Université de Montréal

Définition d'une mesure de compatibilité séquence–structure dans les protéines à l'aide de modèles probabilistes graphiques et de réseaux de neurones artificiels

par

Daniel St-Arnaud

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

Décembre 2000

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé:

Définition d'une mesure de compatibilité séquence–structure dans les protéines à l'aide de modèles probabilistes graphiques et de réseaux de neurones artificiels

présenté par

Daniel St-Arnaud

a été évalué par un jury composé des personnes suivantes:

Pierre L'Ecuyer

(Président-rapporteur)

François Major

(Directeur de recherche)

Jean-Yves Potvin

(Membre du jury)

Mémoire accepté le 5 avril 2001

SOMMAIRE

Dans ce mémoire, on a défini une nouvelle mesure de compatibilité séquence–structure utilisable pour la reconnaissance des motifs de repliement des protéines (*protein fold recognition*). La mesure proposée a été définie dans un cadre mathématique rigoureux et repose sur la construction d'un nouveau type de modèle stochastique pour les séquences d'acides aminés — possiblement non-homologues — qui partagent un même motif de repliement.

Le modèle stochastique proposé est basé sur la théorie des réseaux de Bayes et est paramétrisé à l'aide de réseaux de neurones artificiels. Contrairement aux mesures de compatibilité séquence–structure traditionnelles (les potentiels pseudo-énergétiques), l'approche proposée permet la modélisation d'interactions d'ordre supérieur entre les acides aminés. À cause de l'entassement très compact des acides aminés à l'intérieur des protéines globulaires, on soupçonne que ce type d'interaction est courant dans les protéines.

Le modèle proposé s'avère plus performant qu'un modèle alternatif représentatif des approches traditionnelles. Cependant, la supériorité du modèle proposé est principalement due à une meilleure représentation de l'environnement structural autour des acides aminés. On constate que la modélisation d'interactions d'ordre supérieur ne permet pas une meilleure reconnaissance d'homologies structurales lointaines lorsqu'on se limite à des modèles structuraux "rigides" qui n'encodent pas les variations structurales observées entre les protéines homologues.

TABLE DES MATIÈRES

LISTE DES TABLES

LISTE DES FIGURES

Chapitre 1

INTRODUCTION

Un organisme vivant évolué, comme l'homme, contiendrait plus de 100,000 composés chimiques différents. Un grand nombre de ces molécules sont des *protéines*. Celles-ci assument une grande variété de rôles à tous les niveaux de la vie.

Avec l'arrivée à terme de la phase de séquençage du projet du génôme humain, [1] les séquences de beaucoup de protéines sont maintenant connues. Cependant, l'obtention de la séquence d'une protéine n'est que le premier pas vers la caractérisation de sa fonction. La fonction d'une protéine est intimement liée à sa structure moléculaire : c'est elle qui détermine la nature de ses interactions avec d'autres biomolécules. Connaître la structure des protéines est donc essentiel à la compréhension des mécanismes moléculaires. C'est aussi un pré-requis à toute manipulation de la fonction des protéines, par exemple pour le traitement de maladies.

À l'heure actuelle, les moyens expérimentaux les plus efficaces pour déterminer la structure des protéines sont la cristallographie aux rayons X ou la spectroscopie par résonance magnétique nucléaire (RMN). Bien qu'elles permettent l'obtention de modèles très détaillés, ces techniques sont laborieuses et difficiles à appliquer à grande échelle.

En réponse à cette situation, des efforts considérables ont été investis vers la mise au point de méthodes computationnelles. À ce jour cependant, la prédiction de la conformation des protéines à partir de leur séquence demeure un des grands défis de la

[1] Le séquençage du génôme humain devrait être complété au cours de l'année 2000. Cependant, les génômes complets de plusieurs organismes (e.g. de nombreux virus et bactéries) sont déjà connus.

biologie moléculaire moderne et les méthodes computationnelles ne génèrent souvent qu'une esquisse grossière du motif de repliement (*fold*) des protéines étudiées.

Les techniques les plus fiables procèdent par analyse comparative de séquence. L'idée est d'inférer la structure d'une nouvelle protéine à partir d'une protéine homologue (une protéine dont la séquence est similaire) et dont la structure a déjà été caractérisée. Évidemment, ceci ne fonctionne que lorsqu'une protéine homologue peut être identifiée (environ 50% des cas [1]).

Dans le cas contraire, on peut se rabattre sur une seconde catégorie de techniques, qui visent à "reconnaître" un motif de repliement compatible à partir d'une banque de motifs connus (*protein fold recognition*). Ces méthodes tentent de tirer parti du fait que même des protéines non-homologues ont souvent des architectures semblables [2].

1.1 Reconnaissance des motifs de repliement des protéines

Le travail décrit dans ce mémoire est lié à la seconde catégorie de techniques évoquées ci-dessus. Ce type d'approche permet d'inférer la conformation d'une séquence d'acides aminés comme suit. Premièrement, on sélectionne un ensemble de motifs de repliement représentatif de la diversité structurale observée chez les protéines. Ensuite, on estime l'énergie requise pour contraindre la séquence dans chaque motif. Plus cette quantité est faible, plus l'affinité de la séquence pour le motif est grande. Finalement, si l'affinité de la séquence pour un motif est suffisamment élevée, on suppose que la séquence adopte une conformation similaire à ce motif.

La mise au point d'une bonne mesure d'affinité séquence–motif est critique au succès de cette démarche. Les mesures les plus connues (e.g. [3, 4, 5, 6]) sont définies à partir d'une analyse statistique des protéines connues. Parce qu'elles ne reposent pas sur de véritables principes physiques, elles sont appelées "potentiels pseudo-énergétiques" (*pseudo-energy potential*, *potential of mean force*, etc.).

Bien que leur formulation exacte soit variable, la plupart de ces "potentiels" peuvent être ramenés à la forme générale

$$\mathcal{E} = \sum_i \mathcal{E}_i + \sum_{i,j} \mathcal{E}_{i,j} \tag{1.1}$$

où on compte un terme par acide aminé (les termes d'ordre 1), et un terme par paire d'acides aminés en contact dans le motif (les termes d'ordre 2). Les termes d'ordre 1 permettent de modéliser la compatibilité des acides aminés pour l'environnement structural auquel ils sont contraints dans le motif. Les termes d'ordre 2 permettent de représenter des interactions entre acides aminés.

Les paramètres de ces potentiels sont typiquement estimés à l'aide de tables des fréquences d'occurence (TFO) : les contributions d'ordre 1 (resp. d'ordre 2) sont définies à partir des fréquences d'occurence des acides aminés (resp. paires d'acides aminés en contact) dans différents environnements.

1.2 Objectifs de ce travail

À part quelques exceptions bien précises [7, 8], aucun des potentiels pseudo-énergétiques décrits dans la littérature ne tient compte des interactions d'ordre supérieur qui surviennent lorsque trois acides aminés ou plus sont simultanément en contact. Bien que certains potentiels assument explicitement que les contacts sont indépendants (e.g. [5]), l'absence de termes d'ordre supérieur s'explique principalement par le fait que les données disponibles ne permettent pas de dériver des paramètres utiles à l'aide de TFO. Cependant, les conflits stériques, les contraintes volumiques et les interactions non-locales qui surgissent suite à l'entassement compact des acides aminés à l'intérieur des protéines globulaires suggèrent que les interactions d'ordre supérieur sont courantes dans les protéines [7, 8].

L'objectif principal de ce travail est de définir une nouvelle mesure de compatibilité séquence–motif qui permet la prise en compte des interactions d'ordre supérieur entre

les acides aminés et qui ne repose pas sur des TFO. Un second objectif est de définir cette mesure à l'intérieur d'un cadre mathématique solide.

1.3 Définition d'un nouveau type de modèle stochastique pour les protéines qui partagent une même architecture

Pour rencontrer les objectifs mentionnés ci-dessus, on se basera sur une formulation probabiliste de la compatibilité séquence–structure proposée par White *et al* [6]. Cette formulation repose sur la construction de modèles stochastiques pour les protéines qui adoptent un même motif de repliement.

En plus de reposer sur des bases mathématiques solides, le formalisme de White *et al* [6] ne requiert aucune hypothèse *a priori* quant à la nature des interactions importantes entre les acides aminés (e.g. interactions binaires seulement, interactions d'ordre supérieur, etc.). Il suffit de modéliser les acides aminés en contact à l'aide de variables aléatoires conjointement dépendantes et les interactions pertinentes pourront éventuellement être "apprises" à partir des données.

Cette dernière observation nous conduira à la définition d'un nouveau type de modèle stochastique pour les protéines — possiblement non-homologues — qui partagent une même architecture. La structure du modèle que nous proposons repose sur la théorie des réseaux de Bayes et il est paramétrisé à l'aide de réseaux de neurones artificiels. Ce modèle constitue la principale contribution originale du présent mémoire et *contrairement à toutes les autres approches disponibles, il permet la représentation des interactions d'ordre supérieur qui surviennent dans le coeur des protéines globulaires.*

1.4 Organisation de ce mémoire

Le reste de ce mémoire est organisé comme suit. Les chapitres 2, 3 et 4 introduisent divers concepts nécéssaires à la compréhension de ce travail et résument la

littérature pertinente. Le chapitre 2 rappelle les bases de la biologie des protéines. On y parle en particulier de leur structure moléculaire, de leur évolution et de la correspondance séquence–structure. Ensuite, le chapitre 3 traite de la prédiction de la structure des protéines à partir de leurs séquences. On y parle d'analyse comparative des séquences et de la reconnaissance des motifs de repliement (*fold recognition*) à l'aide d'alignements séquence–structure (*threading*). Finalement, le chapitre 4 présente le vocabulaire et la notation utilisés, ainsi que les modèles probabilistes graphiques et les réseaux de neurones artificiels. Ces derniers sont à la base du modèle stochastique proposé dans ce mémoire.

Les chapitres 5, 6 et 7 constituent le noyau de ce mémoire. Le chapitre 5 présente en détail la formulation probabiliste de la compatibilité séquence–structure proposée par White *et al* [6]. On y présente aussi une représentation simplifiée pour les motifs de repliement. Cette représentation repose sur la notion de *graphe de contact* et constitue un bon point de départ pour la construction de modèles stochastiques pour les séquences partageant un même motif de repliement. On introduit ensuite le modèle stochastique proposé par White *et al* [6]. Ce modèle utilise la théorie des champs aléatoires de Markov et permet la définition d'un potentiel pseudo-énergétique de la forme de l'équation 1.1.

Le chapitre 6 constitue la principale contribution originale de ce mémoire. On y présente les détails d'un nouveau type de modèle stochastique pour les protéines partageant une même architecture. On définit d'abord la structure de ce modèle à l'aide d'un réseau de Bayes et on montre ensuite comment il peut être paramétrisé à l'aide de réseaux de neurones artificiels. On discute aussi des propriétés de ce modèle, en particulier sa capacité d'"apprendre" des dépendances d'ordre supérieur entre les acides aminés en contact.

Enfin, dans le chapitre 7, on évalue différentes variantes du modèle proposé et on les compare au modèle de White *et al* [6]. On en tire quelques conclusions intéressantes quant à la nature de l'information qui est apprise par les deux modèles.

Chapitre 2

STRUCTURE ET FONCTION DES PROTÉINES

Les êtres vivants sont des entités incroyablements complexes. Un organisme unicellulaire (e.g. une bactérie) contient plusieurs milliers de composés chimiques différents. Un organisme plus évolué, comme l'homme, contiendrait plus de 100,000 biomolécules distinctes dont une fraction seulement a été étudiée.

Un grand nombre de ces molécules sont des *protéines*. Celles-ci assument une grande variété de fonctions à tous les niveaux de la vie. Le but de ce chapitre est d'introduire le lecteur à différents aspects de la biologie des protéines : structure, fonction, évolution et correspondance séquence–structure. Pour plus de détails, le lecteur est invité à consulter un texte général de biologie moléculaire ou un ouvrage traitant de la structure des protéines (voir par exemple [9, 10]).

2.1 Les biopolymères

Les deux types de biomolécules les plus connus sont les acides nucléiques (ADN et ARN) et les protéines.

L'ADN

L'ADN (acide désoxyribonucléique) agit comme support de l'information génétique dans la cellule. Un brin d'ADN est formé par concaténation de quatre sous-unités de base, les nucléotides A, T, G et C et c'est la séquence de ces nucléotides qui encode l'information génétique. L'ADN chromosomal est formé d'une seule molécule d'ADN. Celle-ci est composée de deux brins complémentaires et adopte en général la fameuse structure en double hélice déterminée par Watson et Crick en 1953 [11, 12].

Un *gène* est un fragment d'ADN responsable de l'encodage d'une protéine. La fonction d'un gène est donc intimement liée à celle de la protéine qu'il encode. L'ensemble des gènes définissant une espèce est appelé *génome*. L'ensemble des protéines correspondantes est appelé *protéome*.

À l'heure actuelle, les génomes de plusieurs espèces, en particulier l'humain, ont été séquencés. Les séquences d'ADN sont entreposées dans des bases de données publiques telles que GenBank [13]. Grâce à la mise au point de nouvelles techniques de séquençage et à la pression exercée par le projet du génome humain, GenBank a connu une croissance exponentielle au cours des 20 dernières années. La version courante de GenBank contient plus de 9 milliards de nucléotides, regroupés en plus de 8 millions de séquences.

L'ARN

Bien que construites à partir de quatre sous-unités semblables à celles qui constituent l'ADN, les molécules d'ARN (acide ribonucléique) jouent des rôles beaucoup plus diversifiés. On note, entre autres : transfert de l'information génétique du noyau cellulaire vers le cytoplasme (ARN messager), synthèse de protéines (ARN de transfert, ARN ribosomal), catalyse de réactions chimiques (ribozymes), etc. Afin d'assumer ces fonctions variées, les molécules d'ARN adoptent une grande variété de conformations.

Les protéines

Sur l'échelle de la diversité, tant fonctionnelle que conformationnelle, ce sont les protéines qui occupent le premier rang. Les protéines sont des molécules linéaires, formées par concaténation de 20 acides aminés (voir le tableau 2.1). La *séquence* des acides aminés qui composent une protéine est déterminée par la séquence des nucléotides du gène correspondant.

Comme les séquences génomiques, les séquences des protéines connues sont entreposées dans des bases de données publiques telles que SWISS-PROT [14] ou PIR [15]. Celles-ci ont également connu une croissance exponentielle au cours des dernières années, mais à un rythme beaucoup plus lent. Ceci s'explique par la difficulté d'identifier et d'annoter correctement les régions codantes des séquences d'ADN. SWISS-PROT contient les séquences d'environ 80000 protéines.

Acide aminé	id	classe	Acide aminé	id	classe
Alanine	A/ALA	hydrophobe	Leucine	L/LEU	hydrophobe
Arginine	R/ARG	chargé	Lysine	K/LYS	chargé
Asparagine	N/ASN	polaire	Methionine	M/MET	hydrophobe
Acide aspartique	D/ASP	chargé	Phenylalanine	F/PHE	hydrophobe
Cysteine	C/CYS	polaire	Proline	P/PRO	hydrophobe
Glutamine	Q/GLN	polaire	Serine	S/SER	polaire
Acide glutamique	E/GLU	chargé	Threonine	T/THR	polaire
Glycine	G/GLY	—	Tryptophan	W/TRP	polaire
Histidine	H/HIS	polaire	Tyrosine	Y/TYR	polaire
Isoleucine	I/ILE	hydrophobe	Valine	V/VAL	hydrophobe

TAB. 2.1 – Les 20 acides aminés. La colonne "id" indique les identificateurs habituellement utilisés. Les acides aminés sont regroupés en 3 classes en fonction des propriétés de leurs chaînes latérales : hydrophobes, chargées ou polaires. L'acide aminé Glycine, dont la chaîne latérale n'est formée que d'un atome d'hydrogène, est classifié à part.

En solution, les protéines sont repliées de manière complexe. C'est la structure repliée d'une protéine qui détermine sa fonction et la nature de ses interactions avec d'autres biomolécules. Les fonctions assumées par les protéines sont très variées : catalyse de la plupart des réactions chimiques cellulaires (e.g. réplication de l'ADN), transport (e.g. transport de l'oxygène par le sang), échanges trans-membranaires,

signaux intra-cellulaires, encapsidation (virus), etc.

Les protéines dont les structures repliées ont été déterminées sont représentées dans la Protein Data Bank (PDB) [16]. À cause de la complexité et des coûts associés à la détermination de la structure d'une protéine par des moyens expérimentaux (crystallographie aux rayons X et spectroscopie par résonance magnétique nucléaire), seule une faible proportion des protéines connues sont représentées dans la PDB. Au 14 septembre 2000, la PDB contenait 13154 structures.

2.2 Structure des Protéines

C'est la structure moléculaire d'une protéine qui détermine la nature de ses interactions avec d'autres biomolécules et la spécificité de sa fonction. La caractérisation de la fonction d'une protéine passe donc par l'étude de sa structure moléculaire. La structure des protéines est habituellement étudiée selon quatre niveaux d'organisation respectivement appelés structure primaire, secondaire, tertiaire et quaternaire.

La structure primaire : la séquence

Le premier niveau d'organisation, ou structure primaire, est celui de la *séquence* des acides aminés. Les acides aminés composant une protéine sont appelés *résidus* (voir fig. 2.1). C'est la séquence, par le biais des propriétés particulières de chaque acide aminé, qui détermine le repliement de la chaîne en solution — la *conformation* de la protéine — et la spécificité fonctionnelle de celle-ci. La conformation (en général unique) qu'adopte une chaîne est celle qui minimise son énergie libre, par exemple par la formation de liaisons hydrogène entre résidus, ou par l'orientation vers l'intérieur des résidus hydrophobes. La figure 2.2 présente la séquence de la protéine ribosomale L1.

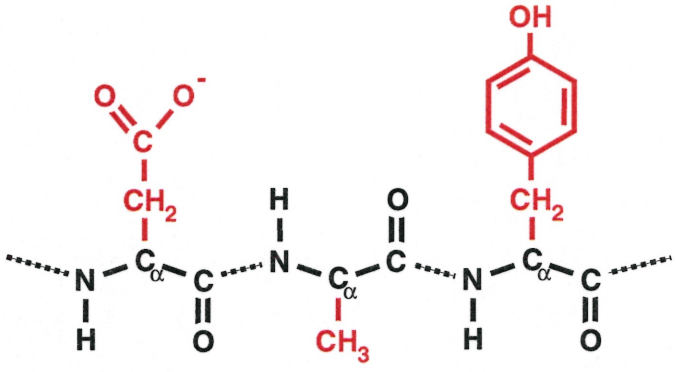

FIG. 2.1 – Fragment d'une séquence d'acides aminés. Les résidus montrés sont ASP, ALA, et TYR. La chaîne principale (ou squelette) est en noir. Les chaînes latérales, différentes pour chaque type d'acide aminé, sont en rouge. Les liaisons peptidiques, qui lient les résidus le long de la chaîne, sont en pointillé.

Structure secondaire

Le second niveau d'organisation est celui de la structure secondaire des protéines. Un *élément de structure secondaire* (ESS) est un sous-ensemble de résidus consécutifs en séquence et agencés selon un patron régulier. On distingue deux types d'ESS : les hélices α et les brins β (voir fig. 2.2 et fig. 2.3). Les brins β sont regroupés à l'intérieur de motifs plus complexes, appelés feuillets β. Les hélices α et les feuillets β sont stabilisés par la formation de liaisons hydrogène entre les groupements NH et CO des résidus impliqués. Les feuillets β sont des structures dites super-secondaires car elles sont stabilisées par des *interactions non-locales* (voir ci-dessous).

Dans les protéines, tous les résidus n'appartiennent pas à des ESS. Les régions qui connectent les ESS les uns aux autres sont appelées *boucles*. Une boucle est un sous-ensemble de résidus consécutifs en séquence mais ne formant pas un ESS (une hélice α ou un brin β).

FIG. 2.2 – Séquence (structure primaire) et structure secondaire de la protéine ribosomale L1. Les hélices α sont indiquées en rouge et les brins β en jaune.

Structure tertiaire

Le troisième niveau d'organisation, la structure tertiaire (ou 3-D), est formée par l'entassement des ESS en une ou plusieurs unités compactes et globulaires, appelées *domaines*. Ceci se produit lorsque la chaîne d'acides aminés se replie sur elle-même en solution.

Le repliement de la chaîne permet des *interactions non-locales* entre acides aminés éloignés en séquence mais spatialement adjacents dans la structure repliée. En plus d'être généralement impliquées dans l'activité d'une protéine, ces interactions sont responsables de la stabilité de sa structure.

La force principale qui guide le repliement d'une protéine en solution aqueuse est l'orientation des chaînes latérales hydrophobes vers l'intérieur de la molécule. Ainsi, la structure repliée se caractérise par un noyau hydrophobe responsable de sa stabilité et par une surface hydrophile assurant sa solubilité. La conformation active d'une protéine est minimale d'un point de vue énergétique et est déterminée de manière unique par sa séquence. La figure 2.3 présente la structure de la protéines ribosomale L1.

On dit des ESS qui composent un domaine qu'ils constituent le *noyau structural* de ce domaine. On dit aussi que deux domaines partagent une même *architecture* si la nature, l'orientation et la position relative des ESS sont préservées (i.e. si leurs noyaux structuraux sont similaires).

FIG. 2.3 – Structure 3-D de la protéine ribosomale L1. Le domaine 2 est en bleu. Pour le domaine 1, les hélices α sont en rouge et les feuillets β sont en jaune.

Structure quaternaire

On note finalement un dernier niveau d'organisation. Celui-ci est présent lorsque la protéine active contient plusieurs domaines agglomérés les uns contre les autres en une structure dite quaternaire.

2.3 Motifs structuraux et motifs de repliement

Un motif structural est une construction régulière présente dans plusieurs protéines. Les hélices α et les feuillets β sont des exemples simples de motifs structuraux mais on s'intéresse souvent à des motifs plus complexes formés de plusieurs éléments de structure secondaire. Ceux-ci sont habituellement associés à des fonctions précises et sont partagés par des protéines qui ont un ancêtre évolutif commun.

Le motif structural défini par l'organisation dans l'espace des ESS qui composent un domaine est appelé *motif de repliement* (*folding motif*) de ce domaine. Les β–

tonneaux constituent un exemple d'un tel motif. Il en existe plusieurs types, associés à des fonctions diverses telles le transport du rétinol (vitamine A). Dans la plupart des cas, le centre du tonneau forme une cavité capable d'accueillir une autre molécule (un *ligand*) [10].

2.4 Similarité structurale et évolution

Il est généralement admis que les chaînes d'acides aminés sont contraintes à un nombre relativement faible d'architectures [2, 17, 18]. Beaucoup de protéines présentent en effet des similarités structurales avec d'autres protéines.

En général, **une forte similarité en séquence implique une similarité structurale significative [10, 19].** Des protéines dont les séquences sont similaires sont appelées *homologues* et possèdent en général un ancêtre évolutif commun. Des homologues sont caractérisés par une fonction commune[1] et la similarité structurale est conséquente à une pression évolutive vers la préservation de cette fonction.

Le principe ci-dessus a donné naissance à *l'analyse comparative des séquences biologiques*, une technique permettant d'inférer la conformation d'une nouvelle séquence d'acides aminés à partir de ses homologues déjà caractérisés. Ceci sera discuté en détail dans la section 3.1.

Par ailleurs, **des chaînes d'acides aminés n'affichant pas une similarité de séquence significative peuvent aussi partager une même architecture [10, 19].** Si deux telles chaînes sont liées dans l'évolution, on parle d'*homologues lointains* et celles-ci ont en général des fonctions semblables. Dans le cas contraire, la similarité structurale est le résultat d'une convergence évolutive[2] et on parle de

[1] Il existe cependant certaines exceptions, où des gènes sont recrutés à un moment de l'évolution pour produire des protéines aux fonctions très différentes. On note par exemple le cas d'une protéine du cristallin affichant une forte homologie avec un enzyme impliqué dans le métabolisme du lactose (voir [10], p249).

[2] La nature tend en effet à préserver et à réutiliser des constructions efficaces (e.g. l'architecture

protéines *analogues*. Des protéines analogues ont rarement des fonctions semblables.

Face à une séquence n'ayant pas d'homologues connus, l'énoncé précédent suggère la prédiction de la conformation d'une séquence S en détectant un motif structural pour lequel S montre une grande affinité. Ceci peut être fait à l'aide d'alignements séquence–structure (voir section 3.2).

Les protéines peuvent être regroupées de manière hiérarchique en fonction du degré de similarité structurale qu'elles partagent. Des protéines homologues constituent une *famille*. Des familles liées dans l'évolution forment une *super-famille* (un ensemble d'homologues lointains). Des super-familles partageant une même architecture définissent un *fold* (un ensemble d'homologues et d'analogues). Les folds sont regroupés en *classes architecturales* selon leur contenu en élements α et β (α, β, $\alpha+\beta$, α/β, etc.).

Quelques classifications structurales des protéines de la PDB sont disponibles. Les plus connues sont SCOP (Structural Classification Of Proteins [20]) et CATH (Class, Architecture, Topology, Homology [21, 22]). La version 1.50 de SCOP contient 24186 domaines définissant 7 classes structurales, 548 folds, 820 super-familles et 1296 familles.

---

d'une protéine) à travers l'évolution.

Chapitre 3

APPROCHES THÉORIQUES POUR LA DÉTERMINATION DE LA STRUCTURE DES PROTÉINES

Déterminer la structure moléculaire d'une protéine par des moyens expérimentaux (e.g. cristallographie aux rayons X ou spectroscopie par résonance magnétique nucléaire (RMN)) est un processus coûteux et laborieux et des efforts considérables ont été investis vers la mise au point de méthodes théoriques. À ce jour cependant, la prédiction de la conformation des protéines à partir de leur séquence demeure un des grands défis de la biologie moléculaire moderne.

Le but de ce chapitre est d'introduire le lecteur aux différentes méthodes disponibles. Celles-ci peuvent être regroupées en trois grandes catégories : l'analyse comparative de séquences, la reconnaissance d'homologie structurale (*fold recognition*), et la prédiction *ab initio*.

Les techniques d'analyse comparative de séquences permettent d'inférer la structure d'une nouvelle protéine à partir de la structure d'une protéine homologue — une protéine dont la séquence est similaire — déjà caractérisée. Comme il s'agit des techniques les plus fiables et les plus abondamment utilisées, celles-ci sont introduites dans la section 3.1 ci-dessous.

Cependant, l'analyse comparative de séquences ne fonctionne que lorsqu'une protéine homologue peut être identifiée. Si ce n'est pas le cas, la recherche d'homologies structurales constitue une alternative intéressante. Les techniques de cette catégorie visent à "reconnaître" l'architecture d'une nouvelle protéine à partir d'une banque de motifs structuraux connus. Les alignements séquence–structure sont un exemple

populaire d'une telle technique. Ceux-ci s'apparentent à l'approche proposée dans ce mémoire et sont présentés dans la section 3.2.

Malgré quelques percées intéressantes et des solutions acceptables pour certains sous–problèmes (e.g. la prédiction de la structure secondaire), les méthodes *ab initio*, qui reposent uniquement sur l'utilisation de principes physico-chimiques et/ou de règles empiriques, n'ont jusqu'à maintenant obtenu que des succès limités [23] et ne seront pas abordées ici.

3.1 Analyse comparative des séquences biologiques

L'analyse comparative de séquence permet d'obtenir rapidement des indices sur la structure et la fonction d'une nouvelle protéine si celle-ci est homologue à des protéines déjà étudiées. Dans le cas de séquences très similaires, une relation d'homologie peut être trivialement détectée à l'aide d'alignements binaires (section 3.1.1). Dans le cas d'homologues lointains, on doit recourir à des techniques plus évoluées basées sur l'analyse simultanée de plusieurs séquences (section 3.1.3).

L'introduction qui suit s'inspire de [19, 24]. Le lecteur intéressé trouvera dans ces ouvrages un traitement détaillé de la théorie des chaînes de caractères, le support formel de l'analyse comparative des séquences biologiques.

3.1.1 Alignement de deux séquences

La *similarité* entre deux séquences d'acides aminés S_1 et S_2 est définie en fonction de l'alignement optimal de S_1 et S_2.

Soient S_1 et S_2 des chaînes d'acides aminés de longueurs $|S_1|$ et $|S_2|$. [1] On définit les chaînes S_1' et S_2' satisfaisant $|S_1'| = |S_2'| = l$ et qui désignent les chaînes S_1 et S_2 après ajout d'*espaces*. L'introduction d'espaces permet la représentation d'insertions

[1] Bien qu'on s'intéresse ici aux chaînes d'acides aminés (protéines), les concepts discutés dans cette section s'appliquent également à l'analyse des séquences de nucléotides (ADN, ARN).

et de délétions (appelés *gaps*) dans les alignements. Une insertion (resp. deletion) correspondent à l'apparition (resp. disparition) d'un fragment de gène dans l'évolution.

La matrice $A_{2\times l}$ de 2 rangées et l colonnes satisfaisant $A_{i,j} = S'_i(j)$ est un *alignement (binaire)* de S_1 et S_2 (voir fig. 3.1). Le *score* de cet alignement est défini par

$$\Phi(A) \equiv \sum_{i=1}^{l} s(S'_1(i), S'_2(i)), \tag{3.1}$$

où $s(x, y)$ est le score de l'alignement des acides aminés x et y. [2]

```
>>d1mba__ 1.1.1.1.5 Myoglobin (Sea hare (Aplysia limacin  (147 aa)
 initn:  96 init1:  96 opt: 136  Z-score: 181.6  bits: 39.6 E(): 7.5e-05
Smith-Waterman score: 136;  31.579% identity in 95 aa overlap (2-93:3-96)

d1mbn_  VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE ...
         :: .: .:. . :: : :.  ..: :.:. ::.. :.. . :  ::  :. :..:::
d1mba_ XSLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKG-KSVADIKASP ...
```

FIG. 3.1 – Alignement optimal des séquences de la myoglobine du cachalot et de la myoglobine de la limace de mer. L'alignement a été effectué avec fasta33 [25, 26] en utilisant la paramétrisation par défaut. Le caractère "-" correspond aux gaps. 31.579% des acides aminés sont conservés.

Le choix de la fonction de score s est critique. Celle-ci est en général définie sous la forme d'une matrice de substitution S. Le score $s(x, y)$ (donné par l'entrée $S_{x,y}$ de S) reflète la probabilité de substitution de l'acide aminé x par l'acide aminé y dans des séquences liées évolutivement. Dans le cas des matrices de la célèbre série PAM [27], ces probabilités sont estimées à partir d'alignements manuels. Les matrices de la série BLOSUM [28], conçues récemment pour la détection d'homologies lointaines,

[2] En pratique, on utilise souvent la variante suivante lorsque des gaps sont présents dans un alignement : le score d'une séquence de n espaces consécutifs (un gap) est donné par $f(n) \equiv a + b(n-1)$, avec $a > b$ et où a est appelé *pénalité d'ouverture* et b *pénalité d'extension* (voir par exemple [24]).

sont construites à partir de patrons de séquence hautement conservés et associés à une fonction biologique reconnue.

3.1.2 Recherche de l'alignement optimal

L'*alignement optimal* $A^\star$ de deux chaînes d'acides aminés S_1 et S_2 est l'alignement qui maximise Φ et le score $\Phi(A^\star)$ de cet alignement est la *similarité* de S_1 et S_2. Le problème d'optimisation correspondant est appelé *problème d'alignement global.* La première solution à ce problème fut proposé par Needleman and Wunsch [29] en 1970.

En pratique, les séquences de protéines issues de familles différentes sont *similaires par regions* et on s'intéresse plutôt au *problème d'alignement local* (PAL). Dans ce cas, on cherche des *sous-chaînes* de S_1 et S_2 dont le score d'alignement est maximal. Celles-ci correspondent en général à des motifs structuraux communs. Les alignements locaux permettent aussi d'identifier des caractéristiques locales fortement conservées (e.g. le site actif d'une famille d'enzymes) dans des séquences autrement peu apparentées.

Le PAL fut introduit dans le cadre de la biologie moléculaire par Smith et Waterman en 1981 [30]. Leur solution au problème (l'algorithme SW) utilise la programmation dynamique et s'éxécute en temps quadratique [24, 19]. [3]

Bien que l'algorithme SW demeure la solution exacte au PAL la plus utilisée de nos jours, l'augmentation rapide de la taille des banques de séquences a conduit au développement de solutions heuristiques plus efficaces. Les méthodes heuristiques les plus connues sont BLAST (pour "basic local alignment search tool" [31]), et FASTA (pour "fast-all" [25, 26]).

En pratique, ces algorithmes doivent être paramétrisés en fonction du problème étudié (e.g. acides nucléiques *vs* protéines). On pense par exemple au choix des fonc-

[3] Si S_1 et S_2 sont des chaînes d'acides aminés de longueur n et m respectivement, la solution au PAL est obtenue en temps $O(mn)$ a l'aide de l'algorithme SW [24].

tions de score ou au pré-traitement des banques de séquences afin d'en réduire la redondance. Il est également primordial d'interpréter les résultats d'une recherche à l'intérieur d'un cadre statistique fiable.

Le lecteur intéressé peut se référer à [32] pour une discussion de la paramétrisation optimale des différentes procédures d'alignements. Les aspects statistiques sont traités d'une manière accessible dans [33]. Pour une analyse comparative des méthodes disponibles, on peut se référer à [34] ou [35].

3.1.3 Analyse de multiples séquences

A première vue, l'analyse simultanée de plusieurs chaînes d'acides aminés peut paraître une simple généralisation de la comparaison de deux séquences introduite dans la section précédente. Il s'agit en fait d'un outil beaucoup plus puissant qui permet l'extraction et la représentation des propriétés clef d'une famille de séquences. L'objectif est de découvrir les caractéristiques communes subtiles qui déterminent la stabilité et l'efficacité de l'architecture partagée par cette famille.

```
d1qpwb_  ... SNADAVMGNPKVKAHGKKVLQSFSDGLKHLD--NLKGTFAKLSELHCDQLHVDPENFRLL ...
d1baba_  ... S-----HGSAQVKGHGKKVADALTNAVAHVD--DMPNALSALSDLHAHKLRVDPVNFKLL ...
d1mbn__  ... KTEAEMKASEDLKKHGVTVLTALGAILKKKG--HHEAELKPLAQSHATKHKIPIKYLEFI ...
d2lhb__  ... TTADELKKSADVRWHAERIINAVDDAVASMD--DTEKMSMKLRN--------------- ...
d1hlb__  ... S-ASQLRSSRQMQAHAIRVSSIMSEYVEELDS-DILPELLATLARTHDLNKVGADHYNLF ...
d3sdha_  ... S---QGMANDKLRGHSITLMYALQNFIDQLDNPDDLVCVVEKFAVNHITRKISAAEFGKI ...
```

FIG. 3.2 – Alignement de quelques séquences de la famille des globines. L'alignement a été effectué avec ClustalW [36].

L'approche classique à l'analyse de multiples séquences est basée sur la construction d'un *alignement multiple* (fig. 3.2).

Soit $S = \{S_1, S_2, \cdots, S_k\}$, un ensemble de séquences d'acides aminés de longueurs $|S_1|, |S_2|, \cdots, |S_k|$ respectivement. On construit $S' = \{S_1{}', S_2{}', \cdots, S_k{}'\}$ satisfaisant $|S_1{}'| = |S_2{}'| = \cdots = |S_k{}'| = l$, où $S_i{}'$ désigne S_i après ajout d'espaces. La matrice

$A_{k\times l}$ de k rangées et l colonnes satisfaisant $M_{i,j} = S_i^{'}(j)$ est l'*alignement multiple* correspondant à $S^{'}$. Une sous-matrice de deux rangées de A est un *alignement binaire induit par M.*

D'une manière analogue au cas de l'alignement de deux séquences, on s'intéresse à l'alignement multiple *optimal* selon une certaine fonction objectif. Un exemple simple est la fonction *sum-of-pairs* (SP [37]), où l'alignement multiple optimal est celui qui maximise la somme des scores de tous les alignements binaires induits. Des alternatives utilisent les notions de *séquence consensus* [19] ou d'*alignement phylogénétique* [38, 39]. On note que contrairement au cas de l'alignement de deux séquences, il n'y a pas consensus quant au choix de la "meilleure" fonction objectif. Pour un traitement détaillé, le lecteur est référé à [19, 24].

3.1.4 Modélisation probabiliste des familles de séquences

Un alignement multiple permet de définir un *modèle probabiliste* (ou *stochastique*) pour une famille de séquences. Un exemple simple est un *profil* : Soit $S = \{S_1, S_2, \cdots, S_k\}$ un ensemble de séquences et soit $A_{k\times l}$ un alignement multiple de $S_1, S_2, \cdots, S_k$. La matrice $M_{21\times l}$, où M_{ij} est la probabilité d'occurence de l'acide aminé i (ou d'un espace) en position j pour $s \in S$, est le *profil* correspondant à A.

Un profil peut être vu comme un *générateur* de séquences. Soit $T = t_1t_2t_3\cdots t_n$ une séquence et soit $T^{'}$ représentant T après ajout d'espaces. $T^{'}$ induit un alignement a de T contre le profil. La probabilité $P(T \mid M, a)$ que T soit *générée* par M sachant que l'alignement correct est a est donnée par

$$P(T \mid M, a) = \prod_{i=1}^{l} M(T^{'}(i), i) \tag{3.2}$$

La probabilité $P(T \mid M)$ que T soit générée par M implique une somme sur tous les alignements possibles et est donnée par

$$P(T \mid M) \;=\; \sum_{a} \frac{P(T, M, a)}{P(M)}$$

$$= \sum_{a} P(T \mid M, a) P(a \mid M) \tag{3.3}$$

Cette quantité est la *vraisemblance* de T dans le profil et peut être utilisée comme critère d'appartenance à la famille représentée par M.

Un profil peut être généralisé à l'aide d'un modèle de Markov caché (HMM) [40]. Il s'agit d'un modèle stochastique plus puissant, couramment utilisé pour la détection d'homologies lointaines [41].

3.1.5 Discussion

Une des premières applications concluantes de l'analyse comparative des séquences biologiques fut rapportée en 1983 par Doolittle *et al* [42]. Ces chercheurs suspectaient le rétro–virus *simian sarcoma* d'induire une forme de cancer en déréglant le processus normal de croissance cellulaire. En identifiant une forte homologie de séquence entre un gène viral et un gène encodant un important facteur de croissance chez les animaux, ils ont pu démontrer que c'est l'expression incontrolée du gène viral *v-sis* par la cellule hôte qui cause la maladie en stimulant exagérément la croissance cellulaire.

De nos jours, les découvertes basées sur l'analyse comparative des séquences biologiques sont presque devenues routinières. Les outils d'analyse (e.g. la suite de programmes BLAST) sont offerts avec des interfaces Web accessibles et les séquences des protéines connues peuvent être obtenues facilement à partir de bases de données publiques telles SWISS-PROT [14] ou PIR [15], ou par traduction des régions codantes des séquences de nucléotides (ADN et ARN) disponibles dans GenBank [13]. Dans environ 50% des cas, une comparaison mécanique avec ces banques de séquences permet la détection de similarités suffisantes pour suggérer la fonction enzymatique ou structurale d'une protéine inconnue [1].

Ces performances remarquables sont en grande partie dues au développement de techniques faisant usage de multiples séquences telles les HMMs ou PSI–BLAST. Celles-ci permettent en effet la détection de trois fois plus d'homologues lointains que

les méthodes traditionnelles restreintes aux alignements binaires [43].

3.2 *Reconnaissance de l'architecture des protéines à l'aide d'alignements séquence–structure*

L'alignement séquence–structure, ou *threading*, est une importante approche à la reconnaissance de l'architecture des protéines. Contrairement à l'analyse comparative de séquences, ce type d'approche permet d'obtenir des indices sur l'architecture d'une nouvelle protéine même si celle-ci ne possède aucun homologue dont la structure a déjà été caractérisée.

On procède en général comme suit : Premièrement, on définit un ensemble de *patrons structuraux* (*folding templates*) représentatif de la diversité structurale observée chez les protéines. Ensuite, on estime l'énergie requise pour contraindre la séquence dans chaque patron. Plus cette quantité est faible, plus l'affinité de la séquence pour le patron est grande. Finalement, si l'affinité de la séquence pour un patron est suffisamment élevée, on suppose que la séquence adopte une conformation similaire à celle du patron.

Il existe une abondante littérature traitant d'alignement séquence–structure. L'introduction qui suit s'inspire de [44, 45, 46, 6, 5].

3.2.1 *Définition des patrons structuraux*

La banque de patrons se doit d'être construite de facon à bien représenter les différents motifs de repliement observés chez les protéines. Chaque patron correspond typiquement au noyau structural conservé dans une famille de protéines apparentées et encode des caractéristiques physico–chimiques jugées importantes (e.g. l'exposition locale au solvant, le type de structure secondaire, les contacts entre acides aminés, etc.).

Pour les besoins de la discussion qui suit, un patron est simplement défini par

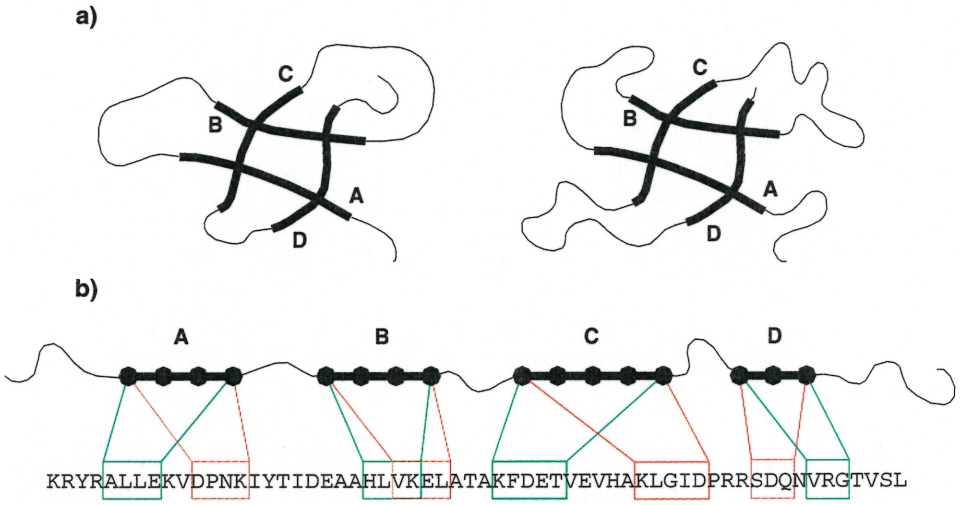

FIG. 3.3 – Alignement séquence–structure. a) Deux protéines partageant un noyau structural similaire (en gras). b) Deux alignements possibles d'une séquence dans le patron défini par le noyau structural en a).

une séquence de *segments structuraux (SS)*. Chaque SS contient un nombre fixe de *positions spatiales* représentant des acides aminés et est associé à un des éléments de structure secondaire qui composent le noyau structural représenté par le patron (voir fig. 3.3a). Les SS sont connectés par des régions de longueur variable appelées *boucles*.

3.2.2 Alignements

Il existe plusieurs façons de contraindre une séquence dans un patron. Une configuration valable (un *alignement*) est construite en assignant progressivement les acides aminés de la séquence aux positions spatiales du patron, de façon à ce que des positions adjacentes d'un même SS soient occupées par des acides aminés adjacents dans la séquence. Les acides aminés non-assignés à une position sont assignés aux boucles (voir fig. 3.3b). Ceci est équivalent à confiner les insertions et délétions (les *gaps*, section 3.1.1) aux régions ne faisant pas partie du noyau (les boucles). On remarque

qu'en permettant des boucles de longueur variable, le nombre d'alignements possibles croît exponentiellement [44].

La compatibilité séquence–patron est définie en fonction de l'énergie de l'alignement optimal de la séquence dans le patron. L'alignement optimal est celui qui minimise une mesure de l'énergie requise pour contraindre la séquence dans cette configuration.

3.2.3 Mesures d'affinité séquence–patron

La littérature regorge d'exemples de mesures développées pour approximer l'énergie d'un alignement séquence–structure (e.g. [3, 4, 5, 6], voir [45] pour une revue). Ces mesures portent le nom de *potentiels pseudo-énergétiques* (*pseudo-energy potential*, *potential of mean force*, etc.).

Un potentiel pseudo-énergétique est habituellement de forme

$$\mathcal{E}(a) = \mathcal{E}_C(a_C) + \mathcal{E}_L(a_L). \tag{3.4}$$

Ici, $a = a_1 a_2 \cdots a_n$ désigne la séquences d'acides aminés alignée, a_C et a_L désignent respectivement les acides aminés assignés au noyau et aux boucles, et $\mathcal{E}_C$ et $\mathcal{E}_L$ correspondent respectivement aux contributions du noyau et des boucles.

Pour les potentiels les plus couramment utilisés, $\mathcal{E}_C$ peut être ramené à la forme générale suivante

$$\begin{aligned} \mathcal{E}_C(a_C) &= g^{(1)}(a_C) + g^{(2)}(a_C) + g^{(3)}(a_C) + \cdots \\ &= \sum_i g_i(a_i) + \sum_i \sum_{j>i} g_{i,j}(a_i, a_j) \\ &\quad + \sum_i \sum_{j>i} \sum_{k>j} g_{i,j,k}(a_i, a_j, a_k) + \cdots \end{aligned} \tag{3.5}$$

Les termes d'ordre 1 permettent de modéliser la compatibilité des acides aminés pour l'environnement structural à chaque position. Les termes d'ordre supérieur permettent de modéliser les interactions entre les acides aminés assignés à des positions voisines

dans l'espace (i.e. on compte un terme $g_{i,j}$ non-nul (resp. $g_{i,j,k}$, $g_{i,j,k,l}$,...) pour chaque paire (resp. triplet, quadruplet, ...) d'acides aminés en contact).

Un potentiel qui n'inclut que des termes d'ordre 1 est parfois appelé *3D-1D potential* [4, 47]. Dans ce cas, les patrons structuraux sont appelés *profils 3D* (*3D-profiles*) par analogie aux profils de séquences (section 3.1.3). Lorsque des termes d'ordre 2 ou plus sont considérés, on parle d'un *contact potential.*

Les paramètres de $\mathcal{E}$ peuvent être fixés à partir de connaissances *a priori* (*knowledge-based potentials*) ou estimés par analyse statistique de la banque de patrons (*empirical potentials*) Dans le second cas, les termes d'ordre 1 (resp. 2, 3, ...) sont typiquement définis à partir des fréquences d'occurence des acides aminés (resp. paires d'acides aminés, triplets d'acides aminés, ...) dans différents contextes structuraux. Cependant, à cause du faible nombre d'exemples disponibles, cette approche est impraticable pour des termes d'ordre 3 ou plus et la somme dans l'équation 3.5 est habituellement tronquée après les termes d'ordre 2 [3, 5, 6] [4].

Un exemple, le potentiel de [6], sera présenté en détail dans le chapitre 5.

3.2.4 Recherche de l'alignement optimal

Lorsque seuls des termes d'ordre 1 sont considérés, l'alignement optimal peut être obtenu en temps polynômial par programmation dynamique [4, 47].

En pratique, la prise en compte des interactions inter-positions (termes d'ordre 2) améliore significativement l'efficacité d'un potentiel pseudo-énergétique [45] pour la reconnaissance d'homologies lointaines. Dans ce cas cependant, si on permet des patrons avec des boucles de longueur variable, la recherche de l'alignement de moindre énergie est un problème NP-difficile [48]. Un algorithme de type *branch and bound* garantissant l'optimalité et performant bien en pratique est introduit dans [44]. Des

[4] On note cependant la fonction de score de [7] qui compte un terme d'ordre 3. L'approche de [8] permet de modéliser les interactions entre un nombre arbitraire de voisins (termes d'ordre 2 et plus) mais n'a pas été appliquée à la définition d'une mesure globale de compatibilité séquence–structure.

approches heuristiques sont présentées dans [7, 49].

3.2.5 Discussion

Dans la littérature, les différentes fonctions de score sont souvent évaluées à l'aide de l'épreuve *d'auto–reconnaissance*, où un ensemble de séquences sont alignées sur une banque de patrons contenant leur propre structure. Il s'avère que tout pseudo–potentiel *raisonnable*, quel que soit sa complexité, réussit à détecter le "bon" patron lors d'une telle épreuve. Thomas et Dill [50] ont expliqué ceci en montrant que tout pseudo–potentiel encode principalement l'effet hydrophobique — la préférence des chaînes latérales hydrophobes pour l'intérieur de la molécule — plutôt que de véritable interactions entre acides aminés.

Comme il y a typiquement plusieurs façons de positionner une séquence à l'intérieur d'un patron de manière à ce que les résidus hydrophobes soient protégés du solvant, on observe souvent de mauvais alignements même dans le cas idéal de l'auto–reconnaissance.

Pour la reconnaissance d'homologues/analogues — l'alignement d'une séquence sur une banque de patrons contenant un homologue/analogue — on observe que la qualité des alignements diminue rapidement lorsque le "bon" patron diffère significativement de la véritable structure [51, 52, 53]. Les propriétés structurales détaillées — en particulier l'accessibilité locale au solvant — sont en effet rarement conservées entre des protéines similaires [54].

La discussion ci-dessus suggère que l'affinement des pseudo-potentiels existants par ajout de paramètres énergétiques plus complexes — particulièrement si ceux-ci sont optimisés pour le problème d'auto-reconnaissance — n'est pas la voie à adopter pour améliorer la reconnaissance d'homologue/analogue [45].

Comme pour l'analyse de séquence, où les patrons de séquence (profils, séquences consensus, HMMs ; section 3.1.3), la mise au point de patrons structuraux *consensus* qui encodent la nature et le degré de variation structurale observés à l'intérieur de

familles de protéines constitue une solution beaucoup plus prometteuse [45, 55, 56].

Jones et Thornton [45] suggèrent que la recherche d'un pseudo-potentiel capable de distinguer l'architecture et l'alignement corrects parmi les milliers d'alternatives possibles relève de l'utopie. En fait, l'existence de *processus de repliement* (*folding pathways* [57, 58]) dans l'espace des conformations est un indice que même la "vraie" fonction d'énergie est incapable de réussir cet exploit. Un objectif plus réaliste serait d'arriver à réduire la banque de patrons à une courte liste de conformations possibles.

Les résultats du dernier concours CASP (*Critical Assessment of methods of protein Structure Prediction* [23]) semblent appuyer l'observation ci-dessus. Dans ce concours, les différentes techniques de prédiction sont confrontées à une liste de protéines (les "cibles") dont les structures sont sur le point d'être rendues publiques. Les prédictions sont ensuite comparées aux structures déterminées expérimentalement. Dans CASP3, les techniques de threading on été les plus performantes sur les cibles affichant peu ou pas d'homologie de séquence avec des protéines connues. Bien que les architectures d'une majorité des 23 cibles de cette catégorie aient été identifiées par l'une ou l'autre des équipes participantes, les prédictions soumises ne capturent au mieux que l'architecture grossière des véritables structures [59].

Chapitre 4

VOCABULAIRE, NOTATION, MODÈLES PROBABILISTES GRAPHIQUES ET RÉSEAUX DE NEURONES ARTIFICIELS

Dans les chapitres 5 et 6, on présentera deux types de modèles stochastiques pour la représentation des protéines qui partagent un même motif de repliement. Le premier modèle, dû à White *et al* [6] (chapitre 5), repose sur la notion de champ aléatoire de Markov. Le second modèle, qui constitue le principale contribution du présent travail (chapitre 6), est basé sur la théorie des réseaux de Bayes et est paramétrisé à l'aide de réseaux de neurones artificiels.

Le but premier de ce chapitre est de présenter les modèles probabilistes graphiques, en particulier les champs aléatoires de Markov et les réseaux de Bayes, ainsi que les réseaux de neurones artificiels, en particulier les *multi-layer perceptron networks*. Le lecteur déjà familier avec ces concepts peut donc se contenter de parcourir la section 4.1, qui présente le vocabulaire et la notation employés dans ce mémoire, avant de passer directement au chapitre 5.

4.1 Vocabulaire et notation

4.1.1 Variables aléatoires et probabilités

Dans ce travail, une lettre en majuscule désigne une variable aléatoire (VA) ou un ensemble de variables aléatoires (exemples : la VA A, ou l'ensemble de VA $Z \equiv \{Z_1, \ldots, Z_n\}$). Une lettre en minuscule désigne une valeur possible pour une VA, ou un ensemble de valeurs possibles pour un ensemble de VA (exemples : a désigne une valeur possible pour A, $z \equiv \{z_1, \ldots, z_n\}$ désigne un ensemble de valeurs possibles

pour $Z_1, \ldots, Z_n$).

L'expression $P(A = a)$ désigne la probabilité que la VA A prennent la valeur a. L'expression $P(Z = z) \equiv P(Z_1 = z_1, \ldots, Z_n = z_n)$ désigne la probabilité que les VA $Z_1, \ldots, Z_n$ prennent respectivement les valeurs $z_1, \ldots, z_n$. Afin d'alléger le texte, ceci est simplement noté par $P(z) \equiv P(z_1, \ldots, z_n)$ lorsqu'aucune confusion n'est possible.

La loi de probabilité de A, notée p_A, est définie par $p_A(a) = P(A = a)$. La loi de probabilité conjointe de $Z \equiv Z_1, \ldots, Z_n$, notée $p_Z \equiv p_{Z_1,\ldots,Z_n}$, est définie par $p_Z(z) = P(Z_1 = z_1, \ldots, Z_n = z_n)$.

4.1.2 Probabilités conditionnelles et indépendance

Soient A, B et C trois ensembles disjoints de variables aléatoires. L'indépendance de A et B est notée par $A \perp B$. Ceci est équivalent à affirmer que $P(a, b) = P(a)P(b)$. L'expression $A \perp B \mid C$ dénote l'indépendance conditionnelle à C. Ceci est équivalent à affirmer que $P(a, b \mid c) = P(a \mid c)P(b \mid c)$.

La loi de probabilité de A conditionnelle à C, notée $p_{A|C}$, est définie par $p_{A|C}(a \mid c) = P(A = a \mid C = c)$.

4.1.3 Graphes d'indépendance

Soit $Z \equiv \{Z_1, \ldots, Z_n\}$ un ensemble de VA et soit $G = (V, E)$ un graphe. L'ensemble $V = \{V_1, ..., V_n\}$ des sommets de G compte un sommet pour chaque VA de Z (le sommet V_i est associé à Z_i). Dans les graphes considérés, il n'y a au maximum qu'un arc entre deux sommets et il n'y a aucun arc d'un sommet à lui-même. Dans le cas où les arcs sont dirigés, on considère uniquement des graphes où il n'y a pas de cycles orientés. Un tel graphe définit un *ordre partiel* sur Z. L'*absence* d'un arc (V_i, V_j) dans l'ensemble des arcs E indique une hypothèse d'indépendance (au sens probabiliste) entre Z_i et Z_j. La nature exacte des hypothèses d'indépendance associées à un certain graphe G dépend du type de graphe considéré (dirigé ou non) et celles-ci sont

appelées *propriétés de Markov* de G (voir plus loin). On dénote par $\mathcal{P}(G)$ la famille de lois de probabilité qui satisfont les propriétés de Markov de G.

Dans le cas non-dirigé, une telle structure est appelée *graphe d'indépendance non-dirigé* (GIND). Dans le cas dirigé, on parle de *graphe d'indépendance dirigé* (GID).

Dans un graphe non-dirigé, $N(V_i)$ représente les *voisins* du sommet V_i (i.e. tous les sommets de G liés à V_i par un arc). De façon similaire, $C(V_i)$ est l'ensemble des sommets connectés a V_i par un chemin dans G. Dans le cas dirigé, $N^-(V_i)$ et $N^+(V_i)$ dénotent respectivement les *parents* et les *enfants* de V_i. Également, $C^-(V_i)$ et $C^+(V_i)$ désignent respectivement les *ancêtres* et les *descendants* de V_i. L'ensemble des variables aléatoires associées aux sommets de $N(V_i)$ (resp. $C(V_i)$, $N^-(V_i)$, $N^+(V_i)$, $C^-(V_i)$, $C^+(V_i)$) est noté $N_i \equiv N(Z_i)$ (resp. $C_i \equiv C(Z_i)$, $N_i^- \equiv N^-(Z_i)$, $N_i^+ \equiv N^+(Z_i)$, $C_i^- \equiv C^-(Z_i)$, $C_i^+ \equiv C^+(Z_i)$).

Un graphe est *complet* s'il existe un arc entre chaque paire de sommets. Dans un graphe non-dirigé G, une *clique* C est un sous-graphe complet maximal (i.e. C augmenté d'un arc de G n'appartenant pas déjà à C n'est pas un sous-graphe complet). Le *graphe des cliques* G^C de G est un graphe qui contient un sommet pour chaque clique de G, et où il y a un arc entre deux sommets si et seulement si l'intersection des deux cliques correspondantes est non-vide. L'intersection de deux cliques adjacentes dans G^C — i.e. deux sommets adjacents de G^C — est un *séparateur*.

Dans un graphe non-dirigé G, un cycle est *sans corde* s'il n'existe pas d'arc entre des sommets non-consécutifs dans le cycle. On dit que G est *triangulé* s'il ne contient aucun cycle sans corde de taille supérieure à 3.

4.2 Modèles probabilistes graphiques

Un *modèle graphique*, aussi appelé *graphe d'indépendance* (GI), est une représentation graphique des relations d'indépendance entre un ensemble $Z = \{Z_1, \ldots, Z_n\}$ de variables aléatoires. Ces relations d'indépendance peuvent être établies à partir

de connaissances *a priori* ou apprises à partir des données [60, 61] et induisent une *factorisation* de la loi de probabilité conjointe de $Z_1, \ldots, Z_n$.

Il existe deux classes de modèles graphiques, les graphes d'indépendances non-dirigés (GIND) et les graphes d'indépendance dirigés (GID). Ceux-ci sont présentés ci-dessous. Il existe une abondante littérature sur les modèles graphiques. Cette section se veut un survol de la théorie sous-jacente et s'inspire partiellement de [62] et de [63]. Pour un traitement plus détaillé et d'autres pointeurs dans la littérature, on peut se référer à [64, 61, 65, 66, 67].

4.2.1 *Les graphes d'indépendance non-dirigés*

Les graphes d'indépendance non-dirigés (GIND) sont utilisés lorsque les relations entre les variables sont symétriques. On entend ici que les relations correspondent à une notion de corrélation plutôt que de causalité. Selon le contexte, ce type de modèle est connu sous le nom de *champ aléatoire de Markov* (MRF), *réseau de Markov* ou *machine de Boltzmann* [68, 69]. Les GIND sont par exemple utilisés en mécanique statistique ou en traitement d'image [70].

Propriétés de Markov

Soit un graphe non-dirigé $G = (V, E)$ et défini pour $Z = \{Z_1, \cdots, Z_n\}$ comme dans la section 4.1. *G est un GIND pour Z si et seulement si, pour tout $i, j \in \{1, \cdots, n\}$ satisfaisant $j \neq i$ et $V_j \notin N(V_i)$, la relation d'indépendance conditionnelle $Z_i \perp Z_j \mid N_i$ tient dans Z.* Cette propriété est appelée *propriété de Markov locale* et on peut montrer qu'elle est équivalente à l'énoncé : $\forall i \in \{1, ..., n\}, P(Z_i = z_i \mid Z - \{Z_i\} = z - \{z_i\}) = P(Z_i = z_i \mid N_i = n_i)$. Des formulations alternatives sont présentées dans [62, 63] mais on peut montrer qu'elles sont toutes équivalentes.

Factorisation induite

Les lois de probabilité conditionnelles $p_{Z_i|N_i}$ sont appelées *caractéristiques locales* et déterminent p_Z d'une manière unique [71, 62], quoique d'une manière en général complexe. On peut montrer que G est un GIND pour $Z = \{Z_1, \cdots, Z_n\}$ si et seulement si il existe un ensemble de fonctions positives f_C pour chaque *clique* C de G et satisfaisant

$$p_Z(z) = \frac{e^{-U(z)}}{\Lambda} = \frac{e^{-\sum_c f_c(z_c)}}{\Lambda} \tag{4.1}$$

où $\Lambda = \sum_x e^{-U(x)}$ est un terme de normalisation appelé *fonction de partition* [71]. Dans ce cas, la loi de probabilité de Z est une *loi de Boltzmann-Gibbs*. Bien qu'il soit facile de dériver les caractéristiques locales à partir des fonctions de clique (les f_C), il n'existe pas en général un ensemble *unique* de fonctions de clique compatibles avec les caractéristiques locales (voir la discussion dans [71], chap. 2). Dans plusieurs applications des GIND (par exemple la modélisation du noyau hydrophobe des protéines [6] et le traitement d'images [70]), p_Z est modélisée à partir du design *ad hoc* des fonctions de clique.

Cependant, dans le cas particulier où G est *triangulé*, il existe une décomposition simple de $p_Z(z)$. Dans ce cas, on a

$$p_Z(z) = \frac{\prod_c p_{Z_c}(z_c)}{\prod_s p_{Z_s}(z_s)} \tag{4.2}$$

où c désigne toutes les cliques et s tous les séparateurs qu'on retrouve dans un *arbre de jonction* [63, 65] de G^C, le graphe des cliques de G.

4.2.2 Les graphes d'indépendance dirigés

Dans les graphes d'indépendance dirigés (GID), les arcs sont habituellement associés à des relation irréversibles (par exemple dans le temps) entre les variables. Les GIDs sont aussi connus sous le nom de *réseaux bayesiens, champs de Markov dirigés, réseaux de croyances ("belief networks")* ou *diagrammes d'influence*. Ce type de modèle est par exemple employé dans les systèmes experts [72, 73].

Propriétés de Markov

Tout comme pour les GIND, on peut préciser la sémantique des relations d'indépendance conditionnelle impliquées par un GID par le biais de ses propriétés de Markov. Soit $G = (V, E)$, un graphe dirigé et acyclique, défini pour $Z = \{Z_1, \cdots, Z_n\}$ comme dans la section 4.1. *G est un GID pour Z si et seulement si, pour tout $i, j \in \{1, \cdots, n\}$ satisfaisant $j \neq i$ et $V_j \notin C^+(V_i)$, la relation d'indépendance conditionnelle $Z_i \perp Z_j \mid N_i^-$ tient dans Z.* Ceci est la *propriété de Markov locale* pour les GID et, similairement au cas non-dirigé, on peut montrer qu'elle est équivalente à l'énoncé : $\forall i \in \{1, \cdots, n\}, P(Z_i = z_i \mid C_i^- = c_i^-) = P(Z_i = z_i \mid N_i^- = n_i^-)$. Des formulation alternatives sont aussi présentées dans [62, 63] et on peut montrer qu'elles sont toutes équivalentes.

Factorisation induite

Les lois de probabilité conditionnelles $p_{Z_i|N_i^-}$ sont appelées *caractéristiques locales* et contrairement au cas des GIND, elles déterminent *directement* la loi conjointe p_Z [62, 63]. On obtient en effet l'expression suivante pour p_Z :

$$p_Z(z) = \prod_i p_{Z_i|N_i^-}(z_i \mid n_i^-). \tag{4.3}$$

Ceci implique également que toute factorisation de p_Z à l'aide de la règle de conditionnement induit un GID.

L'important est de noter qu'*on peut obtenir un modèle pour p_Z par combinaison de modèles obtenus indépendamment pour chacune des lois de probabilité locales $p_{Z_i|N_i^-}$.*

4.2.3 Distinction entre GIND et GID

Il est important de noter que les GIND et les GID ne représentent pas la même classe de lois de probabilité. Cependant, si un GID satisfait certaines conditions, on peut trivialement construire un GIND possédant les mêmes propriétés de Markov (et

inversement). Ceci est résumé par le théorème d'équivalence suivant : *Le GIND G obtenu d'un GID H en éliminant les directions sur tous les arcs de H satisfait les mêmes relations de Markov que H — i.e. P(G) = P(H) — si et seulement si H ne contient aucun sous-graphe dont un sommet a deux parents non-adjacents ou plus. De plus, si un GIND G est triangulé, alors il existe un GID H tel que G et H satisfont les mêmes propriétés de Markov [66, 74].*

On note finalement que de nombreuses lois de probabilité peuvent être représentées à l'aide de l'une ou l'autre classe de modèles. En particulier, les modèles de Markov cachés (HMM) peuvent être vus comme un cas particulier des GIND ou des GID [63]. Les HMM constituent un bon exemple de structures contenant des variables non-observées ou cachées. Avec ce type de structure, on s'intéresse à l'état des variables cachées conditionnellement à une série d'observations sur les variables observables. Dans le cadre des GI, cette démarche est connue sous le nom de *propagation des probabilités.* On peut montrer que les algorithmes de propagation dans les GI [72, 75, 76] sont des généralisations d'algorithmes d'inférence bien connus pour les HMM, tels l'algorithme de Viterbi ou l'algorithme F-B [63]. Ces questions ne seront pas discutées en détail ici puisque dans ce mémoire, on ne s'intéresse qu'au cas des modèles graphiques où toutes les variables sont toujours observées.

4.3 Réseaux de neurones artificiels et apprentissage supervisé

Il existe plusieurs types de réseaux de neurones artificiels (ANN). Dans ce mémoire, on s'intéresse aux *multi-layer perceptron networks* (MLP) aussi appelés *feed-forward multi-layer neural networks.* L'introduction qui suit s'inspire de [77, 62].

4.3.1 Les MLP

Un ANN est composé d'unités de calcul simples et inter-connectées (les *neurones*). Les connections sont directionnelles et possèdent un *poids.* Dans un MLP, les neu-

rones sont organisées en couches successives et les connections sont telles que chaque neurone d'une couche est connecté à tous les neurones de la couche suivante. Chaque connection possède également un *poids*.

On compte au moins deux couches de neurones, une *couche d'entrée* et une *couche de sortie*. Les autres couches sont appelées *couches cachées*. Un MLP représente une *fonction* $y = f(x)$ de son *vecteur d'entrée* x. Celui-ci est encodé par la couche d'entrée et la valeur y de la fonction est encodée par la couche de sortie. La complexité de la fonction représentée par le MLP est déterminée par la taille et l'organisation des couches cachées. La figure 4.1 présente un exemple de MLP.

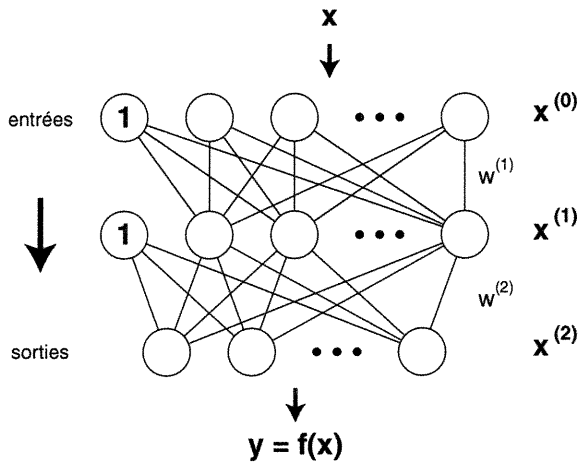

FIG. 4.1 – Exemple d'un réseau de neurones à une couche cachée. Les biais sont explicitement représentés à l'aide de neurones supplémentaires dont la valeur est fixée à 1.

La fonction calculée par un MLP peut être décrite en fonction du comportement individuel des neurones. Le neurone j de la k–ième couche cachée est responsable du calcul de

$$x_j^{(k)} = s^{(k)}\left(\sum_{i=1}^{n^{(k-1)}} w_{ji}^{(k)} x_i^{(k-1)} + b_j^{(k)}\right). \tag{4.4}$$

Ici, les $x_i^{(k-1)}$ sont les $n^{(k-1)}$ entrées du neurone (les neurones de la couche précédente),

$w_{ji}^{(k)}$ est le poids de la connection quittant le i–ième neurone de la couche précédente et $b_j^{(k)}$ est un *biais*. $x_j^{(k)}$ est appelé *valeur* du neurone j. La fonction $s^{(k)}$ est appelée *fonction d'activation* de la k–ième couche.

Les biais peuvent être explicitement représentés par l'ajout d'un neurone supplémentaire $x_0^{(k)}$ à chaque couche et dont la valeur est fixée à $x_0^{(k)} = 1$ (voir fig. 4.1). On obtient l'expression simplifiée suivante pour l'équation 4.4 :

$$x_j^{(k)} = s^{(k)}(\sum_{i=0}^{n^{(k-1)}} w_{ji}^{(k)} x_i^{(k-1)}). \tag{4.5}$$

Pour un MLP à une couche cachée comme celui de la figure 4.1, on obtient l'expression suivante pour la valeur y_l de la l–ième sortie :

$$y_l \equiv x_l^{(2)} = s^{(2)}(\sum_{j=0}^{n^{(1)}} w_{lj}^{(2)} s^{(1)}(\sum_{i=0}^{n^{(0)}} w_{ji}^{(1)} x_i^{(0)})). \tag{4.6}$$

4.3.2 Fonctions d'activation

Pour des raisons pratiques, des fonctions d'activation continues et différentiables sont préférables. Dans le cas des unités cachées, les fonctions d'activation les plus utilisées sont la *sigmoïde* et la *tangente hyperbolique* (*tanh*). Pour la couche de sortie, le choix de la fonction d'activation est lié à l'interprétation des sorties : la sigmoïde et la *tanh* sont utilisées pour des problèmes de classification ou d'estimation de densité alors que la fonction identité est utilisée pour des problèmes de régression.

Lorsque les sorties sont interprétées comme les probabilités d'occurence de n événements mutuellement exclusifs, une *fonction exponentielle normalisée* (ou $softmax$) de forme

$$softmax(x_i) \equiv \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{4.7}$$

est appropriée. Avec une telle fonction, toutes les sorties sont dans l'intervalle $[0, 1]$ et leur somme est 1.

4.3.3 Propriété d'approximation universelle

Les MLP sont des *approximateurs universels*. On peut en effet montrer qu'en autant qu'on dispose d'un nombre suffisant d'unités cachées, toute fonction raisonnable peut être approximée avec une précision arbitraire à l'aide d'un MLP à une couche cachée [77, 62].

Cette propriété est perdue si on élimine les unités cachées. On peut montrer que seules des fonctions *linéairement séparables* peuvent être représentées à l'aide de MLP sans couche cachée [77].

4.3.4 Recherche du meilleur approximateur

Soit g une fonction quelconque. On ne connaît g que par un ensemble $D \equiv \{z_1 = (x_1, y_1), ..., z_n = (x_n, y_n)\}$ de couples entrée–sortie appelées *exemples d'apprentissage* et qu'on suppose tirés indépendamment d'une même distribution de probabilité p_Z.

Si les poids d'un MLP sont vus comme des paramètres adaptatifs, celui-ci définit une classe de fonctions paramétriques $F \equiv \{f(w)\}$. Le problème auquel on s'intéresse est de trouver le *meilleur approximateur* dans F pour g. Ce processus est appelé *apprentissage.*

Si on dispose d'une *mesure d'erreur* $e(f(w), z)$ mesurant la qualité de l'approximation sur l'exemple $z = (x, y)$, le meilleur approximateur est la fonction $f(w^\star) \in F$ qui minimise

$$E_g(f(w)) \equiv \int_z e(f(w), z) p_Z(z) dz. \tag{4.8}$$

Cette quantité est appelée *risque espéré*, ou *erreur de généralisation.*

La fonction d'erreur e peut être dérivée à partir du critère du *maximum de vraisemblance* (voir section 6.2 et [77, 62]). Pour la régression, on utilise souvent la fonction d'erreur suivante

$$e(f(w), z) = \frac{1}{2}(f(w, x) - y)^2, \tag{4.9}$$

appelée *erreur quadratique.*

En pratique, la loi de probabilité p_Z est inconnue et on minimise plutôt le *risque empirique*, ou *erreur d'apprentissage*, donné par

$$E_a(f(w), D) \equiv \frac{1}{|D|} \sum_i e(f(w), z_i). \tag{4.10}$$

Pour $f(w)$ et D choisis indépendamment, $E_a(f(w), D)$ est un *estimateur bruité mais non-biaisé* du risque espéré $E_g(f(w))$.

4.3.5 *Algorithmes d'optimisation*

La recherche du meilleur approximateur $f(w^\star) \in F$ est un problème d'optimisation sur l'espace des poids. Pour les MLP, les poids sont généralement optimisés par descente de gradient. À l'aide de la règle de chaîne, les poids sont mis à jour couche par couche, de la couche de sortie vers la couche d'entrée, par propagation d'un signal d'erreur le long des connections entre les neurones. On appelle cette procédure *algorithme de rétro–propagation* [77].

Dans la version stochastique de cet algorithme, les poids sont mis à jour après la présentation de chaque exemple d'apprentissage à l'aide de l'équation

$$w_{ji}^{(k)} = -\eta \frac{\partial e}{\partial w_{ji}^{(k)}} \tag{4.11}$$

où l'hyper-paramètre η est appelé *taux d'apprentissage.*

Si on définit

$$\alpha_j^{(k)} \equiv \sum_{i=0}^{n^{(k-1)}} w_{ji}^{(k)} x_i^{(k-1)} \tag{4.12}$$

la dérivée dans 4.11 est donnée par

$$\frac{\partial e}{\partial w_{ji}^{(k)}} = \frac{\partial e}{\partial x_j^{(k)}} \frac{\partial x_j^{(k)}}{\partial w_{ji}^{(k)}} = \frac{\partial e}{\partial x_j^{(k)}} \, s^{(k)'}(\alpha_j^{(k)}) x_i^{(k-1)} \tag{4.13}$$

où la quantité

$$\epsilon_j^{(k)} \equiv \frac{\partial e}{\partial x_j^{(k)}} \, s^{(k)'}(\alpha_j^{(k)}) \tag{4.14}$$

est appelée *erreur propagée.*

Pour la couche de sortie, la dérivée $\partial e/\partial x_j^{(k)}$ peut être évaluée directement. Pour les couches cachées, elle peut être calculée recursivement à partir des erreurs propagées de la couche suivante et on obtient

$$\begin{aligned} \frac{\partial e}{\partial x_j^{(k)}} &= \sum_i \frac{\partial e}{\partial x_i^{(k+1)}} \frac{\partial x_i^{(k+1)}}{\partial x_j^{(k)}} \\ &= \sum_i \epsilon_i^{(k+1)} \, w_{ij}^{(k+1)} \end{aligned} \tag{4.15}$$

4.3.6 Considérations pratiques

On a vu qu'un MLP représente une classe de fonctions paramétriques $F = \{f(w)\}$. L'objectif de l'apprentissage est de sélectionner une solution $f(w^\star) \in F$ qui minimise le risque espéré (eq. 4.8). Cette section présente quelques éléments critiques pour le succès de cette démarche.

Estimation du risque espéré

Afin d'évaluer la qualité de la solution obtenue, on doit pouvoir estimer son erreur de généralisation.

Il est avant tout important de noter que $f(w^\star)$ est *biaisée* vers l'ensemble d'apprentissage D : $f(w^\star)$ est systématiquement "meilleure" sur un exemple tiré de D que sur un nouvel exemple tiré au hasard de la véritable loi de probabilité p_Z des exemples. L'erreur d'apprentissage $E_a(f(w^\star), D)$ de $f(w^\star)$ sur D est donc un *estimé biaisé* de l'erreur de généralisation.

Un estimé non-biaisé peut être obtenu en partitionnant D en un ensemble d'apprentissage D_a et en un ensemble de test D_t. Si on sélectionne $f(w^\star)$ par minimisation du risque empirique sur D_a, le risque empirique $E_a(f(w^\star), D_t)$ évalué sur D_t est un estimé *non-biaisé* du risque espéré (section 4.3.4).

Capacité

La *capacité h(F)* d'un ensemble de fonctions est une mesure de sa *diversité* et de sa *complexité*. Pour les MLP, $h(F)$ est déterminée par le nombre d'unités cachées.

Si la capacité est insuffisante, la fonction optimale $f(w^\star) \in F$ sera une approximation trop grossière. Dans le cas contraire, $f(w^\star)$ accordera trop d'importance à la structure particulière de l'ensemble d'apprentissage et généralisera mal sur des cas hors–échantillon. Ceci est présenté dans la figure 4.2a.

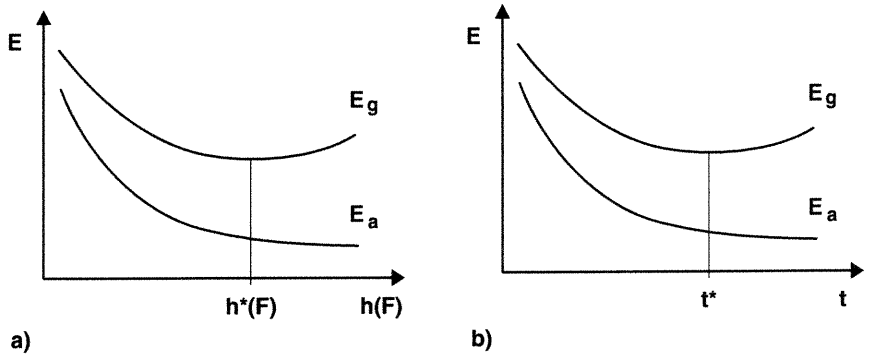

FIG. 4.2 – a) Influence de la capacité $h(F)$ sur l'erreur de généralisation. $h^\star(F)$ indique la capacité optimale. b) Comportement de E_a et E_g en fonction du nombre d'itérations allouées à l'optimisation. $t^\star$ indique le nombre optimal d'itérations.

Condition d'arrêt

On a vu que les poids peuvent être estimés itérativement à l'aide de l'algorithme de rétro–propagation. À chaque itération, le risque empirique diminue progressivement jusqu'à l'obtention d'un minimum. Si la capacité est suffisante, un minimum du risque espéré sera atteint *avant* le minimum du risque empirique, après quoi il y a *sur–entraînement* et le risque espéré augmente (fig. 4.2b).

Ceci suggère un critère d'arrêt pour la procédure d'optimisation : on stoppe dès qu'une suite de α iterations successives n'apporte aucune amélioration du risque

espéré. On appelle ceci *critère d'arrêt prématuré* (*early stopping* [77]).

Taux d'apprentissage

Le taux d'apprentissage η (eq. 4.11) détermine la largeur des "pas" effectués par la procédure d'optimisation en direction d'un minimum local du risque empirique. Si η est choisi trop grand, on risque d'osciller autour d'un minimum sans l'atteindre. Inversement, si η est trop petit, on risque de rester emprisonné dans un minimum local peu important et la convergence sera très lente.

On peut éviter ce problème en diminuant progressivement la valeur de η à chaque itération. On peut par exemple utiliser l'équation

$$\eta_{t+1} = \kappa \eta_t. \tag{4.16}$$

Ici, η_t dénote la valeur du taux d'apprentissage à l'itération t et κ est choisi dans $[0, 1]$. On obtient une procédure plus "sautilleuse" au départ (donc capable d'éviter des minimums peu importants) mais dont la convergence est assurée en un temps raisonnable [77].

Choix des hyper-paramètres

La capacité h, le nombre α d'itérations sans amélioration déterminant l'arrêt, le taux d'apprentissage initial η_0 et la contante κ sont des *hyper-paramètres* qui doivent être optimisés en fonction du nombre d'exemples disponibles et de la complexité du problème considéré.

Soit $E_g(\lambda)$ l'erreur de généralisation du meilleur modèle obtenu à l'aide d'une hyper-paramétrisation λ. On peut obtenir un estimateur $\widehat{E}_g(\lambda)$ pour $E_g(\lambda)$ à l'aide d'un ensemble de test tel que discuté plus haut. L'hyper–paramétrisation optimale $\lambda^\star$ est alors donnée par

$$\lambda^\star = \arg\min_\lambda(\widehat{E}_g(\lambda)). \tag{4.17}$$

Lorsque peu d'exemples sont disponibles, la technique de *validation croisée* permet l'obtention d'un meilleur estimateur pour $E_g(\lambda)$. Ici, on partitionne d'abord D en k sous-ensembles disjoints. On entraîne ensuite un MLP sur l'ensemble E_a défini par l'union de $k-1$ sous-ensembles et on estime $E_g(\lambda)$ sur le sous-ensemble restant. On répète pour chacun des k choix possibles pour E_a. L'estimateur résultant pour $E_g(\lambda)$ est donné par la moyenne des estimés obtenus à chaque expérience [77].

4.3.7 Applications en biologie moléculaire

Les réseaux de neurones, les MLP en particulier, ont été appliqués à un grand nombre de problèmes en biologie moléculaire. Un exemple bien connu est la prédiction de la structure secondaire des protéines à partir de leurs séquences [62, 78]. D'autres exemples sont la prédiction des sites de clivage des peptide de signalisation ou la prédiction des sites d'épissage [62]. On peut consulter [79] pour une revue et des exemples additionnels.

Chapitre 5

DÉFINITION DE MODÈLES STOCHASTIQUES POUR LES PROTÉINES PARTAGEANT UN MÊME MOTIF DE REPLIEMENT

Dans le chapitre 3, on a discuté des techniques disponibles pour "reconnaître" le motif de repliement d'une protéine à partir d'une banque de conformations possibles. Le but premier de ce chapitre est de présenter une formulation probabiliste de la compatibilité séquence–structure dans les protéines. Cette formulation est due à White *et al* [6] et permet d'unifier le problème général de la reconnaissance séquence–structure et le problème plus spécifique de l'alignement séquence-structure à l'intérieur d'un même cadre formel [46]. Cette formulation constitue également une excellente base pour l'application des puissantes techniques d'apprentissage (*machine learning*) au problème de la reconnaissace des motifs de repliement des protéines.

Le reste de ce chapitre est organisé comme suit : Dans la section 5.1, on discute en termes généraux de la formulation probabiliste proposée par White *et al* [6]. Ensuite, dans la section 5.2, on présente une représentation simplifiée pour les motifs de repliement des protéines. Cette représentation, formalisée par White *et al* [6], repose sur la notion de *graphe de contact* et constitue un bon point de départ pour la construction de modèles stochastiques pour les séquences partageant un même motif de repliement. Finalement, dans la section 5.3, on présente un modèle stochastique pour la représentation des protéines qui adoptent un même motif de repliement. Ce modèle est également dû à White *et al* [6] et repose sur la notion de champ aléatoire de Markov.

5.1 *Formulation probabiliste de la compatibilité séquence–structure dans les protéines*

On sait que des séquences d'acides aminés très différentes, apparentées ou non dans l'évolution, peuvent partager une même architecture. Par opposition, des changements en apparence mineurs dans la séquence d'une protéine peuvent empêcher son repliement correct. En outre, on sait que certaines protéines, les protéines prions ont la possibilite d'adopter plus d'une conformation. Ces observations illustrent la nature intrinsèquement bruitée et incertaine de l'information structurale contenue dans les séquences d'acides aminés. Ceci est compréhensible quand on sait que les séquences biologiques ne sont en fait que le produit d'un grand nombre d'évènements aléatoires amplifiés à travers l'évolution. Sachant cela, il apparaît naturel de représenter les relations séquence–structure dans les protéines à l'intérieur d'un cadre probabiliste.

Dans le formalisme proposé par White *et al* [6], un motif de repliement, par exemple l'architecture commune d'un ensemble de protéines, est associé à une séquence de variables aléatoires $Z = Z_1 \dots Z_n$ représentant des séquences d'acides aminés. La séquence Z représente toutes les séquences — possiblement non-homologues — qui adoptent une conformation similaire au motif considéré.

La loi de probabilité p_Z de Z est *a priori* inconnue. On sait seulement que la probabilité $P(Z = s)$ doit réfléter l'affinité structurale de la séquence s pour le motif considéré et on ne peut estimer la loi de probabilité de Z qu'à partir des valeurs observées de Z, i.e. les séquences d'acides aminés connues qui adoptent une conformation similaire au motif de repliement considéré. L'objectif est donc de trouver un bon *modèle* pour p_Z à partir des valeurs observées de Z. [1]

[1] Par opposition, dans la formulation traditionnelle présentée dans la section 3.2, on cherche plutôt un potentiel pseudo-énergétique qui attribue une bonne "énergie" aux séquences qui sont compatibles avec le motif considéré.

5.1.1 Reconnaître le "bon" motif de repliement

Si on dispose d'un bon modèle M pour p_Z, la vraisemblance $P(Z = z \mid M)$ sous M peut être utilisée pour détecter de nouvelles séquences qui adoptent le motif de repliement considéré. Par contre, qu'en est-il lorsqu'on cherche plutôt à "reconnaître" le motif de repliement adopté par une séquence parmi un certain nombre de motifs possibles ? Dans ce cas, on doit d'abord construire (et entraîner) un modèle stochastique pour chaque motif considéré. On peut ensuite reconnaître le bon motif en comparant les probabilités *a posteriori*

$$P(M_i \mid z) = \frac{P(M_i)}{P(z)} P(z \mid M_i) \tag{5.1}$$

des modèles stochastiques entraînés $M_1, M_2, \ldots$ associés aux différentes alternatives possibles [46]. [2]

5.1.2 Mise au point d'un bon modèle pour p_Z

L'obtention d'un bon modèle pour p_Z peut être décomposée en deux étapes : la *définition du modèle* et l'*estimation des paramètres du modèle* (ou *apprentissage*). La première étape correspond à la définition d'une forme paramétrique $M(\theta)$ pour M. La seconde étape correspond à la recherche de la meilleure paramétrisation possible pour M : on cherche $\theta^\star$ qui permet l'approximation la plus fidèle de p_Z à l'aide de $M(\theta^\star)$.

Cependant, pour un motif d'une certaine complexité, p_Z est une loi de probabilité en haute dimension : pour un motif de taille n, p_Z définit une distribution de probabilité sur 20^n séquences (ou, d'une manière équivalente, la distribution jointe de n variables aléatoires discrètes, chacune prenant pour valeur un des 20 acides aminés).

[2] Comme les modèles $M_1, M_2, \ldots$ ne sont pas des variables aléatoires, les quantités $P(M_i)$ et $P(M_i \mid z)$ ne représentent pas vraiment des probabilités. On devrait plutôt parler de *mesures de confiance* (*a priori* et *a posteriori*) envers les différents modèles.

Ceci surpasse largement la taille de l'ensemble des données disponibles pour l'apprentissage (l'ensemble des protéines connues qui adoptent le motif de repliement associé à Z) et on ne peut espérer représenter p_Z de manière exacte à l'aide de M.

Pour contrer ce problème, aussi appelé *malédiction de la dimensionalité*, deux types d'approches sont possibles. Une première idée, qui consiste à ne pas modéliser toutes les dépendances entre les variables, est formalisée par les *modèles graphiques*, ou graphes d'indépendance (voir le chapitre 4). Le modèle stochastique proposé par White *et al* [6] (section 5.3) repose sur cette idée.

Une seconde approche consiste à approximer la forme mathématique de p_Z. Les réseaux de neurones artificiels d'estimation de densité [80] et les approximations polynômiales [81, 64] sont basés sur cette idée. Dans les approximations polynômiales, on se restreint à une forme qui ne tient compte que des dépendances d'ordre inférieur (typiquement d'ordre 1 ou 2). Le modèle stochastique qui constitue la principale contribution du présent travail (chapitre 6) repose sur les deux types d'approches : sa structure est définie à l'aide d'un réseau de Bayes et il est paramétrisé à l'aide de réseaux de neurones artificiels.

5.2 Représentation des motifs de repliement des protéines

Dans cette section, on s'intéresse à la représentation formelle des motifs de repliement des protéines. La représentation présentée ci-dessous est celle qu'on adoptée White *et al* dans [6].

5.2.1 Définition de patrons structuraux

Un motif de repliement a été défini au chapitre 2 comme un assemblage d'éléments de structure secondaire présent dans plusieurs protéines (apparentées ou non) et souvent associé à une fonction précise. Un motif de repliement peut être formellement représenté à l'aide d'un *patron structural*. Un patron structural est défini à l'aide d'une

séquence de *segments structuraux*, de longueur fixe, et connectés par des segments de longueur variables appelées *boucles*.

Les segments structuraux correspondent aux éléments de structure secondaire (hélices–α et brins–β) qui sont conservés dans toutes les protéines qui adoptent le motif considéré. On entend ici que ces éléments sont présents dans toutes ces protéines et que leur longueur, leur position et leur orientation relative sont préservées (ces éléments forment ce qu'on appelle un *noyau structural conservé*).

Par opposition, les boucles correspondent aux régions structurales mal conservées dans les protéines qui adoptent le motif. Ces régions sont typiquement de longueur variable et sont non-structurées (i.e. ce ne sont ni des hélices–α, ni des brins–β).

5.2.2 Représentation du noyau structural conservé

Le patron décrit jusqu'ici ne permet de décrire que la position relative sur la séquence des éléments qui constituent le noyau structural conservé. Cependant, on aimerait aussi encoder la position et l'orientation relative de ces éléments dans l'espace 3-D. Ceci peut être réalisé à l'aide d'un graphe non-dirigé $G = (V_C, E)$, appelé *graphe de contact* (voir la fig. 5.1).

L'ensemble $V_C = \{V_1, V_2, \ldots, V_m\}$ des sommets de G peut être interprété comme un ensemble de *positions spatiales*. Un nombre fixe de positions sont associées à chaque segment structural du patron. Les positions spatiales correspondent à la position des acides aminés appartenant au noyau structural conservé.

Chaque sommet $V_i \in V_C$ est également étiqueté par un ensemble Q_i de k *attributs environnementaux* (ou propriétés). Ces propriétés permettent de préciser l'environnement structural autour de chaque position (e.g. le type de structure secondaire, l'accessibilité au solvant, etc.). On peut exprimer formellement ceci à l'aide d'une fonction $Q : V_C \rightarrow \mathbb{R}^k$ qui associe l'ensemble de propriétés $Q(V_i) = Q_i = \{{q_i}^{(1)}, \cdots, {q_i}^{(k)}\}$ à chaque sommet $V_i \in V_C$ (voir fig. 5.2). Dans ce qui suit, on dira que deux sommets $V_i \in V_C$ et $V_j \in V_C$ sont du même *type* si et seulement si $Q(V_i) = Q(V_j)$.

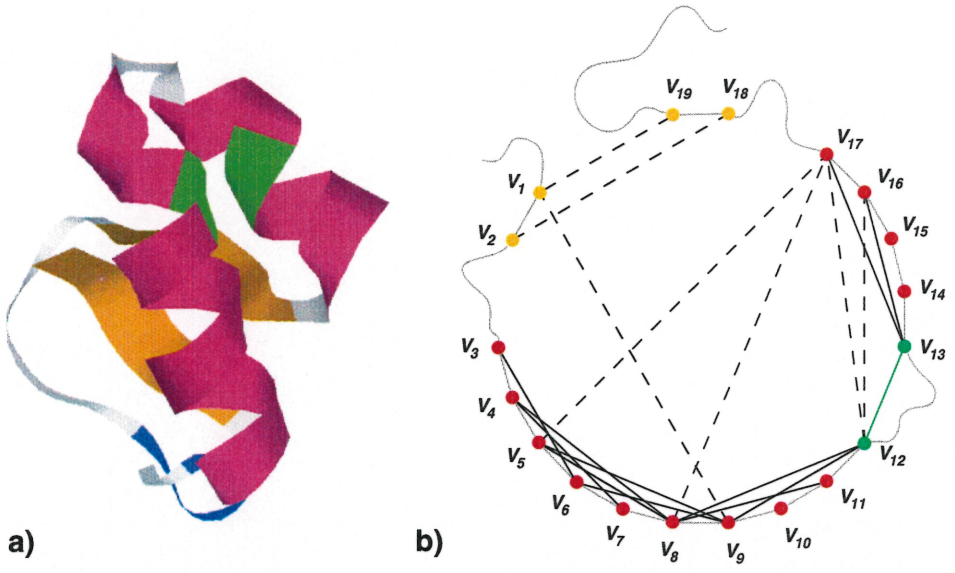

FIG. 5.1 – Graphe de contact pour le noyau structural de la protéine CRAMBIN, présente dans les graines des plantes. Les sommets correspondent aux éléments de structure secondaire (ESS) (hélice–α, en rouge, et feuillet–β, en jaune). Les arcs correspondent à des interactions entre acides aminés (ligne pointillée : contacts inter-ESS, ligne mince : contacts intra-ESS).

L'ensemble E des arcs de G permet la représentation d'interactions physico–chimiques (ou *contacts*) entre les acides aminés occupant des positions spatiales voisines dans l'espace 3-D. Chaque arc $(V_i, V_j) \in E$ est également caractérisé par un ensemble $R_{i,j}$ de propriétés environnementales. Ces propriétés permettent de préciser la nature des différents contacts (e.g. contact entre deux positions d'un même élément de structure secondaire, distance entre les positions en contact, etc.). On peut exprimer ceci formellement à l'aide d'une fonction $R : E \rightarrow \mathbb{R}^l$ qui associe l'ensemble de *propriétés* $R(V_i, V_j) = R_{i,j} = \{r_{i,j}{}^{(1)}, ..., r_{i,j}{}^{(l)}\}$ à chaque arc $(V_i, V_j) \in E$ (voir fig. 5.3). Il est physiquement difficile d'associer une notion de directionalité aux contacts et on considère que ceux-ci sont *symétriques*. Ceci est équivalent à supposer que $R_{i,j} = R_{j,i}$ dès que $i \neq j$. Finalement, on dira que deux arcs $(V_i, V_j) \in E$ et

$$
\begin{array}{|l|}
\hline
{q_i}^{(1)} = \begin{cases} 0 & \text{si } V_i \text{ appartient à une hélice } \alpha \\ 1 & \text{si } V_i \text{ appartient à un feuillet } \beta \end{cases} \\
\hline
{q_i}^{(2)} = \begin{cases} 0 & \text{si } V_i \text{ est une position exposée au solvant} \\ 1 & \text{si } V_i \text{ est une position protégée du solvant} \end{cases} \\
\hline
\end{array}
$$

FIG. 5.2 – Exemple d'une fonction $Q : V_C \rightarrow \mathbb{R}^2$ assignant 2 attributs environnementaux aux $V_i \in V_C$.

$(V_k, V_l) \in E$ sont du même *type* si et seulement si on a $Q(V_i) = Q(V_k)$, $Q(V_j) = Q(V_l)$ et $R(V_i, V_j) = R(V_k, V_l)$.

$$
\begin{array}{|l|}
\hline
{r_{i,j}}^{(1)} = \begin{cases} 0 & \text{contact des chaînes latérales à l'intérieur d'un ESS} \\ 1 & \text{contact des chaînes latérales entre deux ESS} \end{cases} \\
\hline
\end{array}
$$

FIG. 5.3 – Exemple d'une fonction $R : E \rightarrow \mathbb{R}^1$ assignant des propriétés environnementales à chaque arc de E.

5.2.3 Discussion

Bien que les graphes de contact puissent être utilisés pour décrire la structure de protéines individuelles (e.g. comme dans la fig. 5.1), il est important de noter qu'un patron structural vise à représenter les caractéristiques structurales conservées dans un ensemble de protéines partageant une même architecture (e.g. structure secondaire, interactions responsables de la stabilité de la structure) plutôt que des détails structuraux propres à une seule protéine (e.g. la séquence d'acides aminés). Le type de patron décrit ci-dessus constitue donc un bon point de départ pour la mise au point de modèles stochastiques pour les séquences (possiblement non-homologues) qui partagent un même motif de repliement.

5.3 *Un modèle stochastique pour des protéines partageant un même motif de repliement*

Soit $Z = Z_1 Z_2 \dots Z_n$ une séquence de variables aléatoires représentant des séquences d'acides aminés de longueur n et partageant un même motif de repliement. On suppose disposer d'un patron structural pour ce motif (défini comme dans la section 5.2), ainsi que d'un graphe de contact $G = (V_C, E)$ représentant le noyau structural conservé. L'objectif est d'obtenir un modèle pour la loi de probabilité p_Z, sachant que la probabilité $P(Z = s)$ doit réfléter l'affinité structurale de la séquence s pour le motif de repliement considéré. White *et al* [6] ont proposé un modèle basé sur les champs aléatoires de Markov.

5.3.1 *Structure du modèle de White*

Avant de commencer, il est utile de différencier les variables aléatoires représentant les acides aminés du noyau structural conservé et les variables aléatoires représentant les acides aminés des régions non-conservées (les boucles). Pour ce faire, on définit l'ensemble $C \subset \{1, 2, \dots, n\}$, qui désigne les positions de la séquence Z associées au noyau structural, ainsi que l'ensemble $Z_C \subset Z$ correspondant. On peut donc établir une correspondance un-à-un entre les variables aléatoires de Z_C et les sommets du graphe de contact $G = (V_C, E)$. On définit également l'ensemble $L \equiv \{1, 2, \dots, n\} - C$, qui désigne les positions de Z associées aux boucles, ainsi que l'ensemble $Z_L \equiv Z - Z_C$ correspondant.

Dans le modèle de White *et al* [6], les acides aminés constituant le noyau structural sont modélisés indépendamment des acides aminés associés aux boucles. Ceci correspond à l'hypothèse

$$p_Z(z) = p_{Z_C}(z_C) p_{Z_L}(z_L) \tag{5.2}$$

et est analogue à l'équation 3.4 de la section 3.2.3. De plus, les variables aléatoires affectées aux boucles sont supposées indépendantes et indentiquement distribuées

selon une loi de probabilité p_L. La contribution des boucles est donc donnée par

$$p_{Z_L}(z_L) = \prod_{i \in L} p_L(z_i). \tag{5.3}$$

Cependant, comme la plupart des interactions qui stabilisent une protéine surviennent entre des acides aminés situés dans son noyau, on ne peut modéliser ceux-ci à l'aide de variables aléatoires indépendantes. Pour les protéines, une hypothèse raisonnable est d'assumer que la présence d'un acide aminé à une position donnée ne dépend que des acides aminés avec lesquels celui-ci est physiquement susceptible d'interagir — les acides aminés associés à des positions voisines dans le graphe de contact — et de l'environnement physico-chimique à ces positions. Ceci correspond à assumer une relation d'indépendance entre Z_i et Z_j dès que les sommets correspondants V_i et V_j sont non-connectés dans le graphe de contact.

Comme les interactions entre acides aminés sont considérées symétriques (non-directionnelles), ceci a naturellement conduit White *et al* [6] vers la définition d'un graphe d'indépendance non-dirigé (GIND), ou champ aléatoire de Markov, pour Z_C : ceux-ci supposent que le graphe de contact $G = (V_C, E)$ définit un GIND pour Z_C (voir fig. 5.4), c'est-à-dire qu'ils supposent que $p_{Z_C} \in \mathcal{P}(G)$. Les hypothèses d'indépendance évoquées ci-dessus correspondent donc aux propriétés de Markov de G : on a que $Z_i \perp Z_j \mid N_i$ dès que $i \neq j$ et $V_j \notin N(V_i)$.

Comme le graphe de contact G ne satisfait pas en général la condition de triangulation (section 4.2.1), on ne peut exprimer p_{Z_C} de façon simple à l'aide de l'équation 4.2. Cependant, par un design *ad hoc* de *fonctions de cliques* pour G (revoir la section 4.2.1), White *et al* [6] obtiennent l'expression suivante pour p_{Z_C} :

$$p_{Z_C}(s_C) = \frac{1}{\Lambda} \prod_{i \in C} p_i(s_i) \prod_{(i,j) \in E} \frac{p_{i,j}(s_i, s_j)}{p_i(s_i) p_j(s_j)}. \tag{5.4}$$

Ici, $p_i \equiv p_{Z_i}$ est la densité marginale à la position V_i, $p_{i,j} \equiv P_{Z_i, Z_j}$ est la densité jointe aux positions V_i et V_j et Λ est la fonction de partition de l'équation 4.1.

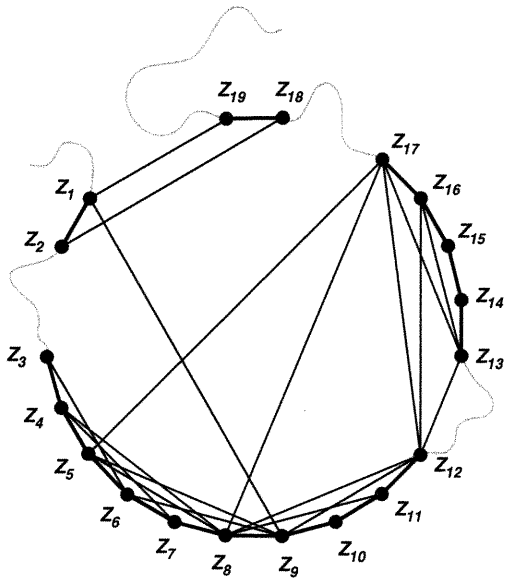

FIG. 5.4 – Approximation à l'aide d'un GIND de la loi de probabilité des séquences partageant le noyau structural de la protéine CRAMBIN. Ce GIND est défini à partir du graphe de contact de la figure 5.1. Le sommet Z_i est une variable aléatoire associée à la position spatiale V_i du graphe de contact. L'absence d'un arc (Z_i, Z_j) indique une hypothèse d'indépendance entre Z_i et Z_j.

5.3.2 Paramétrisation du modèle de White

La probabilité $p_i(x)$ (resp. $p_{i,j}(x,y)$) est estimée par la *fréquence d'occurence* de l'acide aminé x (resp. la paire d'acides aminés (x,y)) à l'intérieur d'*environnements équivalents* dans toutes les protéines connues. Plus précisément, on suppose $p_i^{(A)}(x) = p_k^{(B)}(x) \quad \forall\, x, i, k, A, B$ dès que les positions $V_i^{(A)}$ d'un motif A et $V_k^{(B)}$ d'un motif B sont *du même type* (section 5.2) et on suppose $p_{i,j}^{(A)}(x,y) = p_{k,l}^{(B)}(x,y) \quad \forall\, x, y, i, j, k, l, A, B$ dès que les interactions $(V_i^{(A)}, V_j^{(A)})$ et $(V_k^{(B)}, V_l^{(B)})$ sont *du même type*. Une table des fréquences d'occurence (TFO), ou histogramme, est ainsi construite pour chaque type de position et pour chaque type d'interaction [6].

5.3.3 *Un modèle pour* toutes *les séquences qui adoptent le même motif de repliement*

La discussion ci-dessus a passé sous silence un détail important. Il importe de noter qu'il y a plus d'un choix valable pour C, l'ensemble des positions de séquence associées au noyau structural. Un choix valide c pour C peut être construit en associant progressivement certains des Z_i aux positions spatiales du noyau, de façon à ce que des positions adjacentes d'un même segment structural soient associées à des variables aléatoires adjacentes dans la séquence Z. Un choix particulier fixe la position sur la séquence des régions correspondant au noyau structural conservé, ainsi que la position et la longueur des régions correspondant aux boucles. Or la position et la longueur de ces régions n'est pas nécéssairement la même pour deux séquences de même longueur qui partagent une même architecture (voir fig. 5.5a). Un choix particulier c pour C n'est donc approprié que pour représenter un certain sous-ensemble des protéines de longueur n qui adoptent le motif de repliement considéré (voir fig. 5.5b).

Il est important de remarquer que le modèle proposé par White *et al* dans [6] est un modèle pour $p_{Z|C}$, la loi de probabilité de Z lorsque la position des régions de séquence associées au noyau est fixée. Cependant, un modèle complet pour p_Z devrait tenir compte de toutes les positions possibles pour ces régions. Un tel modèle peut facilement être dérivé à partir du modèle de White en remarquant que

$$p_Z(z) = \sum_c p_{Z|C}(z \mid c) p_C(c), \tag{5.5}$$

où la somme s'effectue sur toutes les positions possibles des régions associées au noyau [46].

5.3.4 *Discussion*

Il est intéressant de noter que le calcul de la somme dans l'équation 5.5 est analogue à l'évaluation de tous les alignements possibles d'une séquence dans un patron

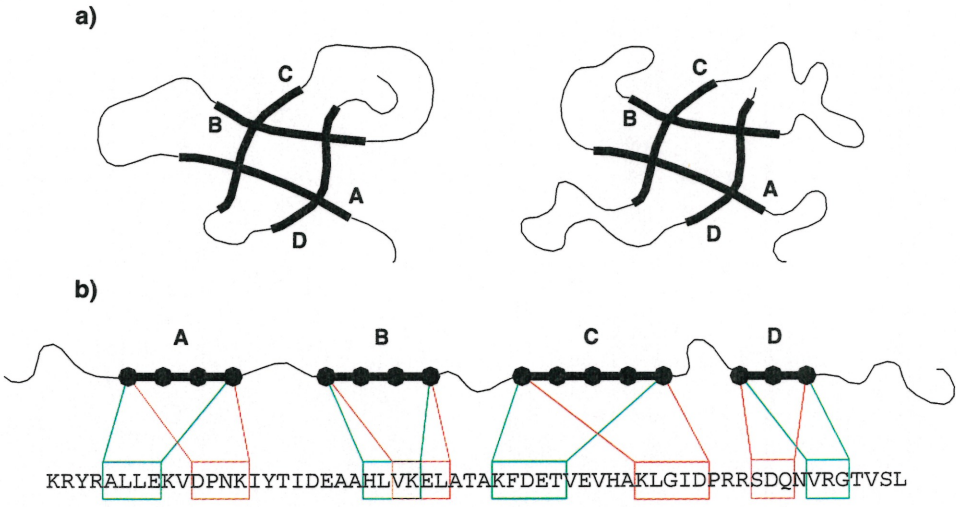

FIG. 5.5 – a) Représentation schématique du noyau structural conservé (en gras) entre deux protéines partageant une même architecture. b) Positions possibles sur une séquence non-caractérisée des régions correspondant au noyau structural conservé.

structural à l'aide d'un potentiel pseudo-énergétique traditionnel [46]. [3]

À cet égard, White *et al* [6] argumentent que leur modèle permet la définition d'un potentiel pseudo-énergétique utilisable pour l'alignement séquence–structure. Supposons qu'on dispose d'un patron structural défini comme dans la section 5.2 et dont le noyau structural est représenté à l'aide du graphe de contact $G = (V_C, E)$. On désigne par a_C et a_L les acides aminés respectivement assignés au noyau structural et aux boucles lors de l'alignement d'une séquence $a = a_1 a_2 \ldots a_n$ dans ce patron (dans ce cas, la position des régions de séquence correspondant au noyau stuctural et aux boucles est fixée par l'alignement).

[3] Il existe cependant une différence : dans les alignements séquence–structure, la compatibilité séquence–patron est définie en fonction de l'"énergie" de l'alignement optimal de la séquence dans la patron (section 3.2.2). Par contre, dans l'équation 5.5, la compatibilité séquence–patron combine les contributions de tous les alignements possibles.

Dans leur article, White *et al* proposent d'utiliser la log-probabilité

$$-\log P(Z = a \mid C)$$

comme mesure de l'"énergie" $\mathcal{E}(a)$ de cet alignement. Lorsque cette quantité est estimée à l'aide de leur modèle stochastique, on obtient

$$\begin{aligned}
\mathcal{E}(a) &\equiv -\log p_{Z|C}(a \mid C) \\
&= -\log p_{Z_C}(a_C) - \log p_{Z_L}(a_L) \\
&= -\sum_{i \in C} \log p_i(a_i) - \sum_{(i,j) \in E} \log \frac{p_{i,j}(a_i, a_j)}{p_i(a_i) p_j(a_j)} - \log \Lambda \\
&\quad -\sum_{i \in L} \log p_L(Z_i)
\end{aligned} \tag{5.6}$$

en combinant les équations 5.2, 5.3 et 5.4.

Chapitre 6

DÉFINITION D'UN NOUVEAU TYPE DE MODÈLE STOCHASTIQUE POUR LES PROTÉINES PARTAGEANT UN MÊME MOTIF DE REPLIEMENT

Dans le chapitre 5, on a présenté une formulation probabiliste de la compatibilité séquence–structure chez les protéines. Cette formulation est due à White *et al* [6]. On a également présenté un modèle stochastique pour la représentation des séquences d'acides aminés qui adoptent un même motif de repliement. Ce modèle est également dû à White *et al* [6] et repose sur la notion de champ aléatoire de Markov.

Même si la formulation introduite par White *et al* [6] est originale, le modèle stochastique qu'ils proposent a plusieurs points faibles. En premier lieu, on a vu que l'approche de White *et al* [6] est équivalente à l'utilisation d'un potentiel pseudo-énergétique traditionnel de forme

$$\mathcal{E}(a) = \sum_i g_i(a_i) + \sum_{i,j} g_{i,j}(a_i, a_j) \tag{6.1}$$

où on compte un terme par acide aminé et un terme par paire d'acides aminés en contact (voir l'équation 5.6). Le modèle de White ne permet donc pas la représentation d'interactions d'ordre supérieur entre les acides aminés. De plus, de façon similaire à la plupart des potentiels pseudo-énergétiques traditionnels, les paramètres du modèle de White sont estimés à l'aide de *tables des fréquences d'occurence* (TFO). On a déjà discuté que cette technique fonctionne très mal si le nombre d'exemples disponibles pour l'estimation des paramètres est faible (ce qui est typiquement le cas pour les protéines). C'est d'ailleurs ceci qui explique l'impossibilité d'inclure des termes d'ordre supérieur à 2 dans l'équation 5.6. Également, l'utilisation de TFO restreint fortement le nombre et le choix des propriétés environnementales dans la définition des patrons

structuraux (section 5.2.2) car le nombre de TFO requis croît exponentiellement avec le nombre de propriétés retenues. Finalement, il est important de noter qu'avec le modèle de White, l'obtention de la probabilité $P(z_C)$ — par exemple pour le calcul de $P(z)$ à l'aide de l'équation 5.5 — requiert le calcul d'un terme de normalisation non–trivial (le Λ de l'équation 5.4).

Le but du présent chapitre est de proposer un nouveau type de modèle stochastique qui contourne toutes ces limitations. Par opposition au modèle de White, paramétrisé à l'aide de TFO et dont la structure est définie à l'aide d'un champ aléatoire de Markov, la structure du modèle proposé repose sur un réseau de Bayes et il est paramétrisé à l'aide de réseaux de neurones artificiels. Il est important de noter que le modèle stochastique proposé ci-dessous constitue la principale contribution originale du présent mémoire. Le contenu de ce chapitre est donc dû en totalité à l'auteur de ce mémoire.

La suite de ce chapitre est organisée comme suit. Premièrement, dans la section 6.1, on définit la structure générale du modèle proposé. Celle-ci induit une factorisation de p_{Z_C} en un produit de lois de probabilités locales. Deuxièmement, dans la section 6.2, on définit une forme paramétrique pour les lois locales à l'aide de réseaux de neurones artificiels. Troisièment, dans la section 6.3, on traite du partage de paramètres entre les lois locales. Finalement, dans la section 6.4, on établit l'architecture détaillée des réseaux de neurones utilisés.

6.1 Structure du modèle proposé

Soit $Z = Z_1 Z_2 \ldots Z_n$ une séquence de variables aléatoires représentant des séquences d'acides aminés de longueur n et partageant un même motif de repliement. On suppose disposer d'un patron structural pour ce motif (défini comme dans la section 5.2), ainsi que d'un graphe de contact $G = (V_C, E)$ représentant le noyau structural conservé.

Comme dans le chapitre 5, l'objectif est d'obtenir un modèle pour la loi de probabilité p_Z, sachant que la probabilité $P(Z = s)$ doit réfléter l'affinité structurale de la séquence s pour le motif de repliement considéré.

Comme dans le cas du modèle de White, on suppose d'abord que les acides aminés du noyau structural peuvent être modélisés indépendamment des acides aminés des boucles. On se rappelle que ceci correspond à l'hypothèse

$$p_Z(z) = p_{Z_C}(z_C)p_{Z_L}(z_L) \tag{6.2}$$

de l'équation 5.2. On suppose également que les variables aléatoires associées aux boucles sont indépendantes et identiquement distribuées selon une loi de probabilité p_L, ce qui correspond à l'hypothèse

$$P_{Z_L}(z_L) = \prod_{i \in L} p_L(z_i). \tag{6.3}$$

de l'équation 5.3. La différence entre le modèle de White et le modèle proposé réside donc dans la modélisation de la loi de probabilité de Z_C.

6.1.1 Modélisation de p_{Z_C} à l'aide d'un réseau de Bayes

Dans cette section, notre objectif est de définir une factorisation acceptable de p_{Z_C} en construisant un GID (plutôt qu'un GIND) pour Z_C à partir du graphe de contact $G = (V_C, E)$.

Définissons $\pi(Z_C)$, un ordre sur Z_C (i.e. une permutation des $Z_i \in Z_C$). On dira que $Z_i < Z_j$ si et seulement si Z_i est choisi avant Z_j dans $\pi(Z_C)$. En utilisant la règle de conditionnement, on obtient

$$P(Z_C = z_C) = \prod_{i \in C} P(Z_i = z_i \mid A(Z_i) = a), \text{ où } A(Z_i) \equiv \{Z_j \mid Z_j < Z_i\}. \tag{6.4}$$

Comme dans la section 5.3, *on peut supposer que la présence d'un acide aminé à une position donnée ne dépend que des acides aminés occupant des positions voisines dans*

le graphe de contact. Plus précisément, si on définit N_i^- comme l'ensemble des Z_j tels que $Z_j < Z_i$ et $(V_i, V_j) \in E$, et si on assume des hypothèses d'indépendance de forme $Z_i \perp Z_j \mid N_i^-$ dès que $Z_j < Z_i$ et $Z_j \notin N_i^-$, on obtient l'expression

$$P(Z_C = z_C) = \prod_{i \in C} P(Z_i = z_i \mid N_i^- = n_i^-) \tag{6.5}$$

pour la probabilité $P(Z_C = z_C)$. L'expression correspondante pour p_{Z_C} est donnée par

$$p_{Z_C}(z_C) = \prod_{i \in C} p_{Z_i|N_i^-}(z_i \mid n_i^-). \tag{6.6}$$

Ceci correspond à l'hypothèse que *le graphe dirigé obtenu de G par ajout d'une direction aux arcs de G selon l'ordre induit par $\pi(Z_C)$ définit un GID pour Z_C* (on remplace tout arc $(V_i, V_j) \in E$ par l'arc dirigé $(V_i \rightarrow V_j)$ si et seulement si $Z_i < Z_j$). Ceci est illustré dans la figure 6.1.

En plus de n'imposer aucune restriction sur la définition des graphes de contact, cette factorisation s'avère plus simple qu'une décomposition en cliques et est automatiquement normalisée. De plus, on élimine une part d'arbitraire inhérente au design des fonctions de clique. Dans le cas particulier d'un graphe de contact triangulé, il est possible d'obtenir un GID qui possède les mêmes propriétés de Markov que le GIND correspondant. Dans les autres cas, le GID obtenu a des propriétés de Markov différentes mais raisonnables.

6.1.2 Choix d'une permutation

Le choix d'une permutation détermine une factorisation de p_{Z_C} en un produit de lois de probabilité locales (équation 6.6). Comme les interactions entre acides aminés sont symétriques d'un point de vue probabiliste, on peut supposer que toutes les permutations $\pi(Z_C)$ des $Z_i \in Z_C$ sont a priori équivalentes. Cependant, les lois locales $p_{Z_i|N_i^-}$ devront être estimées à partir des données et un choix facilitant le processus d'apprentissage est souhaitable. L'objectif est ici de réduire la dimensionalité individuelle des problèmes d'apprentissage des $p_{Z_i|N_i^-}$. Un choix intéressant en ce sens

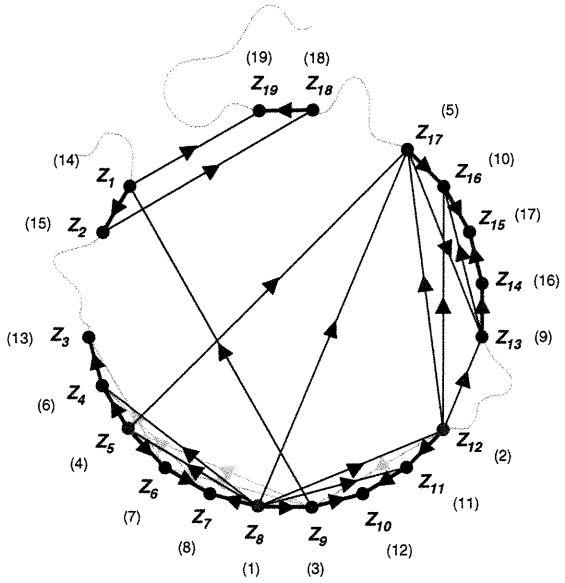

FIG. 6.1 – Approximation à l'aide d'un GID de la loi de probabilité des séquences d'acides aminés à l'intérieur du noyau structural de la protéine CRAMBIN. Ce GID est défini à partir du graphe de contact de la figure 5.1. Le sommet Z_i est une variable aléatoire associée à la position V_i. Les chiffres en parenthèses correspondent à la permutation $\pi(Z_C)$ calculée à l'aide de l'algorithme 6.1 (voir plus loin). Le voisinage N_{17}^- est indiqué en rouge. Ce GID n'est pas équivalent (au sens probabiliste) au GIND de la figure 5.4.

correspond à une permutation qui minimise la cardinalité maximale des *voisinages* N_i^-.

Pour les séquences d'acides aminés, un choix naturel pourrait être celui de la séquence. En pratique cependant, la distribution de la cardinalité des N_i^- obtenue avec cette permutation est comparable à celle qu'on obtient à l'aide d'une permutation choisie au hasard (fig. 6.2). Intuitivement, on préférerait une distribution moins étalée, particulièrement du côté des grandes cardinalités — mieux vaut une majorité de voisinages de cardinalité moyenne que quelques voisinages de trop grande cardinalité.

Algorithme 6.1

- *Soit $G = (V, E)$, le graphe de contact pour un motif structural.*
- *On définit W, l'ensemble (initialement vide) des sommets $V_i \in V$ déjà sélectionnés.*
- *On définit $d^-(i)$, le nombre d'arcs qui arrivent à V_i en provenance des sommets $V_j \in W$.*
- *On définit $d^+(i)$, le nombre d'arcs qui quittent V_i en direction des sommets $V_j \in V - W$.*
- *On répète jusqu'à ce que tous les sommets de V soient choisis :*
  - *On calcule $d = \max_{k \in V-W} d^-(k)$.*
  - *Parmi tous les sommets V_k tels que $d^-(k) = d$, on sélectionne V_j tel que $d^+(j) = \max_k d^+(k)$.*
  - *On met W à jour.*

Peut-on facilement obtenir une permutation satisfaisant cette condition ? On s'intéresse ici aux permutations de n objets, avec n variant de quelques dizaines à quelques centaines, et il est hors de question d'évaluer les $n!$ possibilités. Une "bonne" permutation peut être construite en choisissant itérativement les sommets du graphe de contact $G = (V_C, E)$ à l'aide de l'algorithme 6.1. Il s'agit d'une heuristique vorace qui s'éxécute en temps $O(n^2)$. La permutation $\pi(Z_C)$ retenue correspond à l'ordre de sélection des sommets de G. Ce choix est comparé à un choix arbitraire dans la figure 6.2.

Notons que la discussion ci-dessus s'applique également dans d'autres contextes où les réseaux de Bayes sont utilisés pour factoriser des lois de probabilité en haute dimension. Dans [80], on utilise un ordre arbitraire. Il serait intéressant de mesurer l'effet d'une heuristique telle l'algorithme 6.1 sur des problèmes de référence. Dans le chapitre 7, nous tâcherons d'en quantifier l'effet sur le problème d'estimation de densité étudié dans ce mémoire.

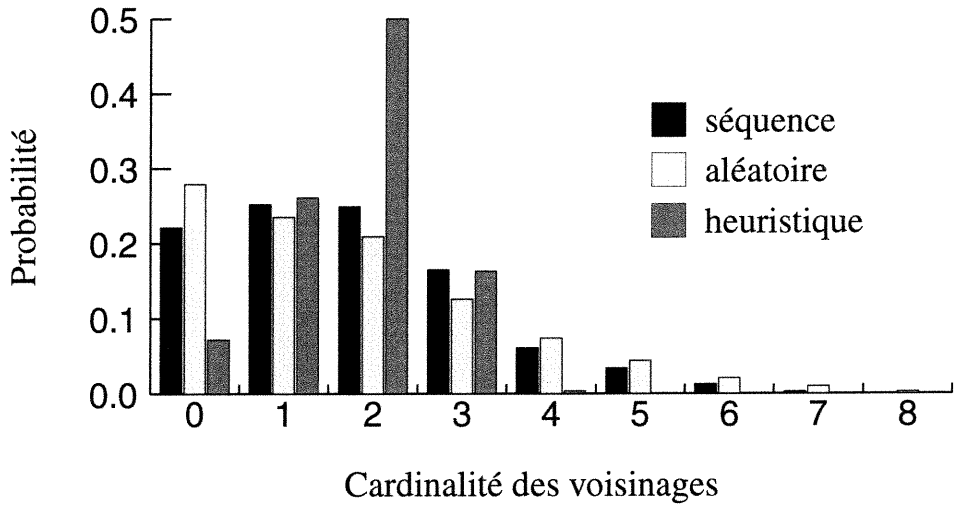

FIG. 6.2 – Distribution de la cardinalité des voisinages $N^-(Z_i^{M_j})$ obtenus pour différentes factorisations des lois de probabilité $p_{Z_C^{M_j}}$ et estimée sur un ensemble $M \equiv \{M_j\}$ de 57 motifs structuraux. Les distributions obtenues en utilisant l'ordre de la séquence, une permutation aléatoire, et la permutation choisie par l'algorithme 6.1 sont respectivement en noir, gris clair et gris.

6.2 Paramétrisation du modèle proposé

Dans la section précédente, on a jeté les bases d'un nouveau type de modèle pour la loi de probabilité de Z_C. Ce modèle repose sur une factorisation de p_{Z_C} dérivée à partir d'un graphe de contact représentant le noyau structural conservé. La factorisation proposée est un produit de lois de probabilité conditionnelles locales de forme

$$p_{Z_C}(z_C) = \prod_i p_{Z_i|N_i^-}(z_i \mid n_i^-). \tag{6.7}$$

Les lois de probabilité locales déterminent p_{Z_C} de façon unique et le but de cette section est de définir une forme paramétrique pour $p_{Z_i|N_i^-}$. L'approche proposée s'inspire de [80] et fait usage de réseaux de neurones artificiels (ANN) d'estimation de densité. Ceux-ci définissent une forme paramétrique pour les lois locales et la struc-

ture globale résultante (BN+ANN) est un *réseau bayesien paramétrique.*

6.2.1 Définition d'une forme paramétrique pour $p_{Z_i|N_i^-}$

La loi de probabilité locale $p_{Z_i|N_i^-}$ peut être approximée à l'aide d'un *multi-layer perceptron network* (MLP) représentant la fonction à valeur vectorielle

$$\begin{aligned} f_i(w_i, n_i^-) &= (f_i^{(1)}(w_i, n_i^-), \ldots, f_i^{(20)}(w_i, n_i^-)) \\ &\approx (P(Z_i = a_1 \mid N_i^- = n_i^-), \ldots, P(Z_i = a_{20} \mid N_i^- = n_i^-)). \end{aligned} \quad (6.8)$$

Ici, w_i correspond aux poids du MLP, $f_i^{(x)}(w_i, n_i^-)$ représente la probabilité conditionnelle $P(Z_i = a_x \mid N_i^- = n_i^-)$, et a_x désigne le $x$ième acide aminé (voir le tableau 2.1). Pour contraindre les sorties du MLP à des valeurs positives de somme 1, il suffit d'utiliser une fonction d'activation de type *softmax* (eq. 4.7) à la couche de sortie. Ceci s'inspire de [80, 77] et est illustré dans la figure 6.3.

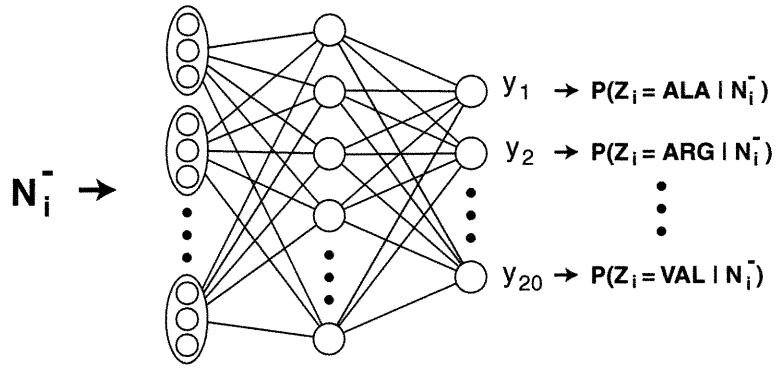

FIG. 6.3 – Paramétrisation des lois de probabilité locales à l'aide de MLP à une couche cachée. On utilise un groupe d'entrées pour l'encodage de chacun des $Z_j \in N_i^-$. Les sorties représentent les paramètres de la loi de probabilité conditionnelle $p_{Z_i|N_i^-}$.

Les MLP définissent une forme paramétrique pour les lois de probabilité locales et la structure globale résultante (fig. 6.4) est un *réseau bayesien paramétrique* [64] de paramètres $w \equiv \cup_i w_i$.

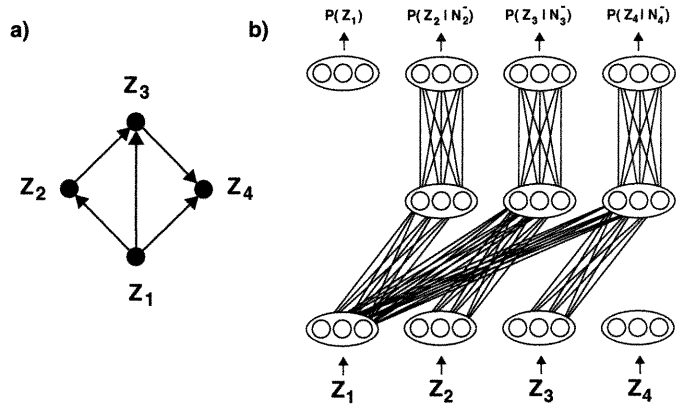

FIG. 6.4 – a) Représentation graphique des dépendances entre les variables aléatoires Z_1, Z_2, Z_3 et Z_4. La probabilité conjointe $P(z_C)$ est donnée par $P(z_C) = P(z_1, z_2, z_3, z_4) = P(z_4 \mid z_3, z_1) \times P(z_3 \mid z_2, z_1) \times P(z_2 \mid z_1) \times P(z_1)$. b) Assemblage de MLP permettant le calcul de la probabilité $P(Z_C = z_C)$ (inspiré de [80]). Les valeurs des Z_i sont encodées dans les unités d'entrée.

6.2.2 Optimisation des paramètres

Soit $s = \{s^{(1)}, ..., s^{(m)}\}$ un ensemble de séquences adoptant un même motif de repliement. Notre objectif est de construire un réseau bayesien paramétrique de paramètres w pour représenter les régions $s_C = \{s_C^{(1)}, ..., s_C^{(m)}\}$ de ces séquences qui correspondent au noyau structural conservé. Dans un cadre bayesien, les paramètres optimaux $w^\star$ sont ceux qui maximisent la probabilité

$$P(w \mid s_C) = \frac{P(s_C \mid w)P(w)}{P(s_C)}. \tag{6.9}$$

En supposant les $s_C^{(i)}$ tirées indépendamment d'une même distribution (la loi de probabilité de Z_C), l'équation 6.9 devient

$$\begin{aligned} P(w \mid s_C) &= \frac{P(w)}{P(s_C)} \prod_i P(Z_C = s_C^{(i)} \mid w) \\ &= \frac{P(w)}{P(s_C)} \prod_i \prod_{j \in C} P(Z_j = s_j^{(i)} \mid n_j^-, w_j). \end{aligned} \tag{6.10}$$

Maximiser cette expression est équivalent à minimiser

$$\begin{aligned} -\log P(w \mid s_C) &= -\sum_i \sum_{j \in C} \log P(Z_j = s_j^{(i)} \mid n_j^-, w_j) \\ &\quad - \log P(w) + \log P(s_C). \end{aligned} \tag{6.11}$$

Le terme $P(s_C)$ ne dépend pas de w et on peut l'ignorer pour l'optimisation. Le terme $P(w)$ est un *régulariseur* permettant de contraindre les paramètres *a priori*. Si $P(w)$ est uniforme sur l'espace des poids, le problème d'optimisation se réduit à minimiser la *log-vraisemblance*

$$\mathcal{L} = -\sum_i \sum_{j \in C} \log P(Z_j = s_j^{(i)} \mid n_j^-, w_j). \tag{6.12}$$

On peut intervertir les deux sommations dans l'équation 6.12 et on remarque que les termes correspondant à chaque loi de probabilité locale peuvent être minimisés indépendamment. On peut donc optimiser indépendamment les MLP (possiblement en parallèle) ; il suffit d'utiliser la log-vraisemblance locale

$$\mathcal{L}_j = -\log P(z_j \mid n_j^-, w_j) \tag{6.13}$$

comme mesure d'erreur pour l'optimisation du MLP représentant $p_{Z_j|N_j^-}$. L'erreur effectuée par celui-ci sur l'exemple $(Z_j = x, N_j^- = n_j^-)$ est alors donnée par

$$\mathcal{L}_j(x, n_j^-) = -\log f_j^{(x)}(w_j, n_j^-). \tag{6.14}$$

6.2.3 Dépendances d'ordre supérieur

A priori, les conflits stériques, les contraintes volumiques et les interactions non-locales qui surgissent suite à l'entassement compact des acides aminés à l'intérieur du noyau des protéines globulaires suggèrent que les interactions d'ordre supérieur sont courantes dans les protéines [7, 8].

La structure décrite ci-dessus peut efficacement capturer des dépendances d'ordre supérieur entre les Z_i. Les dépendances importantes sont "choisies" d'après les données

pendant l'apprentissage et c'est la capacité individuelle des MLP (déterminée par le nombre d'unités cachées) qui décide du nombre de dépendances pouvant être capturées. [1]

Par opposition, on a déjà discuté que le modèle stochastique de White *et al* [6] et les potentiels pseudo-énergétiques traditionnels (section 3.2) ne peuvent capturer que des dépendances d'ordre 2 ou moins.

6.2.4 Paramètres libres

Si toutes les lois de probabilité locales sont estimées indépendamment, le nombre total de paramètres libres pour la structure globale est $O(nmh)$, où n est la cardinalité de Z, $m \equiv max_i|N_i^-|$ est la cardinalité maximale des N_i^-, et $h \equiv max_i h_i$ est le nombre maximal d'unités cachées calculé sur tous les MLP. À titre de comparaison, une approximation polynômiale [81, 64] d'ordre k pour p_{Z_C} comporte $O(n^k)$ paramètres libres (voir [80] pour une discussion plus élaborée).

6.3 Partage de paramètres entre les lois de probabilité locales

En théorie, on peut apprendre indépendamment chacune des lois de probabilité locales, pour chacun des modèle stochastique qu'on souhaite définir. Cependant, la plupart des motifs de repliement d'intérêt n'ont pas un nombre suffisant de représentants dans les banques de données de structures pour que ceci soit possible. De plus, les représentants connus d'un motif donné ne constituent pas un échantillon uniforme de l'espace des séquences partageant ce motif. Les seuls représentants connus d'un motif structural sont souvent les membres d'une même famille de séquences fortement apparentées. Or, le but poursuivi ici n'est pas l'obtention d'une mesure d'affinité structurale fortement biaisée vers cette famille. On recherche plutôt une mesure capable

[1] Dans le cas d'un MLP sans couche cachée, seule des dépendances d'ordre 2 et moins peuvent être capturées.

d'identifier des homologues lointains et des analogues structuraux.

Pour contourner cette difficulté, *on supposera que les paramètres des lois locales peuvent être partagés entre des environnements structuraux équivalents.* Plus précisément, si les environnements structuraux en position i d'un noyau structural conservé A et en position j d'un noyau structural conservé B sont équivalents, on supposera que

$$p_{Z_i^A|N^-(Z_i^A)}(x \mid y) = p_{Z_j^B|N^-(Z_j^B)}(x \mid y). \tag{6.15}$$

Les lois de probabilité $p_{Z_i^A|N^-(Z_i^A)}$ et $p_{Z_j^B|N^-(Z_j^B)}$ peuvent alors être représentées à l'aide d'un même MLP construit comme dans la section 6.2.

6.3.1 Définition d'environnements structuraux équivalents

Dans ce travail, les noyaux structuraux conservés sont représentés à l'aide de graphes de contact. Sous cette représentation, l'environnement structural local autour d'une position i est encodé par le sous-graphe sous-tendu par le sommet V_i et ses voisins $N(V_i)$ dans le graphe de contact. *Des environnements structuraux sont donc considérés équivalents si et seulement si leurs sous-graphes associés sont isomorphes et les propriétés des sommets et des arcs correspondants sont les mêmes* (voir fig. 6.5).

Ainsi, pour la suite de ce travail, on supposera que $Z_i \mid N_i^-$ et $Z_j \mid N_j^-$ sont identiquement distribuées dès que la structure locale du graphe de contact est similaire autour des position i et j.

6.3.2 Modification du modèle probabiliste global

Il est plus facile d'intégrer l'hypothèse de l'équation 6.15 en représentant explicitement l'environnement structural local. Pour ce faire, il suffit d'introduire de nouvelles variables aléatoires dans le modèle probabiliste global : à chaque position V_i du noyau structural conservé, on associe une nouvelle variable aléatoire G_i résumant l'environ-

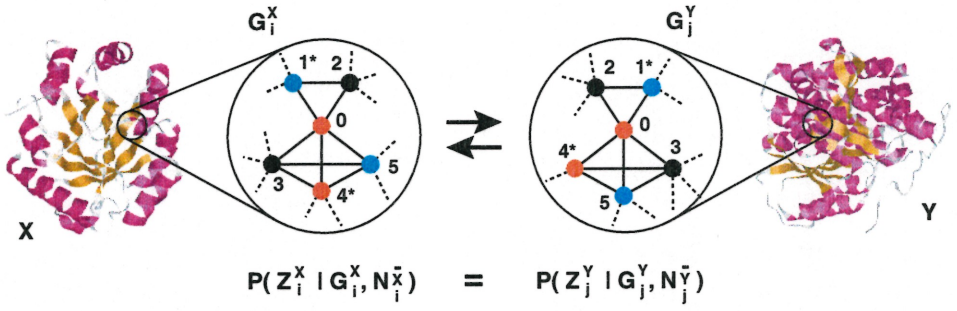

FIG. 6.5 – Partage des paramètres des lois de probabilité locales entre deux environnements structuraux (ES) équivalents. Des ES sont considérés équivalents ssi leurs sous-graphes associés sont isomorphes et les propriétés des sommets et des arcs (représentées par leur couleur) correspondants sont les mêmes. Dans l'illustration, les sommets correspondant ont le même numéro et les sommets étiquetés par une étoile définissent les ensembles N_i^- et N_j^-.

nement structural à cette position et influençant l'assignation d'acides aminés à cette position (voir la fig. 6.6). En accord avec la discussion de la section précédente, la valeur de G_i correspond à la structure graphique locale autour de la position i. Cette formulation rend explicite l'influence de l'environnement structural local sur la compatibilité des acides aminés pour les différentes positions structurales.

Avant de continuer, il est utile de définir $G \equiv \{G_1, \ldots, G_n\}$. On peut interpréter G comme une variable aléatoire représentant la structure du noyau structural conservé. Il est important de noter que dans ce mémoire, on s'intéresse à $p_{Z_C|G}$, la loi de probabilité de Z_C lorsque la structure du noyau structural conservé est fixée. Lorsqu'on utilise le modèle proposé à la section 6.1.1, la probabilité conditionnelle $P(Z_C = z_C \mid G = g)$ est donnée par

$$\begin{aligned} P(z_C \mid g) &= P(z_1, \ldots, z_n \mid g_1, \ldots, g_n) \\ &= \prod_{i \in C} P(z_i \mid g_i, n_i^-). \end{aligned} \tag{6.16}$$

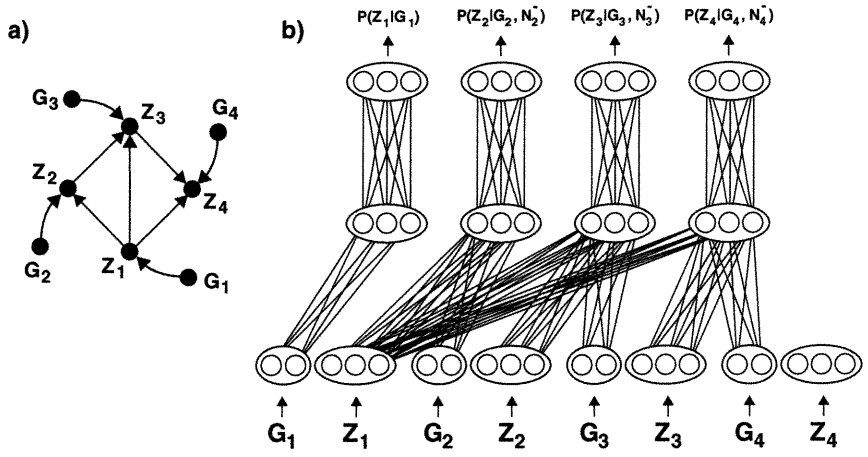

FIG. 6.6 – a) Ajout de variables aléatoires environnementales dans le graphe d'indépendance de la fig. 6.4a. La probabilité conjointe conditionnelle est donnée par $P(z_C \mid g) = P(z_4 \mid g_4, z_3, z_1) \times P(z_3 \mid g_3, z_2, z_1) \times P(z_2 \mid g_2, z_1) \times P(z_1 \mid g_1)$. b) Assemblage de MLP permettant le calcul de la probabilité $P(Z_C = z_C \mid G = g)$.

Les lois de probabilité locales d'intérêt sont alors de forme $p_{Z_i|G_i,N_i^-}$ et elles peuvent toutes être approximées à l'aide d'un unique MLP construit comme dans la figure 6.3 mais utilisant un groupe d'entrées supplémentaires pour l'encodage de G_i (voir la section 6.4).

Lorsqu'on dispose d'un modèle pour $p_{Z_C|G}$, on peut facilement obtenir un modèle pour p_{Z_C} en remarquant que

$$p_{Z_C}(z_C) = \sum_g p_{Z_C|G}(z_c \mid g) p_G(g). \tag{6.17}$$

Dans ce mémoire, les motifs structuraux sont représentés d'une manière déterministe et "rigide" à l'aide de graphes de contact. Ceci est équivalent à limiter G à une valeur unique g^0 et dans ce cas, on a toujours

$$\begin{aligned} p_{Z_C}(z_C) &= \sum_g p_{Z_C|G}(z_c \mid g) p_G(g) \\ &= p_{Z_C|G}(z_C, g^0). \end{aligned} \tag{6.18}$$

Cependant, en accord avec la discussion de la section 3.2.5, l'intégration au présent travail d'un modèle stochastique pour G capable d'encoder les variations structurales observées entre les protéines homologues constitue une avenue de recherche à privilégier.

6.4 Architecture détaillée des MLP

Tel qu'expliqué dans la section précédente, les lois de probabilité locales qui nous intéressent maintenant sont de forme $p_{Z_i|G_i,N_i^-}$, où la valeur de la variable aléatoire G_i correspond à la la structure locale du graphe de contact autour de la position i.

On a déjà mentionné que les $p_{(Z_i|G_i,N_i^-}$ peuvent être approximées à l'aide de MLP construits comme dans la figure 6.3 mais avec un groupe d'entrées supplémentaires affecté à l'encodage de G_i. En fait, toutes les lois locales, dans tous les environnements structuraux, pourraient être représentées à l'aide d'un unique MLP. Pour des raisons pratiques cependant, *on représentera les lois locales* $p_{Z_i|G_i,N_i^-}$ *et* $p_{Z_j \mid G_j,N_j^-}$ *par le même MLP si et seulement si* G_i *et* G_j *sont isomorphes et chaque acide aminé dans* N_i^- *correspond à un acide aminé dans* N_j^- *(et inversement).* Dans ce cas, G_i et G_j contiennent un même nombre de sommets et d'arcs et les ensembles N_i^- et N_j^- sont de même taille. On requiert donc, en plus d'un groupe d'entrées pour chaque acide aminé dans N_i^- et N_j^-, un groupe d'entrées pour chaque propriété de chaque sommet et de chaque arc de G_i et G_j. *L'important est d'utiliser les mêmes unités d'entrée pour l'encodage des propriétés de sommets et d'arcs correspondants de* G_i *et* G_j (il en va de même pour les acides aminés correspondants de N_i^- et N_j^-). Ceci est illustré dans la figure 6.7.

La stratégie ci-dessus permet de spécialiser les MLP en fonction de la densité et de la taille des environnements structuraux considérés. Par la même occasion, on contourne le problème du débalancement des exemples entre des environnements de complexité différente (voir la section 6.4.3).

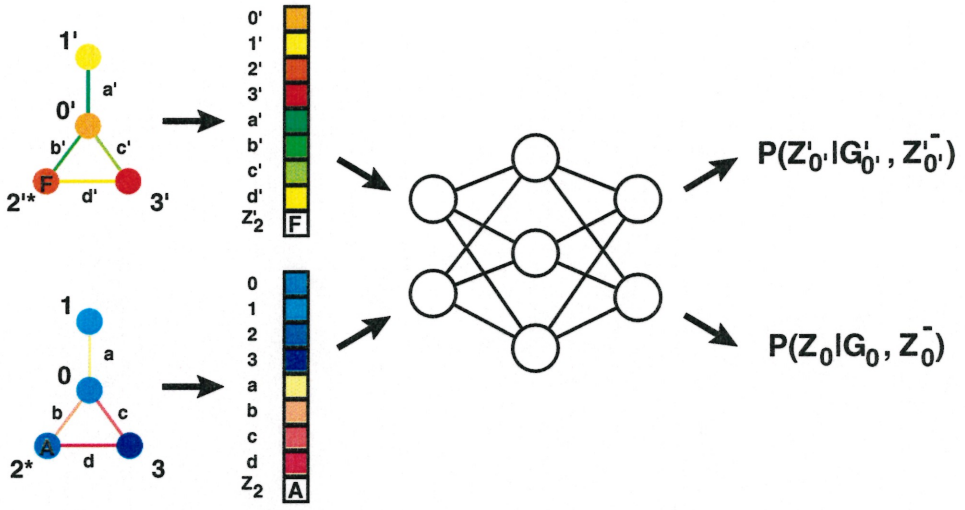

FIG. 6.7 – Représentation de $p_{Z_i|G_i,N_i^-}$ et $p_{Z_j|G_j,N_j^-}$ à l'aide du même MLP. G_i et G_j sont isomorphes et chaque acide aminé dans N_i^- correspond à un acide aminé dans N_j^- (et inversement). Les propriétés des sommets et des arcs (représentées par leur couleur) correspondants sont encodées dans les mêmes unités d'entrées. Les sommets étiquetés par une étoile définissent les ensembles N_i^- et N_j^-.

6.4.1 Encodage des acides aminés

L'encodage le plus simple pour des acides aminés (AA) est l'encodage dit *orthogonal*. Dans ce cas, chaque AA est encodé sur un vecteur de 20 bits (un bit pour chaque type d'AA). Tous prennent la valeur 0, sauf le bit correspondant au type de l'AA encodé, qui prend la valeur 1.

Riss et Krogh [78] ont proposé un encodage beaucoup plus compact basé sur la projection des codes orthogonaux dans un espace continu de plus faible dimension, par exemple le cube $[-1, 1]^3$. Dans ce dernier cas, on ne requiert que 3 neurones en entrée (plutôt que 20) pour l'encodage de chaque AA. La projection optimale peut être apprise en même temps que les paramètres du MLP. Il suffit d'en modifier l'architecture en introduisant, pour chaque AA en entrée, un nouveau module de

calcul formé de 20 neurones affectés à l'encodage orthogonal de cet AA, et connectés uniquement aux neurones assurant son encodage compact. Ces modules calculent les projections individuelles des codes orthogonaux sur $[-1, 1]^3$. On s'assure d'utiliser la même projection pour chaque AA en partageant les poids entre les modules. Le reste de l'architecture demeure inchangé. Pour plus de détails, le lecteur est invité à consulter [78] ou la section 6.2.3 du livre de Baldi et Brunak [62].

En plus de réduire de façon importante le nombre de paramètres effectifs des MLP, cette technique a pour effet de projeter le problème considéré dans un espace où sa solution est plus évidente (en regroupant adaptativement les différents types d'AA en fonction de propriétés communes pertinentes). Pour la prédiction de la structure secondaire des protéines, on pourrait par exemple supposer que les bons "formeurs" d'hélice α seront projetés dans une même région de $[-1, 1]^3$.

En combinaison avec d'autres artifices, l'approche ci-dessus a permis d'améliorer la prédiction de la structure secondaire des protéines à l'aide de MLP [78]. Pour ce problème, le choix d'un espace à trois dimensions pour la projection s'est avéré optimal.

Dans les expériences présentées au chapitre 7, les deux types d'encodage – orthogonal et adaptatif – seront comparés.

6.4.2 Encodage des propriétés environnementales

Dans ce travail, on utilisera un encodage orthogonal pour les propriétés environnementales discrètes. Cependant, contrairement à l'estimation de densité avec une méthode simple comme les tables des fréquences d'occurence (TFO) du modèle de White (section 5.3), les propriétés environnementales continues (e.g. l'indice d'exposition au solvant ou la distance entre des acides aminés) n'ont pas à être discrétisés.

6.4.3 Représentation simplifiée des environnements structuraux locaux

En pratique, le nombre espéré d'exemples d'un environnement structural dans les banques de structures devient insignifiant lorsque sa complexité (proportionnelle au nombre de sommets et au nombre d'arcs du sous-graphe correspondant) augmente. Les contraintes volumiques à l'intérieur des protéines limitent le nombre et la proximité des acides aminés autour d'une position et expliquent en partie cette observation. [2] Par ailleurs, pour des graphes de taille k, le nombre de classes d'isomorphisme croît exponentiellement avec k et on doit s'attendre à ce que le nombre d'exemples par classe diminue rapidement avec k.

En conséquence, une représentation simplifiée pourrait être préférable pour les environnements structuraux complexes. Ceci correspondrait à assumer le partage des paramètres des lois de probabilité locales entre un plus grand nombre d'environnements. Différentes alternatives (e.g. le partage des paramètres entre tous les environnements de même taille) seront évaluées dans le chapitre suivant.

[2] Cet argument suggère que les distributions d'acides aminés sont fortement contraintes – donc plus "faciles" à apprendre – dans les environnements structuraux denses.

Chapitre 7

ÉVALUATION DU MODÈLE STOCHASTIQUE PROPOSÉ

Dans le chapitre 6, on a défini un nouveau type de modèle stochastique pour les séquences d'acides aminés partageant un même motif architecture. On a vu que la vraisemblance d'une séquence dans un tel modèle peut être interprétée comme une mesure de compatiblité séquence–structure dans les protéines.

L'objectif premier de ce chapitre est d'évaluer le modèle proposé et de s'assurer de sa pertinence et de son utilité en regard des données biologiques disponibles. Le modèle proposé sera comparé au modèle de White *et al* [6]. L'approche de White est représentative de celles qui sont généralement employées pour la reconnaissance de la conformation des protéines, en particulier à l'aide d'alignements séquence–structure.

Ce chapitre est organisé comme suit. Les données utilisées pour la validation et l'entraînement ainsi que les mesures de comparaison employées sont d'abord présentées dans la section 7.1. Les expériences sont présentées dans la section 7.2. Le modèle proposé — tel que défini dans le chapitre 6 — est évalué dans la section 7.2.1. Différentes variantes sont évaluées dans la section 7.2.2. Ensuite, dans la section 7.2.3, on évalue l'importance des différents types d'interactions entre acides aminés qui sont prises en compte dans les mesures d'affinité séquence–structure. Dans les sections 7.2.4 et 7.2.5, on évalue respectivement l'utilité de l'encodage adaptatif des acides aminés (section 6.4.1) et de la procédure d'optimisation des graphes d'indépendance (section 6.1.2). Ensuite, dans la section 7.2.6, on observe le comportement du modèle proposé lorsqu'on raffine les patrons structuraux en ajoutant des attributs environnementaux aux sommets et aux arcs des graphes de contact. Finalement, dans la section 7.2.7, on évalue l'influence de la "mémoire de séquence" sur le modèle proposé et on démontre son utilité dans une situation proche d'un véritable problème de prédiction.

7.1 Préliminaires

Le rôle de cette section est de clarifier le cadre expérimental adopté dans ce chapitre. Les questions discutées ici sont le choix des données (section 7.1.1), la construction de patrons structuraux à partir des données (section 7.1.2), la construction et l'entraînement de modèles stochastiques à partir des patrons structuraux (section 7.1.3) ainsi que le choix de mesures pour la comparaison des modèles stochastiques (section 7.1.4).

7.1.1 Les données

Les données choisies pour la validation sont tirées de la banque de domaines SCOP–1.50 (voir [20] et la section 2.4). SCOP–1.50 contient les séquences de 24186 domaines regroupés en 7 classes structurales, 548 motifs de repliement (folds), 820 super-familles et 1296 familles.

SCOP ne constitue pas un échantillon uniforme de l'espace des séquences de protéines. Certains motifs de repliement auxquels appartiennent des familles très étudiées sont sur-représentés (e.g. les globines). Plus grave, les seuls représentants connus d'un motif de repliement sont souvent les membres d'une unique famille de séquences très similaires. On peut obtenir un échantillon plus uniforme en se restreignant à un sous-ensemble de SCOP ne contenant que des séquences peu ou pas apparentées. Dans ce travail, *on utilisera un sous-ensemble de 2873 séquences obtenues par filtrage à 40% d'identité des séquences de* SCOP–1.50 *à l'aide de l'algorithme* ASTRAL [82]. [1] Les structures 3-D correspondantes sont obtenues de la Protein Data Bank [16].

Les 2873 paires séquence–structure ci-dessus sont ensuite regroupées en deux ensembles : un ensemble de validation, contenant 56 protéines, et un ensemble d'en-

[1] Il s'agit d'une forme de *clustering* utilisant la similarité de séquence (section 3.1.1) comme métrique. Des séquences représentatives de chaque *cluster* sont ensuite sélectionnées en fonction de la qualité du meilleur modèle disponible pour leur structure 3-D.

traînement, contenant 2817 protéines. Le choix de l'ensemble de validation est inspiré de [44] et se veut représentatif de la diversité structurale observée dans les protéines. Les séquences retenues sont principalement tirées de protéines qui ne comptent qu'un seul domaine et qui ne forment pas de multimères. Les protéines de l'ensemble de validation sont énumérées dans le tableau 7.1.

7.1.2 Annotation des structures 3-D et construction des patrons structuraux

Pour chacun des 2873 domaines retenus dans la section 7.1.1, on construit un patron structural à partir de l'analyse de sa structure tridimensionnelle détaillée. Les patrons structuraux sont définis à l'aide de graphes de contact comme dans la section 5.2. La longueur des boucles est fixe et correspond aux valeurs observées dans les structures 3-D. [2]

Cette section présente les détails de la procédure d'annotation (identification des contacts et valeur des attributs environnementaux). Les attributs environnementaux considérés dans ce mémoire sont l'exposition au solvant et le type de structure secondaire.

Les structures 3-D sont analysées avec le programme DSSP [83]. Les éléments de structure secondaire (hélices–α et brins–β) son identifiées avec ce programme.

L'exposition au solvant à chaque position est définie comme la surface accessible au solvant de l'acide aminé occupant cette position (calculée par DSSP) normalisée par sa surface totale. Cette valeur normalisée est appelée *indice d'accessiblité au solvant* (IAS).

Deux acides aminés appartenant à une même hélice–α ou à un même feuillet–

[2] On a vu que l'utilisation de boucles de longueur variable permet d'accomoder des séquences de longueur variable dans un unique patron représentant leur motif de repliement commun. Ceci ne sera pas nécessaire ici dans la mesure où on ne considèrera qu'une seule séquence par patron et qu'on connaît *a priori* la configuration optimale de chaque séquence dans le patron correspondant.

β sont considérés en contact si la distance entre leurs atomes C_β (voir fig. 2.1) est inférieure à 6.5Å et si l'angle entre leurs liens $C_\alpha - C_\beta$ est dans l'intervalle $[-90°, 90°]$ (i.e. si leurs chaînes latérales sont parallèles). Deux acides aminés appartenant à deux éléments de structure secondaire différents (sauf deux brins-β d'un même feuillet) sont considérés en contact si la distance entre leurs atomes C_β est inférieure à 6.5Å et si l'angle entre leurs liens $C_\alpha - C_\beta$ est dans l'intervalle $[90°, 270°]$ (i.e. si leurs chaînes latérales se font face).

7.1.3 Construction et entraînement des modèles stochastiques

À partir des 56 graphes de contacts extraits de l'ensemble de validation (section 7.1.2), on construit 56 modèles stochastiques représentant les séquences adoptant chaque motif. Les paramètres de ceux-ci sont ensuite estimés sur l'ensemble d'entraînement. Pour le modèle proposé, ceci correspond à l'estimation des paramètres des lois de probabilité locales. Par exemple, si on utilise un MLP pour la paramétrisation d'une loi de probabilité locale dans un certain environnement structural local (section 6.3), les exemples d'apprentissage correspondent à toutes les occurences de cet environnement dans l'ensemble d'entraînement.

Dans les expériences décrites plus loin, l'architecture des MLP est telle que définie dans les sections 6.2 et 6.4. On utilise une activation de type *tanh* pour les couches cachées et une activation de type $softmax$ pour la couche de sortie. Les poids incidents à chaque neurone sont initialisés aléatoirement dans l'intervalle $[-1/\sqrt{F}, 1/\sqrt{F}]$, où F est le nombre de connections incidentes à ce neurone. Les poids sont optimisés selon le principe du maximum de vraisemblance (section 6.2.2) et à l'aide de la méthode du gradient stochastique (section 4.3.5).

Les hyper-paramètres sont choisis par validation croisée avec $k = 5$ (section 4.3.6) sur l'ensemble d'entraînement. Le nombre d'unités cachées est choisi dans l'intervalle $[2, 20]$. Le taux d'apprentissage est choisi dans l'intervalle $[0.0001, 0.2]$. Le critère d'arrêt α est choisi dans l'intervalle $[2, 25]$. Les intervalles retenus pour chaque hyper-

paramètre ont été choisis par tâtonnement et de façon à ce que le temps de calcul requis par la procédure de validation croisée soit raisonnable. [3]

7.1.4 Évaluation et comparaison des modèles stochastiques

L'objectif des expériences présentées plus loin est d'évaluer différentes variantes du modèle stochastique proposé et de les comparer au MRF de White *et al* (voir [6] et la section 5.3). *La mesure de comparaison qui sera utilisée est la log-vraisemblance des séquences de l'ensemble de validation dans les modèles stochastiques construits à partir de leurs propres structures 3-D.* Pour une séquence s et un modèle stochastique M, cette quantité est donnée par

$$\mathcal{L}(s) \equiv \log P(s \mid M) \tag{7.1}$$

où $P(s \mid M)$ est la probabilité $P(Z = s)$ des chapitres 5 et 6 lorsqu'estimée à l'aide du modèle stochastique M. Cette mesure sera utilisée pour comparer entre elles différentes variantes du modèle proposé.

Dans le cas du MRF de White la probabilité $P(Z = s)$ est de forme

$$P(Z = s) = \frac{1}{\Lambda} Q(s) = \frac{Q(s)}{\sum_t Q(t)} , \tag{7.2}$$

où $Q(s)$ est obtenu en combinant les équations 5.2, 5.3 et 5.4, et la log-vraisemblance $\mathcal{L}(s)$ est donnée par

$$\mathcal{L}(s) = \log Q(s) - \log \Lambda . \tag{7.3}$$

Le calcul du terme de normalisation Λ requiert une somme sur toutes les séquences possibles. Il s'agit d'une tâche très coûteuse qu'on aimerait éviter.

[3] Il est important de noter que le nombre de valeurs explorées pour chaque hyper-paramètre influence drastiquement le temps de calcul requis pour la validation croisée. Le choix des valeurs à explorer devrait donc tenir compte de la puissance de calcul disponible. Pour ce travail, tous les calculs ont été effectués en parallèle sur un *cluster* de 20 ordinateurs Pentium II 400Mhz et un temps de calcul d'environ 2 heures par machine a été jugé raisonnable.

Considérons l'ensemble $\mathcal{S}$ des séquences de même longueur et de même composition en acides aminés que s. On peut raisonnablement supposer que la distribution de $\mathcal{L}(t)$, pour t tirée de $\mathcal{S}$, est approximativement normale de moyenne $\mu_{\mathcal{L}}$ et de variance $\sigma^2_{\mathcal{L}}$. [4]

Sachant qu'on peut obtenir un échantillon $\widehat{\mathcal{S}} = \{s_1, s_2, \ldots, s_n\} \subset \mathcal{S}$ en permutant s aléatoirement n fois, on peut aisément estimer $\mu_{\mathcal{L}}$ et $\sigma^2_{\mathcal{L}}$ à l'aide de la moyenne empirique

$$\widehat{\mu}_{\mathcal{L}} = \frac{1}{n}\sum_i \mathcal{L}(s_i) \tag{7.4}$$

et de la variance empirique

$$\widehat{\sigma}^2_{\mathcal{L}} = \frac{1}{n}\sum_i (\mathcal{L}(s_i) - \widehat{\mu}_{\mathcal{L}})^2 \tag{7.5}$$

calculées sur $\widehat{\mathcal{S}}$.

Dans ce cas, le *z-score*

$$\mathcal{Z}(s) = \frac{\mathcal{L}(s) - \mu_{\mathcal{L}}}{\sigma_{\mathcal{L}}} \approx \frac{\mathcal{L}(s) - \widehat{\mu}_{\mathcal{L}}}{\widehat{\sigma}_{\mathcal{L}}} \tag{7.6}$$

est une mesure normalisée de l'écart entre $\mathcal{L}(s)$ et la valeur espérée de $\mathcal{L}$ pour une séquence de même longueur et de même composition en acides aminés.

Contrairement à $\mathcal{L}(s)$, $\mathcal{Z}(s)$ peut facilement être utilisé comme mesure de comparaison entre le modèle proposé et le MRF de White. Dans le cas de ce dernier, le calcul de $\mathcal{Z}(s)$ ne requiert pas l'estimation du terme de normalisation Λ. En effet, en combinant 7.3 et 7.4, on obtient

$$\begin{aligned} \widehat{\mu}_{\mathcal{L}} &= \frac{1}{n}\sum_i \mathcal{L}(s_i) \\ &= \frac{1}{n}\sum_i (\log Q(s_i) - \log \Lambda) \\ &= \frac{1}{n}\sum_i \log Q(s_i) - \log \Lambda \ . \end{aligned} \tag{7.7}$$

[4] Cette hypothèse a été vérifiée empiriquement. Elle est cependant fausse si on permet des boucles de longueur variable dans les patrons structuraux (voir la discussion plus loin).

$\mathcal{Z}(s)$ est donc donné par

$$\mathcal{Z}(s) = \frac{\mathcal{L}(s) - \widehat{\mu}_{\mathcal{L}}}{\widehat{\sigma}_{\mathcal{L}}} = \frac{\log Q(s) - \frac{1}{n}\sum_i \log Q(s_i)}{\widehat{\sigma}_{\mathcal{L}}} , \tag{7.8}$$

où on a

$$\widehat{\sigma}^2_{\mathcal{L}} = \frac{1}{n}\sum_i (\log Q(s_i) - \frac{1}{n}\sum_i \log Q(s_i))^2 . \tag{7.9}$$

Discussion

Bien qu'on verra dans la section 7.2 que $\mathcal{L}$ et $\mathcal{Z}$ sont des mesures de comparaison généralement consistantes (i.e. un modèle stochastique déclaré supérieur selon $\mathcal{L}$ l'est aussi selon $\mathcal{Z}$, et inversement), il est important de noter qu'il ne s'agit pas de mesures équivalentes. Alors qu'une valeur élevée de $\mathcal{L}$ indique que le modèle considéré accorde une forte probabilité aux "bonnes" séquences, une valeur élevée de $\mathcal{Z}$ indique plutôt que ce modèle discrimine bien entre les "bonnes" et les "mauvaises" séquences.

Utilisés avec la distribution normale standard, les z-scores permettent de vérifier que l'écart entre la log-vraisemblance espérée pour une séquence d'une certaine composition et la valeur observée est bien significatif.

Il importe de mentionner que les z-scores (tels que définis par l'équation 7.6) ne sont appropriés que pour des patrons structuraux "rigides". On a vu que pour bien représenter toutes les séquences partageant un noyau structural commun, il est nécessaire d'inclure une certaine souplesse dans les patrons structuraux sous la forme de boucles de longueur variable (section 5.3.3). Dans ce cas, le calcul de la log-vraisemblance requiert une somme sur tous les alignements possibles dans le patron et on ne peut pas supposer que la log-vraisemblance des séquences permutées est normalement distribuée. Des travaux récents semblent indiquer qu'il s'agit plutôt d'une distribution de la valeur extrême [84]. Dans ce cas, les z-scores n'ont pas la forme de l'équation 7.6 et il est plus compliqué de vérifier que l'écart entre la log-vraisemblance espérée pour une séquence d'une certaine composition et la valeur observée est signi-

ficatif. [5]

7.2 Résultats

L'objectif des expériences présentées dans cette section est d'évaluer différentes variantes du modèle stochastique proposé et de les comparer au MRF de White *et al* (voir [6] et la section 5.3). On aimerait en particulier évaluer les différentes hypothèses de ce travail (hypothèses d'indépendance conditionnelles (section 6.1.1), équivalence et choix des permutations (6.1.2), importance des interactions d'ordre supérieur entre acides aminés (6.2.3), partage des paramètres des lois de probabilité locales entre des environnements structuraux locaux équivalents (6.3), représentation des environnements structuraux (6.3.1)) et voir l'influence des attributs environnementaux dans les graphes de contact.

7.2.1 Validation du modèle proposé au chapitre 6

Le but de l'expérience décrite dans cette section est de comparer le modèle proposé (tel que défini au chapitre 6) au MRF de White.

Il est important de noter que pour cette expérience, on n'a associé aucun attribut environnemental aux sommets et aux arcs des graphes de contact. Il n'y a donc qu'un seul type de sommet et un seul type d'arc dans les graphes de contact.

Cette expérience a nécessité la construction de 112 modèles stochastiques, deux pour chacun des 56 motifs de repliement de l'ensemble de validation. Les premiers ont été construits à l'aide de l'approche proposée au chapitre 6 (réseau de Bayes + MLP), alors que les deuxièmes ont été construits à l'aide de l'approche de White *et*

[5] Dans un problème de prédiction, cette distribution doit être estimée pour chaque paire séquence–patron considérée. Sachant qu'une approche empirique (e.g. par l'alignement de n séquences aléatoires dans chaque patron) exige des calculs considérables, une approche analytique à ce problème est requise. Une ébauche de solution en ce sens a été récemment proposée par Mirny *et al* [84].

protéine		BN+MLP		MRF	
id	(#AA)	nll	z-score	nll*	z-score
256ba	(106)	292.33	4.8684	294.79*	2.486
2end	(137)	393.34	3.2205	397.54*	2.1673
1rcb	(129)	361.83	5.8673	368.54*	3.1884
2mhr	(118)	341.08	4.1529	343.88*	3.1684
451c	(82)	226.14	4.5076	233.13*	2.5295
1bt0a	(73)	196.67	4.4637	198.63*	3.2799
1a6m	(151)	417.45	5.4879	426.14*	3.2454
2lisa	(131)	382.2	3.4829	387.73*	2.0659
1aep	(153)	418.24	5.8077	422.91*	2.7571
1hoe	(74)	211.63	2.8458	208.11*	3.172
1ptf	(87)	237.94	3.9714	238.9*	3.2784
1ycc	(108)	310.63	3.2397	316.47*	1.7358
1poa	(118)	351.61	3.9971	338.82*	4.7646
1aba	(87)	246.43	4.3243	250.56*	3.3194
1cewi	(108)	312.19	3.9948	301.5*	4.8721
2pvba	(107)	292.27	3.6725	294.56*	2.7111
1noa	(113)	305.88	5.2083	308.84*	3.5599
7fd1a	(106)	311.47	3.7193	309.94*	3.4761
1plc	(99)	269.49	5.3437	268.06*	4.7699
1b9oa	(123)	367.94	1.7658	365.2*	1.9979
1bkf	(107)	298.45	4.4894	298.28*	3.89
7rsa	(124)	361.16	4.762	363.26*	3.3295
5nul	(138)	382.11	6.5976	386.88*	4.6425
1sty	(137)	377.63	5.405	386.59*	3.818
1ifc	(131)	356.22	6.1582	364.65*	4.1749
119l	(162)	449.51	5.7683	455.8*	4.273
3chy	(128)	344.13	5.5641	343.5*	5.0091
1pkp_1	(71)	199.62	2.3077	200.22*	1.9229
2aak	(150)	424.82	6.2815	428.53*	5.4678
8dfr	(186)	524.05	5.91	525.98*	4.9954
2gar	(188)	583.85	5.9183	580.67*	5.4771
1qf9a	(194)	542.5	7.0514	550.4*	5.044
1rec	(185)	514.03	8.523	534.53*	4.9379
1ryc	(291)	820.83	7.1072	830.67*	5.3151
1gpr	(158)	437.35	4.7479	437.28*	4.193
1bfg	(126)	360.07	3.3232	358.15*	3.3546
1cnsa	(243)	690.73	6.0522	699.61*	4.1071
1ppn	(212)	603	6.0286	601.61*	4.9474
1hdr	(236)	647.77	8.3313	655.23*	5.9764
1c22a	(262)	741.06	6.7774	747.21*	4.9092
1avwb	(171)	486.64	4.1595	487.57*	3.5554
1elt	(236)	667.93	7.1107	665.2*	5.9869
2cba	(258)	722.83	7.315	728.98*	5.6657
2ayh	(214)	603.38	7.5503	607.45*	5.5433
1d6aa	(262)	732.38	7.485	742.43*	5.3145
3tgl	(265)	752.68	6.0145	764.22*	3.5749
8tlne1	(161)	445.92	5.6935	455.54*	3.3724
1led	(242)	674.78	6.602	677.62*	5.1395
1nar	(289)	812.71	7.7071	826.72*	5.2442
1gci	(269)	735.77	6.7067	735.08*	4.9955
2ctc	(307)	870.26	7.2463	876.49*	5.1102
1as4a1	(336)	920.46	8.9756	933.61*	6.4059
1phb	(405)	1149.9	7.9122	1164.5*	5.7272
7taa_2	(381)	1090.6	7.3316	1094.5*	6.0383
total		26572	298.9	26783*	224.0

TAB. 7.1 – Comparaison du modèle proposé (BN+MLP) au MRF de White. Les valeurs présentées sont la log-vraisemblance négative (nll) et le z-score obtenus par les 54 séquences de l'ensemble de validation dans le modèle stochastique associé à leur propre motif de repliement. Les protéines sont identifiées par leur code de la PDB. Le nombre d'acides aminés de chaque séquence est indiqué entre parenthèses. Une valeur plus faible du nll indique un meilleur modèle. Pour le MRF, la log-vraisemblance négative (nll*) ne correspond pas à une probabilité normalisée (voir le texte). Un valeur plus élevé du z-score indique un modèle plus discriminant. La probabilité que la différence observée des z-scores survienne par hasard en supposant que les deux modèles soient équivalents est inférieure à 10^{-9}.

al [6] décrite au chapitre 5 (MRF).

Dans le premier cas, 426 MLP ont été requis pour la paramétrisation des 56 modèles stochastiques. Les paramètres de ces 426 MLP ont été optimisés par validation croisée à l'aide de la procédure décrite à la section 7.1.3. Les 426 procédures d'optimisation ont été lancées en parallèle sur un *cluster* de 20 ordinateurs de type Pentium II 400Mhz. Ceci a nécessité environ 1.8 heures de calcul par machine, pour un total d'environ 36 heures de calcul. [6] En guise de comparaison, l'estimation des paramètres des 56 modèles construits à l'aide de l'approche de White a nécessité moins de 10 secondes de calcul sur la même architecture.

Les valeurs présentées dans le tableau 7.1 sont la log-vraisemblance négative (définie par $-\mathcal{L}$) et le z-score $\mathcal{Z}$ obtenus par les 54 séquences de l'ensemble de validation dans le modèle stochastique associé à leur propre motif de repliement. Dans le cas du MRF de White, la log-vraisemblance négative n'est pas normalisée. Les valeurs normalisées pourraient être obtenues en additionnant le terme $\log \Lambda$ de l'équation 7.3 aux valeurs du tableau 7.1.

On observe que le modèle proposé performe mieux que celui de White dans presque tous les cas de test et on peut vérifier que la différence est significative. Comme les z-scores obtenus sur chaque cas de test sont indépendants, on peut assumer que la *différence moyenne* entre les z-scores est normalement distribuée, avec une variance estimée sans biais par la variance de la différence des z-scores sur les cas de test et divisée par le nombre de cas de test. En utilisant un test-t de Student bilatéral, on peut calculer la probabilité (valeur-p) que la différence moyenne observée survienne par hasard sous l'hypothèse nulle que l'espérance de la différence est 0. Ceci correspond à la probabilité que les deux modèles soient équivalents et que la différence de performance observée soit due au bruit d'échantillonnage. Pour l'expérience considérée

[6] On peut drastiquement réduire le temps de calcul en restreignant le nombre de valeurs explorées pour chaque hyper-paramètre pendant la validation croisée. Évidemment, ceci s'accompagne d'une certaine perte de qualité au niveau des modèles obtenus.

ici, on obtient une valeur-p inférieure à 1e-9, ce qui indique hors de tout doute que la différence observée est significative.

On remarque finalement que la log-vraisemblance négative totale et le z-score total (définis par la somme des valeurs obtenues sur chaque cas de test) sont des mesures de la performance moyenne sur l'ensemble de validation. Dans les expériences des sections suivantes, seules ces valeurs seront présentées.

7.2.2 Évaluation de différentes stratégies pour le partage des paramètres des lois de probabilité locales

L'hypothèse que les paramètres des lois de probabilités locales peuvent être partagés à l'intérieur d'environnements structuraux équivalents est un élément clef du modèle proposé. L'objectif des expériences décrites ici est d'évaluer différentes stratégies pour le partage des paramètres des lois de probabilité locales. Comme dans la section 7.2.1, on n'associe aucun attribut environnemental aux sommets et aux arcs des graphes de contact. Il n'y a donc qu'un seul type de sommet et un seul type d'arc dans les graphes de contact.

La première stratégie évaluée est celle du chapitre 6 : on suppose que $P(z_i \mid n_i^-) = P(z_j \mid n_j^-)$ dès que les sous-graphes G_i et G_j sous-tendus respectivement par $V_i \cup N(V_i)$ et $V_j \cup N(V_j)$ dans le graphe de contact sont isomorphes et que les propriétés des sommets et des arcs correspondants sont les mêmes (section 6.3.1).

Bien qu'on ait déjà vu que cette stratégie permette l'obtention d'un modèle supérieur au MRF de White (section 7.2.1), on constate que les lois de probabilités locales associées à certains environnements structuraux peu représentés dans les données d'apprentissage (voir la section 6.4.3) s'avèrent difficiles à paramétriser. On soupçonne également que cette stratégie est inutilement trop complexe et que les paramètres des lois de probabilités locales pourraient être partagés entre un plus grand nombre d'environnements structuraux.

La deuxième stratégie évaluée — appelée m0 — consiste à assumer que $P(z_i \mid n_i^-) =$

	nll	z-score (val-p)		nll	z-score (val-p)
chap. 6	26572	298.9 (< 1e-9)	m4	26527	310.9 (< 1e-9)
m6	26505	308.3 (< 1e-9)	m0	26551	308.0 (< 1e-9)
m5	26507	311.5 (< 1e-9)	MRF	—	224.0

TAB. 7.2 – Comparaison de différentes stratégies pour le partage des paramètres des lois de probabilités locales (voir le texte) dans le modèle proposé. Les valeurs présentées sont la log-vraisemblance négative totale (nll) et le z-score total obtenus sur l'ensemble de validation. Pour les 5 alternatives évaluées, le modèle proposé performe mieux que le MRF et la différence est toujours significative (valeur-p < 1e-9). La différence entre la meilleure approche (m5) et l'approche du chapitre 6 est significative (valeur-p < 2e-4).

$P(z_j \mid n_j^-)$ dès que les positions V_i et V_j ont un même nombre de sommets voisins avec les mêmes propriétés. Dans ce cas, toute information sur la position et l'orientation relative des acides aminés autour des positions V_i et V_j est perdue. On constate dans le tableau 7.2 que la stratégie m0 est supérieure à la stratégie du chapitre 6 (valeur-p < 0.002).

Afin de s'assurer que cette différence n'est pas uniquement due au manque d'exemples pour certaines catégories de lois de probabilité locales, trois approches hybrides ont été évaluées. L'approche m4 (resp. m5, m6), utilise la stratégie du chapitre 6 dès que $N(V_i)$ compte 4 (resp. 5, 6) acides aminés ou moins et la stratégie m0 sinon. Les résultats sont présentés dans le tableau 7.2.

En comparant la meilleure approche (m5) à la stratégie m0 (valeur-p < 0.09), on constate que le gain de performance obtenu en tenant compte du détail de la structure graphique locale est négligeable en regard des données disponibles. Ceci peut également indiquer deux choses : 1- la structure locale *détaillée* (position et orientation relative des acides aminés en contact) n'est pas une information pertinente

au problème considéré ici, ou 2- le graphe de contact encode mal cette information.

7.2.3 *Importance des différents types d'interactions entre les acides aminés en contact*

Une hypothèse généralement admise dans le domaine de la reconnaissance de l'architecture des protéines (et dans ce travail) est l'importance de modéliser les interactions entre acides aminés lors de la définition de mesures de pseudo-énergie (voir la section 3.2 et [45]). On entend ici que l'affinité d'un acide aminé pour une position structurale donnée dépend non seulement de variables comme le degré d'enfouissement dans la protéine ou le type de structure secondaire, mais également de la nature des acides aminés occupant des positions voisines dans la protéine. L'objectif des expériences présentées ici est de vérifier l'importance des différents types d'interactions entre les acides aminés en contact (e.g. interactions "binaires", interactions d'ordre supérieur, etc).

On procède d'abord en éliminant toutes les dépendances entre les Z_i dans le modèle global : on assume que

$$P(z) = \prod_i P(z_i). \tag{7.10}$$

On construit ensuite un modèle pour p_Z tel que proposé dans le chapitre 6, c'est-à-dire en supposant que $P(z_i) = P(z_j)$ dès que la structure locale du graphe de contact est similaire autour des positions V_i et V_j. La variable aléatoire Z_i ne dépend donc que de la structure locale du graphe de contact en position V_i. Le modèle stochastique résultant est analogue aux *3D-1D potentials* évoqués dans la section 3.2.4.

Les résultats obtenus pour trois des stratégies de la section 7.2.2 sont présentés dans le tableau 7.3. On constate que les modèles obtenus en éliminant toutes les dépendances entre les Z_i performent moins bien que leurs contreparties qui en tiennent compte et la différence est significative. Ceci indique clairement que la modélisation d'interactions entre les acides aminés en contact est appropriée.

	nll	z-score
chap. 6	26572	298.9
m5	26507	311.5
m0	26551	308.0
m5–h0	26542	303.3
MRF	—	224.0

	nll	z-score (val-p)
profile	26696	269.2 (< 1e-4)
m5–prof	26637	277.9 (< 1e-4)
m0–prof	26685	273.2 (< 1e-4)

TAB. 7.3 – Pour trois des stratégies de la section 7.2.2, on évalue l'importance de modéliser des interactions entre acides aminés voisins. Les modèles obtenus en éliminant toutes les dépendances entre les Z_i (resp. profile, m5–prof, m0–prof) performent moins bien que leurs contreparties (resp. chap.6, m5, m0) et la différence est toujours significative (valeur-p < 1e-4). Le modèle n'utilisant que des MLP sans unité cachée (m5–h0) est moins performant que son homologue m5 (valeur-p < 1e-4).

Cependant, un des objectifs de ce mémoire est de modéliser des interactions d'ordre supérieur qui surgissent lorsque trois acides aminés ou plus sont simultanément en contact. Afin d'évaluer l'importance des interactions d'ordre supérieur entre les acides aminés, on construit également un modèle stochastique en n'utilisant que des MLP sans unité cachée (m5–h0). Ces MLP sont par définition incapables de capturer des dépendances d'ordre supérieur entre les Z_i. Le modèle stochastique résultant apparaît moins efficace que son homologue avec unités cachées (m5) et la différence est significative (valeur-p < 1e-4). Ceci est un indice de l'utilité de modéliser des interactions d'ordre supérieur entre les acides aminés en contact.

Par ailleurs, on remarque que le modèle sans unité cachée est significativement plus efficace que le MRF de White (valeur-p < 1e-4). En fait, même les modèles obtenus en éliminant toute dépendance entre les Z_i performent mieux que le MRF de White (valeur-p < 1e-4) ! Ceci est *a priori* surprenant, sachant que ces modèles ne peuvent par construction modéliser aucune interaction entre les acides aminés en

contact (ce qui n'est pas le cas du MRF, qui capture toutes les dépendances d'ordre 2 entre les Z_i).

Pour comprendre cela, considérons le modèle m0–prof. On se souvient qu'avec la stratégie m0, toute information sur la position relative des acides aminés est perdue. Dans ce cas, lorsqu'on élimine les dépendances entre les Z_i, la seule variable influençant l'affinité d'un acide aminé pour une position est le *nombre* de contacts impliquant cette position. On doit en déduire que le nombre local de contacts est une quantité très informative.

On peut expliquer ceci en se remémorant le résultat de Thomas et Dill [50], à savoir que tout pseudo-potentiel encode principalement l'effet hydrophobique — la préférence des acides aminés hydrophobes pour les positions structurales enfouies dans les protéines — plutôt que de véritables interactions entre acides aminés. En remarquant que le nombre de contacts est un bon indice du degré d'enfouissement d'une position structurale à l'intérieur d'une protéine, on comprend que le modèle proposé approxime très efficacement l'effet hydrophobique. Par opposition, le MRF de White extrapole l'hydrophocité autour d'une position uniquement en fonction du nombre d'acides aminés hydrophobes présents dans le voisinage de cette position.

7.2.4 Encodage des acides aminés

Dans les expériences effectuées jusqu'ici, on s'est contenté d'utiliser un encodage orthogonal pour les acides aminés. L'objectif poursuivi ici est de voir si l'encodage adaptatif décrit dans la section 6.4.1 permet d'améliorer les performances du modèle proposé. On s'attend à ce que l'utilisation de cet encodage permette au modèle proposé de capturer plus efficacement les dépendances entre les acides aminés en contact, particulièrement dans le cas d'environnements structuraux pour lesquels on ne dispose que de peu d'exemples.

On se rappelle que l'encodage adaptatif repose sur la projection des codes orthogonaux dans un espace continu de plus faible dimension. Dans cette expérience, chaque

	nll	z-score (val-p)
chap. 6	26572	298.9
p1	26603	293.5 (0.01)
p3	26576	299.5 (0.84)
p5	26571	302.3 (0.06)
p7	26581	298.1 (0.76)

	nll	z-score (val-p)
m5	26507	311.5
p1	26540	308.0 (0.05)
p3	26517	310.8 (0.76)
p5	26511	314.2 (0.13)
p7	26542	308.4 (0.09)

TAB. 7.4 – Effet de l'encodage adaptatif des acides aminés lorsqu'utilisé avec le modèle du chapitre 6 (table de gauche) et la stratégie m5 de la section 7.2.2 (table de droite). Les résultats obtenus avec l'encodage orthogonal sont dans la première rangée. Les résultats obtenus avec une projection des codes orthogonaux sur $[-1,1]^k$ sont dans la rangée pk. Les valeurs-p sont calculées par rapport au modèle utilisant l'encodage orthogonal.

MLP "apprend" indépendamment la projection optimale pour le sous-problème auquel il est affecté.

On teste ici des projections sur le cube $[-1,1]^k$, avec $k \in \{1,3,5,7\}$. Les résultats sont dans le tableau 7.4. Les meilleurs résultats sont obtenus avec $k = 5$ mais le gain en performance est peu significatif (valeur-p : 0.06). Malgré cela, l'utilisation d'un encodage adaptatif réduit significativement le nombre de paramètres libres des MLP, facilitant ainsi la procédure d'apprentissage.

Par ailleurs, on peut vérifier que les MLP apprennent bien à regrouper les acides aminés en fonction de propriétés communes pertinentes : en accord avec le résultat de Thomas et Dill [50], à savoir que tout pseudo-potentiel encode principalement l'effet hydrophobique, on observe que la corrélation moyenne entre les codes résultants d'une projection sur $[-1,1]^1$ et l'hydrophobicité des acides aminés encodés est 0.71.

7.2.5 *Optimisation du graphe d'indépendance*

On a vu au chapitre 6 que le choix du graphe d'indépendance dirigé (GID) à associer a un motif structural n'est pas unique. Bien qu'*a priori* équivalentes d'un point de vue probabiliste, les factorisations de p_Z résultantes ne sont pas équivalentes du point de vue de l'apprentissage.

	nll	z-score (val-p)
m5	26507	311.5
pas opt.	26524	307.6 (0.06)

TAB. 7.5 – Comparaison du meilleur modèle (m5) à un modèle similaire obtenu sans optimisation du graphe d'indépendance. On utilise pour ce dernier la factorisation de p_Z induite par l'ordre de la séquence des acides aminés.

Dans la section 6.1.2, on a proposé un algorithme pour le choix d'une factorisation qui facilite au maximum l'estimation des paramètres de p_Z. Cet algorithme est évalué dans le tableau 7.5.

Bien que la différence de performance observée ne soit pas très significative, l'optimisation du graphe d'indépendance a pour effet de réduire le nombre total de MLP requis pour la paramétrisation des lois de probabilités locales. L'apprentissage est donc plus rapide.

On note finalement que la technique testée ici est applicable dans tous les contextes où des GID sont utilisés pour factoriser des lois de probabilité en haute dimension. Il serait donc intéressant de l'appliquer à d'autres problèmes.

7.2.6 *Ajout d'attributs environnementaux dans les graphes de contact*

Dans toutes les expériences décrites jusqu'ici, l'information structurale a été résumée uniquement à l'aide de la notion de contact. On a vu qu'il est possible de

représenter plus finement un noyau structural conservé en étiquetant les sommets et les arcs des graphes de contact à l'aide d'attributs environnementaux. Ceux-ci permettent la représentation de caractéristiques structurales jugées importantes mais qui ne sont pas encodées explicitement par le graphe de contact.

	nll	z-score
chap. 6	26013	363.4
m5	25946	366.1
m0	25947	358.9
m5–p5	25961	369.7
MRF	—	325.6

	nll	z-score
profile	26041	341.6
m5–prof	25960	343.8
m0–prof	25962	338.7

TAB. 7.6 – Comparaison du modèle proposé au MRF de White *et al* [6] après ajout d'attributs environnementaux (exposition au solvant, type de structure secondaire, contacts inter/intra ESS) dans les graphes de contact (voir le texte). Les modèles évalués sont ceux du tableau 7.3 et performent tous mieux que le MRF (valeur-p : $<$ 1e-4). On teste également l'encodage adaptatif avec une projection sur $[-1, 1]^5$ (m5-p5). Ceci semble améliorer la discrimination (valeur-p : 0.016). Les modèles obtenus en éliminant toutes les dépendances entre les Z_i (resp. profile, m5–prof, m0–prof) performent moins bien que leurs contreparties (resp. chap.6, m5, m0) et la différence est toujours significative (valeur-p : $<$ 1e-4).

Dans l'article où White *et al* définissent leur modèle [6], les propriétés associées à chaque sommet sont l'exposition au solvant et le type de structure secondaire. Une seule propriété est associée aux arcs. Elle indique si l'arc représente un contact entre deux positions situées dans un même élément de structure secondaire (ESS) ou dans deux éléments différents. Dans ce cas, il y a 4 types de positions structurales (exposée + α, non-exposée + α, exposée + β, non-exposée + β) et 16 types d'interactions. On a vu dans la section 7.2.3 que le MRF extrapole mal l'accessibilité du solvant à

chaque position à partir de l'information de contact. On s'attend donc à ce que le MRF performe beaucoup mieux lorsque cette quantité est représentée explicitement.

Le tableau 7.6 compare les 6 modèles de la section 7.2.3 au MRF de White lorsqu'on utilise la définition exacte de [6] pour les patrons structuraux. [7] En comparant au tableau 7.3, on constate que tous les modèles évalués performent beaucoup mieux avec l'introduction des propriétés environnementales. Bien que les écarts (mesurés sur les z-scores) entre les différents modèles se soient resserrés, on observe que le classement relatif établi dans le tableau 7.3 demeure inchangé.

7.2.7 Affinité pour les structures de protéines homologues/analogues

Lors de la définition d'un patron structural, il est difficile de ne retenir que les caractéristiques structurales vraiment conservées dans toutes les protéines qui adoptent une conformation similaire à celle que représente ce patron. On entend ici que le patron sera toujours biaisé vers la ou les protéines qui ont servies à sa construction. En conséquence, comme les modèles stochastiques sont construits directement à partir du graphe de contact, on s'attend aussi à ce qu'ils soient biaisés vers les séquences des protéines ayant servies à la construction des graphes de contact. Dans la littérature, on appelle ceci la "mémoire de séquence".

La principale conséquence de cette situation est que l'"énergie" d'une séquence dans le patron construit à partir de sa propre structure 3-D sera toujours plus faible que celle qu'obtiendrait une séquence homologue ou analogue. Ainsi, la performance du modèle proposé dans les expériences précédentes est un estimé optimiste de son utilité pour la reconnaissance d'homologies structurales lointaines.

---

[7] Dans [6], l'exposition au solvant est résumée par une valeur binaire de type exposé/non-exposé. Bien que ce ne soit pas le cas pour le modèle proposé, le MRF de [6] requiert, par sa construction, que les propriétés environnementales soit discrétisées (voir la section 5.3). Dans ce travail, une position sera considérée exposée au solvant si son indice d'accessiblité au solvant (section 7.1.2) est supérieur à 0.2. Elle sera considérée non-exposée sinon.

Des expériences supplémentaires sont donc requises afin de s'assurer que le modèle proposé est supérieur au MRF de White dans une situation plus proche d'un "vrai" problème de prédiction.

Pour chaque protéine S de l'ensemble de validation, on identifie une protéine $\mathcal{H}(S)$ qui partage grossièrement la même architecture à l'aide de la banque de données FSSP [85]. Pour éliminer la "mémoire de séquence", on requiert que l'identité de séquence entre S et $\mathcal{H}(S)$ soit inférieure à 20% et que la RMSD[8] mesurée sur les segments de structure secondaire conservés soit supérieure à 2.0Å. On a pu trouver $\mathcal{H}(S)$ satisfaisant cette condition pour 34 des 54 séquences de l'ensemble de validation. Ces 34 séquences constituent l'ensemble de validation pour les deux séries d'expériences décrites dans cette section.

Pour chaque paire $\{S, \mathcal{H}(S)\}$, on construit un patron structural représentant leur architecture commune. Ce patron est défini à l'aide d'un graphe de contact comme dans la section 5.2. Le graphe de contact correspond aux segments de structure secondaire conservés entre S et $\mathcal{H}(S)$, mais extraits de la structure 3-D détaillée de $\mathcal{H}(S)$. [9] La longueur des boucles est ensuite fixée de façon à ce que S soit correctement alignée à l'intérieur du patron.

Si on associe ensuite un modèle stochastique à ce patron, la log-vraisemblance de S dans ce modèle est un estimé réaliste du comportement de celui-ci dans un "vrai" problème de reconnaissance de séquences.

Pour la première série d'expériences, on n'associe aucun attribut environnemental aux sommets et aux arc des graphes de contact. Pour la seconde, on ajoute les attributs environnementaux de la section 7.2.6. Les résultats sont présentés dans les tableaux 7.7 et 7.8.

---

[8] Root Mean Square Deviation.

[9] FSSP contient les alignements des structures de toutes les protéines de PDB qui partagent une même architecture. Ces alignements permettent d'identifier facilement les segments de structure secondaire conservés dans les protéines alignées.

	nll (auto)	nll (hom)	z-score (auto)	z-score (hom)
m5	17479	17860	202.2	138.4 (< 1e-9)
m5–h0	17499	17855	197.1	137.9 (< 1e-9)
m5–p5	17480	17849	204.5	143.4 (< 1e-9)
MRF	—	—	142.9	98.9

TAB. 7.7 – Affinité de 34 séquences de l'ensemble de validation pour les structures de protéines homologues/analogues mesurée à l'aide du meilleur modèle du tableau 7.2 et du MRF de White. On évalue également un modèle sans unités cachées (m5–h0) ainsi qu'un modèle utilisant l'encodage adaptatif des acides aminés avec une projection sur $[-1, 1]^5$ (m5–p5). Les valeurs présentées sont la log-vraisemblance totale (nll) et le z-score total obtenus par les 34 séquences dans les modèles stochastiques construits avec leurs propres structures 3-D (auto) et avec les structures 3-D de protéines homologues/analogues (hom). Les modèles évalués performent mieux que le MRF dans tous les cas (valeur-p (hom) : < 1e-9). La différence observée entre m5 et m5–h0 n'est pas significative (valeur-p (hom) : 0.68) mais la différence observée entre m5 et m5–p5 semble significative (valeur-p (hom) : 0.002).

Dans les deux cas, les variantes évaluées sont les meilleurs modèles des sections 7.2.2 (m5) et 7.2.4 (m5-p5) ainsi que le MRF de White. On évalue également un modèle n'utilisant que des MLP sans unités cachées (m5–h0). La mesure de comparaison utilisée est la log-vraisemblance (et les z-scores) des 34 séquences de l'ensemble de validation dans les modèles stochastiques construits à partir des structures 3-D de leurs homologues dans FSSP. On voit très bien l'influence de la "mémoire de séquence" en comparant avec les valeurs obtenues lorsque les séquences sont alignées dans leur propre structure 3-D.

Pour les deux séries d'expériences, on observe que le classement établi dans les sections 7.2.2 et 7.2.6 est préservé : le modèle proposé (m5) apparaît supérieur au

	nll (auto)	nll (hom)	z-score (auto)	z-score (hom)
m5	17075	17793	239.3	152.0 (< 1e-4)
m5–h0	17114	17776	239.7	154.9 (< 1e-4)
m5–p5	17086	17812	241.3	152.5 (< 1e-4)
MRF	—	—	211.5	138.0

TAB. 7.8 – Répétition des expériences du tableau 7.7 après ajout des attributs environnementaux de la section 7.2.6 (exposition au solvant, type de structure secondaire, contacts inter/intra SSE) dans les graphes de contact. Les modèles évalués performent mieux que le MRF dans tous les cas (valeur-p : < 1e-4). La différence observée entre m5 et m5–h0 n'est pas significative (valeur-p (hom) : 0.07).

MRF de White et la différence est toujours significative. Comme dans le tableau 7.6, l'encodage adaptatif des acides aminés (m5–p5) semble améliorer la discrimination de manière significative. Par contre, la modélisation des interactions d'ordre supérieur ne semble apporter aucune amélioration. En fait, dans la seconde série d'expériences (tableau 7.8), cela semble même contre-productif. Ceci semble indiquer que les interactions d'ordre supérieur sont mal conservées entre les protéines homologues et qu'il est inutile d'en tenir compte si on se restreint à des patrons structuraux "rigides" construits à partir de la structure 3-D d'une seule protéine.

7.3 Discussion

Dans ce chapitre, on a montré que le modèle stochastique proposé dans ce mémoire est supérieur au modèle de White *et al* [6]. On a expliqué dans la section 7.2.3 que ceci est principalement dû au fait que le modèle proposé approxime mieux l'effet hydrophobique en classifiant des distributions d'acides aminés locales en fonction de la structure locale du graphe de contact. On a aussi vu que les interactions d'ordre

supérieur constituent une source d'information utile.

Également, dans la section 7.2.4, on a démontré que l'encodage adaptatif des acides aminés permet de réduire la taille des MLP sans affecter la performance du modèle proposé. En fait, l'utilisation de cet encodage semble améliorer la discrimination entre les "bonnes" et les "mauvaises" séquences (sections 7.2.6 et 7.2.7).

On a aussi montré que le modèle proposé demeure supérieur au MRF de White lorsqu'on raffine les patrons structuraux en ajoutant des attributs environnementaux aux sommets et aux arcs des graphes de contact (section 7.2.6). On a vérifié cela en utilisant la même définition pour les patrons que White *et al* [6] lors de l'élaboration du MRF.

Finalement, dans la section 7.2.7, en mesurant l'affinité de séquences pour les structures 3-D de protéines homologues, on a montré que le modèle proposé permet une meilleure représentation des homologies structurales dans les protéines. On a cependant noté que ceci n'est pas dû à la modélisation d'interactions d'ordre supérieur entre les acides aminés.

Chapitre 8

CONCLUSION

Dans ce mémoire, on a défini une nouvelle mesure de compatibilité séquence–structure utilisable pour la reconnaissance des motifs de repliement des protéines (*protein fold recognition*). La mesure proposée a été définie dans un cadre mathématique rigoureux inspiré de [6] et repose sur la construction d'un nouveau type de modèle stochastique pour les séquences d'acides aminés — possiblement non-homologues — qui partagent un même motif de repliement.

Dans le chapitre 7, on a démontré que le modèle stochastique introduit dans ce mémoire est supérieur au modèle proposé par White *et al* [6]. Une comparaison avec ce modèle s'est avérée intéressante pour trois raisons. Premièrement, le modèle de White est défini à partir d'un formalisme similaire a celui que nous avons adopté dans ce travail. On a donc pu comparer les deux modèles d'une façon fiable à l'intérieur d'un cadre probabiliste. Deuxièmement, le modèle de White et le modèle proposé sont construits à partir d'une même représentation des motifs de repliement des protéines (les graphes de contact). On a donc pu éviter tout biais dû au choix de représentations différentes pour les motifs de repliement. Troisièment, le modèle de White induit une mesure de compatibilité séquence-structure analogue aux potentiels pseudo-energétiques traditionnels. On peut donc raisonnablement généraliser les observations faites par rapport à l'approche de White au cas des approches traditionnelles.

Pour bien comprendre les différences entre le modèle proposé et le modèle de White, il est utile de se rappeler que les mesures de compatibilité séquence-structure

peuvent être exprimées sous la forme générale

$$\mathcal{E}(s) = \sum_i \mathcal{E}_i(s_i) + \sum_{i,j} \mathcal{E}_{i,j}(s_i, s_j) + \sum_{i,j,k} \mathcal{E}_{i,j,k}(s_i, s_j, s_k) + \cdots \tag{8.1}$$

où on compte un terme par acide aminé et un terme par paire (resp. triplet, quadruplet, ...) d'acides aminés en contact. On se souvient que les termes d'ordre 1 permettent de modéliser la compatibilité des acides aminés pour la position structurale à laquelle ils sont contraints dans un motif structural alors que les contributions d'ordre 2 et plus permettent de modéliser des interactions entre les acides aminés en contact.

Avec le modèle stochastique de White (et avec les potentiels pseudo-énergétiques traditionnels), on se rappelle que seules des interactions d'ordre 2 sont modélisées. Par contre, celles-ci sont toutes modélisées (dans le langage de l'équation 8.1, ceci veut dire que tous les termes d'ordre 3 ou plus sont nuls mais que tous les termes d'ordre 2 et moins sont non-nuls). Par opposition, l'approche proposée est capable de "sélectionner" toutes les interactions pour lesquelles les données permettent de dériver des paramètres énergétiques utiles. Ceci n'inclut pas nécessairement toutes les interactions d'ordre 2 mais inclut potentiellement des interactions d'ordre supérieur. Il s'agit d'une différence importante entre l'approche proposée et les approches plus traditionnelles et on a montré au chapitre 7 que ceci permet de mieux modéliser l'affinité des séquences d'acides aminés pour leurs propres structures 3-D. Cependant, on a aussi vu que lorsque les modèles stochastiques sont construits en fonction de la structure 3-D d'une seule protéine chacun, l'inclusion de contributions d'ordre supérieur ne permet pas une meilleure reconnaissance d'homologies structurales. Cette observation semble indiquer que les interactions d'ordre supérieur sont mal conservées entre des protéines homologues. Il apparaît donc inutile d'en tenir compte si on se limite à des modèles structuraux "rigides" qui n'encodent pas les variations structurales observées entre les protéines homologues.

Sachant cela, comment expliquer que le modèle proposé performe mieux que le

modèle de White ? Il est important de noter que les deux modèles diffèrent également quant au type d'information qu'ils utilisent pour le calcul des différentes contributions dans l'équation 8.1.

Dans le modèle de White, on assume que la contribution d'ordre 1 associée à une position structurale ne dépend que des propriétés qui étiquettent cette position dans le graphe de contact (e.g. type de structure secondaire, exposition au solvant, etc.). On assume également que la contribution associée à une interaction ne dépend que des propriétés des positions en contact. Par opposition, avec l'approche proposée, toutes les contributions impliquant une même position (i.e. $\mathcal{E}_x$, $\mathcal{E}_{x,j}$, $\mathcal{E}_{x,j,k}$, ...) sont calculées *simultanément* à l'aide d'un réseau de neurones artificiel en fonction de la structure locale du graphe de contact autour de cette position. En conséquence, toutes les contributions (d'ordre 1, 2, 3, ...) impliquant une position dépendent de la structure locale du graphe de contact autour de cette position.

Il s'agit d'une distinction importante par rapport au modèle de White et on a vu au chapitre 7 qu'elle contribue de façon drastique à la différence de performance entre les deux méthodes. On a en effet remarqué que la structure locale du graphe de contact — ou plus simplement, le nombre de contacts — autour d'une position est un bon indice du degré d'enfouissement de cette position dans une protéine. On a expliqué que cette information permet à l'approche proposée d'approximer très efficacement l'effet hydrophobique (la préférence des acides aminés hydrophobes pour les position structurales enfouies à l'intérieur de la molécule).

L'observation ci-dessus souligne l'importance d'utiliser une bonne représentation de l'environnement structural autour de chaque position. Cette représentation devrait en particulier résumer efficacement le degré d'enfouissement des différentes positions à l'intérieur des protéines. Dans l'approche de White, le degré d'enfouissement de chaque position est approximé à l'aide d'un indice d'exposition au solvant binaire de type exposé/non-exposé. Les résultats du chapitre 7 indiquent clairement qu'une représentation plus sophistiquée, basée par exemple sur la structure locale du graphe

de contact, serait préférable.

RÉFÉRENCES

[1] C. Caskey, R. Eisenberg, E. Lander, and J. Strauss. Hugo statement on patenting of DNA. *Genome Digest*, 2 :6, 1995.

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

[3] S. Miyazawa and R.L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures : quasi-chemical approximation. *Macromolecules*, 18 :534–552, 1985.

[4] J.U. Bowie, R. Lüthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253 :164–170, 1991.

[5] S.H. Bryant and C.E. Lawrence. An empirical energy function for threading protein sequence through the folding motif. *Proteins : Structure, Function and Genetics*, 16 :92–112, 1993.

[6] J. White, I. Muchnick, and T.F. Smith. Modeling protein cores with Markov random fields. *Math. Biosci.*, 124 :149–179, 1994.

[7] A. Godzik, A. Kolinski, and J. Skolnick. Topology fingerprint approach to the inverse protein folding problem. *Journal of Molecular Biology*, 227(1) :227–38, 1992.

[8] T. Grossman, R. Farber, and A. Lapedes. Neural net representations of empirical protein potentials. In C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak, editors, *Proceedings of the 3rd International Conference on Intelligent Systems for Molecular Biology*, pages 154–161. AAAI Press, Menlo Park, CA, 1995.

[9] D. Voet and J.G. Voet. *Biochemistry*. John Willey and Sons Inc., New York, 1995.

[10] C. Branden and J. Tooze. *Introduction to Protein Structure.* Garland Publishing Inc., New York, 1991.

[11] J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids. *Nature*, 171 :737–738, 1953.

[12] J.D. Watson and F.H.C. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171 :964–967, 1953.

[13] D.A. Benson, M.S. Boguski, O.J. Lipman, J. Ostell, B.F. Ouellette, B.A. Rapp, and D.L. Wheeler. Genbank. *Nucleic Acids Research*, 27(1) :12, 1999.

[14] A. Bairoch and R. Apweiler. The swiss-prot protein sequence data bank and its supplement trembl in 1999. *Nucleic Acids Research*, 27 :49, 1999.

[15] Barker WC, Garavelli JS, McGarvey PB, Marzec CR, Orcutt BC, Srinivasarao GY, Yeh LS, Ledley RS, Mewes HW, Pfeiffer F, Tsugita A, and Wu C. The pir-international protein sequence database. *Nucleic Acids Research*, 27 :39–43, 1999.

[16] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, H.Weissig T.N. Bhat, I.N.Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28 :235–242, 2000.

[17] P. Green, D. Lipman, L. Hillier, R. Waterston, D. States, and J.-M. Claverie. Ancient conserved regions in new gene sequences and the protein databases. *Science*, 259 :1711–1716, 1993.

[18] C.A. Orengo, D.T. Jones, and J.M. Thornton. Protein superfamilies and domain superfolds. *Nature*, 372 :631–634, 1994.

[19] D. Gusfield. *Algorithms on Strings, Trees, and Sequences : Computer Science and Computational Biology.* Cambridge University Press, New York, 1997.

[20] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. SCOP : a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247 :536–540, 1995.

[21] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton. Cath – a hierarchic classification of protein domain structures. *Structure*, 5(8) :1093–1108, 1997.

[22] F.M.G Pearl, D. Lee, J.E. Bray, I. Sillitoe, A.E. Todd, A.P. Harrison, J.M. Thornton, and C.A. Orengo. Assigning genomic sequences to cath. *Nucleic Acids Research*, 28(1) :277–282, 2000.

[23] J. Moult, T. Hubbard, K. Fidelis, and J.T. Pedersen. Critical assessment of methods of protein structure prediction (casp) : Round III. *Proteins : Structure, Function, and Genetics*, 3 (suppl.) :2–6, 1999.

[24] M.S. Waterman. *Introduction to Computational Biology.* Chapman & Hall, London, 1996.

[25] D.J. Lipman and W.R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227 :1435–1441, 1985.

[26] W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences (USA)*, 85 :2444–2448, 1988.

[27] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5 :345–352, 1978.

[28] S. Henikoff and J.G. Henikoff. Amino acid substition matrices from protein blocks. *Proceedings of the National Academy of Sciences (USA)*, 89(10) :915–919, 1991.

[29] S.B. Needleman and C.D. Wunsh. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48 :443–53, 1970.

[30] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147 :195–197, 1981.

[31] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215 :403–410, 1990.

[32] W.R. Pearson. Effective protein sequence comparison. *Methods Enzymol.*, 266 :227–258, 1996.

[33] S.F. Altschul, M.S. Boguski, W. Gish, and J.C. Wootton. Issues in searching molecular sequence databases. *Nature Genetics*, 6 :119–129, 1994.

[34] W.R. Pearson. Comparison of methods for searching protein sequence databases. *Protein Science*, 4(6) :1145–1160, 1995.

[35] S.E. Brenner, C. Chothia, and T.J.P. Hubbard. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of the National Academy of Sciences (USA)*, 95 :6073–6078, 1998.

[36] J.D. Thompson, D.G. Higgins D.G., and T.J. Gibson. Clustal w : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22 :4673–4680, 1994.

[37] H. Carrillo and D. Lipman. The multiple alignment problem in biology. *SIAM Journal of Applied Mathematics*, 48 :1073–1082, 1988.

[38] D. Sankoff. Minimal mutation trees of sequences. *SIAM Journal of Applied Mathematics*, 28 :35–42, 1975.

[39] D. Sankoff and R. Cedergreen. Simultaneous comparisons of three or more sequences related by a tree. In W.R. Pearson, editor, *Protein sequence comparison and protein evolution, tutorial T6 at ISMB95*, pages 253–264. Cambridge, UK, 1995.

[40] A. Krogh, M. Brown, I.S. Mian, K. Sjolander, and D. Haussler. Hidden Markov models in computational biology : application to protein modeling. *Journal of Molecular Biology*, 235 :1501–1531, 1994.

[41] K. Karplus, C. Barrett, and R. Hughey. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14(10) :846–856, 1998.

[42] R.F. Doolittle, M. Hunkapiller, L.E. Hood, S. Devare, K. Robbins, S. Aaronson, and H. Antoniades. Simian sarcoma virus oncogene v-sis is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science*, 221 :275, 1983.

[43] J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia. Sequence comparison using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology*, 284 :1201–1210, 1998.

[44] R.H. Lathrop and T.F. Smith. Global optimum protein threading with gapped alignment and empirical pair score functions. *Journal of Molecular Biology*, 255 :641–665, 1996.

[45] D.T. Jones and J.M. Thornton. Potential energy functions for threading. *Proceedings of the National Academy of Sciences (USA)*, 6 :210–216, 1996.

[46] R.H. Lathrop, R.G. Rogers, and T.F. Smith. A Bayes–optimal sequence structure theory that unifies protein sequence–structure recognition and alignment. *Bulletin of Mathematical Biology*, 60 :1039–1071, 1998.

[47] R. Lüthy, J.U. Bowie, and D. Eisenberg. Assessment of protein models with three-dimensional profiles. *Nature*, 356 :83–85, 1992.

[48] R.H. Lathrop. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein engineering*, 7 :1059–1068, 1994.

[49] D.T. Jones, J.M. Thornton, and W.R. Taylor. A new approach to protein fold recognition. *Nature*, 358 :86–89, 1992.

[50] P.D. Thomas and K.A. Dill. Statistical potentials extracted from protein structures : hom accurate are they? *Journal of Molecular Biology*, 257 :457–469, 1996.

[51] S.H. Bryant. Evaluation of threading specificity and accuracy. *Proteins : Structure, Function and Genetics*, 26 :172–185, 1996.

[52] A. Marchler-Bauer and S.H. Bryant. Measures of threading specificity and accuracy. *Proteins : Structure, Function and Genetics*, S1 :74–82, 1997.

[53] A. Marchler-Bauer and S.H. Bryant. A measure of progress in fold recognition ? *Proteins : Structure, Function and Genetics*, S3 :218–225, 1999.

[54] R.B. Russell and G.J. Barton. Structural features can unconserved in proteins with similar folds — an analysis of sidechain to sidechain contacts, secondary structure and accessibility. *Journal of Molecular Biology*, 244 :332–350, 1994.

[55] T. Madej, J.-F. Gibrat, and S.H. Bryant. Threading a database of protein cores. *Proteins : Structure, Function and Genetics*, 23 :356–369, 1995.

[56] M. Gerstein and R.B. Altman. Using a measure of structural variation to define a core for the globins. *Comput. Appl. Biosci.*, 11 :633–644, 1995.

[57] K.A. Dill and H.S. Chan. From levinthal to pathways to funnels. *Nature Structural Biology*, 4 :10–19, 1997.

[58] B. Honig. Protein folding : from the levinthal paradox to structure prediction. *Journal of Molecular Biology*, 293 :283–293, 1999.

[59] A.G. Murzin. Structure classification-based assessment of casp3 predictions for the fold recognition targets. *Proteins : Structure, Function and Genetics*, S3 :88–103, 1999.

[60] D. Heckerman. A tutorial on learning with Bayesian networks, technical report msr-tr-95-06. Technical report, Microsoft Research, Redmond, WA, 1995.

[61] W. Buntime. A guide to the litterature on learning probabilistic networks from data. *IEEE Transactions on knowledge and data engineering*, 8(2), 1996.

[62] P. Baldi and S. Brunak. *Bioinformatics : The Machine Learning Approach.* MIT Press, Cambridge, MA, 1998.

[63] P. Smyth, D. Heckerman, and M. I. Jordan. Probabilistic independance networks for hidden Markov probability models. *Neural Computation*, 9 :227–269, 1997.

[64] B. Frey. *Graphical models for machine learning and digital communication.* MIT Press, Cambridge, MA, 1998.

[65] J. Pearl. *Probabilistic reasonning in intelligen systems.* Morgan Kaufmann, San Mateo, CA, 1988.

[66] J. Whittaker. *Graphical models in applied multivariate statistics.* John Wiley & Sons, New York, 1990.

[67] S.L. Lauritzen. *Graphical Models.* Oxford University Press, Oxford, 1996.

[68] D. H. Ackley, G. E. Hinton, and T.J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9 :147–169, 1985.

[69] G. E. Hinton and T. J. Sejnowski. Learning and relearning in boltzmann machines. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing : Explorations in the microstructure of cognition*, volume 1, chapter 7. MIT Press, Cambridge, MA, 1986.

[70] J. W. Modestino and J. Zhang. A Markov random field model-based approach to image interpretation. In R. Chellappa and A. Jain, editors, *Markov Random Fields : Theory and Application*, pages 369–408. Academic Press, San Diego, CA, 1993.

[71] R. Kindermann and J.L. Snell. *Markov Random Fields and their Applications.* American Mathematical Society, Providence, RI, 1980.

[72] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. Ser. B*, 50 :157–224, 1988.

[73] T.A. Hutchinson D.J. Spiegelhalter, A.P. Dawid and R.G. Cowell. Probabilistic expert systems and graphical modeling : A case study in drug safety. *Phil. Trans. R. Soc. Lond. A*, 337 :387–405, 1991.

[74] J. Pearl, D. Geiger, and T. Verma. The logic of influence diagrams. In R.M. Oliver and J.Q. Smith, editors, *Influence Diagrams, Belief Nets, and Decision Analysis*, pages 67–83. John Wiley, Chichester, UK, 1990.

[75] F.V. Jensen, S.L. Lauritzen, and K.G. Olesen. Bayesian updating in recursive graphical models by local computations. *Computational Statistical Quaterly*, 4 :269–282, 1990.

[76] A.P. Dawid. Application of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, 2 :25–36, 1992.

[77] C.M. Bishop. *Neural networks for pattern recognition.* Clarendon Press, Oxford, 1995.

[78] S.K. Riss and A. Krogh. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *Journal of computationnal biology*, 3(1) :163–183, 1996.

[79] C.H. Wu. Artificial neural networks for molecular sequence analysis. *Computers and chemistry*, 21(4) :237–256, 1997.

[80] S. Bengio and Y. Bengio. Modeling high-dimensional discrete data with multi-layer neural networks. *IEEE Transactions on Neural Networks special issue on data mining and knowledge discovery*, 2000.

[81] R.R. Bahadur. A representation of the joint distribution of responses to n dichotomous items. In H. Solomon, editor, *Studies in item analysis and prediction*, pages 158–168. Standford University Press, California, 1961.

[82] S.E. Brenner, P. Koehl, and M. Levitt. The astral compendium for protein structure and sequence analysis. *Nucleic Acids Research*, 28(1) :254–256, 2000.

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

[84] L.A. Mirny, A.V. Finkelstein, and E.I. Shakhnovich. Statistical significance of protein structure prediction by threading. *Proceedings of the National Academy of Sciences (USA)*, 97 :9978–9983, 2000.

[85] L. Holm and C. Sander. Mapping the protein universe. *Science*, 273 :595–603, 1996.

REMERCIEMENTS

Je tiens d'abord à remercier les membres de ma famille, Thérèse, Alain, Nicolas et Charles, pour leur soutien constant au cours de ce travail, ainsi que Marie-Claude, qui fut une source fidèle d'ensoleillement et de bonne humeur. Je tiens également à remercier les membres du Laboratoire de biologie informatique et théorique (LBIT), en particulier Guylaine Poisson, Nancy Bourassa, Sébastien Lemieux, Patrick Gendron et Dominic Lambert, en compagnie de qui j'ai passé quatre agréables années. Je dois aussi des remerciements particuliers à Laurent David et à Yoshua Bengio pour leurs commentaires toujours pertinents. J'aimerais finalement remercier mon directeur de recherche, François Major, ainsi que le Conseil de recherches en sciences naturelles et en génie (CRSNG) pour leur généreux appui financier.