler plot, but with separate correlation for residues in the amino-terminus β-hairpin (56-72) (●), residues located between α1 and β5 (e.g. L91-E124, but V98) (Δ), and the remaining residues (○). Outlier mutants in the β-hairpin (I58A and L62A) are indicated (●), but neglected in the fitting presented on this panel. It these mutants are nevertheless included in the β-hairpin subgroup the correlation obtained is less good and the slope extrapolated is dramatically changed (y= 6.10+0.62x; R= 0.84). $\Delta\Delta G_{F-U}{}^{1.6M}$ and $\Delta\Delta G_{F-U}{}^{8M}$ are calculated from thermodynamic data using the same formalism than the equation used for calculating $\Delta\Delta G_{F-U}{}^{5.8M}$ (Materials and Methods). Despite an apparent reduction in βf for the three subgroups, the results are overall similar to the kinetic data discussed in details in Results section (Figure 5 (b)).

Figure S4. (a) Correlation of Φ-values for Raf RBD (Φ_F^{kin}) and ubiquitin (Φ_{avge}) (Materials and Methods). The correlation obtained is very good (y= 0.031+0.51x, R= 0.63), considering the difference in packing of α1 over the β-sheet and the changes this bring to the arrangement of the hydrophobic core. There are 5 major outliers (Raf RBD-ubiquitin): L62A-T7A (0.4 vs 0.7), L82A-K27A (0.45 vs 0), V98A-L43A (0.5 vs 0.1), L112A-L50A (0.26 vs 0), L126A-L67A (0.45 vs 0). By removing the most deviant pair (L62A-T7A) R is increased to 0.75. Fitting is further improved (R= 0.84), if residues of the inner core in α1 are realigned (e.g., L82 with V26 and L86 with I30 for RBD and ubiquitin, respectively). In any of these cases, the slope stays constant. (b) $\Delta\Delta G_{U-‡}$ for matching residues of ubiquitin and Raf RBD obtained according to their secondary structure alignment (Figure 6). See Materials and Methods for details on the calculation of $\Delta\Delta G_{U-‡}$. The two stars (**) indicate that the mutation at residue of ubiquitin corresponding to L86 of Raf RBD was not reported. (c) $\Delta\Delta G_{U-‡}$rel of Raf RBD and ubiquitin normalized by the change in side-chain volume upon mutation. (d) Structural comparison of the importance of residues for TS stabilization according to $\Delta\Delta G_{U-‡}$rel normalized by the change in side-chain volume or not. Raf RBD and ubiquitin are in the top and bottom panels, respectively. The major drawback of volume normalization in comparison with $\Delta\Delta G_{F-U}$ in Φ-value calculation is the exaggeration of the importance of smaller residues (A85 and A118). If these residues are excluded the results are comparable to the non-normalized parameter.

Table S1. $\Delta\Delta G_{F\text{-}U}$ and Φ_F obtained from extrapolated equilibrium and kinetic parameters.

	$\Delta\Delta G_{F\text{-}U}^{ext\ a}$ (kJ mol^{-1})	$\Phi_F^{ext\ b}$
Wt	nsap	Nsap
N56M	-1.0	*
I58A	12.3	0.60
I58L	0.8	*
I58F	4.7	0.03
R59A	0.9	*
V60A	10.9	0.63
L62A	15.3	0.39
P63A	6.8	0.50
N64A	4.1	1.30
H2	-3.3	0.41
H2_F62L	5.6	-0.01
Q66A	2.5	0.35
T68A	0.8	*
V69A	4.1	0.72
V70A	6.7	0.90
V72A	3.0	1.01
V72I	3.2	-0.84
M76A	7.6	0.25
S77A	6.0	0.72
S77T	-3.5	0.09
L78A	14.9	0.29
D80A	4.1	0.56
C81A	-0.1	*
C81I	-0.8	*
L82A	10.4	0.39
A85G	6.4	0.56
L86A	12.5	0.44
R89L	-5.4	1.20
L91A	5.7	0.10
P93A	-0.7	*
C95A	-2.1	*
C96A	9.5	-0.10
C96L	0.8	*
C96M	-1.4	*
A97G	1.9	*
V98A	10.4	0.48
R100A	2.6	0.21
E104A	-0.6	*
K109A	0.7	*
L112A	10.1	0.41
D117A	4.3	0.20
A118G	7.8	0.39
A118L	10.0	0.11
L121A	9.6	0.26
E124A	-1.7	*
E125A	2.3	0.86
L126A	13.1	0.45
V128A	13.1	0.42
D129A	1.9	0.38
Δ104-6	-3.9	0.93
Δ101-8+AG	8.0	0.13

* Mutations with $\Delta\Delta G_{F\text{-}U}^{ext} < |2|$ kJ mol^{-1}.

[a] $\Delta\Delta G_{F\text{-}U}^{ext} = \Delta G_{F\text{-}U}^{wt} - \Delta G_{F\text{-}U}^{mut}$. $\Delta G_{F\text{-}U}$ are obtained by extrapolation from urea melting curves realized at equilibrium (see companion article).

[b] $\Phi_F^{ext} = -RT\ln(k_f^{H2O}wt - k_f^{H2O}mut)/(\Delta\Delta G_{F\text{-}U}^{ext})$

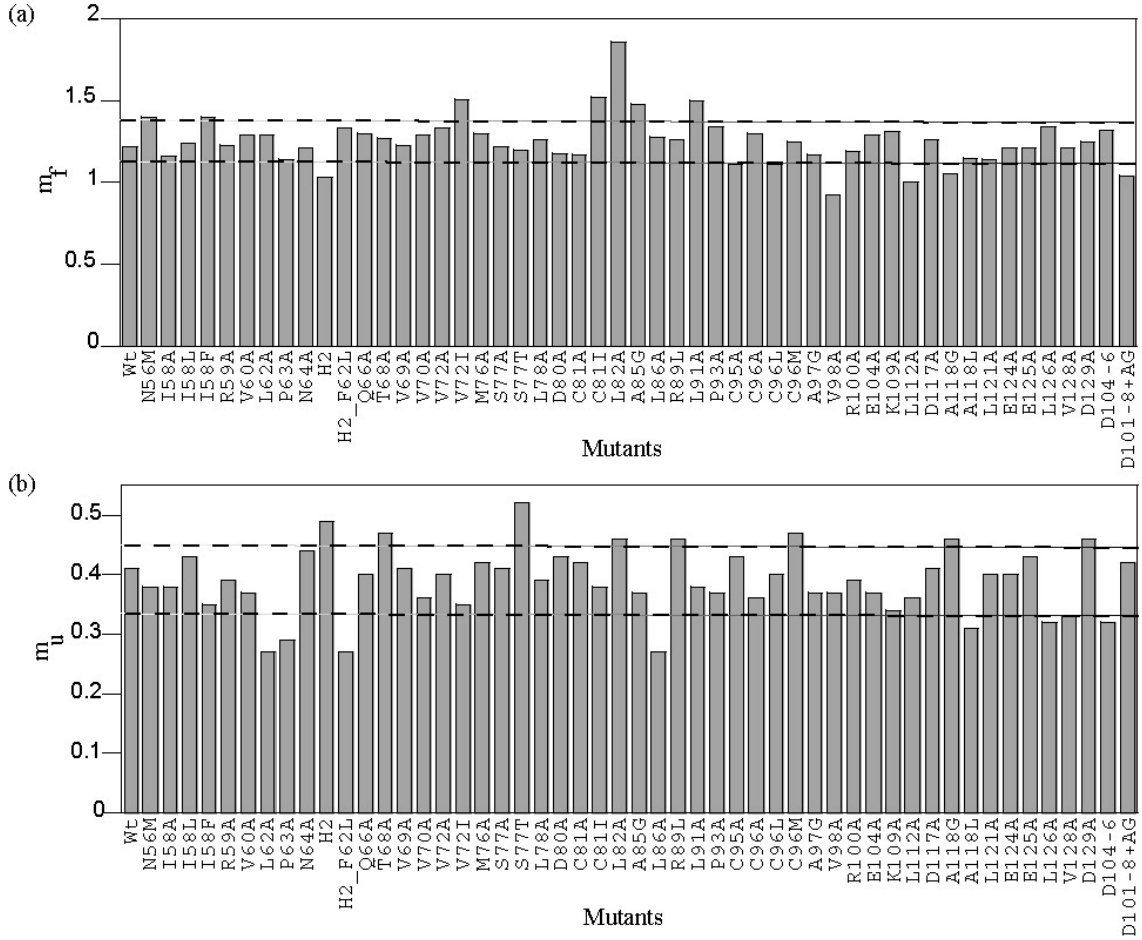

(a)
m_f
2
1.5
1
0.5
0
Wt
N56M
I58A
I58L
I58F
R59A
V60A
L62A
P63A
N64A
H2
H2_F62L
Q66A
T68A
V69A
V70A
V72A
V72I
M76A
S77A
S77T
L78A
D80A
C81A
C81I
L82A
A85G
L86A
R89L
L91A
P93A
C95A
C96A
C96L
C96M
A97G
V98A
R100A
E104A
K109A
L112A
D117A
A118G
A118L
L121A
E124A
E125A
L126A
V128A
D129A
D104-6
D101-8+AG
Mutants
(b)
m_u
0.5
0.4
0.3
0.2
0.1
0
Wt
N56M
I58A
I58L
I58F
R59A
V60A
L62A
P63A
N64A
H2
H2_F62L
Q66A
T68A
V69A
V70A
V72A
V72I
M76A
S77A
S77T
L78A
D80A
C81A
C81I
L82A
A85G
L86A
R89L
L91A
P93A
C95A
C96A
C96L
C96M
A97G
V98A
R100A
E104A
K109A
L112A
D117A
A118G
A118L
L121A
E124A
E125A
L126A
V128A
D129A
D104-6
D101-8+AG
Mutants

Fig. S1

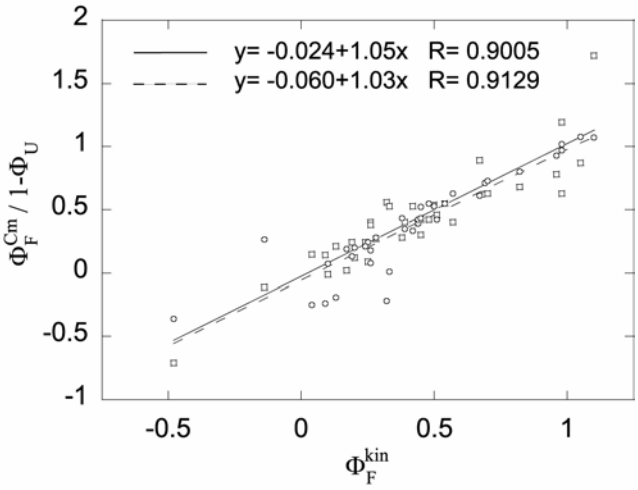

y= -0.024+1.05x R= 0.9005
y= -0.060+1.03x R= 0.9129
2
1.5
1
0.5
0
-0.5
-1
-0.5
0
0.5
1
Φ_F^{Cm} / 1-Φ_U
Φ_F^{kin}

Fig. S2

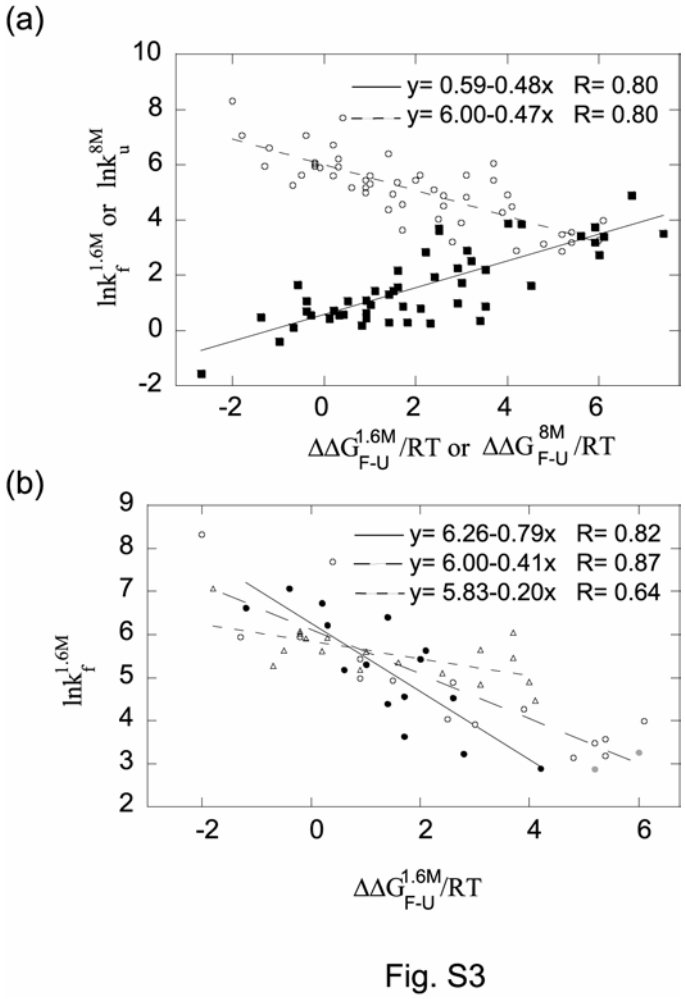

(a)
y= 0.59-0.48x R= 0.80
y= 6.00-0.47x R= 0.80
$\ln k_f^{1.6M}$ or $\ln k_u^{8M}$
$\Delta\Delta G_{F-U}^{1.6M}/RT$ or $\Delta\Delta G_{F-U}^{8M}/RT$
(b)
y= 6.26-0.79x R= 0.82
y= 6.00-0.41x R= 0.87
y= 5.83-0.20x R= 0.64
$\ln k_f^{1.6M}$
$\Delta\Delta G_{F-U}^{1.6M}/RT$

Fig. S3

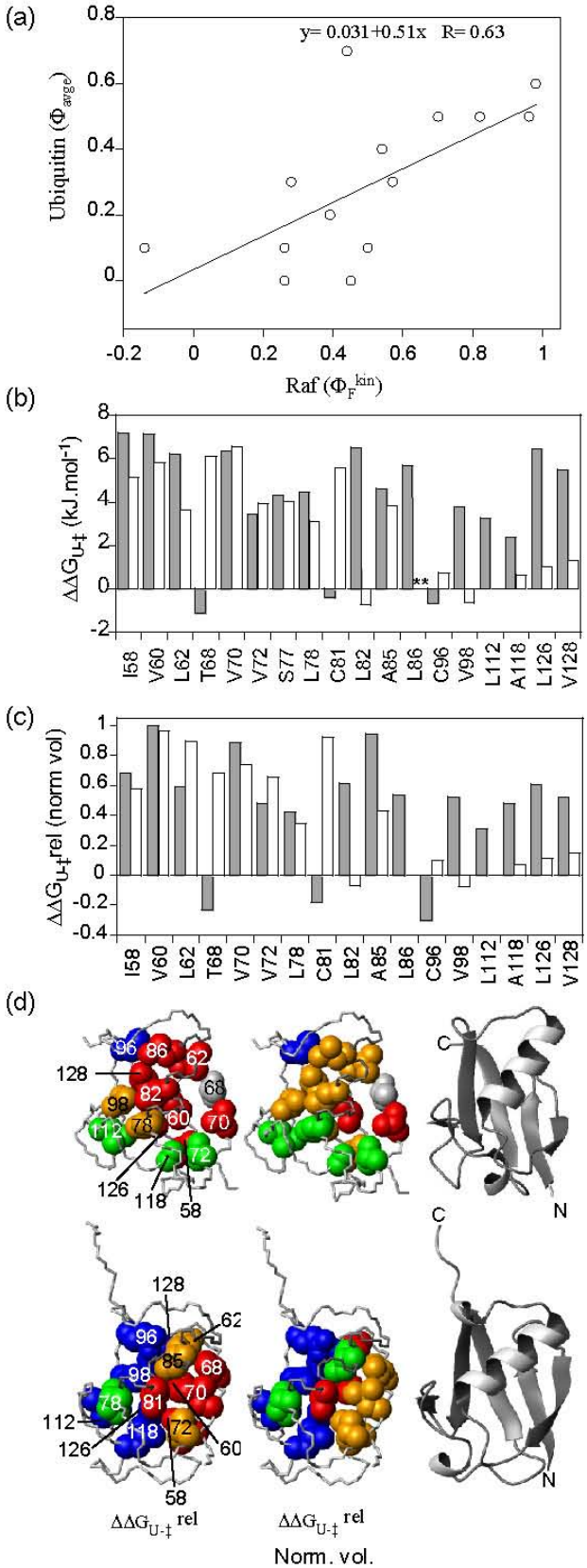

(a)
y= 0.031+0.51x R= 0.63
Ubiquitin (Φ_{avge})
Raf (Φ_F^{kin})
(b)
$\Delta\Delta G_{U-‡}$ (kJ.mol^{-1})
I58 V60 L62 T68 V70 V72 S77 L78 C81 L82 A85 L86 C96 V98 L112 A118 L126 V128
(c)
$\Delta\Delta G_{U-‡}$rel (norm vol)
I58 V60 L62 T68 V70 V72 L78 C81 L82 A85 L86 C96 V98 L112 A118 L126 V128
(d)
C
N
$\Delta\Delta G_{U-‡}$ rel
$\Delta\Delta G_{U-‡}$ rel
Norm. vol.

Fig. S4

Conclusion et Perspectives

Dans cette section, je compte faire un bref retour sur les résultats les plus importants de mes travaux ainsi que suggérer quelques pistes intéressantes à explorer dans le cadre d'études ultérieures et/ou hypothétiques. En second lieu, je présenterai succinctement quelques uns des thèmes émergeants et des perspectives d'avenir de la recherche en biologie structurale, plus particulièrement à propos de l'étude du repliement, le design et la prédiction de structure. Je ne prétends pas pouvoir présenter une position consensuelle en ce qui concerne ces sujets. Vous devrez vous contenter d'un point de vue qui n'engage que moi.

Retour sur mes travaux

Les résultats de l'expérience de perturbation de séquence (**Article 2**) (*22*) ont démontré qu'il est possible en utilisant des librairies segmentaires de codons dégénérées d'obtenir à partir d'une protéine modèle, dans ce cas-ci le DLR de Raf, une diversité de séquences comparable à la majorité des positions variées à celle observée dans un alignement d'analogues structuraux naturels de la topologie d'ubiquitine. Il y a principalement deux classes de résidus qui contredisent cette observation générale. Bien sûr, il y a les résidus dont la conservation est spécifique à la fonction et à des détails structuraux propres au DLR de Raf. En tout et pour tout, il y aurait 12 positions où ces facteurs entrent en ligne de compte de manière significative. La seconde catégorie est constituée principalement des résidus du cœur hydrophobe interne dont la tolérance à la variation est la plus susceptible d'être contrainte dans l'approche par dégénérescence segmentaire de la séquence que nous avons adoptée (déduction logique reposant sur le fait que notre stratégie ne permet pas la co-variation concomitante de l'ensembles des résidus du cœur hydrophobe, mais aussi expérimentalement démontré dans d'autres travaux de perturbation de séquence (*54*)). Cette limitation potentielle n'excluant pas la précédente,

l'importance au maintien de la fonction de liaison de la conservation de certains résidus du cœur hydrophobe, spécifiquement ceux qui stabilisent l'état de transition, a été démontrée directement chez le domaine SH3 de fyn (*147*). D'autre part, la diversité des acides aminés hydrophobes tolérée aux positions du cœur hydrophobe interne à travers les superfamilles reliées à l'ubiquitine a souligné le nombre relativement grand d'arrangements compatibles avec cette topologie (**Article 3**). Pour réconcilier ces observations, la co-variation concomitante des résidus du cœur hydrophobe interne et externe devrait être réalisée sur le DLR de Raf en utilisant l'essai de liaison à *ras* que nous avons mis au point pour la sélection des séquences. Pour des fins de comparaison, il serait intéressant de faire la sélection de ces librairies ou des librairies segmentaires en utilisant un moyen de sélection alternatif qui ne reposerait pas sur la liaison du DLR de Raf à *ras*. Les résultats de ces expériences permettraient de bonifier l'information de séquence obtenue lors des expériences initiales en décortiquant plus avant les liens structure-fonction. Spécifiquement, en combinant ces nouvelles informations avec d'autres approches expérimentales, cela permettrait de créer une matrice artificielle des variations de la séquence qui sont permises dans le DLR de Raf et de faire des analyses de couplage énergétique des résidus, telles que décrites dans des articles portant sur la co-variation naturelle des résidus chez les domaines d'homologie à « PSD-95 large discs/ZO-1 » (PDZ), des protéases, et un récepteur couplé aux GTPases trimériques (*247;248*). Malgré les réserves que je viens de décrire à propos de notre démarche expérimentale, je crois que nous avons pu démontrer dans le cadre de nos articles que l'approche des librairies segmentaires est satisfaisante pour obtenir une vue d'ensemble de l'espace de séquence accessible à une structure et donc, augmenter utilement l'information de séquence. Par extension, la diversité de séquence comparable (i.e. entropie de séquence) entre nos données sur le DLR de Raf et la diversité de séquences naturelles dans la topologie d'ubiquitine suggère que notre approche puisse être appliquée aux topologies peu populeuses afin d'améliorer notre compréhension du fonctionnement de ces protéines et en conséquence, la prédiction de leur structure et de leurs activités biologiques ainsi que leur design.

À ce jour, la prédominance de l'un des divers facteurs, tel que la stabilité, les caractéristiques de la réaction de repliement (soit k_f ou la structure de l'état de transition) ou bien la fonction, sur la conservation évolutive de la séquence est le sujet de vifs débats. Des évidences sérieuses obtenues d'expériences et de simulations informatiques pointent dans la direction de la stabilité, du moins à la majorité des positions qui ne sont pas impliquées directement dans le maintien de la fonction, et particulièrement pour les résidus du cœur hydrophobe chez des domaines SH3 (*146;147*). Nos résultats suggèrent aussi la prédominance de la stabilité chez le DLR de Raf et ce, à toutes les positions testées, exception faite de quelques résidus, représentant moins de 10% de la séquence (**Article 3**). Par conséquent, l'observation originale citée précédemment a pu être étendue aux résidus hors du cœur hydrophobe. Par ailleurs, la corrélation de l'entropie de séquence avec le taux de repliement, bien qu'inférieure à celle obtenue avec la stabilité, est significative. De manière remarquable l'extrapolation des deux corrélations obtenues pour le DLR de Raf permettent d'obtenir un estimé précis de k_f et de la stabilité du ts. En conclusion, il faudrait vérifier les mêmes relations investiguées sur le DLR de Raf chez une protéine possédant un état de transition plus polarisé que celui-ci, afin de définir clairement la relation entre ces paramètres biophysiques et la conservation de la séquence. Dans le cadre de cette étude, nous avons aussi découvert des mutations qui stabilisent le DLR de Raf de manière très significative (**Article 3**). En effet, nous avons démontré que des mutations affectant 5 résidus seulement sont suffisantes pour réduire le ΔG_{F-U} de l'ordre de 150%. Deux de ces mutations, soit R89L et H2, abolissent et réduisent respectivement l'affinité envers *ras*. De plus en plus de preuves expérimentales démontrent que des compromis au niveau de la stabilité de l'état natif doivent être effectués afin de permettre l'établissement d'une fonction enzymatique ou de liaison efficace (*4;5*). Cependant, le mutant stabilisant le plus intéressant que nous ayons obtenu correspond à une délétion de trois résidus observée naturellement chez a-Raf et b-Raf et qui se produit dans une boucle située entre les brins-β 3 et 4. Ce mutant dénommé Δ104-6 liant *ras* avec une affinité comparable au ts, il serait intéressant de vérifier son impact sur la fonction cellulaire accomplie par les divers gènes Raf.

Dans la dernière étude présentée dans le cadre de ma thèse nous avons pu démontrer la similarité entre l'état de transition du DLR de Raf et de l'ubiquitine en utilisant la méthode d'ingénierie des protéines, l'analyse des valeurs-Φ et la comparaison d'autres paramètres cinétiques (**Article 4**). Les caractéristiques générales de la structure de l'état de transition du DLR de Raf apparaissent à mi-chemin entre le modèle polarisé et diffus, soit respectivement entre les protéines L et CI2, leur prototype respectif. D'autre part, l'état de transition d'ubiquitine apparaît beaucoup plus polarisé. Nous avons aussi obtenu quelques preuves indirectes de la présence de structures dans l'état dénaturé. De plus, d'autres évidences expérimentales suggèrent la présence d'une seconde transition sur une large barrière d'énergie dans le cas d'ubiquitine et du DLR de Raf, ajoutant ainsi une autre propriété commune. En ce sens, l'étude sur le mécanisme de repliement du DLR de Raf s'inscrit bien dans la littérature scientifique contemporaine qui contribue à redéfinir actuellement les théories décrivant et modélisant le mécanisme de repliement (voir section **L'état de transition**, **Intermédiaires** et **L'état déplié et/ou dénaturé**). Il m'apparaît clair qu'après un grand détour l'on se rapproche d'une description empirique d'un mécanisme de repliement, bien que l'énoncé d'une théorie et d'équations mathématiques globales soient encore loin d'être réalisables (voir section **Vers un modèle global de la réaction de repliement**).

L'organisation de l'empaquetage dans le coeur hydrophobe au sein des superfamilles reliées à l'ubiquitine et plus largement de la topologie d'ubiquitine ont révélé de nombreuses similarités organisationnelles (**Figure 7** et **Article 3**). En particulier, les contacts établis dans le cœur hydrophobe interne définissent des réseaux qui sont conservés dans certaines superfamilles entre lesquelles des relations évolutives sont suspectées. Nous avons proposé que l'utilisation de schémas basés sur la théorie de graphe et illustrant ces liens architecturaux (*249;250*) pourraient être utilisés afin de représenter de façon simplifiée et non-ambiguë la structure des protéines, mais surtout d'établir des liens évolutifs et structuraux entre des protéines appartenant à la même topologie ou entre les topologies. En combinaison avec des études classiques où la cinétique de repliement est

déterminée, ces nouveaux outils pourraient mener à l'énoncé d'hypothèses éclairantes à propos de l'interrelation entre l'évolution, la sélection et la distribution dans la nature de la structure primaire et tertiaire des protéines. Par exemple, il serait intéressant de vérifier si ce genre d'approche permet de confirmer les sous-classifications au sein de la topologie d'ubiquitine et de prédire les similarités du repliement de ces membres. Allons encore plus loin. Les protéines globulaires dont la topologie est composée de deux couches distinctes d'éléments de structures secondaires démontrent des caractéristiques grossières communes selon l'analyse d'Efimov (*251*). La construction d'un arbre organisationnel structural[31] qui dresserait les liens entre les topologies des protéines bicouches sur la base de graphes décrits ci-haut constituerait une autre manière d'évaluer le mérite de la proposition d'Efimov. En plus, peut-être serait-il alors possible d'établir des liens évolutifs entre les topologies et de relier cela à leur mécanisme de repliement et à la structure de leur état de transition?

Depuis le début de cette section, j'ai déjà commencé à souligner quelques idées intéressantes que j'ai tirées de l'analyse de mes données. Voici plus de détails concernant certaines de ces avenues ainsi que d'autres qui vaudraient la peine d'être investiguées si je devais faire un autre doctorat à la suite de ce projet (par pitié non!) :

- J'ai évoqué plus tôt dans cette section l'intérêt de faire co-varier les résidus du cœur hydrophobe qui ne sont pas situés dans des segments contigus de séquence afin d'obtenir de l'information plus complète sur la diversité de séquence tolérée à ces sites et leurs interrelations ou couplage. D'autre part, nous avons maintenant en main une cartographie assez complète des résidus cruciaux à la formation et à la stabilisation de la structure du DLR de Raf (**Article 3** et **4**). Je propose donc d'utiliser l'approche des librairies segmentaires (**Article 1** et **2**) afin d'étudier les processus d'accommodements à l'insertion de mutants ponctuels déstabilisant, soit par le biais d'une augmentation ou d'une réduction du volume de la chaîne latérale. Par exemple, les résidus correspondant à l'hélice-α pourraient être dégénérés dans le contexte du mutant V60A et la conservation des acides aminés comparée à ce que nous avions observé dans la librairie segmentaire équivalente obtenue chez le ts. La réalisation de plusieurs combinaisons (i.e., à partir de divers mutants ponctuels et

[31] Par similarité avec les arbres phylogénétiques.

librairies segmentaires) de ce genre permettrait de définir certains mécanismes de compensation et de couplages prenant place dans la structure du DLR de Raf. Une approche analogue serait applicable aux résidus dont la conservation semble limitée par des contraintes fonctionnelles ou structurales spécifiques aux DLR de type Raf. Cela permettrait de définir des résidus impliqués dans l'établissement d'un complexe stable avec *ras* et de définir l'interrelation entre les résidus d'un monomère ou de l'hétérodimère.

- Durant l'expérience de perturbation de séquence, nous avons noté la sélection prédominante de certains acides aminés non-ts à une dizaine de positions. Nous avons testé l'effet des mutations les plus fortement sélectionnées sur la formation et la stabilisation de la structure et dans certains cas sur la fonction de liaison à *ras* (**Article 3** et **4**). Dans un cas (i.e., pour le mutant S77T), nous avons montré une augmentation significative de la stabilité. Il serait intéressant de vérifier l'effet d'autres mutants que nous n'avons pas encore testés soit : N64D, C81T, M83E, C96T, D113Q, S120D/E et E125K. Les mutations S120D/E sont particulièrement intéressantes. Le résidu S120 est prédit comme un site potentiel de phosphorylation par la caséine kinase I (http://scansite.mit.edu/, http://www.cbs.dtu.dk/services/NetPhos/). C'est en ce sens que la sélection des acides aminés Asp et Glu, généralement reconnus comme des mimétiques du groupement phosphate, est digne d'intérêt. Comme l'essai de sélection est effectué hors de son contexte cellulaire normal, soit chez la bactérie *E. coli*, on peut supposer que ces mutations stabiliseraient le DLR de Raf ou augmenteraient son affinité envers *ras*. Dans la même optique, il serait amusant de vérifier si la structure du DLR de Raf pourrait être stabilisée au-delà de ce que nous avons observée chez le mutant Δ104-6/S77T/R89L (**Article 3**). Afin de réaliser cet objectif et en accord avec l'effet stabilisant observé chez R89L, l'optimisation de l'empaquetage des résidus impliqués à l'interface de dimérisation (i.e. dans les segments K65-V69 et K84-V88) constituerait une stratégie adéquate. Cela pourrait contribuer à notre compréhension des liens structure/fonction établis au sein de la structure du DLR de Raf. Par exemple, pourquoi le maintien de la fonction de liaison est-elle incompatible à l'optimisation de la stabilité, soit dans une perspective structurale globale ou de dynamique locale?

- Les mécanismes de repliement du DLR de Raf et d'ubiquitine sont similaires (**Article 4**). Par ailleurs, l'ubiquitine a été soumise à une pléthore d'études qui ont permis de décrire certaines propriétés de sa structure et la présence d'intermédiaires potentiels sur sa voie de repliement (*171-175;186-192;252*). Il serait intéressant de pousser l'analyse des similitudes entre le DLR de Raf et l'ubiquitine plus avant, entre autre en réalisant des expériences équivalentes sur le DLR de Raf. En particulier, une expérience d'échange hydrogène/deutérium suivie par RMN pour le DLR de Raf permettrait de délimiter les réarrangements de la chaîne polypeptidique durant le processus de repliement. En second lieu, l'analyse des valeurs-Φ sur

ubiquitine, qui a été récemment publiée (*97*), devrait être complétée afin d'obtenir des points de comparaison similaires et aussi nombreux que ceux du DLR de Raf. Finalement, l'approche de l'insertion des tryptophanes dans la séquence polypeptidique du DLR de Raf et d'ubiquitine mise au point par A. Vallée-Belisle va aussi dans le sens de l'objectif décrit ici, en permettant éventuellement de comparer des intermédiaires potentiels sur leur voie de repliement.

- Dans le cadre de mes travaux, j'ai apporté des preuves supplémentaires des relations variables qui sont établies entre les diverses superfamilles de la topologie d'ubiquitine sur la base principalement de similitudes structurales (**Article 3** et **4**). Ainsi, nous avons confirmé que la superfamille d'ubiquitine et 5 superfamilles qui y seraient reliées évolutivement seraient plus similaires entre elles qu'aux 7 autres superfamilles (**Figure 6**). Il serait intéressant d'explorer les conséquences de ces observations en déterminant les caractéristiques thermodynamiques et cinétiques d'au moins un membre de chaque superfamille (il en reste 10 à caractériser en éliminant la superfamille d'ubiquitine et celle des protéines liant les immunoglobulines, qui est représentée par les protéines-L et G). Il n'y a pas eu jusqu'à maintenant d'étude de repliement rapportant des résultats sur une topologie comportant autant de superfamilles et en particulier sur les caractéristiques communes ou spécifiques des protéines appartenant à chacun des sous-groupes.

- Le mutant stabilisé Δ104-6 correspond à une délétion de trois résidus dans une boucle relativement flexible et distante des régions impliquées dans la stabilisation de l'état de transition et de l'état natif. De plus, la mutation E104A, n'a aucun effet significatif sur $\Delta G_{F\text{-}U}$ et k_f. Or, le mutant Δ104-106 affecte à la fois le k_f et le k_u. Une hypothèse crédible est que les acides aminés E104-K106 déstabilisent le tour-β situé dans cette région et que leur délétion lève une contrainte sur la voie de repliement. Il serait donc intéressant de vérifier cette hypothèse en déterminant comment cette délétion affecte le mécanisme de repliement du DLR de Raf. Pour y parvenir l'effet de la mutation en alanine des résidus H105 et K106, la délétion progressive ou simultanée de chacun des acides aminés de la région Δ104-106 (i.e. Δ104, Δ105, Δ106, Δ104-5, etc.) ainsi que l'insertion du tripeptide E104-K106 dans les protéines a-Raf et b-Raf sur la stabilité et les cinétiques de repliement/dépliement des DLR devraient être monitorées. L'insertion de divers mutants dans le contexte de ce variant et la détermination de leur valeur-Φ pourrait aussi permettre d'approfondir cette question.

- Il serait intéressant de confirmer la déviation apparente dans le graphe de Leffler, en ce qui concerne le motif épingle à cheveux du DLR de Raf (N56-V72). Pour y arriver, il faudrait obtenir beaucoup plus de mutants localisés dans cette région et démontrant un $\Delta\Delta G_{F\text{-}U}/RT > 4$ (**Article 4**). Comme, il y a peu de mutations ponctuelles additionnelles susceptibles de répondre à ce critère, on pourrait utiliser des mutations combinatoires comme cela a déjà été fait dans le cas de la barnase et

de CI2 (*122*). Étant donné que des déviations de ce genre sont habituellement attribuées à la présence de voies de repliement parallèles, la confirmation d'une telle anomalie dans le graphe de Leffler du DLR de Raf ne manquerait pas d'engendrer une série d'expériences ayant comme but de caractériser les voies alternatives suspectées.

Perspectives d'avenir, synergie entre les domaines de recherche et établissement de nouveaux paradigmes

Dans les sections qui suivent, je compte aborder spécifiquement, les études et les progrès qui émergent à l'heure actuelle en biologie structurale et spécifiquement dans l'étude du repliement, en d'autres mots, ce qui nous attend dans un futur pas si lointain. En relation avec cela, je vais brosser un tableau rapide des progrès des idées et concepts mis en place afin d'élaborer un modèle global de la réaction de repliement. En dernier lieu, je vais décrire quelques sujets qui se situent à l'interface entre d'une part, les études classiques du repliement *in vitro* et d'autre part, le repliement *in vivo*, l'implications des chaperonnes dans ce processus ainsi que l'organisation cellulaire normale et pathogénique.

Perspectives sur l'avenir de la recherche en biologie structurale

Quel avenir pour la recherche en biologie structurale? Je vais essayer de répondre à cette question en mettant en exergue les éléments que je trouve les plus prometteurs et intéressants.

Ironiquement, à l'ère post-génomique dans laquelle nous sommes supposément entrée, il n'y a jamais eu autant de projet de nature génomique. La biologie structurale n'a pas échappé à cette tendance et les cinq dernières années ont vu la mise sur pied de projets et de consortia de génomique structurale (http://www.rcsb.org/pdb/strucgen.html, statistiques à http://www.strgen.org/ et http://www.jcsg.org/, par exemple). À terme, l'objectif de ce champ d'études est de mettre au jour l'ensemble des structures qui sont retrouvées dans la nature et d'augmenter l'information de séquence pour certaines topologies peu fréquentes, en ayant comme cible l'objectif d'améliorer nos capacités de

prédire la structure et la fonction des protéines à partir de la séquence polypeptidique. Aux USA, cela s'est traduit par l'instauration du projet « protein structure initiative » (PSI) par le « National Institute of Health » (NIH), dont l'objectif initial était d'obtenir la structure de 10 000 protéines en 10 ans (*253*). À la suite d'une première phase de rodage, les efforts à consentir demeurent gigantesques. C'est quelques 65 millions $US sur 5 ans qui seront investis strictement de la part du NIH dans 5 centres de recherche qui produiront les structures et le développement technologique à cette entreprise. Ceci inclut l'automatisation de chaque étape du clonage à l'exposition aux rayons X en passant par la cristallisation qui a requis particulièrement d'attentions pour la conception de robots spécialisés. Jusqu'à maintenant les structures d'environ 500 protéines ont été obtenues et leur coût de production grâce à ces progrès techniques devrait sous peu être meilleur marché que dans les laboratoires traditionnels. Par ailleurs, il y a beaucoup de débats précisément autour de la question du perfectionnement technologique, et de la pertinence d'un support financier substantiel à cette question pour d'atteindre les objectifs initiaux du projet, et ce au moment où les budgets de la deuxième phase quinquennal du projet sont discutés. Du côté scientifiques, le nombre de nouvelles structures découvertes dans le cadre de ces projet est faible (entre 10-36%), malgré que les procédures expérimentales aient été mises au point afin de maximiser le ratio de nouvelles structures[32] (*254*). Le lecteur attentif de cette thèse n'en sera pas si surpris compte tenu du haut niveau de dégénérescence du code encrypté par la séquence polypeptidique que j'ai évoqué à plusieurs reprises. En dépit de cela, ces projets de grande envergure devraient fortement influencer le cours de la recherche dans les années qui viennent. Par exemple, dans le développement d'algorithmes de prédiction de structure, dont les performances peuvent être sans doute bonifiées par l'expansion de notre connaissance de l'espace de séquence et structurale occupée par l'univers protéique. Dans cette mouvance, l'intérêt pour le développement d'outils informatiques applicables à la prédiction de la structure des protéines est en forte hausse, cette discipline étant de surcroît couplée au développement d'outils en design de structure.

---

[32] Par exemple, en sélectionnant des gènes ne possédant pas d'homologie de séquence significative avec les protéines dont la structure est déjà connue.

Tous les deux ans depuis les dix dernières années, un concours dénommé le « critical assessment of structure prediction » (CASP) permet de mesurer les progrès des divers algorithmes de prédiction de structure. Il y a deux types principaux d'algorithmes de prédiction de structure selon leur mode opératoire, c'est-à-dire que la recherche de la conformation native soit faite *de novo* (modélisation libre, i.e. sans gabarit) ou bien par homologie de séquence (discussion des dernières avancées en prédiction de structure par l'optique de l'initiative CASP discuter dans (*255-259*)). La dernière approche est habituellement la plus précise. Il reste encore des progrès immenses à réaliser en ce qui concerne les protéines de plus de 200 acides aminés, en particulier à multi domaine et dans les cas où l'homologie de séquence est faible. En bref, cette initiative a contribué à mousser les efforts de la communauté scientifique et a émulé le développement de nouvelles approches basées sur l'expansion de notre connaissance globale des structures protéiques façonnées par l'évolution. On doit donc s'attendre à des progrès spectaculaires en ce domaine de recherche. D'autre part, bien que les méthodes de prédiction de structure *de novo* aient pour l'instant un retard significatif sur les approches par homologie de séquence, la solution idéale pourrait bien se retrouver dans la combinaison des deux types d'approches. En effet, leur combinaison pourrait permettre d'optimiser et de rationaliser les méthodes de prédiction de structure en les intégrant dans des outils bioinformatiques efficaces, précis et polyvalents. Par ailleurs, la méthode *de novo* appelée Rosette (*2;3*) se démarque des autres méthodes de prédiction de structure de son groupe par son efficacité et ses progrès remarquables (*260;261*), de même que par son application extrêmement impressionnante au design de structure.

Tel que je l'ai discuté à d'autres endroits dans ma thèse, la méthode Rosette a été appliquée avec succès au design de 9 protéines en se servant comme gabarit topologique de leur contrepartie naturelle (*4*). Pour six de ces exemples, les modèles artificiels s'avérèrent plus stables que leurs homologues de ts. Encore plus marquante est l'étude qui rapporta le design complet d'une structure adoptant une topologie non observée jusqu'à maintenant dans la nature (*5*). Sur un autre front, des progrès dans le design d'activités enzymatiques

nouvelles (*262;263*) ou de liaison (*6*) permettant l'élaboration d'appareil bioélectronique sophistiqué de détection de produits chimiques (*264*) ont été rapportés. Dans ce genre d'approches, au lieu de reformater intégralement la séquence, l'activité enzymatique ou la nouvelle fonction est intégrée à un endroit propice d'une protéine modèle à la structure connue. La prochaine étape sur le chemin du progrès consistera à combiner les outils de design de structure et d'activité enzymatique. En ce sens, les résultats du design de domaine WW basé sur l'utilisation d'alignement de séquence et sur l'identification des résidus couplés évolutivement suggèrent un moyen d'intégrer ces informations pour produire des variants protéiques non seulement structurés, mais ayant retenu en plus leur capacité de liaison naturelle (*155;156*). À la vue de ces exemples, il est clair que les progrès dans le champ d'études de la génomique structurale ainsi que dans la performance des outils de prédiction de la structure et de design seront couplés de manière synergique.

Attardons nous maintenant spécifiquement aux approches informatiques qui permettent l'étude dynamique des changements de conformation qui prennent place dans les protéines. Les simulations moléculaires peuvent modéliser les mouvements de tous les atomes de la chaîne polypeptidique en utilisant la structure native et des paramètres biophysiques en tant que contraintes. Par ces approches, il est possible d'obtenir un aperçu détaillé de la dynamique des conformations d'une protéine en conditions normales ou autres (par exemple à température ou pression élevée). Au cours des dernières années, les développements technologiques ont permis d'approcher le problème du dépliement et même du repliement des protéines (réviser dans (*265*)). À cause de la complexité inhérente au repliement, ce dernier a souvent été étudié en utilisant des modèles simplifiés sur treillis ou non. Par contre, les simulations de la réaction de dépliement peuvent prendre en compte explicitement tous les atomes, car elle est moins gourmande informatiquement[33] entre autre, parce que l'initiation de la réaction se fait à partir de l'état natif, qui expérimente

[33] Je rappelle que selon le principe de micro-réversibilité (voir section **De la dénaturation des protéines : perspectives historiques**), les voie de repliement et de dépliement sont postulées être inverses l'une de l'autre.

moins de changements conformationnels que l'état dénaturé, et par la capacité d'accélérer la réaction à haute température. Néanmoins à ce stade de mise au point, les simulations informatiques doivent encore être systématiquement évaluées par les expériences de cinétique et l'analyse des valeurs-Φ. Ainsi, ces deux catégories d'approches se renforcent mutuellement (réviser dans (*266;267*)). Entre autre le maillage entre la théorie et les expériences ont permis de décrire les voies de repliement et les caractéristiques des états dénaturés de CI2 et de barnase (*96;268-270*) ainsi que les propriété de l'état dénaturé de quelques autres protéines (*176*), d'obtenir une description de l'ensemble de l'état de transition pour AcP en se servant de l'analyse des valeurs-Φ comme contrainte (*271*), de prédire la présence d'un intermédiaire sur une voie de repliement (*272;273*), de décrire en détail et de comparer la voie de repliement de plusieurs homéodomaines et du domaine B de la protéine-A (*102;273;274*), ces dernières protéines se repliant suffisamment rapidement pour que ce processus soit modélisable en entier. À propos d'une question plus pointue dont j'ai traitée brièvement dans l'**Article 4**, les simulations moléculaires devraient permettre de déterminer le rôle respectif des contacts non-natifs versus natifs dans la stabilisation de l'état de transition (*244*). En abordant la question de l'angle des objectifs de ma thèse, la comparaison des simulations moléculaires du dépliement pour trois membres de la superfamille des homéodomaines est particulièrement intéressante. Dans cette étude, Gianni et coll. ont dénoté des changements dans les caractéristiques de la voie de repliement, c'est-à-dire de la nature des états de transition formés. En effet, ceux-ci passeraient des caractéristiques propres à la théorie de nucléation-condensation vers celles de la théorie de la charpente chez les trois homéodomaines étudiés (*274*). Nous reviendrons plus loin sur les retombées de cette observation sur notre perception actuelle et future du processus de repliement.

L'amélioration de ces méthodes entre autre par leur capacité à modéliser les réactions de repliement/dépliement sur des laps de temps prolongés pour des protéines plus complexes, comme pour l'amélioration des autres méthodes informatiques mentionnées ci-dessus, est tributaire de l'amélioration des capacités informatiques (i.e. la puissance des

processeurs et de leur coût à l'achat). Malgré les progrès importants des dernières années, les progrès techniques, qui restent à accomplir pour améliorer substantiellement la performance informatique en ces domaines, sont colossaux. Par exemple, dans le cadre du progrès Blue Gene™, IBM a mis sur pied une équipe multidisciplinaire afin de mettre au point un ordinateur géant capable de simuler la réaction de repliement. (http://www.research.ibm.com/ thinkresearch/pages/2001/20011105_protein.shtml). Cet ordinateur pourrait être composé de 30,000 des micro puces les plus puissantes qu'il prendrait environ 45 jours pour simuler le repliement d'une seule molécule d'une petite protéine. Cela donne une bonne idée de l'ordre de grandeur du saut technologique à accomplir et donc que bien du temps pourrait s'écouler avant que les simulations informatiques puissent être utilisées de manière étendue directement à la simulation directe du repliement des protéines. D'autre part, le projet « folding @ home » (http://folding.stanford.edu/) permet de tirer parti de l'imposante communauté d'internautes répartis à travers le globe pour impartir des tâches informatiques et ainsi augmenter la capacité et la durée des simulations. Jusqu'à maintenant, les simulations moléculaires entreprises grâce à ce système ont été d'envergure limitée. Par ailleurs, la qualité principale d'un modèle en pratique est sa prédictibilité, i.e. son utilité pour poser des hypothèses valables sur un processus vérifiable expérimentalement. À cet égard, Fersht et coll. dans plusieurs des publications ci-dessus et dans d'autres ont démontré pour plusieurs protéines la corrélation entre les résultats de l'analyse des valeurs-Φ et les simulations, en dépit même des limitations technologiques. Une objection qui est souvent opposée à l'encontre des simulations moléculaires, c'est le fait que les résultats obtenus sont représentatifs d'un petit nombre de simulations indépendantes du dépliement d'une seule molécule protéique à la fois (*275*). C'est précisément à la comparaison du repliement d'un échantillon de protéine par les méthodes classiques versus une molécule unique que de nouvelles approches expérimentales, contre toute attente, permettent de s'attaquer.

Deux types d'approches – la spectroscopie de force mécanique et le transfert d'énergie de résonance de Förster (FRET) – permettent maintenant de suivre le repliement

d'une seule molécule protéique à la fois (réviser dans (*276*)). Les techniques de FRET adaptées à l'étude du repliement des protéines (ou dans n'importe quel cas où l'on s'intéresse aux remaniements intramoléculaires d'une protéine) sont basées sur l'insertion d'un fluorophore accepteur et d'un donneur, habituellement situés à longue distance dans la structure primaire, mais à proximité dans la native. Dans ce cas de figure, le niveau de FRET est maximal à l'état natif, puisqu'il dépend de la distance entre les deux fluorophores. Par conséquent, des montages expérimentaux divers – les plus intéressantes portant sur l'étude de protéine en diffusion libre (*277-281*) – ont permis de faire des expériences à l'équilibre et cinétiques dans lesquelles le signal de fluorescence de chaque molécule protéique est évalué individuellement. L'avantage comparatif de l'étude du repliement fondée sur le suivi du changement des caractéristiques des molécules protéiques au niveau unimoléculaire par rapport aux approches classiques, c'est la possibilité du moins en théorie de déterminer directement la distribution des molécules en sous-populations (i.e. état dénaturé, intermédiaire et natif) en fonction des conditions de dénaturation. Dans les premières études qui ont été réalisées à l'équilibre des protéines ayant une réaction de repliement deux-états – CI2 et une protéine du choc thermique froid – se comporte comme prévu par les études classiques en ne démontrant durant l'expérience de FRET que deux populations distinctes, assimilables aux états dénaturé et natif (*277;279*). Avec un nouvel assemblage permettant la mesure du taux de repliement en fonction de la concentration de dénaturant, le même groupe de recherche à démontrer la protéine du choc thermique froid suivait une transition correspondant à une fonction exponentielle simple équivalente en tout point à celle observée par l'intermédiaire d'un montage expérimental classique (*280*).

La spectroscopie de force mécanique permet de déplier une molécule de protéine directement par son étirement et suite au relâchement de cette force, d'observer son repliement (*192;252;282;283;* réviser dans (*284;285*)). J'ai déjà discuté de son application à la détermination des propriétés d'extension et de stabilité des chaînes de poly-ubiquitine et des concordances de ces données avec des preuves de la présence d'intermédiaire(s) chez les monomères d'ubiquitine qui ont été obtenues par divers types d'expériences de RMN

(discuté respectivement dans cette section et **Données à l'équilibre recueillies sur ubiquitine et suggérant la présence d'un intermédiaire tardif**). Par contre, les données sur la cinétique de repliement d'ubiquitine suite à la dénaturation par force mécanique ont cependant de quoi rendre perplexe, avec leur transition qui ne correspond pas à une transition typique tout ou rien (*252*). Il y a plusieurs problèmes techniques inhérents à cette approche qui pourrait expliquer ce résultat et qui devront être résolus avant de généraliser son application à l'étude du repliement et d'en tirer des conclusions qui remettraient en cause la théorie générale du repliement des protéines. Les mêmes précautions s'appliquent bien entendu à toutes ces nouvelles approches. D'autre part, lorsque ces méthodes seront bien maîtrisées, elles devraient apporter des informations intéressantes, différentes et probablement inaccessibles par les méthodes classiques à propos de la réaction de repliement, particulièrement en ce qui concerne les méthodes basées sur le FRET. La comparabilité et la synergie de ces méthodes avec les simulations moléculaires seront intéressantes à vérifier.

L'étude comparative du repliement de protéines extrêmement dissemblables au niveau de la séquence, mais adoptant des structures analogues devrait s'intensifier afin de faire progresser notre compréhension de la réaction de repliement, de la conservation de la topologie et de l'évolution de la structure des protéines. À ce stade-ci, moins d'une dizaine de paires de protéines peuvent être comparées. Dorénavant, il faudrait veiller à sélectionner des topologies structurales variées afin de couvrir de façon plus homogène l'univers protéique. Il faudrait aussi hausser le niveau de rigueur avec laquelle les comparaisons de séquences et de structures ainsi que les choix des modèles sont effectués dans le cadre d'études comparatives. Il est possible d'y arriver en utilisant de manière systématique les banques de données structurales tel que SCOP, CATH et FSSP afin de choisir des sujets d'études possédant une identité de séquence relativement faible et des similarités structurales plus ou moins grandes. J'ai donné un exemple précédemment de la manière dont cette façon de procéder pourrait s'appliquer à la topologie d'ubiquitine (voir section **Retour sur mes travaux**).

Il est évident que les approches informatiques seront de plus en plus utilisées en recherche biologique. Pour des raisons intrinsèques et historiques, la biologie structurale a toujours été en avance sur son époque à ce niveau. En ce sens, nous verrons assurément ce domaine de recherche accorder à moyen terme une place encore plus grande aux approches théoriques et informatiques.

Vers un modèle global de la réaction de repliement

Notre vision du repliement est en phase de passer de théories contradictoires et spécifiques selon les protéines étudiées à un modèle unificateur. Les efforts de plusieurs groupes ont contribué à cela. Je voudrais souligner ici l'importance particulière les travaux des groupes d'A. Fersht et V. Daggett et de D. Baker. Comme j'ai tenté de le décrire dans l'introduction, il y a à peu près 10 ans, deux théories principales s'affrontaient soit la théorie de nucléation-condensation et de la charpente, reposant respectivement sur des données et des analyses riches obtenues expérimentalement sur CI2 et barnase. La théorie de nucléation-condensation en particulier rompait avec plusieurs des aspects sur lesquelles l'emphase est mise dans la théorie du repliement séquentiel. À cet égard, il est instructif de rappeler quelques unes des implications de ce modèle (*89*) pour ensuite les confronter à la vision actuelle de son principal concepteur :

> « (i) A nucleation mechanism and variations thereon in which the nucleation site occurs only flickeringly in the denatured state and no folding intermediates accumulate is an efficient folding pathway of a small protein.
>
> (ii) Evolutionary pressure therefore opposes the accumulation of nucleation site in the denatured state. Sites should become stable only after interacting with other parts of the structure. This discourages searches for structures in isolated fragments by experimentalists or for initiation sites in intact proteins by theoreticians using methods that disregard long distance interactions.
>
> (iii) Evolutionary pressure should minimize nonnative hydrophobic interactions in denatured state since such interactions lower the energies of the denatured states. »

Plusieurs de ces implications ont été remises en cause à la lumière des conclusions basées sur des études couplant des simulations moléculaires et des expériences dans la dernière décennie (voir section précédente et **L'état déplié et/ou dénaturé**).

Justement, voyons l'effet du passage du temps sur la conception qu'expose Daggett et Fersht dans un article de revue de la littérature, entre autre sur les intermédiaires (*266*): « It is our opinion that at the molecular level, intermediates are always present ». Cette citation est typique d'un courant de pensées qui prend de l'ampleur en ce moment et qui mène à la marginalisation de l'opinion contraire qui apparaissait dominante, il n'y a encore pas très longtemps. La structure obtenue pour l'état dénaturé/intermédiaire de l'homéodomaine d'engrailed (*168*) (voir section **L'état déplié et/ou dénaturé**), la comparaison du processus repliement des protéines Im7 et Im9 (*104*) (voir section **Intermédiaires**) ou l'analyse des données de cinétiques du repliement/dépliement de plusieurs protéines répertoriées dans la littérature (*125;127*) sont pleinement concordantes avec cette mouvance. Toujours à propos de Daggett et Fersht, quelle est leur vision actuelle de l'état dénaturé et de l'initiation de la réaction de repliement (*266*) : « The intrinsic conformational properties of the secondary structure and particularly persistent tertiary interactions can help to direct this search by starting the process from denatured states with residual structure ». Il est réconfortant en regard de l'honnêteté de la démarche intellectuelle de ces scientifiques de noter que le groupe de recherche qui a contribué à mettre sur pied la théorie de nucléation-condensation réduise lui-même l'ampleur de la distinction entre les deux théories descriptives principales du processus de repliement. Pour poursuivre cette discussion, je souhaite revenir à l'étude présentant la comparaison de la voie de repliement de trois membre de la superfamille des homéodomaines par l'analyse des valeurs-Φ et des simulations (*274*). Le point clé de cette étude, tel que mentionné avant, est la diversité des voies de repliement observée chez cette topologie : les modèles de nucléation-condensation et de charpente sont distinctement observables ou bien s'hybrident selon l'analogue structural en question, alors que par ailleurs, leur état de transition est très similairement placé par rapport à l'état natif (**Figure 23**). Ce schéma révèle que des

protéines peuvent avoir des divergences dans les détails des états rencontrés et observables sur leur voie de repliement tout en ayant un état de transition similaire (d'autres exemples révisés dans (*136*)). Plus fondamentalement, cela suggère qu'il n'y a pas de différence majeure entre les diverses théories de repliement et qu'il y aurait un continuum entre les modèles qui décriraient le mieux telle ou telle autre réaction de repliement, les distinctions apparaissant plutôt comme un épiphénomène. En quelques sorte, les théories descriptives du mécanisme repliement sont en phase d'être réconciliées, du moins empiriquement.

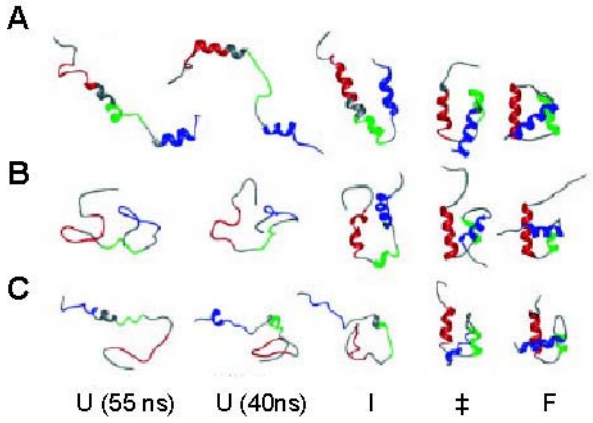


Figure 23. Comparaison des divers états rencontrés sur la voie de repliement de trois membres de la superfamille des homéodomaines.
Les états rencontrés sur les voies de repliement ont été modélisés par des simulations moléculaires sur les homéodomaines ci-joints. **A**, Engrailed. **B**, c-Myb. **C**, hTRF1. Notez les similitudes dans la structure de l'état de transition. En comparaison, il y a plus de différences entre les états qui le précédent. Engrailed **(A)** se replierait en accord ave le modèle de la charpente alors que hTRF1 **(C)** le ferait via le mécanisme de nucléation-condensation. c-Myb semble adopter un mécanisme intermédiaire (adapter de (*274*) avec la permission d'un des auteurs responsables).

En résumé, dans l'esprit de cette nouvelle vision, l'adoption de l'un des deux des modèles susmentionnés par une protéine donnée dépendrait simplement du niveau de propension de sa séquence à stabiliser des éléments des structures secondaire versus tertiaire.

La plupart des états de transition caractérisés à date sont très structurés[34], comme l'indique leur placement généralement plus près de l'état natif. En ce sens, la conformation de l'état natif joue un rôle crucial dans la détermination du processus de repliement, en

[34] A l'exception notable des protéines se repliant très rapidement, entre autre les faisceux d'hélices-α.

particulier de k_f et expliquerait d'une façon simple la variation de 6 ordres de grandeur observée pour le k_f des petites protéines (réviser dans (*286*)) (voir section **Topologie et ordre de contact**). Par contre, certaines protéines ne suivent pas cette corrélation (voir l'**Article 4** pour une discussion sur ce sujet). Il serait intéressant de déterminer les raisons de la déviation de ce comportement pour certaines protéines. La présence d'un fort contenu en acides aminés hydrophobes, de structures résiduelles dans l'état dénaturé ou d'états intermédiaires productifs sont des facteurs qui pourraient accélérer le taux de repliement de la chaîne polypeptidique.

Un autre trait d'union tracé dans les dernières années a été obtenu par la comparaison du mécanisme de repliement d'analogues structuraux en tirant profit de la méthode d'ingénie des protéines[35]. Ces travaux ont mené à la constatation que les protéines partageant la même topologie ne se replient pas nécessairement via un état de transition semblable en particulier logiquement lorsque leur homologie de séquence est très faible[36] (voir section **Les protéines partageant la même topologie se replient-elles par un mécanisme identique : oui et non** et réviser dans (*136*)). Une étude de design qui permit de changer la position du nucléus polarisé de l'état de transition de la protéine-G se révéla très convaincante à la validation de cette idée (*135*). Ainsi, en ce qui me concerne la quadrature du cercle était accomplie, la dégénérescence du message encodé par la structure primaire se répercutait en quelque sorte sur les caractéristiques de l'état de transition chez des analogues structuraux.

La désolvatation de l'intérieur des protéines au cours du repliement constituent un élément propre au repliement des protéines globulaires. Conséquemment, une avenue que je trouve particulièrement intéressante est la prise en compte explicite des molécules de solvant dans les simulations, en particulier l'effet de la désolvatation sur les contacts

---

[35] Je rappelle que sept paires de protéines ont été comparées à ce jour par analyse des valeurs-Φ, mais un peu plus si l'on considère les données cinétiques brutes des protéines de ts (*136*).

[36] L'état de transition d'ubiquitine et du DLR de Raf (identité de séquence < 12%) dont il est question dans l'**Article 4**, AcP et ADA2H ainsi que deux domaines immunoglobuline constituent des exceptions notoires.

établis entre les résidus. Il y a eu très peu d'études à ce sujet, mais celle présentée par Cheung et coll. suggère que ce type de considérations identifierait des lieux communs fondamentaux dans la transition d'énergie entre les états dénaturé et natif (*287*), particulièrement tout juste avant ou après l'état de transition, qui incluraient un processus menant à l'expulsion coopérative des molécules d'H_2O du cœur hydrophobe de la protéine.

Finalement, il sera intéressant de voir comment les modèles du repliement des petites protéines pourront être utilisés et adaptés aux protéines de haut poids moléculaire qui sont plus courantes dans les cellules eucaryotes. Outre que les protéines plus lourdes soient souvent composées de nombreuses unités de repliement indépendantes ou non, qu'elles puissent donc nécessiter l'intervention de chaperonnes moléculaires *in vivo*, d'autres propriétés fondamentales de leur empaquetage (*20*) et du milieu cellulaire laisse supposer des divergences moins attendues.

Relations entre le processus de repliement *in vitro* et *in vivo*, la structure native, la fonction, la biologie cellulaire et l'organisation de la vie

À plusieurs égards les processus cellulaires et biologiques étudiés *in vitro* sont susceptibles de se distinguer des processus équivalents tels qu'ils prennent place dans leur contexte physiologique normal. Alors que les conditions de température et de pH sont habituellement prudemment ajustées, l'un des facteurs qui est ignoré trop souvent dans la préparation des expériences réalisées *in vitro* est le phénomène dit d'encombrement moléculaire (« molecular crowding ») ou l'effet du volume exclus (réviser dans (*288-291*)) (**Figure 24**). En effet, à l'intérieur d'une cellule l'encombrement moléculaire est plus élevé que dans les tampons habituellement utilisés pour diluer les protéines d'intérêt lors des expériences réalisées *in vitro*. Cela découle de la concentration supérieure en macromolécules de toutes sortes incluant les acides nucléiques, les sucres, les lipides, les protéines et le cytosquelette que l'on retrouve dans une cellule. En chiffres clairs, la concentration de protéine lors des expériences qui nous ont permis de déterminer le taux de repliement du DLR de Raf était de 0.04 g/*l* (4 μM), alors que la concentration des protéines

varie de 200 à 400 g/*l* selon le type cellulaire[37]. L'encombrement moléculaire à l'intérieur des cellules a pour effet de restreindre le volume réellement disponible à une macromolécule[38] dans un volume donné et donc d'augmenter sa concentration effective par rapport à une solution diluée préparée par purification. Il est clair que des processus biologiques incluant les interactions moléculaires[39] (i.e. protéine-protéine, protéine-acide nucléique, etc.), la diffusion passive des molécules, les activités enzymatiques et plus fondamentalement la stabilité des protéines et leur réaction de repliement peuvent être affectées par ce phénomène.

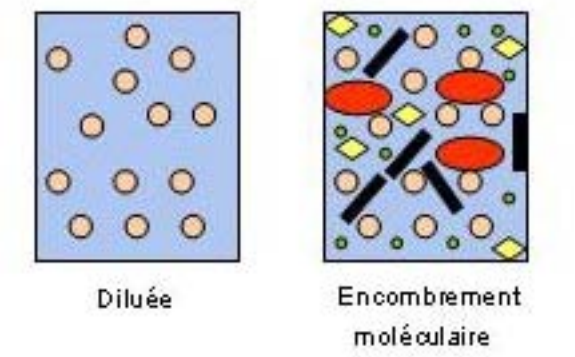


Figure 24. Schéma expliquant l'effet d'une solution de macromolécules concentrée sur le volume exclus et la concentration effective.
La concentration d'une macromolécule d'intérêt (cercle saumon) dépend bien sûr du nombre de ces molécules en solution dans un volume donné. Dans les deux panneaux ci-dessus, la concentration de cette macromolécule est égale. Par contre, la concentration effective est plus grande dans le panneau de droite. En effet, la quantité de solvant libre (bleu) diminue en fonction du nombre de macromolécules dissoutes. Cela illustre les variations de conditions entre les solutions diluées habituellement utilisées *in vitro* et la concentration des macromolécules à l'intérieur des cellules.

D'autre part, des études de repliement du lysozyme dont les ponts dissulfures avaient été réduits au préalable ont démontré une accélération du repliement correct et une

---

[37] Le DLR de Raf étant un très petit domaine, la concentration tel qu'exprimée (g/*l*) n'est pas représentative de la majorité des protéines. Cependant, même pour les protéines plus grosses, la différence devrait être d'au moins deux ordres de grandeur entre la concentration totale des protéines *in vitro* et *in vivo*.
[38] L'effet de l'encombrement moléculaire est hautement dépendant de la taille de la molécule d'intérêt et est significatif principalement pour les macromolécules.
[39] Par exemple, la constante d'équilibre d'une dimérisation pourrait augmenter d'un facteur de 8-40 fois; la tétramérisation de 10^3-10^5 fois dans le cytoplasme bactérien versus les protocoles classiques d'expérimentation *in vitro* (*289*).

tendance accrue à l'agrégation qui pouvaient être compensées *in vivo* par l'intervention des chaperonnes cellulaires appropriées (*292;293*). Par ailleurs, des expériences de simulations informatiques reposant sur un modèle de protéine globulaire simplifié ont indiqué que le taux de repliement est accéléré par une augmentation de l'encombrement moléculaire probablement via une déstabilisation de l'état dénaturé dont la structure étendue est défavorisée dans ce contexte (*294*). Or, il existe des moyens expérimentaux simples afin de mimer l'encombrement moléculaire cellulaire qui réside dans l'ajout de certains composés chimiques tel que le dextran 70 et le Ficoll 70. Leur utilisation généralisée dans les tampons de renaturation/dénaturation des modèles protéiques étudiés par les méthodes spectroscopiques régulières permettrait de vérifier l'importance du volume exclus directement sur les taux et le processus de repliement déterminés expérimentalement.

L'encombrement moléculaire via le phénomène d'agrégation noté au paragraphe précédent pourrait avoir favorisé la sélection au cours de l'évolution des propriétés oligomériques de la structure de plusieurs protéines (*295*), le processus d'oligomérisation étant fréquemment couplé chez les oligomères obligatoires[40] à la réaction de repliement en tant que tel (*118;119*). Ainsi, les propriétés de la formation et de la stabilisation de la structure des polypeptides auraient modelé l'organisation de la vie.

Un autre processus biologique pour lequel l'étude du repliement pourrait apporter un éclairage intéressant est l'allostérie. L'allostérie a été étudiée initialement dans des enzymes et chez l'héméglobine principalement. Des cas d'allostérie chez des protéines ne possédant pas d'activité enzymatique ou *a priori* sous forme monomérique commencent à être répertoriés. D'autre part, des régions flexibles ou dynamiques dans les structures RMN pourraient indiquer des régions impliquées dans des phénomènes d'allostérie (*297*), de telle sorte que celle-ci pourrait bien être une propriété intrinsèque du moins sous la forme de potentiel pour toutes les protéines (*298*). Les régions les plus dynamiques d'une structure donnée pourraient théoriquement être identifiées aussi par des expériences de repliement,

car elles adopteraient leur conformation native tardivement ou auraient peu d'impact sur la stabilité et le taux de repliement. L'émergence de méga complexe protéique impliqué entre autre chose dans les voies de signalisation (i.e. signalosome) suggère que le phénomène d'allostérie serait plus étendu qu'initialement estimé (réviser dans (*296*)). Par exemple, des cas nouveaux et très intéressants d'allostérie ont été décrits pour certaines protéines impliquées dans la transmission de signaux intracellulaires. Les fonctions respectives de la protéine du syndrome de Wiskott-Aldrich neural (N-WASP) et de la protéine tyrosine phosphatase 1b (PTP1b) peuvent en effet être régulées par la liaison de petites molécules chimiques respectivement, en périphérie du domaine de liaison au GTPase (GBD) et du site actif (*299;300*). Dans chacun de ces cas, les inhibiteurs chimiques bloquent ces protéines dans une conformation inactive. Dans le cas de PTP1b la liaison de l'inhibiteur induit de nombreux remaniements structuraux notamment dans une boucle située à proximité du site de liaison, cette nouvelle conformation correspondant à la conformation inactive naturelle qui avait déjà été étudiée. En ce qui concerne N-WASP les changements sont plus dramatiques, le domaine GBD se repliant au moment de son interaction avec l'inhibiteur, stabilisant ainsi la structure auto inhibitrice aussi décrite auparavant (*301*). Ce mécanisme d'allostérie et d'auto inhibition pour un domaine non-enzymatique s'assimile parfaitement bien à la vague d'intérêt suscitée par ces protéines ou sous-régions de protéines qui ne forment pas de structures stables, plus précisément sur la sélection de cette propriété pour l'accomplissement de certaines fonctions cellulaires chez plusieurs protéines (réviser dans (*1*)). En effet, la combinaison de l'obtention de séquence par l'entremise des divers projets génomiques, d'expériences de RMN et de prédiction de structure a permis de souligner l'importance numérique jusqu'à présent sous-estimée des protéines n'adoptant pas de structure native bien définie. Parmi les avantages biologiques des protéines ou domaines non-structurés, il y a la grande affinité et versatilité des surfaces d'interactions potentielles. Pour illustrer cela, mentionnons à nouveau N-WASP, dont la section amino-terminale du domaine GBD adopte des conformations distinctes en fonction de sa liaison à la GTPase ou

---

[40] En opposition aux monomères formant des oligomères induits (voir section **Connaissances de base sur la structure native des protéines : structure primaire, secondaire, tertiaire et quaternaire**).

bien en conformation inhibée (*301*). Pour prendre un exemple concret plus près du sujet de cette thèse, la région carboxy-terminale de l'ubiquitine qui semble se replier tardivement pourrait représenter un exemple assimilable conceptuellement à un principe élargi d'allostérie (*188-190;192*). En effet, il est connu que la formation de chaîne poly-ubiquitine à partir d'une lysine dans cette région cible[41] la protéine marquée vers le protéasome. On pourrait supposer que la flexibilité structurale autour du point d'ancrage dans ce type de chaîne poly-ubiquitine agit comme un signal moléculaire qui permet la reconnaissance par la machinerie protéolytique et donc la dégradation, et que ce message est discernable de ceux convoyés par les autres types de chaînes poly-ubiquitine. Dans le cas du DLR de Raf, plusieurs indices suggèrent la présence de régions dynamiques. La plus sérieuse est la différence perceptible dans l'hélice-α qui semble se tordre dans le complexe formé avec Rap1A (*218;219*). D'autres régions telles le tour-β précédent l'hélice-α et ceux autour de la région E94-C96 et E104-K108 démontrent des conformations variables dans la structure obtenue par RMN (*217*). Il sera intéressant de voir évoluer au cours des prochaines années notre compréhension de l'effet des régions non-structurées et des boucles non seulement dans la fonction des protéines multi domaines, mais aussi sur la réaction de repliement étudiée *in vitro* et la fonction des protéines modèle étudiées.

Notre compréhension du mécanisme de repliement des protéines et de son adaptation évolutive au contexte cellulaire passera par une meilleure intégration de notre connaissance de ce processus *in vitro* avec la fonction des chaperonnes moléculaires et des molécules impliquées dans la synthèse protéique dans les cellules. Or, la complexité de l'intégration des ces données réside précisément dans les différences notables entre les réactions de repliement prenant place dans ces deux contextes. D'abord, le repliement *in vivo*, du moins dans le cytoplasme des eucaryotes, s'opère de manière co-traductionnelle, i.e. bien avant que la traduction de l'ARN_m par le ribosome soit complétée. En effet, cela est indubitable considérant la comparaison entre le taux d'élongation de la chaîne

[41] L'ubiquitine des mammifères possède 7 lysines dont 4 ont été formellement identifiées dans des chaînes polyubiqutine.

polypeptidique qui est de 2-8 acides aminés par secondes chez les eucaryotes et le taux de repliement d'un domaine de 100 acides aminés qui est de l'ordre des millisecondes à quelques secondes. Par conséquent, il est clair que des éléments de structures locaux auront été formés avant la fin de la synthèse protéique quel que soit la taille de la chaîne polypeptidique considérée. Premièrement, on peut déduire que la réaction de repliement apparaîtrait plus séquentielle, si on pouvait la suivre de façon détaillée dans ce contexte-ci, *in vivo* qu'*in vitro*. Par ailleurs, le moment du déclenchement du repliement et le rôle du ribosome n'est toujours pas clairement établi : dans quelle mesure le repliement est-il enclenché à l'intérieur du ribosome? Est-ce que le ribosome agit comme une chaperonne? Etc. Par la suite, l'ampleur du rôle des chaperonnes est toujours floue. Dans la voie de sécrétion, ces dernières seraient tout simplement essentielles au repliement des protéines complexes, qui sont modifiées substantiellement par des processus post-traductionnels dans ce compartiment cellulaire (i.e. principalement des glycosylations et la formation de ponts dissulfures spécifiques) (réviser dans (*302*)). D'autre part, dans le cytoplasme un nombre restreint de protéines nécessiterait le concours des chaperonnes afin de se replier vers leur structure native (rôle des chaperonnes cytoplasmiques révisé dans (*303*)). Cette prédiction est confortée par l'abondance de protéines se repliant de manière autonome *in vitro*, bien qu'il faille préciser que nos connaissances à cet égard sont limitées généralement à des protéines de petite taille (< 250 acides aminés). Revenons à la notion de l'encombrement moléculaire, et son effet prédictible sur le repliement des protéines dans le contexte cellulaire. Le rôle que les chaperonnes pourrait jouer pour compenser ce phénomène est largement méconnu pour l'instant, bien que l'impact du volume exclus sur le repliement du lysozyme indique que son impact général pourrait être non-négligeable (voir plus haut dans cette section). Chez la bactérie *E. coli*, l'essentialité du rôle de la chaperonne cytosolique GroEL/GroES dans le repliement des protéines a été démontrée pour seulement 84 des 2400 protéines exprimées dans ce compartiment en utilisant une approche qui combinait des méthodes protéomiques et génétiques (*304*). Parmi les protéines requérant absolument GroEL/GroES, 13 sont nécessaires à la viabilité des souches bactériennes. Une second classe inclut 126 protéines supplémentaires qui démontrent une dépendance moins grande

envers cette chaperonne. Tout au plus 150 autres protéines se rajoutent à cette liste, si l'expérience est dirigée à partir d'une souche ou d'autres chaperonnes majeures ne sont pas exprimées. Comme la chaperonne GroEL/GroES est la seule à être essentielle à la croissance en toutes conditions, ces chiffres fournissent une bonne estimation du rôle des chaperonnes dans le repliement des protéines chez les bactéries dont la niche écologique correspond à des conditions relativement douces de température et de milieu. À cause de la taille moyenne plus élevée de leurs protéines et du ratio substantiel (environ 30%) de celles-ci intégrant la voie de sécrétion, les eucaryotes devraient certainement avoir un ratio plus élevé de leur protéome qui dépendrait des chaperonnes pour se replier. En bref, il semblerait que les chaperonnes moléculaires sont importantes pour le repliement d'un nombre relativement faible des protéines cytosoliques. Elles y arrivent par deux mécanismes distincts, soit en isolant le polypeptide non-natif du milieu cellulaire ou bien en le maintenant dans un état compatible avec le repliement par une interaction directe induite par des cycles d'hydrolyse des nucléotides, principalement l'ATP. Qu'elles appartiennent à l'une ou l'autre de ces classes, les chaperonnes préviennent l'agrégation des protéines incorrectement repliées. Cette fonction joue un rôle prépondérant dans la résistance à l'accumulation de structures amyloïdes[42] caractéristiques d'une vingtaine de conditions pathologiques dégénératives qui affectent en particulier le système nerveux central, tel que l'Alzheimer, les maladies à prions ainsi que le parkinson. C'est dans le but de mieux comprendre et de combattre ces maladies que le processus d'agrégation et les mécanismes de résistance sont étudiés *in vitro* sur des modèles protéiques simplifiés à l'aide des méthodes conventionnelles de l'étude du repliement.

Au cours des dernières années, l'étude du processus d'agrégation qui est impliqué dans plusieurs maladies dégénératives a connu une véritable révolution, lorsqu'il a été démontré que la capacité de formation des plaques et des fibrilles étaient une propriété propre à des protéines aux structures diverses et non-amyloïdes à l'état basal, et contre

[42] Les caractéristiques biophysiques des structures amyloïdes indiquent une organisation commune composée de feuillets-β principalement.

toute attente que la cytotoxicité de ces molécules était équivalente à leurs analogues naturellement agrégeant et pathogéniques (réviser dans (*305;306*)). Même de petites protéines se repliant via un mécanisme deux-états forment des structures amyloïdes. Parmi les premières protéines de ce type dont les propriétés d'agrégation ont été étudiées, mentionnons le domaine SH3 de la sous-unité p85alpha de la 3-kinase du phosphatidylinositol, le module 9 de la fibronectine de type III et l'AcP (*307-309*). En particulier les données sur l'AcP ont permis de définir en profondeur les déterminants de séquence de son agrégation (*310;311*). D'après ces données, il semblerait que c'est l'espèce susceptible à l'agrégation et non les plaques en tant que telles qui serait responsable de la toxicité dans ces maladies. Cela découlerait du fait que la première espèce est plus réactive et peut conséquemment entraîner plusieurs protéines ou macromolécules dans les agrégats déposés. Cela est intéressant parce que cette espèce devrait être plus facile à cibler au niveau pharmacologique et qu'accessoirement cela rehausse l'intérêt de s'intéresser à la dynamique de ce phénomène au niveau plus fondamental de la recherche. A ce propos, des résultats récents suggèrent que le titrage des chaperonnes moléculaires par les protéines défectueuses serait responsable de la cytotoxicité cellulaire en surchargeant le mécanisme de contrôle de la qualité. Dans les prochaines années, ces nouveaux modèles du processus de formation des plaques amyloïdes et des fibrilles en combinaison avec les simulations informatiques devraient nous en apprendre plus sur l'agrégation pathogénique des protéines, en particulier sur leurs caractéristiques structurales. Cela devrait nous permettre éventuellement de contrecarrer efficacement ces processus dégénératifs.

Si je devais résumer en une phrase ce que je voulais accomplir en décrivant les exemples cités dans cette section, je dirais que nous devrions observer au cours des prochaines années une intégration accrue de nos connaissances sur le phénomène de repliement *in vitro* et en biologie structurale à celles sur le fonctionnement et le rôle des protéines dans le contexte cellulaire normal et/ou pathologique. À l'instar de plusieurs autres domaines en sciences biologiques l'ère de l'union des savoirs est bien entamée. Je pense qu'il y aura là matière à excitation neuronale !

Bibliographie

1. Dyson, H. J. and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions, *Nat. Rev. Mol. Cell Biol. 6*, 197-208.
2. Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions, *J. Mol. Biol. 268*, 209-225.
3. Simons, K. T., Strauss, C., and Baker, D. (2001) Prospects for ab initio protein structural genomics, *J. Mol. Biol. 306*, 1191-1199.
4. Dantas, G., Kuhlman, B., Callender, D., Wong, M., and Baker, D. (2003) A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins, *J. Mol. Biol. 332*, 449-460.
5. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy, *Science 302*, 1364-1368.
6. Dwyer, M. A., Looger, L. L., and Hellinga, H. W. (2004) Computational design of a biologically active enzyme, *Science 304*, 1967-1971.
7. Soding, J. and Lupas, A. N. (2003) More than the sum of their parts: on the evolution of proteins from peptides, *Bioessays 25*, 837-846.
8. Panchenko, A. R., Luthey-Schulten, Z., Cole, R., and Wolynes, P. G. (1997) The foldon universe: a survey of structural similarity and self-recognition of independently folding units, *J. Mol. Biol. 272*, 95-105.
9. Sanger, F. (1952) The arrangement of amino acids in proteins, *Adv. Protein Chem. 7*, 1-67.
10. Anfinsen, C. B. and Redfield, R. R. (1956) Protein structure in relation to function and biosynthesis, *Adv. Protein Chem. 48*, 1-100.
11. Anfinsen, C. B. and Scheraga, H. A. (1975) Experimental and theoretical aspects of protein folding, *Adv. Protein Chem. 29*, 205-300.
12. Voet, D. and Voet, J. G. (1994) Amino Acids, in *Biochemistry* 2nd ed., pp 56-70, Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore.
13. Voet, D. and Voet, J. G. (1994) Three-Dimensionnal Structures of Proteins, in *Biochemistry* 2nd ed., pp 141-190, John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore.
14. Hutchinson, E. G. and Thornton, J. M. (1994) A revised set of potentials for beta-turn formation in proteins, *Protein Sci. 3*, 2207-2216.
15. Chou, K. C. and Blinn, J. R. (1997) Classification and prediction of beta-turn types, *J. Protein Chem. 16*, 575-595.
16. Maity, H., Maity, M., and Englander, S. W. (2004) How cytochrome c folds, and why: submolecular foldon units and their stepwise sequential stabilization, *J. Mol. Biol. 343*, 223-233.
17. Tanford, C. (1978) The hydrophobic effect and the organization of living matter, *Science 200*, 1012-1018.
18. Voet, D. and Voet, J. G. (1994) Covalent Structures of Proteins, in *Biochemistry* 2nd ed., pp 105-140, Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore.
19. Kauzmann, W. (1959) Some factors in the interpretation of protein denaturation, *Adv. Protein Chem. 14*, 1-63.
20. Liang, J. and Dill, K. A. (2001) Are proteins well-packed?, *Biophys. J. 81*, 751-766.
21. Bonneau, R., Strauss, C. E., and Baker, D. (2001) Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation, *Proteins 43*, 1-11.
22. Campbell-Valois, F. X., Tarassov, K., and Michnick, S. W. (2005) Massive Sequence Perturbation of a Small Protein, *Proc. Natl. Acad. Sci. U.S.A. 102*, 14988-14993.
23. Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis, *Nature 181*, 662-666.

24. Perutz, M. F., Kendrew, J. C., and Watson, H. C. (1965) Structure and function of haemoglobin, *J. Mol. Biol. 13*, 669-678.
25. Rossmann, M. G. and Argos, P. (1976) Exploring structural homology of proteins, *J. Mol. Biol. 105*, 75-95.
26. Richardson, J. S. (1977) beta-Sheet topology and the relatedness of proteins, *Nature 268*, 495-500.
27. Richardson, J. S. (1981) The anatomy and taxonomy of protein structure, *Adv. Protein Chem. 34*, 167-339.
28. Lesk, A. M. and Chothia, C. (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins, *J. Mol. Biol. 136*, 225-270.
29. Ponting, C. P. and Benjamin, D. R. (1996) A novel family of Ras-binding domains, *Trends Biochem. Sci. 21*, 422-425.
30. Michnick, S. W. and Shakhnovich, E. (1998) A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies, *Fold. Des 3*, 239-251.
31. Larson, S. M. and Davidson, A. R. (2000) The identification of conserved interactions within the SH3 domain by alignment of sequences and structures, *Protein Sci. 9* , 2170-2180.
32. Larson, S. M., Di Nardo, A. A., and Davidson, A. R. (2000) Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions, *J. Mol. Biol. 303* , 433-446.
33. Hill, E. E., Morea, V., and Chothia, C. (2002) Sequence conservation in families whose members have little or no sequence similarity: the four-helical cytokines and cytochromes, *J. Mol. Biol. 322*, 205-233.
34. « Conserved Key Amino Acids in Protein sequences » (CKAAPs), banque de données, http://ckaaps.sdsc.edu/perl/browser.pl.
35. « Family of Structurally Similar Proteins » (FSSP), banque de données, http://srs6.ebi.ac.uk/srsbin/cgi-bin/wgetz.
36. Kiel, C. and Serrano, L. (2005) The Ubiquitin Domain Superfold: Structure-based Sequence Alignments and Characterization of Binding Epitopes, *J. Mol. Biol.*, publié à l'avance sur Internet.
37. Expasy Swiss-Prot TrEMBL, banque de données, http://ca.expasy.org/sprot/.
38. Valdar, W. S. (2002) Scoring residue conservation, *Proteins 48*, 227-241.
39. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997) CATH--a hierarchic classification of protein domain structures, *Structure. 5*, 1093-1108.
40. « Class, Architecture, Topology and Homologous superfamily » (CATH) , banque de données, http://www.biochem.ucl.ac.uk/bsm/cath/.
41. Holm, L. and Sander, C. (1994) The FSSP database of structurally aligned protein fold families, *Nucleic Acids Res. 22*, 3600-3609.
42. Holm, L. and Sander, C. (1996) The FSSP database: fold classification based on structure-structure alignment of proteins, *Nucleic Acids Res. 24*, 206-209.
43. Reddy, B. V., Li, W. W., Shindyalov, I. N., and Bourne, P. E. (2001) Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins, *Proteins 42*, 148-163.
44. Li, W. W., Reddy, B. V., Shindyalov, I. N., and Bourne, P. E. (2001) CKAAPs DB: a conserved key amino acid positions database, *Nucleic Acids Res. 29*, 329-331.
45. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol. 247*, 536-540.
46. Hubbard, T. J., Murzin, A. G., Brenner, S. E., and Chothia, C. (1997) SCOP: a structural classification of proteins database, *Nucleic Acids Res. 25*, 236-239.
47. Hubbard, T. J., Ailey, B., Brenner, S. E., Murzin, A. G., and Chothia, C. (1999) SCOP: a Structural Classification of Proteins database, *Nucleic Acids Res. 27*, 254-256.
48. Lo, C. L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G., and Chothia, C. (2000) SCOP: a structural classification of proteins database, *Nucleic Acids Res. 28*, 257-259.
49. « Structural Classification of Proteins » (SCOP), banque de données, http://scop.mrc-lmb.cam.ac.uk/scop/.

50. Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure, *J. Mol. Biol. 313*, 903-919.
51. « Simple Module Architecture Research Tool » (SMART), banque de données, http://smart.embl-heidelberg.de/.
52. « Protein FAMilies » (PFAM), banque de données, http://www.sanger.ac.uk/Software/Pfam/.
53. Reidhaar-Olson, J. F. and Sauer, R. T. (1988) Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences, *Science 241*, 53-57.
54. Lim, W. A. and Sauer, R. T. (1989) Alternative packing arrangements in the hydrophobic core of lambda repressor, *Nature 339*, 31-36.
55. Bowie, J. U. and Sauer, R. T. (1989) Identifying determinants of folding and activity for a protein of unknown structure, *Proc. Natl. Acad. Sci. U.S.A. 86*, 2152-2156.
56. Axe, D. D., Foster, N. W., and Fersht, A. R. (1996) Active barnase variants with completely random hydrophobic cores, *Proc. Natl. Acad. Sci. U.S.A. 93*, 5590-5594.
57. Finucane, M. D. and Woolfson, D. N. (1999) Core-directed protein design. II. Rescue of a multiply mutated and destabilized variant of ubiquitin, *Biochemistry 38*, 11613-11623.
58. Gu, H., Yi, Q., Bray, S. T., Riddle, D. S., Shiau, A. K., and Baker, D. (1995) A phage display system for studying the sequence determinants of protein folding, *Protein Sci. 4*, 1108-1117.
59. Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q., and Baker, D. (1997) Functional rapidly folding proteins from simplified amino acid sequences, *Nat. Struct. Biol. 4*, 805-809.
60. Gu, H., Kim, D., and Baker, D. (1997) Contrasting roles for symmetrically disposed beta-turns in the folding of a small protein, *J. Mol. Biol. 274*, 588-596.
61. Kim, D. E., Gu, H., and Baker, D. (1998) The sequences of small proteins are not extensively optimized for rapid folding by natural selection, *Proc. Natl. Acad. Sci. U.S.A. 95*, 4982-4986.
62. Coco, W. M., Encell, L. P., Levinson, W. E., Crist, M. J., Loomis, A. K., Licato, L. L., Arensdorf, J. J., Sica, N., Pienkos, P. T., and Monticello, D. J. (2002) Growth factor engineering by degenerate homoduplex gene family recombination, *Nat. Biotechnol. 20*, 1246-1250.
63. Sidhu, S. S., Lowman, H. B., Cunningham, B. C., and Wells, J. A. (2000) Phage display for selection of novel binding peptides, *Methods Enzymol. 328*, 333-363.
64. Campbell-Valois, F. X. and Michnick, S. W. (2004) Synthesis of Libraries and Screening with the DHFR PCA, *In Press*.
65. Michnick, S. W., Remy, I., Campbell-Valois, F. X., Vallee-Belisle, A., and Pelletier, J. N. (2000) Detection of protein-protein interactions by protein fragment complementation strategies, *Methods Enzymol. 328*, 208-230.
66. Michnick, S. W. (2001) Exploring protein interactions by interaction-induced folding of proteins from complementary peptide fragments, *Curr. Opin. Struct. Biol. 11*, 472-477.
67. Pelletier, J. N., Arndt, K. M., Pluckthun, A., and Michnick, S. W. (1999) An in vivo library-versus-library selection of optimized protein-protein interactions, *Nat. Biotechnol. 17*, 683-690.
68. Waldo, G. S., Standish, B. M., Berendzen, J., and Terwilliger, T. C. (1999) Rapid protein-folding assay using green fluorescent protein, *Nat. Biotechnol. 17*, 691-695.
69. Waldo, G. S. (2003) Genetic screens and directed evolution for protein solubility, *Curr. Opin. Chem. Biol. 7*, 33-38.
70. Minard, P., Scalley-Kim, M., Watters, A., and Baker, D. (2001) A "loop entropy reduction" phage-display selection for folded amino acid sequences, *Protein Sci. 10*, 129-134.
71. Scalley-Kim, M., Minard, P., and Baker, D. (2003) Low free energy cost of very long loop insertions in proteins, *Protein Sci. 12*, 197-206.
72. Anson, M. L. (1945) Protein Denaturation and the Properties of Protein Groups, *Adv. Protein Chem. 2*, 361-386.
73. Wu, H. (1931) Studies on Denaturation of Proteins XIII. A Theory of Denaturation, *Chinese J. Physiol. 5*, 321-344.

74. Mirsky, A. E. and PAULING, L. (1936) On the Structure of Native, Denatured and Coagulated Protein, *Proc. Natl. Acad. Sci. U.S.A. 22*, 439-447.
75. Tanford, C. (1968) Protein denaturation, *Adv. Protein Chem. 23*, 121-282.
76. Winter, G., Fersht, A. R., Wilkinson, A. J., Zoller, M., and Smith, M. (1982) Redesigning enzyme structure by site-directed mutagenesis: tyrosyl tRNA synthetase and ATP binding, *Nature 299*, 756-758.
77. Fersht, A. R. (1987) Dissection of the structure and activity of the tyrosyl-tRNA synthetase by site-directed mutagenesis, *Biochemistry 26*, 8031-8037.
78. Fersht, A. R., Matouschek, A., and Serrano, L. (1992) The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding, *J. Mol. Biol. 224*, 771-782.
79. Fersht, A. R. (1995) Characterizing transition states in protein folding: an essential step in the puzzle, *Curr. Opin. Struct. Biol. 5*, 79-84.
80. Matouschek, A., Serrano, L., and Fersht, A. R. (1992) The folding of an enzyme. IV. Structure of an intermediate in the refolding of barnase analysed by a protein engineering procedure, *J. Mol. Biol. 224*, 819-835.
81. Matouschek, A., Kellis, J. T., Jr., Serrano, L., and Fersht, A. R. (1989) Mapping the transition state and pathway of protein folding by protein engineering, *Nature 340*, 122-126.
82. Serrano, L., Matouschek, A., and Fersht, A. R. (1992) The folding of an enzyme. III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure, *J. Mol. Biol. 224*, 805-818.
83. Otzen, D. E., Itzhaki, L. S., elMasry, N. F., Jackson, S. E., and Fersht, A. R. (1994) Structure of the transition state for the folding/unfolding of the barley chymotrypsin inhibitor 2 and its implications for mechanisms of protein folding, *Proc. Natl. Acad. Sci. U.S.A. 91*, 10422-10425.
84. Itzhaki, L. S., Neira, J. L., Ruiz-Sanz, J., de Prat, G. G., and Fersht, A. R. (1995) Search for nucleation sites in smaller fragments of chymotrypsin inhibitor 2, *J. Mol. Biol. 254*, 289-304.
85. Jackson, S. E. and Fersht, A. R. (1991) Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition, *Biochemistry 30*, 10428-10435.
86. Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. (1995) The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding, *J. Mol. Biol. 254*, 260-288.
87. Abkevich, V. I., Gutin, A. M., and Shakhnovich, E. I. (1994) Specific nucleus as the transition state for protein folding: evidence from the lattice model, *Biochemistry 33*, 10026-10036.
88. Wetlaufer, D. B. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins, *Proc. Natl. Acad. Sci. U.S.A. 70*, 697-701.
89. Fersht, A. R. (1995) Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications, *Proc. Natl. Acad. Sci. U.S.A. 92*, 10869-10873.
90. Lopez-Hernandez, E. and Serrano, L. (1996) Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2, *Fold. Des 1*, 43-55.
91. Villegas, V., Martinez, J. C., Aviles, F. X., and Serrano, L. (1998) Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain, *J. Mol. Biol. 283*, 1027-1036.
92. Fulton, K. F., Main, E. R., Daggett, V., and Jackson, S. E. (1999) Mapping the interactions present in the transition state for unfolding/folding of FKBP12, *J. Mol. Biol. 291*, 445-461.
93. Hamill, S. J., Steward, A., and Clarke, J. (2000) The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology, *J. Mol. Biol. 297*, 165-178.
94. Cota, E., Steward, A., Fowler, S. B., and Clarke, J. (2001) The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold, *J. Mol. Biol. 305*, 1185-1194.
95. Fowler, S. B. and Clarke, J. (2001) Mapping the folding pathway of an immunoglobulin domain: structural detail from Phi value analysis and movement of the transition state, *Structure. (Camb.) 9*, 355-366.

96. Kazmirski, S. L., Wong, K. B., Freund, S. M., Tan, Y. J., Fersht, A. R., and Daggett, V. (2001) Protein folding from a highly disordered denatured state: the folding pathway of chymotrypsin inhibitor 2 at atomic resolution, *Proc. Natl. Acad. Sci. U.S.A. 98*, 4349-4354.
97. Went, H. M. and Jackson, S. E. (2005) Ubiquitin folds through a highly polarized transition state, *Protein Eng Des Sel 18*, 229-237.
98. Kim, D. E., Fisher, C., and Baker, D. (2000) A breakdown of symmetry in the folding transition state of protein L, *J. Mol. Biol. 298*, 971-984.
99. McCallister, E. L., Alm, E., and Baker, D. (2000) Critical role of beta-hairpin formation in protein G folding, *Nat. Struct. Biol. 7*, 669-673.
100. Anil, B., Sato, S., Cho, J. H., and Raleigh, D. P. (2005) Fine structure analysis of a protein folding transition state; distinguishing between hydrophobic stabilization and specific packing, *J. Mol. Biol. 354*, 693-705.
101. Garcia-Mira, M. M., Boehringer, D., and Schmid, F. X. (2004) The folding transition state of the cold shock protein is strongly polarized, *J. Mol. Biol. 339*, 555-569.
102. Sato, S., Religa, T. L., Daggett, V., and Fersht, A. R. (2004) From The Cover: Testing protein-folding simulations by experiment: B domain of protein A, *Proc. Natl. Acad. Sci. U.S.A. 101*, 6952-6956.
103. Capaldi, A. P., Kleanthous, C., and Radford, S. E. (2002) Im7 folding mechanism: misfolding on a path to the native state, *Nat. Struct. Biol. 9*, 209-216.
104. Friel, C. T., Capaldi, A. P., and Radford, S. E. (2003) Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: similarities and differences in the folding of homologous proteins, *J. Mol. Biol. 326*, 293-305.
105. Kragelund, B. B., Osmark, P., Neergaard, T. B., Schiodt, J., Kristiansen, K., Knudsen, J., and Poulsen, F. M. (1999) The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP, *Nat. Struct. Biol. 6*, 594-601.
106. Jager, M., Nguyen, H., Crane, J. C., Kelly, J. W., and Gruebele, M. (2001) The folding mechanism of a beta-sheet: the WW domain, *J. Mol. Biol. 311*, 373-393.
107. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I., and Baker, D. (1999) Experiment and theory highlight role of native state topology in SH3 folding, *Nat. Struct. Biol. 6*, 1016-1024.
108. Martinez, J. C. and Serrano, L. (1999) The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved, *Nat. Struct. Biol. 6*, 1010-1016.
109. Northey, J. G., Di Nardo, A. A., and Davidson, A. R. (2002) Hydrophobic core packing in the SH3 domain folding transition state, *Nat. Struct. Biol. 9*, 126-130.
110. Guerois, R. and Serrano, L. (2000) The SH3-fold family: experimental evidence and prediction of variations in the folding pathways, *J. Mol. Biol. 304*, 967-982.
111. Lorch, M., Mason, J. M., Clarke, A. R., and Parker, M. J. (1999) Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the I-state, *Biochemistry 38*, 1377-1385.
112. Otzen, D. E. and Oliveberg, M. (2002) Conformational plasticity in folding of the split beta-alpha-beta protein S6: evidence for burst-phase disruption of the native state, *J. Mol. Biol. 317*, 613-627.
113. Ternstrom, T., Mayor, U., Akke, M., and Oliveberg, M. (1999) From snapshot to movie: phi analysis of protein folding transition states taken one step further, *Proc. Natl. Acad. Sci. U.S.A. 96*, 14854-14859.
114. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M., and Dobson, C. M. (1999) Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding, *Nat. Struct. Biol. 6*, 1005-1009.
115. Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. L., and Oas, T. G. (1997) The energy landscape of a fast-folding protein mapped by Ala-->Gly substitutions, *Nat. Struct. Biol. 4*, 305-310.

116. Nolting, B., Golbik, R., Neira, J. L., Soler-Gonzalez, A. S., Schreiber, G., and Fersht, A. R. (1997) The folding pathway of a protein at high resolution from microseconds to seconds, *Proc. Natl. Acad. Sci. U.S.A. 94*, 826-830.
117. Chu, R., Pei, W., Takei, J., and Bai, Y. (2002) Relationship between the native-state hydrogen exchange and folding pathways of a four-helix bundle protein, *Biochemistry 41*, 7998-8003.
118. Milla, M. E., Brown, B. M., Waldburger, C. D., and Sauer, R. T. (1995) P22 Arc repressor: transition state properties inferred from mutational effects on the rates of protein unfolding and refolding, *Biochemistry 34*, 13914-13919.
119. Mateu, M. G., Sanchez Del Pino, M. M., and Fersht, A. R. (1999) Mechanism of folding and assembly of a small tetrameric protein domain from tumor suppressor p53, *Nat. Struct. Biol. 6*, 191-198.
120. Choe, S. E., Li, L., Matsudaira, P. T., Wagner, G., and Shakhnovich, E. I. (2000) Differential stabilization of two hydrophobic cores in the transition state of the villin 14T folding reaction, *J. Mol. Biol. 304*, 99-115.
121. Seeliger, M. A., Breward, S. E., and Itzhaki, L. S. (2003) Weak cooperativity in the core causes a switch in folding mechanism between two proteins of the cks family, *J. Mol. Biol. 325*, 189-199.
122. Fersht, A. R., Itzhaki, L. S., elMasry, N. F., Matthews, J. M., and Otzen, D. E. (1994) Single versus parallel pathways of protein folding and fractional formation of structure in the transition state, *Proc. Natl. Acad. Sci. U.S.A. 91*, 10426-10429.
123. Matthews, J. M. and Fersht, A. R. (1995) Exploring the energy surface of protein folding by structure-reactivity relationships and engineered proteins: observation of Hammond behavior for the gross structure of the transition state and anti-Hammond behavior for structural elements for unfolding/folding of barnase, *Biochemistry 34*, 6805-6814.
124. Sanchez, I. E. and Kiefhaber, T. (2003) Origin of unusual phi-values in protein folding: evidence against specific nucleation sites, *J. Mol. Biol. 334*, 1077-1085.
125. Sanchez, I. E. and Kiefhaber, T. (2003) Hammond behavior versus ground state effects in protein folding: evidence for narrow free energy barriers and residual structure in unfolded states, *J. Mol. Biol. 327*, 867-884.
126. Matouschek, A., Otzen, D. E., Itzhaki, L. S., Jackson, S. E., and Fersht, A. R. (1995) Movement of the position of the transition state in protein folding, *Biochemistry 34*, 13656-13662.
127. Sanchez, I. E. and Kiefhaber, T. (2003) Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding, *J. Mol. Biol. 325*, 367-376.
128. Kragelund, B. B., Hojrup, P., Jensen, M. S., Schjerling, C. K., Juul, E., Knudsen, J., and Poulsen, F. M. (1996) Fast and one-step folding of closely and distantly related homologous proteins of a four-helix bundle family, *J. Mol. Biol. 256*, 187-200.
129. Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M. A., Jaenicke, R., and Schmid, F. X. (1998) Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins, *Nat. Struct. Biol. 5*, 229-235.
130. Sato, S., Xiang, S., and Raleigh, D. P. (2001) On the relationship between protein stability and folding kinetics: a comparative study of the N-terminal domains of RNase HI, E. coli and Bacillus stearothermophilus L9, *J. Mol. Biol. 312*, 569-577.
131. Vallee-Belisle, A., Turcotte, J. F., and Michnick, S. W. (2004) raf RBD and Ubiquitin Proteins Share Similar Folds, Folding Rates and Mechanisms Despite Having Unrelated Amino Acid Sequences, *Biochemistry 43*, 8447-8458.
132. Martinez, J. C., Pisabarro, M. T., and Serrano, L. (1998) Obligatory steps in protein folding and the conformational diversity of the transition state, *Nat. Struct. Biol. 5*, 721-729.
133. Northey, J. G., Maxwell, K. L., and Davidson, A. R. (2002) Protein folding kinetics beyond the phi value: using multiple amino acid substitutions to investigate the structure of the SH3 domain folding transition state, *J. Mol. Biol. 320*, 389-402.
134. Kim, D. E., Yi, Q., Gladwin, S. T., Goldberg, J. M., and Baker, D. (1998) The single helix in protein L is largely disrupted at the rate-limiting step in folding, *J. Mol. Biol. 284*, 807-815.

135. Nauli, S., Kuhlman, B., and Baker, D. (2001) Computer-based redesign of a protein folding pathway, *Nat. Struct. Biol. 8*, 602-605.
136. Zarrine-Afsar, A., Larson, S. M., and Davidson, A. R. (2005) The family feud: do proteins with similar structures fold via the same pathway?, *Curr. Opin. Struct. Biol. 15*, 42-49.
137. Plaxco, K. W., Simons, K. T., and Baker, D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins, *J. Mol. Biol. 277*, 985-994.
138. Plaxco, K. W., Simons, K. T., Ruczinski, I., and Baker, D. (2000) Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics, *Biochemistry 39*, 11177-11183.
139. Baker, D. (2000) A surprising simplicity to protein folding, *Nature 405*, 39-42.
140. Grantcharova, V., Alm, E. J., Baker, D., and Horwich, A. L. (2001) Mechanisms of protein folding, *Curr. Opin. Struct. Biol. 11*, 70-82.
141. Makarov, D. E., Keller, C. A., Plaxco, K. W., and Metiu, H. (2002) How the folding rate constant of simple, single-domain proteins depends on the number of native contacts, *Proc. Natl. Acad. Sci. U.S.A. 99*, 3535-3539.
142. Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D., and Finkelstein, A. V. (2003) Contact order revisited: influence of protein size on the folding rate, *Protein Sci. 12*, 2057-2062.
143. Dill, K. A., Fiebig, K. M., and Chan, H. S. (1993) Cooperativity in protein-folding kinetics, *Proc. Natl. Acad. Sci. U.S.A. 90*, 1942-1946.
144. Weikl, T. R. and Dill, K. A. (2003) Folding rates and low-entropy-loss routes of two-state proteins, *J. Mol. Biol. 329*, 585-598.
145. Lindberg, M. O., Tangrot, J., Otzen, D. E., Dolgikh, D. A., Finkelstein, A. V., and Oliveberg, M. (2001) Folding of circular permutants with decreased contact order: general trend balanced by protein stability, *J. Mol. Biol. 314*, 891-900.
146. Larson, S. M. and Pande, V. S. (2003) Sequence optimization for native state stability determines the evolution and folding kinetics of a small protein, *J. Mol. Biol. 332*, 275-286.
147. Di Nardo, A. A., Larson, S. M., and Davidson, A. R. (2003) The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core, *J. Mol. Biol. 333*, 641-655.
148. Calloni, G., Taddei, N., Plaxco, K. W., Ramponi, G., Stefani, M., and Chiti, F. (2003) Comparison of the folding processes of distantly related proteins. Importance of hydrophobic content in folding, *J. Mol. Biol. 330*, 577-591.
149. Ventura, S., Vega, M. C., Lacroix, E., Angrand, I., Spagnolo, L., and Serrano, L. (2002) Conformational strain in the hydrophobic core and its implications for protein folding and design, *Nat. Struct. Biol. 9*, 485-493.
150. Shakhnovich, E., Abkevich, V., and Ptitsyn, O. (1996) Conserved residues and the mechanism of protein folding, *Nature 379*, 96-98.
151. Mirny, L. A., Abkevich, V. I., and Shakhnovich, E. I. (1998) How evolution makes proteins fold quickly, *Proc. Natl. Acad. Sci. U.S.A. 95*, 4976-4981.
152. Mirny, L. and Shakhnovich, E. (2001) Evolutionary conservation of the folding nucleus, *J. Mol. Biol. 308*, 123-129.
153. Plaxco, K. W., Larson, S., Ruczinski, I., Riddle, D. S., Thayer, E. C., Buchwitz, B., Davidson, A. R., and Baker, D. (2000) Evolutionary conservation in protein folding kinetics, *J. Mol. Biol. 298*, 303-312.
154. Larson, S. M., Ruczinski, I., Davidson, A. R., Baker, D., and Plaxco, K. W. (2002) Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation, *J. Mol. Biol. 316*, 225-233.
155. Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B., and Ranganathan, R. (2005) Natural-like function in artificial WW domains, *Nature 437*, 579-583.
156. Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H., and Ranganathan, R. (2005) Evolutionary information for specifying a protein fold, *Nature 437*, 512-518.

157. Ramachandran, G. N. and Sasisekharan, V. (1968) Conformation of polypeptides and proteins, *Adv. Protein Chem. 23*, 283-438.
158. Levinthal, C. (1969) How to Fold Graciously. In "Mossbauer Spectroscopy in Biological Systems", (Debrunner, P., Tsibris, J. C. M., and Munck, E., Eds.) 67 ed., pp 22-24, University of Illinois Press, Urbanna.
159. Haber, E. and Anfinsen, C. B. (1961) Regeneration of enzyme activity by air oxidation of reduced subtilisin-modified ribonuclease, *J. Biol. Chem. 236*, 422-424.
160. Anfinsen, C. B. and Haber, E. (1961) Studies on the reduction and re-formation of protein disulfide bonds, *J. Biol. Chem. 236*, 1361-1363.
161. Levinthal, C. (1968) Are there Pathways for Protein Folding, *Journal de Chimie Physique et de Physico-Chimie Biologique 65*, 44-45.
162. Kim, P. S. and Baldwin, R. L. (1982) Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding, *Annu. Rev. Biochem. 51* , 459-489.
163. Privalov, P. L. (1979) Stability of proteins: small globular proteins, *Adv. Protein Chem. 33*, 167-241.
164. Privalov, P. L. (1982) Stability of proteins. Proteins which do not present a single cooperative system, *Adv. Protein Chem. 35*, 1-104.
165. Baldwin, R. L. and Rose, G. D. (1999) Is protein folding hierarchic? I. Local structure and peptide folding, *Trends Biochem. Sci. 24*, 26-33.
166. Eaton, W. A., Munoz, V., Hagen, S. J., Jas, G. S., Lapidus, L. J., Henry, E. R., and Hofrichter, J. (2000) Fast kinetics and mechanisms in protein folding, *Annu. Rev. Biophys. Biomol. Struct. 29*, 327-359.
167. Luisi, D. L., Wu, W. J., and Raleigh, D. P. (1999) Conformational analysis of a set of peptides corresponding to the entire primary sequence of the N-terminal domain of the ribosomal protein L9: evidence for stable native-like secondary structure in the unfolded state, *J. Mol. Biol. 287*, 395-407.
168. Religa, T. L., Markson, J. S., Mayor, U., Freund, S. M., and Fersht, A. R. (2005) Solution structure of a protein denatured state and folding intermediate, *Nature 437*, 1053-1056.
169. Munoz, V., Thompson, P. A., Hofrichter, J., and Eaton, W. A. (1997) Folding dynamics and mechanism of beta-hairpin formation, *Nature 390*, 196-199.
170. Srinivasan, R. and Rose, G. D. (1999) A physical basis for protein secondary structure, *Proc. Natl. Acad. Sci. U.S.A. 96*, 14258-14263.
171. Cox, J. P., Evans, P. A., Packman, L. C., Williams, D. H., and Woolfson, D. N. (1993) Dissecting the structure of a partially folded protein. Circular dichroism and nuclear magnetic resonance studies of peptides from ubiquitin, *J. Mol. Biol. 234*, 483-492.
172. Searle, M. S., Williams, D. H., and Packman, L. C. (1995) A short linear peptide derived from the N-terminal sequence of ubiquitin folds into a water-stable non-native beta-hairpin, *Nat. Struct. Biol. 2*, 999-1006.
173. Zerella, R., Evans, P. A., Ionides, J. M., Packman, L. C., Trotter, B. W., Mackay, J. P., and Williams, D. H. (1999) Autonomous folding of a peptide corresponding to the N-terminal beta-hairpin from ubiquitin, *Protein Sci. 8*, 1320-1331.
174. Zerella, R., Chen, P. Y., Evans, P. A., Raine, A., and Williams, D. H. (2000) Structural characterization of a mutant peptide derived from ubiquitin: implications for protein folding, *Protein Sci. 9*, 2142-2150.
175. Jourdan, M., Griffiths-Jones, S. R., and Searle, M. S. (2000) Folding of a beta-hairpin peptide derived from the N-terminus of ubiquitin. Conformational preferences of beta-turn residues dictate non-native beta-strand interactions, *Eur. J. Biochem. 267*, 3539-3548.
176. Zagrovic, B., Snow, C. D., Khaliq, S., Shirts, M. R., and Pande, V. S. (2002) Native-like mean structure in the unfolded ensemble of small proteins, *J. Mol. Biol. 323*, 153-164.
177. Baldwin, R. L. and Rose, G. D. (1999) Is protein folding hierarchic? II. Folding intermediates and transition states, *Trends Biochem. Sci. 24*, 77-83.
178. Karplus, M. and Weaver, D. L. (1976) Protein-folding dynamics, *Nature 260*, 404-406.
179. Dill, K. A. (1985) Theory for the folding and stability of globular proteins, *Biochemistry 24*, 1501-1509.

180. Alonso, D. O. and Dill, K. A. (1991) Solvent denaturation and stabilization of globular proteins, *Biochemistry 30*, 5974-5985.
181. Agashe, V. R., Shastry, M. C., and Udgaonkar, J. B. (1995) Initial hydrophobic collapse in the folding of barstar, *Nature 377*, 754-757.
182. Gutin, A. M., Abkevich, V. I., and Shakhnovich, E. I. (1995) Is burst hydrophobic collapse necessary for protein folding?, *Biochemistry 34*, 3066-3076.
183. Ptitsyn, O. B. (1987) Protein Folding: Hypothesis and Experiments., *J. Prot. Chem. 6*, 273-293.
184. Ptitsyn, O. B. (1995) Molten globule and protein folding, *Adv. Protein Chem. 47*, 83-229.
185. Uversky, V. N. and Fink, A. L. (2004) Conformational constraints for amyloid fibrillation: the importance of being unfolded, *Biochim. Biophys. Acta 1698*, 131-153.
186. Harding, M. M., Williams, D. H., and Woolfson, D. N. (1991) Characterization of a partially denatured state of a protein by two-dimensional NMR: reduction of the hydrophobic interactions in ubiquitin, *Biochemistry 30*, 3120-3128.
187. Stockman, B. J., Euvrard, A., and Scahill, T. A. (1993) Heteronuclear three-dimensional NMR spectroscopy of a partially denatured protein: the A-state of human ubiquitin, *J. Biomol. NMR 3*, 285-296.
188. Kitahara, R., Yamada, H., and Akasaka, K. (2001) Two folded conformers of ubiquitin revealed by high-pressure NMR, *Biochemistry 40*, 13556-13563.
189. Kitahara, R. and Akasaka, K. (2003) Close identity of a pressure-stabilized intermediate with a kinetic intermediate in protein folding, *Proc. Natl. Acad. Sci. U.S.A. 100*, 3167-3172.
190. Briggs, M. S. and Roder, H. (1992) Early hydrogen-bonding events in the folding reaction of ubiquitin, *Proc. Natl. Acad. Sci. U.S.A. 89*, 2017-2021.
191. Varadan, R., Walker, O., Pickart, C., and Fushman, D. (2002) Structural properties of polyubiquitin chains in solution, *J. Mol. Biol. 324*, 637-647.
192. Carrion-Vazquez, M., Li, H., Lu, H., Marszalek, P. E., Oberhauser, A. F., and Fernandez, J. M. (2003) The mechanical stability of ubiquitin is linkage dependent, *Nat. Struct. Biol. 10*, 738-743.
193. Went, H. M., Benitez-Cardoza, C. G., and Jackson, S. E. (2004) Is an intermediate state populated on the folding pathway of ubiquitin?, *FEBS Lett. 567*, 333-338.
194. Krantz, B. A., Mayne, L., Rumbley, J., Englander, S. W., and Sosnick, T. R. (2002) Fast and slow intermediate accumulation and the initial barrier mechanism in protein folding, *J. Mol. Biol. 324*, 359-371.
195. Friel, C. T., Beddard, G. S., and Radford, S. E. (2004) Switching two-state to three-state kinetics in the helical protein Im9 via the optimisation of stabilising non-native interactions by design, *J. Mol. Biol. 342*, 261-273.
196. Khorasanizadeh, S., Peters, I. D., and Roder, H. (1996) Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues, *Nat. Struct. Biol. 3*, 193-205.
197. Krantz, B. A. and Sosnick, T. R. (2000) Distinguishing between two-state and three-state models for ubiquitin folding, *Biochemistry 39*, 11696-11701.
198. Khorasanizadeh, S., Peters, I. D., Butt, T. R., and Roder, H. (1993) Folding and stability of a tryptophan-containing mutant of ubiquitin, *Biochemistry 32*, 7054-7063.
199. Inaba, K., Kobayashi, N., and Fersht, A. R. (2000) Conversion of two-state to multi-state folding kinetics on fusion of two protein foldons, *J. Mol. Biol. 302*, 219-233.
200. Batey, S., Randles, L. G., Steward, A., and Clarke, J. (2005) Cooperative folding in a multi-domain protein, *J. Mol. Biol. 349*, 1045-1059.
201. Martin, A. and Schmid, F. X. (2003) The folding mechanism of a two-domain protein: folding kinetics and domain docking of the gene-3 protein of phage fd, *J. Mol. Biol. 329*, 599-610.
202. Leopold, P. E., Montal, M., and Onuchic, J. N. (1992) Protein folding funnels: a kinetic approach to the sequence-structure relationship, *Proc. Natl. Acad. Sci. U.S.A. 89*, 8721-8725.
203. Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z., and Socci, N. D. (1995) Toward an outline of the topography of a realistic protein-folding funnel, *Proc. Natl. Acad. Sci. U.S.A. 92*, 3626-3630.

204. Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997) Theory of protein folding: the energy landscape perspective, *Annu. Rev. Phys. Chem. 48*, 545-600.
205. Onuchic, J. N. and Wolynes, P. G. (2004) Theory of protein folding, *Curr. Opin. Struct. Biol. 14*, 70-75.
206. Ozkan, S. B., Bahar, I., and Dill, K. A. (2001) Transition states and the meaning of Phi-values in protein folding kinetics, *Nat. Struct. Biol. 8*, 765-769.
207. Lyubovitsky, J. G., Gray, H. B., and Winkler, J. R. (2002) Mapping the cytochrome C folding landscape, *J. Am. Chem. Soc. 124*, 5481-5485.
208. Wright, C. F., Lindorff-Larsen, K., Randles, L. G., and Clarke, J. (2003) Parallel protein-unfolding pathways revealed and mapped, *Nat. Struct. Biol. 10*, 658-662.
209. Dill, K. A. and Chan, H. S. (1997) From Levinthal to pathways to funnels, *Nat. Struct. Biol. 4*, 10-19.
210. Wong, K. B., Clarke, J., Bond, C. J., Neira, J. L., Freund, S. M., Fersht, A. R., and Daggett, V. (2000) Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding, *J. Mol. Biol. 296*, 1257-1282.
211. Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T., Imoto, T., Smith, L. J., Dobson, C. M., and Schwalbe, H. (2002) Long-range interactions within a nonnative protein, *Science 295*, 1719-1722.
212. Kortemme, T., Kelly, M. J., Kay, L. E., Forman-Kay, J., and Serrano, L. (2000) Similarities between the spectrin SH3 domain denatured state and its folding transition state, *J. Mol. Biol. 297*, 1217-1229.
213. Tang, Y., Rigotti, D. J., Fairman, R., and Raleigh, D. P. (2004) Peptide models provide evidence for significant structure in the denatured state of a rapidly folding protein: the villin headpiece subdomain, *Biochemistry 43*, 3264-3272.
214. Wellbrock, C., Karasarides, M., and Marais, R. (2004) The RAF proteins take centre stage, *Nat. Rev. Mol. Cell Biol. 5*, 875-885.
215. Scheffler, J. E., Waugh, D. S., Bekesi, E., Kiefer, S. E., LoSardo, J. E., Neri, A., Prinzo, K. M., Tsao, K. L., Wegrzynski, B., Emerson, S. D., and . (1994) Characterization of a 78-residue fragment of c-Raf-1 that comprises a minimal binding domain for the interaction with Ras-GTP, *J. Biol. Chem. 269*, 22340-22346.
216. Emerson, S. D., Waugh, D. S., Scheffler, J. E., Tsao, K. L., Prinzo, K. M., and Fry, D. C. (1994) Chemical shift assignments and folding topology of the Ras-binding domain of human Raf-1 as determined by heteronuclear three-dimensional NMR spectroscopy, *Biochemistry 33*, 7745-7752.
217. Emerson, S. D., Madison, V. S., Palermo, R. E., Waugh, D. S., Scheffler, J. E., Tsao, K. L., Kiefer, S. E., Liu, S. P., and Fry, D. C. (1995) Solution structure of the Ras-binding domain of c-Raf-1 and identification of its Ras interaction surface, *Biochemistry 34*, 6911-6918.
218. Nassar, N., Horn, G., Herrmann, C., Scherer, A., McCormick, F., and Wittinghofer, A. (1995) The 2.2 A crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with Rap1A and a GTP analogue, *Nature 375*, 554-560.
219. Nassar, N., Horn, G., Herrmann, C., Block, C., Janknecht, R., and Wittinghofer, A. (1996) Ras/Rap effector specificity determined by charge reversal, *Nat. Struct. Biol. 3*, 723-729.
220. Terada, T., Ito, Y., Shirouzu, M., Tateno, M., Hashimoto, K., Kigawa, T., Ebisuzaki, T., Takio, K., Shibata, T., Yokoyama, S., Smith, B. O., Laue, E. D., and Cooper, J. A. (1999) Nuclear magnetic resonance and molecular dynamics studies on the interactions of the Ras-binding domain of Raf-1 with wild-type and mutant Ras proteins, *J. Mol. Biol. 286*, 219-232.
221. Herrmann, C., Martin, G. A., and Wittinghofer, A. (1995) Quantitative analysis of the complex between p21ras and the Ras-binding domain of the human Raf-1 protein kinase, *J. Biol. Chem. 270*, 2901-2905.
222. Block, C., Janknecht, R., Herrmann, C., Nassar, N., and Wittinghofer, A. (1996) Quantitative structure-activity analysis correlating Ras/Raf interaction in vitro to Raf activation in vivo, *Nat. Struct. Biol. 3*, 244-251.

223. Fridman, M., Maruta, H., Gonez, J., Walker, F., Treutlein, H., Zeng, J., and Burgess, A. (2000) Point mutants of c-raf-1 RBD with elevated binding to v-Ha-Ras, *J. Biol. Chem. 275*, 30363-30371.
224. Fridman, M., Walker, F., Catimel, B., Domagala, T., Nice, E., and Burgess, A. (2000) c-Raf-1 RBD associates with a subset of active v-H-Ras, *Biochemistry 39*, 15603-15611.
225. Manor, D. (2000) Measurement of GTPase.effector affinities, *Methods Enzymol. 325:139-49.*, 139-149.
226. Kiel, C., Wohlgemuth, S., Rousseau, F., Schymkowitz, J., Ferkinghoff-Borg, J., Wittinghofer, F., and Serrano, L. (2005) Recognizing and defining true Ras binding domains II: in silico prediction based on homology modelling and energy calculations, *J. Mol. Biol. 348*, 759-775.
227. Wohlgemuth, S., Kiel, C., Kramer, A., Serrano, L., Wittinghofer, F., and Herrmann, C. (2005) Recognizing and defining true Ras binding domains I: biochemical analysis, *J. Mol. Biol. 348*, 741-758.
228. Barnard, D., Sun, H., Baker, L., and Marshall, M. S. (1998) In vitro inhibition of Ras-Raf association by short peptides, *Biochem. Biophys. Res. Commun. 247*, 176-180.
229. Zeng, J., Nheu, T., Zorzet, A., Catimel, B., Nice, E., Maruta, H., Burgess, A. W., and Treutlein, H. R. (2001) Design of inhibitors of Ras--Raf interaction using a computational combinatorial algorithm, *Protein Eng 14*, 39-45.
230. Strumberg, D. and Seeber, S. (2005) Raf kinase inhibitors in oncology, *Onkologie. 28*, 101-107.
231. Maxwell, K. L., Wildes, D., Zarrine-Afsar, A., Los Rios, M. A., Brown, A. G., Friel, C. T., Hedberg, L., Horng, J. C., Bona, D., Miller, E. J., Vallee-Belisle, A., Main, E. R., Bemporad, F., Qiu, L., Teilum, K., Vu, N. D., Edwards, A. M., Ruczinski, I., Poulsen, F. M., Kragelund, B. B., Michnick, S. W., Chiti, F., Bai, Y., Hagen, S. J., Serrano, L., Oliveberg, M., Raleigh, D. P., Wittung-Stafshede, P., Radford, S. E., Jackson, S. E., Sosnick, T. R., Marqusee, S., Davidson, A. R., and Plaxco, K. W. (2005) Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins, *Protein Sci. 14*, 602-616.
232. Sander, C. and Schneider, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment, *Proteins 9*, 56-68.
233. Dowdy, S. M. and Wearden, S. (19820) *Statistics for research* John Wiley & Sons, Inc., New York, Chischester, Brisbane, Toronto, Singapore.
234. Privalov, P. L. and Gill, S. J. (1988) Stability of protein structure and hydrophobic interaction, *Adv. Protein Chem. 39*, 191-234.
235. Makhatadze, G. I. and Privalov, P. L. (1995) Energetics of protein structure, *Adv. Protein Chem. 47*, 307-425.
236. Privalov, G. P. and Privalov, P. L. (2000) Problems and prospects in microcalorimetry of biological macromolecules, *Methods Enzymol. 323*, 31-62.
237. Eftink, M. R. and Shastry, M. C. (1997) Fluorescence methods for studying kinetics of protein-folding reactions, *Methods Enzymol. 278*, 258-286.
238. Matouschek, A. and Fersht, A. R. (1991) Protein engineering in analysis of protein folding pathways and stability, *Methods Enzymol. 202*, 82-112.
239. Stein, R. L. (1993) Mechanism of enzymatic and nonenzymatic prolyl cis-trans isomerization, *Adv. Protein Chem. 44*, 1-24.
240. Schmid, F. X., Mayr, L. M., Mucke, M., and Schonbrunner, E. R. (1993) Prolyl isomerases: role in protein folding, *Adv. Protein Chem. 44*, 25-66.
241. Zarrine-Afsar, A. and Davidson, A. R. (2004) The analysis of protein folding kinetic data produced in protein engineering experiments, *Methods 34*, 41-50.
242. Sosnick, T. R., Dothager, R. S., and Krantz, B. A. (2004) Differences in the folding transition state of ubiquitin indicated by phi and psi analyses, *Proc. Natl. Acad. Sci. U.S.A. 101*, 17377-17382.
243. Feng, H., Vu, N. D., Zhou, Z., and Bai, Y. (2004) Structural examination of phi-value analysis in protein folding, *Biochemistry 43*, 14325-14331.
244. Settanni, G., Rao, F., and Caflisch, A. (2005) Phi-value analysis by molecular dynamics simulations of reversible folding, *Proc. Natl. Acad. Sci. U.S.A. 102*, 628-633.

245. Raleigh, D. P. and Plaxco, K. W. (2005) The protein folding transition state: what are Phi-values really telling us?, *Protein Pept. Lett. 12*, 117-122.
246. Pelletier, J. N., Campbell-Valois, F. X., and Michnick, S. W. (1998) Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments, *Proc. Natl. Acad. Sci. U.S.A. 95*, 12141-12146.
247. Lockless, S. W. and Ranganathan, R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families, *Science 286*, 295-299.
248. Suel, G. M., Lockless, S. W., Wall, M. A., and Ranganathan, R. (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins, *Nat. Struct. Biol. 10*, 59-69.
249. Kannan, N. and Vishveshwara, S. (1999) Identification of side-chain clusters in protein structures by a graph spectral method, *J. Mol. Biol. 292*, 441-464.
250. Lindorff-Larsen, K., Rogen, P., Paci, E., Vendruscolo, M., and Dobson, C. M. (2005) Protein folding and the organization of the protein topology universe, *Trends Biochem. Sci. 30*, 13-19.
251. Efimov, A. V. (1995) Structural similarity between two-layer alpha/beta and beta-proteins, *J. Mol. Biol. 245*, 402-415.
252. Fernandez, J. M. and Li, H. (2004) Force-clamp spectroscopy monitors the folding trajectory of a single protein, *Science 303*, 1674-1678.
253. Service, R. (2005) Structural biology. Structural genomics, round 2, *Science 307*, 1554-1558.
254. Service, R. (2005) Structural biology. A dearth of new folds, *Science 307*, 1555.
255. Moult, J., Fidelis, K., Tramontano, A., Rost, B., and Hubbard, T. (2005) Critical assessment of methods of protein structure prediction (CASP) - round VI, *Proteins*.
256. Moult, J. (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction, *Curr. Opin. Struct. Biol. 15*, 285-289.
257. Tress, M., Ezkurdia, I., Grana, O., Lopez, G., and Valencia, A. (2005) Assessment of predictions submitted for the CASP6 comparative modelling category, *Proteins*.
258. Wang, G., Jin, Y., and Dunbrack, R. L., Jr. (2005) Assessment of fold recognition predictions in CASP6, *Proteins*.
259. Vincent, J. J., Tai, C. H., Sathyanarayana, B. K., and Lee, B. (2005) Assessment of CASP6 predictions for new and nearly new fold targets, *Proteins*.
260. Bradley, P., Misura, K. M., and Baker, D. (2005) Toward high-resolution de novo structure prediction for small proteins, *Science 309*, 1868-1871.
261. Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., and Baker, D. (2005) Progress in modeling of protein structures and interactions, *Science 310*, 638-642.
262. Bolon, D. N. and Mayo, S. L. (2001) Enzyme-like proteins by computational design, *Proc. Natl. Acad. Sci. U.S.A. 98*, 14274-14279.
263. Looger, L. L., Dwyer, M. A., Smith, J. J., and Hellinga, H. W. (2003) Computational design of receptor and sensor proteins with novel functions, *Nature 423*, 185-190.
264. Benson, D. E., Conrad, D. W., de Lorimier, R. M., Trammell, S. A., and Hellinga, H. W. (2001) Design of bioelectronic interfaces by exploiting hinge-bending motions in proteins, *Science 293*, 1641-1644.
265. Karplus, M. and McCammon, J. A. (2002) Molecular dynamics simulations of biomolecules, *Nat. Struct. Biol. 9*, 646-652.
266. Daggett, V. and Fersht, A. (2003) The present view of the mechanism of protein folding, *Nat. Rev. Mol. Cell Biol. 4*, 497-502.
267. Snow, C. D., Sorin, E. J., Rhee, Y. M., and Pande, V. S. (2005) How well can simulation predict protein folding kinetics and thermodynamics?, *Annu. Rev. Biophys. Biomol. Struct. 34*, 43-69.
268. Li, A. and Daggett, V. (1994) Characterization of the transition state of protein unfolding by use of molecular dynamics: chymotrypsin inhibitor 2, *Proc. Natl. Acad. Sci. U.S.A. 91*, 10430-10434.
269. Bond, C. J., Wong, K. B., Clarke, J., Fersht, A. R., and Daggett, V. (1997) Characterization of residual structure in the thermally denatured state of barnase by simulation and experiment: description of the folding pathway, *Proc. Natl. Acad. Sci. U.S.A. 94*, 13409-13413.

270. Li, A. and Daggett, V. (1998) Molecular dynamics simulation of the unfolding of barnase: characterization of the major intermediate, *J. Mol. Biol. 275*, 677-694.
271. Vendruscolo, M., Paci, E., Dobson, C. M., and Karplus, M. (2001) Three key residues form a critical contact network in a protein folding transition state, *Nature 409*, 641-645.
272. Jemth, P., Day, R., Gianni, S., Khan, F., Allen, M., Daggett, V., and Fersht, A. R. (2005) The structure of the major transition state for folding of an FF domain from experiment and simulation, *J. Mol. Biol. 350*, 363-378.
273. Mayor, U., Guydosh, N. R., Johnson, C. M., Grossmann, J. G., Sato, S., Jas, G. S., Freund, S. M., Alonso, D. O., Daggett, V., and Fersht, A. R. (2003) The complete folding pathway of a protein from nanoseconds to microseconds, *Nature 421*, 863-867.
274. Gianni, S., Guydosh, N. R., Khan, F., Caldas, T. D., Mayor, U., White, G. W., DeMarco, M. L., Daggett, V., and Fersht, A. R. (2003) Unifying features in protein-folding mechanisms, *Proc. Natl. Acad. Sci. U.S.A. 100*, 13286-13291.
275. Day, R. and Daggett, V. (2005) Ensemble versus single-molecule protein unfolding, *Proc. Natl. Acad. Sci. U.S.A. 102*, 13445-13450.
276. Zhuang, X. and Rief, M. (2003) Single-molecule folding, *Curr. Opin. Struct. Biol. 13*, 88-97.
277. Deniz, A. A., Laurence, T. A., Beligere, G. S., Dahan, M., Martin, A. B., Chemla, D. S., Dawson, P. E., Schultz, P. G., and Weiss, S. (2000) Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2, *Proc. Natl. Acad. Sci. U.S.A. 97*, 5179-5184.
278. Garcia-Mira, M. M., Sadqi, M., Fischer, N., Sanchez-Ruiz, J. M., and Munoz, V. (2002) Experimental identification of downhill protein folding, *Science 298*, 2191-2195.
279. Schuler, B., Lipman, E. A., and Eaton, W. A. (2002) Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy, *Nature 419*, 743-747.
280. Lipman, E. A., Schuler, B., Bakajin, O., and Eaton, W. A. (2003) Single-molecule measurement of protein folding kinetics, *Science 301*, 1233-1235.
281. Rhoades, E., Gussakovsky, E., and Haran, G. (2003) Watching proteins fold one molecule at a time, *Proc. Natl. Acad. Sci. U.S.A. 100*, 3197-3202.
282. Rief, M., Gautel, M., Oesterhelt, F., Fernandez, J. M., and Gaub, H. E. (1997) Reversible unfolding of individual titin immunoglobulin domains by AFM, *Science 276*, 1109-1112.
283. Cecconi, C., Shank, E. A., Bustamante, C., and Marqusee, S. (2005) Direct observation of the three-state folding of a single protein molecule, *Science 309*, 2057-2060.
284. Fisher, T. E., Marszalek, P. E., and Fernandez, J. M. (2000) Stretching single molecules into novel conformations using the atomic force microscope, *Nat. Struct. Biol. 7*, 719-724.
285. Allison, D. P., Hinterdorfer, P., and Han, W. (2002) Biomolecular force measurements and the atomic force microscope, *Curr. Opin. Biotechnol. 13*, 47-51.
286. Gillespie, B. and Plaxco, K. W. (2004) Using protein folding rates to test protein folding theories, *Annu. Rev. Biochem. 73*, 837-859.
287. Cheung, M. S., Garcia, A. E., and Onuchic, J. N. (2002) Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse, *Proc. Natl. Acad. Sci. U.S.A. 99*, 685-690.
288. Luby-Phelps, K. (2000) Cytoarchitecture and physical properties of cytoplasm: volume, viscosity, diffusion, intracellular surface area, *Int. Rev. Cytol. 192*, 189-221.
289. Ellis, R. J. (2001) Macromolecular crowding: an important but neglected aspect of the intracellular environment, *Curr. Opin. Struct. Biol. 11*, 114-119.
290. Ellis, R. J. (2001) Macromolecular crowding: obvious but underappreciated, *Trends Biochem. Sci. 26*, 597-604.
291. Hall, D. and Minton, A. P. (2003) Macromolecular crowding: qualitative and semiquantitative successes, quantitative challenges, *Biochim. Biophys. Acta 1649*, 127-139.
292. van den, B. B., Ellis, R. J., and Dobson, C. M. (1999) Effects of macromolecular crowding on protein folding and aggregation, *EMBO J. 18*, 6927-6933.

293. van den, B. B., Wain, R., Dobson, C. M., and Ellis, R. J. (2000) Macromolecular crowding perturbs protein refolding kinetics: implications for folding inside the cell, *EMBO J. 19*, 3870-3875.
294. Cheung, M. S., Klimov, D., and Thirumalai, D. (2005) Molecular crowding enhances native state stability and refolding rates of globular proteins, *Proc. Natl. Acad. Sci. U.S.A. 102*, 4753-4758.
295. Goodsell, D. S. and Olson, A. J. (2000) Structural symmetry and protein function, *Annu. Rev. Biophys. Biomol. Struct. 29*, 105-153.
296. Bray, D. and Duke, T. (2004) Conformational spread: the propagation of allosteric states in large multiprotein complexes, *Annu. Rev. Biophys. Biomol. Struct. 33*, 53-73.
297. Kern, D. and Zuiderweg, E. R. (2003) The role of dynamics in allosteric regulation, *Curr. Opin. Struct. Biol. 13*, 748-757.
298. Gunasekaran, K., Ma, B., and Nussinov, R. (2004) Is allostery an intrinsic property of all dynamic proteins?, *Proteins 57*, 433-443.
299. Wiesmann, C., Barr, K. J., Kung, J., Zhu, J., Erlanson, D. A., Shen, W., Fahr, B. J., Zhong, M., Taylor, L., Randal, M., McDowell, R. S., and Hansen, S. K. (2004) Allosteric inhibition of protein tyrosine phosphatase 1B, *Nat. Struct. Mol. Biol. 11*, 730-737.
300. Peterson, J. R., Bickford, L. C., Morgan, D., Kim, A. S., Ouerfelli, O., Kirschner, M. W., and Rosen, M. K. (2004) Chemical inhibition of N-WASP by stabilization of a native autoinhibited conformation, *Nat. Struct. Mol. Biol. 11*, 747-755.
301. Kim, A. S., Kakalis, L. T., Abdul-Manan, N., Liu, G. A., and Rosen, M. K. (2000) Autoinhibition and activation mechanisms of the Wiskott-Aldrich syndrome protein, *Nature 404*, 151-158.
302. Schroder, M. and Kaufman, R. J. (2005) The mammalian unfolded protein response, *Annu. Rev. Biochem. 74*, 739-789.
303. Young, J. C., Agashe, V. R., Siegers, K., and Hartl, F. U. (2004) Pathways of chaperone-mediated protein folding in the cytosol, *Nat. Rev. Mol. Cell Biol. 5*, 781-791.
304. Kerner, M. J., Naylor, D. J., Ishihama, Y., Maier, T., Chang, H. C., Stines, A. P., Georgopoulos, C., Frishman, D., Hayer-Hartl, M., Mann, M., and Hartl, F. U. (2005) Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli, *Cell 122*, 209-220.
305. Stefani, M. and Dobson, C. M. (2003) Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution, *J. Mol. Med. 81*, 678-699.
306. Dobson, C. M. (2003) Protein folding and misfolding, *Nature 426*, 884-890.
307. Guijarro, J. I., Sunde, M., Jones, J. A., Campbell, I. D., and Dobson, C. M. (1998) Amyloid fibril formation by an SH3 domain, *Proc. Natl. Acad. Sci. U.S.A. 95*, 4224-4228.
308. Litvinovich, S. V., Brew, S. A., Aota, S., Akiyama, S. K., Haudenschild, C., and Ingham, K. C. (1998) Formation of amyloid-like fibrils by self-association of a partially unfolded fibronectin type III module, *J. Mol. Biol. 280*, 245-258.
309. Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M., Ramponi, G., and Dobson, C. M. (1999) Designing conditions for in vitro formation of amyloid protofilaments and fibrils, *Proc. Natl. Acad. Sci. U.S.A. 96*, 3590-3594.
310. Chiti, F., Calamai, M., Taddei, N., Stefani, M., Ramponi, G., and Dobson, C. M. (2002) Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases, *Proc. Natl. Acad. Sci. U.S.A. 99 Suppl 4*, 16419-16426.
311. Chiti, F., Taddei, N., Baroni, F., Capanni, C., Stefani, M., Ramponi, G., and Dobson, C. M. (2002) Kinetic partitioning of protein folding and aggregation, *Nat. Struct. Biol. 9*, 137-143.
312. Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol. 157*, 105-132.
313. Koehl, P. and Levitt, M. (1999) Structure-based conformational preferences of amino acids, *Proc. Natl. Acad. Sci. U.S.A. 96*, 12524-12529.
314. Richards, F. M. (1958) On the enzymatic activity of subtilisin-modified ribonuclease, *Proc. Natl. Acad. Sci. U.S.A. 44*, 162-166.
315. Freifelder, D. (1982) Characterization of macromolecules, in *Physical Biochemistry* 2nd ed., W.H.Freeman and Company, San Francisco.

Annexe : informations supplémentaires

Tableau AI. Aperçu des propriétés chimiques des acides aminés naturels

Acide aminé	Classes[a]	pK_R	Hydropathies[b]	Propension hélice-α[c]	Propension Brin-β[c]	Volume[d]
Histidine (H)	aro	6,04	-3,2	-0,11	1,34	76,83
Tryptophane (W)	aro		-0,9	0,21	-0,14	121.41
Tyrosine (Y)	aro	10,46	-1,3	0,05	-0,49	86.33
Phenylalanine (F)	aro		2,8	-0,01	-0,67	93,43
Isoleucine (I)	ali		4,5	-0,26	-0,77	81,98
Valine (V)	ali		4,2	-0,06	-0,7	63,75
Leucine (L)	ali		3,8	-0,38	0,15	81,98
Méthionine (M)	ali		1,9	-0,09	-0,71	82,41
Cystéine (C)	ali	8,37	2,5	0,57	-0,63	38,23
Alanine (A)	ali		1,8	-0,04	-0,12	25,95
Glycine (G)	T		-0,4	1,24	0,76	0
Proline (P)	T		-1,6	3.11	0	54,69
Sérine (S)	P		-0,8	0,15	1,45	31,82
Thréonine (T)	P		-0,7	0,39	-0,7	51,39
Asparagine (N)	P		-3,5	0,25	1,05	54
Glutamine (Q)	P		-3,5	-0,02	1,67	72,23
Aspartate (D)	A	3,90	-3,5	0,27	1,12	49,42
Glutamate (E)	A	4,07	-3,5	-0,33	0,91	67,65
Lysine (K)	B	10,54	-3,9	-0,18	0,29	93,92
Arginine (R)	B	12,48	-4,5	-0,3	0,34	106,56

[a] Aro: aromatique; ali: aliphatique; T: acides aminés aux propriétés conformationnelles distinctes; S: polaire; A: acide; B: base

[b] Hydrophobicité relative selon Kyte et Doolittle (*312*).

[c] Propension pour les hélices-α et les brins-β des divers acides aminés selon Koehl et Levitt (*313*). Les données pour la proline m'ont été communiquées personnellement par P. Koehl.

[d] Volume de la chaîne latérale en $Å^3$ selon Richards (*314*).

Tableau AII. Code génétique et alphabet pour l'encodage des bases dégénérées

	T	C	A	G
T	TTT Phe (F) TTC TTA Leu (L) TTG	TCT Ser (S) TCC TCA TCG	TAT Tyr (Y) TAC TAA arrêt TAG arrêt	TGT Cys (C) TGC TGA arrêt TGG Trp
C	CTT Leu (L) CTC CTA CTG	CCT Pro (P) CCC CCA CCG	CAT His (H) CAC CAA Gln (Q) CAG	CGT Arg (R) CGC CGA CGG
A	ATT Ile (I) ATC ATA ATG Met (M)	ACT Thr (T) ACC ACA ACG	AAT Asn (N) AAC AAA Lys (K) AAG	AGT Ser (S) AGC AGA Arg (R) AGG
G	GTT Val (V) GTC GTA GTG	GCT Ala (A) GCC GCA GCG	GAT Asp (D) GAC GAA Glu (E) GAG	GGT Gly (G) GGC GGA GGG

B= C, G ou T; **D**= A, G ou T; **H**= A, C ou T; **K**= G ou T; **M**= A ou C; **N**= A, C, G ou T; **R**= A ou G; **S**= C ou G; **V**= A, C ou G; **W**= A ou T; **Y**= C ou T.