Université de Montréal

Les feuillets beta dans les protéines. Annotation, comparaison et construction.

par
Marc Parisien

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en Informatique

Août, 2005

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

**Les feuillets beta dans les protéines.
Annotation, comparaison et construction.**

présenté par:

Marc Parisien

a été évalué par un jury composé des personnes suivantes:

Yoshua Bengio
président-rapporteur

François Major
directeur de recherche

Andreea Schmitzer
membre du jury

Mémoire accepté le 20 octobre 2005

RÉSUMÉ

La croissance exponentielle des séquences disponibles, tant pour l'ADN, l'ARN et les protéines, initié par le projet public de séquençage du génome humain, laisse le champ de la détermination tri-dimensionelle des macro-molécules parraître bien pauvre avec ses avancées presque linéaires, mais combien importante puisque la structure du composé lui confère sa fonction. Il devient alors important de considérer toute approche *in sillico* dans le chemin entre la séquence et la structure tertiaire dans le but de comprendre le repliement et accélerer le calcul des modèles 3-D. Le présent mémoire porte essentiellement sur les feuillets beta des protéines à savoir l'annotation, la comparaison et la construction. Il se compose de quatre manuscrits dont certains sont déjà publiés et d'autres sont en preparation de publication.

Un espace de recherche conformationelle est tout d'abord exposé. Les éléments de structures secondaires des protéines, soit les hélices alpha et les brins beta, sont approximés par des composantes à géométrie fixe. Ces composantes sont alors coordonnées dans l'espace à l'aide de relations spatiales encodées dans des matrices de transformations homogènes extraites des structures déjà connues. Nous démontrons alors, par la technique d'échantillonage nommé Jack Knife, que cet espace de recherche contient les structures 3-D des protéines connues, et qu'il est alors possible d'élaborer des modèles plausibles pour des protéines de structures inconnues.

Ensuite, nous introduisons le concept de graphe topologique pour les feuillets beta des protéines, dans lequel les noeuds sont les résidues participant au feuillet, et les arrêtes encodent soient les liens peptidiques, les partenaires beta ou bien les ponts hydrogènes entre les résidues connectés. Une fois le graphe defini on peut alors construire une base de données de structure 3-D de feuillets avec leurs graphes topologiques associés. Un algorithme d'isomorphisme de sous-graphe, adapté à la nature particulière des graphes de topologie rendant possible la comparaison de graphes de tailles appréciables, permet la recherche de motifs topologiques et d'en récupérer les structures 3-D correspondantes.

Ces graphes sont ensuites utilisés pour identifier des fragments de feuillets beta, assemblés dans l'espace afin de générer des modèles 3-D, à partir de la description topologique, ou 2-D, des feuillets. La technique d'échantillonage Jack Knife nous indique qu'il est alors possible de reconstruire les feuillets beta existants, et ce avec une grande précision, à l'aide de fragments tirés d'autres feuillets.

Finalement, l'identification de familles de protéines à partir des feuillets β révèle la faiblesse des outils actuels pour l'annotation des feuillets. Nous avons donc développé une approche basée sur de récents travaux démontrant l'existance de forces stabilisatrices autres que les ponts hydrogènes. Les annotations de feuillets, suite à l'application de notre méthode, diffèrent substantiellement de celles déjà publiées et en change la conception par l'étendue des angles ϕ-ψ et les patrons de ponts hydrogènes observés.

mots-clés : protéine, feuillet beta, graphe, modélisation, énergie.

ABSTRACT

The exponential increase of available sequences, as much for DNA, RNA and proteins, initiated by the public human genome sequencing project, has left the field of macro-molecular three-dimensional structure determination far behind with it's linear advances, but how important since the structure of the compound gives it's function. Then, it becomes important to consider any *in sillico* approaches in the path from sequence to structure in hope to understand the folding processes and accelerate the computations of 3-D models. The present work is essentially about beta-sheets in proteins, that is, the annotation, the comparison and the construction of such beta-sheets. The work is composed of four manuscripts; some of which are already published and others in preparation of.

A conformational search space is first exposed. The secondary structure elements of proteins, let be the alpha-helices and the beta-sheets, are approximated by fixed geometry components. These components are then coordinated in space with the help of spatial relationships, encoded in homogeneous transformation matrices, and extracted from already known structures. We show that, via the Jack-Knife sampling technique, the search space addresses all 3-D structures of known proteins, and is thus possible to elaborate convincing models for proteins of unknown structures.

Next, we introduce the concept of topological graphs for protein beta-sheets, in which the vertexes are the residues part of the beta-sheet, and the edges encode either the peptide bond, the beta-sheet partnership or the hydrogen bonds between the connected residues. Once the graph is defined, we can then build a database of 3-D structures of beta-sheets with their associated topological graphs. A sub-graph isomorphism algorithm, adapted to the particular nature of these topological graphs and now making possible the comparison of graphs of appreciable sizes, paves the way for topological motif searches which returns the corresponding beta-sheet 3-D structures.

These graphs are then used for the identification of beta-sheet fragments, assembled in space in order to generated 3-D models, from the topological description, or 2-D, of the beta-sheet. The Jack-Knife sampling technique shows us that it is possible to rebuild existing beta-sheets, with high precision, using fragments from other beta-sheets.

Finally, the identification of protein families from their beta-sheets reveals the weaknesses of actual tools for the annotation of protein beta-sheets. We have thus developed an approach based on recent findings showing the existence of other stabilizing forces than the hydrogen bonds. The annotations of beta-sheets, following our method, differs largely from those already published, and consequently changes the conception of beta-sheets by the broad span of ϕ-ψ angles and hydrogen bonding motifs observed.

keywords: protein, beta-sheet, graph, modeling, energy.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES APPENDICES

LISTE DES SIGLES

PDB	Protein Data Bank
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
3-D	Three Dimensional
ADN	Acide désoxyribonucléique
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
CSP	Constraint Satisfaction Problem
DSSP	Dictionary of Secondary Structure of Proteins
HTM	Homogenous Transformation Matrix
PAM	Point Accepted Mutation
SCOP	Structural Classification of Proteins
CPU	Central Processing Unit
α	Alpha (de l'alphabet Grec)
β	Beta (de l'alphabet Grec)

A tous ceux et celles qui liront ce mémoire...

REMERCIEMENTS

J'aimerais remercier mon directeur de recherche, le docteur François Major, pour m'avoir plongé dans le monde de la structure des protéines alors que son laboratoire verse dans la structure des ARNs, et ce en me laissant carte blanche quant à mon parcours sub-aquatique. Nos flotteurs sont jamais assez gros lorque vient le temps d'aller au fond des choses...

J'aimerais aussi remercier un collègue, Sébastien Lemieux, pour un jour m'avoir suggéré de générer une image en Postscript directement ! Me voilà donc le nez dans un livre pour apprendre à mieux converser avec mon imprimante. J'ai alors compris que l'Outil donne à l'Homme un net avantage, surtout avec un manuel d'instructions.

À Alain Normandeau pour les beaux jours chez Cerveau et d'avoir mis les bogues informatiques dans mon "Quality World[1]", sans lesquels un programmeur ne peut évoluer par le cycle analyser-comprendre-corriger.

À Laurent Bréhélin pour son enthousiasme inébranlable, même dans les moments les plus brumeux.

Et à ma dulcinée, Lucie Marcil, pour m'avoir donné des ailes et le goût d'écrire ce mémoire.

[1] W Glasser. Choice theory : a new psychology of personal freedom. HarperCollins Publishers Inc. New York. 1998.

CHAPITRE 1

INTRODUCTION

1.1 Les protéines

Les protéines forment l'une des trois classes majeures de polymères au sein des cellules, avec l'acide ribonucléique (ARN) et l'acide désoxyribonucléique (ADN). Une protéine est une chaîne d'acides aminés. Les cellules du corps humain font appels à plus de vingt types d'acides aminés différents dont certains, dits essentiels, sont non-synthétisables par les cellules et doivent, par conséquent, provenir de notre alimentation[1].

Chaque acide aminé se compose d'atomes reliés entre eux par des liens covalents selon des patrons établis. La Figure 1.1 montre les atomes d'un acide aminé. Ces atomes se partitionnent en deux groupes ; ceux de la chaîne principale et ceux de la chaîne latérale. Les atomes de la chaîne principale sont les mêmes pour tous les acides aminés. Les atomes de la chaîne latérale sont propres à chaque type d'acide aminé. Les figures 1.2 à 1.6 montrent les vingt acides aminés les plus courants.

La chaîne latérale peut s'ammarer à la chaîne principale à deux endroits distincts mais toujours sur le même atome, le carbone α. On parle alors d'énantiomères, les uns, de forme L, sont images miroir des autres, de forme D. Le Règne Vivant métabolise la forme stéréochimique L des acides aminés. La forme D peut même être nocive chez l'humain [1]. La Figure 1.7 montre comment identifier les acides aminés de forme L par le truc mnémonique "CORN".

Les protéines sont construites par l'ajout d'acides aminés les uns à la suite des autres. La nature de l'acide aminé à rajouter est déterminée par le code génétique. Comme il existe plus de 20 acides aminés différents, le code génétique, avec son alphabet de quatre symboles ou bases {A,C,G,T}, nécessite un encodage des acides

[1] Le Coke diète, bien qu'il contienne de la phénylalanine, un acide aminé essentiel, ne doit pas servir de source unique d'alimentation.

aminés sur une suite de trois bases consécutives (puisque $4^2 = 16$ seulement). À cette suite de trois bases, ou codon, correspond un anti-codon (par complémentarité des bases A/T et G/C). Une molécule clé, l'ARN de transport, fait correspondre un anti-codon spécifique avec un type d'acide aminé. Le ribosome coordonne alors le décodage de l'ADN en sa protéine correspondante par l'appel successif d'ARN de transport. Les acides aminés ainsi transportés sont mis bout à bout pour former la protéine. La liaison de deux acides aminés par les atomes de la chaîne principale crée le lien peptidique. La Figure 1.8 met en évidence l'emplacement du lien peptidique reliant les acides aminés.

Il est intéressant de mentionner ici que les détracteurs de la théorie de Darwin concernant l'évolution ont pour argument que le problème de l'origine des protéines, avec l'usage exclusif de la forme L des acides aminés et l'enchaînement entre eux par le lien peptidique seulement, ne peut être le fruit du hasard[2].

Lorsqu'on suit la chaîne principale d'une protéine en sautant d'un atome à l'autre par les liens covalents on rencontre, tour à tour, les atomes suivants ; l'azote N, le carbone C_α puis le carbone C. Ce dernier est lié au prochain azote via le lien peptidique. On a donc ainsi trois liens covalents principaux par acide aminé, autour desquels on peut mesurer les angles de torsions ; ϕ, ψ et ω. La Figure 1.9 montre ces angles de torsion dans la chaîne principale.

Le repliement de la chaîne peptidique sur elle-même entraîne la formation de ponts hydrogènes entre les groupements amides N-H et carboxyliques C=O de la chaîne principale. Les helices α, prédites par Pauling, Corey et Branson [2], sont une suite de ponts hydrogènes entre les groupements C=O du résidue i et N-H du résidue i+4 (intéractions locales dans l'espace et en séquence). Les feuillets β, prédits eux aussi par Pauling et Corey [3], sont composés de trois motifs répétés de ponts hydrogènes entre deux chaînes peptidiques adjacentes (intéractions spatiales locales par des parties distantes de la séquence). Les hélices α et les feuillets β definissent les éléments de structures secondaires des protéines [4]. Le détail des

[2]par exemple, voir http ://www.darwinismrefuted.com/molecular_biology_03.html

ponts hydrogènes dans une hélice α est montré dans la Figure 1.10. Le détail des ponts hydrogènes dans un feuillet β est montré dans la Figure 1.13.

La structure tertiaire, ou tri-dimensionelle, de la protéine peut être vue comme l'assemblage dans l'espace des éléments de structures secondaires, i.e. des hélices α et des brins β en feuillets β. La représentation des brins β, de même que les hélices α, en ruban flèché est due à Richardson [5]. Cette représentation en ruban met bien en valeur les feuillets β dans les méandres de la chaîne peptidique. La structure tertiaire de la protéine ubiquitine est montrée dans la Figure 1.11, en représentation de Richardson (les hélices sont en rubans hélicoïdaux, les brins en flèches épaises et les boucles en tubes cylindriques).

Il arrive parfois que les protéines s'assemblent pour former un complexe actif. Par exemple, l'insuline chez l'humain forme un complexe hexamérique (voir Figure 1.12). Les complexes peuvent aussi être formés de protéines différentes. On parle alors de structure quaternaire.

Anfinsen démontra que la suite d'acides aminés composant la protéine, i.e. la séquence primaire (1-D), est suffisante à elle seule pour en encoder la structure tridimensionnelle (3-D) [6]. Et que la structure finale est celle dont l'energie potentielle est minimale (hypothèse thermodynamique) [7]. En 1972, il recoit le prix Nobel pour ses travaux sur les principes gouvernants le repliement des protéines.

Le problème du repliement des protéines ("Protein Folding Problem" [8]), c'est-à-dire la prédiction de la structure tri-dimensionelle des protéines à partir de la séquence, est, à ce jour, encore ouvert et est le sujet de maintes recherches. L'intérêt de ce problème est principalement dû aux applications médicales, à savoir la fabrication simplifiée et précise (non-toxique) de médicaments, et la compréhension du fonctionnement de la cellule.

Ce problème est équivalent à assigner les angles de torsion ϕ et ψ pour chaque acide aminé formant la protéine, puisque les chaînes latérales vont s'accomoder de la position de la chaîne principale. Même en discrétisant ces angles à 7 valeurs possibles [9] l'espace de recherche reste quand même très grand. En assumant 1 millions d'opérations par seconde, un an de calcul serait necessaire pour explorer

tout l'espace de recherche d'une chaîne peptidique de longueur 16 seulement ! Il est à noter que la longueur moyenne des protéines chez *Saccharomyces Cerevisiae*, un eukaryote uni-cellulaire mieux connu sous le nom de "levure du boulanger", est de plus de 450 acides aminés [10].

1.2 Les feuillets β

Les feuillets β se forment lorsque deux bouts de chaînes peptidiques passe l'une à côté de l'autre et tissent alors un réseau de ponts hydrogènes pour stabiliser la conformation. Rien n'empêche par la suite à d'autres bouts de venir s'y coller pour étendre la portée du feuillet. Ainsi, le feuillet se compose de brins β.

Pauling et ses collègues ont déterminés les trois patrons de base de ponts hydrogènes entre paires de brins dans les feuillets β [3,11]. On parlera alors de patrons ou motifs canoniques. Ces motifs sont dépeints en détails dans la Figure 1.13, où l'on y voit les atomes des chaînes principales en intéractions. La Figure 1.14 nous montre comment ces motifs sont utilisés entre paires de brins β, soient parallèles ou anti-parallèles. La Figure 1.15 montre comment les chaînes latérales sont disposées dans un feuillet β ; les partenaires β pointent dans le même sens.

On retrouve des feuillets β dans 80% des protéines ; il existe alors des protéines sans feuillet puis d'autres presqu'exclusivement sous forme de feuillets. On retrouve les feuillets sous plusieurs formes, dont la β-hélice (Figure 1.16), la β-sandwich (Figure 1.17), le β-propulseur (Figure 1.18) et le β-baril (Figure 1.19), où tous se retrouvent dessinés dans la représentation de Richardson.

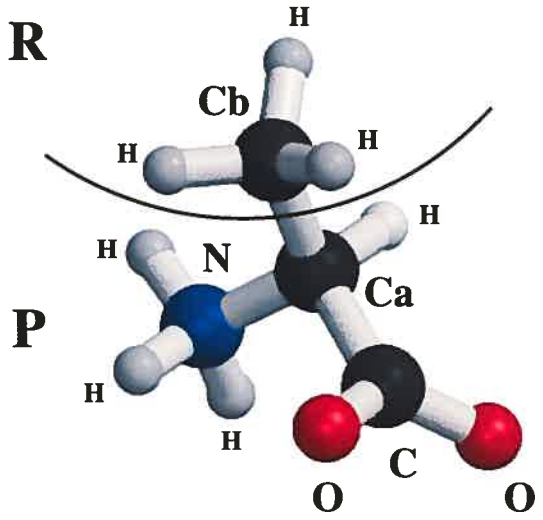
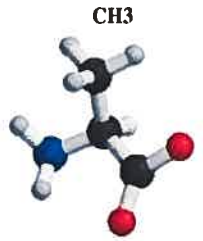
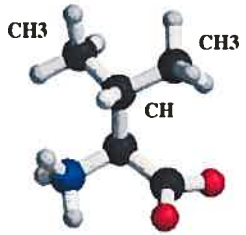


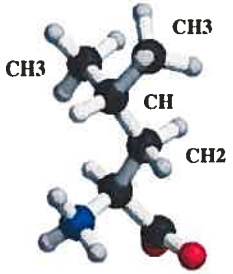
Figure 1.1 – Les atomes d'un acide aminé. La région **P** comprends les atomes de la chaîne principale alors que la région **R** comprends les atomes de la chaîne latérale. Chaque atome est représenté par une sphère de couleur ; l'azone en bleu, l'oxygène en rouge, le carbone en noir et l'hydrogène en gris. Les sphères sont proportionelles aux rayons de van der Waals propres à chaque type d'atome. Les atomes sont liés entre eux par des liens covalents représentés par des cylindres. Les acides aminés se distinguent entre eux par les atomes qui figurent dans la région **R**. **Ca** est le carbone alpha, ou C_α, tandis que **Cb** est le carbone beta, ou C_β. L'atome C_α est celui auquel vient s'attacher la chaîne latérale. L'azote est chargé positivement par la présence d'un atome d'hydrogène en surplus. Les deux atomes d'oxygènes se partagent, en résonance, un double-lien avec l'atome de carbone et une charge négative (par l'absence de lien avec un atome d'hydrogène).



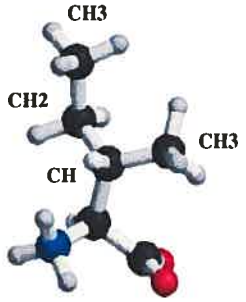
(a) Alanine (ALA :A)



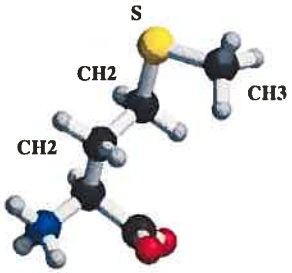
(b) Valine (VAL :V)



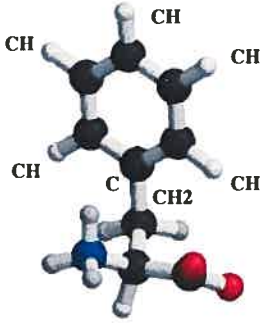
(c) Leucine (LEU :L)



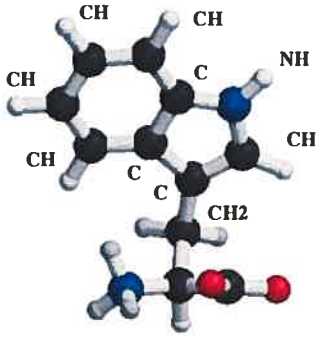
(d) Isoleucine (ILE :I)



(e) Méthionine (MET :M)

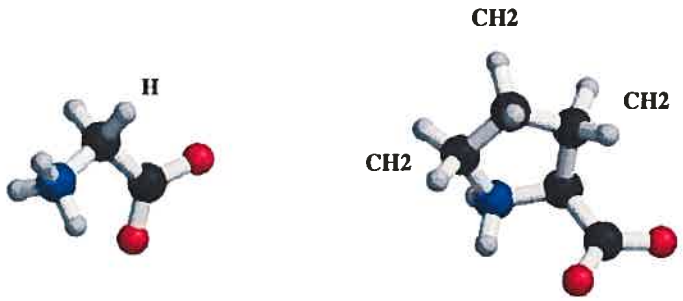


(f) Phénylalanine (PHE :F)



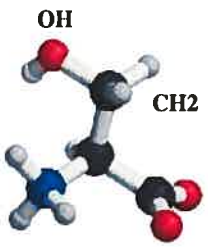
(g) Tryptophane (TRP :W)

FIG. 1.2 – Les acides aminés hydrophobiques. Ces acides aminés ont tendance a fuir l'eau, par l'absence de groupement polaire en chaîne latérale, et se retrouvent donc a l'intérieur des protéines globulaires. Les atomes sont représentés par des sphères de couleurs : l'azote en bleu, l'oxygène en rouge, le souffre en jaune, le carbone en noir et l'hydrogène en gris. Les liens covalents reliants les atomes sont des cylindres gris. Les positions des atomes d'hydrogènes ne sont pas optimales puisqu'ils ont étés placés par une procédure géométrique. Les atomes de la chaîne principale sont en bas des images tandis que les atomes de la chaîne latérale occupent le haut, et leurs groupes fonctionels sont annotés.

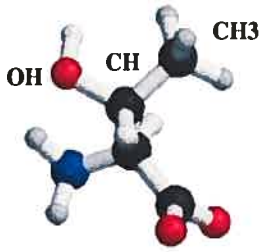


(a) Glycine (GLY :G) (b) Proline (PRO :P)

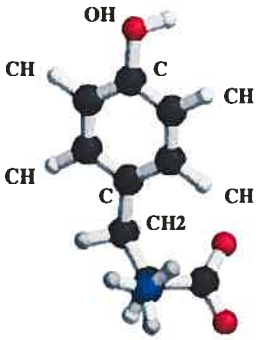
FIG. 1.3 – Les acides aminés spéciaux. La glycine n'a qu'un proton en lieu de chaîne latérale, ce qui lui confère une plus grande liberté conformationelle (ϕ,ψ). La proline est le seul acide aminé dont la chaîne latérale forme un cycle avec la chaîne principale. Contrairement à la glycine, la proline adopte des angles de torsion fixes, et de ce fait est connue pour briser les helices alphas.



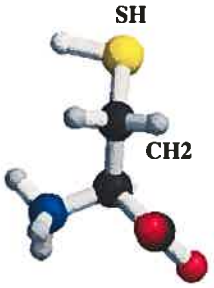
(a) Sérine (SER :S)



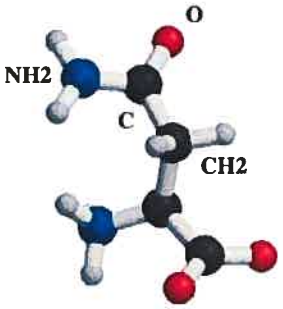
(b) Thréonine (THR :T)



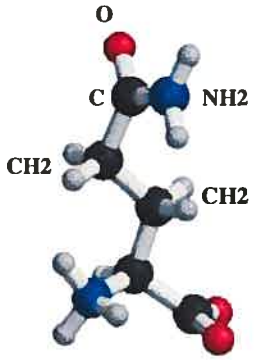
(c) Tyrosine (TYR :Y)



(d) Cystéine (CYS :C)

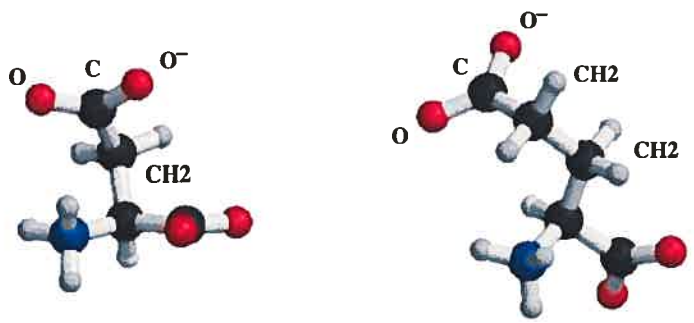


(e) Asparagine (ASN :N)



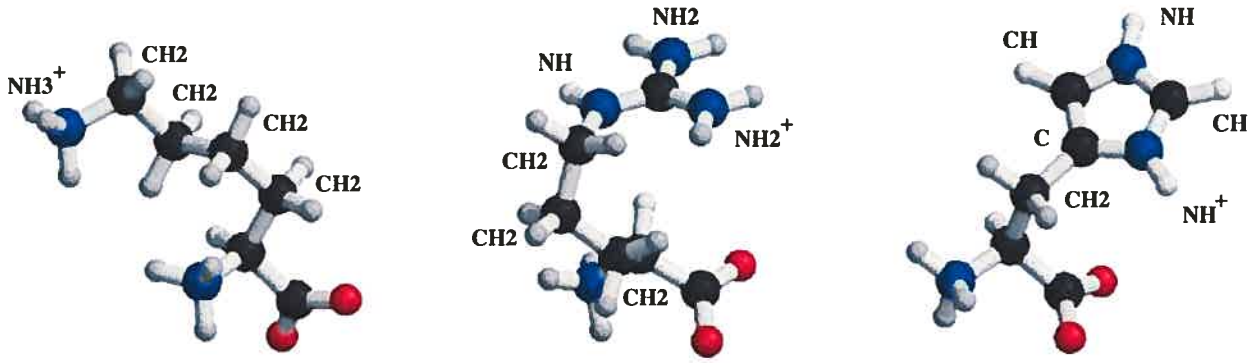
(f) Glutamine (GLN :Q)

FIG. 1.4 – Les acides aminés polaires. La présence de groupement polaires, soit CO, OH, SH ou NH2, en chaîne latérale fait de ces résidus des hydrophiles, et se retrouvent souvent en surface des protéines globulaires.



(a) Aspartate (ASP :D) (b) Glutamate (GLU :E)

FIG. 1.5 – Les acides aminés acides (accepteurs de protons). Ces deux acides aminés sont négativement chargé à un pH de 7.5. La charge négative est en résonance entre les deux groupement CO.



(a) Lysine (LYS :K) (b) Arginine (ARG :R) (c) Histidine (HIS :H)

FIG. 1.6 – Les acides aminés basiques (donneurs de protons). Dans l'arginine, la charge positive est en résonance entre les deux groupements NH2. Dans l'histidine, dont on voit la forme protonée, la charge positive est en résonance entre les deux groupements NH.

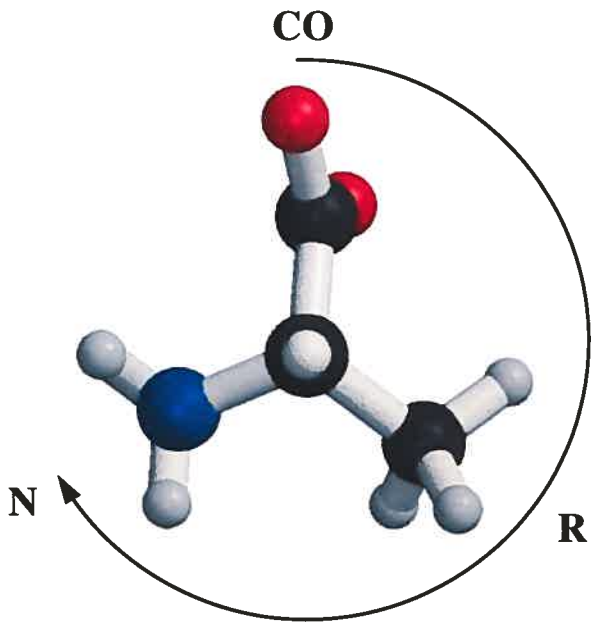
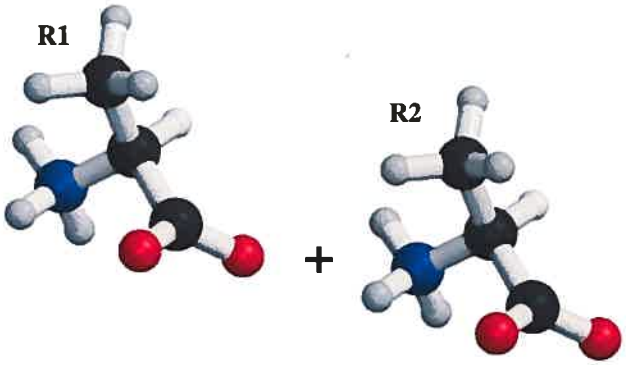
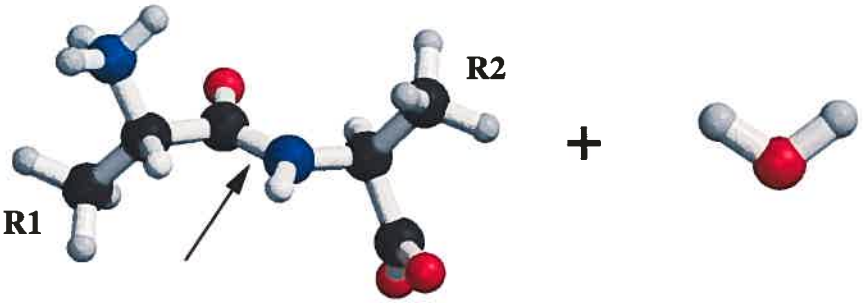


Figure 1.7 – La règle de CORN. Lorsqu'on aligne le carbone alpha derrière son atome d'hydrogène on obtient, en parcourant dans le sens des aiguilles d'une montre, le groupement carboxylique **CO**, la chaîne latérale **R** et le groupement aminé **N**, d'où la règle de CORN. Cette règle s'applique aux acides aminés de la forme stéréo-chimique L.

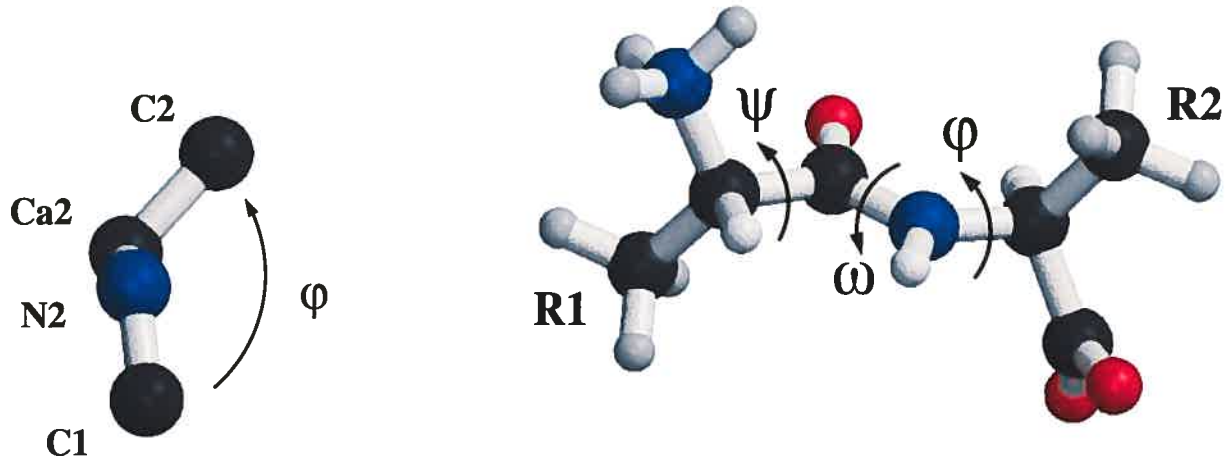


(a)



(b)

Figure 1.8 – Le lien peptidique. Deux acides aminés sont reliés entre eux par le lien peptidique. **a)** Le lien peptidique résulte de l'attaque d'un atome d'oxygène du premier acide aminé sur deux atomes d'hydrogènes du second acide aminé. **b)** Il y a alors formation d'un di-peptide et d'une molécule d'eau (H_2O). Le lien peptidique résultant est indiqué par la flèche. Les chaînes latérales **R1** et **R2** sont identifiées pour suivre la construction.



(a) (b)

Figure 1.9 – Les angles de torsion de la chaîne principale. **a)** Un angle de torsion, ou dihédral, se mesure à partir de quatre atomes ; on aligne l'un sur l'autre les deux atomes centraux puis on mesure l'angle formé entre les deux atomes terminaux. Par exemple, l'angle ϕ est mesuré à partir des atomes C1, N2, Ca2 et C2. **b)** Les trois angles de torsion de la chaîne principale d'une protéine ; ϕ, ψ et ω. La figure montre les angles ψ_1, ω_1 et ϕ_2. Comme l'angle ω a pour valeur fixe -179°, on peut entièrement définir la chaîne principale d'une protéine en spécifiant les angles ϕ et ψ pour chaque acide aminé.

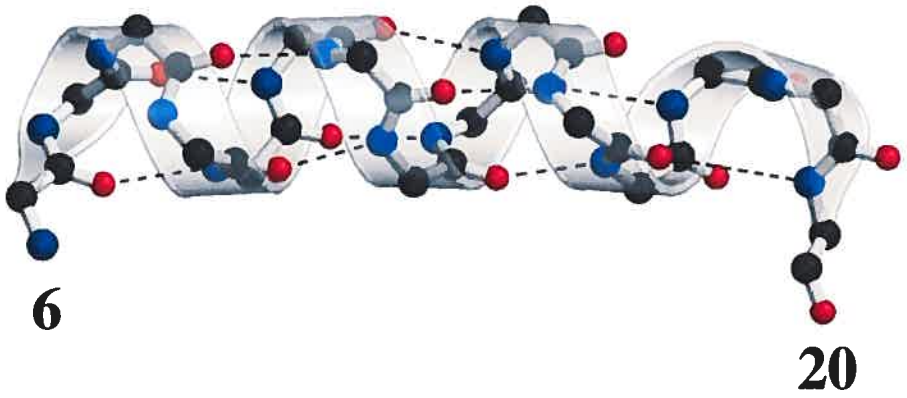


Figure 1.10 – L'hélice alpha. Lorsque la chaîne principale se recroqueville sur elle-même on assiste alors à la formation de ponts hydrogènes entre les atomes du groupement carboxylique C=O de l'acide aminé i et les atomes du groupement aminé H-N de l'acide aminé i+4. La figure montre les atomes de la chaîne principale, sans les atomes d'hydrogènes. Un ruban semi-transparent souligne l'hélicité. Les ponts hydrogènes sont dépeints avec les traits pointillés. L'hélice est celle trouvée dans le fichier PDB 1CRN, entre les résidues 6 et 20.

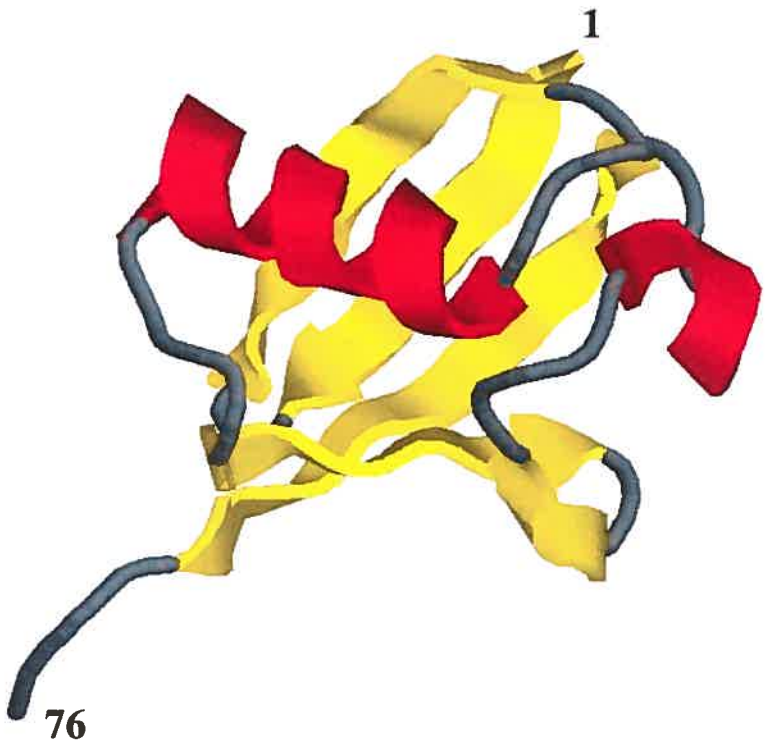


Figure 1.11 – L'ubiquitine comme structure tertiaire. Les hélices α sont en rouge tandis que les brins β sont en jaune. La représentation de la protéine est celle de Richardson. Le premier résidue (1), nommé N-terminal, et le dernier (76), le C-terminal, de la protéine sont annotés. La structure est celle du fichier PDB 1UBI.

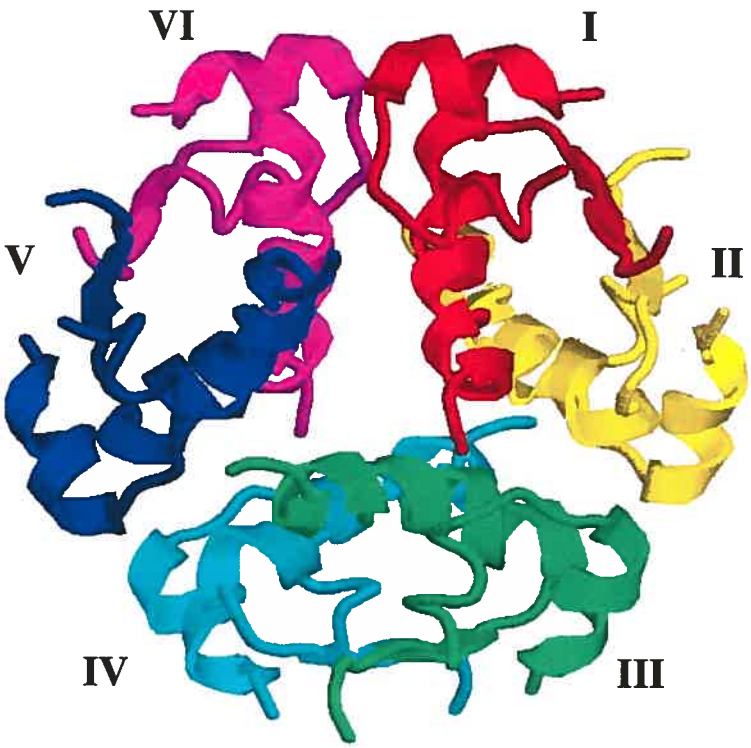


Figure 1.12 – L'insuline comme structure quaternaire. La forme active de l'insuline humaine forme un complexe hexamérique. Chaque monomère est identifié par un nombre romain. La structure est celle du fichier PDB 1AIY.

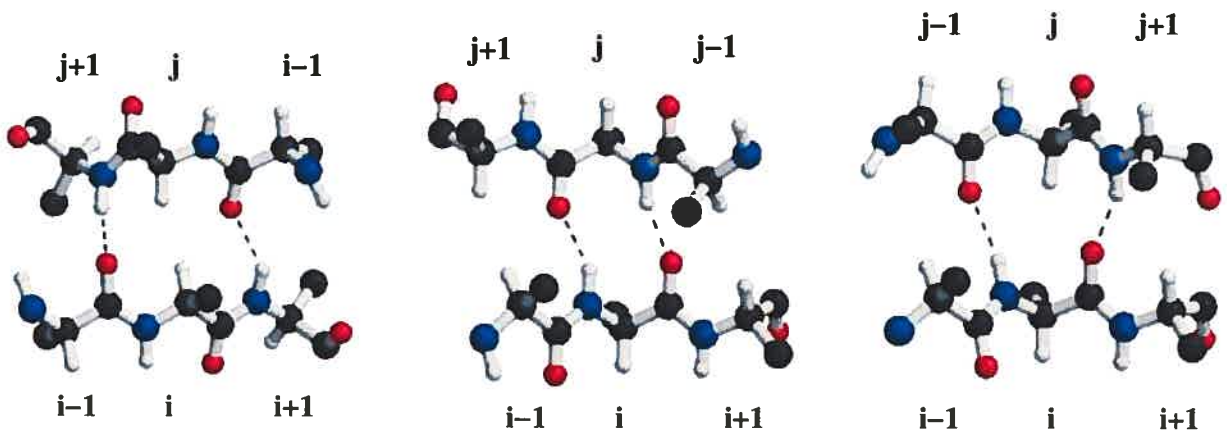


Figure 1.13 – Les motifs canoniques de ponts hydrogènes dans les feuillets β. Les atomes sont représentés par des sphères de couleurs tandis que les liens covalents par des tubes gris. Les atomes de carbones sont en noirs, les oxygènes en rouges, les azotes en bleus et les hydrogènes en gris. Les atomes sont ceux de la chaîne principale plus les carbones β pour indiquer les directions des chaînes latérales. Les ponts hydrogènes entre les groupements C=O et H-N sont représentés par des lignes pointillées. **a)** Motif **I** entre deux chaînes anti-parallèles. Comme les résidues i et j ne sont pas impliqués dans l'échange de ponts hydrogènes on qualifie ce motif d'anneau élargi ("wide ring"). **b)** Motif **II** entre deux chaînes anti-parallèles. Comme les résidues i et j sont impliqués dans l'échange de ponts hydrogènes on qualifie ce motif d'anneau étroit ("narrow ring"). **c)** Motif **III** unique entre deux chaînes parallèles.

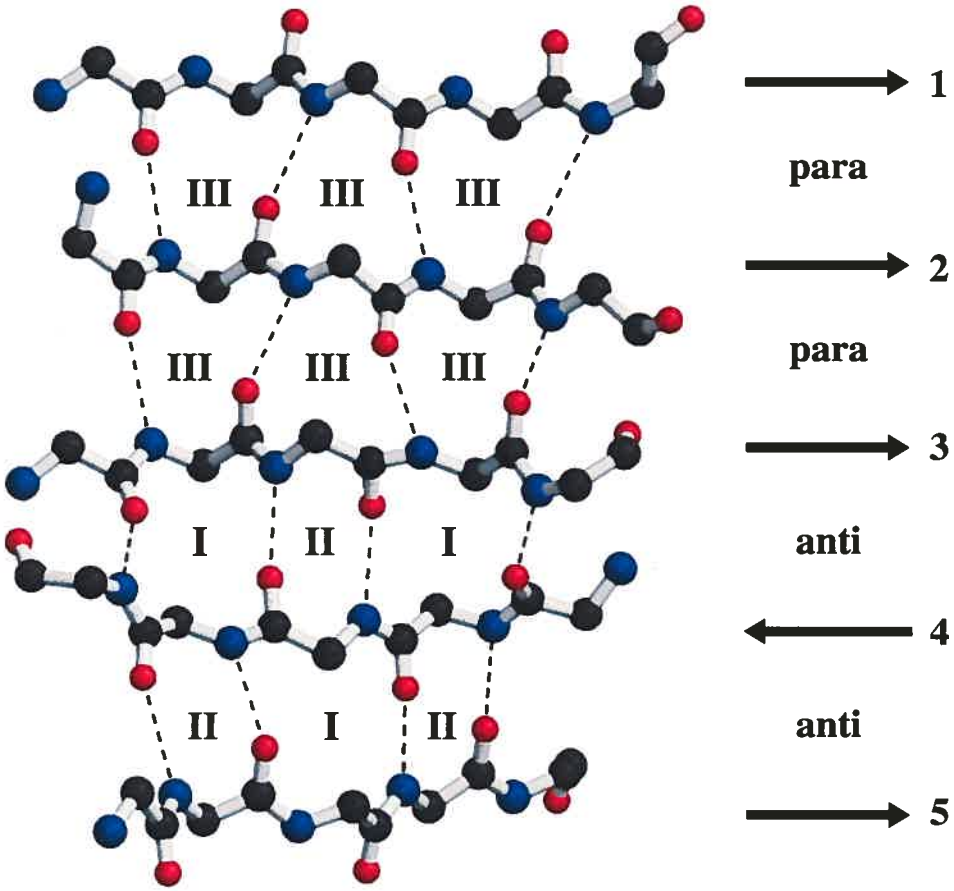
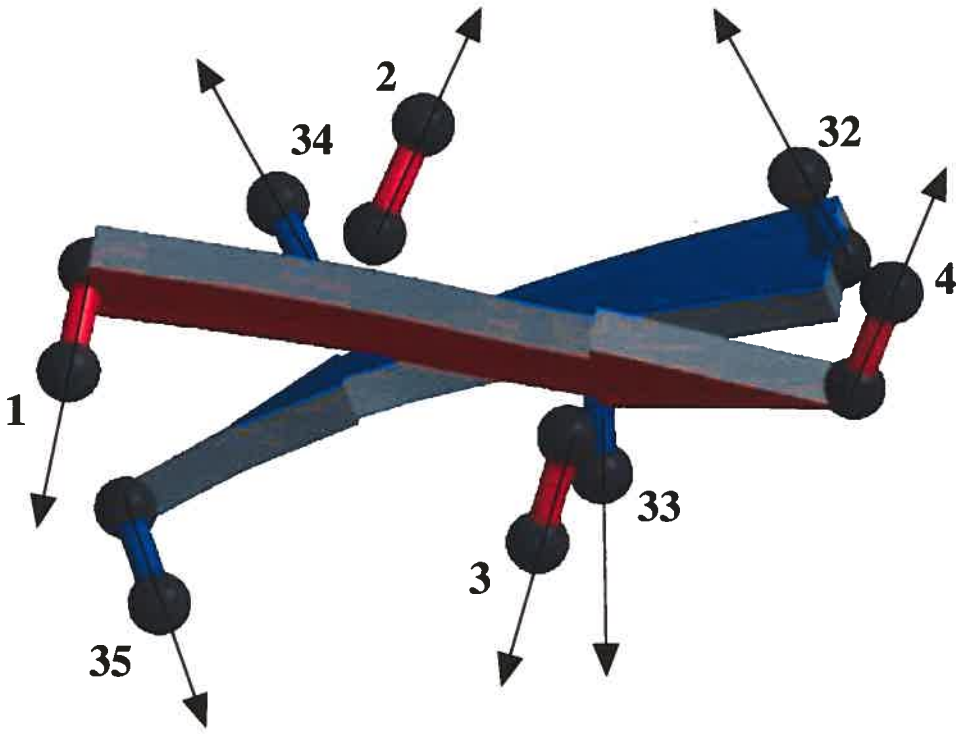


Figure 1.14 – L'agencement des motifs canoniques entre paires de brins parallèles (para) et anti-parallèles (anti). Cinq brins, numérotés de 1 à 5, sont orientés de sorte à offrir toutes les combinaisons d'empilements de paires de brins ; on aura alors para-para, para-anti et anti-anti. Les flèches indiquent le sens de progression des chaînes. Les paires de brins parallèles font appels au motif **III** seulement. Les paires de brins anti-parallèles alternent les motifs **I** et **II**. Seuls les atomes des chaînes principales sont présentés, à l'exclusion des atomes d'hydrogènes. Les atomes d'oxygènes sont en rouges tandis que les atomes d'azotes sont en bleus. Les ponts hydrogènes sont représentés par de minces lignes discontinues reliant l'azote à l'oxygène. Les patrons de ponts hydrognès permettent n'importe quelle combinaison de brins, qu'ils soient parallèles ou anti-parallèles.

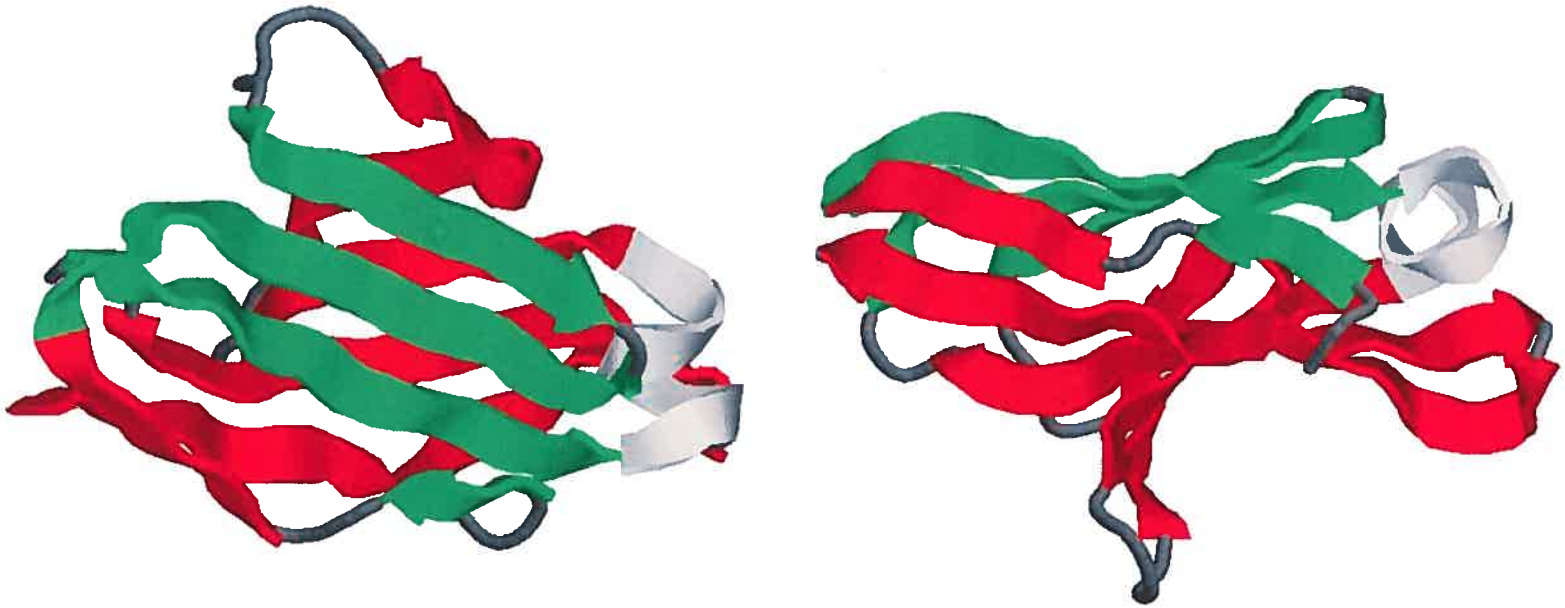


(a)

Figure 1.15 – Disposition des chaînes latérales dans les feuillets β. Le dessin montre le feuillet β tiré du fichier PDB 1CRN. On y voit les deux brins β ; l'un est composé des résidues 1 à 4, où une flèche rouge indique la progression du brin, l'autre de 32 à 35 représenté par la flèche bleu. Les flèches noires partent des carbones α et passent par les carbones β correspondants, et indiquent du coup la position des chaînes latérales. Le dessin met en évidence deux traits ; le premier étant que les résidues de deux brins différents mais côte à côte dans un feuillet ont leurs chaînes latérales qui pointent dans la même direction. Le second trait illustre l'alternance haut/bas des chaînes latérales suivant un brin β.

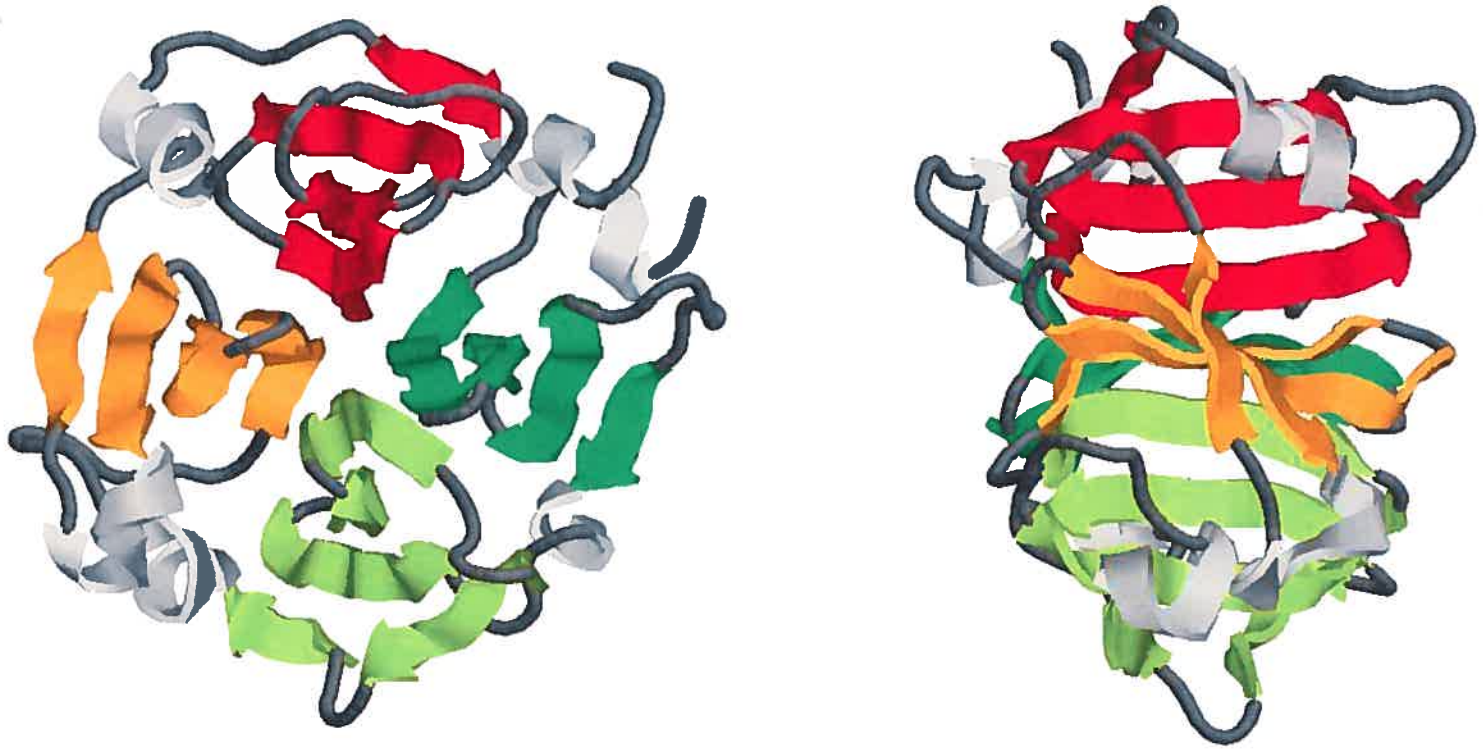
(a) (b)

Figure 1.16 – Exemple de feuillet β : β-hélice. L'hélice est de couleur or. À la différence d'une hélice α avec 3.4 résidues par tour, l'hélice β, ici, en compte 12 résidues par tour. Le code PDB est 1ZEG, chaîne A. Les images ont été produites avec RasMol [12]. Les annotations des feuillets β sont celles de β-Spider [13]. **a)** Une vue de front. **b)** Une vue de coté.



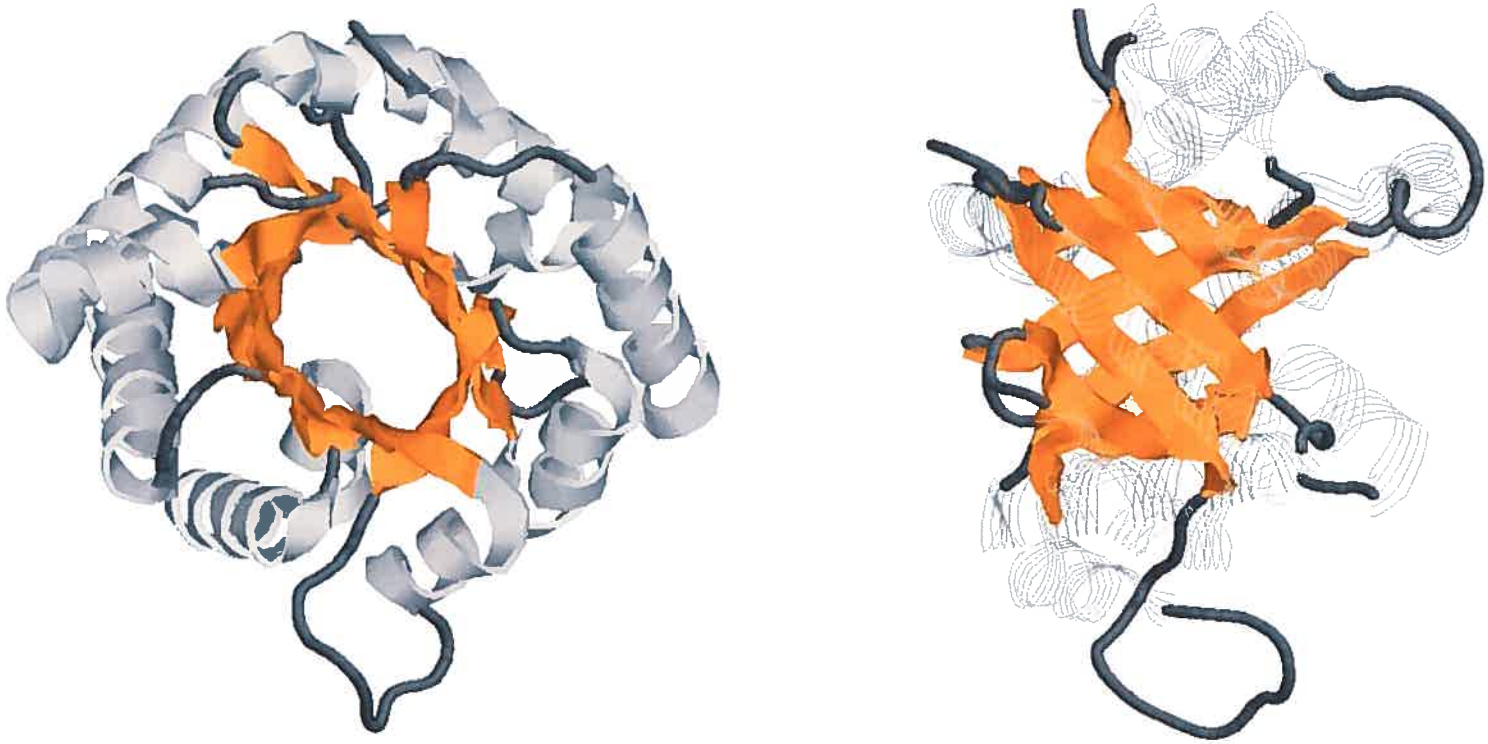
(a) (b)

Figure 1.17 – Exemple de feuillet β : β-sandwich. La sandwich est composée de deux feuillets qui se collent l'un (en rouge) sur l'autre (en vert) avec un certain angle. Le code PDB est 2RHE. **a)** Une vue du dessus. **b)** Une vue de coté.



(a) (b)

Figure 1.18 – Exemple de feuillet β : β-propulseur. Le propulseur est composé de quatre feuillets (aux couleurs rouge, or, vert et lime) qui se placent à angles droits pour adopter la forme d'une turbine. Le code PDB est 1HXN. **a)** Une vue de front. **b)** Une vue de coté.



(a) (b)

Figure 1.19 – Exemple de feuillet β : β-baril. Le baril est composé de huit brins (N=8), tous parallèles, et forment une cavité interne. Lorsqu'on saute d'un brin à l'autre il faut se décaller d'un résidue pour rester à la même hauteur. De là, pour faire le tour du baril et revenir au même résidue il faudra faire 8 décallages, d'où (S=8). Cette configuration de baril avec (N=8,S=8) est appelée "TIM-barrel". Il existe d'autres configurations de barils en variants N et S. Le code PDB est 1TIM, chaîne A. **a)** Une vue du dessus. **b)** Une vue de coté, avec les hélices α rendues semi-transparentes.

1.3 La présentation du mémoire

Ce mémoire est une suite d'articles. Les articles sont tous écrits en anglais soit parce qu'ils ont déjà étés publiés, soit dans un but futur de publication dans des journaux scientifiques. Les articles sont mis dans l'ordre chronologique de rédaction et réflète alors l'ordre et le fil des travaux ici présentés. Nous invitons le lecteur à faire la lecture des annexes jointes dans lesquelles sont développés les termes nécessaires à la bonne compréhension de ce mémoire.

Voici la liste des articles :

1. M Parisien and MC Peitsch and F Major. A protein conformational search space defined by secondary structure contacts. *Pac Symp Biocomput*, 243 :425-436,1998. (c) 1998 World Scientific.
2. M Parisien and F Major. A graph representation for protein β-sheets and it's applications. Non-publié.
3. M Parisien and F Major. A β-sheet conformational search space defined by β-sheet topology graphs. Non-publié.
4. M Parisien and F Major. A new catalog of protein β-sheets. *Proteins*, 61 :545-558, 2005. (c) 2005 Wiley-Liss, Inc.

1.4 Le premier article

L'idée d'utiliser les matrices de transformations homogènes (MTH) pour la construction de modèles tri-dimensionnels de protéines vient du succès de l'approche pour la construction d'acides ribonucléiques, tel que démontré par l'application MC-SYM [14].

Comme MC-SYM utilise des matrices de transformations homogènes (MTH) entre points équivalents dans les bases de nucléotides d'ARN, il fallait alors définir ces mêmes points équivalents pour les protéines. Curieusement, ces points se retrouvent dans la chaîne principale dans le cas des protéines et dans les chaînes latérales pour les ARNs. Comme les points équivalents se retrouvent dans la chaîne

principale des protéines, les MTH résultantes pourront alors encoder l'empilement des chaînes peptidiques les unes sur les autres, incluant en partie l'intéraction entre les chaînes latérales. De plus, pour simplifier l'espace de recherche conformationnelle, les éléments de structures secondaires, à savoir les helices α et les brins β, sont construits à partir d'angles dihèdres fixes et se retrouvent à être les objets transformés par les MTHs.

Nous avons alors démontré qu'il est possible de reconstruire toute protéine à partir de MTHs allochtones, ce qui montre alors que ces MTHs sont interchangeables, et qu'il est alors possible de générer des modèles plausibles pour des protéines à structure inconnue. Pour cette preuve, nous avons utilisés la RMSD comme mesure du succès de la reconstruction de la protéine. Certes, cette mesure introduit un biais par rapport aux modèles généres, mais elle est la seule mesure reconnue par les experts du domaine comme mesure de différence (ou de similarité) entre deux structures tertiaires.

Cet article est paru dans la conférence "Pacific Symposium on Biocomputing" [15]. Les papiers de cette conférence sont revisés par des pairs ("peer-reviewed").

La suite logique de ces travaux était de prédire où sont les contacts entre les éléments de structures secondaires, là où les MTHs sont utilisées pour passer d'un élément à l'autre. Nous nous sommes alors concentré sur les protéines à composantes helicales seulement. Plusieurs approches ont étés utilisées mais en vain, les contacts dissimulent un ordre plus complexe que les modèles tentés.

La reconstruction de protéines avec des éléments de structures secondaires à géométrie fixe montre que les feuillets β sont plus difficile à rebâtir que les hélices α et est la cause, en grande partie, de la fabrication de modèles tri-dimensionnels grossiers par notre méthode. L'amélioration de la construction de feuillets β est alors devenu le sujet d'étude.

1.5 Le deuxième article

Pour pouvoir reconstruire les feuillets β par fragments, il nous fallait inévitablement encoder la topologie des feuillets sous forme de descripteurs mathématiques, lesquels par la suite devraient nous permettre de les comparer entre eux et d'identifier des topologies similaires. La topologie des feuillets β est alors codées dans un graphe, et un algorithme d'isomorphisme de sous-graphes à été développé pour comparer des feuillets de tailles appréciables.

L'idée d'encoder les feuillets β sous forme de graphe n'est pas nouvelle [16–18]. Par contre, ici, les graphes des feuillets ont pour noeuds les résidus tandis que les arêtes ont plusieurs saveurs capturants les relations entre ces résidus. Les relations étant les liens covalents (le lien peptidique entre les résidus), partenaires β (les résidus côte à côte dans les feuillets) et ponts hydrogènes (lorsqu'il y a présence de). L'expressivité de cet encodage nous permet de capturer toutes les caractéristiques des feuillets, notamment les patrons de ponts hydrogènes, l'orientation relative des brins, les "β-bulges" (résidu inséré dans un brin en bordure) et les "β-barrels" (réseau circulaire de brins β).

Pour comparer ces graphes nous avons débutés par un algorithme d'isomorphisme de sous-graphe proposé par Ullmann [19]. La comparaison de graphes de feuillets β de tailles moyennes s'avère insurmontable, du moins dans la forme générale de l'algorithme sur des graphes d'arêtes monochromatiques. Une analyse détaillée de la construction de la matrice initiale d'isomorphisme nous a permis d'exploiter dans celle-ci la topologie particulière des graphes de feuillets β et d'en reduire substantiellement les hypothèses d'isomorphisme à tester.

Une base de données des graphes de feuillets β avec leurs structures 3-D associées à été construite. De là, plusieurs applications sont possibles, notamment la suggestion de structures tri-dimensionnelles (3-D) de feuillets à partir de sa topologie (2-D), l'analyse statistique des acides-aminés autour des "β-bulges", des "β-barrels" ansi que la caractérisation de familles de protéines à partir de leurs feuillets β.

1.6 Le troisième article

Pour voir combien universel sont les graphes de topologies des feuillets β nous avons réalisé une expérience dans laquelle l'isomorphisme de sous-graphe est appliqué sur chaque paire de graphe dans notre base de données. Le résultat étonnant est que les graphes de topologies des feuillets sont, en général, uniques et donc l'isomorphisme avec d'autres graphes est impossible. Cette approche pour la suggestion de structures 3-D de feuillets est sujette à la taille du feuillet à construire, des patrons de ponts hydrogènes entre les brins et aussi à ses défauts, notamment le nombre et l'emplacement des "β-bulges".

Le découpage des graphes de topologie en fragments plus petits augmente le nombre d'isomorphes pour chaque partie, mais encore faut-il rassembler les fragments dans l'espace pour que le produit final ait l'allure d'un vrai feuillet, avec ses courbatures et son réseau de ponts hydrogènes.

Nous avons alors opté pour découper le graphe à construire par paires de brins β. Par exemple, considérons un feuillet avec trois brins ; 1, 2 et 3. On aura alors deux paires ; soit 1-2 et 2-3. Avec l'algorithme d'isomorphisme de sous-graphe, présenté dans l'article précédant, on peut alors trouver des candidats de structures 3-D pour chaque paires. L'assemblage des deux paires dans l'espace passe par le brin commun, ici le brin 2, par lequel on aligne l'un sur l'autre le brin 2 de la première paire avec le même brin 2 de la seconde paire (l'algorithme d'alignement est exactement le même que celui utilisé dans le calcul de la "Root Mean Square Deviation", plus connue sous l'appellation RMSD). Une fois assemblée dans l'espace, la structure finale à deux exemplaires du brin 2 ; il faut donc éliminer des résidus dans ce brin pour retrouver la prescription originale du feuillet. Les résidus écartés sont ceux qui ne sont pas impliqués dans des ponts hydrogènes. Ainsi, on conserve la conformation des ponts hydrogènes en sautant d'une paire de brins à l'autre. Les feuillets β produits par cet algorithme d'assemblage sont d'une qualité exceptionnelle (à l'oeil ils ont l'air "vrais"). De plus, l'expérience en Jackknife nous montre qu'il est possible de reconstruire presque tous les feuillets avec une preci-

sion atomique. L'algorithme explore aussi les limites de flexibilité des feuillets et s'avèrera utile dans la construction *De Novo* de protéines.

1.7 Le quatrième article

Le succès de l'identification de familles de protéines à l'aide des feuillets β isomorphes est fortement lié à la qualité des annotations des feuillets. Or, dans une étude de la famille des ubiquitines, nous avons remarqué que certains membres étaient absents parce que leurs feuillets étaient mal notés. Nous nous sommes alors lancés dans l'identification des feuillets dans le but d'en améliorer les annotations.

Notre première approche fut de considérer tous les patrons possibles de ponts hydrogènes entre paires de tri-peptides, c'est-à-dire entre paires de chaînes peptidiques ayants trois acides aminés. Cette longueur de trois est spécialement choisie pour permettre tous les motifs de ponts hydrogènes canoniques. Comme les annotations resultantes étaient bien meilleures que celles de DSSP [20], la référence en annotations, nous avons même débuté la rédaction d'un article pour en proférer la bonne nouvelle.

Une réflexion sur les causes de l'échec de DSSP de mener à bien sa mission d'annotation des feuillets β et sur notre solution nous a fait croire que l'approche d'énumération exhaustive des patrons de ponts hydrogènes n'explique peut-être pas tout sur la cohésion de chaînes peptidiques adjacentes. En effet, un article clé [21] montre bien qu'il existe d'autres forces, au delà des ponts hydrogènes, qui stabilisent les feuillets β.

Comme les ponts hydrogènes sont en majorité de nature electrostatique, nous avons donc décidé d'utiliser une approche energétique où l'énergie entre deux chaînes seraient évaluée à l'aide d'un champ de force classique (i.e. sans faire appel à la mécanique quantique). Cette approche a maintenant l'avantage de capturer les ponts hydrogènes mais aussi les autres forces en jeu, notamment les dipoles C=O...C=O. La comparaison détaillée entre les résultats de DSSP et notre programme, baptisé β-Spider, montre qu'il est alors possible de croire à une meilleure

classification automatique des protéines par les feuillets β.

Cet article est paru dans la revue scientifique "Proteins. Structure, Function, and Bioinformatics" [13].

CHAPITRE 2

A PROTEIN CONFORMATIONAL SEARCH SPACE DEFINED BY SECONDARY STRUCTURE CONTACTS

M. PARISIEN, F. MAJOR

Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, Québec, Canada H3C 3J7

M. PEITSCH

Glaxo Wellcome Research & Development and Geneva Biomedical Research Institute, CH-1228 Plan-les-Ouates, Switzerland

A conformational search space describing the relative position and orientation of protein secondary structure elements in three-dimensions was defined. These spatial relations were encoded by homogeneous transformation matrices between pairs of residues "in contact" in two different secondary structure elements. A database of all occurrences of spatial relations for five hydrophobic residues was built. The use of one residue contact per pair of secondary structure elements, which were approimated by standard (ϕ, ψ) assignments, was sufficient to reproduce accurately the core structure of proteins with known three-dimensional structures.

M Parisien and MC Peitsch and F Major. A protein conformational search space defined by secondary structure contacts. *Pac Symp Biocomput*, 243 :425-436,1998.

2.1 Introduction

Knowledge about protein three-dimensional structure is key to protein function comprehension and manipulation. Due to difficulties associated with experimental protein structure elucidation, it is not surprising that predictive methods are increasingly gaining popularity. Protein modeling is mainly restricted to comparative methods which only apply to 15 to 20 percent of all known sequences sharing more than 30% identity [22–26]. Consequently to the many genome sequencing projects, an explosion of novel gene discoveries of unknown structure and function is observed [27]. *De novo* protein structure prediction methods are thus needed.

The secondary structure (SS) of a protein can be inferred from its sequence by using statistical methods, such as Markov models [28,29], and neural networks [30, 31]. The SS of a protein can also be determined experimentally, for instance from NMR spectroscopy data. The β-sheet topology of a protein can also be inferred from statistical methods [32], and determined from NMR spectroscopy data [33,34]. Once the sheet topology has been assigned, atomic coordinates of homologous β-sheets in previously determined 3-D structures can be proposed. Thus, α-α and α-β residue contacts can be infered theoretically or determined experimentally [35,36]. These contacts can be translated into geometrical contraints to define a *constraint satisfaction problem* (CSP) to resolve the 3-D structure.

In the search for an acceptable *de novo* modeling scheme, existing methods have been considered and analyzed, and our deire to make use of accumulated structural data led us to consider a protein adaptation of the MC-SYM RNA CSP solver [14, 37, 38]. We thus proove the following scheme for *de novo* protein structure prediction : (i) the definition of the protein SS by existing experimental and theoretical methods ; (ii) the use of SS information to assign β-sheet topologies and α-α and α-β residue contacts to define a CSP ; (iii) the use of MC-SYM to generate consistent core structures ; (iv) the use of existing methods to complete the core structures with loops and side-chains ; and, (v) the refinement and evaluation of the structures using existing energy minimization protocols and potentials [39–41].

In this article, the focus has been put on the implementation of the protein conformational search space, the creation of operators to manipulate protein 3-D core structures, and a best-first search algorithm to demonstrate that the developed conformational search space contains the native x-ray crystal structures. Other conformational search spaces were introduced in the past. The most common methods are based on the sampling of the ϕ-ψ torsion space [9], on theoretical spatial relations of SS elements [35, 42], on the properties of the loops connecting the SS elements [41], and on geometrical sampling of the SS element space [43]. Altough almost the same precision can be reached by the use of these methods, they describe conformational search spaces that are larger than the one introduced in this article, and, in general, require more structural information to converge to the native fold.

2.2 Conformational search space

2.2.1 Definitions

Residue contacts bear side-chain and backbone packing information, that is, the relative position and orientation of the two SS elements which contain the residues in contact. A protein *core structure* is the assembly of its constituent SS elements in 3-D space. Two SS elements are in SS *contact* if they share at least one residue contact.

A *residue contact* between residues A and B forms if their distance is smaller that a certain threshold, $|A, B| < d$, where d is the threshold value and $|\bullet|$ denotes the Euclidean distance. The ensemble of all residue contacts in a given protein constitutes a *residue contact graph*, where each node represents a residue and each edge represents a residue contact (See Figure 2.1).

Similarly, the ensemble of all SS contacts defines the SS *contact graph*, where the nodes represent the SS elements and the edges indicate that at least one residue contact exists between a given pair of residues in the connected SS elements (see Figure 2.2).

Every SS element in a protein is involved in a SS contact. To satisfy this condi-

tion consider the degenerated distance threshold, d = ∞. This makes the SS contact graph *connected*, that is, there is a *path* that connects any pairs of SS elements. A connected SS contact graph that contains no cycle is a *SS contact spanning tree* (see Figure 2.3). There are N^{N-2} spanning trees for a graph that contains N fully connected vertices. The SS spanning tree addresses all SS elements and suggests a construction order in which the SS elements can be introduced. A possible order for the SS spanning tree in Figure 2.3 would be : H2 as the reference SS element; H1 placed from H2; H3 placed from H2; H4 placed from H3; and, H0 placed from H3. It is possible to define a protein *conformational search space* from a SS contact graph. Each residue contact can either be used as a spatial relation operation which positions and orients a SS element from another one (just as in the construction order above), or as a distance constraint. For instance, the contacts dropped in the selection of the SS spanning tree should be replaced by distance constraints that must be satisfied in the final constructions.

2.2.2 Implementation

Homogeneous transformation matrices [44] (HTM) were used to encode the spatial information of residue contacts. HTMs constain translation and rotation information. For instance, the *local referential* of a residue can be represented by an HTM from three right-handed unary orthogonal vectors that can be calculated from three non-colinear atomic coordinates. A local referential indicates the translation and rotation to be applied to the residue coordinates expressed in the canonical referential to obtain its absolute coordinates. Consider for instance the local referential of a residue A, R_A, which can be calculated by using three backbone atoms in A. One of the three is elected as the origin of A while the two others respectively align with the X and Y axes (see Figure 2.4). Backbone atoms, instead of side-chain, were chosen because the backbone characterizes much better the relative orientation and position of the SS elements.

The spatial relation between two residues A and B can also be expressed with an HTM : $T_{A \rightarrow B} = R_A^{-1} \times R_B$. A residue contact between A and B can be reproduced

between any pair of residues, for instance A' and B', by applying $R_{B'}^{-1} \times T_{A \to B} \times R_{A'}$ to the atomic coordinates of B' to position and orient B' with respect to A'. Symetrically, $R_{A'}^{-1} \times T_{A \to B}^{-1} \times R_{B'}$ applied to atomic coordinates of A' positions and orients A' relative to B'.

Any residue contact found in the Protein Data Bank (PDB) [45] can be extracted and used afterwards as a building block of protein 3-D structure to orient and position SS elements. Once a pair of residues have been positioned and oriented from the aplication of HTMs, the extension of each SS element is made by using standard (ϕ,ψ) assignments for the other residues; for instance, (-60°,-40°) for α-helices and (-120°, 140°) for β-strands [46]. In this way, any pair of SS elements involved in SS contact in the database of 3-D structures can accurately be reproduced from a single residue contact. A protein 3-D structure can be built by applying this construction scheme to each of its constituent SS elements. Our hypothesis is that all protein 3-D folds are contained in a conformational search space defined from such SS element contacts.

2.2.3 Transformational sets

A *transformational set* is a set of HTMs associated with a residue contact type, that is, the types of residues and the nature of their host SS elements (α-α, α-β, intra-strand β-β and inter-strand β-β). All possible residue combinations could be part of a residue contact, and thus $1600 = 4 \times 20 \times 20$ transformational sets could be defined. A question that was addressed is whether it would be possible to find a smaller subset of residue contacts that would allow one to define a conformational space containing al protein folds within a desired precision.

Subsets of residues can be identified from *weighted* SS contact graphs where the *weight* of an adge is determined by the minimum *residue contact distance* between the connected SS elements (see Figure 2.5). The residue contact distance is defined by the Euclidean distance between the two closest threading points of two residues, as defined in reference [40]. If only a subset of residues is considered then a subset-specific weighted SS contact graph is defined. The minimum SS spanning tree can

be computed from this graph, as shown in Figure 2.6. The number of contacts in the spanning tree depends on the relative frequencies of the residues and their propensity to make contact. The magnitude in the distances is function of SS element packing and the nature of the contacts.

Consider, for instance, all subsets of five amino acids[1]. There are $\frac{20!}{5!(20-5)!} = 15504$ such subsets. For each subset, consider all contacts and contact distances found in the minimum SS spanning trees obtained from all protein 3-D structures in the PDB Select 25 database (the main characteristic of the PDB Select 25 is that no two structures share more than 25% sequence identity[2]) [47, 48]. The subset that maximizes the number of contacts and returns the smallest mean and median distances is {ALA, ILE, LEU, PHE, VAL}. This result is somewhat not surprising since hydrophobic residues are known to be buried inside proteins and form contacts. This result also confirms our intuition that hydrophobic residues could be best suited for the proposed construction scheme.

A database of transformational sets was built using the subset {ALA, ILE, LEU, PHE, VAL} for α-α, α-β and inter-strand β-β contacts defined by distances smaller than 7.0Å. For the intra-strand β-β contact, hydrogen bonds were used.

2.3 Demonstration

Here we demonstrate that the conformational search space defined by the transformational sets defined above contains any of the known x-ray crystal structures. The demonstration is based on the reproduction of the x-ray crystal structures of 46 proteins randomly chosen from PDB Select 25. Note that the reproductions of those proteins were made by using HTMs extracted from other proteins (Jackknife experiment [49]). To avoid exhaustive exploration, the building procedure was driven by the knowledge of the x-ray crystal structure and, in the consideration of several possible search directions, by exploring first the ones minimizing the

[1] the number five came from an initial intuition that the subset composed of the five most hydrophobic residues could generate good results.

[2] note that a database containing no similar folds would be more appropriate.

root mean square deviation (RMSD) from the x-ray crystal structure; for this reason, the procedure is referred to as the best-first search procedure. This algorithm, however, does not guarantee convergence to optimal construction.

1. Build a pseudo-crystal. Make the reference element the one of the two SS elements that are linked by the largest number of residue contacts and add it to the best-first queue.
2. If the queue is empty then STOP and report the best conformation found so far. Otherwise, select from the best-first queue the partial structure that minimizes the RMSD from the crystal structure divided by the number of SS elements in the partial structure, and superimpose it to the crystal structure.
3. Append to the partial structure selected in step 2 a new SS element according to the spanning tree. All residue contacts from the crystal structure can be used to append the new element. For each contact, compute the spatial relation between the partial structure residue and its partner in the pseudo-crystal structure, T_{best}. Among the transformational set for this contact, determine the best HTM by taking the matrix, from those that differ from T_{best} by less than δÅ in the translation, that minimizes the Euclidean distance of the rotations. Apply the best HTM and the canonical ϕ-ψ assignments to position and orient the SS element in the partial structure. Add the new partial structure to the best-first queue.
4. If the new partial structure from step 3 is complete, that is, all SS elements are present, then compare it to the best completed structure built so far and select the one that has the minimum RMSD with respect to the pseudo-crystal. Remove from the best-first queue all partial structures that would lead to higher RMSDs. A lower bound for the RMSD of partial structures is approximated by adding 0.15Å for each missing SS element. Thus, partial structures with a lower bound RMSD higher than the current best RMSD are eliminated from the queue. Goto step 2.

In the first step of this algorithm, a pseudo-crystal structure is built. The pseudo-crystal represents the x-ray crystal structure from which SS elements were substituted by standard ϕ-ψ assignment SS elements. The building order that was considered is a maximum spanning tree derived from a weighted graph where the nodes represent SS elements and the vertices represent residue contacts (see Figures 2.7 and 2.8). The weights were defined as the number of residue contacts observed in the x-ray crystal structure.

The results of applying the best-first search procedure to 46 proteins are shown in Table 2.1. The results suggest that spatial relations among the five selected hydrophobic residues "in contact" can be used as building blocks of protein 3-D structures. From the RMSD values, the x-ray crystal structures of all tested proteins are clearly accessible from a conformational search space defined by residue contacts.

2.4 Conclusion

A new and efficient representation of protein conformational search space, based on residue contacts, was developed. We have shown that hydrophobic contacts contain information about the relative position and orientation of α-α and α-β SS elements. This is of course a necessary step, not a highly significant result, since evaluating and deciding which fold is the correct one represents the actual difficulty of automated protein structure determination. Nevertheless, the technique presented here shows promises for the developement of a productive protein 3-D modeling scheme. For instance, the technique should allow one to explore a representative small fraction of a protein's conformational space with the use of low resolution data, such as covariation data from multiple sequence analysis and mutagenesis data. Furthermore, a better characterization of residue contacts and their spatial relations should allow us to predict protein 3-D structure from sequence and SS information, a requisite to *de novo* protein deign. The fact that higher RMSD values were measured for proteins that are mainly composed of β-strands

indicates that more efforts should be put on the re-construction of β-sheets than on the re-construction of α-helices. Producing actual predictions is the next step of this research project.

Acknowledgments

We thank S. Lemieux and S. Oldziej for intersting discussions about this work. This work is funded by the Medical Research Council (MRC) of Canada. MP is a NSERC graduate scholar. FM is a MRC fellow.

Protein	#SS		#residues		#PU	RMSD[1] (Å)	RMSD[2] (Å)	RMSD[3] (Å)
	α	β	α	β				
1aak	5	4	44	26	7	0.83	2.47	3.09
1ab2	2	8	20	37	8	1.27	2.88	3.29
1abm	7	5	110	24	11	1.45	3.33	3.26
1atx	0	4	0	22	3	1.52	3.98	4.40
1bab	7	0	108	0	6	0.73	2.54	2.60
1bvh	5	4	52	20	8	0.79	2.44	2.36
1c5a	4	0	49	0	3	0.79	1.56	1.95
1cbn	2	2	21	8	3	0.66	2.17	1.62
1cde	6	7	88	52	12	1.28	2.88	2.85
1chr	12	11	135	57	21	1.04	2.92	3.08
1crl	11	11	116	65	20	1.32	3.26	3.20
1dsb	8	5	112	30	12	1.07	2.97	3.07
1ede	11	8	121	46	18	1.19	3.54	4.26
1erg	1	3	8	14	3	1.15	2.10	2.03
1fas	0	3	0	18	2	1.29	2.07	2.88
1gox	11	8	117	32	18	0.71	2.53	2.77
1hlb	8	0	115	0	7	0.86	2.24	2.39
1l18	9	4	102	15	12	0.69	2.93	2.78
1lga	13	0	125	0	11	0.91	2.49	2.95
1nar	7	9	87	55	14	0.96	2.81	2.92
1ofv	5	6	53	27	10	0.75	2.31	2.60
1pfk	13	11	168	66	21	0.95	2.73	3.06
1phh	12	18	125	81	26	0.86	3.42	3.18
1pii	18	16	157	79	26	0.62	3.35	3.20
1pox	19	20	229	113	32	0.87	3.52	3.59
1rhd	10	10	109	43	19	0.93	3.57	3.38
1sbp	12	11	131	57	20	1.00	3.15	3.62
1sto	7	5	91	34	11	1.75	3.11	3.15
1tca	10	7	89	35	16	1.67	3.04	3.54
1tml	7	7	86	35	12	0.73	2.39	2.38
1ula	7	12	87	68	18	1.08	3.12	3.06
1wsy	11	8	117	50	18	0.78	2.77	3.40
1xya	13	9	170	51	18	0.81	2.83	2.66
2acq	10	8	112	41	16	0.71	3.34	3.22
2atc	9	11	119	81	10	1.41	3.18	3.37
2ctc	9	8	110	45	16	0.74	2.76	2.91
2cyp	12	0	149	0	9	0.99	2.59	2.46
2fal	8	0	112	0	7	0.88	2.16	2.22
2gbp	10	12	139	67	17	1.07	2.64	3.17
2liv	4	7	54	44	9	1.20	2.51	2.53
2pia	5	11	41	70	13	2.55	3.60	3.52
2rn2	4	5	47	43	8	0.91	2.61	2.63
2tmd	16	13	162	69	26	1.00	3.89	3.91
3dfr	3	8	33	59	10	1.58	2.61	3.06
4fxn	4	5	52	37	8	1.33	2.99	3.00
5p21	4	6	51	44	9	1.31	2.86	3.24

Table 2.1 – Results of the best-first search for 46 proteins of the PDB. Proteins are referred to by their PDB mnemonics. The RMSD[1] values indicate the RMSD of the pseudo-crystal structure from the crystal structure. The pseudo-crystal is obtained by substituing the SS elements by canonical elements obtained from standard assignments for the α-helices abd the β-strands. The RMSD[2] values indicate the RMSD of the best found structure, as identified by the best-first search procedure, from the corresponding pseudo-crystal structures. The RMSD[3] values indicate the RMSD of the best found structure, as identified by the best-first search procedure, from the corresponding x-ray crystal structures. The #PU values indicate from how many different proteins the HTMs used in the best structure were extracted. Note that for all proteins composed of N SS elements, $N-1$ HTMs were used for its construction. All RMSD values were calculated for the backbone atoms in the SS elements only.

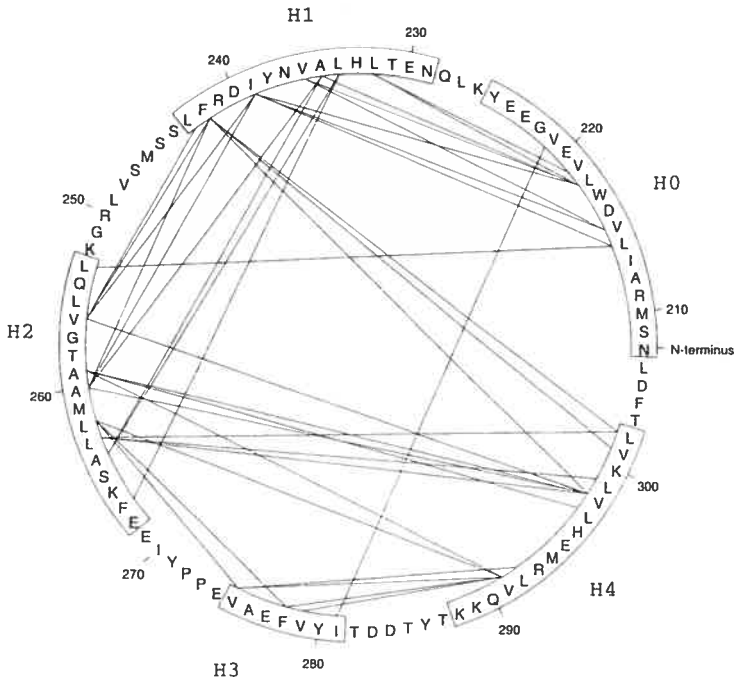


Figure 2.1 – Residue contact graph for the cyclin box (PDB file 1fin). The framed residues are those involved in SS elements. The lines represent residue contacts. The distance threshold was set to 7.0Å between threading points as defined in reference [40].

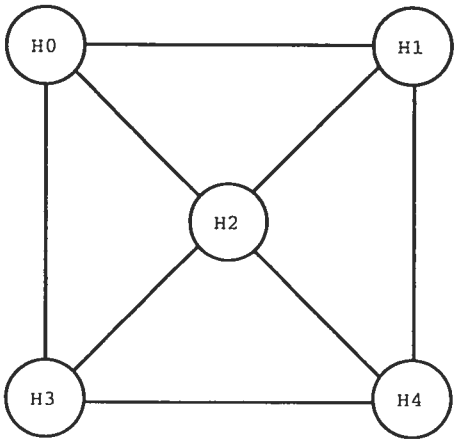


Figure 2.2 – SS contact graph for the cyclin box (PDB file 1fin). The SS elements are circled. An edge was drawn when at least one residue contact was observed between two SS elements (see the residue contact graph in Figure 2.1).

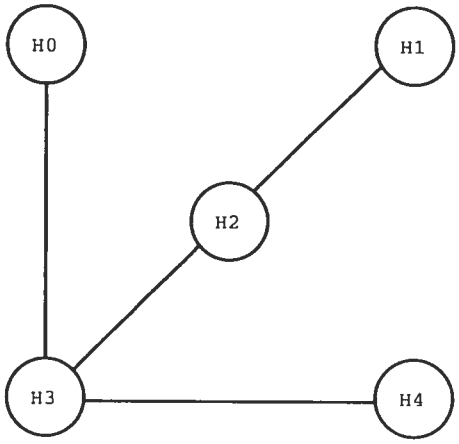


Figure 2.3 – One of the spanning trees for the cyclin box (PDB file 1fin). The SS elements are circled. An edge was drawn when at least one residue contact was observed between two SS elements. The tree, as compared to the graph contains no cycle (see the corresponding SS contact graph in Figure 2.2).

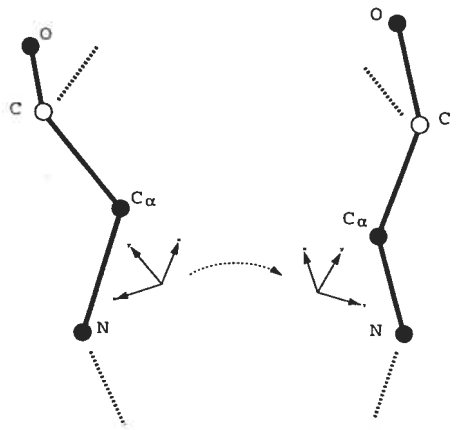


Figure 2.4 – Spatial relation between two residues. The axis systems represent the local referential of the residues. The dotted arrow indicates the transformation of one's referential into the other. The atoms selected to compute the local referential are indicated with black dots. The dotted lines indicate the peptide bonds.

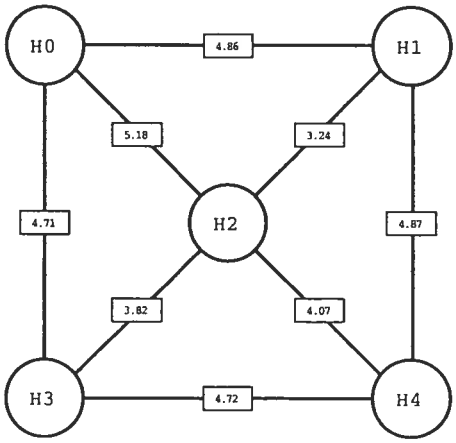


Figure 2.5 – Weigthed SS contact graph for cyclin box (PDB 1fin). The weigths correspond to the minimum distances between two connected SS elements.

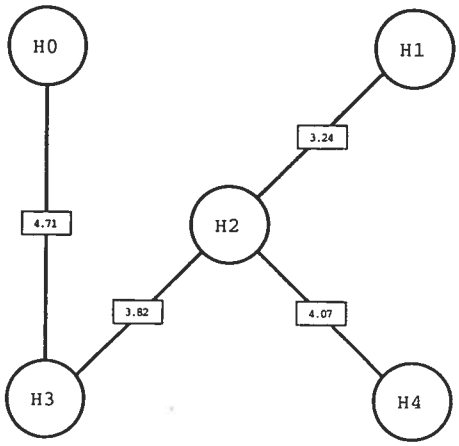


Figure 2.6 – Weighted SS spanning tree for cyclin box (PDB 1fin). This is the minimum spanning tree corresponding to the graph in Figure 2.5.

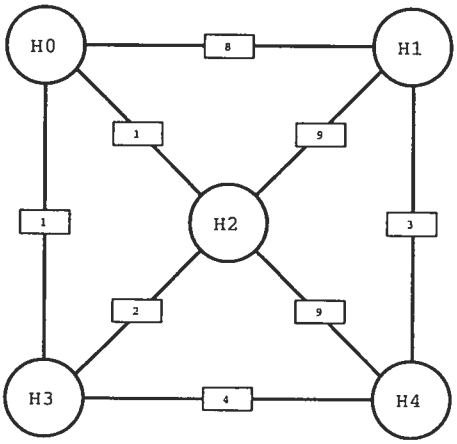


Figure 2.7 – Weigthed SS contact graph for cyclin box (PDB file 1fin). The weights in this case correspond to the number of residue contacts between two connected SS elements.

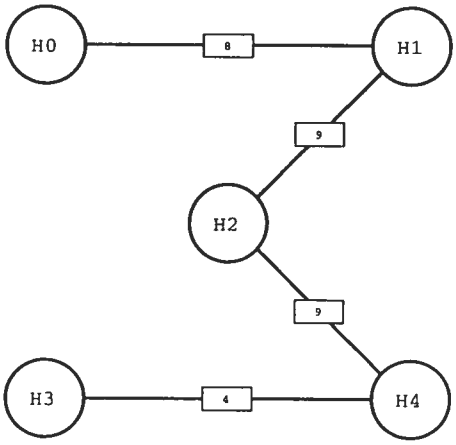


Figure 2.8 – Weighted SS spanning tree for cyclin box (PDB file 1fin). This is the maximum spanning tree corresponding to the graph in Figure 2.7.

CHAPITRE 3

A GRAPH REPRESENTATION FOR PROTEIN β-SHEETS AND IT'S APPLICATIONS

Marc Parisien and François Major
Département d'Informatique et de Recherche Opérationnelle,
Université de Montréal, CP 6128 Succ. Centre-Ville,
Montréal, Québec, Canada H3C 3J7

A graph representation for protein β-sheets is here presented. The residue-level graph encodes all the topological features of β-sheets; peptidic bonds between residues of the same strand, β-sheet partnership and inter-strand H-bonding. The β-sheet topological graph has thus the expressiveness to accurately model all β-sheet features such as standard parallel/anti-parallel H-bond motifs, β-bulges and β-barrels. A database of atomic coordinates of β-sheets and their corresponding β-sheet topology graph has been compiled. A sub-graph isomorphism algorithm has been adapted from Ullman's original form to compare β-sheet topological graphs of appreciable sizes, therefore the β-sheet database can be scrupulously examined for particular motif searches. Applications vary from *De Novo* protein design to β-bulges and β-barrels analysis.

3.1 Introduction

Graphs are powerful mathematical abstractions which enable us to express relations, either quantitative, like distance costs, or qualitative, like pedigrees in genealogical trees, between connected objects. Literature shows that there are several graph representations of proteins, which allow for various tasks such as protein sketching [15], Nuclear Magnetic Resonance (NMR) structure determination [50,51] as well as motifs comparison, similarity search, protein fold identification and classification [16,17,52–54].

β-sheets are present in more than 80% of globular proteins. Proteins can be classified with the help of their β-sheet topology [55–57], i.e. number of strands, relative orientations of strands within the β-sheet. Also, similar β-sheet structures show to have conserved amino-acid type at specific key positions, which qualifies them as folding nucleus residues [58–60].

β-sheet topology graphs can be inferred from low resolution NMR spectroscopy where inter-strand H-bonds can be deduced from slowly exchanging amide resonances. Other cross-strands interactions can be calculated with help of strong and weak Nuclear Overhauser Effect (NOE [61–63]) data [33,34,64–72].

Secondary structure prediction or elucidation in conjunction to strand pairs alignment prediction [28,31,73–81] can also lead to sheet topology graphs. In this case, two hypothesis on the inter-strand H-bonds (and correspondingly, on the orientation up/down of a side-chain for a given residue) must be explored.

Once a graph representation of protein β-sheet can be calculated one can build a database of such graphs, and their corresponding amino-acid atomic 3-D coordinates, in order to compare them with each other or to perform a particular β-sheet motif search within this database. Indeed, several applications of the β-sheet topology graph can be thought of, notably :

1. β-sheet modeling. Given a β-sheet topology graph, each graph matching in the database can supply atomic 3-D coordinates that can be held as models for this β-sheet topology, and will conform in all points to the specified graph.

This can be seen as going from a 2-D representation, i.e. the β-sheet topology graph, to actual 3-D models of β-sheets. It goes further than homology modeling since we do not restrict ourselves to β-sheet coordinates from proteins having similar sequences to the target, thus addressing the conformational search space spanned by β-sheets found in the PDB [82, 83].

2. Sequence analysis. When a particular β-sheet motif is searched for in the database, all the graph-matching solutions can be superimposed, and thus the sequence entropy at each vertex, or residue position in the β-sheet topology graph, can be calculated to reveal key amino-acid types for this particular motif. This can be done for β-bulges, β-barrels and for β-sheets from the same structural family but with distant sequences.

3.2 Method

3.2.1 β-Sheet Topology Graph

A graph $G = (V, E)$ is composed of a finite set of vertices $V = \{x_i\}$ and a finite set of edges $E = \{(x_i \in V, y_i \in V)\}$. The graph is said to be oriented if the edges set has ordered pairs. The in degree of a vertex is the number of oriented edges that is incident to that vertex, while the out degree is the number of edges leaving the vertex. A β-sheet topology graph is a graph G in which the set of vertices V is all the residues contained in the β-sheet, while the set of edges E describes all the various topological relations between the residues of V. This graph is said to be at residue level, because vertices of the graph are residues, compared to other graph representations of proteins, like secondary structure level graphs in which vertices are now secondary structure units. The β-sheet topology graph is oriented; edges start at the lowest residue number to terminate at the highest residue number of the connected residue pair, thus outlining the β-strand progression from N terminal to C terminal, as well as the relative β-strand positions within the β-sheet, and is weakly connected, that is, there is an undirected path between any pair of residues (however it is not strongly connected because some residues may have in degrees

of zero), therefore each residue is attached to the β-sheet graph with either of the topological relations. The topological relations that are censed are :

Type C This relation expresses the backbone $C_i \rightarrow N_{i+1}$ peptidic covalent bond between residues R_i and R_{i+1}.

Type H This relation expresses the presence of at least an H-bond between residues R_i and R_j. H-bonds are defined by a classical Coulomb electrostatic interaction energy as calculated in the DSSP program [20].

Type P This relation expresses a β-sheet partnership between residues R_i and R_j, which are in the same register, thus side-by-side in the β-sheet. This partnership relation is also calculated by DSSP [20].

Type HP This relation is used when types **H** and **P** are simultaneously exhibited.

From there, we can say that the edge set E is composed of the specific edge type sets ; $E = E^C \cup E^H \cup E^P$ (with $E^{HP} = E^H \cap E^P$).

3.2.2 Database

The culled PDB Select 25 database [84], a subset of PDB [85] whose sequences share no more than 25% identity, is used to provide atomic coordinates for the backbone of β-sheets. As of the 7^{th} of February 2004, this database had 1966 chains. The cutoff values are 25% for sequence identity, 2.0Å for resolution and 0.25 for R-Factor. A culled PDB Select 90 database has also been used. Residues that are flagged by DSSP [20] in the 'E' state are considered as part of a β-sheet. The sharing of the same β-sheet identification number by such residues ensures that the corresponding graph is weakly connected. From there, a graph representation of the β-sheet is calculated, in which the vertices of the graph are the residues found within the β-sheet, while the edges encode for the β-sheet topological relations. Every β-sheet feature can be represented in a β-sheet topology graph ; from standard parallel or anti-parallel β-strand pairings to non-canonical inter-strand H-bonding motifs, β-bulges and β-barrels. Figure 3.1 shows a hypothetical mixed parallel/anti-parallel β-sheet with all it's topological relations explicited.

3.2.3 Subgraph Isomorphism

Once a β-sheet topology graph is define we can now tackle the problem of comparing these graphs, or more specifically, to answer the question of whether a graph is included in another one. This problem is known as the subgraph isomorphism problem; is there a subgraph of one graph which is isomorphic to another graph, and is considered NP-complete [86]. Although the fact that β-sheet topology graphs are planar (even for β-barrels), and a polynomial-time algorithm exists [87] for solving such problems when the considered graphs are planar, we adapted Ullman's subgraph isomorphism algorithm [19] to the particular structure of β-sheet topology graphs, thus making possible the comparison of graphs with several hundred residues in few seconds of computation time.

Ullman's subgraph isomorphism algorithm between two graphs, $G^1 = (V^1, E^1)$ and $G^2 = (V^2, E^2)$ with $|V^1| \leq |V^2|$, starts by filling a matrix M^0 of size $|V^1| \times |V^2|$, in which an element in M^0_{ij} is equal to 1 if it is possible, *a priori*, to map the i^{th} vertex of G^1, namely G^1_i, on the j^{th} vertex of G^2, G^2_j, by considering the number and type of incoming and outgoing edges of G^1_i and G^2_j. Let $In(G_i, \mathbf{T})$ be a function that counts the number of incoming edges of a given type $\mathbf{T}$ in a graph G for the i^{th} vertex. Similarly, let $Out(G_i, \mathbf{T})$ count the number of outgoing edges of type $\mathbf{T}$ for the i^{th} vertex in G. Then :

$$M^0_{ij} = \begin{cases} 1 \quad \text{if} & \begin{array}{rcrcll} In(G^2_j, \mathbf{C}) & & & \geq & In(G^1_i, \mathbf{C}) & \text{and} \\ In(G^2_j, \mathbf{H}) & + & In(G^2_j, \mathbf{HP}) & \geq & In(G^1_i, \mathbf{H}) & \text{and} \\ In(G^2_j, \mathbf{P}) & + & In(G^2_j, \mathbf{HP}) & \geq & In(G^1_i, \mathbf{P}) & \text{and} \\ In(G^2_j, \mathbf{HP}) & & & \geq & In(G^1_i, \mathbf{HP}) & \text{and} \\ Out(G^2_j, \mathbf{C}) & & & \geq & Out(G^1_i, \mathbf{C}) & \text{and} \\ Out(G^2_j, \mathbf{H}) & + & Out(G^2_j, \mathbf{HP}) & \geq & Out(G^1_i, \mathbf{H}) & \text{and} \\ Out(G^2_j, \mathbf{P}) & + & Out(G^2_j, \mathbf{HP}) & \geq & Out(G^1_i, \mathbf{P}) & \text{and} \\ Out(G^2_j, \mathbf{HP}) & & & \geq & Out(G^1_i, \mathbf{HP}) & \end{array} \\ 0 \quad \text{otherwise} \end{cases} \tag{3.1}$$

As an example, consider the graphs G^1 and G^2 in Figure 3.2. The resulting M^0 matrix is shown in Table 3.1a.

Because of the special nature of the β-sheet topology graph, when a residue R^1_i of graph G^1 is mapped onto a residue R^2_j of graph G^2 we want also that the

entire strand S_i^1 which includes R_i^1 be also mapped to the corresponding strand S_j^2 in G^2 which contains R_j^2. That requirement makes us consider sub-matrices of M^0, $S_i^1 \times S_j^2$, and zero-out any diagonals of these sub-matrices which admit a null entry. Now, the M^0 matrix contains much less one's, thus reducing the number of possible mappings to check. To pursue our example, the new matrix M^0 is shown in Table 3.1b. Notice that the surviving diagonals have a residue mapping for each residue of each strand of G^1, which happens to be the smallest graph in terms of number of residues.

Ullman's algorithm now generates permutation vectors $V[i] = j$ that tell which residue R_i^1 of G^1 is mapped on which residue R_j^2 of G^2. The permutation vectors have $|V^1|$ entries. Since edges in β-sheet topology graphs are directed from lowest residue number to highest not all generated permutations are valid ; only those that are strictly increasing are taken into account. This requirement has a huge impact on the pruning of the backtrack tree that is used to generate the permutation vectors.

The isomorphism test is then applied for each valid permutation vectors. For each labeled edge $e_{kl}^1 \in V^1 \times V^1$ (we suppose that if $e_{kl}^1 \notin E^1$ then it's label is $\emptyset$ instead of one of **C**, **H**, **P**, **HP**) consider the mapped edge $e_{V[k]V[l]}^2$ linking the two mapped residues $R_{V[k]}^2$ and $R_{V[l]}^2$ in G^2. For the isomorphism test between graphs G^1 and a subgraph of G^2 be found true, each individual label comparison $e_{kl}^1 \leftrightarrow e_{V[k]V[l]}^2$ must be held true, or else G^1 is not an isomorphic subgraph of G^2 relative to the given permutation vector V. The following truth table can be found in Table 3.5. Briefly, this table says that a covalent **C** edge in G^1 must be mapped to a covalent **C** edge in the isomorphic subgraph in G^2, that a partner **P** edge can be mapped to either a partner **P** or an **HP** edge, that an H-bond edge **H** can be also mapped to either a partner **P** or an **HP** edge, that an **HP** edge must map onto an **HP** edge, and finally that an unspecified relation $\emptyset$ can be mapped to any types except the covalent **C** edge.

In our example, a total of four permutation vectors are generated but only two will pass successfully the isomorphism test. The first solution maps residues

$\{111, 112, 113, 121, 122, 123\}$ of G^1 onto $\{211, 212, 213, 222, 223, 224\}$ of G^2, while the second maps onto $\{221, 222, 223, 232, 233, 234\}$. Notice that the unspecified relation between residues R^1_{112} and R^1_{122} in G^1 does not prevent a more specific mapping in G^2 in which type **P** is found.

3.2.4 Distances

The sequence distance between two isomorphic graphs is the sum on all residue positions of the amino-acid substitution similarity score given in the PAM250 matrix [88], as if the β-sheet topological graph would serve as the alignment template. Higher values indicate higher similarity between the sequences of the isomorphic graphs. The RMSD, or root mean square deviation, is calculated in the standard way. At a given β-sheet residue position, one can calculate the sequence variation or entropy using Shannon's [89] celebrated equation $-\sum_C p_i \, log(p_i)$. The Kullback-Leibler distance [90] between two probability distributions, p and q, is $\sum_C p_i \, log(p_i/q_i)$. The last two measures are dependent of the amino-acid partitioning C used; from the two class hydrophobic/polar split (HP model) to the full-fledged 20 symbols break up [91, 92]. A better amino-acid partition could be generated using the methodology of Wang and Wang [93] not on the context-independent MJ matrix [94] but rather on the β-β environment-dependent one [95]. Furthermore, the β-β MJ matrix could be made specific for parallel and anti-parallel strands pairing [96, 97].

3.3 Results and Discussion

3.3.1 Protein Design

Although the graph matching algorithm is not able to generate new β-sheet topologies, it can asses whether or not if a given topology is unheard of. In a recent article published by Baker et al. [98], the group claims to have engineered a novel protein fold expressed as Top7. The protein's β-sheet topology graph has been used to scan the graph database and was found to be no isomorphic subgraph of any

β-sheet in the culled PDB Select 25, it is thus a unique β-sheet topology.

On the other hand, the β-sheet topology of human monocyte chemoattractant protein MCP-1 (PDB code 1DOK [99, 100]) (See Figure 3.3a) has 75 isomorphic siblings in the culled PDB Select 25, not only in the Interleukin 8-like chemokines SCOP [56] family, but also in various other unrelated protein classes, as shown in Table 3.3. The MCP-1 β-sheet has a bulged residue at position ALA26, thus making the corresponding topology graph depart from the ubiquitous canonical anti-parallel 3-stranded sheets (we obtain 411 isomorphic β-sheet if we eliminate residue 25 and make residue 26 a partner **P** to residue 45). Since this β-sheet topology, including the β-bulge, seems to be used in a wide variety of proteins of different functions it can be speculated about the specific role of the β-bulge as to prevent amyloidosis [101], since the bulged strand is exposed to solvent. Figure 3.4 shows a superposition of the heavy backbone atoms of the 75 isomorphic graphs. The models are superimposed on their first strand, which comprises residues 25 to 31. The two closest models have an RMSD of 0.38 Å, while the two farthest have 6.99 Å. These models can serve as template for either *De Novo* protein design or homology modeling.

3.3.2 β-Bulge

Sequence analysis has been performed on the β-bulge region of MCP-1, residues $\{25, 26, 27, 28, 43, 44, 45\}$. This β-bulge pattern is the most common occurring in anti-parallel strand pairs, and is of type C+ (classic), as defined in [102]. A total of 733 isomorphic siblings can be found in the culled PDB Select 25 for this particular residue motif shown in Figure 3.3b. The results are summarized in Table 3.4. Four amino-acid classes were used to partition the findings of the subgraph isomorphism algorithm, and are those taken from [92], Table II. This β-bulge motif has particular amino-acid preferences at key positions. It is noteworthy to mention that our results differ slightly from those published in [102]. In particular, we find that residues in position **1** are not necessarily large hydrophobic residues, but instead, large aromatic residues and cysteine {CFYW} (only 6% occurence) seem to be

forbidden. Also, not only small residues {GPATS} (40%) are favored at position **2**, but also large polar residues {NHQEDRK} (40%), although large hydrophobic residues {MLIV} (only 7%) seem to be denied. Position **X** does not seem to have a particular preference, at least in our amino-acid subdivisions. It is also interesting to note some deviance from the reference amino-acid composition, especially at position 25 where large hydrophobic residues predominate (68%), and at position 45 for which aromatic residues are observed more than normally (27%). Even though the Shannon entropy measures in Table 3.4 doesn't reveal conserved amino-acid positions, exception made at residue position 25 where the entropy is very low due to the high presence of large hydrophobic amino-acids, the Kullback-Leibler distance, on the other hand, is more sensible to departures to the reference amino-acid distribution (from β-sheets of PDB Select 25), and shows that positions 25 (highly large hydrophobic), 27 (predominantly polar or charged residues) and 45 (important aromatic occupancy) are of extremely specific amino-acid distribution, and could prove to be necessary for the C+ β-bulge type fold.

3.3.3 β-Barrel

A β-barrel is a β-sheet in which the first strand is H-bonded to the last one, and makes a barrel-like 3-D structure. Two parameters fully describe the topological features of regular β-barrels; the number of strands, n, taking part in the barrel and the shear number, S, that is the distance separating the starting residue on the first strand and the terminal residue on the same strand after a walk around the barrel in a direction perpendicular to the strand orientations [103]. Figure 3.5 shows two alternatives for (n=8,S=8) β-barrels called β-rings. Here, the relative strands orientation is parallel. The difference between the two graphs is the handedness of the partnership **P** connections.It is interesting to note that these two β-ring motifs yield quite different results as for isomorphic siblings in the culled PDB Select 25 database. Figure 3.5a has 85 solutions while Figure 3.5b has none. This is due to the handedness of the crossover connections between two consecutive parallel strands [4]. The β-barrel encoded in the β-sheet topological graph of Figure 3.5a would

put the α-helices *outside* of the barrel, whereas the one if Figure 3.5b would put them *inside* the barrel to satisfy the connection handedness (almost all examples are found to be right-handed). Within this theoretical framework it is possible to address various values of n and S.

3.3.4 Ubiquitin-Like Fold

Michnick and Shaknovich have presented a strategy for detecting the conservation of the folding nucleus residues in protein superfamilies [58]. They have applied their methodology to some proteins in the Ubiquitin-like superfamily, as defined in SCOP [56], and results in the identification of seven potential residues involved in the folding nucleus, from which six of those are found in β-sheets. From there, a β-sheet topological graph, or β-sheet descriptor, has been defined to contain all the key folding nucleus residues, and shown in Figure 3.6. This descriptor, when searched in the culled PDB Select 90 database in β-sheets with less than 75 residues, produces 34 solutions that can be found in Table 3.5. When more than one match of the descriptor can be applied on a specific β-sheet only the best sequence match is retained. This procedure thus allows for a systematic fold alignment based on the topology of the underlying β-sheets. Even though the descriptor doesn't capture all Ubiquitin-like folds (SCOP version 1.65) in the PDB Select 90, it is able to pick up only β-sheets of the β-grasp fold type, with only one exception in PDB file 1DFU which is a small β-barrel (n=6,S=10). A sequence entropy analysis can be found in Figure 3.7. It is interesting to note the entropy change upon the amino-acid type partitioning choice. With the 6-letter code of [58], residue 16 has a high entropy, whereas with the 6-letter code of [92] Table II, it has a low entropy. This is due to the fact that the later amino-acid partition groups together the charged and bulky polar residues (NHQEDRK), while in the former amino-acid partition these residues are in three different groups, thus enhancing the entropy at this position. Residues at position 3, 5 and 67 have low entropies, in accord with Michnick and Shakhnovich. On the other hand, residues 15, 17 and 69 do not show low entropies even though they have a highly hydrophobic content. An algorithm to compute

the common β-sheet topological graph from a set of given β-sheets of the same family or superfamily should reveal the key positions for the fold, but will prove very sensitive to the quality of the β-sheet annotations.

3.4 Conclusion

A graph representation for protein β-sheets is here presented. The residue-level graph encodes all the topological features of β-sheets; peptidic bonds between residues of the same strand, β-sheet partnership and inter-strand H-bonding. The β-sheet topological graph has thus the expressiveness to accurately model all β-sheet features such as standard parallel/anti-parallel H-bond motifs, β-bulges and β-barrels. A database of atomic coordinates of β-sheets and their corresponding β-sheet topology graph has been compiled. A sub-graph isomorphism algorithm has been adapted from Ullman's original form to compare β-sheet topological graphs of appreciable sizes, therefore the β-sheet database can be scrupulously examined for particular motif searches. Applications vary from *De Novo* protein design to β-bulges and β-barrels analysis. The LINUX version of the program and associated database can be found at the following address : http ://www-lbit.iro.umontreal.ca/bSheet/index.html.

3.5 Acknowledgments

We would like to thank Vincent Devloo for reviewing this manuscript. This work was jointly funded by the Canadian Institutes of Health Research (CIHR MT-14604), Genome Québec and Genome Canada. FM is a CIHR investigator.

	S_1^2				S_2^2				S_3^2			
	1	0	1	0	1	0	1	0	0	0	0	0
S_1^1	0	1	1	0	0	1	1	0	0	1	1	0
	0	0	1	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	1	0	0	0	1	0	0
S_2^1	0	1	1	0	0	1	1	0	0	1	1	0
	0	0	0	0	0	1	0	1	0	1	0	1

(a)

	S_1^2				S_2^2				S_3^2			
	1	0	0	0	1	0	0	0	0	0	0	0
S_1^1	0	1	0	0	0	1	0	0	0	0	0	0
	0	0	1	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	1	0	0	0	1	0	0
S_2^1	0	0	0	0	0	0	1	0	0	0	1	0
	0	0	0	0	0	0	0	1	0	0	0	1

(b)

Table 3.1 – Optimization of the M^0 matrix of the subgraph isomorphism problem between the two graphs depicted in Figure 3.2. S_l^k refers to the l^{th} strand of graph k. **a)** The M^0 matrix without optimization as computed by the original Ullman subgraph isomorphism algorithm. **b)** The M^0 matrix after the zero-out diagonal walk optimization.

		$e^2_{V[k]V[l]}$				
		C	**H**	**P**	**HP**	$\emptyset$
e^1_{kl}	**C**	T				
	H		T		T	
	P			T	T	
	HP				T	
	$\emptyset$		T	T	T	T

Table 3.2 – Truth table used in the subgraph isomorphism algorithm. e^1_{kl} is an edge in G^1 between residues R^1_k and R^1_l, while $e^2_{V[k]V[l]}$ is an edge in G^2 between residues $R^2_{V[k]}$ and $R^2_{V[l]}$. The vector V is one of the mapping vectors generated by the algorithm such that the i^{th} entry in V, $V[i]$, is the residue in G^2 on which i of G^1 is mapped onto. The various topological relations **C**, **P**, **H** and **HP** as well as $\emptyset$ are as defined in the text. Only true entries in the truth table are signaled; all other entries are to be taken as false.

PDB ID	S_1	S_2	S_3	D_{SEQ}	R_{SEQ}	D_{RMSD}	R_{RMSD}	Molecule	Header
1DOK	25A-31A	40A-45A	50A-53A	88	-	0.00	-	MONOCYTE CHEMOATTRACTANT PROTEIN 1	CHEMOATTRACTANT
1Q42	85A-91A	100A-105A	147A-150A	26	0	2.48	61	MRNA TRANSPORT REGULATOR MTR2	TRANSLATION
1TVX	39A-45A	55A-60A	65A-68A	24	1	0.80	4	NEUTROPHIL ACTIVATING PEPTIDE 2 VARIANT	CYTOKINE
1EC7	10A-16A	40A-45A	50A-53A	15	2	1.20	13	GLUCARATE DEHYDRATASE	LYASE
1EE8	44A-50A	53A-58A	63A-66A	14	3	2.24	57	MUTM (FPG) PROTEIN	DNA BINDING PROTEIN
1NH2	66D-72D	75D-80D	112D-115D	13	4	2.11	53	TRANSCRIPTION INITIATION FACTOR TFIID	TRANSCRIPTION/DNA
1C8C	20A-26A	29A-34A	43A-46A	13	5	0.89	5	DNA-BINDING PROTEIN 7A	DNA BINDING PROTEIN/DNA
2HFT	113-119	122-127	174-177	13	6	1.74	16	HUMAN TISSUE FACTOR	COAGULATION FACTOR
1QJ5	356A-362A	365A-370A	396A-399A	12	7	1.61	27	7,8-DIAMINOPELARGONIC ACID SYNTHASE	AMINOTRANSFERASE
1CEW	47I-53I	59I-64I	96I-99I	11	8	2.12	54	CYSTATIN	PROTEINASE INHIBITOR(CYSTEINE)
1JOV	98A-104A	107A-112A	126A-129A	6	9	2.96	64	HI1317	UNKNOWN FUNCTION
1G4M	228A-234A	324A-329A	342A-345A	6	10	1.45	21	BETA-ARRESTIN1	SIGNALING PROTEIN
1GUI	39A-45A	48A-53A	146A-149A	5	11	2.05	45	LAMINARINASE 16A	CARBOHYDRATE BINDING MODULE
1JZ8	222A-228A	241A-246A	288A-291A	5	12	2.14	55	BETA-GALACTOSIDASE	HYDROLASE
1RVK	3A-9A	37A-42A	47A-50A	4	13	2.07	48	ISOMERASE/LACTONIZING ENZYME	UNKNOWN FUNCTION
1LSH	801A-807A	819A-824A	869A-872A	3	14	4.01	70	LIPOVITELLIN (LV-1N, LV-1C)	LIPID BINDING PROTEIN
1QBF	172O-178O	245O-250O	307O-310O	3	15	1.78	36	GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE	GLYCOLYTIC PATHWAY
1MUC	6-12	36-41	46-49	3	16	1.07	8	MANDELATE RACEMASE	ISOMERASE
1LUG	6A-12A	36A-41A	46A-49A	3	17	1.22	15	MUCONATE LACTONIZING ENZYME	ISOMERASE
1B8A	90A-96A	117A-122A	142A-145A	2	18	2.67	63	CARBONIC ANHYDRASE II	LYASE
1E42	29A-31A	34A-39A	47A-50A	2	19	3.14	65	ASPARTYL-TRNA SYNTHETASE	LIGASE
1NJH	875A-881A	884A-889A	902A-905A	2	20	1.14	11	BETA2-ADAPTIN	ENDOCYTOSIS
1MB4	54A-60A	66A-71A	75A-78A	2	21	1.65	30	PROTEIN YOJE	UNKNOWN FUNCTION
1ES5	152A-158A	277A-282A	341A-344A	1	22	1.96	41	ASPARTATE-SEMIALDEHYDE DEHYDROGENASE	OXIDOREDUCTASE
1DZK	209A-215A	223A-228A	236A-239A	1	23	3.98	69	DD-TRANSPEPTIDASE	HYDROLASE
1N9E	39A-45A	50A-55A	67A-70A	-1	24	1.82	28	ODORANT-BINDING PROTEIN	ODORANT BINDING PROTEIN
1BKB	377A-383A	529A-534A	689A-692A	-1	25	2.03	44	GLYOXYLOXIDASE	OXIDOREDUCTASE
1K3X	32-38	47-52	62-65	-1	26	1.37	19	TRANSLATION INITIATION FACTOR 5A	TRANSLATION
1JYI	44A-50A	53A-58A	63A-66A	-1	27	2.09	51	ENDONUCLEASE VIII	HYDROLASE/DNA
1NVM	196A-202A	266A-271A	275A-278A	-2	28	2.17	56	TYROSYL-DNA PHOSPHODIESTERASE	HYDROLASE
1PJJ	150B-156B	210B-215B	270B-273B	-3	20	1.87	38	4-HYDROXY-2-OXOVALERATE ALDOLASE	LYASE/OXIDOREDUCTASE
1J5P	343A-349A	415A-420A	429A-432A	-3	30	2.35	60	INOSITOL-3-PHOSPHATE SYNTHASE	ISOMERASE
1WHO	123A-129A	194A-199A	203A-206A	-3	31	1.66	31	HYPOTHETICAL PROTEIN TM1643	UNKNOWN FUNCTION
1GYT	28-34	65-70	75-78	-3	32	2.03	43	ALLERGEN PHL P 2	ALLERGEN
1UMW	66A-72A	81A-86A	100A-103A	-4	33	2.07	49	NUCLEAR TRANSPORT FACTOR 2	NUCLEAR TRANSPORT
1IJQ	411A-417A	422A-427A	432A-435A	-4	34	1.73	33	SERINE/THREONINE-PROTEIN KINASE PLK	KINASE
1JC4	408A-414A	419A-424A	430A-433A	-4	35	1.28	17	LOW-DENSITY LIPOPROTEIN RECEPTOR	LIPID TRANSPORT
1F2T	35A-41A	46A-51A	65A-68A	-5	36	1.45	22	METHYLMALONYL-COA EPIMERASE	ISOMERASE
1IWL	3A-9A	70A-75A	83A-86A	-5	37	3.15	66	MJ0570 ABC-ATPASE	REPLICATION
1DPG	124A-130A	139A-144A	149A-152A	-6	38	3.45	68	OUTER-MEMBRANE LIPOPROTEINS CARRIER	PROTEIN TRANSPORT
1QLZ	207A-213A	346A-351A	368A-371A	-6	39	1.50	24	GLUCOSE 6-PHOSPHATE DEHYDROGENASE	OXIDOREDUCTASE (CHOH(D) - NAD(P))
1NEB	190A-196A	209A-214A	232A-235A	-7	40	1.03	40	SEMAPHORIN 4D	DEVELOPMENTAL PROTEIN
1JET	15A-21A	37A-42A	89A-92A	-7	41	2.09	52	ADRENAL SECRETORY PROTEIN E1	LIPID BINDING PROTEIN
1D7U	61A-67A	71A-76A	143A-146A	-8	42	1.01	7	OLIGO-PEPTIDE BINDING PROTEIN	PEPTIDE TRANSPORT
1N9E	351A-357A	360A-365A	404A-407A	-8	43	1.66	32	2,2-DIALKYLGLYCINE DECARBOXYLASE (PYRUVATE)	LYASE
1MW9	95A-101A	122A-127A	139A-142A	-9	44	1.87	37	LYSYL OXIDASE	OXIDOREDUCTASE
1K8K	188-194	370-375	1010-1013	-9	45	1.59	26	LEUKOTRIENE A-4 HYDROLASE	HYDROLASE
1EYQ	268C-274C	410C-415C	423C-426C	-9	46	1.77	35	DNA TOPOISOMERASE I	ISOMERASE
1NH2	55C-61C	66C-71C	76C-79C	-9	47	1.31	18	ACIN-LIKE PROTEIN 3	STRUCTURAL PROTEIN
1QLZ	30A-36A	48A-53A	93A-96A	-11	48	2.02	42	CHALCONE-FLAVONONE ISOMERASE 1	ISOMERASE
1F9Z	399A-405A	417A-422A	427A-430A	-11	50	0.67	0	TRANSCRIPTION INITIATION FACTOR TFIID	TRANSCRIPTION/DNA
7AHL	28A-34A	39A-44A	56A-59A	-11	51	1.41	20	SEMAPHORIN 4D	DEVELOPMENTAL PROTEIN
1KAF	81A-87A	248A-253A	275A-278A	-12	52	1.46	23	ALPHA-HEMOLYSIN	LYASE
1Q7F	128A-134A	137A-142A	151A-154A	-12	53	3.31	67	TRANSCRIPTION REGULATORY PROTEIN MOTA	CYTOLYTIC PROTEIN
1JZ8	1004A-1010A	1014A-1019A	1024A-1027A	-12	54	0.77	2	BRAIN TUMOR CG10719-PA	TRANSCRIPTION
1JKG	822A-828A	835A-840A	853A-856A	-13	55	2.32	59	BETA-GALACTOSIDASE	TRANSLATION
1KKO	72A-78A	92A-97A	110A-113A	-13	56	1.91	39	P18	HYDROLASE
1KSH	3A-9A	54A-59A	64A-67A	-13	57	0.97	6	3-METHYLASPARTATE AMMONIA-LYASE	TRANSPORT PROTEIN
1H2W	17B-23B	60B-65B	104B-107B	-15	58	1.56	25	ARF-LIKE PROTEIN 2	LYASE
1R17	338A-344A	348A-353A	361A-364A	-15	59	0.78	3	PROLYL ENDOPEPTIDASE	SIGNALING PROTEIN/HYDROLASE
1J8D	285A-291A	310A-315A	386A-389A	-16	60	2.06	46	FIBRINOGEN-BINDING PROTEIN SDRG	HYDROLASE
1N2E	109A-115A	118A-123A	158A-161A	-17	61	2.06	47	RIBOFLAVIN SYNTHASE	CELL ADHESION
1EW9	247A-253A	267A-272A	280A-283A	-17	62	2.08	50	PANTOTHENATE SYNTHETASE	TRANSFERASE
1OXX	49A-55A	73A-78A	91A-94A	-18	63	2.50	62	CHORISMATE LYASE	LIGASE
1KCM	301R-307K	310R-315K	325R-328K	-18	64	1.09	9	ABC TRANSPORTER, ATP BINDING PROTEIN	LYASE
1FW9	42A-48A	56A-61A	87A-90A	-19	65	2.32	58	PHOSPHATIDYLINOSITOL TRANSFER PROTEIN ALPHA	TRANSPORT PROTEIN
1QJZ	126A-132A	139A-144A	152A-155A	-21	66	1.73	34	CHORISMATE LYASE	LIPID BINDING PROTEIN
1NWW	272A-278A	287A-292A	304A-307A	-25	67	0.71	1	SEMAPHORIN 4D	LYASE
1A12	85A-91A	94A-99A	117A-120A	-25	68	1.16	12	LIMONENE-1,2-EPOXIDE HYDROLASE	DEVELOPMENTAL PROTEIN
1DQI	242A-248A	251A-256A	261A-264A	-27	69	1.10	10	REGULATOR OF CHROMOSOME CONDENSATION 1	HYDROLASE
	49A-55A	106A-111A	116A-119A	-32	70	1.21	14	SUPEROXIDE REDUCTASE	NUCLEOTIDE EXCHANGE FACTOR
									OXIDOREDUCTASE

Table 3.3 – Subgraph isomorphic β-sheets of the monocyte chemoattractant protein 1 (MCP-1) [99, 100]. The source β-sheet topological graph has been extracted from the PDB code 1DOK structure as annotated by the DSSP secondary structure assignment algorithm [20]. In the table, the PDB ID column refers to the PDB code of the isomorphic β-sheet. The S_1 column identifies which residues are part of the first strand mapping of the β-sheet, S_2 for the second strand and S_3 for the third strand. Residues are identified by their PDB residue number in addition to the optional chain identifier. The D_{SEQ} column is the sequence distance between the reference amino-acid sequence on the β-sheet of 1DOK with the isomorphic sibling, as calculated by the PAM250 substitution matrix [88], and higher values signify higher homology. The maximum D_{SEQ} value is 88 for the comparison of the β-sheet sequence of 1DOK with itself. The R_{SEQ} column is the sorted sequence distance rank. The D_{RMSD} column shows the RMSD distance between the crystal structure of the β-sheet of 1DOK compared to the isomorphic sibling. The RMSD is taken after strands S_1 of both structures have been aligned and is calculated from all heavy backbone atoms. The R_{RMSD} column is the sorted structure distance rank. The Molecule column is extracted from the PDB file under the "COMPND MOLECULE" keyword. The Header column is also taken from the PDB file but under the "HEADER" keyword.

#	T	SE	KL	CFYW		MLIV		GPATS		NHQEDRK	
				N	%	N	%	N	%	N	%
E		1.34e+00		13728	14	34433	36	24198	25	23762	25
25		9.61e-01	2.46e-01	58	8	497	**68**	120	16	58	**8**
26	1	1.21e+00	5.41e-02	46	**6**	341	47	157	21	189	26
27	2	1.19e+00	3.53e-01	94	13	53	**7**	291	**40**	295	**40**
28		1.28e+00	1.38e-02	86	12	321	44	170	23	156	21
43		1.26e+00	5.96e-02	119	16	342	47	179	24	93	**13**
44	X	1.29e+00	9.17e-03	82	11	305	42	169	22	177	24
45		1.29e+00	1.20e-01	200	**27**	280	38	182	25	71	**10**

Table 3.4 – C+ class [102] β-bulge sequence analysis. This β-bulge is found in the monocyte chemoattractant protein 1 (MCP-1) [99, 100]. There are 733 isomorphic siblings to this β-bulge. The '#' column refers to the residue sequence number as found in PDB code 1DOK. The T column is the residue tags of the C+ β-bulge motif found in [102], Figure 2. The SE column is the calculated Shannon [89] entropy for a given residue position. The KL column is the Kullback-Leibler distance [90] of the observed amino-acid distribution at a particular residue position compare to the reference observed amino-acid distribution found in β-sheets in general. The following columns are the 4-class amino-acid partitioning used for the sequence analysis, and comes from the works of [92], Table II. Amino-acids are mentioned by their 1-letter code. Classes are cysteine and aromatics {CFYW}, large hydrophobic {MLIV}, small {GPATS} and large polar or charged {NHQEDRK}. Each amino-acid class has an absolute occurrence count, N, as well as a relative occurrence count, %. The E row is the reference amino-acid distribution as found in β-sheets of the culled PDB Select 25 [84]. Entries which differ significantly from the reference distribution are underlined and boldfaced.

1PGB	1OGW	1OEY	1NDD	1MG4	1EM7
1BT0	2IGD	1HZ6	1EUV	1OQQ	1EF1
1KH0	1H4R	1AYF	1A70	1FRR	1CZP
1IUE	1AWD	1M4V	1OFF	1BXT	1LM8
1DOI	1KRH	1ET9	1DFU	1FRD	1Q16
1KLU	1FNU	1ENF	3SEB		

Table 3.5 – The 34 solutions identified by the β-sheet descriptor of the Ubiquitin-like superfamily of fold on the culled PDB Select 90 database [84]. All of them are under the Ubiquitin-like superfamily (SCOP version 1.65 [56]) except PDB code 1DFU [104] which is a β-barrel (n=6, S=10).

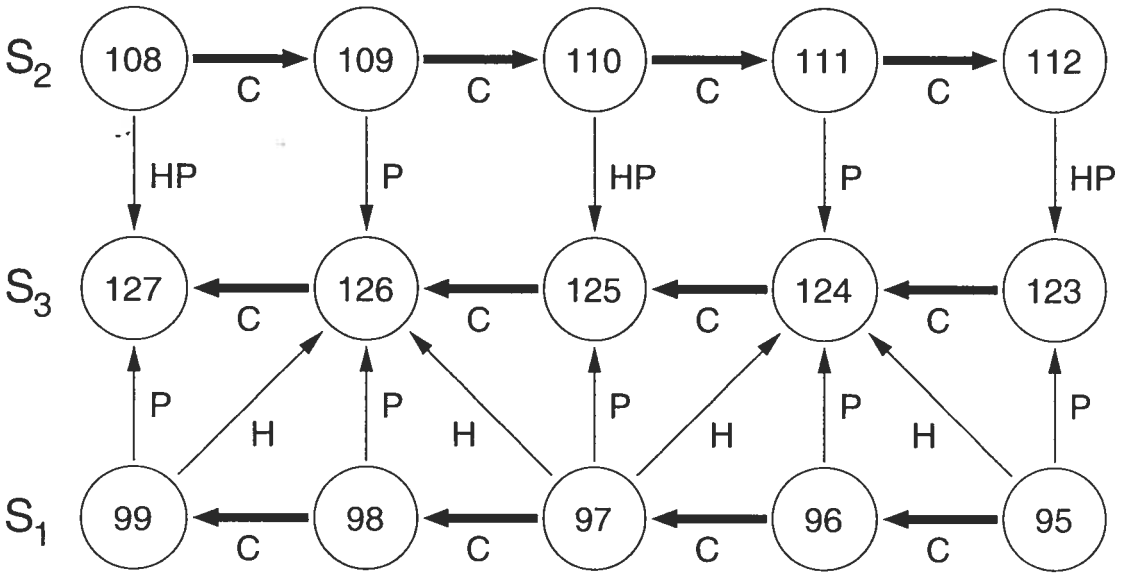


Figure 3.1 – Hypothetical β-sheet topological graph. Strands are S_1 $[95, 99]$, S_2 $[108, 112]$ and S_3 $[123, 127]$. Strand S_1 is parallel to strand S_3, while S_2 is anti-parallel to S_3. Notice the difference in the H-bonding patterns between the parallel strand pair and the anti-parallel one. The various topological relations **C**, **P**, **H** and **HP** are displayed in the picture.

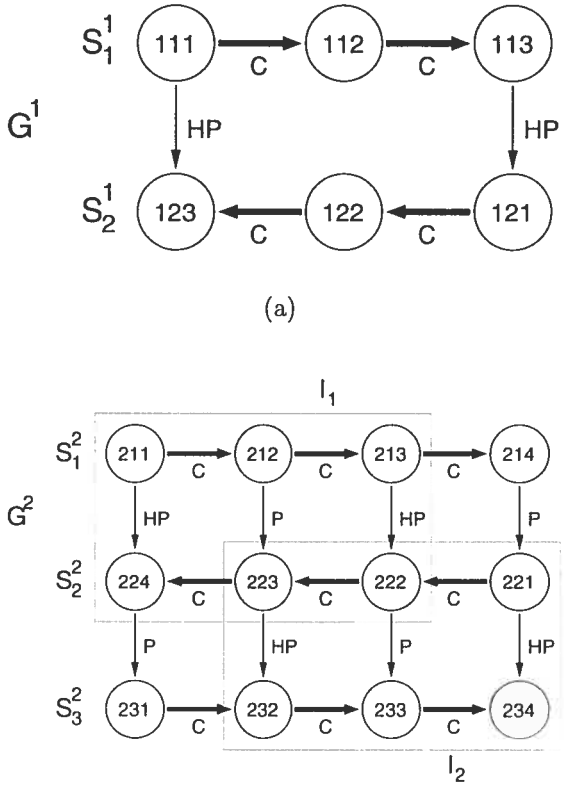
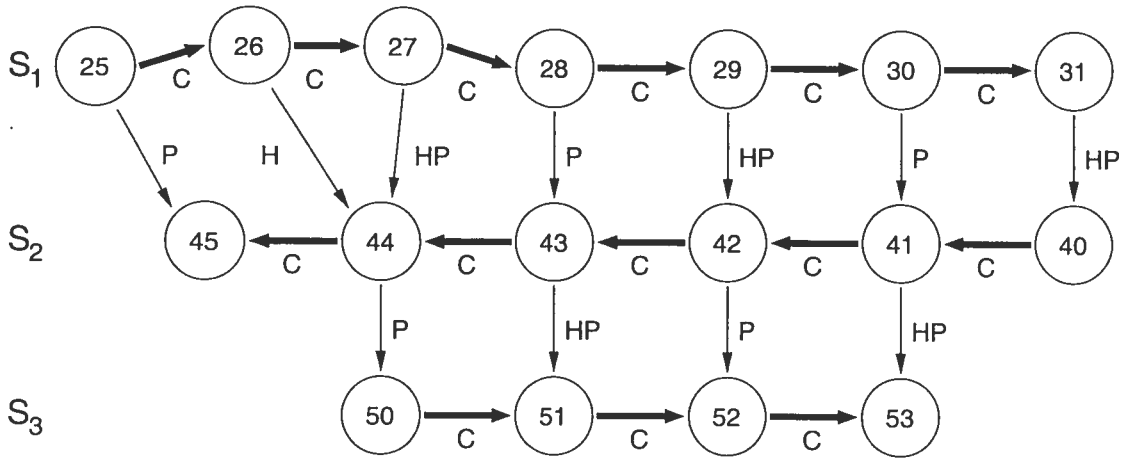
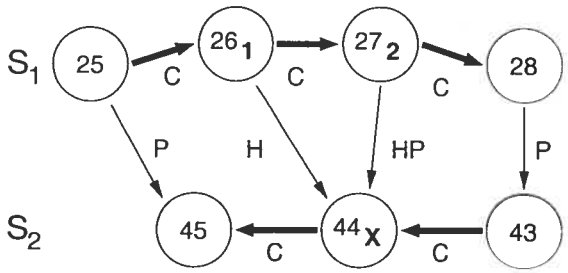


Figure 3.2 – The two β-sheet topology graphs, G^1 and G^2, used in Ullman's subgraph isomorphism algorithm. In S^i_j, the superscript i refers to the graph number while the subscript j refers to the strand number within the graph. **a)** The graph G^1 which serves as the β-sheet topology to search for in G^2. It is a two-stranded anti-parallel β-sheet with strands of length three. The partnership relation, **P**, between residues R^1_{112} and R^1_{122} has been omitted from the graph to show that it does not prevent a mapping in G^2. **b)** The graph G^2, a canonical three-stranded anti-parallel β-sheet with strands of length four. The two isomorphic solutions of G^1 in G^2, I^1 and I^2 are highlighted in Grey boxes.

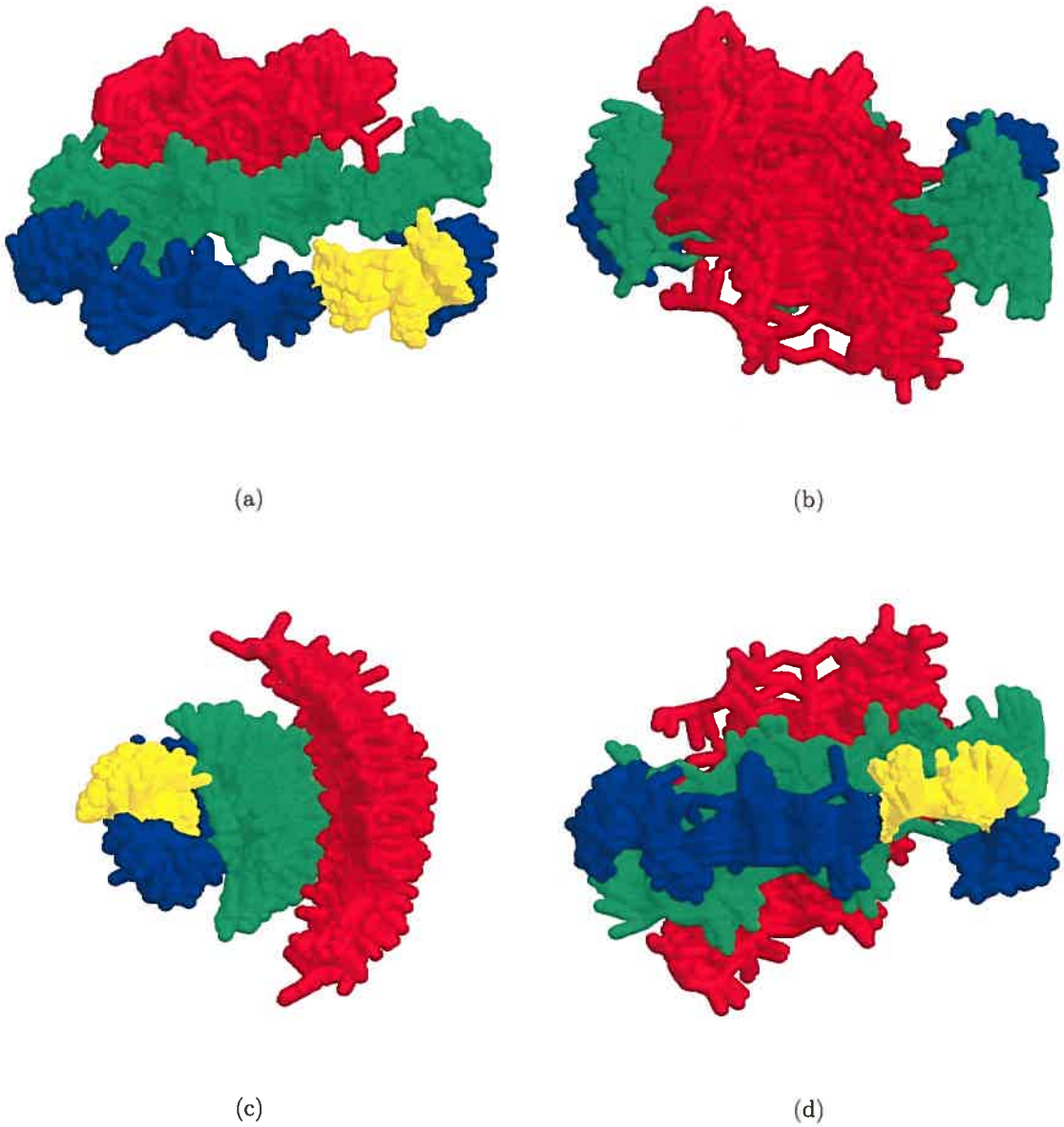


(a)



(b)

Figure 3.3 – Observed β-sheet topology graph of monocyte chemoattractant protein 1 (MCP-1) (PDB code 1DOK) [99, 100]. Topological features, namely covalent link **C**, β-sheet partnership **P** and H-bonding **H**, are calculated by the DSSP algorithm [20]. **a)** The complete β-sheet topology graph of MCP-1.**b)** The β-sheet topology graph of the C+ type β-bulge [102] used for sequence analysis of the C+ motif, and found in MCP-1. Subscripts i to residue sequence number R_i refer to Thornton's bulged residue nomenclature found in [102], Figure 2.



(a) (b)

(c) (d)

Figure 3.4 – The 75 superimposed models of the β-sheet topological graph of monocyte chemoattractant protein 1 (MCP-1) (PDB code 1DOK) [99,100]. The input β-sheet topology graph is as Figure 3.3a. These models are those of Table 3.3. The models have been superimposed on their first strand, which spawns residues 25 to 31. The two closest models have an RMSD of 0.38 Å, while the two farthest have 6.99 Å. Colors are blue for strand S_1, green for strand S_2 and red for strand S_3. The bulged residues, 26 and 27, are colored in yellow. The images were produced by RasMol [12]. **a)** Top view. **b)** Side view 1. **c)** Front view. **d)** Side view 2.

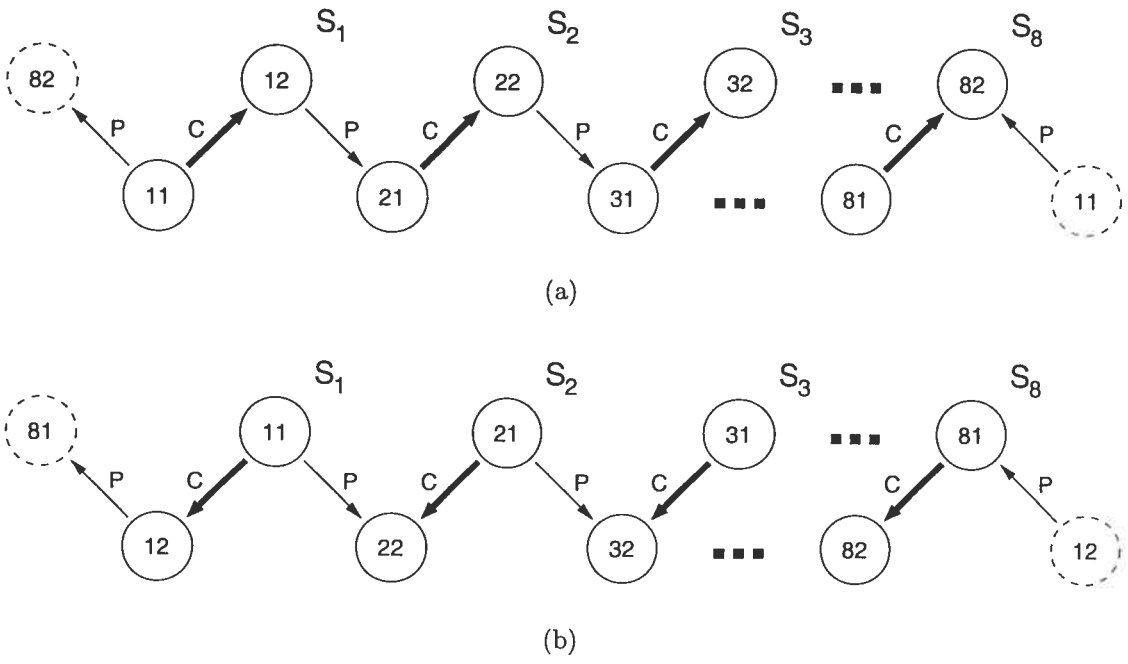


Figure 3.5 – β-barrel rings for β-strands with (n=8,S=8). In that configuration only two residues per strands are needed to express the ring in a topological graph. The dashed residues are used to show the connection between the first and last strands. **a)** A β-ring which would result in right-handed crossover connections for them to pass outside the β-barrel, which are the prevalent connection types observed in solved protein structures [4]. **b)** A β-ring which would result in left-handed crossover connections for them to pass outside the β-barrel. A right-handed connection would have to go by the inside of the barrel, which is impossible given the number of strands.

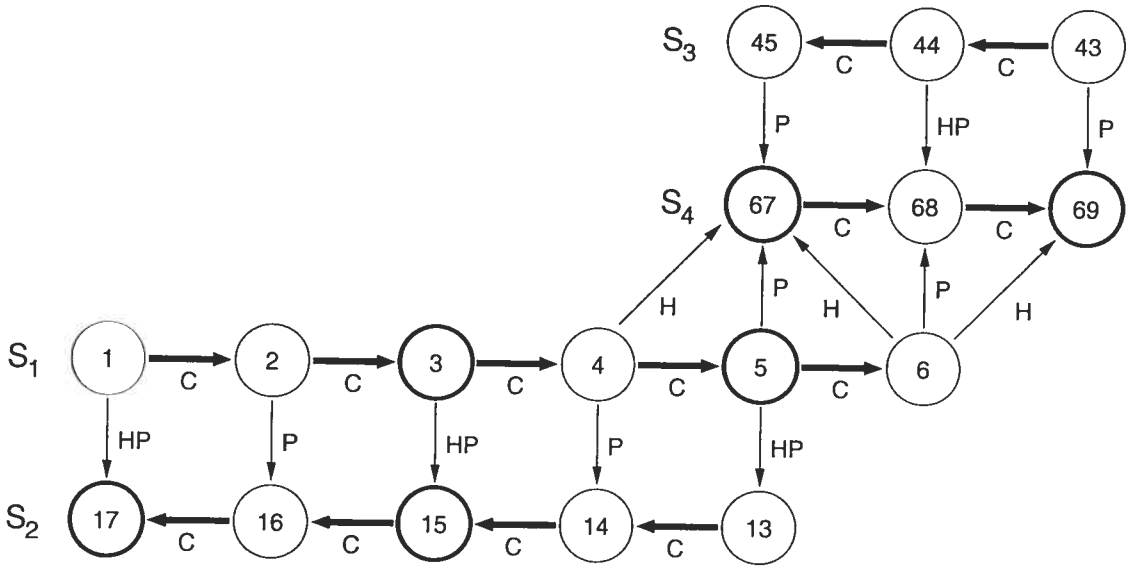


Figure 3.6 – β-sheet descriptor for the identification of the Ubiquitin-like superfamily of fold (SCOP version 1.65 [56]). This graph contains the folding nucleus residues as identified in [58], and are circled in bold. The residue numbering follows that of protein 1UBI [105].

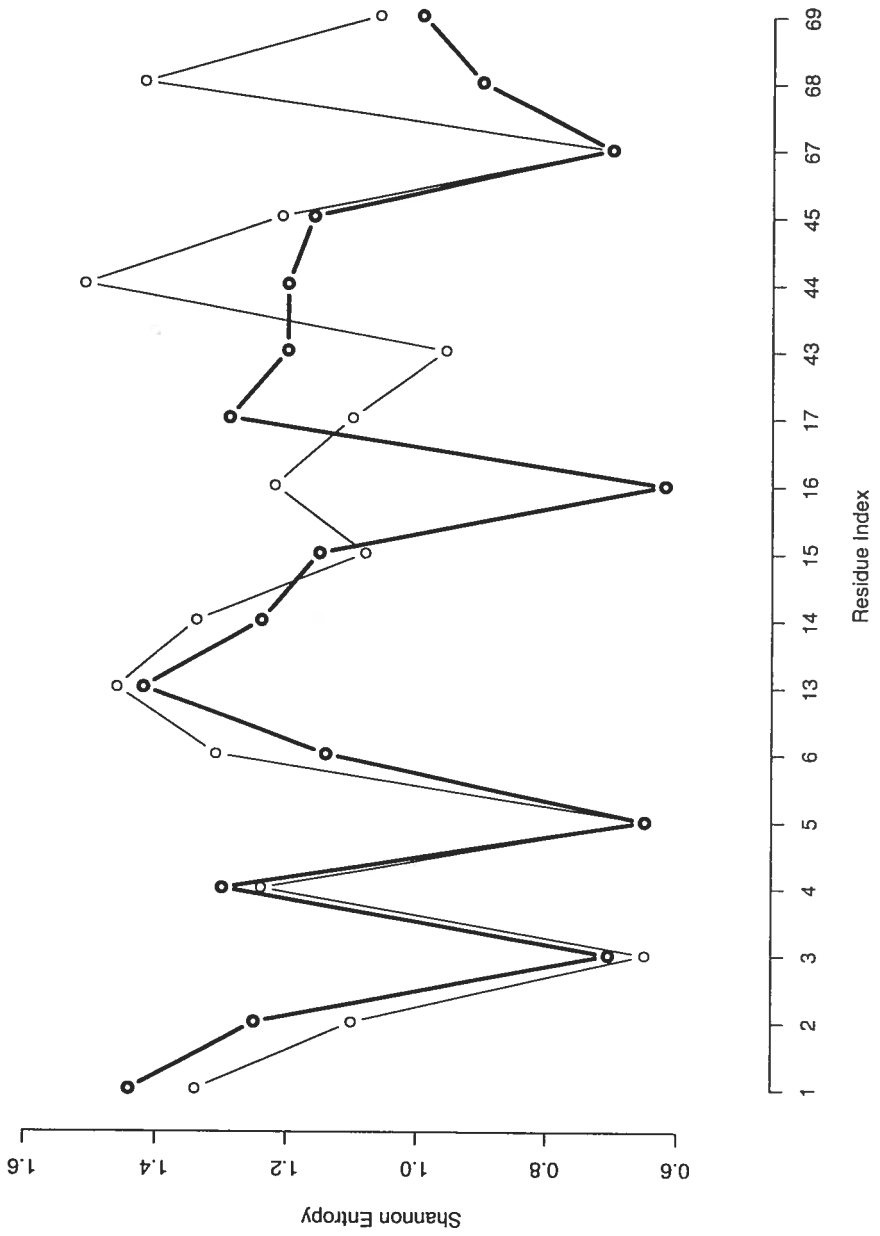


Figure 3.7 – Shannon entropy [89] of each residue position in the Ubiquitin-like β-sheet descriptor of Figure 3.6 for the 34 solutions found in the culled PDB Select 90 database [84]. The bold line is the entropy calculated with the 6-letter alphabet of [92] Table II ({CFYW}, {MLIV}, {G}, {P}, {ATS}, {NHQEDRK}). The thin line is the entropy calculated with the 6-letter code of [58] ({MLIV}, {FYW}, {AGPSTQN}, {HKR}, {DE}, {C}). Residue indexes are those found in protein 1UBI [105].

CHAPITRE 4

A β-SHEET CONFORMATIONAL SEARCH SPACE DEFINED BY β-SHEET TOPOLOGY GRAPHS

Marc Parisien and François Major
Département d'Informatique et de Recherche Opérationnelle,
Université de Montréal, CP 6128 Succ. Centre-Ville,
Montréal, Québec, Canada H3C 3J7

A β-sheet conformational search space defined as β-sheet topology graphs is presented. The β-sheet topology graph is an enhanced 2D descriptor in which H-bonds between adjacent strands are specified, thus enabling the expression of β-sheet defects like β-bulges and non-canonical H-bond ladders. A database of β-sheet 3D coordinates with their corresponding topology graph as been built, and used by a new computer program, the β-sheet builder, in a Jackknife [49] experiment to show that 1) 70% of β-sheets can be rebuilt within less than 3 Å of RMSD (precision), and 2) the extent of the physically permitable conformational flexibility of β-sheets (flexibility).

4.1 Introduction

β-sheets are present in 77% of proteins from the PDB select 25 [47] version of the Protein Data Bank [106]. When present, as much as 18% (standard deviation 15%) of the residues are found in the sheets. Thus, a proper β-sheet construction scheme is essential. Unlike helices, where a single tuple (ϕ,ψ) = (-60°,-40°) is sufficient to reproduce the 3.6 residues/turn ratio and the O_i to N_{i+4} backbone H-bonds [4], β-sheets exhibit far more complex three-dimensional spanning, as expressed by the curl of individual strands and the twist, pleat and arch in the assembly of strands into β-sheets [107]. Hence, a β-sheet construction algorithm is needed to address and handle all these features, including β-sheet defects like β-bulges and non-canonical H-bond ladders between adjacent strands. In hope of finding the native fold, the β-sheet 3D generative method should be precise, i.e. be able to spawn accurately native β-sheet folds, and flexible to explore the entire span of β-sheet conformational space.

β-sheet modeling has been attempted by using a simple mathematical description via helical surfaces [108]. Parallel strands can be created by a unique (ϕ,ψ) tuple, while anti-parallel strands, with their small (subscript s) and large (subscript l) H-bond rings, need (ϕ_l,ψ_l) and (ϕ_s,ψ_s) [109–111]. Homogeneous transformation matrices between (ϕ,ψ) = (-120°,+140°) extended strands is also an essay to β-sheet construction [14,15]. More refined procedures are FOLDTRAJ [112,113], which uses a Ramachandran probability distribution for each of three-state secondary structure in an off-lattice random walk. ROSETTA [114–117] proceeds by a simulated annealing on fragment libraries biased by a scoring function to drive strands into sheets. UNRES [118–121] combined with CSA [122–124] employ a sophisticated mix of genetic algorithm on united-residue random polypeptide chains and energy minimisation steps. LINUS [125–127] makes usage of a Metropolis Monte Carlo procedure in which extended chains are folded hierarchically, without any force field. TOUCHSTONE [128,129] uses an on-lattice Monte Carlo exploration guided by threading-based tertiary restraints.

The major underlying theme behind all these procedures is the formation of β-sheets by concatenation in 3D space of single strands. Also, many of them are unable to express β-sheet defects like β-bulges or non-canonical H-bond ladders between strands. Our method is a radical departure from previous cited methods in that we do not attempt to build the β-sheet by trying to assemble strands together but, instead, we exploit the network of inter-strand H-bonded residues, as observed in native protein folds, thus assembling the sheet in a strand-perpendicular fashion. The method of construction of β-sheets presented here is an effort to enhance the quality of 3D structures produced by MC-Sym [14, 15], and posseses both quality of precision and flexibility, as showned by our Jackknife experiment.

4.2 Methods

4.2.1 β-sheet topology graph

The β-sheet topology graph, $G = (V, E)$, is a directed planar graph where V (vertices) is the set of residues forming the β-sheet and E (edges) the set of relations encoding for covalent and inter-strand H-bonds. Edges originate from the N-terminal residue and end at the C-terminal residue. Edges are of four types : Type **C** is the covalent bond. Type **H** is the presence of at least 1 H-bond. Type **P** is a β-partner relation (without H-bond). Type **HP** is a β-partner relation plus the presence of at least an H-bond (i.e. type **H** + type **P**). The planarity of the β-sheet topology graph has a major argument ; crossings between either two covalent bonds, two H-bonds, or a covalent bond and an H-bond is sterically impossible, and β-sheets adopting the Moebius strip shape have not been observed yet. Example of β-sheet topology graph is given in Figure 4.1.

4.2.2 Sub-graph isomorphism algorithm

Graph isomorphism is defined as a one-to-one mapping between the vertices of graph G_1 and G_2. Sub-graph isomorphism is finding all isomorphisms of graph G_1 in sub-graphs of G_2. If graph G_1 is a sub-graph isomorph of graph G_2 then we say

that G_1 is included in G_2, and we write $G_1 \subseteq G_2$.

Sub-graph isomorphism is thought to be a NP-complete problem [86], and therefore a computationally expensive procedure. We decided to implement and adapt the original Ullmann's algorithm [19] because of it's algorithmic simplicity of its brute force approach. Direct utilisation of Ullmann's algorithm leads to exponential calculation time such that even small isomorphism problems with 20 vertices, or β-sheet residues, are beyond reach. Several optimisations were included in the algorithm for efficiency and problem size issues. Ullmann's algorithm starts by filling an $M \times N$ matrix M_o, where M is the size $|V_p|$ of the projected graph G_p and N the size $|V_h|$ of the host graph G_h. A 1 at position [i][j] in M_o indicates that, a-priori, considering only the out and in degrees of G_p and G_h for all types of edges, residue i of G_p can be mapped to residue j in G_h. Diagonals in this M_o matrix can be interpreted as the mapping of strands from G_p onto the strands of G_h. Thus, as a first optimisation, if a diagonal contains at least a 0, that is, there are no possible mapping of residue i of G_p on residue j of G_h, then we zero-out the full strand mapping in M_o, thus reducing the number of possible a-priori residue mapping. The next optimisation takes into account the fact that edges in the topology graph are from N-terminal to C-terminal; any partial permutation of graph mapping that leads to an edge from C-terminus to N-terminus is discarded, thus pruning entire sub-trees in the mapping search space. These two optimisations open the possibility of considering sub-graph isomorphism problems with graphs that contain hundreds of residues in few seconds of CPU time. An example of the application of the sub-graph isomorphism algorithm is given in Figure 4.2.

4.2.3 β-sheet databases

A database of all β-sheet backbone atoms 3D coordinates with their corresponding β-sheet topology graph has been built. Secondary structure assignments were made using the DSSP program [20]. Source proteins come from the PDB select 25 (april 2002 version). Residues from the same β-sheet are identified by DSSP, and thus our definition of β-sheet is the same as DSSP. The covalent type relation

(type **C**) is inferred from the consecutive PDB residue numbers. An H-bond (type **H**) is present if it's value is lower or equal to the arbitrary cut-off of -1.5 kcal/mol. H-bond energies are given by DSSP. The β-sheet partners relation (type **P**) are also given by DSSP. Finally, if type **H** and **P** are observed for a given pair of residues then type **HP** is used. The β-sheet database contains 2773 β-sheets. From the β-sheet database each pair of H-bonded strands are extracted to form the β-sheet strand pairs database. This database holds 8318 pairs of strands. One should note that pairs of strands are sub-β-sheets, and thus are also expressed in the topology graph paradigm presented here.

4.2.4 β-sheet Builder

A computer program has been made to explore the conformational flexibility of β-sheets. It takes as input a β-sheet topology graph, like the one as in Figure 4.1, and produces β-sheets, that respect all the prescribed relations (covalent bonds, H-bonds, β-sheet partners, etc.) from the input graph, as 3D atomic coordinates for all heavy backbone atoms. Here are the steps taken by the program from input to output :

Step I **Read the β-sheet topology graph.** The graph is encoded as an adjacency matrix, using the proper type definitions (**C**, **H**, **P**, **HP**) introduced earlier. The sequence is also specified for each amino-acid in the β-sheet for use in the output PDB files. For example, suppose that the β-sheet builder program has been given the graph of Figure 4.1 as input, which is a mixed parallel/anti-parallel β-sheet.

Step II **Split the β-sheet topology graph in pairs of H-bonded strands.** A β-sheet of N strands will have N-1 pairs of strands, except for the special case of β-barrels, where there are N pairs (note that β-barrels are not treated by the program). We thus obtain N-1 sub-graphs, G_i, one for each pair of strands. Figure 4.3 is an example of a split from the graph of Figure 4.1. One should note that each pair of strands is anchored to another pair by sharing a common strand. In a β-sheet of N strands there will be N-2 shared strands.

The strand S3 of Figure 4.3 stands as the common shared strand.

Step III **Read the database of β-sheet strand pairs. For each pair G_j read, use the modified sub-graph isomorphism algorithm to see if $G_i \subseteq G_j$.** When the condition $G_i \subseteq G_j$ is met, that means we can use the 3D atomic coordinates of the β-sheet strand pair G_j as a solution for the strand pair G_i. For each of the N-1 graphs G_i we collect all G_j, and their corresponding 3D atomic coordinates, such that $G_i \subseteq G_j$, thus filling a set $\{G_i\} = \{G_i^1, G_i^2, \ldots\}$ of 3D structures. A 3D structure for a given pair of strands G_i is inserted in the set $\{G_i\}$ only if it is different enough, in RMSD, from all the structures already in $\{G_i\}$. This ensures that the program does not waste time in building sheets that are similar in RMSD. The set sizes can also be fixed to an arbitrary value, say 100. For example, the program might assign a 3D structure for pair P1 of the graph of Figure 4.3 like the residues A103-A107 and A194-A198 of PDB code 1KLQ (see Figure 4.4a). The 3D structure of pair P2 could be like the residues G184-G188 and G141-G145 of PDB code 1QSG (see Figure 4.4b).

Step IV **Backtrack to generate all possible $\{G_1\} \times \{G_2\} \times \cdots \times \{G_{N-1}\}$ 3D configurations.** Given that the strand pairs G_i and G_{i+1} share a common strand, each configuration will require N-2 strand pair fusion, namely G_1 with G_2, G_2 with G_3, $\cdots$, G_{N-2} with G_{N-1}. Once all strand pairs of the input β-sheet have been assembled, the final 3D coordinates are printed in the PDB file format. The fusion process needs further explanations : for example, if the program proceeds to a fusion of pair P1 to the pair P2 of Figure 4.3, it first aligns S3 of P1 onto S3 of P2. The alignment that uses both 3D structures selected at Step III, is a procedure in which a spatial transformation is calculated to minimize the RMSD of all implicated atoms. The assembly is rejected if the RMSD of the aligned strands is higher than a threshold. Here, the RMSD of the alignment is 0.48 Å (C_α only). The final assembly has too many residues in S3 ; those from P1 and those from P2. We keep in the final structure the pair of residues that are H-bonded. Thus, the

following residues are kept to form the strand S3, residues 123 to 127 are, respectively, A194, G185, A196, G187, A198 (residues from chain A are from PDB code 1KLQ, those on chain G are from 1QSG). The final assembly is showned in Figure 4.5. Residues can be renumbered to the specifications of the final graph (Figure 4.1). Since the fusion process is fairly CPU intensive and that it is done within a backtrack, this step is by far the most time consuming.

4.2.5 Peptidic bond

The common strand fusion process of the β-sheet builder introduces deformation of the peptidic bonds along the resulting fused strand. The deformation affects the peptidic bond length and valence and torsion angles. For example, in Figure 4.5, peptide bond lengths are 1.16 Å from A194 to G185, 1.04 Å from G185 to A196, 1.60 Å from A196 to G187 and 1.35 Å from G187 to A198. The β-sheet program defines a cut-off value over which the RMSD of the resulting fused 3D structure is rejected, thus controling the departure from ideal peptidic bond length, valence and torsion angles. An energy minimisation step should be performed once all side-chains have been added to the β-sheet along with the loops and helices to complete the protein model.

4.2.6 Jackknife

The β-sheet database has a total of 1687 β-sheets in which we can find at least 10 residues. All these sheets where subject to reconstruction using the β-sheet builder under a Jackknife condition [49]. The Jackknife forbids the usage of the 3D structure of the β-sheet X when trying to rebuild it, otherwise we would always have a perfect reconstruction. The quality of common strand alignment RMSD has been fixed to 0.5 Å. A time limit of 1 hour is allocated to the backtrack algorithm (Step IV). Backtrack set $\{G_i\}$ sizes where limited to 100; the crystal structure was not used to select candidate 3D coordinates, instead, the first 100

structures selected by the sub-graph isomorphism algorithm were used. This has for consequence that the search tree has at most $10^{2\times(N-1)}$ leaves for a β-sheet with N strands. 3D structures in a given backtrack set $\{G_i\}$ where at least 0.65 Å from each other. The database contains 7783 pairs of strands. As an indication of the conformational space addressed by the β-sheet builder and its database, the worst and best RMSD (all heavy backbone atoms), with respect to the crystal structure, for each β-sheet reconstruction has been kept. Since the backtrack search tree sizes have been reduced from many orders of magnitude for time concerns, a procedure which would not mimic the use of the program in the context of an unknown β-sheet, the best RMSD values could, in theory, be improved, and the worst RMSD worsened. The best RMSD is a measure of how precise (Figure 4.6a) is the β-sheet builder and its database. It is also an indication of how similar are pairs of strands in PDB Select 25 under the Jackknife condition. On the other hand, the worst RMSD is a measure of the flexibility (Figure 4.6b) of β-sheets encoded in pairs of strands, the deviation from the crystal structure is solely 3D since the β-sheet builder always produce β-sheets from the same topology graph, i.e. preserving the relation types **C**, **P**, **H** and **HP**, between pairs of residues.

4.3 Results And Discussion

4.3.1 Rebuilding β-Sheets of the PDB

A total of 6 CPU days were needed to complete the Jackknife task on Intel Pentium III class running at 650 Mhz. From the 1687 β-sheets, 1190 (70.6%) have at least 1 solution. From these, 23 (1.4%) reached the 1 hour run-time limit, but were able to produce solutions. Mean calculation time is 3 minutes 20 seconds (with standard deviation of 9 minutes 30 seconds). 1174 (69.6%) have a best RMSD under 3.0 Å while 16 (0.9%) have a best RMSD greater than 3.0 Å. Thus, 99% of the β-sheets that had solutions are under 3.0 Å from the crystal structure. The mean best RMSD is 0.97 Å with 0.57 Å of standard deviation. A visual inspection of the crystal structures for which the best RMSD was higher than 3.0 Å point to the

fact that several strands have a pronounced (within 1 residue) orientation change greater than 60° (some up to 90°), as in the case of PDB 1BY2 residues 16 to 23, 1FGY residues 267A to 274A, 1NBC residues 74A to 85A, 1HZT residues 62A to 68A, 3VUB residues 30 to 38, 1K0S residues 97A to 105A, 1FJR residues 163A to 171A, 1GLV residues 67 to 76 and 1DFM residues 82A to 87A. Figure 4.7a shows an example of a sharp orientation change. The greatest best RMSD, 5.76 Å, is for the β-sheet of PDB code 1A8D between residues 275 and 424. 3CLA has a strange pair of strands which seems to be partially unzipped (residues 90 to 97 and 144 to 150), as in Figure 4.7b. The Figure 4.9 shows a typical example of precision of the rebuilding method.

Figure 4.8 shows the distribution of best RMSD for the 1190 rebuilt β-sheets under the Jackknife condition. As most as 90% of the distribution lies within the upper bound value of 1.5 Å. 99% of the distribution is reached at 3.0 Å.

No solutions were obtained for 497 (29.4%) β-sheets. Several reasons explain why some β-sheets cannot be rebuilt :

1. A pair of strands in the β-sheet is unique, either by the length of the strands, inter-strand H-bond pattern, β-bulges, etc. This yields a backtrack tree size of 0, since this pair will have no other associated 3D structure. 168 (10.0%) β-sheets have unique pairs in the database.
2. A pair of strands in the β-sheet is rare, thus resulting in few 3D structures for that pair and eventually leading to zero solutions. 246 (14.6%) of β-sheets have at least a particular pair.
3. The backtrack algorithm was allowed a 1 hour run-time. 1 (0.06%) β-sheet didn't produce solutions within the 1 hour limit.
4. The β-sheet builder does not build β-barrels. 82 (4.9%) β-sheets in the database are β-barrels.

4.3.2 Rebuilding a β-Sheet of Novel Topology

In order to asses the power of our method we have proceeded to the rebuilding of a β-sheet of novel topology from a protein called Top7 [98]. Indeed, we have not found in our database of β-sheets this topology. Since the β-sheet builder decomposes the whole sheet to rebuild in pairs of strands, the global topology of the sheet does not restrict the performances of the builder. Figure 4.10 shows the 579 three-dimensional models for the β-sheet of the Top7 protein. The closest solution from the crystal structure is 1.00Å. Each solution is different from one another of at least 1Å. The initial backtrack search tree size was 8×10^{13} and took approximately 13 hours to complete on an AMD64 class running at 2.2 gigahertz.

4.4 Conclusion

A database of all β-sheets backbone heavy atom 3D atomic coordinates along with their corresponding β-sheet topology graph has been made. From it, a database of all β-sheet strand pairs backbone heavy atom 3D atomic coordinates along with their corresponding β-sheet topology graph has also been made.

A β-sheet topology graph is then fed to the β-sheet builder program. The β-sheet topology graph is first divided in pairs of adjacent strands. Each pair, which are themselves graphs, is then associated to a backtrack set of 3D structures that are selected through the modified Ullmann's sub-graph isomorphism algorithm. With the help of a backtrack algorithm, the 3D structures of each pairs are assembled in space along their common strand. A cut-off quality parameter can be specified to reject assemblies that have a high RMSD between the 3D solutions of the shared strand. Special care is taken to preserve the correct H-bond ladder between adjacent strands by retaining H-bonded residues of each pair to form a unique common strand.

For PDB select 25, 70% of β-sheets, with 10 or more residues, have been rebuilt under the Jackknife condition, of which 90% are under 1.5 Å from the crystal structure. Thus, our method of construction seems appropriately accurate for *De*

Novo protein construction and sketching. On the other hand, the worst RMSD Figures indicate the extent of the conformational search space addressed by the method.

4.5 Acknowledgments

We thank L. Brehelin, P. Gendron and S. Oldziej for reviewing this manuscript and for interesting discussions. This work is funded by the Canadian Institutes of Health Research (CIHR). FM is a CIHR investigator.

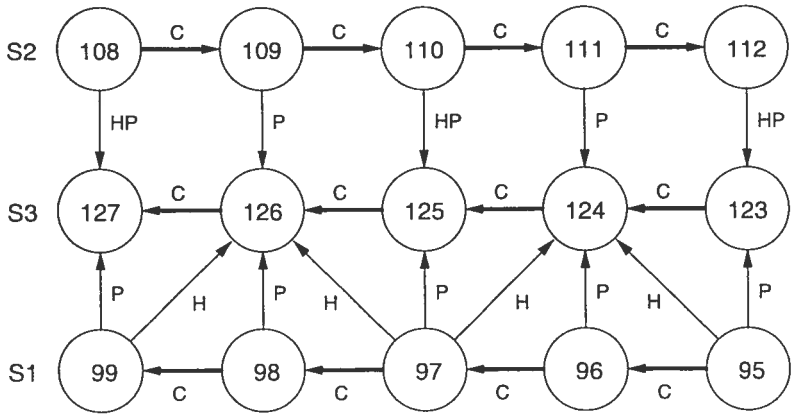


Figure 4.1 – Hypothetical β-sheet topology graph. Strands are S1 [95, 99], S2 [108, 112] and S3 [123, 127]. Strand S1 is parallel to S3 while S2 is anti-parallel to S3. Covalent bonds are depicted with type **C** links, H-bonds with type **H**, β-partner with type **P** and type **HP** for type **H** + type **P**. Notice the difference in the H-bond patterns between the parallel strands and the anti-parallel ones.

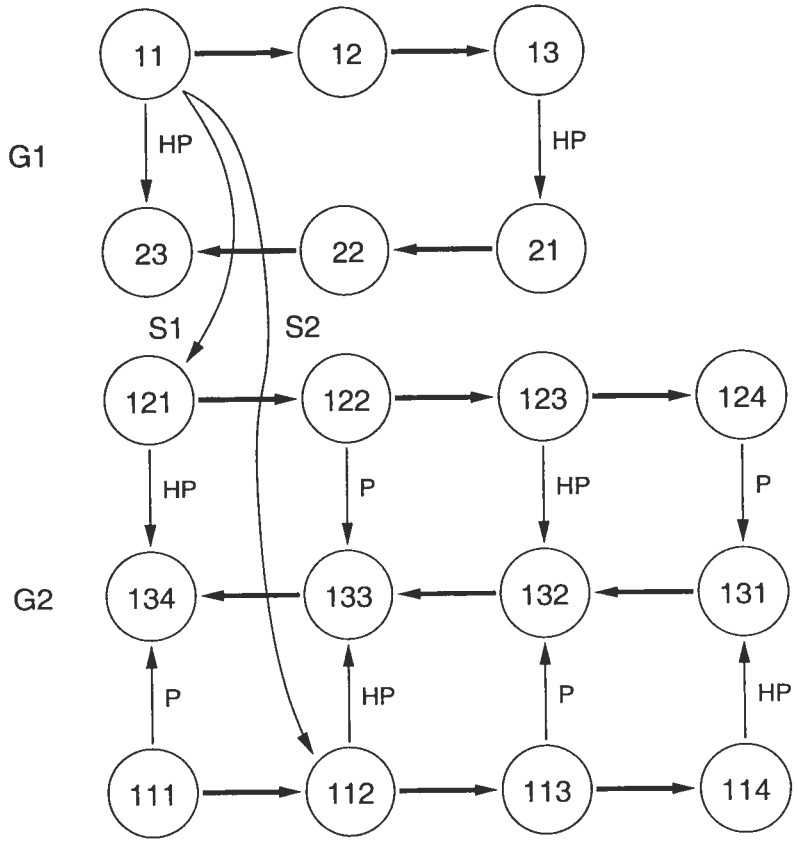


Figure 4.2 – Graph G_1 has 2 isomorphic solutions in G_2. Solution 1 maps residues {11, 12, 13, 21, 22, 23} of G_1 onto residues {121, 122, 123, 132, 133, 134} of G_2 respectively. Solution 2 maps residues {11, 12, 13, 21, 22, 23} of G_1 onto residues {112, 113, 114, 131, 132, 133} of G_2 respectively. Note that vertices of type **P** in G_2 do not prevent the isomorphism algorithm of finding solutions.

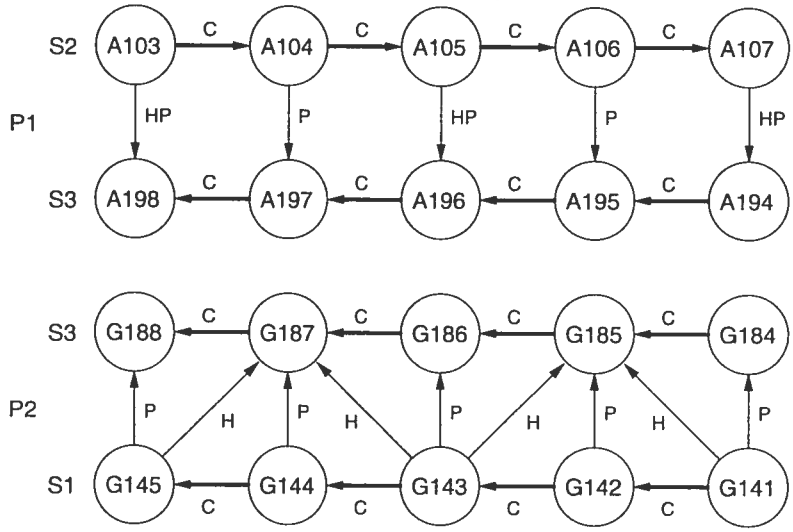
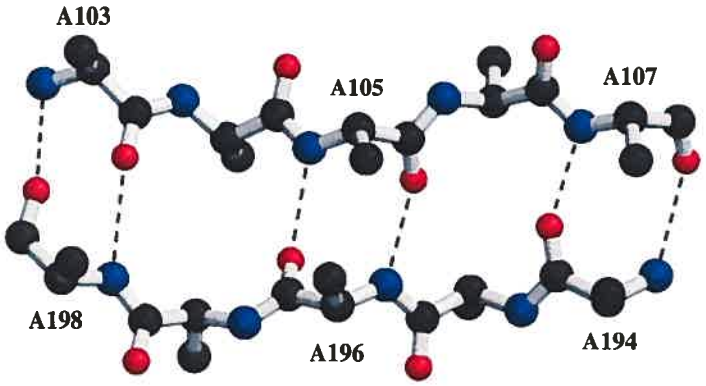
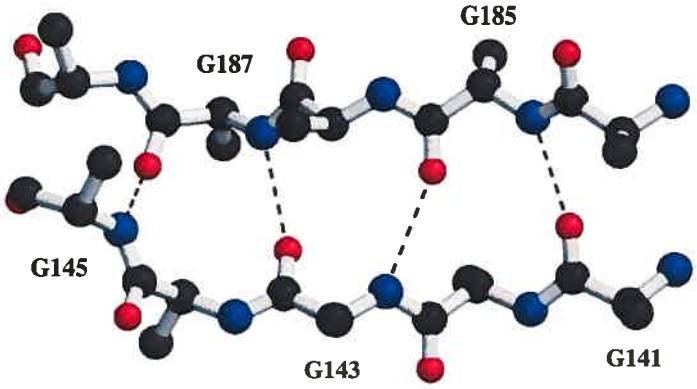


Figure 4.3 – The split of the β-sheet topology graph of Figure 4.1 as described in step II. Two pairs of strands, P1 and P2, are created. Pairs P1 and P2 share a common strand, S3.



(a)



(b)

Figure 4.4 – 3D structures associated to the β-sheet topology graphs. H-bonds are depicted with dashed lines. Atomic colouring scheme is as follow : blue for nitrogen, red for oxigen and black for carbon. Figures were produced with Molscript [130] and Raster3D [131]. **a)** The 3D structure of the pair P1 chosen at step III. Residues A103-A107 and A194-A198 are from PDB code 1KLQ. **b)** The 3D structure of the pair P2 chosen at step III. Residues G184-G188 and G141-G145 are from PDB code 1QSG.

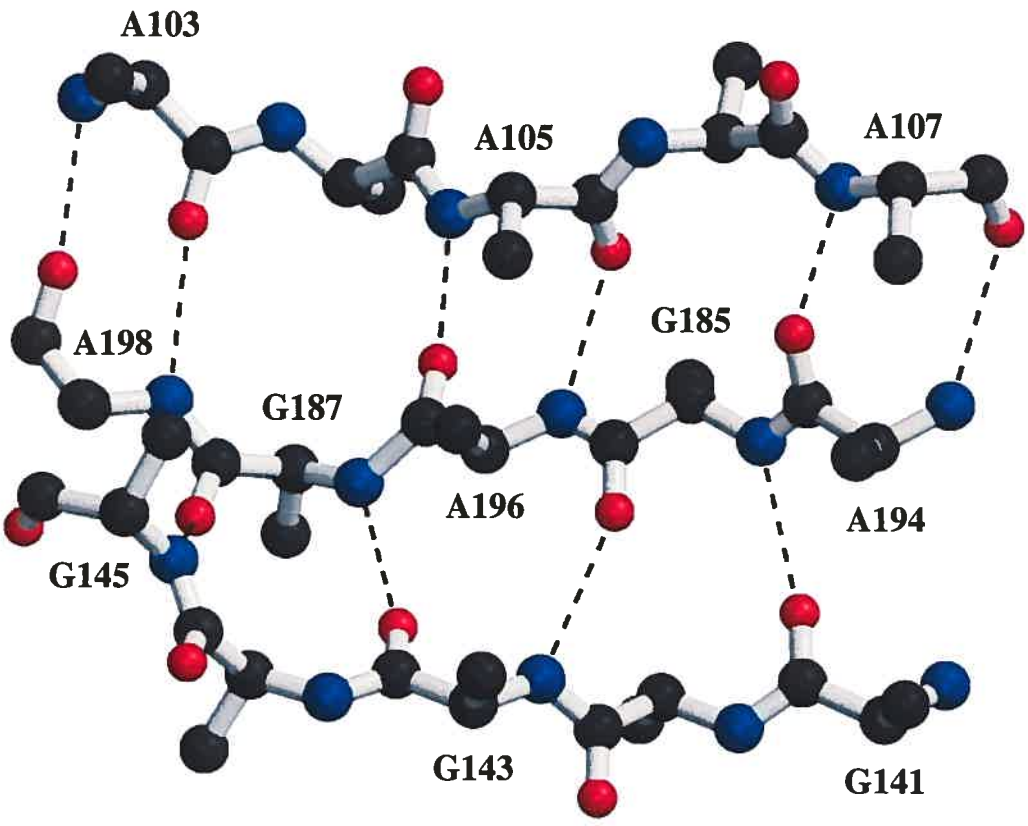
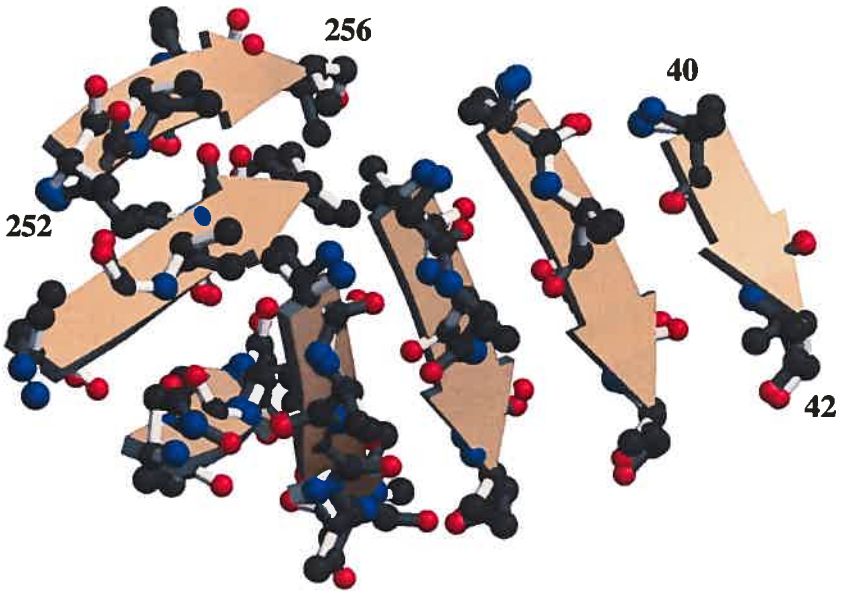
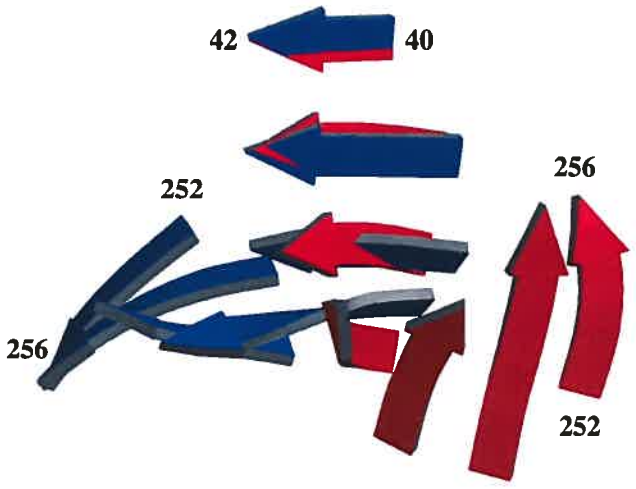


Figure 4.5 – The 3D structure of the β-sheet after of the final assembly, following step IV.



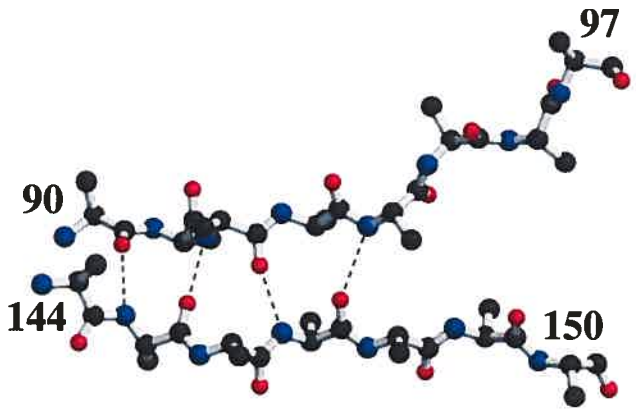
(a)



(b)

Figure 4.6 – Display of the precision and flexibility of the β-sheet builder. Target β-sheet is from PDB code 1TML, between residues 40 and 256. This β-sheet has 30 residues distributed in 7 strands. **a)** Example of precision. Crystal structure has light grey cylinders while best RMSD (0.83 Å) structure has dark grey ones. Strand ribbons are pictured for the crystal structure. **b)** Example of flexibility. Crystal structure is in red while the worst RMSD rebuilt structure is in blue. Both structures are aligned along the strand 40 to 42. The high RMSD (7.88 Å) comes from the fact that the rebuilt structure chooses a different path as soon as the third strand from the top.

(a)



(b)

Figure 4.7 – Two examples of β-sheet irregularities that account for high RMSD in rebuilt structures. **a)** Side view of strand residues 16 to 23 of PDB code 1BY2. This strand has a sharp orientation change at residue 18. **b)** Top view of strands residues 90 to 97 and 144 to 150 of PDB code 3CLA. The strands seem partially unzipped, as depicted by the interruption of the H-bond ladder (dashed lines).

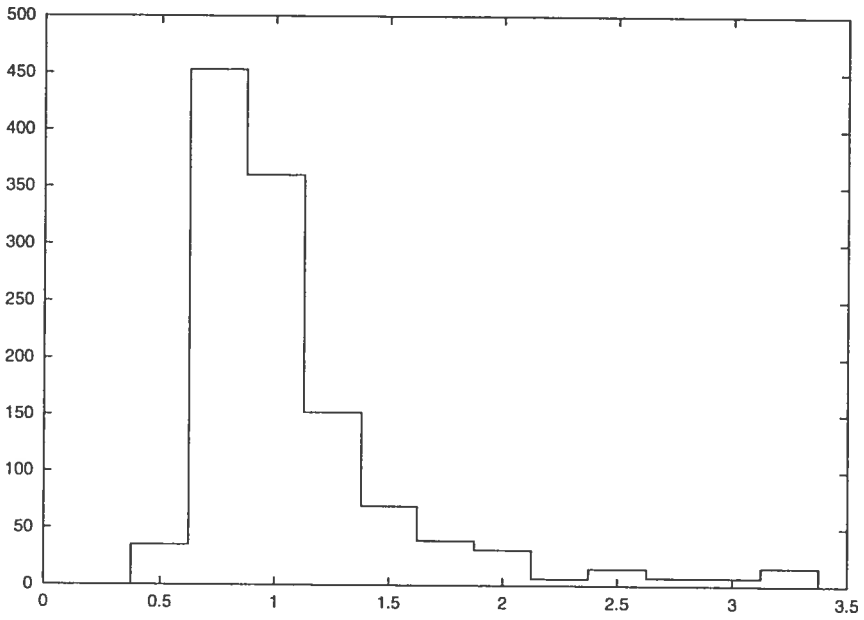


Figure 4.8 – Distribution of the best RMSD for the 1190 rebuilt β-sheets under the Jackknife condition. Histogram classes are 0.25 Å wide. Values in abscissa are the upper bound for each class. The last class at 3.25 Å contains the count of β-sheets with best RMSD $>$ 3.00 Å. As most as 90% of the distribution lies within the upper bound value of 1.5 Å. 99% of the distribution is reached at 3.0 Å

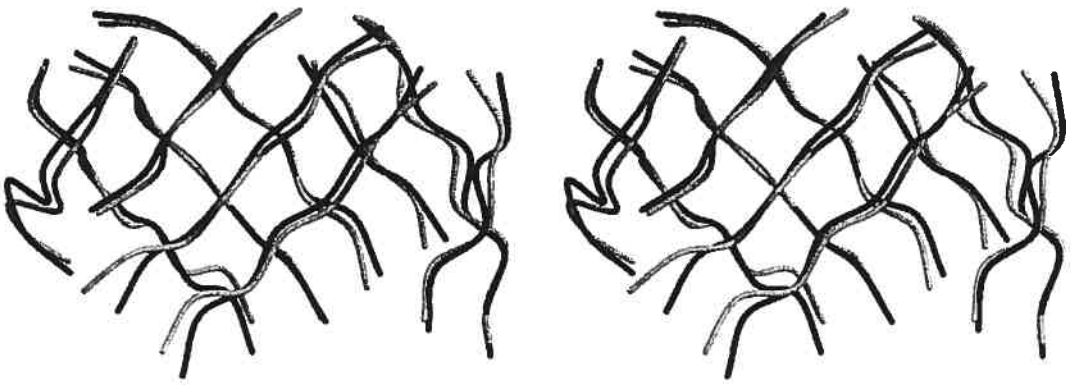
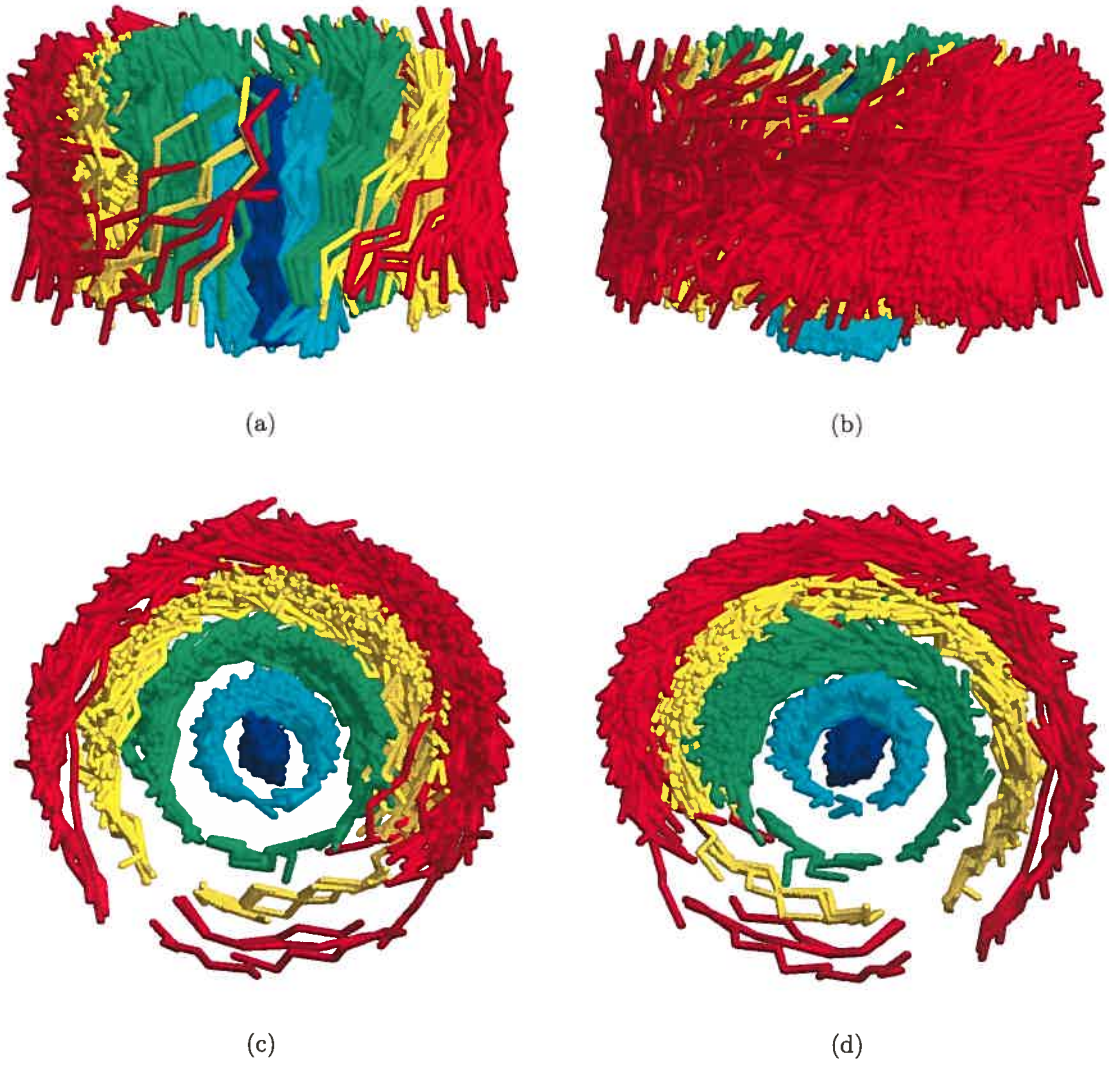


Figure 4.9 – Stereo view of the C_α trace of a rebuilt β-sheet using the Jackknife method. The β-sheet is from PDB code 1EIO. The β-sheet is composed of 74 residues distributed in 10 strands, in a quasi-barrel fashion. The crystal structure is in black while the rebuilt structure is in light grey. The RMSD is 1.34 Å for all heavy backbone atoms. The Figure was produced with the help of Molscript [130] and the strereo3d program of the Raster3D [131] package.



(a) (b)

(c) (d)

Figure 4.10 – 3-D models of the β-sheet in protein Top7. The 579 solutions obtained by the β-sheet builder are shown in α-carbon traces. Each sheet is distant from one another of at least 1Å. Using the residue numbering scheme of PDB file 1QYS, the strands are 6-12 in cyan, 15-21 in blue, 47-53 in green, 78-84 in red and 87-93 in yellow. All β-sheets are aligned on the strand 15-21, which is the N-terminal border strand. The other border strand, the C-terminal 78-84, is in red. **a)** A view from the top where the N-terminal strand can be viewed. **b)** A view from the top, 180° from a. **c)** A front view, where residue 15 is closest to the viewer. **d)** A front view, where residue 21 is closest to the viewer.

CHAPITRE 5

A NEW CATALOG OF PROTEIN β-SHEETS

Marc Parisien and François Major
Département d'Informatique et de Recherche Opérationnelle,
Université de Montréal, CP 6128 Succ. Centre-Ville,
Montréal, Québec, Canada H3C 3J7

Systematic protein folding studies depend on protein three-dimensional structure annotation, the assignment of amino acid structural types from atomic coordinates. Significant stabilizing factors between adjacent β-sheet peptide chains have recently been characterized and were not considered during the development of previously published annotation methods. To produce an accurate β-sheet domain catalog and to encompass the full β-sheet spectacle, we developed a method, β-Spider, which evaluates a packing energy between adjacent peptide chains in accordance with the newly discovered stabilizing factors. While considering important energetic factors, our approach also minimizes the use of subjective criteria, such as (ϕ, ψ) boundaries and sets of H-bonding motifs that are used in other existing methods. As a result of the application of β-Spider to a set of available high-resolution X-ray crystal structures, we present here a new β-sheet catalog that differs considerably from the one produced by the most acclaimed DSSP method. The catalog includes new H-bonding motifs that were never reported.
M Parisien and F Major. A new catalog of protein β-sheets. *Proteins*, 61 :545-558, 2005.

5.1 Introduction

Accurate secondary structure assignment (annotation) from protein atomic coordinates is of utmost importance for the understanding of the protein folding phenomenon. In the hopes that one day secondary structure prediction from sequence data will be a problem of the past, secondary structure annotation is widely used in the development of secondary structure prediction methods [132–135] (cf. *Ab Initio* methods, such as UNRES [118–120] and ROSETTA [114–116], and protein classification systems, such as CATH [136,137] and SCOP [56]).

Particularly difficult is the annotation of β-sheets. The formation of hydrogen bonds (H-bonds) between the backbone C=O and N-H groups has led to the theoretical prediction by Pauling and Corey [11] of β-sheets, a regular network of H-bonds between adjacent peptide chains. The prediction of β-sheets was proposed shortly after the prediction, by the same group and Branson [2], of two H-bonding motifs within a single chain : the 3.7-residue helix (α-helix) and the 5.1-residue helix (π-helix). The β-sheet model has been confirmed experimentally more than a decade after its publication [138], but the model did not include the twist and shear that were observed in β-sheet folds.

Among the several attempts to delineate protein β-sheet regions from atomic coordinates, one can find purely geometrical methods based on constraints defined by C_α-C_α distances and (ϕ, ψ) patterns, such as DEFINE [139], P-SEA [140], xtlsstr [141], STICK [142] and VoTAP [143], mathematical approaches such as P-Curve [144], pattern matching methods based on the identification of geometrical H-bonding templates, such as DSSP [20] and DSSPcont [145,146], or hybrid methods based on both geometrical constraints and H-bonding templates, such as STRIDE [147], in which the authors use the (ϕ,ψ) angles as a geometric guide.

Protein β-sheets exhibit a great geometrical flexibility [83], expressed in the sheet twist and strand shear [4,82] and curl. Their precise annotation is thus far more complex than that of α-helices, and is more subject to discordant and inaccurate findings [148]. β-sheets depart from planarity, and thus may be invisible to

automated identification methods that rely solely on geometrical features. As a result, currently available secondary structure annotation methods produce improper β-sheet information, as they yield to fragmented domains and erroneous edges, and they often even miss complete β-sheets. Automated β-sheet annotation methods are so inaccurate that many researchers avoid their use in protein structure studies. The DSSP algorithm was explicitely mentionned as the secondary structure determination method in only 14 out of the 811 protein chains in the culled PDB Select 25 database.

The individual amino-acid β-sheet propensity has been shown to be due to local steric interactions between the side and main chains [149], which set the backbone ϕ and ψ angles so that their carbonyl oxygen and amide hydrogen atoms get exposed to the formation of H-bonds. The distribution of amino-acids in β-sheet regions identified with the canonical H-bonding motifs differs from that of amino-acids in proteins [96, 97], which is the foundation of the Chou-Fasman secondary structure prediction algorithm [132]. However, the inclusion of non-canonical H-bonding motifs in β-sheet regions steers the distribution of amino-acids in β-sheets towards that of amino-acids in proteins, which profoundly impacts β-sheet prediction from sequence, and partially explains the relative failure of the Chou-Fasman and alike algorithms.

The results of recent studies indicated that more than the backbone H-bonds contributes to the stability of adjacent peptide chains in the formation of a β-sheet. Notably, the C=O$\cdots$C=O electrostatic dipole energy has been estimated to be at the same order of magnitude than that of an H-bond [21, 150]. The bifurcated C_α-$H_\alpha\cdots$O=C H-bonds also contribute to the stabilization of β-sheets [82, 151–155]. In this regard, a significant proportion of the inter-chain backbone energies are not considered in the electrostatic H-bond terms employed in most β-sheet automated annotation methods.

Three hydrogen bonding patterns defined by Pauling et al. [3] (see Figure 5.1) are largely used in the identification of β-bridges, and therefore used in almost all automated β-sheets labeling methods. For instance, the program STRIDE [147], in

addition to identifying these three patterns, considers the anti-parallel bridges of type III H-bonding pattern (see Figures 5.4c and 5.4d in the Frishman and Argos article [147]).

In general, when at least two inter-chain backbone H-bonds are found, one can conclude in the formation of β-sheets. A problem is that many of these patterns contain less than two H-bonds, and others, even though sharing at least two H-bonds, do not match the canonical templates of largely used programs such as DSSP and STRIDE. The usage of a better H-bond definition, as proposed by Wade and coworkers [156] and used in STRIDE, is not the solution as it does not account for the C=O$\cdots$C=O and C_α-$H_\alpha\cdots$O=C binding forces.

We introduce here a new algorithm, β-Spider, which detects protein β-sheet domains from protein atomic coordinates. During the development of β-Spider, we considered the combination of simple geometrical criteria, in addition to a non-bonded energy evaluation that verifies the β-sheet formation propensity of two adjacent peptide chains. The algorithm does not rely on (ϕ,ψ) calculations nor on canonical H-bonding template matching (such as DSSP). The algorithm is implemented in a computer program which processes input PDB formatted files to output the PDB SHEET records that correspond to the findings of the algorithm. An online version is available at : http ://www-lbit.iro.umontreal.ca/bSpider/.

5.2 Results and Discussion

In the development of the β-Spider algorithm, we considered three parameters to decide whether a residue is part of a β-sheet or not : 1) the C_α^i-C_α^j distance (residue centers not too far from each other), 2) the C_β^i-C_α^i-C_α^j-C_β^j torsion angle (residue side-chains point in the same direction), and 3) the non-bonded energy between the adjacent peptide chains (the chains are glued together). The cut-off values for these parameters were chosen at the 99.5 percentile of their distribution, plus an additional 5% to capture the essential sample population. Although each cut-off value is an extremum, the concerted usage of the three provides useful the

identification of beta-sheet partners and domains.

The results described and discussed below are independent of the X-ray resolution, as we applied the method to six different resolution structures. Worth noting is that both computer programs, DSSP and β-Spider, exhibit the same trends as the X-ray resolution varies (see Supplementary Materials). Here, to simplify the discussions, we built a database of the most stringent structures, those at 1.6 Å resolution, and we compared the results obtained with DSSP and β-Spider.

We define that a β-sheet identified by DSSP is mapped by β-Spider if at least one of its residues is found in a β-sheet identified by β-Spider. Using this definition, all DSSP β-sheets are mapped by β-Spider. Similarly, at the strand level, all but one DSSP β-strands are mapped by β-Spider (see residues 16 and 17 in 1MJ4 chain A). Finally, at the residue level, 84 DSSP β-sheet residues are not mapped by β-Spider (see Table VI in Supplementary Materials).

The geometrical cut-off parameters were chosen from their distributions in the three canonical H-bonding motifs I, II and III (see the motif definitions in Figure 5.1, and the distributions in Figure 5.2). For the C^i_α-C^j_α distances we have the means $\mu_I = 4.5$Å, $\mu_{II} = 5.2$Å, and $\mu_{III} = 4.8$Å. The cut-off value was fixed at 6.2Å, which is Z = 4.3 standard deviations from the distribution that has the greatest mean (distribution II) (see Figure 5.2a). For the C^i_β-C^i_α-C^j_α-C^j_β torsion angles we have the means $\mu_I = -36.5$ degrees, $\mu_I = -10.6$, and $\mu_I = -22.9$. The cut-off values were fixed at -128.0 and +68.4 degrees, and the corresponding Z-scores are -3.6 and 4.3 (see Figure 5.2b).

The energy cut-off value of β-Spider has also been chosen from the distributions of the Coulomb electrostatic and van der Waals energies found in the three canonical H-bonding motifs I, II and III (see the motif definitions in Figure 5.1, and the distributions in Figure 5.2). The major energy contribution comes from the Coulomb electrostatics (means $\mu_I = -13.9$, $\mu_{II} = -11.3$ and $\mu_{III} = -12.1$ Kcal/mol for Coulomb (Figure 5.2d) compared to $\mu_I = -6.0$, $\mu_{II} = -5.1$ and $\mu_{III} = -5.4$ for van der Waals (Figure 5.2e)). The cut-off value of -8.2 Kcal/mol is thus at 1.9 standard deviations away from μ_{II} (Figure 5.2c), for which there is no linear cor-

relation between the Coulomb electrostatics and the van der Waals energies either, and as indicated by poor Pearson's correlation coefficients; $R^2_I = 0.06$, $R^2_{II} = 0.57$ and $R^2_{III} = 0.14$. Furthermore, there are no correlation between the total energy E^3_{ij} and the distance between the C_αs; $R^2_I = 0.01$, $R^2_{II} = 0.01$ and $R^2_{III} = 0.04$ (see Figure 2 in Supplementary Materials).

It is worth noting that the values of Motif III, the parallel bridge, are between those of Motif I, the anti-parallel open ring, and those of Motif II, the anti-parallel closed ring, in all dimensions reported in Figure 5.2 (C_α-C_α distances, side-chain directions, Coulomb electrostatics, and van der Waals energies).

The relative energetic contributions of the C=O$\cdots$C=O dipoles, the C=O$\cdots$H-N, and C=O$\cdots$H$_\alpha$-C$_\alpha$ H-bonds have been put into contrast to the total energy E^3_{ij} (see equation 5.12 in Materials and Methods), as shown in Figure 5.3. In all three canonical H-bonding motifs, the C=O$\cdots$C=O energy contributions are greater than zero, and thus participate significantly to the stabilization of adjacent strands (they have the same arithmetic sign as the total energy). Also, the C=O$\cdots$C=O components are in the same order of magnitude as the standard H-bonds C=O$\cdots$H-N, as noted by Milner-White and colleagues [21, 150], even though they used a different van der Waals and partial charges assignment [157, 158], and is particularly true in the parallel motif configuration. For the anti-parallel closed ring the C=O$\cdots$H-N H-bonds component must compensate for the destabilizing C=O$\cdots$H$_\alpha$-C$_\alpha$ bonds, as shown by the high percentages (over 100%) of the former, and the low percentages of the latter (below 0% or opposite sign of the total energy). Figure 5.3d shows the results of the three previously discussed curves for tri-peptides forming H-bonds, independently of the strand orientations.

DSSP finds 38181 residues in 5766 β-strands that form 1700 β-sheets, whereas β-Spider finds 64560 residues (an increase of 69.1%) in 8401 β-strands (an increase of 45.8%) that form 2088 β-sheets (an increase of 22.8%). In comparison, DSSP finds an average number of 22.4% β-residues per protein, whereas β-Spider finds an average of 30.9% β-residues per protein. Although the numbers of β-sheets identified by β-Spider is 388 more than the number of β-sheets found by DSSP,

β-Spider has actually found 827 new β-sheets, i.e. sheets in which no residues are marked in the state 'E' by DSSP. Given that the number of new β-sheets is much greater than the difference in reported annotated sheets by both methods, we are in face of the fact that several β-sheets reported by DSSP are smaller in size, and were combined in single larger ones by β-Spider; in other words, DSSP fragments the actual β-sheets in many independent ones. Among the 64560 residues found in β-sheets by β-Spider, 10.3% are β-edges and 4.5% are β-bulges. β-edges are β-neighbors that satisfy the geometrical criteria, but not the energetic one. β-edge residues are often found at the beginning and end of β-strands, as two strands merge into, or split from, a β-sheet. Note that our definition of β-bulges differs from that of DSSP (see Step 2 of the algorithm in Supplementary Materials).

The variety of H-bonding patterns between adjacent tri-peptides is shown in Table 5.1. 80% of all observed motifs adopt one of the three canonical motifs, defined in DSSP. However, more than 10% are missed by DSSP because they adopt non-canonical anti-parallel motifs, and more than 17% are missed by DSSP because they adopt parallel non-canonical motifs. As much as 45% more β-strands were uncovered by β-Spider compared to DSSP. 2783 among 8404 (33%) β-strands are new β-strands, i.e. strands in which no residue is found in the 'E' state by DSSP, and thus contribute to the formation of newly identified β-sheets or define new β-sheet edges. Figure 5.4 shows new β-strands that were identified by β-Spider in β-sheets that were identified by DSSP. Table 5.2 shows the longest β-strands that were not detected by DSSP, but adds to β-sheets already detected by DSSP. The largest new β-sheet found by β-Spider is in the chain A of 1G66, and contains 25 residues distributed in three β-strands : {134,138}, {148,155}, and {173,184}. Many β-strands that were added to the β-sheets captured by DSSP now define new β-sheet edges. We would find instructive to revisit the works described by Richardson [101] and Westhead [135].

Consider the β-helix in protein 1K5C (chain A), where we noticed a significant difference in the numbers of residues identified by DSSP, 177, versus β-Spider, 275. The main difference comes from the fact that β-Spider annotates the whole β-helix

as a β-strand, whereas DSSP captures only the residues on the flat faces of the helix. Interestingly, the β-sheets of DSSP hosting the greatest number of strands are β-helices (see for instance 1K5C and 1DAB)

Our β-sheet catalog contains β-strands that, on average, are more than one residue longer than the DSSP catalog, whereas the β-sheets, on average, contain almost eight additional residues. In our database of high-resolution structures, for DSSP, the average number of residues per β-sheet is 22.4 residues, whereas it is 30.9 for β-Spider (an increase of near 38%). Note that when analyzing the β-strand lengths, β-sheet sizes, and β-sheet residue content (the fraction of residues involved in β-sheet per protein), we observe the same distribution shapes between DSSP and β-spider, however a small shift up can be noticed in the β-Spider curves (see Supplementary Materials).

The detection of new β-strands has a profound impact on automatic protein fold identification. Consider the high-resolution X-ray crystal structures 1JUB and 1H16 (Figure 5.4c and 5.4d). The fold of 1JUB and 1H16 is the β-barrel. However, this fold could not have been detected as such by using the assignments made by DSSP. Note that in both CATH (version 2.5.1 - released January 2004) [136, 137]) and SCOP (version 1.65 - December 2003) [56]), the PDB file 1JUB has not yet been classified, whereas 1H16 has been correctly classified as containing the β-barrel.

Chou and Fasman noted that the amino-acid content of α-helices and β-sheets differ, which constitutes the basis of secondary structure prediction [132]. Then, Lifson and Sander observed that the amino-acid content in parallel and anti-parallel β-strands also differ [96]. In the context of our new β-sheet catalog, the amino-acid distributions were revisited. A total of eleven distributions were compared and are summarized in Table 5.3. The PDB Select 25 distribution can now be compared to those found in the β-sheets annotated by DSSP and β-Spider.

Figure 5.5 shows a dendrogram resulting from clustering all distributions based on the Kullback-Leibler amino-acid distance matrix (see Materials and Methods). Worth noting is the anti-parallel closed ring (Motif II), which with a distinctive twist is found distant from the open ring (Motif I), also in the anti-parallel configu-

ration. The closed ring does not allow for the appearance of the proline amino-acid, as the motif requires two backbone H-bonds whereas proline contains only one acceptor. In one hand, the DSSP motifs are distinct and abound in the same direction as previously noticed by Lifson and Sander [96]. On the other hand, the parallel and anti-parallel amino-acid distributions determined by β-Spider are much closer than those of DSSP. Also of great interest is the position of the amino-acid distribution of PDB Select 25 (i.e. of proteins in general) which is rooted under the β-Spider tree and thus blur the specificity of amino-acids to β-sheets. The amino-acid distributions from DSSP and β-Spider are irreconcilable, given that a high barrier partitions the laters from the formers (consider that all DSSP distributions are under one root). The β-edge distribution, with a high content of PRO and GLY residues, is distinct and can be the subject of further studies.

The relationship between the amino-acid distributions and the (ϕ,ψ) plot is clear, as shown in Figure 5.6. First, the three canonical H-bonding motifs (Figure 5.6abc), which with their stringent constraints to form at least two backbone H-bonds, populate almost exclusively the β regions of the Ramachandran plot, and have distinct amino-acid distributions. As discussed by Street and Mayo [149], it is the side-chain steric hindrance that sorts the backbone (ϕ,ψ) angles resulting in the exposure of the polar atoms of the main chain to form H-bonds. If we consider the C=O$\cdots$H-N energies and all backbone atom contributions, we get the (ϕ,ψ) plot in Figure 5.6f (β-partners only), where the α-helix and the L-α-helix regions are more populated (the greyscale intensities have been log-scaled). The (ϕ,ψ) angles in β-sheets, as defined by β-Spider, are now less indicative of the secondary structure type.

5.3 Materials and Methods

5.3.1 Backbone Hydrogen Atoms

Two backbone H-atoms are needed for the various energy calculations : H_N is the H connected to the backbone amide, and H_α is the H attached to the C_α. H-

atoms are invisible to X-rays. We positioned them geometrically in the following way. For H_N :

$$\vec{H_{Ni}} = \vec{N_i} + 1.01 \times \left(\frac{\vec{C_{i-1}} - \vec{O_{i-1}}}{\left| \vec{C_{i-1}} - \vec{O_{i-1}} \right|} \right) \tag{5.1}$$

where the $|\ |$ operator is the vector length. For H_α, we have the following :

$$\vec{P_1} = \left(\frac{\vec{C_{\alpha_i}} - \vec{N_i}}{\left| \vec{C_{\alpha_i}} - \vec{N_i} \right|} \right) \tag{5.2}$$

$$\vec{P_2} = \left(\frac{\vec{C_{\alpha_i}} - \vec{C_i}}{\left| \vec{C_{\alpha_i}} - \vec{C_i} \right|} \right) \tag{5.3}$$

$$\vec{P_x} = 1.09 \times \sin(109.5°/2) \times \left(\frac{\vec{P_1} \times \vec{P_2}}{\left| \vec{P_1} \times \vec{P_2} \right|} \right) \tag{5.4}$$

$$\vec{P_y} = 1.09 \times \cos(109.5°/2) \times \left(\frac{\vec{P_1} + \vec{P_2}}{\left| \vec{P_1} + \vec{P_2} \right|} \right) \tag{5.5}$$

$$\vec{H_{\alpha i}} = \vec{C_{\alpha i}} - \vec{P_x} + \vec{P_y} \tag{5.6}$$

where the $\times$ operator between two vectors is the cross-product.

There are two special cases to consider; the proline amino-acid which has no H_N atom, and the glycine amino-acid which has two H_α atoms :

$$\vec{H_{1\alpha i}} = \vec{C_{\alpha i}} - \vec{P_x} + \vec{P_y} \tag{5.7}$$

$$\vec{H_{2\alpha i}} = \vec{C_{\alpha i}} + \vec{P_x} + \vec{P_y} \tag{5.8}$$

The positioning of the H_N atom is fairly simple given that the nitrogen atom has a sp^2 orbital configuration. On the other hand, the tetrahedral atomic configuration around the sp^3 C_α atom and it's stereo-specificity require a more sophisticated geo-

metrical H_α positioning. The hydrogen covalent bond lengths, valence and torsion angles are taken from Amber [159]. For simplicity, we assumed the amide hydrogen bond i parallel to the previous carbonyl bond i-1 (Equation 5.1). Consequently, the torsion angle $\langle H_i, N_i, C_{i-1}, O_{i-1} \rangle$ has a value of 180°, as prescribed by Amber, but the valence angles $\langle H_i, N_i, C_{i-1} \rangle$ and $\langle H_i, N_i, C_{\alpha i} \rangle$ are not exactly as prescribed by the force-field. It is noteworthy to mention that this amide hydrogen bond configuration, except for the bond length, is also used in DSSP.

5.3.2 Energy Evaluation

The non-bonded energy has the same functional form as in the Amber force field [159]. It uses a Lennard-Jones 12-6 for the van der Waals interactions and a simple Coulomb electrostatics for the charged interactions. Thus, the non-bonded energy E_{ij}^A between atoms i and j separated by a distance R_{ij} (in Å) is given by :

$$E_{ij}^A = \epsilon_{ij}^* \frac{R_{ij}^{*12}}{R_{ij}^{12}} - 2\epsilon_{ij}^* \frac{R_{ij}^{*6}}{R_{ij}^6} + 332 \frac{Q_i Q_j}{R_{ij}} \tag{5.9}$$

The mixing rules are $\epsilon_{ij}^* = \sqrt{\epsilon_i^* \epsilon_j^*}$, $R_{ij}^* = R_i^* + R_j^*$. The parameters Q_i, ϵ_i^* and R_i^* are taken from Amber [159]. Since our force-field may not be the same as the one employed in the protein structure refinement process, and the location of hydrogen atoms are based on a geometrical, not an energetic, procedure, we reset the distance R_{ij} to the optimal distance of approach, R_{ij}^*, whenever $R_{ij} \leq R_{ij}^*$. Thus, the van der Waals energy of a pair of atoms that are too close in space will not neutralize the attractive Coulombic part.

From there, the non-bonded energy E_{ij}^R between residues i and j is the sum over all backbone atom interactions :

$$E_{ij}^R = \sum_{\{N,C_\alpha,C,O,H_N,H_\alpha\}}^{i} \quad \sum_{\{N,C_\alpha,C,O,H_N,H_\alpha\}}^{j} E_{ij}^A \tag{5.10}$$

which can be developed into :

$$
\begin{aligned}
\mathrm{E}_{ij}^{R} \;=\; & \underbrace{\sum_{\{C,O\}}^{i}\sum_{\{C,O\}}^{j}\mathrm{E}_{ij}^{A}}_{1} + \\
& \underbrace{\sum_{\{C,O\}}^{i}\sum_{\{H,N\}}^{j}\mathrm{E}_{ij}^{A} + \sum_{\{N,H\}}^{i}\sum_{\{O,C\}}^{j}\mathrm{E}_{ij}^{A}}_{2} + \\
& \underbrace{\sum_{\{C,O\}}^{i}\sum_{\{H_\alpha,C_\alpha\}}^{j}\mathrm{E}_{ij}^{A} + \sum_{\{C_\alpha,H_\alpha\}}^{i}\sum_{\{O,C\}}^{j}\mathrm{E}_{ij}^{A}}_{3} + \cdots
\end{aligned}
\tag{5.11}
$$

This energy evaluation takes into account the two potential backbone N-H$\cdots$O=C H-bonds (2) formed between the two residues, as well as the C=O$\cdots$C=O dipole (1) and the bifurcated C_α-$H_\alpha\cdots$O=C H-bonds (3). It is important to realize that this energy evaluation is not the same as $\Delta\Delta G$ [160]; here we calculate the energy needed to separate the two adjacent peptide chains to infinity, without regards to side-chain atoms. The N and C terminal blocking groups are not taken into account. For simplicity, the protonated form of histidine, as well as the reduced form of cysteine is used, as they show no major backbone partial charge differences between their variant species. The side-chain atoms are not implicated in the energy evaluations since the β-sheet phenomenon is thought to be the predominant interactions of polar main-chain atoms [161]. Special attention is given to proline which has no H_N atom, and to glycine which has two H_α atoms. Finally, the non-bonded energy E_{ij}^{3} between two sequentially distant tri-peptides $\{i-1,i,i+1\}$ and $\{j-1,j,j+1\}$ is given by :

$$
\mathrm{E}_{ij}^{3} = \sum_{\Delta i=-1}^{+1}\sum_{\Delta j=-1}^{+1}\mathrm{E}_{i+\Delta i\; j+\Delta j}^{R} \tag{5.12}
$$

These two tri-peptides are energetically favored if their non-bonded energy E_{ij}^{3} is smaller, or better, than a certain threshold $\mathrm{E}_{\text{threshold}}^{3}$:

$$E^3_{ij} \leq E^3_{\text{threshold}} \tag{5.13}$$

which has been determined empirically by inspecting the non-bonded energy E^3_{ij} distributions of all residues i and j identified by DSSP rule sets and fixed to -8.2 Kcal/mol (see Figure 5.2).

5.3.3 Geometrical Evaluation

Some authors [97, 162] use a geometrical evaluation procedure to identify the β-sheet regions in proteins. Here, we propose the use of similar geometrical features to look for in potential β-sheet regions. We have :

$$\left|\vec{C_{\alpha i}} - \vec{C_{\alpha j}}\right| \leq 6.2\text{Å} \tag{5.14}$$

that is, the two C_α atoms are close in space and :

$$-128.0° \leq \text{torsion}\left\langle \vec{C_{\beta i}}, \vec{C_{\alpha i}}, \vec{C_{\alpha j}}, \vec{C_{\beta j}} \right\rangle \leq 68.4° \tag{5.15}$$

Therefore, if we align the two C_α atoms, the projected angle spanned between the two C_β atoms must be less than a certain amount, thus making sure that these C_β atoms point in the same direction. Since glycine has no C_β atom, if one of the two compared residues i or j is a glycine we make equation 5.15 be true. This is different from the commonly used rule :

$$\left(\vec{C_{\beta i}} - \vec{C_{\alpha i}}\right) \cdot \left(\vec{C_{\beta j}} - \vec{C_{\alpha j}}\right) \geq 0 \tag{5.16}$$

where the $\cdot$ operator is the vector scalar product. The difference between equations 5.15 and 5.16 lies in the fact that equation 5.15 measures a projected angle between the two side-chain vectors, and thus does not captures the contribution along the $C_{\alpha i}$-$C_{\alpha j}$ axis, the local perpendicular to the β-sheet strands direction.

Two residues i and j are geometrically favored if they satisfy both equations 5.14 and 5.15.

5.3.4 β-Sheet Definition

Four types of residues are defined in order to unmask the β-sheet domains found in a given protein 3-D structure. The residue types are used to identify potential β-sheet members. Here are the type definitions :

1. β-Partners(i, j) are two residues i and j for which the energetic (equation 5.13) and geometrical criteria (equations 5.14 and 5.15) are met.
2. β-Edges(i, j) are two residues i and j for which the geometrical criteria (equations 5.14 and 5.15) are met, but are energetically unfavored (equation 5.13 is not verified). This happens often for residues lying on strand edges when the two adjacent peptide chains split to take different routes in the protein fold, thus lowering the associated E_{ij}^3 energy.
3. β-Neighbors(i, j) are two residues i and j that are either β-Partners(i, j) or β-Edges(i, j), as they satisfy the geometrical criteria (equations 5.14 and 5.15) in both cases.
4. β-Bulges are small peptide chains for which both ends are β-Neighbors. On the non-bulged strand, the β-Neighbors must be adjacent in sequence, thus capturing a one-strand insertion of extra residues between two consecutive β-bridges. Bulges of up to three residues are considered.

The β-sheet identification process is divided in three steps. Residues that are part of α-helices or β-turns are not considered as potential β-sheet residues. They are identified in the PDB file under the sections HELIX and TURN respectively.

Step 1) The first step has for goal the identification of β-Partners and β-Edges present in the protein fold (pseudo-code of the algorithm in the Supplementary Materials). Since β-Edges are defined after β-Partners are identified, the only resort of a strand to be part of a β-sheet is to have at least one of its residues to be energetically and geometrically favored (equations 5.13, 5.14 and 5.15).

Step 2) The second step identifies the β-bulges, as it needs the completion of the previous step for its success.

Step 3) The third step identifies the β-sheet domains by recursively visiting adjacent β-Neighbors and β-bulges, and also the N and C-terminal residues $(i \pm 1)$. In order to remove all single-residue strands from the β-sheets we must apply the two following rules until no further residues are deleted from the β-sheets :

Rule 1) Single-residue strands are removed from the β-sheets, as we expect β-strands to be composed of at least two consecutive residues. The β-Neighbors of these single-residue strands are thus updated.

Rule 2) Any residues that are left neighbor-less from the application of Rule 1 are also removed from the β-sheets.

β-Spider substitutes the PDB SHEET record from the input file with the newly identified and correctly determined β-sheet domains.

5.3.5 Amino-Acid Distributions

In order to compare the different amino-acid distributions found in β-sheets by either DSSP or β-Spider we use the Kullback-Leibler distance $\mathrm{KL}(I, J)$ [90] between two amino-acid partitions I and J, which is defined as :

$$\mathrm{KL}(I, J) = \sum_{k}^{20} \mathbb{P}_k^I \log_2 \frac{\mathbb{P}_k^I}{\mathbb{P}_k^J} \tag{5.17}$$

where $\mathbb{P}_k^I$ is the fraction of amino-acid type k in distribution I. The sum is over all amino-acid types which sum to 20. Units are in bits. The Kullback-Leibler distance has the following properties :

$$\mathrm{KL}(I, J) \neq \mathrm{KL}(J, I) \tag{5.18}$$

$$\mathrm{KL}(I, J) \geq 0 \tag{5.19}$$

$$\mathrm{KL}(I, J) = 0 \longleftrightarrow I = J \tag{5.20}$$

The Kullback-Leibler distance is not necessarily symmetric (equation 5.18), always positive definite (equation 5.19) and the KL distance will be 0 if and only if the two distributions I and J are the same. The construction of the symmetric distance matrix $\mathcal{M}_{ij}$ needed for the clustering is simply defined as :

$$\mathcal{M}_{ij} = \max(\mathrm{KL}(I, J), \mathrm{KL}(J, I)) \tag{5.21}$$

5.3.6 Protein Databases

A subset of the Protein Data Bank (PDB) [85] with no more than 25% of identity in sequence called PDB Select 25 [47] has been used. Six different versions of the PDB Select 25 compiled by the Pisces server [84], in date of the 27$^{\text{th}}$ of April 2004, at various X-ray resolutions, are exploited to rule out the effect of X-ray resolutions on the annotation quality. The sets of protein X-ray structures vary in resolution from 1.6 to 3.0 Å. There are 811 chains in the most stringent set while 3083 are present in the most relaxed set. Otherwise mentioned, the strictest data set at 1.6Å of resolution has been used for the various statistics. In order to establish a common foundation among the protein files and accelerate the annotation, the helical residues (α, π, 3-10) identified by DSSP were not considered during β-sheet identification. We used the computer program dssp2pdb (http ://keres.colorado.edu/dssp2pdb/) to replace the secondary structure annotations of the original PDB files.

5.3.7 H-Bonding Nomenclature

Multiple patterns of H-bonds can be found across two contiguous peptide backbone chains. A nomenclature has been derived to describe unambiguously these H-bond motifs between two tri-peptides. H-bonds are directional and run from the H-bond acceptor, the backbone C=O group, to the H-bond donor, the backbone H-N group. From there, a given C=O group in a residue could potentially be involved in at most three H-bonds with any three residues on the other peptide chain (we exclude intra-strand H-bonds). Thus, using a 0-1 notation for the absence or presence

of an H-bond respectively, each residue require three digits to display all the possible C=O to N-H links. Since a pair of tri-peptides contains six residues, we need a grand total of 18 digits for all possible H-bond motifs. The digits are arranged into six groups of three with each group representing a residue. For example, consider the parallel ring in which the middle residue of one strand is H-bonded to the first and last residue in the other strand (Figure 5.1c). For the first strand we would have the following digit pattern : 000-001-000, which indicates that the second residue (second group of digits) accepts an H-bond from the third residue (third digit has a value of 1) on the other strand. The other strand would be represented by the pattern : 010-000-000, that is, the first residue accepts an H-bond from the second residue in the first strand. The resulting pattern is the adjunction of the strand patterns : 000-001-000-010-000-000. The number of H-bonds is readily available from the descriptor and is the sum of 1's occurring in the motif. Both parallel and anti-parallel relative strand orientations are subject to symmetry-related digit patterns. For example, the following descriptor 010-000-000-000-001-000 is also a parallel ring. Therefore, in order to obtain a unique descriptor for symmetry-related motifs we consider both the original descriptor and its sibling by swapping the residues {i-1,i,i+1} $\leftrightarrow$ {j-1,j,j+1} and keep the lexicographically smaller descriptor. The H-bond definition is the same as used in DSSP.

5.4 Conclusion

As Street and Mayo demonstrated, the amino-acid side-chain has a direct influence on the conformation of the backbone [149], which confirmed the observation, by Chou and Fasman [132], of preferred ϕ-ψ torsion patterns in α-helices and β-bridges (motifs I, II and III). The favored amino acid conformations expose the amide hydrogen and carbonyl oxygen atoms to adjacent chains, promoting the formation of networks of hydrogen bonds in β-sheets [11]. However, β-sheets adopt a variety of shapes, with sheared and curled strands, which are often less extended than β-bridges. Consequently, it is not a surprise to observe, in β-sheets,

amino acid and torsion angle distributions that differ from that of DSSP. In addition to the C=O$\cdots$H-N hydrogen bonds, β-sheets are composed of bifurcated C_α-$H_\alpha\cdots$O=C hydrogen bonds and C=O$\cdots$C=O electrostatic dipoles, as was also shown by Milner-White and coworkers [21, 150].

We therefore introduced a new algorithm, β-Spider, to annotate the β-sheets in proteins. The algorithm is not biased towards specific H-bonding or ϕ-ψ patterns. The algorithm includes all non-bonded stabilizing/packing factors between adjacent β-sheet peptide main chains. The resulting β-sheet catalog differs considerably from the one obtained using DSSP or any other existing algorithm. The new catalog contains H-bonding motifs that were never observed prior to β-Spider. The new β-sheet amino acid and ϕ-ψ distributions are less indicative of the secondary structure type than was previously believed.

5.5 Acknowledgments

The authors thank the reviewers for providing pertinent comments and suggestions. This work was funded by the Canadian Institutes of Health Research (CIHR MT-14604). FM is a CIHR investigator.

Anti-parallel

DSSP	N	%
Motif I (open ring)		
001-000-000-001-000-000	532	3.40
001-000-000-001-000-100	2914	18.65
001-000-100-001-000-100	3816	24.43
Motif II (closed ring)		
000-010-000-000-010-000	6429	41.15
000-010-000-000-010-100	23	0.15
000-010-000-000-110-000	142	0.91
	13856	88.7

β-Spider	N	%
000-000-000-000-000-000	45	0.29
000-000-000-000-000-010	10	0.06
000-000-000-000-000-100	19	0.12
000-000-000-000-001-000	27	0.17
000-000-000-000-010-000	319	2.04
000-000-000-000-011-000	24	0.15
000-000-000-001-000-000	114	0.73
000-000-000-001-000-100	22	0.14
000-000-000-010-000-000	19	0.12
000-000-010-000-001-000	49	0.31
000-000-010-000-010-000	47	0.30
000-000-010-000-011-000	137	0.88
000-000-100-000-001-000	15	0.10
000-000-100-000-010-000	46	0.29
000-000-100-001-000-000	425	2.72
000-000-100-001-000-100	75	0.48
000-000-100-001-010-000	12	0.08
000-000-100-010-000-000	73	0.47
000-000-100-011-000-000	17	0.11
000-000-110-000-001-000	25	0.16
000-000-110-000-011-000	20	0.13
000-010-000-000-100-000	226	1.45
	1766	11.3

Total entries : 15622

(a)

Parallel

DSSP	N	%
Motif III (parallel ring)		
000-001-000-010-000-000	6462	80.68
000-001-000-110-000-000	143	1.78
000-001-001-010-000-000	12	0.15
	6617	82.6

β-Spider	N	%
000-000-000-000-000-000	45	0.56
000-000-000-000-000-001	40	0.50
000-000-000-000-001-000	370	4.62
000-000-000-000-010-000	66	0.82
000-000-000-000-011-000	65	0.81
000-000-000-001-000-000	35	0.44
000-000-000-010-000-000	477	5.96
000-000-000-010-000-001	14	0.17
000-000-000-010-001-000	85	1.06
000-000-000-010-011-000	12	0.15
000-000-000-011-000-000	17	0.21
000-000-000-100-000-000	14	0.17
000-000-001-010-000-000	43	0.54
000-000-001-011-000-000	12	0.15
000-001-000-100-000-000	87	1.09
001-000-000-100-000-000	10	0.12
	1392	17.4

Total entries : 8009

(b)

Table 5.1 – The repertoire of H-bonding motifs occuring in the 1.6Å resolution PDB Select 25 database. The residues in the middle cross-strand pair are β-Partners, as they satisfy both the geometric and energectic criteria. The residues in the flanking cross-strand pairs are β-Edges, as they satisfy at least the geometric criteria. See Materials and Methods for the definition of the H-bonding motif descriptor (the 01 vectors). H-bonding motifs under the DSSP sections are identified by DSSP and β-Spider, whereas those under the β-Spider section are only detected by β-Spider. N : the number of examples found in the database. In the motif descriptors, the H-bonds sought by DSSP are shown in bold.

PDB	Strand	Length
1JUBA	37-53	17
1UCSA	41-56	16
1QV9A	262-276	15
1GPIA	244-253	10
1OC7A	372-381	10
1FQTA	46-54	9
1H16A	270-278	9
1HX0A	63-71	9
1JNDA	296-304	9
1OCYA	388-396	9
1GP0A	1570-1577	8
1ISUA	43-50	8
1L5OA	307-314	8
1MK0A	83-90	8
1N1PA	314-321	8
1NC5A	204-211	8
1OF8A	339-346	8
1QHQA	4-11	8
1QHVA	498-505	8
1SX5A	177-184	8
2ENG	127-134	8
1A8D	256-262	7
1AQUA	165-171	7
1BQCA	267-273	7
1CRUA	259-265	7
1DCIA	112-118	7
1G5AA	149-155	7
1GA6A	330-336	7
1GTVA	30-36	7
1H12A	260-266	7
1HT6A	53-59	7
1JNRA	136-142	7
1JNRA	441-447	7
1JU3A	398-404	7
1JUBA	69-75	7
1JZ8A	274-280	7

(a)

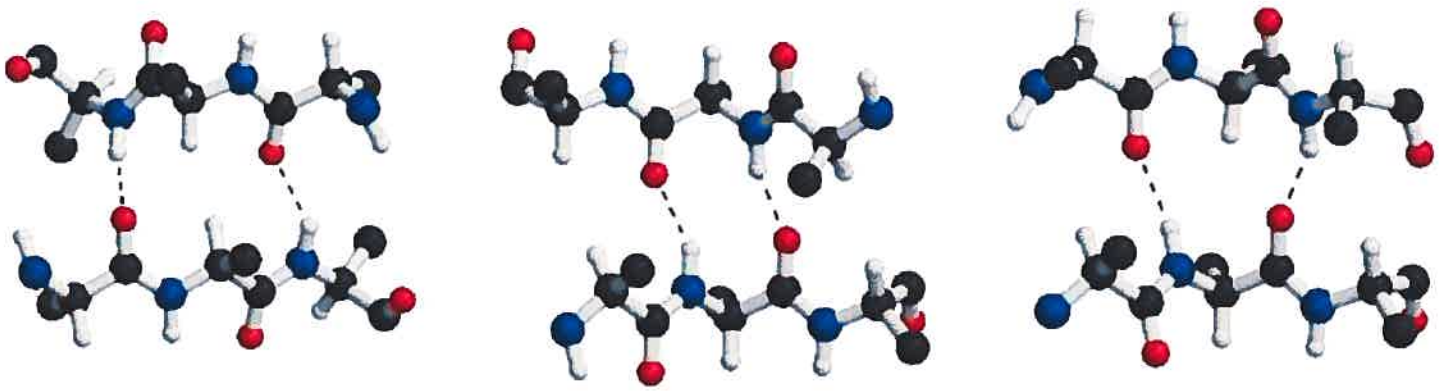
PDB	Strand	Length
1KWGA	363-369	7
1LLFA	48-54	7
1N1PA	447-453	7
1PO5A	36-42	7
1QHVA	409-415	7
1RK6A	436-442	7
2ENG	144-150	7
1A8D	249-254	6
1GA6A	339-344	6
1GCI	264-269	6
1GKLA	1037-1042	6
1GKPA	447-452	6
1GSOA	348-353	6
1H12A	253-258	6
1H16A	345-350	6
1HX0A	45-50	6
1I71A	12-17	6
1JU3A	413-418	6
1JZ8A	498-503	6
1JZ8A	604-609	6
1KA1A	321-326	6
1LUCA	48-53	6
1LUGA	227-232	6
1LUGA	238-243	6
1M1NB	180-185	6
1NTHA	418-423	6
1NYKA	91-96	6
1O7IA	109-114	6
1O7NA	133-138	6
1OYGA	95-100	6
1QQ9A	238-243	6
1RK6A	473-478	6
1V6SA	228-233	6
1YGE	205-210	6
3SEB	40-45	6

(b)

Table 5.2 – New β-strands found in DSSP's β-sheets on the 1.6Å resolution PDB Select 25 database. For each PDB structure, the inclusive β-strand residue numbers, as extended by β-Spider, are given (Strand). **a)** New β-strands of seven residues or more. **b)** New β-strands of six and seven residues.

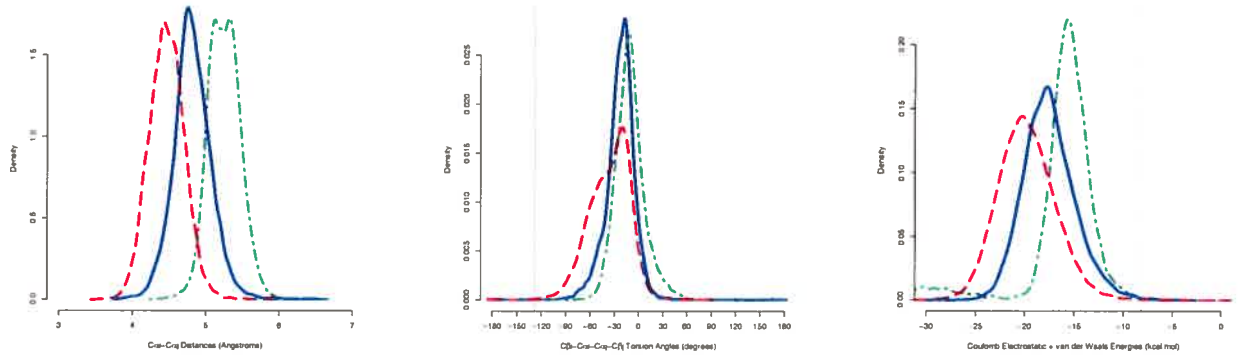
	PDB	DSSP				β-Spider					
	Select 25	β	I	II	III	β	Edge	Bulge	Partner	Para	Anti
ALA	7.7	6.7	6.4	6.8	6.9	6.9	6.6	7.7	6.9	7.4	6.4
ARG	4.9	4.4	4.6	5.1	2.8	4.4	4.1	5.4	4.4	3.1	4.5
ASN	4.6	2.8	2.6	2.3	2.3	4.2	4.7	5.2	4.1	4.1	3.9
ASP	5.8	3.1	2.6	2.8	2.5	4.7	6.2	7.5	4.4	4.2	4.3
CYS	1.8	1.6	1.8	1.7	1.6	1.7	1.9	1.1	1.7	1.8	1.7
GLN	3.9	2.8	2.9	3.1	1.9	2.9	2.4	3.8	2.9	2.2	2.9
GLU	6.7	4.3	4.5	4.6	2.7	4.2	3.8	7.2	4.1	2.7	4.0
GLY	7.1	5.2	3.4	6.0	4.7	9.2	11.8	6.9	9.0	13.9	13.7
HIS	2.3	2.4	2.1	2.6	2.3	2.5	2.3	2.8	2.5	2.5	2.2
ILE	5.7	9.6	9.0	9.0	13.3	7.0	4.7	4.7	7.4	8.7	6.3
LEU	8.8	10.0	10.5	9.0	12.8	8.3	6.9	7.4	8.5	9.2	7.8
LYS	6.3	4.5	4.5	4.7	2.7	4.6	4.2	6.1	4.6	3.2	4.5
MET	2.2	2.1	2.3	2.4	2.5	2.0	2.1	1.2	2.0	2.1	1.9
PHE	4.1	5.8	6.0	6.3	6.2	4.8	3.4	4.0	5.0	4.9	4.9
PRO	4.5	1.9	3.2	0.0	0.8	3.8	9.2	4.1	3.2	2.6	2.7
SER	6.0	5.0	4.8	5.3	3.8	5.8	6.7	7.0	5.6	4.8	5.4
THR	5.7	6.9	7.6	6.7	5.4	6.8	5.9	6.5	6.9	5.4	6.7
TRP	1.5	2.0	2.4	2.2	1.5	1.7	1.4	1.4	1.7	1.4	2.0
TYR	3.6	5.2	5.9	6.3	4.5	4.3	3.0	3.8	4.5	3.7	4.8
VAL	6.8	13.7	12.8	13.1	18.7	10.2	8.7	6.2	10.6	12.1	9.4
	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 5.3 – Comparison of amino-acid distributions in PDB Select 25 versus those found in β-sheets by DSSP and β-Spider (β columns). For DSSP, columns I, II and III refer, respectively, to the three canonical H-bonding motifs. For β-Spider, the Edge, Bulge and Partner columns refer respectively to the residues in the β-sheets that are either classified as β-Edges, β-Bulges, or β-Partners (see Materials and Methods for their formal definition). For two given β-Edge or β-Partner residues, the Para and Anti distributions relate to the relative strand orientations. The data were obtained from the analysis of the 1.6Å resolution PDB Select 25 database.

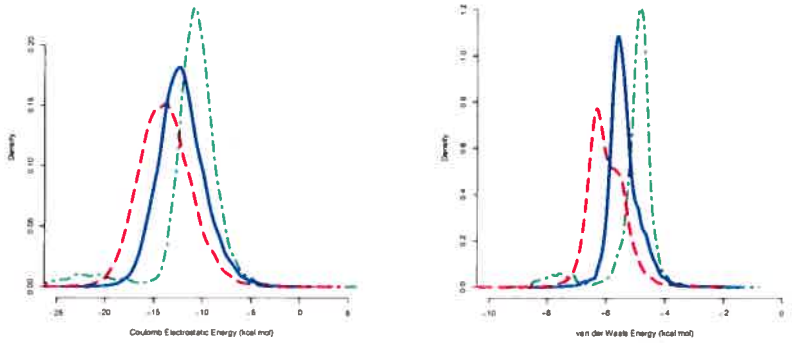


(a) (b) (c)

Figure 5.1 – The three β-sheet canonical H-bonding motifs. The N atoms are colored in blue, O in red, C in black, and H in gray. The covalent bonds are represented by gray cylinders. The H-Bonds between the backbone C=O$\cdots$H-N atoms are shown using dashed lines. The H-atoms were positioned using a procedure described in the Methods. The N-terminal residue is located left of the lower strand. The images were produced using RasMol [12], MolScript [130] and Raster3D [131]. **a)** Motif I, as found in anti-parallel strands, and often referred to as the wide or open ring. The shown example was extracted from the PDB file 1A8D, residues 286-288 and 419-421. **b)** Motif II, as found in anti-parallel strand pairs, and often referred to as the narrow or close ring. The shown example was extracted from the PDB file 1A3C, residues 4-6 and 176-178. **c)** Motif III, as found in parallel strands. The example was extracted from the PDB file 1A3C, residues 34-36 and 100-102.



(a) (b) (c)



(d) (e)

Figure 5.2 – β-Spider cut-off parameters. The dashed red lines correspond to motif I, the dashed-dotted green lines to motif II, and the continuous blue lines to motif III. The cut-off values are marked by the vertical gray lines. The area under the curves equals one. **a)** Distribution of the C^i_α-C^j_α distances when residues i and j are β-sheet partners according to DSSP. The cut-off value is 6.2Å. **b)** Distribution of the C^i_β-C^i_α-C^j_α-C^j_β torsion angles when residues i and j are β-sheet partners according to DSSP. The cut-off values are -128.0 and +68.4 degrees. **c)** Distribution of the total binding energy E^3_{ij} (equation 5.12) when residues i and j are β-sheet partners according to DSSP. The cut-off value is set -8.2 Kcal/mol. **d)** Contribution of the Coulomb electrostatic energy. **e)** Contribution of the van der Waals energy.

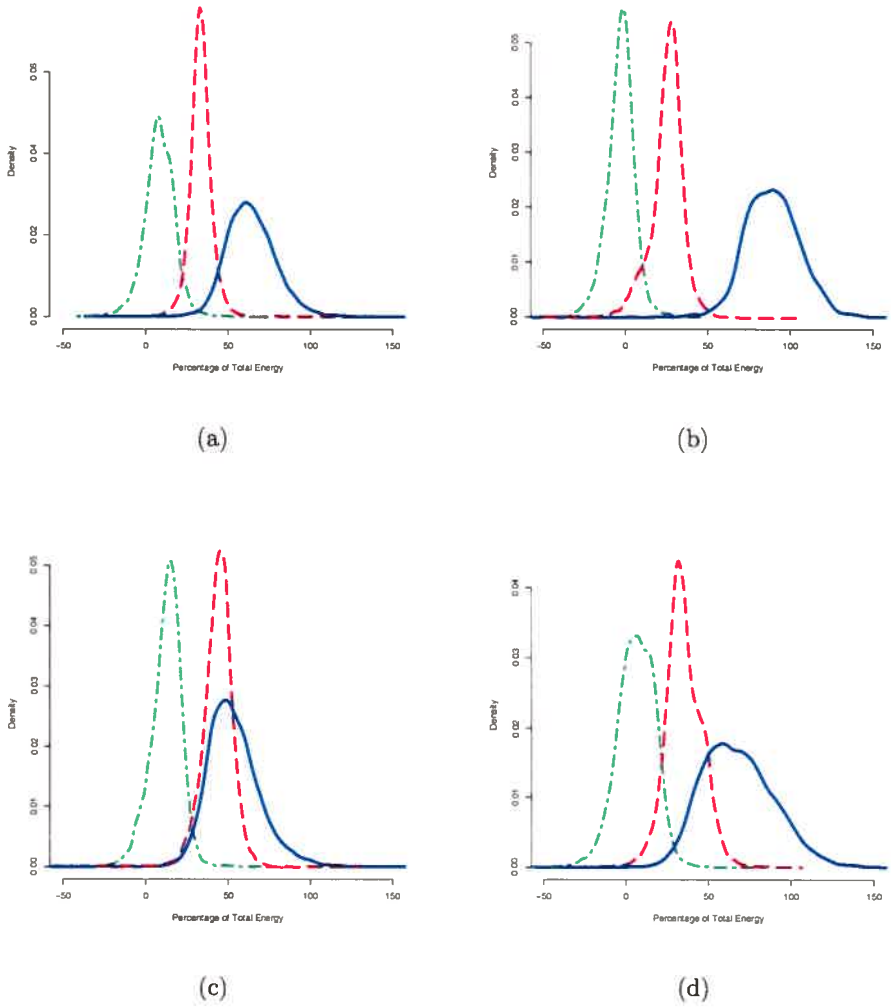


Figure 5.3 – Relative energetic contributions. The C=O$\cdots$H-N contributions are shown using continuous blue lines (equation 5.11 (2)), and correspond to the standard H-bond definition used in DSSP, C=O$\cdots$C=O are shown using dashed red lines (equation 5.11 (1)), and the C=O$\cdots$H$_\alpha$-C$_\alpha$ are shown using dotted-dashed green lines (equation 5.11 (3)). The area under the curves equals one. The total energy refers to E^3_{ij} (equation 5.12). Contributions higher than 100% indicate that the total energy, E^3_{ij}, is smaller than the specific energy considered, as it has been lowered by unfavorable Coulomb interactions. Contributions smaller than zero indicate that the specific energy considered is in the opposite sign of the total energy, and thus correspond to a destabilization. The data were extracted from the 1.6Å resolution PDB Select 25 database. **a)** Pairs of tri-peptides displaying the H-bonding motif I. **b)** Pairs of tri-peptides displaying the H-bonding motif II. **c)** Pairs of tri-peptides displaying the H-bonding motif III. **d)** A combination of the three graphs.

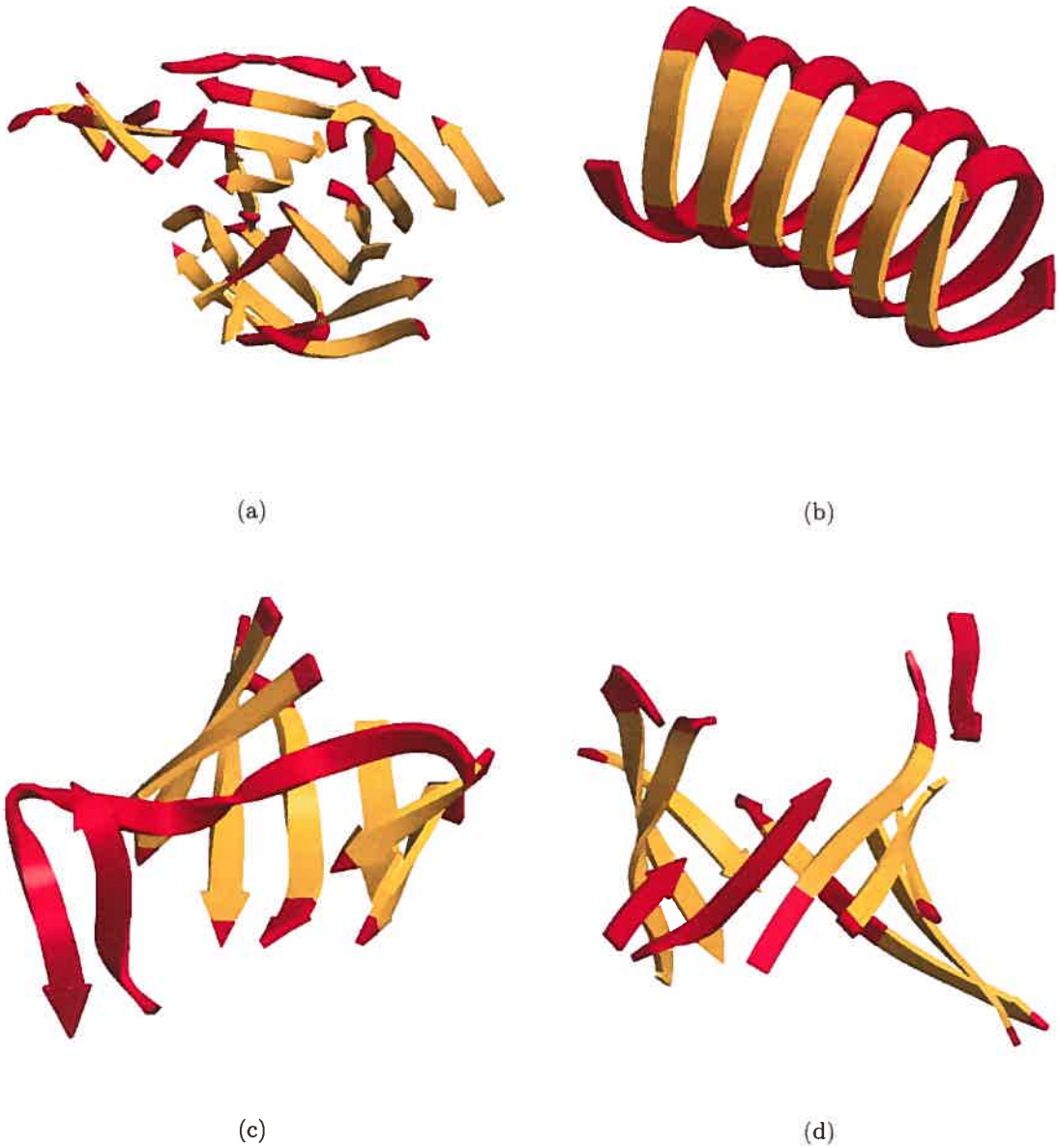


Figure 5.4 – β-Spider vs. DSSP. The DSSP assignments are shown in gold, and β-Spider in red. The images were produced using RasMol [12], MolScript [130], and Raster3D [131]. **a)** Chain A in PDB file 1GPI. At the top, a new strand identified by β-Spider changes the β-sheet border spanned by residues 244 to 253. **b)** The β-helix in chain A in PDB file 1EZG. The β-helix is defined by residues 4 to 82. **c)** A β-barrel in the chain A in PDB file 1JUB. The strand identified by β-Spider, from residues 37 to 53, confirms the β-barrel domain. **d)** A β-barrel in the chain A in PDB file 1H16. Two strands detected by β-Spider were needed to complete the β-barrel domain.

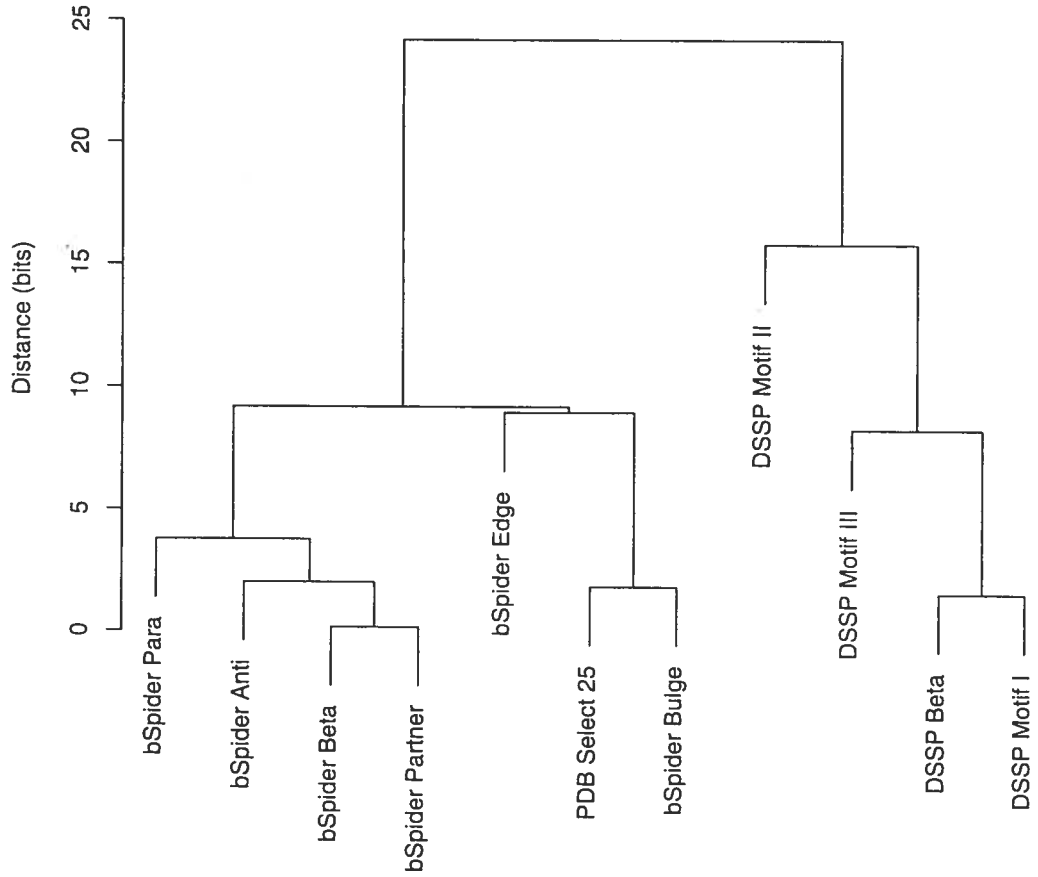


Figure 5.5 – Clustering of various amino-acid distributions. Each of the amino-acid distributions have been compiled and clustered from the distance matrix as defined in Materials and Methods (equation 5.21). The amino-acid distributions are those of Table 5.3. The distance units are bits. The average-merge clustering has been performed by the R statistical package [163].

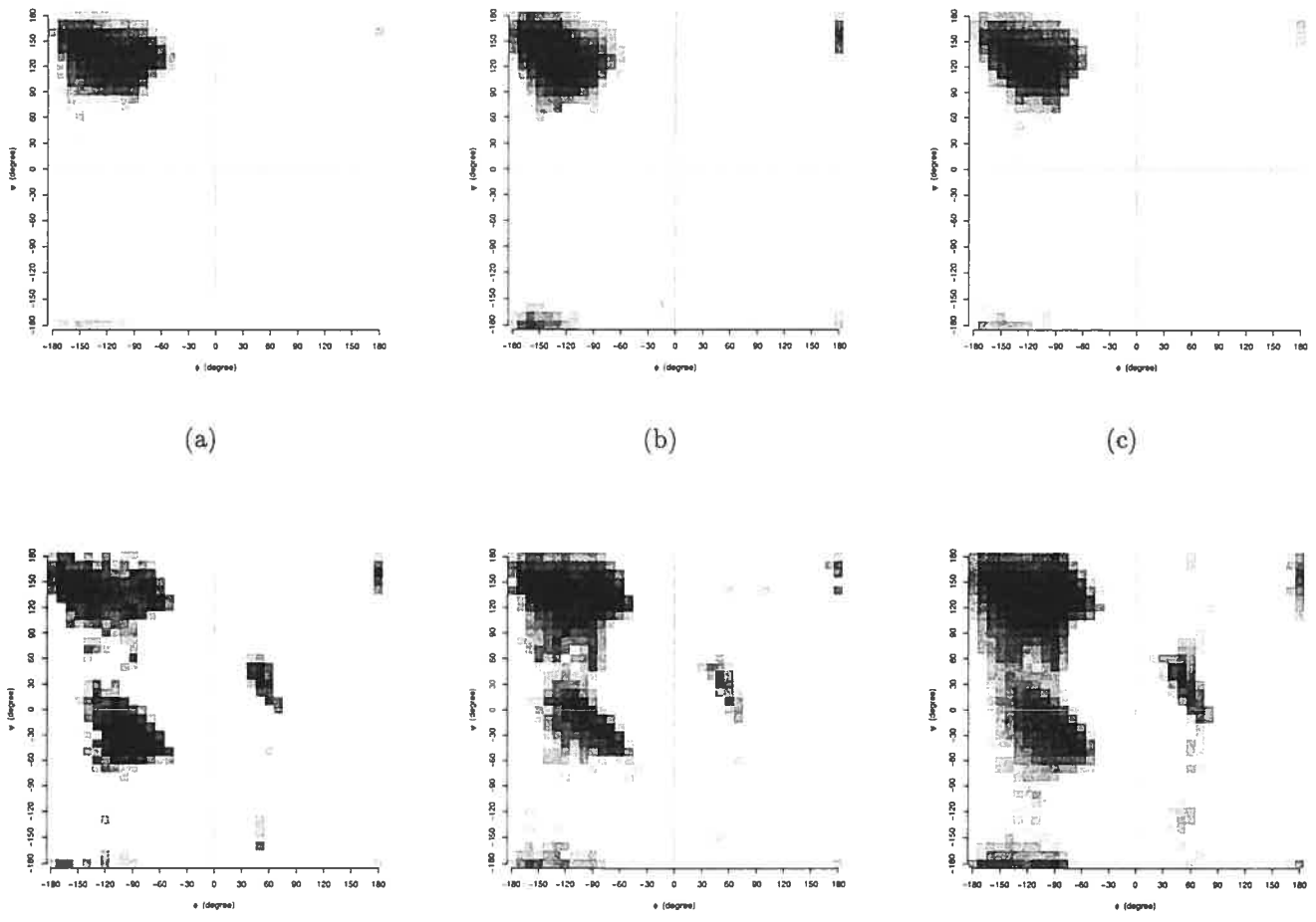


Figure 5.6 – Ramachandran plots of DSSP versus β-Spider. The ϕ/ψ plots of the three canonical H-bonding motifs are displayed above and against the three topological features found in β-sheets by β-Spider (below). For the canonical H-bonding motifs, each time a motif has been identified the ϕ/ψ values of both residues forming the β-bridge were plotted. The ϕ angle, in abscissa, and the ψ angle, in ordinate, are shown in degrees. The values were binned to 10° in both axes. The darker the pixel, the larger is the population at the particular ϕ/ψ value. For clarity, the $\log_e$(count) has been used to adapt to the linear gray scale. Glycine and proline residues have not been included in the ϕ/ψ plots due to their particular steric characteristics. The data were extracted from the analysis of the 1.6Å resolution PDB Select 25 database. **a)** ϕ/ψ plot of the H-bonding motif I of Figure 5.1. **b)** ϕ/ψ plot of the H-bonding motif II. **c)** ϕ/ψ plot of the H-bonding motif III. **d)** ϕ/ψ plot of the β-Bulges identified by β-Spider. **e)** ϕ/ψ plot of the β-Edges identified by β-Spider. **f)** ϕ/ψ plot of the β-Partners identified by β-Spider.

CHAPITRE 6

A NEW CATALOG OF PROTEIN β-SHEETS ; SUPPLEMENTARY MATERIALS

Marc Parisien and François Major
Département d'Informatique et de Recherche Opérationnelle,
Université de Montréal, CP 6128 Succ. Centre-Ville,
Montréal, Québec, Canada H3C 3J7

M Parisien and F Major. A new catalog of protein β-sheets. *Proteins*, 61 :545-558, 2005.

Version	Sequence Identity (%)	Resolution (Å)	R-factor	Number of Chains
A)	25	1.6	0.25	811
B)	25	1.8	0.25	1406
C)	25	2.0	0.25	2055
D)	25	2.2	1.00	2396
E)	25	2.5	1.00	2784
F)	25	3.0	1.00	3083

Table 6.1 – The culled PDB Select 25 databases.The six databases have been compiled by the Pisces server [84] on the 27^{th} of April 2004, and contain only X-ray determined structures. Unless mentioned, the A) version is used throughout the article.

1A3C	1A4IA	1A62	1A6M	1A8D	1A8E	1ABA	1AH7	1AHO	1AIE
1AJSA	1AOP	1AQUA	1ARB	1ATG	1B16A	1B3AA	1B5EA	1B6A	1B6G
1B8OA	1B8ZA	1BD0A	1BGF	1BKF	1BKRA	1BQCA	1BRT	1BTEA	1BTKA
1BX4A	1BX7	1BYI	1BYQA	1C1KA	1C4QA	1C52	1C5EA	1C75A	1C7KA
1C8CA	1C9OA	1CC8A	1CCWA	1CCWB	1CEX	1CIPA	1CRUA	1CS1A	1CSEI
1CXQA	1CY5A	1CZPA	1D2SA	1D4OA	1D4TA	1D5TA	1D7PM	1D8WA	1DBWA
1DCIA	1DCS	1DD9A	1DF4A	1DFMA	1DG6A	1DI6A	1DJ0A	1DK8A	1DKIA
1DL2A	1DLWA	1DP7P	1DQZA	1DS1A	1DY5A	1DYPA	1DZKA	1E19A	1E29A
1E2WA	1E30A	1E58A	1E5KA	1E6UA	1E7LA	1E85A	1E87A	1E9GA	1EAJA
1EAQA	1EB6A	1ECA	1EDG	1EDMB	1EGWA	1EJ0A	1EJ8A	1EJDA	1EJGA
1EK6A	1EKQA	1ELKA	1ELUA	1EN2A	1ES5A	1ES9A	1ET1A	1EU1A	1EUVA
1EUVB	1EUWA	1EVLA	1EW4A	1EY4A	1EYHA	1EYVA	1EZGA	1EZM	1F0IA
1F1EA	1F1UA	1F2TA	1F2TB	1F46A	1F74A	1F7LA	1F86A	1F8EA	1F94A
1F9VA	1F9YA	1F9ZA	1FCQA	1FCYA	1FD3A	1FF4A	1FG7A	1FGYA	1FIUA
1FK5A	1FL0A	1FLMA	1FM0D	1FM0E	1FO8A	1FP2A	1FQTA	1FR2A	1FR2B
1FS7A	1FSGA	1FT5A	1FW9A	1FX2A	1FYEA	1G12A	1G2BA	1G2RA	1G2YA
1G3P	1G4YB	1G5AA	1G61A	1G66A	1G6GA	1G6SA	1G6UA	1G6XA	1G8QA
1GA6A	1GBS	1GCI	1GD0A	1GHEA	1GJ7A	1GK7A	1GK8A	1GK8I	1GK9A
1GK9B	1GKLA	1GKMA	1GKPA	1GMUA	1GMXA	1GNLA	1GP0A	1GPIA	1GPPA
1GPQA	1GQIA	1GS5A	1GSOA	1GTVA	1GU2A	1GUTA	1GV9A	1GVDA	1GVEA
1GVFA	1GVJA	1GVOA	1GVP	1GWEA	1GWMA	1GWUA	1GXMA	1GXUA	1GY7A
1GYXA	1GZ8A	1H05A	1H10A	1H12A	1H16A	1H1NA	1H2CA	1H2WA	1H32A
1H4AX	1H4GA	1H4XA	1H80A	1H8DL	1H97A	1H99A	1HBNA	1HBNB	1HBNC
1HD2A	1HDHA	1HDOA	1HFES	1HNJA	1HQ1A	1HQKA	1HQSA	1HT6A	1HW1A
1HX0A	1HXHA	1HXIA	1HYOA	1HZ4A	1HZTA	1I0DA	1I0RA	1I0VA	1I12A
1I1JA	1I1WA	1I24A	1I27A	1I2TA	1I40A	1I4UA	1I52A	1I58A	1I60A
1I71A	1I88A	1ID0A	1IDPA	1IE9A	1IFC	1IFRA	1IHRA	1II5A	1IJQA
1IJVA	1IJYA	1IKPA	1INLA	1IO0A	1IO7A	1IOMA	1IOOA	1IQ6A	1IQZA
1IRDB	1IRQA	1IS3A	1ISPA	1ISUA	1IT2A	1ITXA	1IU8A	1IUAA	1IUPA
1IUQA	1IV3A	1IW0A	1IX9A	1IXH	1J0PA	1J1NA	1J2JB	1J2RA	1J31A
1J34C	1J3AA	1J3WA	1J77A	1J98A	1J9BA	1JATA	1JB3A	1JBEA	1JBOA
1JBOB	1JCDA	1JCLA	1JEKA	1JEKB	1JER	1JETA	1JF3A	1JF8A	1JFBA
1JG1A	1JHGA	1JHJA	1JI1A	1JI7A	1JIGA	1JK3A	1JKEA	1JKXA	1JL0A
1JL1A	1JM1A	1JNDA	1JNIA	1JNRA	1JO0A	1JOVA	1JR8A	1JTVA	1JU2A
1JU3A	1JUBA	1JUHA	1JX6A	1JY2N	1JY2O	1JY2P	1JYKA	1JZ8A	1K0MA
1K20A	1K3IA	1K3XA	1K3YA	1K4IA	1K4NA	1K55A	1K5CA	1K5NA	1K5NB
1K6FA	1K6XA	1K7CA	1K7JA	1K8UA	1K92A	1KA1A	1KAFA	1KB0A	1KGDA
1KICA	1KJQA	1KKOA	1KLLA	1KM4A	1KMTA	1KMVA	1KNGA	1KNMA	1KOE
1KOIA	1KPF	1KQ1A	1KQ3A	1KQ6A	1KQFA	1KQFC	1KQPA	1KQRA	1KR4A
1KS8A	1KT6A	1KUGA	1KV7A	1KW3B	1KWFA	1KWGA	1KYFA	1KYHA	1L0IA
1L2HA	1L3KA	1L5OA	1L6RA	1L7AA	1L7LA	1L7MA	1L9LA	1L9XA	1LAM
1LB3A	1LC0A	1LC5A	1LF7A	1LKKA	1LLFA	1LLMC	1LLNA	1LM4A	1LM1A
1LNIA	1LO7A	1LQ9A	1LQAA	1LQTA	1LQVA	1LRIA	1LS1A	1LS9A	1LTZA
1LU0A	1LU4A	1LUCA	1LUGA	1LUQA	1LV7A	1LWBA	1LYCA	1LYQA	1LYVA
1LZLA	1M0KA	1M15A	1M1FA	1M1NA	1M1NB	1M1QA	1M22A	1M2DA	1M2KA
1M2XA	1M40A	1M4JA	1M4LA	1M55A	1M65A	1M7GA	1M7YA	1M9ZA	1MB3A
1MC2A	1MCTI	1ME4A	1MF7A	1MFGA	1MFMA	1MG4A	1MG7A	1MH1	1MJ4A
1MJ5A	1MK0A	1MKKA	1MLA	1MN8A	1MNNA	1MOQ	1MPLA	1MQKH	1MQKL
1MSOA	1MSOB	1MTPA	1MTPB	1MUN	1MUWA	1MWQA	1MXRA	1MY7A	1N08A
1N0QA	1N13A	1N13B	1N1PA	1N2EA	1N3LA	1N40A	1N55A	1N57A	1N62A
1N62B	1N62C	1N7OA	1N7SA	1N7SB	1N7SC	1N7SD	1N8KA	1N8VA	1NA3A
1NBUA	1NC5A	1NC7A	1NF9A	1NFP	1NG6A	1NH0A	1NK4A	1NKD	1NKIA
1NLNA	1NLQA	1NLS	1NM8A	1NN5A	1NNFA	1NNXA	1NOFA	1NOGA	1NOX
1NQJA	1NTEA	1NTHA	1NTVA	1NU0A	1NUYA	1NWAA	1NWPA	1NWWA	1NWZA
1NXB	1NXMA	1NYCA	1NYKA	1NYTA	1NZ0A	1NZJA	1O06A	1O08A	1O1ZA
1O2DA	1O4YA	1O6VA	1O7IA	1O7JA	1O7NA	1O7NB	1O7QA	1O82A	1O8BA
1O8XA	1O97C	1O97D	1O98A	1O9GA	1O9IA	1O9RA	1OAA	1OAFA	1OAIA
1OBDA	1OBOA	1OC7A	1OCYA	1OD3A	1OD6A	1ODMA	1ODZA	1OE1A	1OEWA
1OF8A	1OFNA	1OFZA	1OGAE	1OGWA	1OH0A	1OH4A	1OHLA	1OI0A	1OI7A
1OJRA	1OK0A	1OOHA	1OPD	1OQJA	1OQQA	1OQVA	1ORC	1ORRA	1OS6A
1OU8A	1OUWA	1OW4A	1OX0A	1OXCA	1OXDA	1OXXK	1OYGA	1OZ2A	1OZNA
1P0HA	1P0ZA	1P1MA	1P36A	1P4CA	1P4OA	1P5DX	1P5FA	1P5ZB	1P6OA
1P9HA	1P9IA	1PA1A	1PA7A	1PAZ	1PB7A	1PBJA	1PE9A	1PFBA	1PIDA
1PINA	1PKHA	1PKOA	1PLC	1PO5A	1PQ7A	1PQHA	1PSRA	1PVMA	1PWAA
1PWBA	1PWMA	1PZ4A	1PZ7A	1PZGA	1Q0RA	1Q1AA	1Q35A	1Q5YA	1Q6OA
1Q7EA	1Q7LA	1Q7LB	1Q92A	1QAUA	1QB7A	1QDDA	1QE3A	1QFTA	1QG8A
1QGIA	1QGVA	1QH5A	1QHQA	1QHVA	1QJ4A	1QKSA	1QL0A	1QLWA	1QMGA
1QMQA	1QNRA	1QOPB	1QQ4A	1QQ5A	1QQ9A	1QQFA	1QQQA	1QREA	1QS1A
1QTNA	1QTNB	1QTWA	1QU9A	1QV9A	1QW2A	1QW9A	1QWGA	1QWNA	1QWYA
1QXYA	1QZ5A	1R0RI	1R26A	1R29A	1R2MA	1R2QA	1R5LA	1R5RA	1R5YA
1R6DA	1R6XA	1R8SA	1R8SE	1R9LA	1RB9	1RDQE	1RDQI	1RGZA	1RHS
1RK6A	1RKQA	1RKUA	1ROCA	1RQWA	1RTQA	1RU4A	1RWHA	1S1DA	1S29A
1S9RA	1SFSA	1SHUX	1SQSA	1SVFA	1SVFB	1SX5A	1T1DA	1TCA	1THX
1UASA	1UBKL	1UBKS	1UCAA	1UCRA	1UCSA	1UFOA	1UFYA	1UG6A	1UGIA
1UGXB	1UI0A	1UJPA	1UKFA	1UOYA	1UQ5A	1URSA	1US5A	1USCA	1USGA
1USMA	1UTEA	1UTG	1UUQA	1UZBA	1V6SA	1V7RA	1V7ZA	1V8CA	1VCC
1VDWA	1VFYA	1VH5A	1VHNA	1VHUA	1VHWA	1VIAA	1VIMA	1VIOA	1VJUA
1WER	1WFBA	1WHI	1YCC	1YGE	256BA	2A0B	2ARCA	2AYH	2CPGA
2ENG	2ERL	2FDN	2IGD	2ILK	2LISA	2MCM	2MHR	2NLRA	2PTH
2PVBA	2TNFA	2TPSA	3CAOA	3CHBD	3EZMA	3LZT	3NUL	3PVIA	3SDHA
3SEB	3SIL	3VUB	4EUGA	4UBPA	4UBPB	6RLXA	6RLXB	7A3HA	7ODCA
8ABP									

Table 6.2 – The 811 protein chains of the 1.6Å resolution database labeled by their four-letter PDB codes. The fifth letter in the suffix is the PDB chain identifier. Underlined chains are those whose secondary structure elements were determined using DSSP.

Amino-Acid	Atom	Charge (Coulomb)
ALA		
	N	-0.4157
	H	+0.2719
	CA	+0.0337
	HA	+0.0823
	C	+0.5973
	O	-0.5679
ARG		
	N	-0.3479
	H	+0.2747
	CA	-0.2637
	HA	+0.1560
	C	+0.7341
	O	-0.5894
ASN		
	N	-0.4157
	H	+0.2719
	CA	+0.0143
	HA	+0.1048
	C	+0.5973
	O	-0.5679
ASP		
	N	-0.5163
	H	+0.2936
	CA	+0.0381
	HA	+0.0880
	C	+0.5366
	O	-0.5819
CYS		
	N	-0.4157
	H	+0.2719
	CA	+0.0213
	HA	+0.1124
	C	+0.5973
	O	-0.5679
GLN		
	N	-0.4157
	H	+0.2719
	CA	-0.0031
	HA	+0.0850
	C	+0.5973
	O	-0.5679
GLU		
	N	-0.5163
	H	+0.2936
	CA	+0.0397
	HA	+0.1105
	C	+0.5366
	O	-0.5819
GLY		
	N	-0.4157
	H	+0.2719
	CA	-0.0252
	1HA	+0.0698
	2HA	+0.0698
	C	+0.5973
	O	-0.5679
HIS		
	N	-0.3479
	H	+0.2747
	CA	-0.1354
	HA	+0.1212
	C	+0.7341
	O	-0.5894
ILE		
	N	-0.4157
	H	+0.2719
	CA	-0.0597
	HA	+0.0869
	C	+0.5973
	O	-0.5679

(a)

Amino-Acid	Atom	Charge (Coulomb)
LEU		
	N	-0.4157
	H	+0.2719
	CA	-0.0518
	HA	+0.0922
	C	+0.5973
	O	-0.5679
LYS		
	N	-0.3479
	H	+0.2747
	CA	-0.2400
	HA	+0.1426
	C	+0.7341
	O	-0.5894
MET		
	N	-0.4157
	H	+0.2719
	CA	-0.0237
	HA	+0.0880
	C	+0.5973
	O	-0.5679
PHE		
	N	-0.4157
	H	+0.2719
	CA	-0.0024
	HA	+0.0978
	C	+0.5973
	O	-0.5679
PRO		
	N	-0.2548
	CA	-0.0266
	HA	+0.0641
	C	+0.5896
	O	-0.5748
SER		
	N	-0.4157
	H	+0.2719
	CA	-0.0249
	HA	+0.0843
	C	+0.5973
	O	-0.5679
THR		
	N	-0.4157
	H	+0.2719
	CA	-0.0389
	HA	+0.1007
	C	+0.5973
	O	-0.5679
TRP		
	N	-0.4157
	H	+0.2719
	CA	-0.0275
	HA	+0.1123
	C	+0.5973
	O	-0.5679
TYR		
	N	-0.4157
	H	+0.2719
	CA	-0.0014
	HA	+0.0876
	C	+0.5973
	O	-0.5679
VAL		
	N	-0.4157
	H	+0.2719
	CA	-0.0875
	HA	+0.0969
	C	+0.5973
	O	-0.5679

(b)

Atom	Radius R^* (Å)	Energy ϵ^* (kcal/mol)
N	1.8240	0.1700
H	0.6000	0.0157
CA	1.9080	0.1094
HA	1.3870	0.0157
1HA	1.3870	0.0157
2HA	1.3870	0.0157
C	1.9080	0.0860
O	1.6612	0.2100

(c)

Table 6.3 – Force-field parameters taken from Amber [159], reproduced here as requested. **a and b)** Atomic partial charges. The protonated form of His was used. **c)** Van der Waals radii R^* and well depth energies ϵ^*.

	DSSP						β-Spider					
	Resolution (Å)						Resolution (Å)					
	1.6	1.8	2.0	2.2	2.5	3.0	1.6	1.8	2.0	2.2	2.5	3.0
β content (%)	21.4	21.9	21.7	21.4	21.0	20.7	35.2	35.7	34.9	34.4	33.7	33.3
Anti/Para	1.49	1.49	1.61	1.66	1.69	1.75	1.54	1.56	1.63	1.67	1.68	1.71
Residues												
N	38181	68766	104222	122280	141198	155034	64560	114949	171934	200949	231818	254691
%†							169.1	167.2	165.0	164.3	164.2	164.3
Strands												
N	5766	10293	15363	17944	20618	22557	8404	14323	21357	24967	28795	31676
%†							145.8	139.2	139.0	139.1	139.6	140.4
Sheets												
N	1700	2981	4466	5246	6072	6739	2088	3681	5467	6409	7481	8341
%†							122.8	123.5	122.4	122.2	123.2	123.8
New strands												
N							2783	4781	6974	8108	9418	10484
%‡							33.1	33.4	32.6	32.5	32.7	33.1
New sheets												
N							827	1458	2108	2464	2888	3234
%‡							39.6	39.6	38.6	38.4	38.6	38.8

Table 6.4 – DSSP versus β-Spider at various X-ray resolutions. The number of residues, β-strands and β-sheets at various X-ray resolutions are compiled for the β-sheet domains found in the 1.6Å resolution PDB Select 25 database (Table 6.1, version A) identified by either DSSP or β-Spider. The percentage rows † are those calculated from the DSSP reference counts to 100%. The percentage rows ‡ are those calculated from the β-Spider reference counts to 100%. New strands are β-strands detected by β-Spider in which none of it's composing residues are found in the extended (E) state by DSSP. A β-sheet composed entirely of new β-strands is thus called a new β-sheet. The ratios of anti-parallel strands to parallel, as well as the β-sheet content percentages are also given.

	DSSP		β-Spider	
	Resolution (Å)		Resolution (Å)	
	1.6	3.0	1.6	3.0
Sheet Size (in number of residues)				
Mean	22.4	23.0	30.9	30.5
Std. Dev.	17.9	20.6	36.4	36.8
Median	19	20	12	13
Mode	4	4	4	4
Maximum	177	341	275	441
Max. PDB Code	1K5CA	1QFGA	1K5CA	1QFGA
Sheet Size (in number of strands)				
Mean	4.3	4.3	4.5	4.4
Std. Dev.	2.4	2.5	3.6	3.4
Median	4	4	3	3
Mode	2	2	2	2
Maximum	29	46	27	33
Max. PDB Code	1K5CA	1DABA	1JU3A	1N9EA
Strand Length (in number of residues)				
Mean	5.2	5.4	6.9	7.0
Std. Dev.	2.6	2.8	4.8	4.9
Median	5	5	6	6
Mode	5	4	2	2
Maximum	27	33	102	125
Max. PDB Code	1H4GA	1PREA	1P9HA	1VH4A

Table 6.5 – Detailed statistics for DSSP versus β-Spider. The means, standard deviations (Std. Dev.), medians, modes and maximums for the β-sheet sizes in number of residues and strands, as well as the β-strand lengths in number of residues are compiled for DSSP and β-Spider, at the two extreme X-ray resolutions considered (Table 6.1, versions A and F).

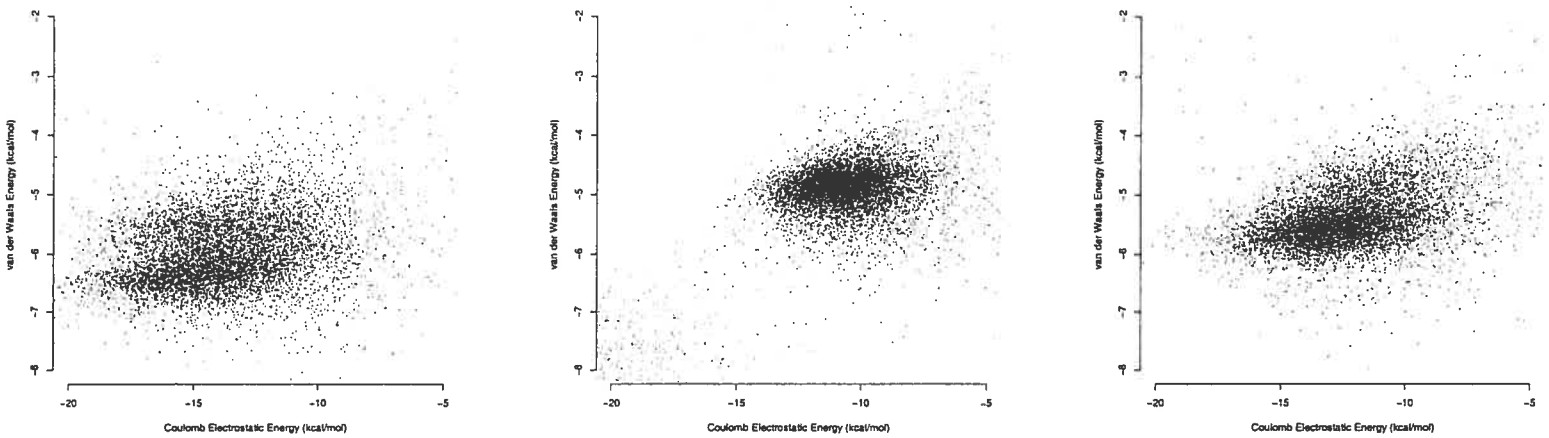
PDB	Residue(s)
1C7KA	9-10
1C9OA	23-24
1CIPA	239
1EJ0A	155-156
1EU1A	236
1F7LA	95-97
1FL0A	173
1G3P	21
1GY7A	31-32
1H16A	75-77
1HQKA	128
1HW1A	64
1HW1A	67-68
1INLA	197-198
1IUAA	51-52
1J2RA	165
1J31A	246
1JFBA	398
1JI1A	229
1JOVA	253-254
1JU2A	249-250
1JZ8A	202
1JZ8A	819
1K3IA	274
1KV7A	264-266
1L7AA	172-173

(a)

PDB	Residue(s)
1LLFA	200-201
1LZLA	151-152
1M2KA	119
1M2XA	61
1M40A	43
1MJ4A	16-17
1MJ5A	133
1NF9A	175
1NKIA	90
1NLS	150
1O7NB	560-562
1O7QA	281
1OEWA	70-71
1OH0A	34
1OZ2A	253
1PZ7A	103-105
1QE3A	180-181
1QFTA	74-76
1QJ4A	36
1QQ9A	77
1RQWA	25-26
1RTQA	70
1RTQA	73-75
1S1DA	138-139
1URSA	203
1VJUA	128-129

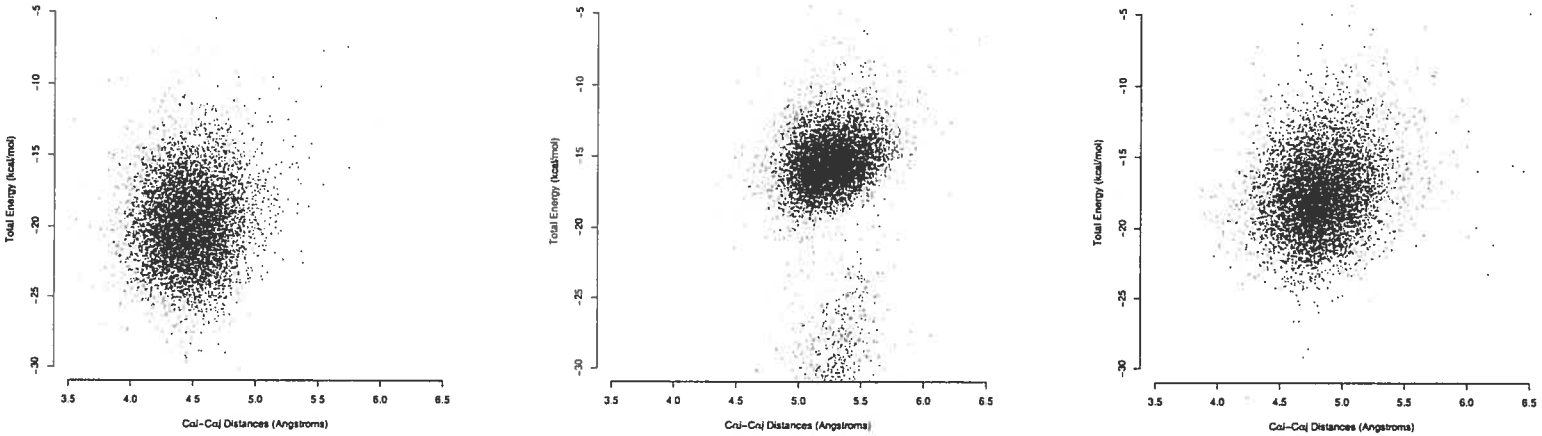
(b)

Table 6.6 – Residues that are not in β-sheets, but were annotated as so by DSSP in the 1.6Å resolution PDB Select 25 protein database. The residues are identified by the four-letter PDB codes. The fifth letter is used as the chain identifier.



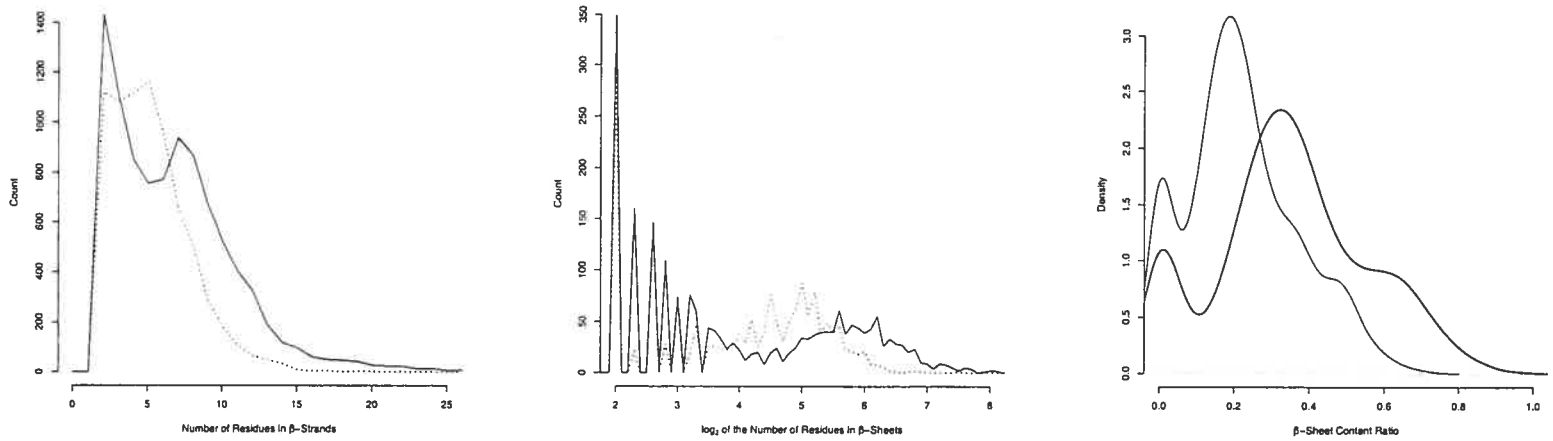
(a) (b) (c)

Figure 6.1 – Correlation between Coulomb electrostatic and van der Waals. For each of the three canonical H-bonding motifs, the Coulomb electrostatic energy along with its van der Waals have been compiled. A linear x-y plot is then produced to show any linear correlations between the two energy contributions. In each plot the Coulomb electrostatic is found on the abscissa while the van der Waals is on the ordinate. The data is extracted from the 1.6Å resolution PDB Select 25 database. **a)** The plot for the H-bonding motif I. The Pearson's R^2 correlation coefficient is 0.06. **b)** The plot for the H-bonding motif II. The Pearson's R^2 correlation coefficient is 0.57. **c)** The plot for the H-bonding motif III. The Pearson's R^2 correlation coefficient is 0.14.



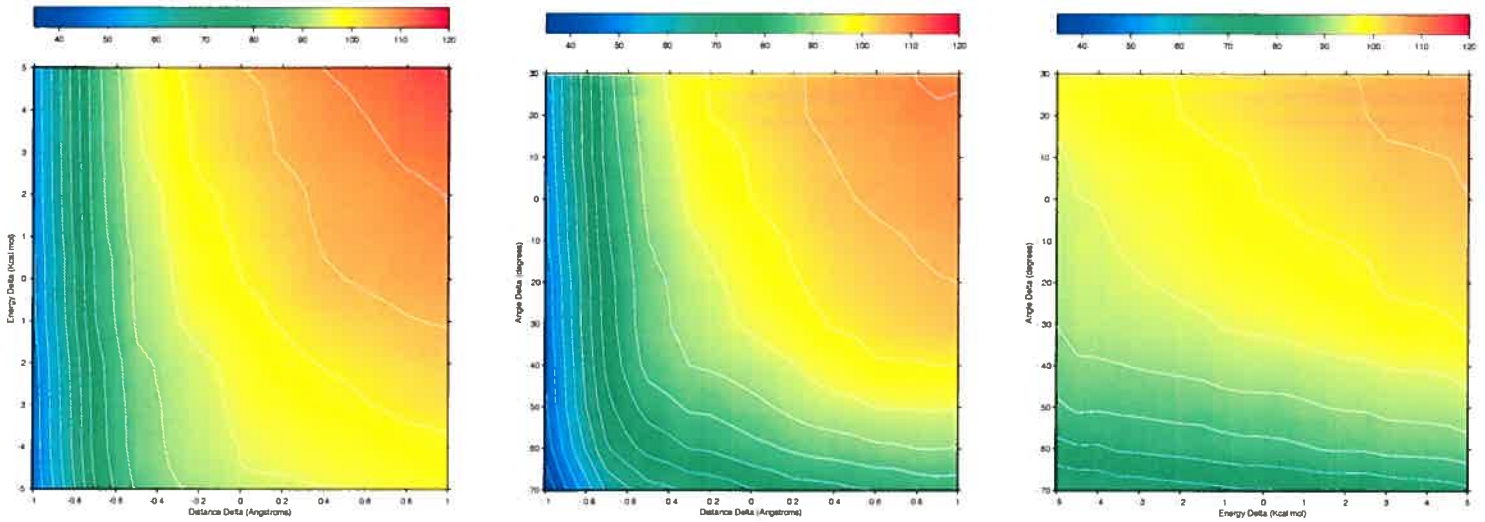
(a) (b) (c)

Figure 6.2 – Correlation between the $C_{\alpha i}$-$C_{\alpha j}$ distance and the total energy E^3_{ij}. For each of the three canonical H-bonding motifs, the $C_{\alpha i}$-$C_{\alpha j}$ distance, for β-sheet partners i and j, along with its total energy E^3_{ij} have been compiled. A linear x-y plot is then produced to show any linear correlations between the two distributions. The distance is on the abscissa, whereas the energy is on the ordinate. The data is extracted from the 1.6Å resolution PDB Select 25 database. **a)** H-bonding motif I. The Pearson's R^2 correlation coefficient is 0.01. **b)** H-bonding motif II. The Pearson's R^2 correlation coefficient is 0.01. **c)** H-bonding motif III. The Pearson's R^2 correlation coefficient is 0.04.



(a) (b) (c)

Figure 6.3 – β-strand lengths, β-sheet sizes and content ratios. The β-strand lengths and β-sheet sizes in number of residues for DSSP and β-Spider are compared. Also, the β-sheet content ratio is shown. In each plot the solid or bold line is used for β-Spider while the dotted or thin line is used for DSSP. The data is extracted from the 1.6Å resolution PDB Select 25 database. **a)** Distribution of β-strand lengths in number of residues. For DSSP (dotted line) the mean β-strand length is 5.2 residues with a standard deviation of 2.6. The median is 5 as is also the mode. The two peaks are at N=2 and N=5. For β-Spider (solid line) the mean β-strand length is 6.8 residues with a standard deviation of 3.6. The median is 6 as is also the mode. The two peaks are at N=3 and N=7. **b)** Distribution of β-sheet sizes in number of residues. For DSSP (dotted line) the mean β-sheet size is 22.4 residues with a standard deviation of 17.9. The median is 19 while the mode is 4. For β-Spider (solid line) the mean β-sheet size is 30.9 residues with a standard deviation of 36.3. The median is 12 while the mode is 4. For DSSP, the first major peaks are at $\log_2(N = 4) = 2.0$, $\log_2(N = 6) = 2.6$ and $\log_2(N = 8) = 3.0$, while for DSSP, the first major peaks are at $\log_2(N = 4) = 2.0$, $\log_2(N = 5) = 2.3$ and $\log_2(N = 6) = 2.6$ **c)** Density of the β-sheet content as a function of the content ratio. The ratio ranges from 0 (no residues in β-sheets) to 1 (all residues in β-sheets). The area under each curve is equal to unity. The mean β-sheet content ratio for DSSP is 0.22 (22%) while is 0.35 (35%) for β-Spider.



(a) (b) (c)

Figure 6.4 – Numerical stability of the β-Spider algorithm. The three parameters have been compared against one another near the selected cut-off values. In each graphic a total of 441 (21 times 21) grid points have been evaluated to the relative number of residues in β-sheets from the reference value of 100% at the cut-off values. The distance steps are of 0.1Å, the energy steps are of 0.5 Kcal/mol, and the torsion angle steps are of 5 degrees. Isobar lines are traced at each 5% level. The abscissa and ordinate values are relative to the cut-off values. Ten randomly chosen protein chains have been selected for the reference value (giving 1041 β-sheet residues), and were used throughout the grid point evaluations. The stability is assessed by the inter-space between isobars in each dimension near the cut-off values, i.e. when the delta amplitude is zero. **a)** The E^3_{ij} energy against the C^i_α-C^j_α distance. The torsion angle cut-offs are fixed at -128.0 and +68.4 degrees. **b)** The C^i_β-C^i_α-C^j_α-C^j_β torsion angle against the C^i_α-C^j_α distance. The energy cut-off is fixed at -8.2 kcal/mol. **c)** The C^i_β-C^i_α-C^j_α-C^j_β torsion angle against the E^3_{ij} energy. The C^i_α-C^j_α distance cut-off is fixed at 6.2Å. These images were produced with the help of gri (http ://gri.sourceforge.net/).

Step 1 Identification of β-Partners and β-Edges. The β-Partners and β-Edges are identified by enumerating all possible pairs of tri-peptides (lines 1 and 4) not part of an α-helix or β-turn identified in the PDB file under the HELIX and TURN records. When a pair of residues i and j satisfy both the energy criterion E(i,j) and the geometrical criteria G(i,j) (line 7) we make them β-Partners (line 10). Then, both flanking β-bridges are made temporarily β-Edges, considering that the strand pairs are parallel (lines 15 and 18) or anti-parallel (lines 21 and 24), as long as they satisfy the geometrical criteria and are not previously marked as β-Partners (lines 14, 17, 20 and 23). Ulteriorly, these β-Edges can be promoted to β-Partners after an energy evaluation (line 7).

```
1: for all residue i such that
   {i − 1,i,i + 1} ⊈ HELIX ∪ TURN do
2:   // consider the tri-peptide {i-1,i,i+1}
3:
4:   for all residue j such that
     (j > i + 5) and {j − 1,j,j + 1} ⊈ HELIX ∪ TURN do
5:     // consider the tri-peptide {j-1,j,j+1}
6:
7:     if E(i, j) and G(i, j) then
8:
9:       // assign the β-Partners
10:      let β-Partners(i, j) ← true
11:      let β-Edges(i, j) ← false
12:
13:      // assign the β-Edges
14:      if G(i − 1, j − 1) and not β-Partners(i − 1, j − 1) then
15:        let β-Edges(i − 1, j − 1) ← true
16:      end if
17:      if G(i + 1, j + 1) and not β-Partners(i + 1, j + 1) then
18:        let β-Edges(i + 1, j + 1) ← true
19:      end if
20:      if G(i − 1, j + 1) and not β-Partners(i − 1, j + 1) then
21:        let β-Edges(i − 1, j + 1) ← true
22:      end if
23:      if G(i + 1, j − 1) and not β-Partners(i + 1, j − 1) then
24:        let β-Edges(i + 1, j − 1) ← true
25:      end if
26:    end if
27:  end for
28: end for
```

Step 2 Identification of β-Bulges. The β-Bulges are identified by enumerating all possible residue pairs i and j (lines 1 and 2) which are β-Neighbors (line 4) and for which an adjacent residue $i+1$ on the non-bulged strand is also a β-Neighbor of a not too far residue $j+\Delta j$ on the bulged strand (line 10). The residues between j and $j+\Delta j$ are them marked as bulged (lines 12 and 13). The pseudo-code outlined here finds the β-Bulges on the C-terminal strand of a parallel pair of strands. For the detection of β-Bulges on the N-terminal strand simply exchange the i and j indexes in the code. For the anti-parallel case, use $j-\Delta j$ (line 10) and $j-jj$ (line 13) instead.

```
 1: for all residue i do
 2:   for all residue j such that j > i do
 3:
 4:     if β-Neighbors(i, j) then
 5:
 6:       // check bulges of 1 to 3 residues
 7:       for  Δj = 2 to 4  do
 8:
 9:         // j-bulged residues are flanked by β-Neighbors
10:         if β-Neighbors(i + 1, j + Δj) then
11:
12:           for  jj = 1 to Δj − 1  do
13:             let β-Bulge(j + jj) ← true
14:           end for
15:
16:         end if
17:       end for
18:     end if
19:   end for
20: end for
```

Step 3 Identification of β-Sheet domains. β-sheet domains are recursively identified by painting a β-Neighbored residue (line 4) and it's surrounding siblings (lines 24-26 and 35-37). Each domain will be painted with a distinct color (lines 6 and 7). β-Bulges are painted by following the strands N and C terminal directions (lines 21 and 32). β-Neighbors are painted in such a way that we impose that for a given pair of residues i and $i \pm 1$ have β-Neighbors k and l on the same strand, to within a length of a β-Bulge (lines 23 and 34). Then, these β-Neighbors are painted recursively (lines 24-26 and 35-37).

```
 1: Procedure PaintSheet
 2: let color ← 0
 3: for all residue i do
 4:   if not Painted(i) and β-Neighbor(i) then
 5:     // paint this residue
 6:     let color ← color+1
 7:     PaintSheetRec(color, i)
 8:   end if
 9: end for
10:
11: Procedure PaintSheetRec(color, residue i)
12: if not Painted(i) then
13:   if β-Neighbor(i) or β-Bulge(i) then
14:
15:     // paint this residue
16:     let Color(i) ← color
17:     let Painted(i) ← true
18:
19:     // paint C-terminal residue of i, that is residue i + 1
20:     if β-Bulge(i + 1) then
21:       PaintSheetRec(color, i + 1)
22:     else
23:       for all residues k, l such that
          |k − l| ≤ 4 and β-Neighbors(i, k) and β-Neighbors(i + 1, l) do
24:         PaintSheetRec(color, i + 1)
25:         PaintSheetRec(color, k)
26:         PaintSheetRec(color, l)
27:       end for
28:     end if
29:
30:     // paint N-terminal residue of i, that is residue i − 1
31:     if β-Bulge(i − 1) then
32:       PaintSheetRec(color, i − 1)
33:     else
34:       for all residues k, l such that
          |k − l| ≤ 4 and β-Neighbors(i, k) and β-Neighbors(i − 1, l) do
35:         PaintSheetRec(color, i − 1)
36:         PaintSheetRec(color, k)
37:         PaintSheetRec(color, l)
38:       end for
39:     end if
40:   end if
41: end if
```

CHAPITRE 7

CONCLUSION

La prédiction de la structure tri-dimensionelle d'une protéine à partir de sa séquence seule est l'un des problèmes majeurs encore ouverts en bio-informatique. Aucune combinaison de logiciel/matériel n'est capable, à ce jour, de traiter en calculs prédictifs des protéines de taille moyenne avec 450 acides-aminés. On doit cependant souligner les efforts notoires comme Folding@Home [164] et Blue Gene [165]. Toute approche informatique au chemin menant de la séquence à la structure tridimensionelle est alors bienvenue.

Dans un premier temps, nous avons proposé un espace de recherche conformationel dans lequel les éléments de structures secondaires sont approximés en utilisant des angles dihèdres fixes. Ensuite, ces éléments sont placés dans l'espace, les uns par rapports aux autres, par des relations spatiales extraites des base de données de structures déjà existantes, puis reproduites à l'aide des matrices de transformations homogènes. L'expérience de Jackknife montre que cet espace de recherche contient toutes les structures de protéines connues présentement.

En second lieu, nous avons défini les graphes de feuillets β qui encodent chaque acide-aminé en noeud et les relations de liens peptidiques, ponts hydrogène et partenaires β dans des arêtes aux couleurs distinguées. Nous avons développé un algorithme d'isomorphisme de sous-graphe, adapté à la topologie particulière des graphes de feuillets β, et nous permet alors de comparer ces graphes entre-eux. Entres autres, ceci nous permets de proposer des structures tri-dimensionelles de feuillets à partir de son graphe topologique (*de novo* protein design), d'analyser le contenu en acides-aminés pour chaque position dans un motif topologique particulier (β-bulges et β-barrels), et de comparer les graphes de feuillets de membres d'une même famille de protéine.

En troisième partie, nous avons développé une méthode permettant de reconstruire un feuillet β, non pas à partir de son graphe topologique en entier mais avec

un partitionement justicieux de ce graphe en plus petit morceaux. L'algorithme d'isomorphisme de sous-graphe précédemmant décrit est utilisé pour récupérer des structures 3-D de fragments du feuillets, qui sont ensuite rassemblés dans l'espace en une structure final qui satisfait en tout point le graphe topologique original. L'expérience de Jackknife nous montre qu'il est possible de reconstruire presque tous les feuillets β avec une précision atomique (i.e. de l'ordre de quelques Angstroms de différence).

Dans un quatrième mouvement, nous nous sommes penchés sur l'annotation des feuillets β dans les protéines, étant insatisfaits des résultats obtenus avec les programmes existants. En tenant compte de toutes les forces en jeu dans la stabilisation de deux chaînes peptidiques adjacentes, nous avons utilisés un champ de force classique pour l'estimation de cette énergie de cohésion. Il en résulte que les feuillets β sont maintenants plus étendus qu'on le croyait, avec un contenu en acides-aminés fort différent de ceux annotés par les programmes existants, et les paramètres structuraux (les angles ϕ et ψ) maintenant moins caractéristiques de cette classe d'élément de structure secondaire. Les conséquences pour la prédiction de la structure 3-D, du moins celle des brins β, sont catastrophiques.

Enfin, dans un dernier élan, nous avons comparés entres eux les différents facteurs d'influence dans la formation des feuillets β pour en déterminer la force motrice principale. Ces facteurs sont l'hydrophobicité des faces du feuillet, le nombre de ponts hydrogène, la fréquence des acides-aminés dans les brins ainsi que la préférence d'appariement des acides-aminés en partenaires β. Une fonction de score à alors été esquissée pour capturer ces facteurs. L'optimisation de cette fonction de score pour distinguer les topologies natives parmies toutes les topologies alternatives révèle que la capacité d'un feuillet à construire une face hydrophobique est de la plus haute importance. Tandis que l'appariement des acides-aminés au sein du feuillet semble sans influence.

BIBLIOGRAPHIE

[1] M Friedman. Formation, nutritional value, and safety of d-amino acids. *Adv Exp Med Biol*, 289 :447–481, 1991.

[2] L Pauling, RB Corey, and HR Branson. The structure of proteins : Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci (USA)*, 37 :205–211, 1951.

[3] L Pauling and RB Corey. The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci (USA)*, 37 :251–256, 1951.

[4] JF Richardson and DC Richardson. Principles and patterns of protein conformation. In GD Fasman, editor, *Prediction of protein structure and the principles of protein conformation*, pages 1–98. Plenum Press, New York, 1989.

[5] JS Richardson. Schematic drawings of protein structures. *Methods Enzymol*, 115 :359–380, 1985.

[6] CB Anfinsen, RR Redfield, WL Choate, and WR Carroll. Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *J Biol Chem*, 207 :201–210, 1954.

[7] CB Anfinsen, E Haber, M Sela, and FH White. The kinematics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci (USA)*, 47(9) :1309–1314, september 1961.

[8] HS Chan and KA Dill. The protein folding problem. *Physics Today*, 46 :24–32, 1993.

[9] MJ Rooman, JPA Kocher, and SJ Wodak. Prediction of protein backbone conformations based on seven structure assignments. *J Mol Biol*, 221, 1991.

[10] J Zhang. Protein-length distributions for the three domains of life. *Trends Genet*, 16 :107–109, 2000.

[11] L Pauling and RB Corey. Configurations of polypeptide chains with favored orientations around single bonds : Two new pleated sheets. *Proc Natl Acad Sci (USA)*, 37 :729–740, 1951.

[12] RA Sayle and EJ Milner-White. Rasmol : Biomolecular graphics for all. *Trends Biol Sci*, 20 :374–376, 1995.

[13] M Parisien and F Major. A new catalog of protein β-sheets. *Proteins*, 61 :545–558, 2005.

[14] F Major, M Turcotte, D Gautheret, G Lapalme, E Fillion, and R Cedergren. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science*, 253 :1255–1260, 1991.

[15] M Parisien, MC Peitsch, and F Major. A protein conformational search space defined by secondary structure contacts. *Pac Symp Biocomput*, 243 :425–436, 1998.

[16] EM Mitchell, PJ Artymiuk, DW Rice, and P Willett. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J Mol Biol*, 212 :151–166, 1990.

[17] I Koch, F Kaden, and J Selbig. Analysis of protein sheet topologies by graph theoretical methods. *Proteins*, 12 :314–323, 1992.

[18] PJ Artymiuk, HM Grindley, AR Poirrette, DW Rice, EC Ujah, and P Willett. Identification of β-sheet motifs, of ψ-loops, and of patterns of amino acid residues in three-dimensional protein structures using a subgraph-isomorphism algorithm. *J Chem Inf Comput Sci*, 34 :54–62, 1994.

[19] JR Ullmann. An algorithm for subgraph isomorphism. *Journal of the ACM*, 23 :31–42, 1976.

[20] W Kabsch and C Sander. Dictionary of protein secondary structure : pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22 :2577–2637, 1983.

[21] PH Maccallum, R Poet, and EJ Milner-White. Coulombic interactions between partially charged main-chain atoms not hydrogen-bonded to each other influence the conformations of alpha-helices and antiparallel beta-sheet. a new method for analysing the forces between hydrogen bonding groups in proteins includes all the coulombic interactions. *J Mol Biol*, 248 :361–373, 1995.

[22] RF Doolittle. Similar amino acid sequences : chance or common ancestry ? *Science*, 214 :149–159, 1981.

[23] P Bork, C Ouzounis, C Sander, M Scharf, R Schneider, and E Sonnhammer. What's in a genome. *Nature*, 358 :287, 1992.

[24] C Chothia. One thousand folds for the molecular biologist. *Nature*, 357 :543–544, 1992.

[25] G Casari, MA Andrade, P Bork, J Boyle, A Daruvar, C Ouzounis, R Schneider, J Tamames, A Valencia, and C Sander. Challenging times for bioinformatics. *Nature*, 376 :647–648, 1995.

[26] J Moult. The current state of the art in protein structure prediction. *Curr. Opin. Biotech.*, 7 :422–427, 1996.

[27] Ncbi-genbank. http ://www.ncbi.nlm.nih.gov/, 1997.

[28] K Asai, S Hayamizu, and K Handa. Prediction of protein secondary structure by the hidden markov model. *CABIOS*, 9 :141–146, 1993.

[29] A Krogh, M Brown, IS Mian, K Sjolander, and D Haussler. Hidden markov models in computational biology. applications to protein modeling. *J Mol Biol*, 235 :1501–1531, 1994.

[30] N Quian and TJ Sejnowski. Predicting the secondary tructure of globular proteins using neural network models. *J Mol Biol*, 202 :865, 1988.

[31] LH Holley and M Karplus. Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci (USA)*, 86 :152–156, 1989.

[32] TJ Hubbard. Use of b-strand interaction pseudo-potentials in protein structure prediction and modelling. In R.H. Lathrop, editor, *Protein Structure Prediction MiniTrack of the 27th HICSS*, pages 336–354. IEEE Computer Society Press, 1994.

[33] W Zhang, TE Smithgall, and WH Gmeiner. Sequential assignment and secondary structure determination for the Src homology2 domain of hematopoietic cellular kinase. *FEBS Lett*, 406 :131–135, 1997.

[34] H Ponstingl and G Otting. NMR assignments, secondary structure and hydration of oxidized *escherichia coli* flavodoxin. *Eur J Biochem*, 244 :384–399, 1997.

[35] FE Cohen, MJE Sternberg, and WR Taylor. Analysis and prediction of the packing of α-helices against a β-sheet in tertiary structure of globular pproteins. *J Mol Biol*, 156 :821–862, 1982.

[36] KC Chou, G Nemethy, S Rumsey, RW Tuttle, and HA Scheraga. Interactions between an alpha-helix and a beta-sheet. energetics of alpha/beta packing in proteins. *J Mol Biol*, 186 :591–609, 1985.

[37] D Gautheret, F Major, and R Cedergren. Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *J Mol Biol*, 229 :1049–1064, 1993.

[38] F Major, D Gautheret, and R Cedergren. Reproducing the three-dimensional structure of a transfer RNA molecule from structural constraints. *Proc Natl Acad Sci (USA)*, 90 :9408–9412, 1993.

[39] K Yue and KA Dill. Folding proteins with a simple energy function and extensive conformational searching. *Protein Sci*, 5 :254–261, 1996.

[40] SH Bryant and CE Lawrence. An empirical energy function for threading protein sequence through the folding motif. *Proteins*, 16 :92–112, 1993.

[41] JR Gunn. Sampling protein conformations using segment libraries and a genetic algorithm. *J. Chem. Phys.*, 106 :4270–4281, 1997.

[42] TR Defay and FE Cohen. Protein modeling. In Robert A. Meyers ed., editor, *Encyclopedia of Molecular Biology and Molecular Medicine*, volume 5, pages 158–169. VCH Publishers Inc., New-York, NY, 1996.

[43] P Herzyk and RE Hubbard. An automated method for modelling seven-helix transmembrane receptors from experimental data. *Biophysical J.*, 69 :2419–2442, 1995.

[44] RP Paul. *Robot Manipulators : Mathematics, Programming, and Control.* MIT Press, Cambridge, 1981.

[45] FC Bernstein, TF Koetzle, GJB Williams, EF Meyer Jr., MD Brice, JR Rodgers, O Kennard, T Shimanouchi, and M Tasumi. The protein data-bank : A computer based archival file for molecular structures. *Eur. J. Biochem.*, 80 :319–324, 1977.

[46] JS Richardson. The Anatomy and Taxonomy of Protein Structure. *Advances in Proteins Chemistry*, 34 :167–339, 1981.

[47] U Hobohm, M Scharf, R Schneider, and C Sander. Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Science*, 1 :409–417, 1992.

[48] U Hobohm and C Sander. Enlarged representative set of protein structures. *Protein Science*, 3 :522, 1994.

[49] B. Efron. *The Jack Knife, the bootsrap and other resampling plans.* Society for industrial and applied mathematics, Philadelphia, PA, 1982.

[50] EC van Geerestein-Ujah, M Mariani, H Vis, R Boelens, and R Kaptein. Use of graph theory for secondary structure recognition and sequential assignment in heteronuclear (13C, 15N) NMR spectra : application to HU protein from Bacillus stearothermophilus. *Biopolymers*, 7 :691–707, 1996.

[51] C Bailey-Kellogg, A Widge, JJ Kelley, MJ Berardi, JH Bushweller, and BR Donald. The NOESY jigsaw : automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J Comput Biol*, 7 :537–558, 2000.

[52] HM Grindley, PJ Artymiuk, DW Rice DW, and Willett P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J Mol Biol*, 229 :707–721, 1993.

[53] PJ Artymiuk, AR Poirrette, HM Grindley, DW Rice, and P Willett. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J Mol Biol*, 243 :327–344, 1994.

[54] A Harrison, F Pearl, I Sillitoe, T Slidel, R Mott, J Thornton, and C Orengo. Recognizing the fold of a protein structure. *Bioinformatics*, 19 :1748–1759, 2002.

[55] JS Richardson. beta-sheet topology and the relatedness of proteins. *Nature*, 268 :495–500, 1977.

[56] AG Murzin, SE Brenner, T Hubbard, and C Chothia. SCOP : A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247 :536–540, 1995.

[57] AV Efimov. A structural tree for proteins containing s-like beta-sheets. *FEBS Lett*, 437 :246–250, 1998.

[58] SW Michnick and E Shakhnovich. A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Fold Des*, 3 :239–251, 1998.

[59] N Kannan, S Selvaraj, MM Gromiha, and S Vishveshwara. Clusters in alpha/beta barrel proteins : implications for protein structure, function, and folding : a graph theoretical approach. *Proteins*, 43 :103–112, 2001.

[60] R Qamra, B Taneja, and SC Mande. Identification of conserved residue patterns in small beta-barrel proteins. *Protein Eng*, 15 :967–977, 2002.

[61] AW Overhauser. Paramagnetic relaxation in metals. *Phys Rev*, 89 :689–700, 1953.

[62] AW Overhauser. Polarization of nuclei in metals. *Phys Rev*, 92 :411–415, 1953.

[63] I Solomon. Relaxation processes in a system of two spins. *Phys Rev*, 99 :559–565, 1955.

[64] M Delepierre, CM Dobson, and FM Poulsen. Studies of beta-sheet structure in lysozyme by proton nuclear magnetic resonance. assignments and analysis of spin-spin coupling constants. *Biochemistry*, 21 :4756–4761, 1982.

[65] F Inagaki, NJ Clayden, N Tamiya, and RJ Williams. Individual assignments of the amide proton resonances involved in the triple-stranded antiparallel pleated beta-sheet structure of a long neurotoxin, Laticauda semifasciata III from Laticauda semifasciata. *Eur J Biochem*, 123 :99–104, 1982.

[66] KH Mayo. Epidermal growth factor from the mouse. physical evidence for a tiered beta-sheet domain : two-dimensional NMR correlated spectroscopy and nuclear Overhauser experiments on backbone amide protons. *Biochemistry*, 24 :3783–3794, 1985.

[67] GT Montelione, K Wuthrich, EC Nice, AW Burgess, and HA Scheraga. Identification of two anti-parallel beta-sheet conformations in the solution structure of murine epidermal growth factor by proton magnetic resonance. *Proc Natl Acad Sci (USA)*, 83 :8594–8598, 1986.

[68] PL Weber, SC Brown, and L Mueller. Sequential ^{1}H NMR assignments and secondary structure identification of human ubiquitin. *Biochemistry*, 26 :7282–7290, 1987.

[69] BJ Stockman, AM Krezel, JL Markley, KG Leonhardt, and NA Straus. Hydrogen-1, carbon-13, and nitrogen-15 NMR spectroscopy of Anabaena 7120 flavodoxin : assignment of beta-sheet and flavin binding site resonances and analysis of protein-flavin interactions. *Biochemistry*, 29 :9600–9609, 1990.

[70] TC Pochapsky and XM Ye. 1H NMR identification of a beta-sheet structure and description of folding topology in putidaredoxin. *Biochemistry*, 30 :3850–3856, 1991.

[71] CG Ullman, PI Haris, KF Smith, RB Sim, VC Emery, and SJ Perkins. Beta-sheet secondary structure of an LDL receptor domain from complement factor

I by consensus structure predictions and spectroscopy. *FEBS Lett*, 371 :199–203, 1995.

[72] CG Ullman, PI Harris, DA Galloway, VC Emery, and SJ Perkins. Predicted alpha-helix/beta-sheet secondary structures for the zinc-binding motifs of human papillomavirus E7 and E6 proteins by consensus prediction averaging and spectroscopic studies of E7. *Biochem J*, 319 :229–239, 1996.

[73] K Maruyama, Y Itoh, and F Arisaka. Circular dichroism spectra show abundance of beta-sheet structure in connectin, a muscle elastic protein. *FEBS Lett*, 202 :353–355, 1986.

[74] A Perczel, K Park, and GD Fasman. Deconvolution of the circular dichroism spectra of proteins : the circular dichroism spectra of the antiparallel beta-sheet in proteins. *Proteins*, 13 :57–69, 1992.

[75] D Frishman and P Argos. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng*, 9 :133–142, 1996.

[76] JT Yang. Prediction of protein secondary structure from amino acid sequence. *J Protein Chem*, 15 :185–191, 1996.

[77] A Galat. A note on circular-dichroic-constrained prediction of protein secondary structure. *Eur J Biochem*, 236 :428–435, 1996.

[78] M Asogawa. Beta-sheet prediction using inter-strand residue pairs and refinement with Hopfield neural network. *ISMB*, 5 :48–51, 1997.

[79] V Di Francesco, P McQueen, J Garnier, and PJ Munson. Incorporating global information into secondary structure prediction with hidden markov models of protein folds. *ISMB*, 5 :100–103, 1997.

[80] M Clementi, S Clementi, G Cruciani, M Pastor, AM Davis, and DR Flower. Robust multivariate statistics and the prediction of protein secondary structure content. *Protein Eng*, 10 :747–749, 1997.

[81] M Ito, Y Matsuo, and K Nishikawa. Prediction of protein secondary structure using the 3D-1D compatibility algorithm. *CABIOS*, 4 :415–424, 1997.

[82] BK Ho and PM Curmi. Twist and shear in beta-sheets and beta-ribbons. *J Mol Biol*, 317 :291–308, 2002.

[83] EG Emberly, R Mukhopadhyay, C Tang, and NS Wingreen. Flexibility of beta-sheets : principal component analysis of database protein structures. *Proteins*, 55 :91–98, 2004.

[84] G Wang and RL Dunbrack Jr. PISCES : a protein sequence culling server. *Bioinformatics*, 19 :1589–1591, 2003.

[85] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The Protein Data Bank. *Nucl Acids Res*, 28 :235–242, 2000.

[86] MR Garey and DS Johnson. *Computers and Intractability : A Guide to the Theory of NP-Completeness.* Freeman & co., New York, 1979.

[87] J Hopcroft and J Wong. Linear time algorithm for isomorphism of planar graphs. In *6th Annual ACM Symposium on Theory of Computing*, pages 172–184, 1974.

[88] MO Dayhoff, RM Schwartz, and BC Orcutt. A model of evolutionary change in proteins. In M.O. Dayhoff, editor, *Atlas of Protein Science and Structure*, volume 5, Suppl. 3, pages 345–352, Silver Spring, MD, 1978. National Biomedical Research Foundation.

[89] CE Shannon and W Weaver. *The mathematical theory of communication.* University of Illinois Press, Urbana, IL, 1949.

[90] S Kullback and RA Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22 :79–86, 1951.

[91] K Fan and W Wang. What is the minimum number of letters required to fold a protein? *J Mol Biol*, 328 :921–926, 2003.

[92] T Li, K Fan, J Wang, and W Wang. Reduction of protein sequence complexity by residue grouping. *Protein Eng*, 16 :323–330, 2003.

[93] J Wang and W Wang. Grouping of residues based on their contact interactions. *Phys Rev E*, 65 :041911, 2002.

[94] S Miyazawa and RL Jernigan. Residue - residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, 256 :623–644, 1996.

[95] C Zhang and SH Kim. Environment-dependent residue contact energies for proteins. *Proc Natl Acad Sci (USA)*, 97 :2550–2555, 2000.

[96] S Lifson and C Sander. Antiparallel and parallel β-strands differ in amino acid residue preferences. *Nature*, 282 :109–111, 1979.

[97] S Lifson and C Sander. Specific recognition in the tertiary structure of beta-sheets of proteins. *J Mol Biol*, 139 :627–639, 1980.

[98] B Kuhlman, G Dantas, GC Ireton, G Varani, BL Stoddard, and D Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302 :1364–1368, 2003.

[99] TM Handel and PJ Domaille. Heteronuclear (1H, 13C, 15N) NMR assignments and solution structure of the monocyte chemoattractant protein-1 (MCP-1) dimer. *Biochemistry*, 35 :6569–6584, 1996.

[100] J Lubkowski, G Bujacz, L Boque, PJ Domaille, TM Handel, and A Wlodawer. The structure of MCP-1 in two crystal forms provides a rare example of variable quaternary interactions. *Nat Struct Biol*, 4 :64–69, 1997.

[101] JS Richardson and DC Richardson. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci (USA)*, 99 :2754–2759, 2002.

[102] AW Chan, EG Hutchinson, D Harris, and JM Thornton. Identification, classification, and analysis of beta-bulges in proteins. *Protein Sci*, 2 :1574–1590, 1993.

[103] AD McLachlan. Gene duplications in the structural evolution of chymotrypsin. *J Mol Biol*, 128 :49–79, 1979.

[104] M Lu and TA Steitz. Structure of escherichia coli ribosomal protein l25 complexed with a 5s rrna fragment at 1.8-a resolution. *Proc Natl Acad Sci (USA)*, 97 :2023–2028, 2000.

[105] S Vijay-Kumar, CE Bugg, and WJ Cook. Structure of ubiquitin refined at 1.8 a resolution. *J Mol Biol*, 194 :531–544, 1987.

[106] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28 :235–242, 2000.

[107] C Chothia. Conformation of twisted b-pleated sheets in proteins. *J Mol Biol*, 75 :295–302, 1973.

[108] D Znamenskiy, KL Tuan, A Poupon, J Chomilier, and JP Mormon. Beta-sheet modeling by helical surfaces. *Protein Engineering*, 13 :407–412, 2000.

[109] K Nishikawa and HA Scheraga. Geometrical criteria for formation of coiled-coil structures of polypeptide chains. *Macromolecules*, 9 :395–407, 1976.

[110] FR Salemme and DW Weatherford. Conformational and geometrical properties of beta-sheets in proteins. I. parallel beta-sheets. *J Mol Biol*, 146 :101–117, 1981.

[111] FR Salemme and DW Weatherford. Conformational and geometrical properties of beta-sheets in proteins. II. antiparallel and mixed beta-sheets. *J Mol Biol*, 146 :119–141, 1981.

[112] HJ Feldman and CW Hogue. A fast method to sample real protein conformational space. *Proteins*, 39 :112–131, 2000.

[113] HJ Feldman and CW Hogue. Probabilistic sampling of protein conformations : new hope for brute force? *Proteins*, 46 :8–23, 2002.

[114] KT Simons, C Kooperberg, E Huang, and D Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*, 268 :209–225, 1997.

[115] KT Simons, I Ruczinski, C Kooperberg, BA Fox, C Bystroff, and D Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, 34 :82–95, 1999.

[116] KT Simons, R Bonneau, I Ruczinski, and D Baker. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins*, Suppl 3 :171–176, 1999.

[117] I Ruczinski, C Kooberberg, R Bonneau, and D Baker. Distributions of beta-sheets in proteins with application to structure prediction. *Proteins*, 48 :85–97, 2002.

[118] A Liwo, S Oldziej, MR Pincus, RJ Wawak, S Rackovsky, and HA Scheraga. A united-residue force field for off-lattice protein-structure simulations. I. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J Comp Chem*, 18 :849–873, 1997.

[119] A Liwo, MR Pincus, RJ Wawak, S Rackovsky, S Oldziej, and HA Scheraga. A united-residue force field for off-lattice protein-structure simulations. II. parameterization of local interactions and determination of the weights of energy terms by Z-score optimization. *J Comp Chem*, 18 :874–887, 1997.

[120] A Liwo, R Kazmierkiewicz, C Czaplewski, M Groth, S Oldziej, RJ Wawak, S Rackovsky, MR Pincus, and HA Scheraga. United-residue force field for off-lattice protein-structure simulations; III. origin of backbone hydrogen-bonding cooperativity in united-residue potentials. *J Comp Chem*, 19 :259–276, 1998.

[121] A Liwo, P Arlukowicz, C Czaplewski, S Oldziej, J Pillardy, and HA Scheraga. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape : application to the UNRES force field. *Proc Natl Acad Sci (USA)*, 99 :1937–1942, 2002.

[122] J Lee, HA Scheraga, and S Rackovsky. New optimization method for conformational energy calculations on polypeptides : Conformational space annealing. *J Comp Chem*, 18 :1222–1232, 1997.

[123] J Lee, HA Scheraga, and S Rackovsky. Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. *Biopolymers*, 46 :103–115, 1998.

[124] J Lee, A Liwo, and HA Scheraga. Energy-based *de-novo* protein folding by conformational space annealing and an off-lattice united-residue force field : Application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K. *Proc Natl Acad Sci (USA)*, 96 :2025–2030, 1999.

[125] R Srinivasan and GD Rose. LINUS : A simple algorithm to predict the fold of a protein. *Proteins*, 22 :81–99, 1995.

[126] R Srinivasan and GD Rose. A physical basis for protein secondary structure. *Proc Natl Acad Sci (USA)*, 96 :14258–14263, 1999.

[127] R Srinivasan and GD Rose. Ab initio prediction of protein structure using LINUS. *Proteins*, 47 :489–495, 2002.

[128] D Kihara, H Lu, A Kolinski, and J Skolnick. TOUCHSTONE : an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci (USA)*, 98 :10125–10130, 2001.

[129] D Kihara, Y Zhang, H Lu, A Kolinski, and J Skolnick. Ab initio protein structure prediction on a genomic scale : application to the Mycoplasma genitalium genome. *Proc Natl Acad Sci (USA)*, 99 :5993–5998, 2002.

[130] PJ Kraulis. Molscript : A program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography*, 24 :946–950, 1991.

[131] EA Merritt and DJ Bacon. Raster3D : Photorealistic Molecular Graphics. *Methods in Enzymology*, 277 :505–524, 1997.

[132] PY Chou and GD Fasman. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*, 13 :211, 1974. (Publication No.922 from the Graduate Departement of Biochemistry, Brandeis University).

[133] J Garnier, DJ Osguthorpe, and B Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol*, 120 :97–120, 1978.

[134] G Pollastri, D Przybylski, B Rost, and P Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, 47 :228–235, 2002.

[135] JA Siepen, SE Radford, and DR Westhead. Beta edge strands in protein structure prediction and aggregation. *Protein Sci*, 12 :2348–2359, 2003.

[136] CA Orengo, AD Michie, S Jones, DT Jones, MB Swindells, and JM Thornton. CATH - a hierarchic classification of protein domain structures. *Structure*, 5 :1093–1108, 1997.

[137] FMG Pearl, D Lee, JE Bray, I Sillitoe, AE Todd, AP Harrison, JM Thornton, and CA Orengo. Assigning genomic sequences to CATH. *Nucl Acids Res*, 28 :277–282, 2000.

[138] CC Blake, DF Koenig, GA Mair, AC North, DC Phillips, and VR Sarma. Structure of hen egg-white lysozyme. Ax three-dimensional fourier synthesis at 2 angstrom resolution. *Nature*, 206 :757–761, 1965.

[139] FM Richards and CE Kundrot. Identification of structural motifs from protein coordinate data : secondary structure and first-level super secondary structure. *Proteins*, 3 :71–84, 1988.

[140] G Labesse, N Colloc'h, J Pothier, and JP Mornon. P-SEA : a new efficient assignment of secondary structure from c alpha trace of proteins. *Comput Appl Biosci*, 13 :291–295, 1997.

[141] SM King and WC Johnson. Assigning secondary structure from protein coordinate data. *Proteins*, 35 :313–320, 1999.

[142] WR Taylor. Defining linear segments in protein structure. *J Mol Biol*, 310 :1135–1150, 2001.

[143] F Dupuis, JF Sadoc, and JP Mornon. Protein secondary structure assignment through Voronoi tessellation. *Proteins*, 55 :519–528, 2004.

[144] H Sklenar, C Etchebest, and R Lavery. Describing protein structure : a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins*, 6 :46–60, 1989.

[145] CA Andersen, AG Palmer, S Brunak, and B Rost. Continuum secondary structure captures protein flexibility. *Structure (Camb)*, 10 :175–184, 2002.

[146] P Carter, CAF Andersen, and B Rost. DSSPcont : continuous secondary structure assignments for proteins. *Nucl Acids Res*, 31 :3293–3295, 2003.

[147] D Frishman and P Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23 :566–579, 1995.

[148] N Colloc'h, C Etchebest, E Thoreau, B Henrissat, and JP Mornon. Comparison of three algorithms for the assignment of secondary structure in proteins : the advantages of a consensus assignment. *Prot Eng*, 6 :377–382, 1993.

[149] AG Street and SL Mayo. Intrinsic beta-sheet propensities result from van der waals interactions between side chains and the local backbone. *Proc Natl Acad Sci (USA)*, 96 :9074–9076, 1999.

[150] PH Maccallum, R Poet, and EJ Milner-White. Coulombic attractions between partially charged main-chain atoms stabilise the right-handed twist found in most beta-strands. *J Mol Biol*, 248 :374–384, 1995.

[151] R Taylor and O Kennard. Crystallographic evidence for the existence of C-H...O, C-H...N, and C-H...Cl hydrogen bonds. *J Am Chem Soc*, 104 :5063–5070, 1982.

[152] R Preissner, U Egner, and W Saenger. Occurrence of bifurcated three-center hydrogen bonds in proteins. *FEBS Lett*, 288 :192–196, 1991.

[153] ZS Derewenda, L Lee, and U Derewenda. The occurrence of C-H...O hydrogen bonds in proteins. *J Mol Biol*, 252 :248–262, 1995.

[154] R Vargas, J Garza, DA Dixon, and BP Hay. How strong is the Calpha-H...O=C hydrogen bond ? *J Am Chem Soc*, 122 :4750–4755, 2000.

[155] A Senes, I Ubarretxena-Belandia, and DM Engelman. The Calpha-H...O hydrogen bond : a determinant of stability and specificity in transmembrane helix interactions. *Proc Natl Acad Sci (USA)*, 98 :9056–9061, 2001.

[156] DN Boobbyer, PJ Goodford, PM McWhinnie, and RC Wade. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J Med Chem*, 32 :1083–1094, 1989.

[157] AT Hagler, E Huler, and S Lifson. Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *J Am Chem Soc*, 96 :5319–5327, 1974.

[158] AT Hagler and S Lifson. Energy functions for peptides and proteins. II. The amide hydrogen bond and calculation of amide crystal properties. *J Am Chem Soc*, 96 :5327–5335, 1974.

[159] WD Cornell, P Cieplak, CI Bayly, IR Gould, KM Merz, DM Ferguson, DC Spellmeyer, T Fox, JW Caldwell, and PA Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc*, 117 :5179–5197, 1995.

[160] AS Yang and B Honig. Free energy determinants of secondary structure formation : II. Antiparallel beta-sheets. *J Mol Biol*, 252 :366–376, 1995.

[161] F Avbelj and J Moult. Role of electrostatic screening in determining protein main chain conformational preferences. *Biochemistry*, 34 :755–764, 1995.

[162] M Levitt and J Greer. Automatic identification of secondary structure in globular proteins. *J Mol Biol*, 114 :181–239, 1977.

[163] R Ihaka and R Gentleman. R : a language for data analysis and graphics. *J Comp and Graphical Statistics*, 5(3) :299–314, 1996.

[164] MR Shirts and VS Pande. Screen savers of the World, unite! *Science*, 290 :1903–1904, 2000.

[165] F Allen, P Coteus, P Crumley, A Curioni, M Denneau, W Donath, M Eleftheriou, B Fitch, B Fleischer, CJ Georgiou, R Germain, G Almasi, M Giampapa, D Gresh, M Gupta, R Haring, H Ho, P Hochschild, S Hummel, T Jonas, D Lieber, G Martyna, W Andreoni, K Maturu, J Moreira, D Newns, M Newton, R Philhower, T Picunko, J Pitera, M Pitman, R Rand, A Royyuru, D Beece, V Salapura, A Sanomiya, R Shah, Y Sham, S Singh, M Snir, F Suits, R Swetz, WC Swope, N Vishnumurthy, BJ Berne, TJC Ward, H Warren,

R Zhou, A Bright, J Brunheroto, C Cascaval, and J Castanos. Blue Gene : a vision for protein science using a petaflop supercomputer. *IBM Systems Journal*, 40 :310–327, 2001.

Annexe I

Matrice de Transformation Homogène

I.1 La matrice

Un point V dans $\Re^3$ tel que $V = (x, y, z)$ peut être déplacé dans cet espace. Une translation $Tr = (T_x, T_y, T_z)$ amènera le point V au point V' de cette façon :

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} T_x \\ T_y \\ T_z \end{pmatrix} \tag{I.1}$$

Ou, sous forme compacte :

$$V' = V + Tr \tag{I.2}$$

Similairement, une rotation Rot déplacera le point V au point V' par :

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{pmatrix} \times \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{I.3}$$

Ou, sous forme compacte :

$$V' = Rot \times V \tag{I.4}$$

La matrice de rotation Rot a la particularité de voir son déterminant toujours égal à 1 peut importe les magnitudes des rotations. Dans le cas où la rotation est nulle cette matrice devient la matrice identité I.

Les deux opérations, la translation et la rotation, peuvent être combinées en un mouvement plus complexe :

$$V' = Tr + Rot \times V \tag{I.5}$$

Notez que dans cette dernière si la translation est nulle on revient à l'équation I.4, tandis que si la rotation est nulle on revient à l'équation I.2, puisque maintenant on aura $Rot = I$, la matrice identité.

Réecrivons l'équation I.5 en n'utilisant qu'une seule matrice :

$$\begin{pmatrix} x' \\ y' \\ z' \\ 1 \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} & R_{13} & T_x \\ R_{21} & R_{22} & R_{23} & T_y \\ R_{31} & R_{32} & R_{33} & T_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{I.6}$$

Sous forme compacte :

$$V' = M \times V \tag{I.7}$$

Où M est la matrice de transformation homogène (MTH), et encode la rotation et la translation pour passer d'un point V à un point V'. Le livre de Paul [44] parle en détails des MTHs. Il est important de noter ici que le produit de MTHs n'est pas commutatif mais associatif (comme un produit de matrices ordinaires).

I.2 Un référentiel

Un référentiel R dans $\Re^3$ se définit par trois vecteurs orthonormés $\vec{R_x}, \vec{R_y}, \vec{R_z}$ dont la particularité est que le produit vectoriel $\vec{R_x}$ par $\vec{R_y}$ donne $\vec{R_z}$ (la règle de la main droite). Un Référentiel R peut être encodé dans une matrice de transformation homogène M_R en supposant un référentiel global R_G (voir Figure I.1).

I.3 Les opérations

Plusieurs opérations avec les matrices de transformations homogènes sont alors possibles. L'ultime but est recréer la relation spatiale (translation et rotation) entre

deux référentiels R'_1 et R'_2 telle qu'observée entre les référentiels R_1 et R_2.

Avec la relations suivantes :

$$M_{R_i}{}^{-1} \times M_{R_i} = I \tag{I.8}$$

$$M_{R_1} \times M_{R_{1\to 2}} = M_{R_2} \tag{I.9}$$

On a alors que :

$$M_{R_{1\to 2}} = M_{R_1}{}^{-1} \times M_{R_2} \tag{I.10}$$

La Figure I.1 montre bien l'équation I.9 où le chemin M_{R_2} menant à R_2 est celui composé des chemins M_{R_1} vers R_1 puis $M_{R_{1\to 2}}$ vers R_2.

De là, en déplaçant R'_2 vers R''_2 par rapport à R'_1 on imite la relation de R_2 par rapport à R_1. Le produit matriciel suivant s'applique alors aux coordonnées de R'_2 :

$$M_{R''_2} = \left(M_{R'_1} \times M_{R_{1\to 2}} \times M_{R'_2}{}^{-1}\right) \times M_{R'_2} \tag{I.11}$$

Par exemple, si on fait correspondre R'_1 à R_1 on aura :

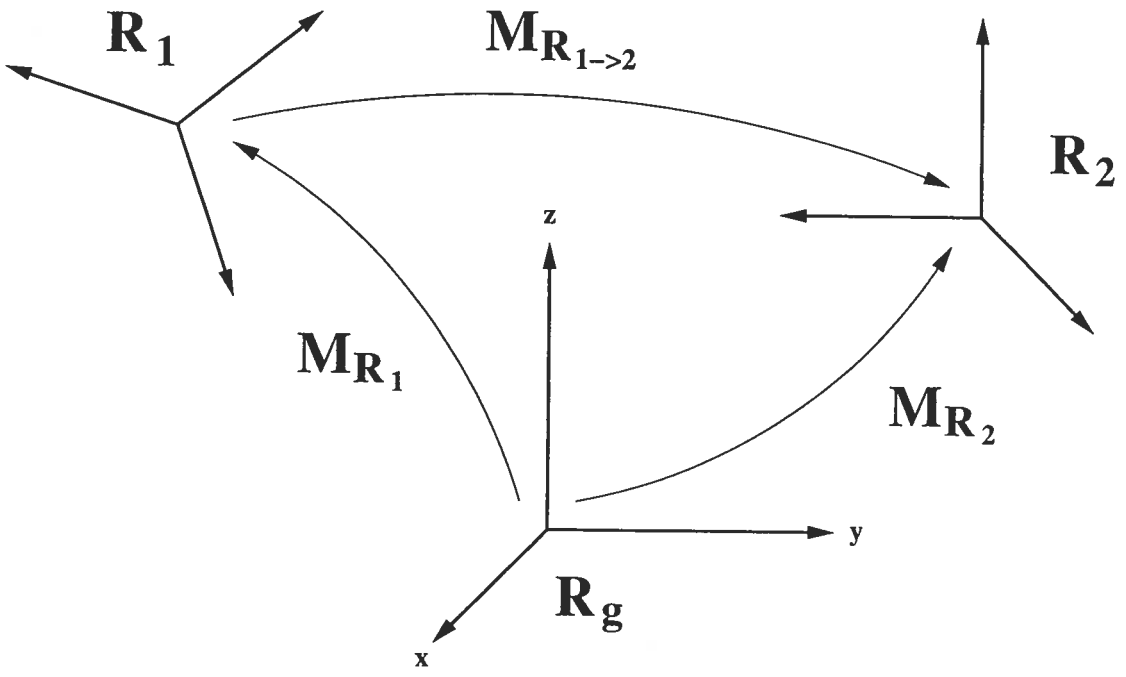
$$M_{R''_2} = \left(M_{R'_1} \times M_{R_{1\to 2}} \times M_{R'_2}{}^{-1}\right) \times M_{R'_2} \tag{I.12}$$

$$= \left(M_{R_1} \times M_{R_{1\to 2}} \times M_{R'_2}{}^{-1}\right) \times M_{R'_2} \tag{I.13}$$

$$= \left(M_{R_2} \times M_{R'_2}{}^{-1}\right) \times M_{R'_2} \tag{I.14}$$

$$= M_{R_2} \tag{I.15}$$

Tout simplement, puisque $M_{R'_1} = M_{R_1}$, $M_{R_1} \times M_{R_{1\to 2}} = M_{R_2}$ (Equation I.9) et $M_{R'_2}{}^{-1} \times M_{R'_2} = I$ (Equation I.8).



(a)

Figure I.1 – Référentiels et matrices de transformations homogènes. La figure montre trois référentiels ; le référentiel global R_G, R_1 et R_2. Trois matrices de transformations homogènes sont aussi indiquées ; M_{R_1} pour passer du référentiel global R_G à R_1, M_{R_2} pour passer du référentiel global R_G à R_2, puis $M_{R_{1\to 2}}$ pour passer du référentiel R_1 à R_2. Les vecteurs ortho-normés définissants R_G sont $\vec{x}$, $\vec{y}$ et $\vec{z}$.