

THESE DE DOCTORAT
DE L'UNIVERSITE LIBRE DE BRUXELLES

ULB

présentée par

Anne PAJON

pour obtenir le grade de DOCTEUR en SCIENCES

SUJET DE LA THÈSE :

« Création et analyse de Pescador, une base de données de peptides en solution et analyse de motifs structuraux de protéines et de leurs caractéristiques expérimentales. »

soutenance privée le mardi 21 mai 2002

soutenance publique le vendredi 21 juin 2002

DEVANT LE JURY COMPOSÉ DE :

Président : Victor STALON
Secrétaire : Georges HUEZ
Promoteur : Shoshana WODAK
Rapporteur : Vincent RAUSSENS

et

Marc COLET
Johan WOUTERS
Manuel RICO

Je tiens à remercier en premier lieu Shoshana WODAK de m'avoir acceptée au sein de son laboratoire « Service de Conformation des Macromolécules Biologiques ». *Je vous remercie de la confiance que vous m'avez témoignée tout au long de ces quatre années.*

Un très grand merci à Wim VRANKEN. *Merci d'abord d'avoir lu et relu ce manuscrit. Merci pour les nombreuses discussions scientifiques, merci pour l'année passée à travailler à tes côtés, merci pour ta disponibilité et ta gentillesse. Ces quelques lignes ne sont bien sûr pas à la mesure de l'aide apportée. Sincèrement, merci.*

Je remercie Marc COLET, Georges HUEZ, Victor STALON, Vincent RAUSSENS, Manuel RICO et Johan WOUTERS pour avoir accepté de faire partie de mon jury.

Je souhaite remercier Christian LEMER pour ses conseils avisés en informatique qui ont contribué à l'orientation et à la réalisation de ce travail.

J'exprime ici mes remerciements à Jean RICHELLE pour son aide et sa disponibilité.

Un merci tout particulier à Stéphanie HÉRY pour son amitié et pour les nombreuses discussions qui ont su ensoleiller certaines journées difficiles.

Merci à Vincent DEVLOO pour la bonne ambiance des déjeuners et des pauses café passés à refaire le monde.

Une grande reconnaissance à Fernanda SIROTA LEITE pour sa gentillesse et pour avoir partagé mon quotidien ces derniers mois.

Je n'oublie pas Nathalie LEFÈVRE pour sa gentillesse et sa disponibilité.

Je tiens également à remercier tous ceux que j'ai eu l'occasion de rencontrer durant cette thèse au sein du laboratoire : Isabel COUTINHO, Didier CROES, Ujjwal DAS, Leonardo DE MARIA, Benoît DESSAILLY, Mohammed ENNASAR, Cédric GOVAERTS, Alfonso JARAMILLO, Raphael LEPLAE, Raul MENDEZ, Maria-Elena OCHAGAVIA ROQUE, Koji OGATA, Ernesto PEREZ RUEDA, Nicolas SIMONIS, Isabel TOMÀS OLIVEIRA, Ariane TOUSSAINT, Daniel VAN BELLE et Jacques VAN HELDEN.

Puis je remercie tout particulièrement Françoise VOVELLE, sans qui je ne serais pas là aujourd'hui. *Cette thèse est aussi un peu la vôtre. Merci d'avoir guidé mes premiers pas dans le monde scientifique et plus particulièrement dans celui des structures de protéines.*

Un chaleureux merci à mon amie de longue date, Sandra JAVOY, pour tous les bons moments passés, présents et à venir. *Merci d'avoir accompagné, encouragé et égayé mes années à la fac et celles qui ont suivies.*

Que Vincent MITTERRAND trouve ici le témoignage de toute ma reconnaissance pour le soutien qu'il m'a apporté durant ces quatre années. *Un grand merci aussi pour avoir relu avec tellement d'attention ce manuscrit. Merci d'avoir cru en moi.*

Un grand merci à mes parents, Jeannine et Jacques PAJON, et à mes grands-parents, Anne-Marie et Roland PAJON, pour la relecture attentive de ce manuscrit, pour leur soutien moral et matériel ainsi que pour leurs sincères encouragements.

A Vincent...

Résumé

Ces dernières années, un nombre important de données a été obtenu à partir d'expériences de Résonance Magnétique Nucléaire et de Dichroïsme Circulaire, sur l'influence de la séquence en acides aminés ainsi que sur l'influence d'autres types de paramètres sur la conformation des peptides en solution. Au regard du nombre de ces données, nous avons entrepris le développement de Pescador 'PEptides in Solution ConformAtion Database : Online Resource', une base de données de peptides en solution, afin de collecter ces données dans le but de les analyser. Pescador contient des données de RMN et de Dichroïsme Circulaire sur des peptides en solution, ainsi que des paramètres structuraux dérivés de ces données expérimentales. Un système de collecte des données par l'intermédiaire d'une interface web a été réalisé ainsi qu'un accès aux données stockées. Pour illustrer l'intérêt d'une telle base de données dans l'analyse des conformations des peptides, nous présentons une analyse des déplacements chimiques des protons alpha stockés dans Pescador, obtenus par RMN pour différents peptides, et provenant de différents laboratoires. A partir de cette analyse, nous proposons de nouvelles valeurs de déplacements chimiques pour les conformations non structurées, valeurs de références pour l'étude de la conformation de peptides en solution. Tout d'abord, nous montrons que ces valeurs sont similaires à celles obtenues expérimentalement sur des peptides modèles, et que leur variation en fonction de la concentration de trifluoroéthanol est similaire à celle reportée dans la littérature. Nous présentons aussi que les valeurs des déplacements chimiques sont utilisés pour dériver des facteurs de correction qui prennent en compte l'effet des résidus voisins. Ces facteurs sont comparables à ceux récemment dérivés à partir d'une série de peptides GGXGG [51]. Ces résultats encourageants suggèrent que plus le nombre de données déposées dans Pescador sera important, plus les analyses de ces données offriront des paramètres clés significatifs pour l'analyse de la conformation des peptides.

La deuxième partie de ce travail réside dans l'analyse de motifs structuraux de protéines et de leurs caractéristiques expérimentales. Nous présentons une méthode de classification permettant de passer des 97 protéines obtenues par RMN à des groupes de motifs structuraux. L'analyse de ces motifs et de leurs caractéristiques expérimentales met en évidence des particularités spécifiques en terme de patterns de contraintes nOe. La confirmation des observations faites dans le cadre de cette analyse avec des informations issues de la littérature, permet d'offrir des perspectives intéressantes à ce type d'étude.

Table des matières

<i>Abréviations</i>	3
Introduction générale	4
I Pescador : « The PEptides in Solution ConformAtion Database : Online Resource. »	6
1 Introduction	7
2 Méthodes	9
2.1 La Résonance Magnétique Nucléaire	9
2.1.1 Les déplacements chimiques	10
2.1.2 Valeurs de référence de Merutka et al.	11
2.1.3 Facteurs de correction de Schwarzinger et al.	12
2.2 Statistiques utilisées	13
2.2.1 Médiane et intervalle interquartile	14
2.2.2 Histogramme	15
2.2.3 Tests de normalité	15
3 Description de la base de données	17
3.1 Introduction	17
3.2 Schéma et organisation	17
3.3 Intégrité	20
3.4 L'interface web : l'outil de dépôt des données	22
3.5 Le traitement des données	25
3.6 L'acquisition des données	25
3.7 L'extraction et l'analyse des données	26
4 Analyses de Pescador	28
4.1 Introduction	28
4.2 Vue générale des données disponibles	29
4.3 Définition du sous-ensemble restreint de données	30
4.4 Les déplacements chimiques : de nouvelles valeurs de référence	32
4.5 Les facteurs de correction des valeurs δ des protons alpha	36
4.6 L'influence des résidus voisins sur les déplacements chimiques	37
4.7 Application des facteurs de correction	41
4.8 L'influence du TFE sur les δ des protons amides	43

5 Discussion	46
II Étude de motifs structuraux récurrents de protéines en relation avec leurs empreintes expérimentales	49
1 Introduction	50
2 Description de la méthode de clustering	54
2.1 Introduction	54
2.2 Clustering et sélection des modèles représentatifs	54
2.3 Familles ayant des signatures de Ramachandran identiques	57
2.4 Clustering hiérarchique	58
3 Aspects informatiques et outils développés	61
3.1 Introduction	61
3.2 Description du module décrivant une molécule	62
3.3 L'extraction des données	62
3.4 Le clustering des données	64
3.5 L'analyse des patterns de contraintes nOe	65
4 Analyses des motifs structuraux et de leurs caractéristiques expérimentales	66
4.1 Introduction	66
4.2 Analyse générale des résultats	66
4.3 Les différents motifs obtenus	70
4.4 Les patterns de contraintes nOe dans les structures régulières	72
4.4.1 L'hélice α	72
4.4.2 Le brin β	74
4.5 Analyse détaillée des motifs en tournant	77
4.5.1 motifs α - α	77
4.5.2 motifs β - β	78
4.5.3 motifs α - β	79
4.5.4 motifs β - α	81
4.6 Qualité locale des structures	97
4.7 Améliorations de la méthode et perspectives d'analyses	97
5 Discussion	101
Conclusion générale	104
III Annexes	106
1 Matériels et Méthodes	107
1.1 Zones favorables de la carte de Ramachandran	107
1.2 Les valeurs des déplacements chimiques des trois peptides étudiés	108
1.3 Liste des 97 protéines	112
2 Publication	114
Bibliographie	119

Abréviations

BMRB	BioMagResBank
HTML	HyperText Markup Language
nOe	nuclear Overhauser effect
PDB	Protein Data Bank
RMN	Résonance Magnétique Nucléaire
rmsd	root mean-square deviation
ppm	parties par million
SQL	Structured Query Language
TFE	TriFluoroEthanol
NMR-STAR	Nuclear Magnetic Resonance Self-Defining Text Archival and Retrieval

Introduction générale

Depuis plus de vingt ans, l'informatique s'est révélé être un apport fondamental pour la biologie moléculaire. Ce domaine, qu'est la bioinformatique, a fait son apparition dans les années 1980 avec les premières banques de biomolécules (EMBL [55] et GenBank [5]). Elle propose en effet des méthodes et des logiciels qui permettent de gérer, d'organiser, de comparer, d'analyser, et d'explorer l'information génétique et génomique stockée dans les bases de données. Grâce à la bioinformatique, il est ainsi possible de prédire et de produire des connaissances nouvelles ainsi que d'élaborer de nouveaux concepts.

La compilation et l'organisation des données est un des principaux aspects de ce domaine notamment avec la création de bases de données. Elles sont généralement construites autour de thèmes précis comme celui des informations expérimentales provenant de la Résonance Magnétique Nucléaire, ou celui des structures tridimensionnelles des biomolécules. La plus connue est bien entendu celle de la Protein Data Bank [6] dans laquelle aujourd'hui plus de 2600 structures déterminées par RMN sont disponibles. La base de données BioMagResBank [52] (BMRB) contient quant à elle essentiellement les valeurs des déplacements chimiques de protéines.

Incontestablement, toutes ces banques de données constituent une source de connaissance d'une grande richesse, exploitable dans le développement de méthodes d'analyse ou de prédiction. En effet, l'importance des données expérimentales structurales comme celles provenant de la Résonance Magnétique Nucléaire peuvent nous permettre de déduire des informations importantes allant des statistiques sur les valeurs des déplacements chimiques à la structure tridimensionnelle des macromolécules biologiques.

L'évaluation des différentes approches dans le but de validation donne naissance à un ensemble d'outils, d'études ou de méthodes qui convergent vers un but commun.

Ce travail se décompose en deux parties et présente deux aspects de l'analyse de données biologiques.

Dans la première partie, la base de données de peptides en solution, Pescador, est présentée et analysée. Cette étude porte sur l'analyse de données expérimentales issues d'expériences RMN stockées dans la base données et sur l'influence de certaines conditions expérimentales sur ces données.

Dans la seconde partie, nous allons montrer une méthode de classification de structures de protéines obtenues par RMN afin d'obtenir des groupes de motifs structuraux. Les analyses des caractéristiques expérimentales, comme les contraintes nOe, ont permis de mettre en évidence des propriétés communes à ces motifs.

Première partie

**Pescador : « The PEptides in
Solution Conformation Database :
Online Resource. »**

Chapitre 1

Introduction

Depuis plus d'une dizaine d'années, un grand nombre d'études a porté sur l'analyse des peptides, de leurs préférences conformationnelles et de leurs interactions en solution. Beaucoup de ces peptides ont été construits dans le but d'élucider les facteurs gouvernant la formation des structures secondaires. Une partie importante de ce travail a été consacrée à la formation des hélices α et à leurs équilibres conformationnels [8, 46, 4, 43]. Avec l'intérêt grandissant du phénomène d'agrégation, et des transitions conformationnelles de structures en hélices vers des structures étendues [32, 33, 1, 56], le centre d'intérêt s'est déplacé vers la formation des épingles à cheveux β et des feuillets β [16, 17, 29, 44, 18, 49, 66, 50]. D'autres études sur des peptides ont été réalisées afin de mettre en évidence la dépendance entre la séquence locale en acides aminés et des paramètres de RMN, comme les déplacements chimiques [7, 35, 61, 51] et les constantes de couplages [36, 19]. Ces paramètres fournissent des informations sur la structure secondaire, les populations relatives, ainsi que l'influence de la séquence locale.

Ces données de Résonance Magnétique Nucléaire et de Dichroïsme Circulaire sur les peptides sont nombreuses. Elles proviennent de différents laboratoires de recherche où elles ont été étudiées, mais elles ne sont malheureusement pas disponibles au public. De plus, une très faible proportion d'entre elles sont publiées dans les journaux scientifiques.

Dans la première partie de ce travail, nous allons décrire comment les données que nous avons étudiées sont obtenues lors d'expériences de Résonance Magnétique Nucléaire. Nous

présentons aussi l'étude faite par Schwarzinger et al. [51] sur les facteurs de correction. Une brève présentation des statistiques utilisées pour l'analyse des données de Pescador termine le chapitre introductif des méthodes.

Nous allons ensuite expliquer la construction proprement dite de Pescador, une base de données de peptides en solution. Puis, nous allons décrire l'outil de déposition des données et la procédure de traitement de celles-ci.

Nous illustrerons ensuite comment Pescador peut être utilisé pour dériver des informations utiles pour l'analyse des conformations de peptides. Pour cela, nous allons présenter une analyse des déplacements chimiques des protons alpha mesurés par Résonance Magnétique Nucléaire. Ce type de données, dont l'abondance est importante pour les peptides, est disponible pour pratiquement tous les enregistrements de Pescador, ce qui n'est pas le cas des autres types de données. De plus, de nombreuses analyses sur les peptides et leurs conformations se sont concentrées sur les déplacements chimiques des protons. Ces valeurs sont donc un excellent point de départ pour évaluer l'intérêt de Pescador.

Chapitre 2

Méthodes

2.1 La Résonance Magnétique Nucléaire

La Résonance Magnétique Nucléaire ou RMN en milieu liquide est une technique utilisée pour l'analyse des structures de nombreuses molécules chimiques. Elle sert principalement à la détermination structurale des composés organiques. Les principaux noyaux étudiés sont le proton 1H , le carbone ^{13}C et le phosphore ^{31}P . La méthode repose sur le magnétisme nucléaire. Les noyaux de certains atomes (1H , ^{13}C , ...) possèdent un moment magnétique nucléaire, c'est à dire qu'ils se comportent comme des aimants microscopiques caractérisés par une grandeur quantique : le spin.

Les deux principaux paramètres de la RMN abordés dans ce travail sont les suivants :

- Les déplacements chimiques qui reflètent l'environnement électronique du proton sont les données que nous avons analysées au cours de cette première partie sur la base de données Pescador.
- Les effets Overhauser nucléaires (nOe), qui reflètent les phénomènes de relaxation dipolaire entre les paires de protons, donnent les informations essentielles pour la détermination de la structure des protéines. Ces effets ont été étudiés dans la deuxième partie de ce travail portant sur l'étude des motifs structuraux récurrents de protéines.

2.1.1 Les déplacements chimiques

Le champ B_0 appliqué extérieurement à l'échantillon ne présente pas nécessairement la même valeur au niveau des différents noyaux et de leurs spins qui appartiennent à une même molécule. Le nuage électronique local peut apporter un effet perturbateur qui se traduit, au niveau du noyau, par une valeur du champ légèrement différente de B_0 et qui s'écrit $B_0(1 - \sigma)$ où σ est appelé constante d'écran. Elle est positive lorsque ce phénomène a pour origine principale la précession du moment cinétique orbital qui est associé au mouvement de rotation du nuage électronique par rapport à B_0 . Le champ magnétique induit par cette précession est de signe opposé à B_0 . Il s'ensuit que la fréquence de résonance n'est plus exactement égale à $\nu = \frac{\gamma}{2\pi}B_0$ mais $\nu = \frac{\gamma}{2\pi}B_0(1 - \sigma) = \nu_0(1 - \sigma)$. Cela revient à corriger le rapport gyromagnétique γ d'un facteur $(1 - \sigma)$ et conduit à une différenciation des fréquences de résonance en fonction de l'environnement électronique du noyau considéré, donc de la nature du groupement chimique auquel il appartient. Cet effet est connu sous le nom de déplacement chimique. Il s'agit d'un effet fin, nombre sans dimension, de l'ordre de 10^{-6} qui ne peut être observé qu'au moyen d'un champ B_0 aussi homogène que possible sur tout le volume utile de l'échantillon.

L'effet de déplacement chimique étant décelable grâce aux corrections d'inhomogénéité, il convient de définir une procédure pour le caractériser. La mesure absolue de ν_0 avec une précision meilleure que 10^{-6} étant impossible, on préfère mesurer ν par rapport à la fréquence de résonance d'une substance de référence ν_R . En outre, pour permettre une comparaison plus simple entre des spectres enregistrés à l'aide de spectromètres opérant à des valeurs différentes de B_0 , on a recours à une échelle delta δ , qui s'exprime en ppm (parties par million) et qui est définie par la relation :

$$\delta_{ppm} = \frac{\nu - \nu_R}{\nu_0} 10^6 \simeq \frac{\nu - \nu_R}{\nu_R} 10^6 \quad (2.1)$$

$$\delta_{ppm} = (\sigma_R - \sigma) 10^6 \quad (2.2)$$

L'échelle δ représente donc bien une mesure relative du coefficient d'écran, indépendante de B_0 . Le facteur 10^6 permet de travailler avec des nombres de l'ordre de l'unité, de

la dizaine ou de la centaine. On choisit généralement comme référence une substance conduisant à un pic unique, facilement soluble et peu sensible aux effets de solvant, et résonant à une extrémité du spectre. Pour la spectroscopie du proton et du carbone-13, le TMS (tétraméthylsilane : $\text{Si}(\text{CH}_3)_4$) est communément utilisé.

Outre les informations structurales de nature intramoléculaire qu'il fournit, le déplacement chimique est sensible à certains effets de nature intermoléculaire, comme la proximité d'un cycle aromatique ou le pH. L'attribution des déplacements chimiques à chaque proton de la protéine constitue l'étape préliminaire incontournable à toute étude structurale.

2.1.2 Valeurs de référence de Merutka et al.

Nous présentons les valeurs de référence issues de la littérature que nous avons utilisées. Ces valeurs de référence pour des conformations aléatoires sont obtenues à partir d'une série de petits peptides composés de glycines. Elles sont utilisées pour déterminer les conformations du squelette dans les peptides et les protéines. Les valeurs des déplacements chimiques des noyaux du squelette observées sont alors comparées à ces valeurs de référence suivant la formule :

$$\Delta\delta = \delta_{\text{observé}} - \delta_{\text{référence}} \quad (2.3)$$

Ces valeurs $\Delta\delta$ servent à décrire la structure secondaire adoptée par le peptide ou le segment de protéine [63]. Une valeur négative indique plutôt une préférence pour une conformation en hélice α et une positive indique une préférence pour une conformation en feuillet β .

Les valeurs de référence (tableau 2.1 page 12) proposées par Merutka et al. sont basées sur une série de peptides linéaires H-Gly-Gly-X-Gly-Gly-OH, où X est l'un des 20 acides aminés naturels. Les valeurs δ des protons alpha ont été obtenues dans un solvant 90% $\text{H}_2\text{O}/10\% \text{D}_2\text{O}$, à 277.2K et à pH 5.0. Ces valeurs de référence, obtenues à partir d'une série de peptides ayant des conformations aléatoires, sont aussi appelées valeurs 'random coil'.

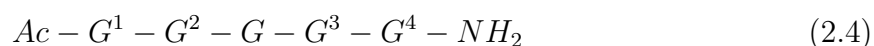
Résidu	$\delta_{H\alpha}^{réf.}$
ALA	4.34
ARG	4.34
ASN	4.76
ASP	4.63
CYS	4.58
GLN	4.36
GLU	4.29
GLY	4.01
HIS	4.77
ILE	4.18
LEU	4.35
LYS	4.32
MET	4.52
PHE	4.62
PRO	4.44
SER	4.49
THR	4.39
TRP	4.67
TYR	4.56
VAL	4.13

Tab. 2.1: Valeurs des déplacements chimiques $\delta_{H\alpha}$ des protons alpha pour chaque résidu X obtenues à partir de la série de peptides GGXGG.

2.1.3 Facteurs de correction de Schwarzinger et al.

Une analyse de l'effet des résidus voisins a été récemment publiée par Schwarzinger et al. [51]. Elle a examiné une série de peptides GGXGG dans le but d'obtenir des facteurs de correction dépendant de la séquence pour les déplacements chimiques utilisés comme référence et basés sur des peptides en conformation aléatoire.

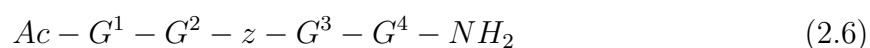
Les facteurs de correction sont calculés à partir d'un peptide de référence ayant pour séquence :



Le facteur de correction A du déplacement chimique pour un résidu de type z est obtenu comme suit :

$$A[z] = \delta(G^1) - \delta(G_{réf.}^1) \quad (2.5)$$

$\delta(G^1)$ est la valeur δ observée pour le résidu G^1 dans la séquence du peptide :



$\delta(G_{réf.}^1)$ est la valeur correspondante pour le résidu G^1 dans le peptide de référence ayant pour séquence 2.4. Les trois autres facteurs de correction sont calculés de la même manière et les résultats sont présentés dans le tableau 2.2 (page 14). Les facteurs de correction permettent de compenser l'effet de la séquence sur les valeurs de référence d'une série de peptides de conformations aléatoires telles que celles de Merutka et al. (tableau 2.1 page 12) évoquées au paragraphe précédent. Cette correction est définie par l'équation :

$$\begin{aligned}\delta_{R\text{corrigé}} &= \delta_{Rréf.} + \Delta\delta_{R-1} + \Delta\delta_{R+1} + \Delta\delta_{R-2} + \Delta\delta_{R+2} \\ &= \delta_{Rréf.} + C_{[z=R-1]} + B_{[z=R+1]} + D_{[z=R-2]} + A_{[z=R+2]}\end{aligned}\quad (2.7)$$

où le résidu R se trouve dans cette sous séquence :

$$\dots - R_{-2} - R_{-1} - R - R_{+1} - R_{+2} - \dots \quad (2.8)$$

Afin de mettre en évidence l'application des facteurs de correction, prenons comme première approximation seulement le tripeptide ayant comme séquence $R_{-1} - R - R_{+1}$. Le déplacement chimique du résidu en question R doit être corrigé par le facteur de correction B correspondant au résidu R_{+1} . Ce facteur est obtenu en déterminant l'effet de R_{+1} sur le déplacement chimique de G^2 dans le peptide $Ac - G^1 - G^2 - R_{+1} - G^3 - G^4 - NH_2$. De façon identique, le $\delta_{Rréf.}$ du résidu R doit être corrigé par le facteur de correction C correspondant au résidu R_{-1} , qui est obtenu en déterminant l'effet de R_{-1} sur le δ de G^3 dans le peptide $Ac - G^1 - G^2 - R_{-1} - G^3 - G^4 - NH_2$.

Les valeurs des facteurs de correction issues de cette étude sont présentées dans le tableau 2.2 (page 14). Seules les valeurs supérieures ou égales à 0.08 ppm sont prises en compte lors de l'application de ces facteurs de correction.

2.2 Statistiques utilisées

Lors de l'analyse des données stockées dans Pescador, nous avons utilisé le programme R, qui est à la fois un environnement et un langage de script pour les statistiques et la création de graphiques [22]. Nous avons implémenté plusieurs scripts en langage R dont voici les principales composantes utilisées.

z	A	B	C	D
ALA	-0.02	-0.03	-0.03	0.00
ARG	-0.02	-0.02	-0.02	0.00
ASN	-0.01	-0.01	-0.02	-0.01
ASP	-0.02	-0.01	-0.02	-0.01
CYS	-0.01	0.02	0.00	0.00
GLN	-0.01	-0.02	-0.01	0.00
GLU	-0.02	-0.02	-0.02	0.00
GLY	0	0	0	0
HIS	-0.03	-0.06	0.01	0.01
ILE	-0.03	-0.02	-0.02	-0.01
LEU	-0.04	-0.03	-0.05	-0.01
LYS	-0.02	-0.02	-0.01	0.00
MET	-0.02	-0.01	-0.01	0.00
PHE	-0.06	-0.09	-0.08	-0.04
PRO	-0.01	0.11	-0.03	-0.01
SER	-0.01	0.02	0.00	-0.01
THR	-0.01	0.05	0.00	-0.01
TRP	-0.08	-0.10	-0.15	-0.16
TYR	-0.05	-0.10	-0.08	-0.04
VAL	-0.02	-0.01	-0.02	-0.01
Moyenne	-0.03	-0.02	-0.03	-0.02

Tab. 2.2: Facteurs de correction de Schwarzinger et al. [51]. Les facteurs de correction supérieurs à 0.08 ppm sont en **gras**.

2.2.1 Médiane et intervalle interquartile

Par définition, la médiane \tilde{x} est la valeur correspondant au milieu de la fonction de répartition d'une variable aléatoire :

$$\tilde{x} : \int_{-\infty}^{\tilde{x}} dF(x) = \frac{1}{2} \quad (2.9)$$

La médiane est un indicateur insensible aux valeurs extrêmes ce qui en fait un outil très intéressant dans le domaine des statistiques robustes.

L'intervalle interquartile est l'étendue de la distribution sur laquelle se trouve concentrée la moitié des éléments dont les valeurs sont les moins différentes de la médiane. On exclut alors de la distribution les 25% des valeurs les plus faibles et les 25% des valeurs les plus fortes.

2.2.2 Histogramme

L'histogramme est analogue à la courbe de densité. L'ordonnée associée à chaque abscisse est égale à la fréquence d'apparition de la valeur dans l'échantillon. Dans le cas d'une variable aléatoire discrète, la construction de l'histogramme ne pose pas de problème. Par contre, pour une variable aléatoire continue, il est nécessaire de résumer les valeurs à reporter sur la courbe en classes.

La détermination du nombre de classes d'un histogramme est délicate et il n'existe pas de règle absolue. Un trop faible nombre de classes fait perdre de l'information et aboutit à gommer les différences pouvant exister entre des groupes de l'ensemble étudié. En revanche, un trop grand nombre de classes aboutit à des graphiques incohérents où certaines classes deviennent vides ou presque car n , la taille de l'échantillon, est un nombre fini.

2.2.3 Tests de normalité

Deux tests ont été utilisés afin de déterminer si la distribution obtenue suit ou non une distribution normale. Dans le but de déterminer s'il était possible ou non d'appliquer une étude statistique à ces données pour les analyser.

Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov [10] est utilisé pour déterminer si un ensemble de données est consistant avec une fonction de distribution spécifique, mais pas seulement la distribution normale. Comme beaucoup de tests généraux, il n'est pas aussi puissant qu'un test spécifiquement construit pour tester la normalité. Cependant il est utilisé ici puisqu'il ne requiert aucune condition particulière, c'est un test non paramétrique. L'interprétation des résultats obtenus par le logiciel R est basé sur la valeur p obtenue à la suite du calcul du test. Cette valeur est la probabilité avec laquelle l'hypothèse nulle est rejetée, plus elle est petite ($p < 0.05$) et plus le rejet de l'hypothèse est convaincante. Si l'hypothèse nulle est rejetée, alors la population étudiée ne représente pas une distribution normale.

Test de Shapiro-Wilk

Le test de Shapiro-Wilk [48] est spécifique pour tester si une distribution est normale ou non, sans nécessiter au départ la valeur de la moyenne ou de la variance de la distribution étudiée. Ce test n'indique pas le type de distribution non normale obtenue. L'hypothèse nulle est aussi rejetée lorsque la valeur p est petite et dans ce cas, la population étudiée ne correspond pas à une distribution normale.

Chapitre 3

Description de la base de données

3.1 Introduction

Dans ce chapitre, nous allons décrire comment nous avons organisé et assemblé les informations dans une base de données dédiée, Pescador : 'PEptides in Solution Conformation Database : Online Resource'. Cette base de données contient des informations provenant d'expériences de Résonance Magnétique Nucléaire et de Dichroïsme Circulaire effectuées sur des peptides en solution, ainsi que des informations structurales dérivant des données expérimentales précédentes.

Dans cette partie, nous allons donc présenter une description détaillée de la structure et de l'organisation de la base de données Pescador. Nous allons aussi décrire l'outil de déposition des données et la procédure de traitement de celles-ci.

Cette base de données est accessible sur internet à l'adresse suivante : <http://www.ucmb.ulb.ac.be/Pescador/>.

3.2 Schéma et organisation

Dans la base de données Pescador, chaque entrée représente une expérience possédant un identifiant unique (section 'experiment', figure 3.1 page 19). L'expérience est caractérisée par un ensemble de conditions expérimentales (section 'experimental conditions') contenant les valeurs du pH, de la température,... (table 'Experimental_conditions'), la liste des solvants ('Solvents' et 'Exp2Solvent') ainsi que

des composants additionnels ('AddComponents' et 'Exp2AddComponents'). Autour de chaque expérience sont associées quatre autres sections :

- Section 1 : la première section, 'global information', contient une description de l'origine et de la provenance des données (références bibliographiques, table 'Refs' et informations sur le laboratoire, table 'Labs').
- Section 2 : 'peptide data' est constituée des informations nécessaires à la caractérisation du peptide sur le plan chimique. La séquence en acides aminés ('PeptideResidues') est présente ainsi que la liste des liaisons peptidiques 'cis' ('CisResidues') et des liaisons covalentes additionnelles ('SupplementaryCovBonds').
- Section 3 : les données expérimentales provenant soit de la RMN, soit du Dichroïsme Circulaire se situent dans la troisième section 'experimental data'. Elles sont mesurées à des conditions expérimentales spécifiques décrites dans la section associée 'experimental conditions'. En fonction de leur type, les valeurs sont stockées pour chaque peptide, chaque résidu ou pour chaque atome. La table 'Chemical-ShiftValues' contient les valeurs des déplacements chimiques, 'CCMethod' et 'CC-Values' les informations sur les constantes de couplage, 'HDEExchange' les valeurs d'expérience d'échange hydrogène/deutérium, 'NH_TemperatureCoeff' les coefficients de température des déplacements chimiques des protons amides, 'NOEs' les valeurs des contraintes nOe et enfin 'CDEperiment' les données expérimentales de Dichroïsme Circulaire.
- Section 4 : la dernière section 'structure data' de la base de données est composée d'informations sur la conformation du peptide comme la population globale ('Global_population'), les structures secondaires ('QualStructAnalValues') et les liaisons hydrogènes ('HBondsValues') obtenues à partir des données expérimentales ainsi que les valeurs des angles ('AverageAnglesValues') et les structures tridimensionnelles ('CalculatedStructures').

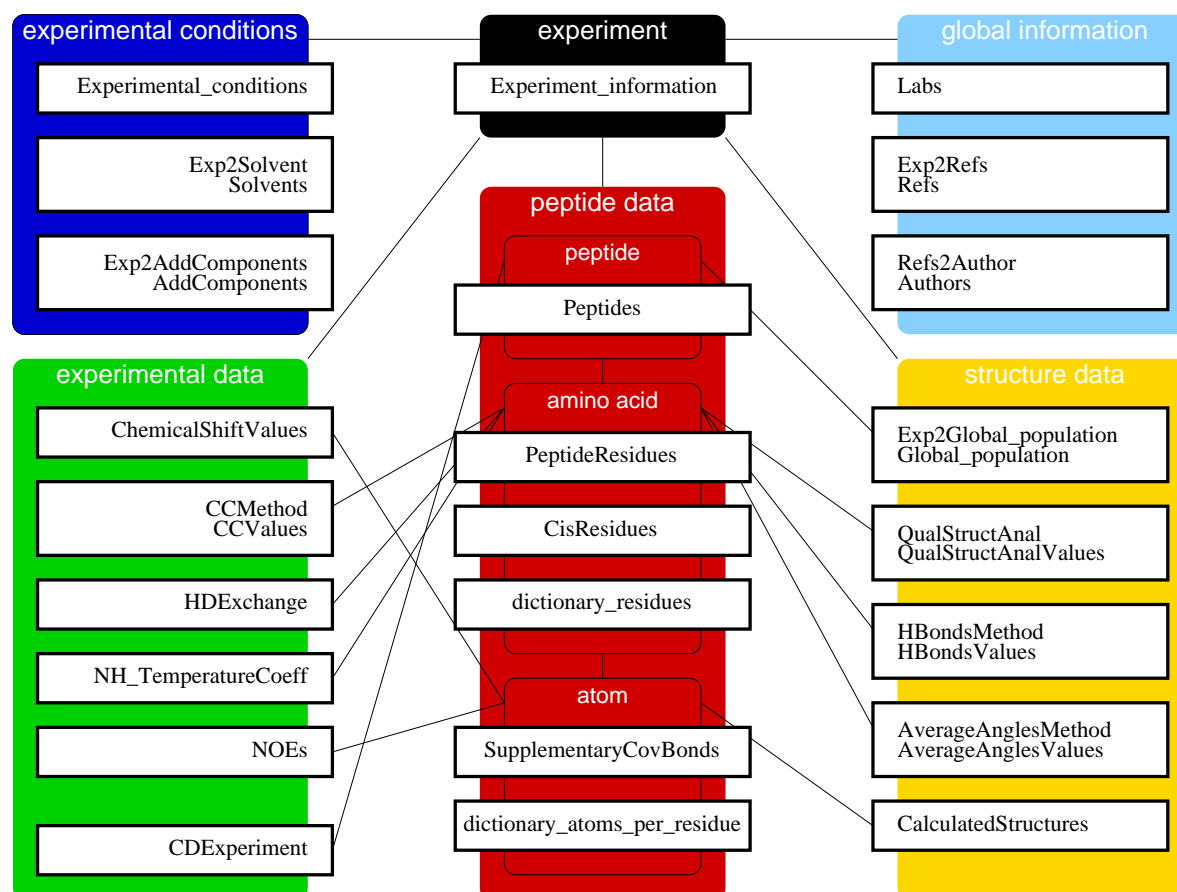


Fig. 3.1: Schéma de la base de données Pescador. Chaque rectangle en gras représente une ou plusieurs tables ayant certaines propriétés ou attributs. Les relations entre ces tables sont représentées par des lignes. Les rectangles aux coins arrondis représentent des sections : groupement virtuel de tables associées par le type des données qu'elles contiennent. Les cinq sections sont centrées autour de la section principale 'experiment' contenant la table 'Experiment_information'.

La base de données Pescador est constituée d'un total de 40 tables et a été implémentée sous SYBASE.

Un exemple concret d'une entrée de Pescador (figure 3.2 page 21) donne une illustration de l'organisation des tables dans la base de données. La table 'Experiment_information' (figure 3.2, au centre) contient une liste d'identifiants uniques. Pour cette entrée, il correspond à 'ExpID=81'. Celui pour la table 'Experimental_conditions' est 'CondID=18', pour la table 'Peptides' est 'PeptID=57' et enfin celui pour la table 'Labs' est 'LabID=3'.

Les informations sur le laboratoire ayant produit les données sont stockées dans la table 'Labs' ayant comme identifiant unique 'LabID=3' (figure 3.2, en haut à droite).

Les données sur la structure chimique du peptide se trouvent stockées dans la section 'peptide data'. La séquence de chaque peptide est archivée dans la table 'Peptides'. Dans 'PeptideResidues' se trouve la liste des résidus constitutifs de la séquence. Cette dernière table permet de créer un lien entre les résidus de la séquence et les données expérimentales entrées au niveau des résidus.

La section des conditions expérimentales est constituée de trois tables. Celle du haut ('Experimental_conditions') contient les paramètres expérimentaux. Les champs ont ici des valeurs uniques, il n'est donc accepté qu'une seule valeur pour chaque paramètre. Par ailleurs, deux tables sont nécessaires pour stocker les valeurs des solvants utilisés pour chaque expérience. En effet, plusieurs solvants peuvent correspondre à une expérience donnée, comme dans la table 'Exp2Solvent' de la figure 3.2 (page 21), et ces solvants peuvent avoir des proportions différentes, comme dans la table 'Solvents' (H₂O 90% et D₂O 10%). L'utilisation de deux tables pour permettre l'entrée de données multiples est aussi nécessaire dans le cadre des données stockées pour les populations de conformation. Ces données sont, comme précédemment stockées dans deux tables distinctes : la première contenant le lien entre l'expérience et les différentes populations, et la seconde contenant la liste des valeurs des populations obtenues par une méthode expérimentale. Dans l'exemple de la figure 3.2 apparaissent les déplacements chimiques des protons alpha.

La section des données expérimentales ne contient que la table des déplacements chimiques 'ChemicalShiftValues'. Elle contient donc les valeurs des déplacements chimiques de la résonance des noyaux. La figure 3.2 (page 21) nous montre le type de données le plus souvent mesuré pour des peptides : les valeurs δ des protons alpha.

3.3 Intégrité

Afin d'assurer la validité des données dans la base de données Pescador, un certain nombre de contraintes ont été ajoutées. Elles sont de trois types :

- La première est la clé primaire de la table. Elle possède un caractère obligatoire et est l'identifiant unique pour chacune des tables (figure 3.3 page 22, rectangle gris).

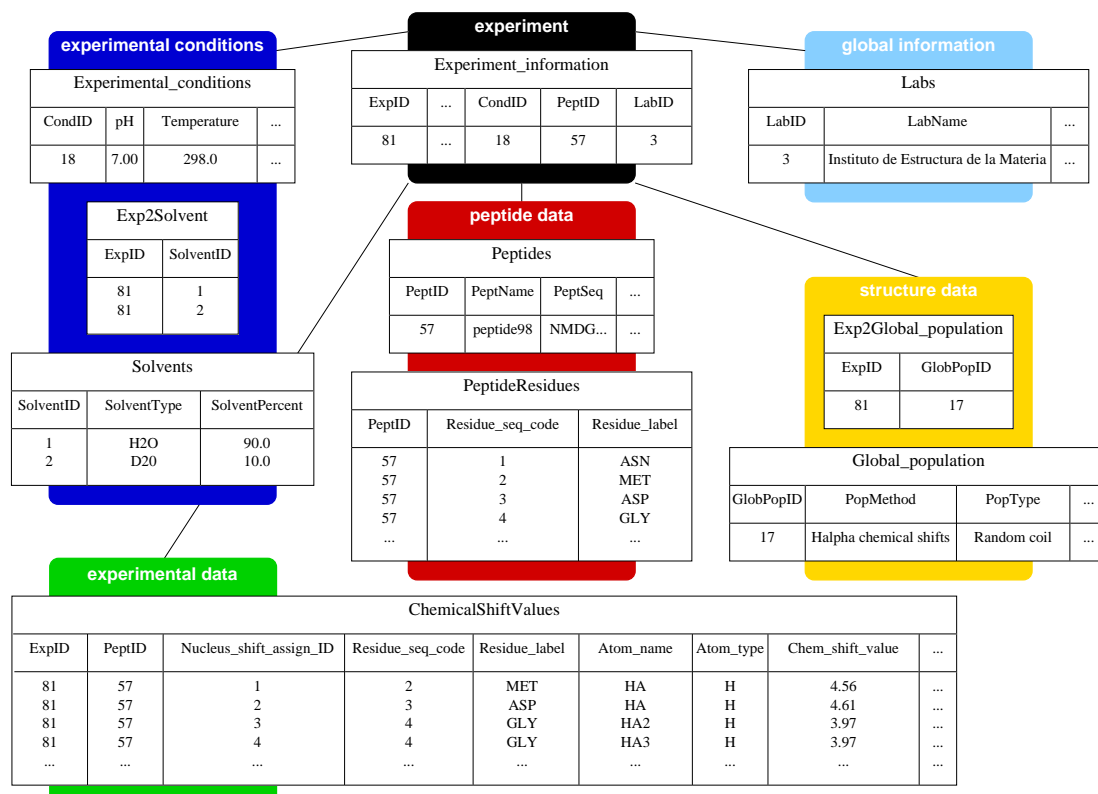


Fig. 3.2: Exemple d'une entrée de la base de données Pescador.

- La seconde est la clé étrangère. Elle introduit un lien entre deux tables où les champs choisis de chaque ligne de la première table correspondent à ceux de la table cible (figure 3.3, lignes pointillées).
- La dernière contrainte est d'un autre type puisqu'il s'agit d'imposer un ensemble de valeurs acceptables pour le champ d'une table. Pour assurer la valeur d'entrée d'un paramètre la valeur devra être comprise dans un intervalle autorisé. Prenons par exemple le cas du pH : cet intervalle sera alors [1, 14].

Afin d'illustrer l'importance de ces contraintes, la figure 3.3 (page 22) présente un exemple construit autour de la table des déplacements chimiques 'ChemicalShiftValues'. Les entrées de cette table sont les valeurs des déplacements chimiques des noyaux de chaque expérience. La clé primaire de cette table est le couple des champs 'ExpID' et 'Nucleus_shift_assign_ID', les valeurs de ces deux champs sont donc différentes à chaque ligne de la table (voir aussi la figure 3.2). Il y a également trois clés étrangères, définissant les liens entre la table 'ChemicalShiftValues' et les tables 'Experiment_information', 'PeptideResidues' et 'dictionary_atoms_per_residue'. La

clé étrangère constituée du couple ('ExpID', 'PeptID') crée un lien vers la table cible 'Experiment_information'. La présence de cette clé assure le bon ordre de saisie des différentes données : les données relatives aux déplacements chimiques ne pourront ainsi être entrées qu'une fois les informations cruciales concernant l'expérience stockées dans la base de données. De même, la présence du nom et du numéro du résidu dans la table 'PeptideResidues' est indispensable pour entrer la valeur d'un déplacement chimique du noyau de celui-ci. L'atome associé à ce résidu doit aussi figurer dans la table 'dictionary_atom_per_residue' contenant la liste des atomes pour chaque résidu.

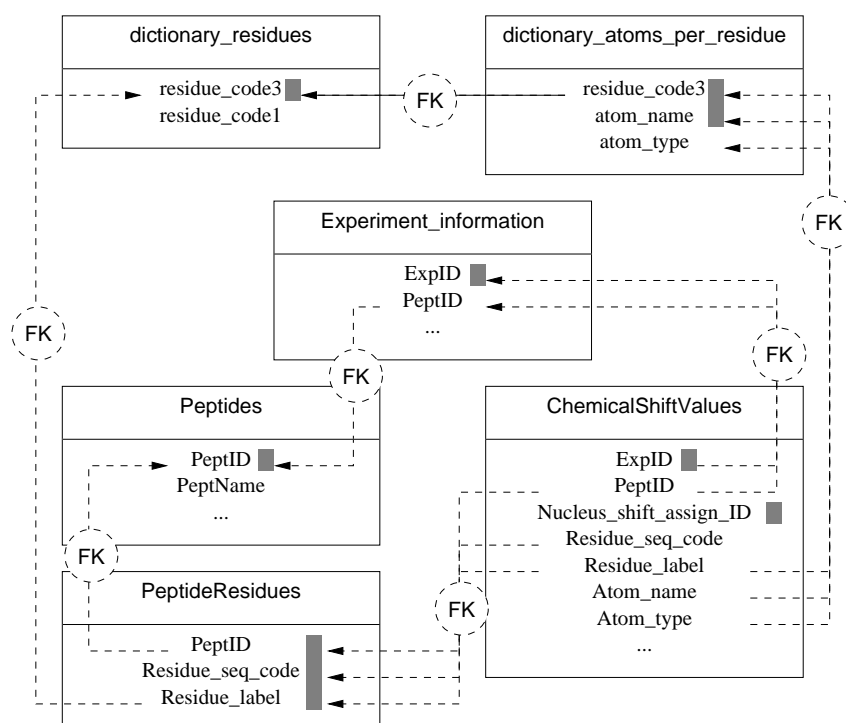


Fig. 3.3: Schéma des contraintes associées à la table des déplacements chimiques 'ChemicalShiftValues'. Chaque rectangle représente une table et les lignes pointillées entre les différentes colonnes de ces tables désignent les clés étrangères (FK : Foreign Key). Les petits rectangles gris représentent les clés primaires de chaque table.

3.4 L'interface web : l'outil de dépôt des données

Un des aspects importants pour la réussite d'une base de données est la simplicité avec laquelle il est possible d'y déposer des données. Afin d'atteindre cet objectif, nous avons développé une interface web (figure 3.4 page 24). Cet outil a été construit dans le

but de minimiser le temps nécessaire au dépôt des données : il est basé sur des formulaires à compléter et des listes à choix multiples. Les différentes sections du formulaire de déposition sont des pages HTML, facilement accessibles par la barre de menu cliquable ('Table of contents') sur la partie gauche de la figure 3.4).

Des vérifications sur les informations saisies sont faites sur l'ordinateur du dépositeur par l'intermédiaire de scripts en langage JavaScript. Elles correspondent à la première étape du processus du traitement des données réalisée au cours de la déposition (rectangle gris en haut de la figure 3.5 page 26). Il s'agit de vérifier si tous les champs obligatoires ont bien été complétés. Par exemple, si le nombre de solvants entrés pour cette expérience est égal à trois (section 'Experimental conditions', figure 3.4) le programme vérifie que l'utilisateur a correctement rempli les trois lignes correspondantes dans le tableau du dessous, les deux colonnes contenant les informations sur le type de solvant, et leurs proportions respectives dans la solution. Quand cette page est complète, la marque correspondant à cette section, à l'origine de couleur rouge, devient verte dans le menu de gauche. Pour certains paramètres, comme les données sur les déplacements chimiques, le dépositeur a la possibilité d'entrer un fichier dans un des formats disponibles afin de lui éviter la fastidieuse tâche de la saisie manuelle.

Plusieurs graphiques concernant les déviations des déplacements chimiques, basés sur les valeurs de référence les plus courantes [7, 35, 61], sont générés automatiquement. Ces graphiques sont calculés à la demande pendant la déposition si les valeurs des déplacements chimiques ont été entrées manuellement, ou sinon à tout moment une fois que le fichier de données a été vérifié.

Un identifiant est assigné à chaque déposition, et l'accès aux données est protégé par un mot de passe durant toute la période de déposition. Afin d'éviter au maximum la perte des données, chaque valeur entrée est directement sauvegardée dans un fichier temporaire.

Quand toutes les marques sont devenues vertes, la page 'Deposition' (en bas du menu gris de la 'Table of content' apparaît afin de soumettre les données. Cela permet de continuer leurs traitements (rectangle gris au milieu de la figure 3.5 page 26).

PEPTIDE CONFORMATION DATABASE - DEPOSITION - Microsoft Internet Explorer

Fichier Edition Affichage Favoris Outils ?

Adresse <http://ucmb.ulb.ac.be/Pescador/Deposit/deposition.cgi?depid=PSCD-355>

PESCADOR - Deposition id : PSCD-355.

Table of contents

- [PESCADOR home](#)
- General information**
- Laboratory of origin
- Non-natural amino acids
- Peptide information**
- Covalent/cis bonds
- Information about the experiment**
- Literature references
- Experimental conditions
- Nuclear Magnetic Resonance**
- Chemical shift data
- Coupling constants
- H/D exchange
- NH chem. shift temp. coeff.
- NOE data
- Structural information from NMR data**
- Global structure populations
- Qualitative structure analysis
- Calculated structures
- Hydrogen bonds
- Average angles
- BioMagResBank data
- Graphs**
- Deposition**

Experimental conditions

EXPERIMENTAL CONDITIONS

Enter the number of solvent components:
(e.g. use '2' for an H₂O/D₂O mix, '3' with additional TFE, ...)

Solvent	Type of solvent used	Respective percentage of solution	
1	H ₂ O	H ₂ O	<input type="text" value="60"/> %
2	H ₂ O	D ₂ O	<input type="text" value="10"/> %
3	Hexafluoroacetone	Hexafluoroacetone	<input type="text" value="30"/> %

Molecules containing water
 Temperature Kelvin
 Type of buffer Acetate
 Concentration of buffer mmol/l

Enter the number of other components present:

Component	Type of component	Concentration of component
1	NaCl	<input type="text" value="15"/> mmol/l
2	EDTA	<input type="text" value="2"/> mmol/l

Concentration of peptide mmol/l
 Does the peptide aggregate in these conditions? No Not sure Yes
 How was the aggregation state determined?
 Solubility data for this peptide (in these or other conditions)
 Additional components in sample or other remarks

[Literature references](#) [Nuclear Magnetic Resonance](#)

Problems, questions or suggestions? Please contact [Wim Vranken \(wim@ucmb.ulb.ac.be\)](mailto:wim@ucmb.ulb.ac.be)

Fig. 3.4: Vue générale d'une des pages de dépôt de Pescador.

3.5 Le traitement des données

Une procédure a été développée pour le traitement des données après qu'elles aient été déposées (rectangles gris du milieu et du bas de la figure 3.5 page 26). Les changements opérés lors de cette procédure sont faits sur une copie séparée du fichier de données. Cette procédure consiste en une standardisation et en une vérification des données dans le but de détecter d'éventuelles erreurs. La standardisation consiste à changer les noms des atomes pour qu'ils soient identiques à ceux servant de référence proposée par l'IUPAC [34]. Puis sont ajoutés des codes d'assignement stéréospécifique afin d'obtenir des données prêtes à être entrées dans les différentes tables de Pescador. Ensuite, un certain nombre de champs (comme le nom des journaux, des laboratoires, le nom des peptides, des solvants, les séquences en acides aminés) sont comparés à un dictionnaire contenant les valeurs standards. Ce dictionnaire correspond aux valeurs déjà stockées une ou plusieurs fois dans la base de données ou ajoutées ici car figurant parmi les valeurs de référence. Les valeurs ne correspondant à aucune entrée du dictionnaire sont examinées par la suite par l'annotateur et dans certains cas par le dépositeur. Les données soumises sont acceptées seulement après accord de l'une, et parfois des deux personnes (rectangle gris du bas, figure 3.5 page 26).

A la fin de la procédure de traitement, les données sont converties en format texte, similaire au format NMR-STAR utilisé par BioMagResBank [52], et elles sont entrées dans les tables de Pescador. Les données sont alors rendues publiques.

3.6 L'acquisition des données

Acquérir un minimum de données est nécessaire à l'élaboration de toute base de données, afin de la valider avant de la déclarer utilisable à la communauté scientifique. Pour atteindre ce but, nous avons entré un ensemble de données RMN sur des peptides préalablement collectés pour développer le programme Agadir [40, 41, 42] de l'EMBL d'Heidelberg. De plus, un ensemble important de données RMN sur des peptides disponibles à

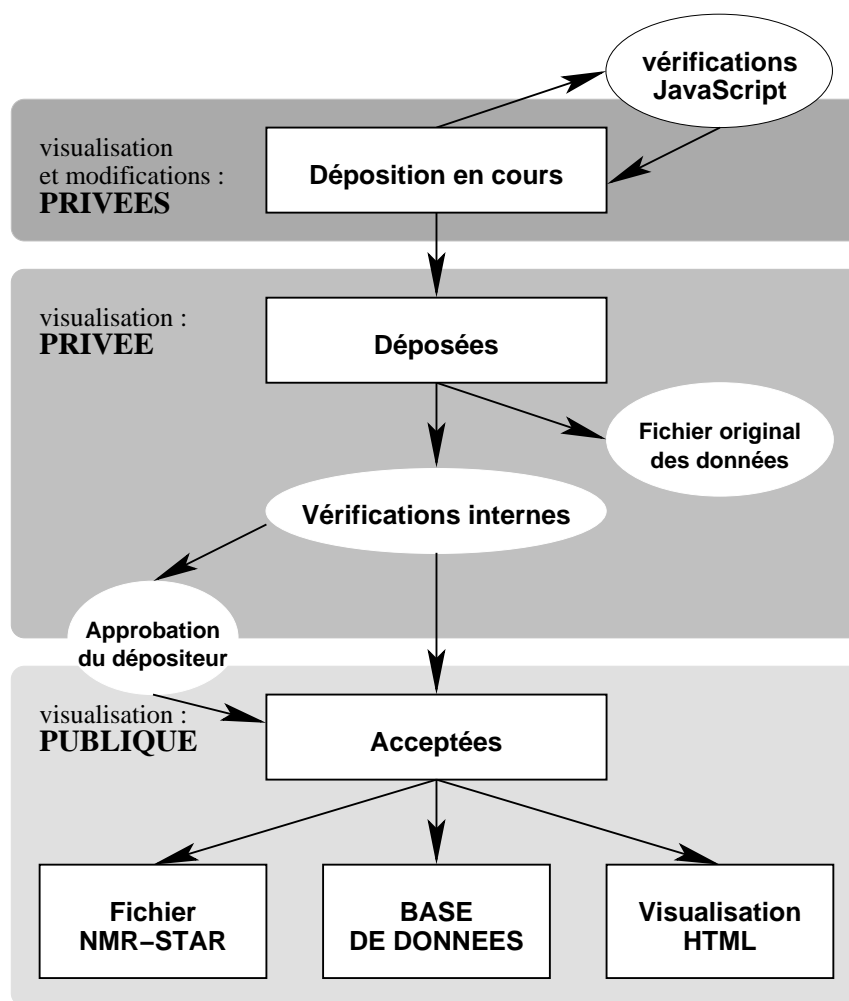


Fig. 3.5: Schéma du traitement des données dans Pescador. Chaque rectangle gris représente les différentes étapes et l'accessibilité privée ou publique aux données. L'état des données est entouré d'un rectangle noir. Les cercles entourés par des flèches changent l'état des données. Les flèches indiquent le sens du traitement des données, de haut en bas.

l'« Instituto de Estructura de la Materia » (M. Rico, Espagne) a été entré et déposé dans Pescador. Celui-ci devra être complété par des données se trouvant dans la littérature. Pescador s'intéresse donc aux peptides et à leurs conformations, une limite de 30 résidus a été fixée comme longueur acceptable des peptides pouvant être déposés dans Pescador.

3.7 L'extraction et l'analyse des données

Afin d'analyser les données déposées, nous avons développé des scripts en langage SQL ('Structured Query Language'). Cela nous a permis d'obtenir automatiquement un ordre de grandeur de plusieurs paramètres, tel que la distribution des valeurs, et un

certain nombre de paramètres caractéristiques comme la moyenne, la médiane, les valeurs minimales et maximales et les déviations standards. De plus, les résultats de l'interrogation des tables à l'aide de ces scripts peuvent être stockés dans la base de données sous forme de tables virtuelles : les vues. Elles sont créées afin de regrouper dans une même table des données comprises dans un certain intervalle ou ayant un certain type. Par exemple, pour examiner les propriétés des peptides n'ayant pas de conformation stable, un sous ensemble de données a été défini. Les peptides possédant une population de structures secondaires globales de plus de 20% ou les peptides ayant des résidus faisant partie d'éléments de structures secondaires ont été exclus de ce sous ensemble. L'élimination des peptides, ayant été étudiée dans des solvants non aqueux (H_2O et/ou D_2O), a aussi été établie grâce à une combinaison judicieuse de tables à l'aide de scripts SQL. Il en est de même pour la suppression des résidus situés aux extrémités de chaque peptide lors d'études de différentes propriétés.

Les résultats des diverses interrogations de Pescador sont générés dans un format directement lisible par le programme R, environnement et langage pour les statistiques et la création de graphiques [22]. Nous avons implémenté plusieurs scripts en langage R afin de générer automatiquement une analyse des résultats, ainsi que des représentations graphiques de ces analyses.

Chapitre 4

Analyses de Pescador

4.1 Introduction

Dans ce chapitre, nous allons tout d'abord faire un tour d'horizon des données disponibles dans Pescador. Nous allons voir comment les valeurs des déplacements chimiques et les informations expérimentales des 145 peptides contenus dans Pescador peuvent être utilisés pour dériver une nouvelle série de valeurs de référence pour les valeurs δ des protons alpha des vingt acides aminés. Cette série est basée sur des peptides ayant des séquences hétérogènes et n'ayant aucune conformation préférentielle connue. De plus, l'effet des résidus voisins a été étudié dans le cadre de cette nouvelle série de référence pour dériver des facteurs de correction. Ces facteurs nous permettent d'obtenir une nouvelle série complète de valeurs de référence, basées sur la séquence, pour des conformations non structurées, plutôt que d'utiliser celles basées sur des conformations aléatoires de peptides polyglycines. Un autre effet analysé est celui de l'influence du TriFluoroEthanol sur les valeurs des déplacements chimiques des protons amides. Les valeurs obtenues et les tendances observées ici sont comparables à celles dérivées de mesures expérimentales sur des peptides individuels, et sont dans les limites attendues. D'après notre connaissance, les valeurs de référence pour les déplacements chimiques obtenues à partir d'un ensemble de données expérimentales provenant de différents peptides sont les premières du genre. Ces résultats encourageants suggèrent que les analyses de ce type peuvent être un précieux moyen pour dériver des paramètres clés pour analyser les conformations préférentielles de peptides en solution.

4.2 Vue générale des données disponibles

La figure 4.1 présente une vue du nombre de données contenues dans chacune des tables de Pescador. La base de données contient actuellement 233 expériences déposées et traitées pour un total de 145 peptides.

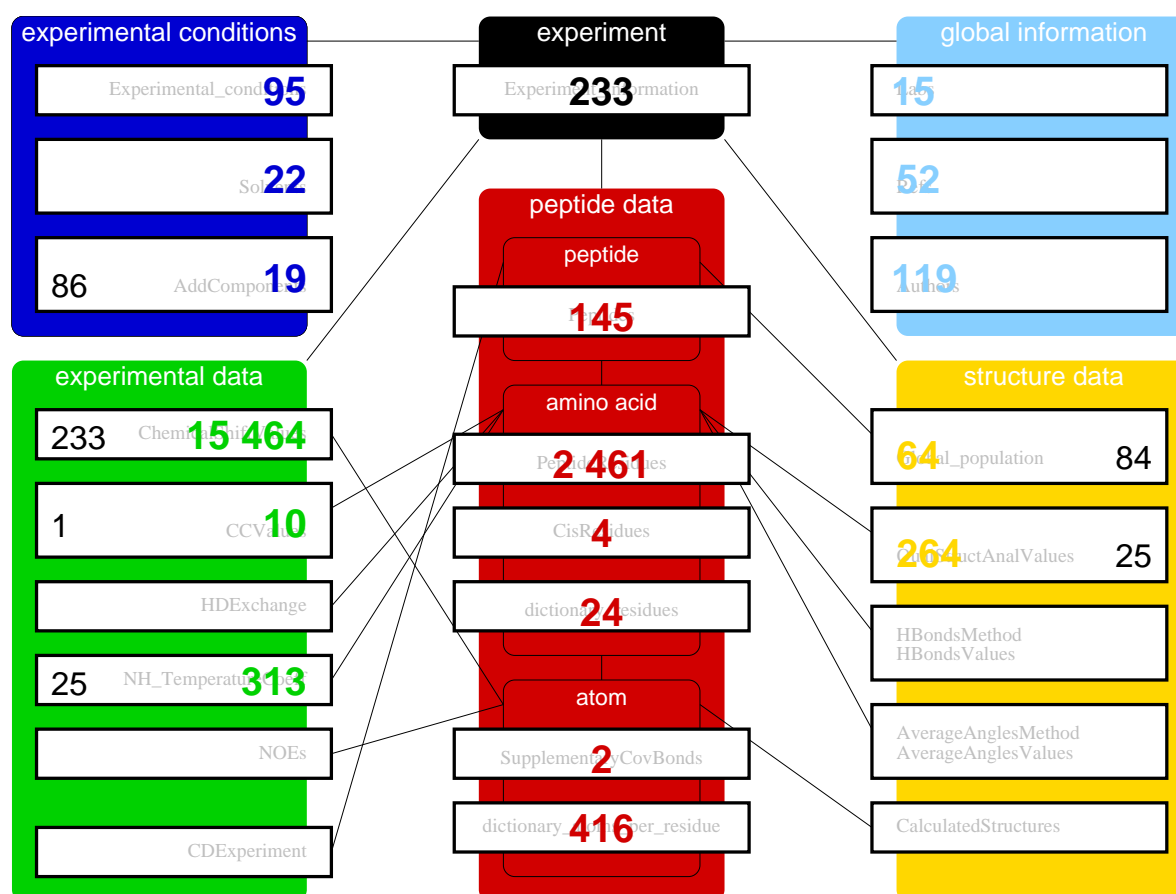


Fig. 4.1: Données contenues dans Pescador. Le nombre total de données par tables est de la couleur de la section. Le nombre en noir indique le nombre d'expériences possédant cette information.

Ces données sont issues, pour un tiers, de celles utilisées pour dériver les paramètres du programme de prédiction Agadir (voir graphique de la répartition du type d'entrée, figure 4.2 page 31), les autres proviennent des dépositions sur le site internet. Ces données proviennent de 15 laboratoires différents, dont 136 dépositions sont originaires de l'« Instituto de Estructura de la Materia » (M. Rico, Espagne). La longueur moyenne des peptides est de 16,97 résidus (graphique de la distribution de la longueur des peptides, figure 4.2 page 31). Le graphique de la répartition des solvants (figure 4.2) montre que plus

de la moitié des expériences ont été réalisées dans un solvant aqueux (H_2O et/ou D_2O) et un tiers dans un solvant ayant une concentration de TFE supérieure à 20%. Le pH de la majorité des expériences est inférieur à 7 (graphique de la distribution du pH). Toutes les expériences déposées possèdent des données sur les déplacements chimiques des protons alpha, seulement une contient des données sur les déplacements chimiques des carbones. Le graphique de la répartition du nombre de valeurs δ indique que plus d'un quart des 15464 valeurs stockées dans la table 'ChemicalShiftValues' sont des valeurs δ issues de protons alpha (HA). 25 expériences possèdent des coefficients de température des déplacements chimiques des protons amides. Les données structurales sont rares, seulement 84 entrées de Pescador ont une population structurale globale et 25 possèdent des données de structure secondaire.

La figure 4.3 (page 32) montre la comparaison de la distribution des acides aminés dans les peptides de notre base de données à celle dans les protéines issues de la base de données SWISS-PROT [3]. Les deux distributions sont assez similaires. Une exception évidente est l'alanine, qui apparaît environ deux fois plus dans Pescador que dans la SWISS-PROT. Cette proportion d'alanine, connue pour sa forte préférence pour les conformations en hélice, est due au fait qu'un nombre important de données représente des peptides ayant des séquences construites pour former des hélices. Ces peptides ont été utilisés pour dériver les paramètres du programme de prédiction Agadir [40, 41, 42, 30]. Le biais introduit par ces peptides sera discuté par la suite. Une autre exception est la cystéine, qui n'apparaît qu'avec une très faible fréquence pour l'ensemble des petits peptides de Pescador.

4.3 Définition du sous-ensemble restreint de données

L'objectif étant de dériver une série de valeurs δ de référence pour les conformations non structurées des peptides, il est nécessaire de définir un sous-ensemble restreint de données qui contient seulement des peptides ayant peu ou pas une conformation définie en solution.

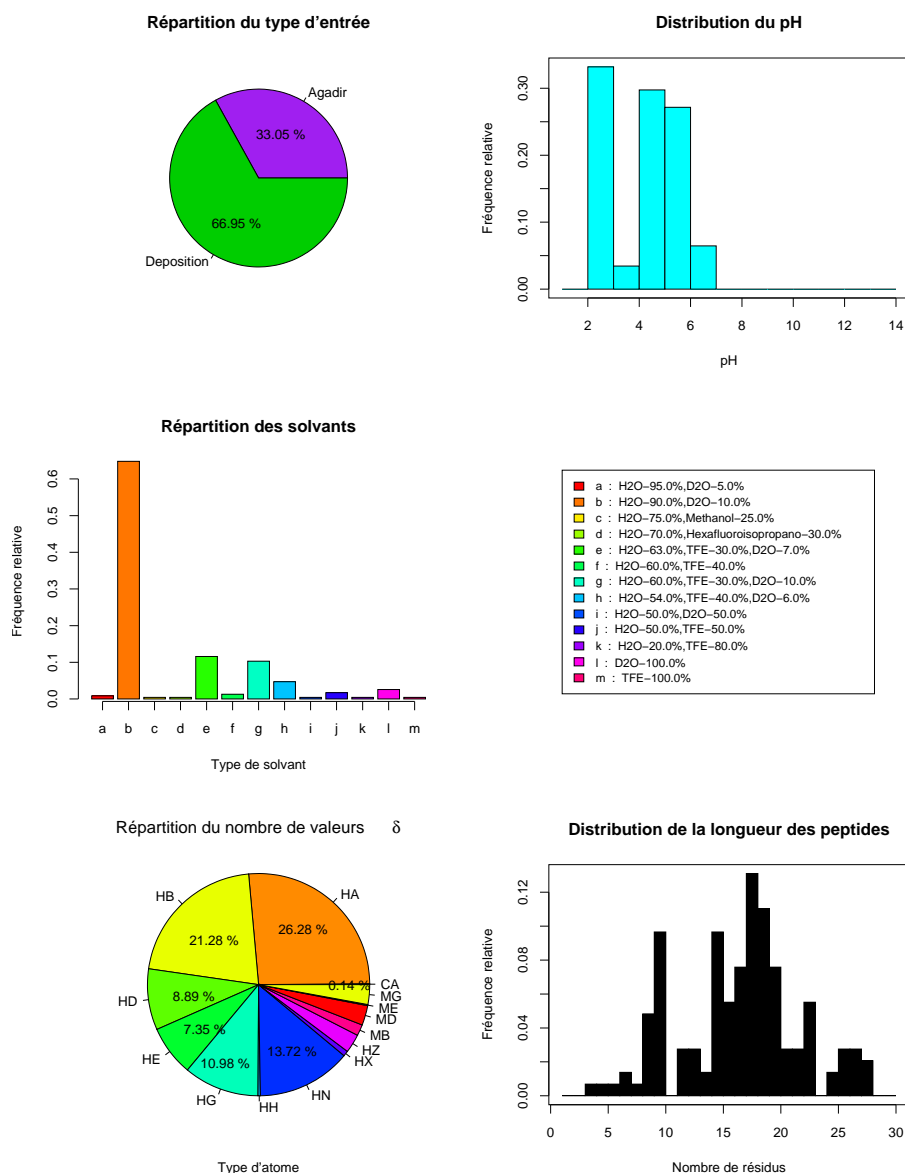


Fig. 4.2: Distribution et répartition de certaines valeurs contenues dans Pescador.

L'ensemble complet des données de Pescador contient d'une part des peptides ayant une structure hélicoïdale ou en feuillet β et d'autre part des peptides non structurés. Nous avons défini l'ensemble restreint en partant de l'ensemble complet et en excluant tous les peptides connus comme adoptant au total plus de 20% de structures secondaires. Ont aussi été exclus, les peptides étudiés dans des solvants non aqueux, et les peptides ayant des résidus se trouvant dans une conformation bien définie ou situés aux extrémités de la séquence. A partir des résultats exposés dans le tableau 4.1 (page 33), nous observons que les moyennes des valeurs δ des protons alpha pour l'ensemble restreint sont similaires à celles de l'ensemble complet, mais en général un peu plus élevées (0.02 ppm en moyenne).

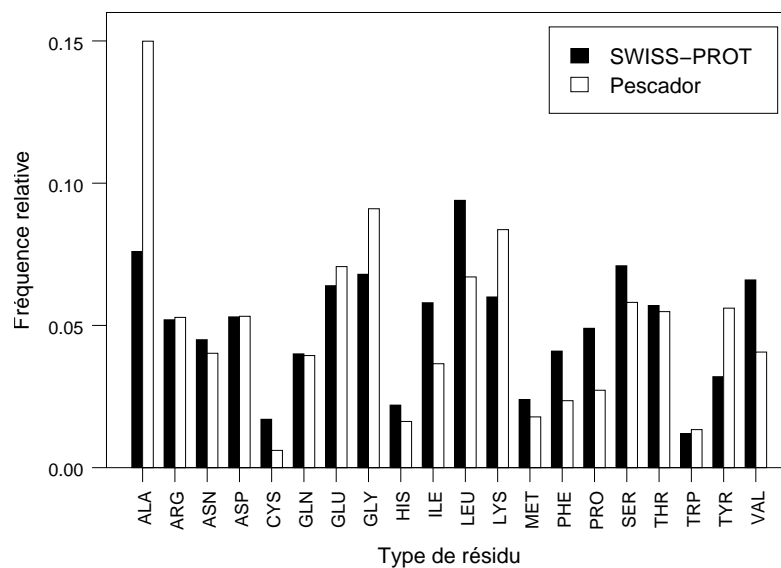


Fig. 4.3: Fréquences des résidus. Les barres blanches représentent ceux déposés dans Pescador (tous les peptides) et les noires sont ceux dans des séquences de protéines de la base de données SWISS-PROT.

Cette différence est due à la présence, dans l'ensemble complet, de nombreux peptides hélicoïdaux provenant de l'analyse d'Agadir. Quelques valeurs δ montrent des différences remarquables. A savoir, les valeurs δ pour les résidus Ile et Met sont respectivement 0.08 ppm et 0.11 ppm plus élevées dans le sous-ensemble, alors que celle pour le résidu Trp est 0.05 ppm plus faible. L'intervalle interquartile pour les valeurs du sous-ensemble est plus petit (en moyenne 0.11 ppm pour l'ensemble restreint à comparer à 0.17 ppm pour l'ensemble complet). Au regard de ces résultats, les résidus sélectionnés ici dans le sous-ensemble représentent mieux des peptides ayant des conformations non structurées. L'ensemble restreint sera donc utilisé pour toutes les autres analyses de la base de données.

4.4 Les déplacements chimiques : de nouvelles valeurs de référence

Typiquement, les valeurs des déplacements chimiques des noyaux du squelette dans les peptides et les protéines sont comparées aux valeurs de référence issues de la littérature, afin d'être utilisées pour déterminer les conformations du squelette. Ces valeurs servent dans la plupart des cas à décrire la structure secondaire adoptée par le peptide ou le

Résidu	Ensemble complet		Ensemble restreint		$\Delta\delta$
	nr	$\delta_{complet}$	nr	$\delta_{restreint}$	
ALA	537	4.24 ± 0.09	164	4.27 ± 0.08	-0.03
ARG	209	4.27 ± 0.17	77	4.30 ± 0.09	-0.03
ASN	168	4.70 ± 0.10	79	4.71 ± 0.11	-0.01
ASP	224	4.64 ± 0.13	96	4.63 ± 0.09	0.01
CYS	13	4.57 ± 0.15	9	4.58 ± 0.31	-0.01
GLN	154	4.28 ± 0.16	70	4.32 ± 0.09	-0.04
GLU	282	4.23 ± 0.15	90	4.28 ± 0.12	-0.05
GLY	504	3.98 ± 0.07	193	3.97 ± 0.06	0.01
HIS	53	4.70 ± 0.11	19	4.71 ± 0.03	-0.01
ILE	150	4.08 ± 0.29	57	4.16 ± 0.07	-0.08
LEU	291	4.28 ± 0.13	116	4.31 ± 0.09	-0.03
LYS	357	4.25 ± 0.16	135	4.28 ± 0.10	-0.03
MET	76	4.36 ± 0.25	19	4.47 ± 0.10	-0.11
PHE	92	4.59 ± 0.16	37	4.60 ± 0.08	-0.01
PRO	92	4.42 ± 0.10	38	4.42 ± 0.10	0.00
SER	203	4.41 ± 0.14	74	4.44 ± 0.12	-0.03
THR	225	4.33 ± 0.17	90	4.34 ± 0.09	-0.01
TRP	46	4.59 ± 0.42	24	4.54 ± 0.20	0.05
TYR	219	4.50 ± 0.17	62	4.54 ± 0.08	-0.04
VAL	163	4.08 ± 0.17	70	4.10 ± 0.10	-0.02
Nombre total	4058		1519		
Moyenne		0.16		0.11	-0.02
Moyenne absolue					0.03

Tab. 4.1: Valeurs des déplacements chimiques des protons alpha pour les 20 acides aminés (nombre de résidu (nr) et médiane ± intervalle interquartile). La dernière colonne ($\Delta\delta$) représente la différence des valeurs δ des protons alpha calculée selon l'équation : $\Delta\delta = \delta_{complet} - \delta_{restreint}$. Les $\Delta\delta$ supérieurs à 0.08 ppm sont en **gras**.

segment de protéine [63]. Afin de procéder de la même manière, il est intéressant tout d'abord de comparer les valeurs δ mesurées sur des peptides ou des protéines et servant de référence d'un côté, avec celles collectées dans Pescador. Ces dernières font partie du sous-ensemble décrit précédemment et correspondent à des peptides non structurés ou ayant une conformation aléatoire. Les valeurs de référence pour des conformations aléatoires sont habituellement obtenues à partir d'une série de petits peptides composés de glycines [7, 35, 61]. La figure 4.4 (page 34) montre une assez large divergence entre les valeurs δ des séries dérivées par les différents auteurs. L'utilisation seule de ces références n'est souvent pas suffisante, l'effet additionnel de la séquence dû aux influences des résidus voisins peut jouer un rôle important. Récemment, une série de facteurs de correction a

été proposée afin de compenser les effets importants de la séquence [51], basée une fois encore sur une série de peptides composés en majorité de glycines.

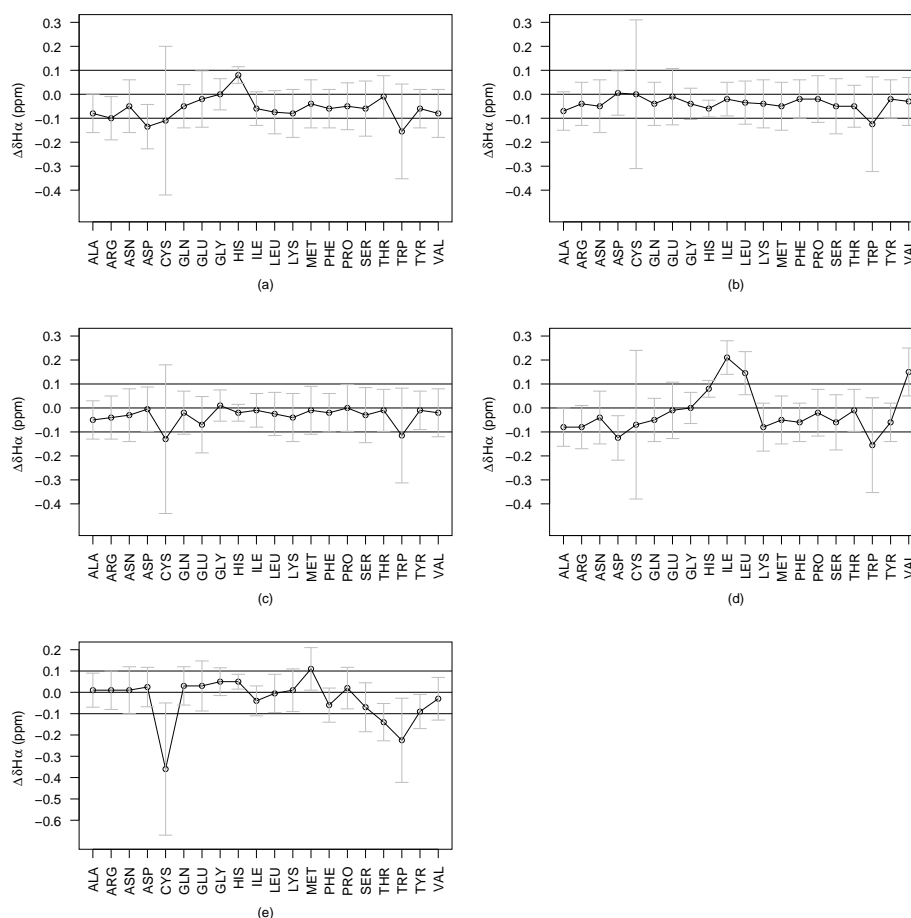


Fig. 4.4: Déviations $\Delta\delta$ des protons alpha pour les 20 acides aminés à partir des valeurs δ issues de la littérature et ayant des conformations aléatoires. Ces différences sont obtenues par l'équation $\Delta\delta = \delta_{\text{pescador}} - \delta_{\text{réf.}}$, soustraction des valeurs du sous-ensemble de Pescador aux valeurs correspondantes provenant des séries : (a) GGXA (308 K, pH 7.0, Bundi and Wüthrich [7]) (b) GGXGG (entre 278 K et 328 K, pH 5.0, Merutka et al. [35]) (c) GGXAGG (298 K, pH 5.0, Wishart et al. [61]) (d) Chemical Shift Index, (Wishart et al. [63]) (e) valeurs moyennes de BMRB. La médiane (milieu de la barre) et l'intervalle interquartile (La hauteur de la barre représente l'intervalle où se trouve 50% des valeurs) sont représentés afin de donner une idée de la distribution des données dans Pescador.

La figure 4.5 (page 35) et le tableau 4.1 (page 33) donnent une vue d'ensemble du nombre et de la distribution des valeurs des déplacements chimiques des protons alpha extraits de l'ensemble restreint des données. La distribution des valeurs δ pour les résidus Ala et Lys suit plus ou moins une distribution normale. Cependant celle pour les résidus Ile et Trp n'est clairement pas normale. En fait, les tests statistiques de normalité

de Kolmogorov-Smirnov [10] et de Shapiro-Wilk [48] établissent qu'aucune des distributions ne suit une distribution normale. Il n'est donc pas possible d'appliquer des tests statistiques standards sur ces données puisque la majorité d'entre eux requière une distribution normale comme une des conditions initiales. Seulement une analyse descriptive basée sur les médianes et les dispersions est donc possible ici. L'absence de distribution normale pour les valeurs δ était attendue à cause de la diversité des données collectées sous des conditions expérimentales diverses, et assujetties à un nombre important de facteurs différents.

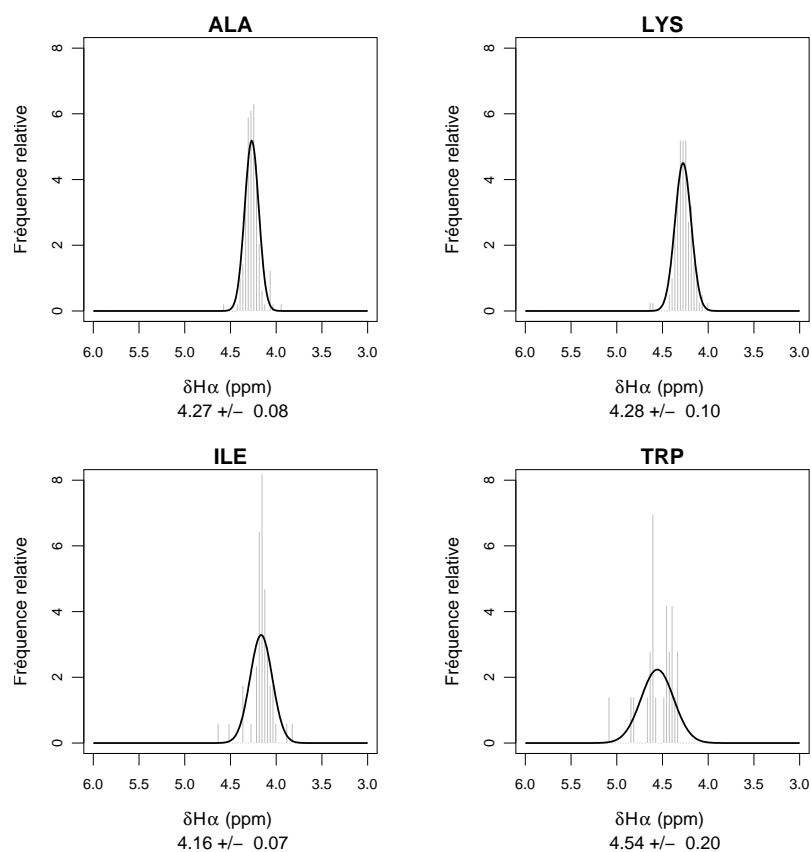


Fig. 4.5: Distribution des valeurs δ des protons alpha du sous-ensemble pour les résidus Ala, Lys, Ile et Trp. La fréquence relative des valeurs δ est représentée par des barres verticales gris clair et la courbe en noir est la distribution normale qui s'ajuste sur la distribution des fréquences observées.

La distribution des déplacements chimiques des protons amides (figure 4.6 page 36) met en évidence des dispersions encore plus importantes. Dans le cas de l'alanine, par exemple, l'intervalle interquartile des valeurs δ pour les protons amides est 3 fois plus important que pour les protons alpha. Cette observation étant prévisible, puisque les δ des protons

amides sont extrêmement plus dépendants du type de solvant, de la température ainsi que du pH que les protons alpha.

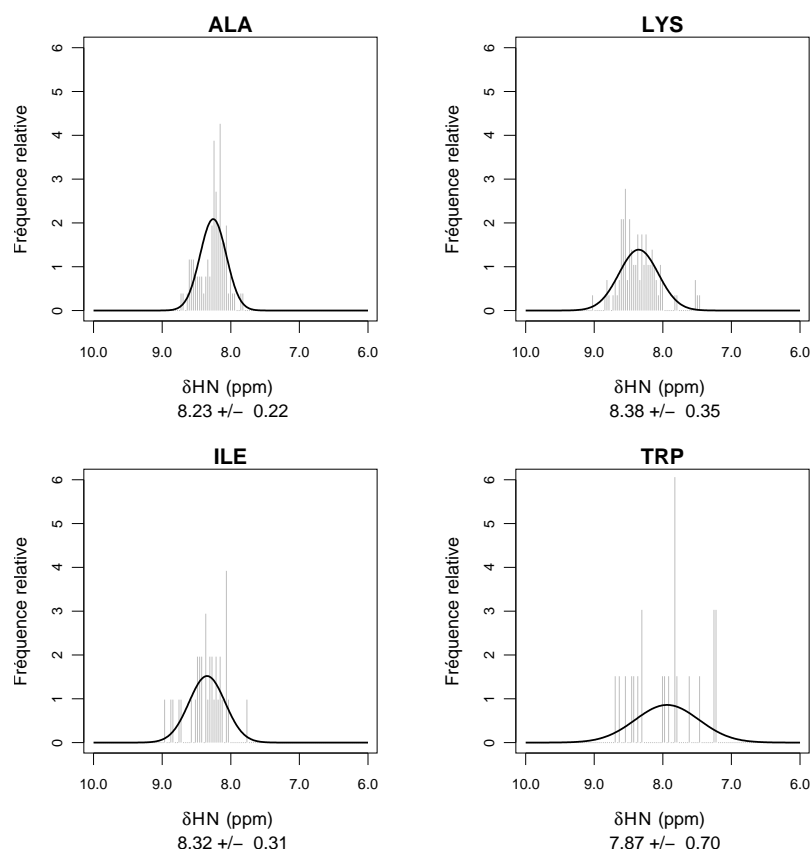


Fig. 4.6: Distribution des valeurs δ des protons amide du sous-ensemble pour les résidus Ala, Lys, Ile et Trp.

4.5 Les facteurs de correction des valeurs δ des protons alpha

Les facteurs de correction sont calculés à partir de la médiane des valeurs δ des protons alpha des 20 acides aminés dérivés de la base de données comme valeurs de référence ($\delta_{restreint}$, tableau 4.1 page 33). Le facteur de correction A du déplacement chimique du proton alpha pour un résidu de type z est obtenu comme suit :

$$A[z] = \frac{\sum_{i=1}^{20} \left(\sum_{j=1}^{n_i} (\delta X_{a_j}[i] - \delta X_{réf.}[i]) \right)}{\sum_{i=1}^{20} n_i} \quad (4.1)$$

$\delta X_{aj}[i]$ est la valeur δ du proton alpha du résidu X_a de type i pour un résidu z spécifique dans la sous séquence :

$$\dots - X_a - X_b - z - X_c - X_d - \dots \quad (4.2)$$

$\delta X_{réf.}[i]$ est ici la médiane des valeurs δ des protons alpha pour un résidu de type i ($\delta_{restreint}$, tableau 4.1 page 33). Une somme sur tous les n_i fragments de séquence est faite ainsi que sur les 20 acides aminés de type i . Les trois autres facteurs de correction (B , C et D) sont calculés de la même manière et les résultats sont présentés dans le tableau 4.2 (page 38). Les facteurs de correction permettent de compenser l'effet de la séquence sur les valeurs médianes des δ calculées dans Pescador et servant de référence. Cette correction est basée sur la formule de Schwarzinger et al. [51] :

$$\begin{aligned} \delta_{R\text{corrigé}} &= \delta_{R\text{réf.}} + \Delta\delta_{R-1} + \Delta\delta_{R+1} + \Delta\delta_{R-2} + \Delta\delta_{R+2} \\ &= \delta_{R\text{réf.}} + C_{[z=R-1]} + B_{[z=R+1]} + D_{[z=R-2]} + A_{[z=R+2]} \end{aligned} \quad (4.3)$$

où le résidu R se trouve dans cette sous séquence :

$$\dots - R_{-2} - R_{-1} - R - R_{+1} - R_{+2} - \dots \quad (4.4)$$

4.6 L'influence des résidus voisins sur les déplacements chimiques

L'effet des résidus voisins sur les valeurs par défaut des δ des résidus du squelette peptidique est connu depuis longtemps, particulièrement l'effet de la proline. La figure 4.7 (page 39) montre une comparaison entre les distributions des valeurs δ des protons alpha des résidus précédant quatre résidus représentatifs (Ala, Lys, Ile et Trp) et celles des résidus précédant Pro. Les distributions sont clairement différentes et mettent en évidence le fait que les données contenues dans Pescador reproduisent cette tendance correctement.

Une analyse détaillée de l'effet des résidus voisins a été récemment publiée par Schwarzinger et al. [51], qui examinait une série de peptides GGXGG dans le but d'obtenir des

z	A	B	C	D	n_i
ALA	-0.01 ± 0.10	-0.04 ± 0.09	-0.03 ± 0.09	-0.02 ± 0.08	109
ARG	0.00 ± 0.07	0.00 ± 0.06	0.00 ± 0.07	-0.01 ± 0.06	48
ASN	0.03 ± 0.08	0.00 ± 0.06	-0.01 ± 0.06	0.00 ± 0.07	45
ASP	0.00 ± 0.08	-0.01 ± 0.07	-0.02 ± 0.09	-0.02 ± 0.10	53
CYS	0.02 ± 0.02	0.00 ± 0.03	0.21 ± 0.11	-0.09 ± 0.03	3
GLN	0.01 ± 0.07	-0.03 ± 0.07	-0.01 ± 0.07	0.00 ± 0.09	37
GLU	-0.02 ± 0.11	-0.04 ± 0.07	-0.04 ± 0.09	0.01 ± 0.12	62
GLY	0.00 ± 0.08	0.00 ± 0.05	0.05 ± 0.10	0.01 ± 0.03	11
HIS	0.02 ± 0.21	-0.08 ± 0.03	0.06 ± 0.07	-0.01 ± 0.02	8
ILE	0.00 ± 0.06	0.01 ± 0.06	0.02 ± 0.09	0.00 ± 0.06	36
LEU	-0.02 ± 0.12	-0.03 ± 0.08	-0.01 ± 0.07	-0.01 ± 0.07	78
LYS	-0.02 ± 0.09	-0.03 ± 0.07	-0.02 ± 0.06	0.01 ± 0.10	71
MET	-0.05 ± 0.05	-0.01 ± 0.08	0.01 ± 0.07	0.03 ± 0.13	12
PHE	-0.03 ± 0.06	-0.06 ± 0.08	-0.04 ± 0.19	-0.07 ± 0.08	24
PRO	0.04 ± 0.13	0.30 ± 0.06	0.00 ± 0.13	0.04 ± 0.06	20
SER	0.01 ± 0.07	0.05 ± 0.04	0.03 ± 0.05	-0.02 ± 0.10	33
THR	0.02 ± 0.05	0.08 ± 0.04	0.03 ± 0.05	0.01 ± 0.11	52
TRP	-0.08 ± 0.16	-0.16 ± 0.11	-0.16 ± 0.08	-0.16 ± 0.09	15
TYR	-0.02 ± 0.07	-0.08 ± 0.11	-0.06 ± 0.08	-0.01 ± 0.06	32
VAL	-0.01 ± 0.12	0.02 ± 0.12	0.03 ± 0.07	0.00 ± 0.09	41
Moyenne	-0.01 ± 0.09	0.00 ± 0.07	0.00 ± 0.08	-0.02 ± 0.08	
	0.02 ± 0.09	0.05 ± 0.07	0.04 ± 0.08	0.03 ± 0.08	

Tab. 4.2: Facteurs de correction des valeurs δ des protons alpha dans la sous séquence 4.2 (page 37), calculés à partir de l'équation 4.1 (page 36) sur l'ensemble restreint. Les facteurs de correction supérieurs à 0.08 ppm sont en **gras**. La ligne du bas du tableau indique la moyenne absolue. n_i est le nombre de segments de séquence par type z de résidu.

facteurs de correction dépendant de la séquence pour les déplacements chimiques utilisés comme référence et basés sur des peptides en conformations aléatoires. Une analyse similaire a été faite ici en utilisant les déplacements chimiques stockés dans Pescador. Afin d'éviter le plus possible une interférence avec l'influence de la structure secondaire sur les valeurs δ , le sous-ensemble de données a été utilisé pour dériver les facteurs de correction. Les résultats présentés dans le tableau 4.2 mettent en évidence une fois de plus que la valeur δ d'un résidu précédant une proline est significativement plus élevée, en moyenne de 0.30 ppm, par rapport aux autres résidus. En outre, des influences de 0.08 ppm et plus sont observées pour les résidus précédant His, Thr, et Tyr et pour les quatre résidus entourant Trp. La tendance observée pour la cystéine est basée sur un nombre trop limité de données (3 valeurs) et ne sera donc pas prise en compte.

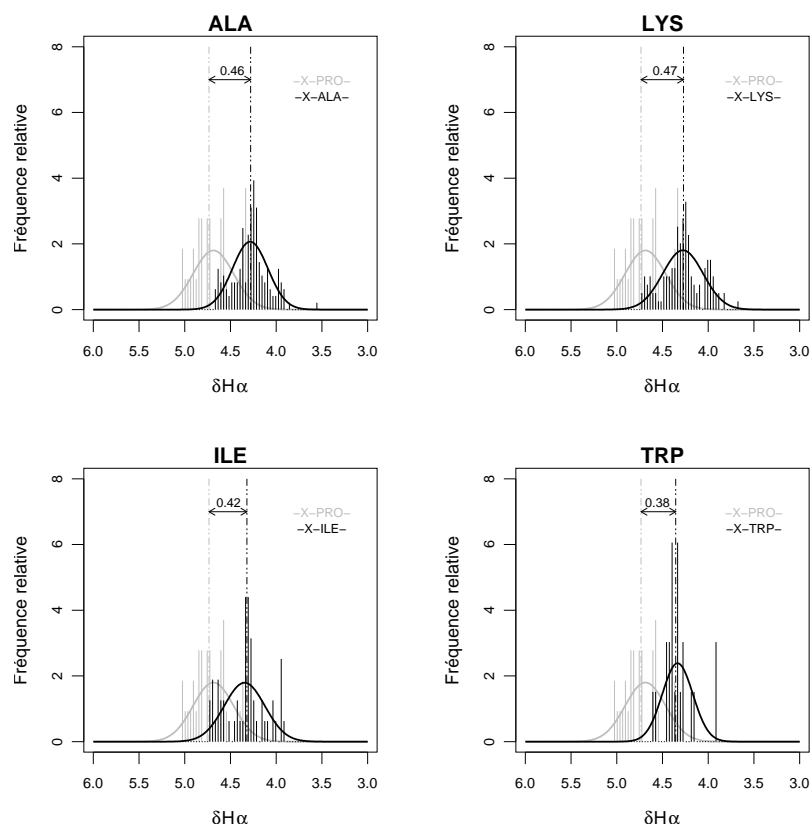


Fig. 4.7: Comparaison entre les distributions des valeurs δ des protons alpha du résidu X suivi par Pro (barres grises) et celles du résidu X suivi par les résidus Ala, Lys, Ile et Trp (barres noires) du sous-ensemble de données. La fréquence relative des valeurs δ est représentée par des barres verticales, la courbe de la distribution normale est ajustée à la distribution des valeurs observées et les lignes verticales pointillées représentent la médiane de chaque distribution. La différence entre les deux médianes est au dessus de la double flèche.

Les tendances observées pour Pro et Trp sont similaires à celles de l'étude faite par Schwarzinger et al. [51] (tableau 2.2 page 14), mais sont plus prononcées. Les valeurs observées pour les résidus précédant His et Thr sont aussi plus importantes (-0.08 ppm et 0.08 ppm respectivement) alors qu'elles n'atteignent pas le seuil significatif pour Schwarzinger et al. [51]. Dans le cas des résidus précédant et suivants Phe et Tyr, les valeurs sont quasiment similaires (de 0.02 à 0.04 ppm de différence avec les valeurs de Schwarzinger et al.). Les différences observées peuvent être attribuées au fait que les données de Pescador dérivent de séquences différentes, alors que celles de la série GGXGG sont biaisées par la présence de résidus Gly. Une explication possible de la divergence des résultats entre la série des peptides GGXGG et ceux issus de Pescador, est l'absence (ou la présence) d'interactions des chaînes latérales des résidus aromatiques Phe, Trp et Tyr avec le squelette. Le fait est que l'existence d'interactions $\text{Tyr}_i\text{-Gly}_{i+2}$ a été démontrée [26, 27]. Les

déviations plus importantes pour les résidus précédant Pro et Thr peuvent être reliées à l'augmentation attendue des interactions stériques pour des résidus ayant des chaînes latérales par rapport au résidu Gly.

Nous avons davantage examiné les valeurs médianes δ des protons alpha du sous-ensemble de Pescador pour les résidus se trouvant avant une proline (tableau 4.3). Les déviations de ces valeurs par rapport aux valeurs de référence de Pescador corrigées par le facteur B pour Pro, sont comprises entre -0.11 et 0.02 ppm suivant le type de résidu. Elles sont comprises entre 0.01 et 0.25 ppm quand elles sont calculées par rapport aux valeurs de référence aléatoires de la série GGXGG de Merutka et al. [35] combinées au facteur de correction B de Schwarzinger et al. [51]. Malgré le nombre très faible de données, les valeurs observées pour chaque type de résidu précédant une proline sont plus proches des valeurs prédites par Pescador (déviation absolue de 0.06) que celles proposées par Schwarzinger et al. (déviation absolue de 0.12). Ces résultats prometteurs mettent en évidence que l'augmentation du nombre de données pourrait permettre de calculer des facteurs de correction spécifiques pour des paires de résidus.

Résidu x-PRO	nr	$\delta_{H\alpha}$	Pescador			Schwarzinger et al.		
			$\delta_{réf.}$	$\delta_{corrigé}^*$	$\Delta\delta$	$\delta_{réf.}$	$\delta_{corrigé}^*$	$\Delta\delta$
ALA	2	4.46	4.27	4.57	-0.11	4.34	4.45	0.01
ARG	3	4.61	4.30	4.60	0.01	4.34	4.45	0.16
ASN	12	4.91	4.71	5.01	-0.10	4.76	4.87	0.04
ASP	2	4.85	4.63	4.93	-0.08	4.63	4.74	0.11
GLU	2	4.56	4.28	4.58	-0.02	4.29	4.31	0.25
ILE	2	4.37	4.16	4.46	-0.09	4.18	4.29	0.08
PRO	2	4.72	4.42	4.72	0.00	4.44	4.55	0.17
SER	4	4.76	4.44	4.74	0.02	4.49	4.61	0.15
VAL	5	4.33	4.10	4.40	-0.07	4.13	4.24	0.09
Moyenne absolue					0.06	0.12		

Tab. 4.3: Déviation $\Delta\delta$ des protons alpha entre les valeurs $\delta_{H\alpha}$ observées dans Pescador pour les résidus x précédant une proline (colonne 3) et les valeurs de référence ($\delta_{réf.}$, colonne 4 correspondant aux valeurs de la colonne $\delta_{restreint}$ dans le tableau 4.1 page 33) corrigées par le facteur de correction B de Pro issus de Pescador ($\delta_{corrigé}^* = \delta_{réf.} + B_{Pro}$, colonne 5) et celles de Merutka et al. [35] (colonne 7 correspondant aux valeurs du tableau 2.1 page 12) corrigées par le facteur de correction B de Schwarzinger et al. [51].

4.7 Application des facteurs de correction

Nous avons testé les facteurs de correction dérivés de Pescador sur des petits peptides ne se trouvant pas dans la base de données : le feuillet β formé par la Carp Granulin 1-30 [57] (voir tableau 1.1 page 109, chapitre 1.2 de la partie III), le fragment V3 en tournant avec un semblant d'hélice [9] (voir tableau 1.2 page 110) et la phosphatase acide lysosomale [13] formant un tournant β (voir tableau 1.3 page 111). Les séquences en acides aminés sont données dans le tableau 4.4 (page 42) . Les facteurs de correction (tableau 4.2 page 38) sont appliqués aux valeurs médianes δ de Pescador pour chaque type de résidu suivant l'équation 4.3 (voir page 37). Ils sont considérés comme étant additifs et toutes les contributions ont été prises en compte (même celles inférieures à 0.08 ppm). Pour les trois peptides étudiés, les déviations absolues $DA = \sum | \Delta\delta | = \sum | \delta_{\text{observé}} - \delta_{\text{corrigé}} |$ entre les valeurs δ observées expérimentalement et celles calculées à partir de Pescador sont inférieures à celles calculées à l'aide des valeurs aléatoires standards de Merutka et al. [35] (tableau 4.4 page 42, colonne 1). Elles restent inférieures même après l'ajout des facteurs de correction de Schwarzingger et al. [51] (tableau 4.4, colonne 2). Il est intéressant de noter que les valeurs δ corrigées, proposées par Schwarzingger et al. (tableau 4.4, colonne 2), se situent à mi-chemin entre les valeurs aléatoires d'origine et les valeurs δ corrigées de Pescador.

La différence entre les trois jeux est encore plus frappante pour la phosphatase acide lysosomale (figure 4.8 page 43) qui n'a pas de conformation de préférence : il possède seulement un tournant β central. Pour ce cas particulier, les $\Delta\delta$ des protons alpha pour chaque résidu sont représentés sur la figure 4.8. Les déviations des valeurs observées par rapport aux valeurs de référence sont globalement moins importantes dans le cas de Pescador (cercle rouge sur la figure 4.8). Ces résultats sont davantage en adéquation avec le fait que la phosphatase acide lysosomale n'est pas structurée en solution. Plus en détail, les différences entre les trois sets sont encore plus marquées pour les résidus Gln4, Pro5 et Pro6. Les déviations sont respectivement de 0.22, 0.26 et -0.03 ppm lorsqu'elles sont

Peptide	Merutka et al. [35]			Merutka et al. [35] + Schwarzinger et al. [51]			Pescador		
	DA	NEG	POS	DA	NEG	POS	DA	NEG	POS
A ¹	0.282	-0.072	0.210	0.283	-0.060	0.222	0.274	-0.061	0.213
B ²	0.075	-0.053	0.021	0.071	-0.045	0.026	0.068	-0.036	0.031
C ³	0.111	-0.082	0.029	0.089	-0.072	0.017	0.069	-0.049	0.020

¹Carp Granulin 1-30, VIHCDAAATICPDGTTCSLSPYGVWYCSPFS

²Fragment en tournant V3, YNKRKRIHIGPGRAFYTTKNIIGC

³Acide Phos. Lysosomal, MQAQPPGYRHVADGEDHA

Tab. 4.4: Déviation globale moyenne des valeurs δ des protons alpha du peptide Carp Granulin 1-30 [57] (A), du fragment en tournant V3 [9] (B), et de la phosphatase acide lysosomale [13] (C). Les déviations absolues (DA), les contributions négatives (NEG) et les positives (POS) sont la somme des différences entre les valeurs δ observées et les valeurs δ aléatoires standards de Merutka et al. [35] (colonne 1), de Merutka et al. [35] avec les facteurs de correction de Schwarzinger et al. [51] (colonne 2) et les valeurs δ de référence de Pescador avec les facteurs de correction (colonne 3).

calculées avec les valeurs de Merutka et al. [35], de 0.11, 0.15 et -0.03 ppm calculées avec les valeurs de Merutka et al. [35] et les facteurs de correction de Schwarzinger et al. [51], et de -0.05, 0.01 et 0.01 avec les valeurs de référence de Pescador corrigées par les facteurs de correction.

Le protocole utilisé par Pescador pour calculer les valeurs de déplacement chimique dépendant de la séquence, produit des résultats équivalents à ceux obtenus à partir des valeurs dérivées de peptides polyglycines, mais des différences sont cependant visibles. Il est clair que ce protocole est ouvert aux améliorations car la simple addition des facteurs de correction implique que l'effet des résidus sur leurs voisins ne dépend ni des uns ni des autres. En outre, une augmentation des données sur des peptides n'ayant pas de conformation de préférence en solution améliorerait l'exactitude des valeurs médianes calculées à partir de l'ensemble restreint de données. Une investigation complète, pour trouver la meilleure façon d'implémenter les facteurs de correction, pourrait alors être envisagée. Mais globalement, une analyse de multiples séquences de peptides semble présenter une vue plus réaliste des conformations de peptides en solution, et doit donc apporter des facteurs de correction plus fiables. Dans l'attente d'un ensemble suffisamment important, certains facteurs de correction peuvent déjà être spécifiques pour chaque type de résidu. Par exemple, le facteur de correction calculé pour un résidu Ala quand il est devant Pro

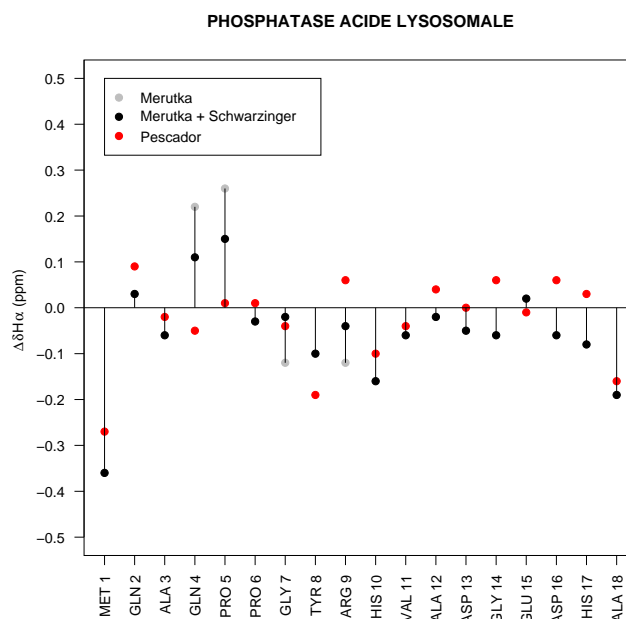


Fig. 4.8: Déviations $\Delta\delta$ entre les valeurs δ des protons alpha observées pour la phosphatase acide lysosomale et les valeurs de référence. Le point gris correspond aux déviations par rapport aux valeurs de Merutka et al. [35], le point noir correspond aux déviations par rapport aux valeurs de Merutka et al. [35] avec les facteurs de correction de Schwarzingger et al. [51] et le point rouge correspond aux déviations par rapport aux valeurs de Pescador avec les facteurs de correction.

n'est pas le même que lorsque le résidu Ala se trouve devant les autres acides aminés. Des combinaisons de paires de résidus peuvent être ainsi faites.

4.8 L'influence du TFE sur les δ des protons amides

L'addition de TFE à une température donnée fait diminuer les valeurs δ des protons amides [35]. Ce résultat est confirmé par l'observation des acides aminés Ile, Leu, Asp, Thr, Arg et Trp des peptides stockés dans Pescador (figure 4.9 page 45) et étudiés à différentes concentrations de TFE à 278K. Malgré le nombre limité de données pour chaque point, la diminution des valeurs δ des protons amides se poursuit jusqu'à 80% de TFE (graphes Leu, Asp et Thr de la figure 4.9). L'observation de cette décroissance reflète sûrement davantage la diminution des liaisons hydrogènes possibles entre les protons amides et le solvant. Un résultat similaire mais moins prononcé a été observé pour les valeurs δ des protons alpha en fonction de la concentration de TFE. Dans ce cas, l'accroissement du nombre de valeurs négatives des $\Delta\delta$ est probablement dû à une augmentation des

conformations hélicoïdales. Par exemple, la valeur δ des protons alpha de Leu diminue en moyenne de 0.10 ppm quand la concentration de TFE augmente de 10% à 80%, tandis que celle pour les protons amides diminue en moyenne de 0.54 ppm.

Cette tendance observée par Merutka et al. [35], même avec un nombre limité de données dans Pescador, est confirmée. Cela illustre clairement les possibilités offertes pour d'autres analyses, comme l'effet du TFE sur les valeurs des déplacements chimiques des protons amides ou sur la structure secondaire, à partir du moment où suffisamment de données pourront être disponibles.

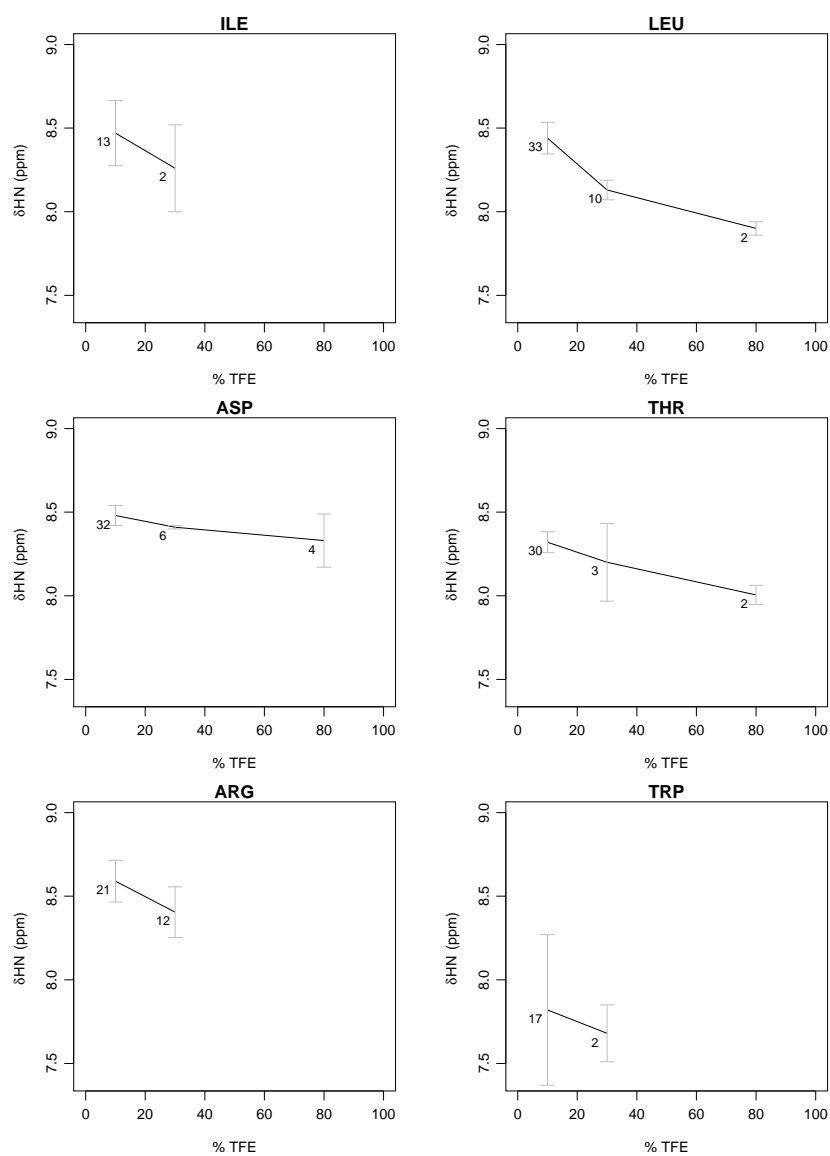


Fig. 4.9: Évolution des valeurs médianes δ des protons amides des résidus sélectionnés en fonction de la concentration de TFE à 278K provenant du sous-ensemble. Les barres claires représentent l'intervalle dans lequel se trouvent 50% des valeurs et le nombre à côté de chaque point indique le nombre des données.

Chapitre 5

Discussion

L'importance de rassembler des données sur des peptides ayant des conformations différentes a déjà été illustrée par le travail sur Agadir, un programme de prédiction d'hélice α pour les peptides [40, 41, 42, 30]. Avec Pescador, la capture des données est étendue à un plus grand nombre de peptides possédant des conformations de préférences différentes. Un des principaux avantages de Pescador est donc d'offrir la possibilité d'analyser les effets de nombreux paramètres, comme la température ou le pH, à partir de caractéristiques expérimentales observées sur les acides aminés, sans devoir recourir à une analyse expérimentale d'une série complète de peptides. De plus, les données de base sur les conformations de peptides proviennent d'un nombre important de sources, l'analyse de ces données réduit donc l'influence des approches spécifiques liées aux laboratoires et peuvent aider à l'obtention de conclusions plus fiables. Mais aussi, moins de conditions sur le comportement des peptides sont nécessaires dans ce cas par rapport à l'examen d'une série limitée de peptides. La souplesse et la diversité de Pescador peuvent en faire un excellent outil pour l'identification des structures secondaires de peptides et de leurs relations à la séquence en acides aminés, ainsi que pour une identification de l'effet des conformations nécessitant davantage de validations expérimentales. Le nombre limité des données dans Pescador permet déjà de reproduire des valeurs et des tendances décrites dans la littérature. Pescador deviendra probablement de plus en plus utile au fur et à mesure que de nouvelles données y seront déposées.

Bien que l'analyse présentée ici soit essentiellement basée sur les déplacements chimiques des protons alpha, elle est directement transférable aux autres types de noyaux qui sont parfois mieux adaptés à la détermination de la structure secondaire. Cependant, les spectres des noyaux hétéronucléaires sont rarement enregistrés pour les peptides, une méthode plus fiable pour déterminer la structure secondaire basée sur les protons alpha serait donc très utile pour les études de conformations de peptides. Dans un peptide, le déplacement chimique d'un proton alpha est principalement défini par la valeur moyenne de l'angle phi du squelette. La valeur phi de préférence pour un résidu dans un tel peptide est déterminée premièrement par le type de la chaîne latérale de ce résidu, interagissant avec le squelette et imposant des contraintes stériques sur sa conformation, et deuxièmement par les effets des résidus voisins. Cependant, deux résidus ayant la même valeur phi de préférence n'ont pas obligatoirement la même valeur δ pour le proton alpha puisque la chaîne latérale du résidu et les interactions avec les résidus voisins peuvent avoir un effet différent sur l'environnement chimique du proton alpha. Les données de Pescador fournissent quant à elles des valeurs δ de protons alpha dépendantes de la séquence, pour un type de résidu donné dans sa conformation de préférence en solution provenant de peptides n'adoptant pas de structure secondaire évidente. Ces nouvelles valeurs pour des conformations non structurées sont intrinsèquement différentes des valeurs des index de déplacement chimique (CSI [62]) obtenues à partir de protéines ayant des structures secondaires connues, ou des valeurs aléatoires obtenues à partir d'une série de peptides polyglycines courts. Les valeurs pour des conformations non structurées fournissent donc un excellent point de départ pour l'évaluation de la formation des structures secondaires dans les peptides. L'angle dièdre phi de préférence étant différent pour chaque type de résidu, la taille de la déviation indiquant la formation de structure secondaire est aussi dépendante du résidu. Si davantage de données, contenant des informations détaillées sur des peptides adoptant une structure secondaire, sont entrées dans Pescador, il sera alors possible de calibrer les déviations à partir des valeurs des conformations non structurées avec des informations sur la structure secondaire.

Pescador contient pour l'instant essentiellement des données RMN de base. L'importance des données de Dichroïsme Circulaire pour la base de données est capitale, cette méthode étant couramment utilisée pour déterminer la population globale de structures secondaires pour les peptides. Ces données seraient particulièrement intéressantes pour les peptides étudiés par RMN, les paramètres expérimentaux basés sur les résidus pourraient alors être corrélés à ceux sur la population globale de structures secondaires.

Deuxième partie

Étude de motifs structuraux
récurrents de protéines en relation
avec leurs empreintes expérimentales

Chapitre 1

Introduction

L'importante accumulation des données spatiales sur les protéines depuis plusieurs années a permis de révéler la présence de motifs structuraux récurrents, comme ceux impliqués dans les régions en tournant. La nomenclature des tournants est basée sur les types de structures secondaires qui les entourent. La lettre α est utilisée pour une hélice α et la lettre β pour un brin β . Une connexion entre une hélice α et un brin β est appelée connexion $\alpha\beta$. L'identification de ces motifs et la description de leurs propriétés structurales et des caractéristiques en terme de séquence font partie d'un domaine actif de recherche depuis de nombreuses années. En effet, il existe sur ce sujet une abondante littérature, notamment sur les connexions $\alpha\alpha$, $\alpha\beta$, $\beta\alpha$ et sur les connexions $\beta\beta$ [53, 37, 38, 14, 47, 59, 15, 60, 45, 64]. La classification des motifs en tournant mentionnée précédemment est basée sur différentes mesures de la similarité entre structures. Parmi ces mesures, les valeurs des angles dièdres [59], la longueur du tournant et les patterns de liaisons hydrogènes [53, 14], et le r.m.s.d. (root mean square deviation) des distances entre les atomes C_α [47] sont utilisés. D'autres combinent plusieurs de ces valeurs comme dans l'étude faite par Wintjens et al. [60] combinant une classification basée sur les angles dièdres puis suivie d'une basée sur le r.m.s.d..

Les différentes classifications des motifs en tournant ont permis de décrire la présence de séquences consensus et d'analyser des propriétés structurales comme les liaisons hydrogènes et les contacts entre résidus. Le but de cette analyse est d'associer pour la première

fois une base de données de fragments issus de protéines obtenues par RMN, à des données expérimentales observées afin de déterminer des propriétés communes parmi les familles de motifs structuraux obtenus.

Les données expérimentales provenant de la RMN que nous allons étudier sont les contraintes de distances nOe. Les effets Overhauser nucléaires (nOe) sont parmi les paramètres les plus importants pour la détermination de la structure des protéines. L'intensité des pics nOe, obtenue à partir des expériences NOESY (Nuclear Overhauser Effect Spectroscopy), contient une information structurale de distance entre les protons éloignés de moins de 5 Å. En effet, le volume du pic nOe entre deux protons est fonction de leur distance. L'intensité η_{ij} entre deux protons i et j dépend de la distance r_{ij} moyenne entre ces deux protons selon la relation : $\eta_{ij} \propto \frac{1}{r_{ij}^6}$.

La RMN fournit ainsi des données structurales directement quantifiables comme le volume des pics nOe. A partir de ce jeu de distances inter-protons dérivé du volume des pics, la modélisation moléculaire a pour but d'exploiter au mieux toutes ces informations afin de construire un modèle tridimensionnel de la molécule. Son principe consiste, en explorant l'espace conformationnel, à obtenir un ensemble de structures qui satisfassent les données expérimentales et qui correspondent à des minima énergétiques. La liste de ces contraintes nOe obtenue est intrinsèquement incomplète. Des informations sont en effet perdues car des pics peuvent se chevaucher. Certains ont une trop faible intensité pour être observés, d'autres se trouvent dans des régions possédant trop de bruit de fond. Par conséquent, ces pics ne sont donc pas tous détectés. Cette étude est donc aussi intéressante pour déterminer quelles seraient les contraintes indispensables permettant de définir la structure d'un cluster de fragments.

Dans le but de mettre en relation des données structurales locales et des données expérimentales de RMN, il est nécessaire de les collecter pour ensuite les étudier. Les données RMN disponibles sont stockées à deux endroits. Les données structurales et les contraintes nOe se trouvent dans la PDB et les données sur les déplacements chimiques

sont déposées dans BioMagResBank. Toutes les structures déterminées par RMN et déposées dans la PDB n'ont pas systématiquement des fichiers de contraintes nOe associés. Quand ces fichiers de contraintes existent, ils sont dans des formats très différents, provenant de divers logiciels de modélisation, et ne pouvant donc pas être utilisés tels quels sans une étape de standardisation et de nettoyage préalable. Cette étape est fastidieuse et compliquée puisque même la PDB n'a pas encore résolu ce problème. L'espoir d'une meilleure organisation des informations vient de BMRB, mais cette base de données est spécialisée pour l'instant dans les déplacements chimiques. Et tous les déplacements chimiques déposés n'ont pas obligatoirement une structure tridimensionnelle existante dans la PDB. Nous avons donc décidé d'utiliser le set de 97 protéines issu du travail de thèse de Doreleijers [12] sur la validation des structures RMN de biomolécules au cours duquel la standardisation et le nettoyage des fichiers a été fait. Depuis, un travail de standardisation des fichiers de contraintes est en cours à la BMRB pour environ 1 280 protéines.

Dans ce chapitre, nous allons présenter les différentes étapes de la classification des fragments de protéines issus d'expériences RMN. Nous allons ainsi comprendre comment nous sommes passés des 97 protéines aux groupes de motifs structuraux, dans le but d'étudier leurs relations avec leurs caractéristiques expérimentales que sont les contraintes nOe.

Nous allons décrire ensuite le modèle utilisé pour modéliser les protéines en objets informatiques, et les étapes symbolisées par différents modules servant à la construction de ces objets. La lecture des fichiers texte au format NMR-STAR [20, 21, 54] contenant les données ainsi que les différentes étapes de la classification ont été programmées en langage Eiffel.

Nous présenterons ensuite les patterns de contraintes nOe observés pour différentes familles de fragments correspondant à des types de motifs différents. Une analyse faite par Wüthrich et al. [65] sur les distances inter protons observées dans des polypeptides polyalanines, offre une base à l'étude de la conformation des protéines obtenues à partir

des contraintes nOe. Une étude statistique de ces distances a aussi été réalisée à partir d'un groupe de 19 protéines étudiées par radiocristallographie. Les patterns de contraintes nOe que nous obtenons pour des motifs ayant une structure secondaire régulière, en hélice α ou en feuillet β , nous ont permis de valider les résultats obtenus et ont servi de base à l'étude des patterns de contraintes dans le cas des motifs en tournant.

Chapitre 2

Description de la méthode de clustering

2.1 Introduction

La méthode de classification employée est basée sur celle de Wintjens et al. [60] et se décompose en trois étapes principales. Cette classification a été adaptée au cas des structures de protéines déterminées par RMN puisqu'elles possèdent dans la plupart des cas des structures multiples appelées modèles. Ces modèles sont le résultat de la multitude de solutions rendues possibles suite aux calculs réalisés par l'intermédiaire de programme de modélisation à partir des données expérimentales de RMN.

2.2 Clustering et sélection des modèles représentatifs

Au cours de cette première étape, le but est d'obtenir un groupe de fragments de protéines représentatif de l'ensemble des modèles de structures, en éliminant les fragments qui ne correspondent pas à des conformations très peuplées. La protéine est tout d'abord découpée sur base de la séquence, en segments se chevauchant (figure 2.1 page 55, section 'découpage'). Quand la protéine est constituée d'un ensemble de structures, alors un ensemble de fragments est obtenu pour chaque segment de séquence. Ce groupe d'ensembles de fragments de longueur L est le point de départ de la classification. Chacun de ces ensembles, associé à un segment de séquence, sera considéré séparément pour la section suivante.

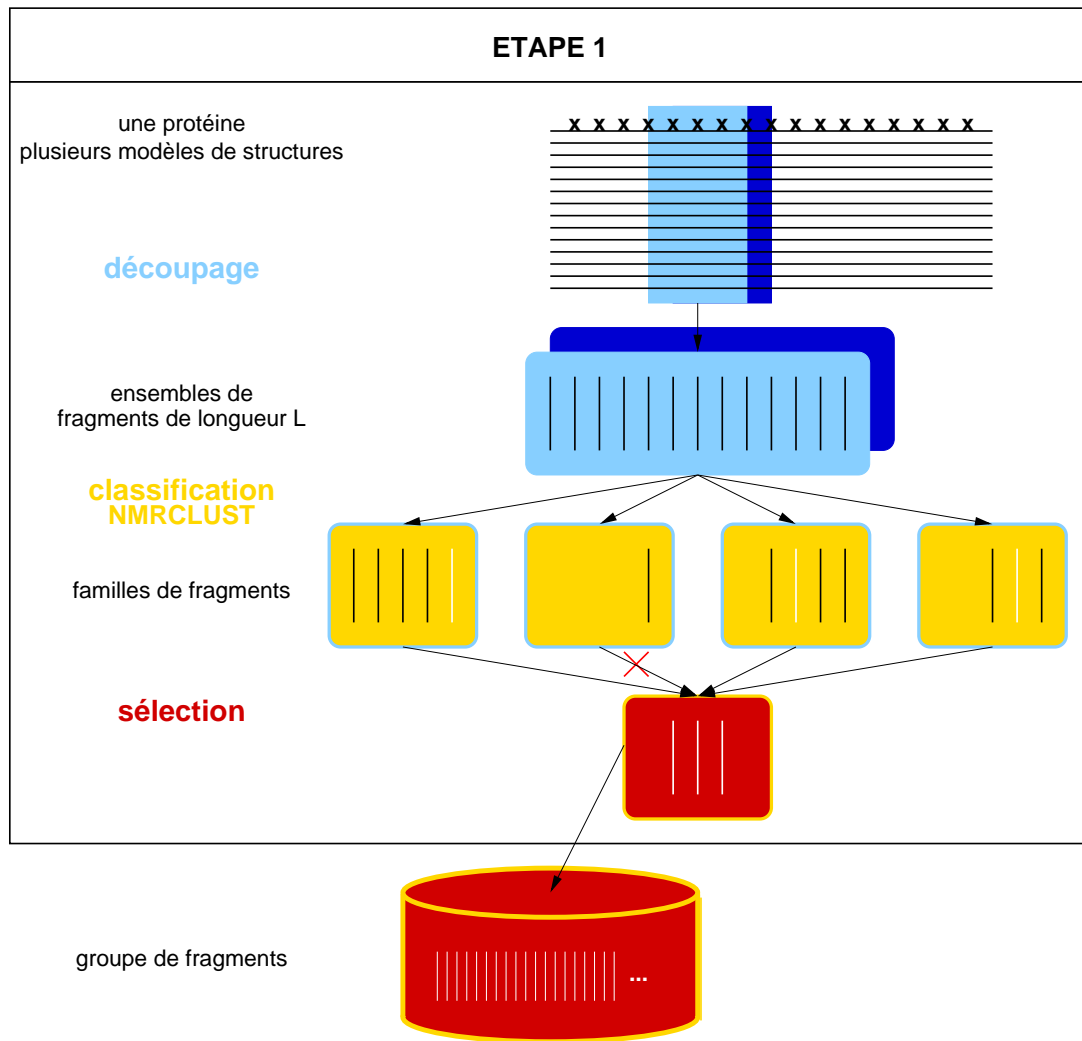


Fig. 2.1: Première étape de la classification. L'encadré noir montre la classification d'une seule protéine en fragments représentatifs. Les rectangles bleus correspondent à la section 'découpage', les jaunes à la section 'classification' et les rouges à la section 'sélection'. Les fragments sélectionnés feront partie du groupe de fragments du bas de la figure.

Pour sélectionner des fragments parmi un groupe, nous avons besoin tout d'abord de mesurer et de chiffrer leur degré de similitude ou de dissimilitude. Dans ce but, nous avons choisi un classique en terme de mesure de distance de dissimilitude entre structures tridimensionnelles : le r.m.s.d.. La distance de dissimilitude d_{ij} entre deux modèles i et j pour chaque atome k du squelette (N, C $_{\alpha}$, C et O) est calculée à l'aide de l'équation suivante :

$$d_{ij} = \sqrt{\frac{1}{N} \sum_{k=1}^N \sum_{r=x,y,z} (r_i^k - r_j^k)^2} \quad (2.1)$$

où N représente le nombre total d'atomes du fragment. Ce calcul est obtenu par l'algorithme de Kabsch du programme U3BEST [23, 24], après superposition des fragments

deux à deux. Une matrice de dissimilitude est alors générée à partir de cette distance calculée entre tous les fragments du sous-ensemble. Cette matrice D est une matrice triangulaire supérieure, ayant le même nombre de lignes et de colonnes, nombre égal au nombre des fragments constituant l'ensemble, égal aussi au nombre des modèles de la protéine :

$$D = \begin{pmatrix} 0 & \cdots & d_{ij} \\ & \ddots & \vdots \\ & & 0 \end{pmatrix} \quad (2.2)$$

A partir de cette matrice, le programme de clustering NMRCLUST [25] (figure 2.1 page 55, section 'classification') permet de regrouper les fragments de l'ensemble en familles de conformation similaire, et de sélectionner parmi ces familles un fragment représentatif (fragment en blanc sur la figure 2.1).

Le programme NMRCLUST développé par Kelley et al. [25] offre une approche automatique pour le clustering d'un ensemble de structures de protéines dérivées de la RMN en sous familles ayant une conformation similaire. La figure 2.2 (page 57) montre les différentes étapes permettant de passer d'une matrice des distances aux structures représentatives. Cette méthode a l'avantage de ne pas nécessiter la définition d'une valeur limite souvent subjective, permettant de couper l'arbre au meilleur endroit puisqu'une valeur de pénalité est calculée. Ce programme accepte en entrée, aussi bien des structures qu'il superpose sur la base d'un ensemble d'atomes défini par l'utilisateur, qu'une matrice des distances déterminée par ailleurs. Le programme NMRCLUST n'est donc utilisé ici qu'à partir de l'étape 2 (figure 2.2 page 57), lors de la classification des motifs structuraux.

Tous les fragments représentatifs de chacune des familles ne sont pas conservés pour la prochaine étape de la classification (figure 2.1 page 55, section 'sélection'). Les familles possédant moins de 20% du nombre total de fragments ont été éliminées, ainsi que les fragments ayant moins de 90% des résidus dans les zones favorables de la carte de Ramachandran (voir figure 1.1 page 107, chapitre 1 de la partie III). Tous les autres fragments, provenant de toutes les protéines étudiées, constituent le nouveau groupe de fragments (en bas de la figure 2.1 page 55), point de départ de l'étape suivante.

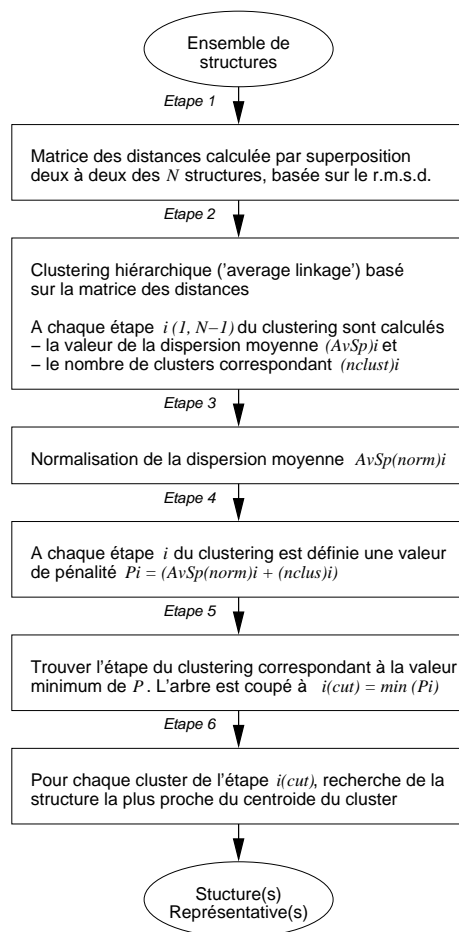


Fig. 2.2: Diagramme illustrant la progression de l'algorithme implémenté dans le programme NMR-CLUST [25].

2.3 Familles ayant des signatures de Ramachandran identiques

A chacun des résidus des fragments issus du groupe précédent est assigné un code basé sur les valeurs des angles ϕ , ψ et ω , correspondant à une région spécifique de la carte de Ramachandran (figure 2.3 page 58). Sept régions [47] ont été définies : six pour les conformations *trans* ($\omega \simeq \pm 180^\circ$, ω étant l'angle de la liaison peptidique précédant le résidu) notées A, C, G, B, E ou P ; et une pour les conformations *cis* ($\omega \simeq 0^\circ$) notée O. A représente les conformations en hélices α droites, C les conformations en hélices 3_{10} droites, G celles en hélices gauches, B les conformations en feuillet β droits, E les conformations en feuillet β gauches et P les conformations étendues. Lorsque les valeurs des angles ϕ , ψ et ω sont en dehors des régions définies, le code assigné au résidu est noté \star .

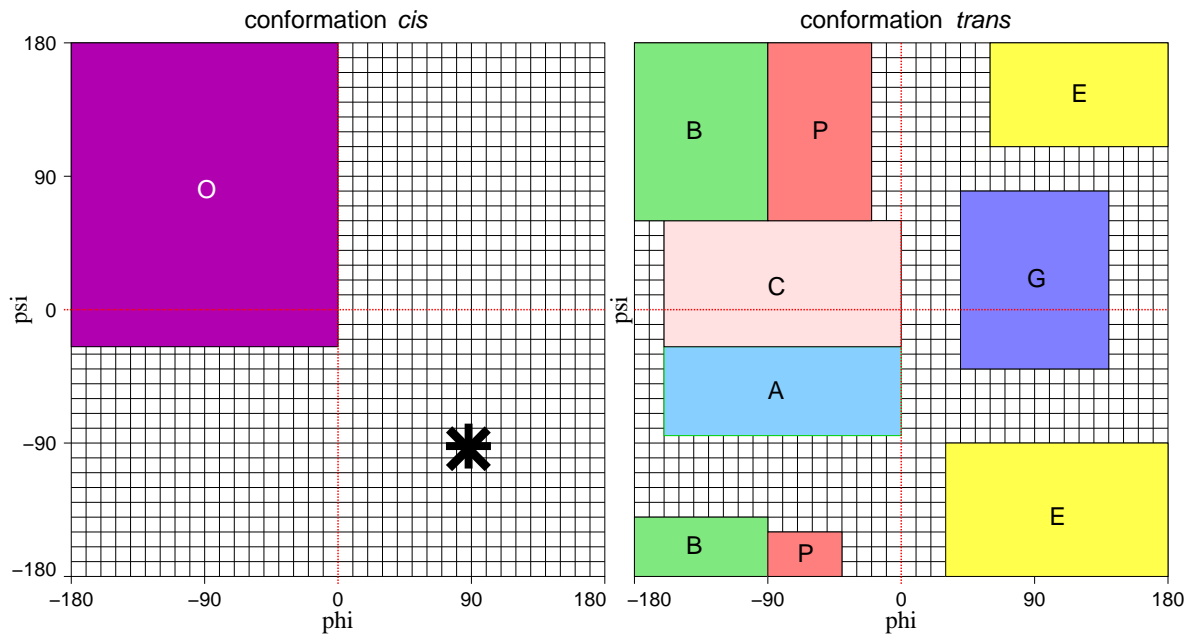


Fig. 2.3: Cartes de Ramachandran définissant les 7 régions [47]. La carte de gauche correspond aux conformations *cis* avec $-30^\circ \leq \omega \leq 30^\circ$, et celle de droite correspond aux conformations *trans* avec $\omega \leq -150^\circ$ ou $\omega \geq 150^\circ$.

La mesure utilisée ici pour chiffrer le degré de similitude entre deux fragments est basée sur les valeurs des angles ϕ , ψ et ω . Deux fragments peuvent être considérés comme ayant des conformations similaires quand les valeurs des angles dans chacun des fragments sont dans les mêmes régions de la carte de Ramachandran. Ces fragments appartiendront alors à la même famille ayant pour caractéristique la signature de Ramachandran, constituée des huit codes différents. La figure 2.4 (page 59) schématise cette étape de la classification des fragments, classification basée sur les angles ϕ , ψ , ω , en reprenant le groupe de fragments de l'étape précédente, afin d'aboutir au groupe de familles d'ensembles de fragments.

2.4 Clustering hiérarchique

La classification hiérarchique est une des méthode utilisée pour regrouper des éléments. Une mesure de distance entre ces éléments permet de générer une matrice de dissimilitude. La classification obtenue est représentée sous forme d'un arbre dont chaque feuille est un cluster d'éléments. Les éléments les plus proches sont agrégés pour former un noeud dont la distance par rapport aux autres éléments est recalculée par une méthode d'agrégation

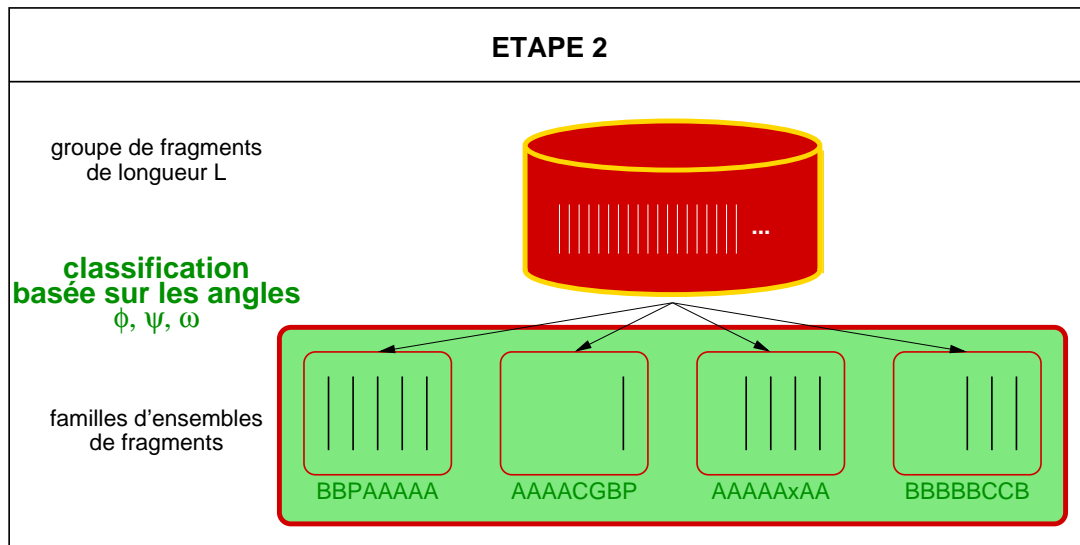


Fig. 2.4: Deuxième étape de la classification.

qui peut-être la moyenne des distances des éléments formant le noeud ('average linkage'), la distance minimale ('single linkage'), ou la distance maximale ('complete linkage'). On procède ainsi jusqu'à ce que les deux derniers noeuds soient agrégés, on est alors à la racine de l'arbre. Le choix de la distance et de la méthode d'agrégation à utiliser est souvent difficile. Un problème courant avec cette méthode consiste à déterminer où couper l'arbre de manière à avoir un nombre de classes maximisant la variabilité inter-classes et minimisant celle intra-classe.

A chacune des familles ayant une signature de Ramachandran identique est appliquée une classification hiérarchique, basée sur la même distance de dissimilitude qu'à la première étape : le r.m.s.d.. L'algorithme crée une hiérarchie des clusters, les éléments les plus similaires sont regroupés dans des clusters aux plus bas niveaux, tandis que les éléments moins similaires sont regroupés dans des clusters aux plus haut niveaux. Celui utilisé ici est un algorithme agglomératif qui tente de regrouper deux clusters en un plus grand.

Tout d'abord, la distance de dissimilitude entre les clusters est calculée. Au début, chaque élément forme un cluster à lui seul. Puis les deux clusters les plus similaires sont regroupés, c'est à dire les deux clusters ayant la plus petite valeur de dissimilitude. Ces deux opérations sont répétées tant que la valeur de dissimilitude est inférieure à une valeur limite θ égale à $\theta = 1.25 + \frac{L}{10}$ où L est la longueur des fragments. A chaque étape,

la distance entre le nouveau cluster et les autres est recalculée, elle correspond à la plus petite distance entre chacun des éléments du nouveau cluster et chaque ancien cluster. La méthode d'agrégation employée ici est donc celle de la distance minimale ('single linkage'). La figure 2.5 montre cette dernière étape de la classification des fragments, le clustering hiérarchique, appliqué à chacune des familles issues de l'étape précédente.

Le choix de la valeur de θ est basé sur celui déterminé lors d'une précédente étude [60]. La présence d'une discontinuité de la valeur moyenne du r.m.s.d. au cours des différentes étapes du clustering n'a pas été observée lors de cette étude, mais une marche sur cette courbe a permis de déterminer cette valeur θ . Cette courbe ayant été obtenue en utilisant un ensemble de fragments test et en lui appliquant seulement la méthode de clustering hiérarchique, il nous a semblé raisonnable d'utiliser cette même valeur dans des conditions similaires.

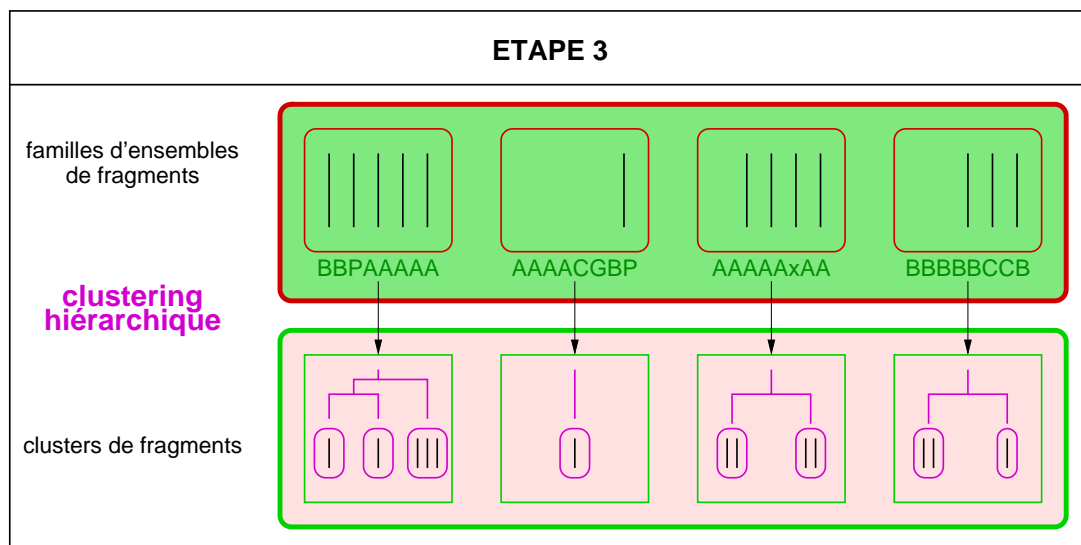


Fig. 2.5: Troisième étape de la classification.

La méthode de classification en trois étapes, basée à la fois sur le r.m.s.d. et sur les valeurs des angles dièdres, est utilisée pour déterminer des motifs structuraux récurrents dans la base de données des 97 protéines, et pour étudier les relations avec leurs empreintes expérimentales. Afin de mettre en évidence ces différentes caractéristiques et de les analyser, nous allons présenter les outils informatiques développés qui nous ont permis d'atteindre cet objectif.

Chapitre 3

Aspects informatiques et outils développés

3.1 Introduction

Le but du développement de ces outils et de l'utilisation d'un langage orienté objet est d'offrir la possibilité d'une réutilisation d'une partie ou de l'intégralité des outils développés. La justification de l'utilisation des objets plutôt qu'une approche traditionnelle utilisant des fonctions comme base de l'architecture des logiciels, est fondée sur les objectifs de qualité du logiciel que sont en particulier l'extensibilité, la réutilisation et la compatibilité.

Les fonctions d'un système ont souvent tendance à changer tandis que les types d'objets manipulés restent les mêmes. En effet quoi qu'il arrive à une structure tridimensionnelle de protéine, le système continuera à manipuler des points, des vecteurs et des listes de coordonnées. L'objectif d'extension du système est dans ce cas envisageable.

La possibilité d'exploiter les ressemblances de différents systèmes logiciels plutôt que de réinventer des solutions à des problèmes qui se sont déjà posés est l'un des objectifs majeur de l'utilisation des types d'objets. Ils fournissent des unités réutilisables mais aussi adaptables à chaque cas particulier.

Un autre facteur de la qualité logiciel, la compatibilité, a été défini comme étant la facilité avec laquelle les modules peuvent être combinés entre eux.

Notre choix s'est porté sur le langage Eiffel car nous y avons retrouvé les aspects de qualité évoqués ci-dessus. De plus, ce langage est un excellent moyen d'apprentissage de la technique orientée objet. Il permet en effet d'aborder les concepts d'héritage de façon simple et claire et d'utiliser le typage statique.

3.2 Description du module décrivant une molécule

Le concept central sur lequel repose l'ensemble de la technologie objet est la classe, type abstrait de données muni d'une implémentation éventuellement partielle. Les objets construits à partir des ces classes sont des instances de classes. Nous allons présenter celles que nous avons construites pour modéliser une molécule tridimensionnelle possédant un ensemble de modèles. Les objets alors créés sont les données de base sur lesquelles la méthode de classification a été appliquée.

Le modèle orienté objet utilisé afin de représenter les protéines issues d'expérience RMN est illustré dans le diagramme de classe de la figure 3.1 (page 63). Celui-ci met en évidence une structure hiérarchique de l'atome à la molécule. L'agrégation, représentée par une flèche en forme de losange, est une relation de type 'ensemble/élément'. Un résidu ('RESIDUE') est un ensemble d'atomes ('ATOM'). L'héritage permet la classification des objets. Si 'ATOM' hérite de 'POINT3D', cela signifie qu'un atome est une sorte particulière de point 3D. Il est représenté par une flèche triangulaire. Une association est une relation statique décrivant des liens entre classes. Elle est symbolisée par une ligne.

3.3 L'extraction des données

La construction des molécules s'est faite en lisant les données dans des fichiers au format NMR-STAR [20, 21, 54]. A partir de la grammaire connue de ce type de fichier, nous avons utiliser *Lex*, un analyseur lexical, et *Yacc*, un analyseur syntaxique pour les lire. Ces outils ont été développés pour pouvoir créer des compilateurs. Un compilateur est un programme qui lit un langage et qui le traduit par un programme équivalent dans

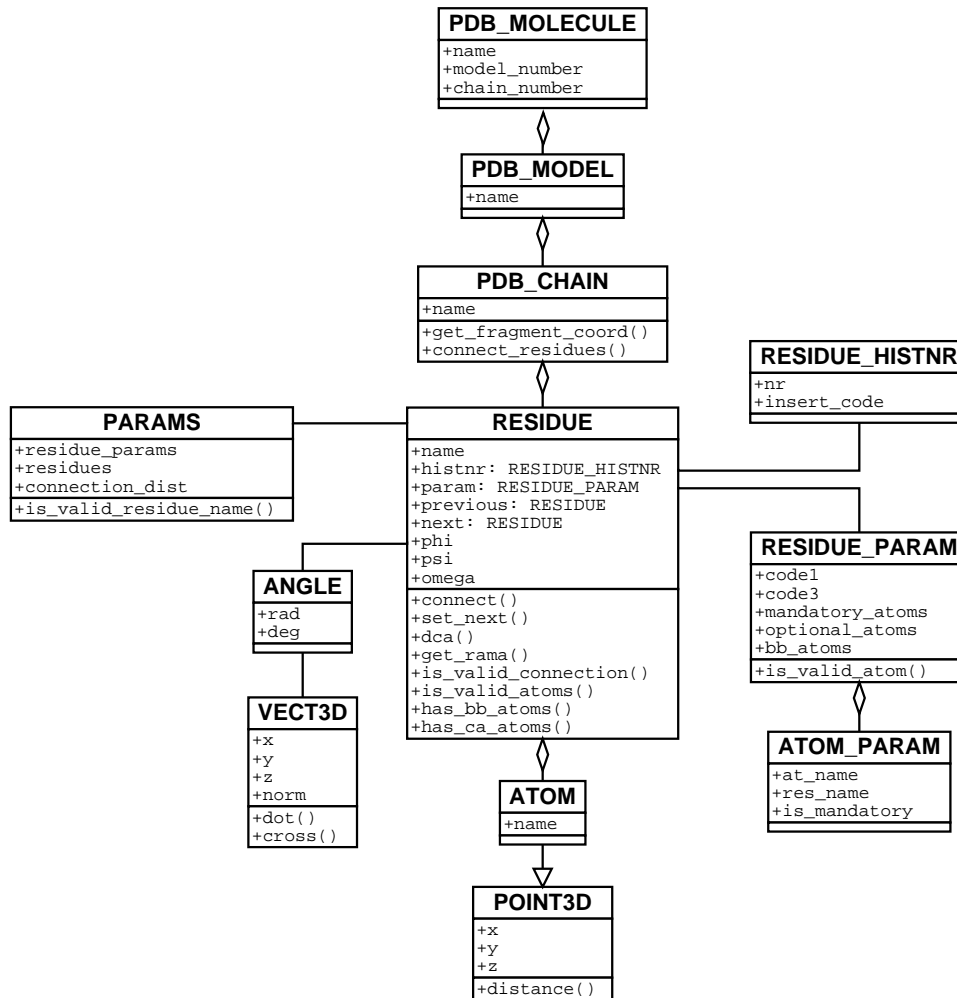


Fig. 3.1: Diagramme de classes. Les rectangles symbolisent les classes constituées de trois parties. La première contient le nom de la classe, la seconde ses attributs et la dernière les opérations associées. La flèche en forme de losange représente une agrégation et celle en forme de triangle une relation d'héritage. Une ligne entre deux classes représente une association.

un autre langage en rapportant les erreurs éventuelles. Ces outils sont donc bien adaptés à notre problème, de plus leur fiabilité et leur rapidité ont été mises à l'épreuve depuis longtemps. Le rôle de l'analyse lexicale est de fournir une suite d'éléments ou de marques à partir d'une suite de caractères en entrée. Par exemple, dans la déclaration suivante (figure 3.2 page 64), l'analyseur lexical va identifier les caractères du fichier et les associer aux marques 'DATA_NAME' et 'TEXT_STRING'. L'analyse syntaxique associée est une analyse hiérarchique du fichier. L'objectif de cette analyse est de regrouper ces marques fournies par l'analyse lexicale en phrases grammaticales et de leur associer des actions. En reprenant l'exemple précédent, nous pouvons construire l'arbre syntaxique abstrait

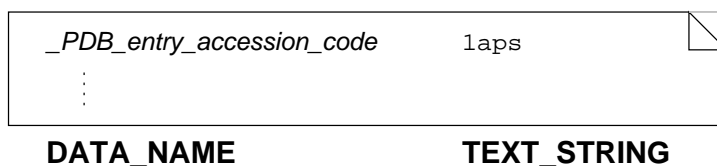


Fig. 3.2: Exemple illustrant l'analyseur lexical. Le rectangle symbolise le fichier contenant un ensemble de caractères. Les mots en gras représentent les marques associées.

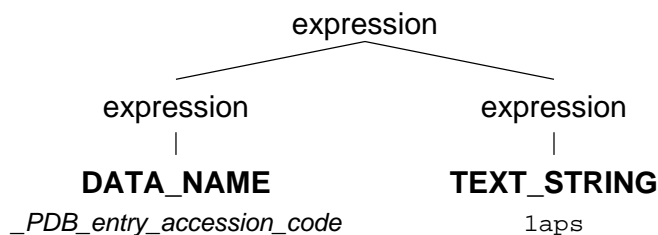


Fig. 3.3: Exemple illustrant l'arbre syntaxique.

présenté sur la figure 3.3. Dans ce cas, l'action associée à cette expression est d'assigner la valeur de la marque 'TEXT_STRING' à une variable.

Nous avons utilisé l'outil Gobo Eiffel Lex (Gelex) pour générer une classe Eiffel équipée des routines pour analyser le fichier texte en entrée et obtenir un ensemble de marques. Quand les expressions régulières sont trouvées, le code Eiffel correspondant est alors exécuté. Puis, de la même manière l'outil Gobo Eiffel Yacc (Geyacc) a été utilisé pour générer une classe Eiffel équipée de routines pour analyser la grammaire constituée par l'ensemble des marques issu de l'analyseur lexical précédent. Cette étape permet de convertir une grammaire en classe Eiffel et d'y associer un groupe d'actions, permettant de créer un atome, puis un résidu et donc de construire les molécules.

3.4 Le clustering des données

A partir des objets sérialisés représentant des molécules biologiques issues d'expériences RMN et qui sont construits à l'aide de l'étape précédente, le clustering de ces données a pu être effectué.

Les trois étapes de la méthode décrites dans le chapitre 2 (partie II) ont été matérialisées par différents modules pouvant être exécutés séparément. Ces modules permettent de

transformer les objets molécules de départ en fragments, puis de les grouper suivant des critères bien précis, définis précédemment lors de la description des étapes de la classification. Donc en partant d'une base de données d'objets 'PDB_MOLECULE', le clustering permet d'obtenir une base de données de fragments de protéines groupés en clusters.

Parmi ces modules, celui du clustering hiérarchique a utilisé une routine existante, pour superposer deux vecteurs de points, qui est à la base du calcul de la matrice de dissimilarité. Elle provient du programme U3BEST écrit en Fortran par Kabsch [23] [24]. Cette routine permet d'obtenir le r.m.s.d. après une superposition optimale des vecteurs de coordonnées des deux fragments considérés. Une interface entre cette routine traduite en langage C et une classe Eiffel a été implémentée afin de permettre son utilisation.

3.5 L'analyse des patterns de contraintes nOe

A chacun des fragments ont été associées les contraintes nOe issues de la lecture des fichiers NMR-STAR. Des fichiers de résultats tabulés ont alors été générés afin d'être directement utilisable par le logiciel R [22]. Les graphiques de résultats représentant les contraintes nOe observées le long de la séquence pour les fragments sélectionnés sont alors obtenus automatiquement.

Chapitre 4

Analyses des motifs structuraux et de leurs caractéristiques expérimentales

4.1 Introduction

La méthode de classification présentée au chapitre 2 (partie II), basée à la fois sur le r.m.s.d. entre les fragments et sur les valeurs des angles ϕ , ψ , ω a permis d'organiser en clusters les différents fragments extraits des 97 protéines de départ.

Nous présentons dans ce chapitre une analyse générale des résultats, afin de mettre en évidence le nombre restreint de fragments utilisés pour les analyses suivantes et leurs éliminations au cours des différentes étapes de la classification. Les clusters obtenus sont ensuite ordonnés par type de motifs et les patterns de contraintes nOe associés aux motifs sont analysés. Une analyse plus détaillée de quatre motifs est présentée montrant les relations entre les motifs étudiés et les contraintes nOe observées.

4.2 Analyse générale des résultats

Le tableau 4.1 (page 69) présente l'évolution du nombre de fragments au cours des différentes étapes de la classification, pour les fragments de 10 résidus de longueur. Le groupe de données de départ est constitué de 97 protéines, découpées en segments de 10 résidus se chevauchant, formant un groupe de 124 831 fragments répartis dans 5 410

ensembles, soit en moyenne 23 fragments par ensemble. Cette valeur est aussi le nombre moyen de modèles par protéine. En moyenne, plus de 3 familles de fragments sont obtenues par ensemble après la classification faite par NMRCLUST, 19 886 fragments représentatifs sont ainsi sélectionnés. Plus de la moitié sont éliminés au cours de l'étape de sélection qui consiste à supprimer les fragments représentatifs de groupes peu peuplés, possédant moins de 20% de la population des ensembles précédents.

Le choix de la limite des 20% est arbitraire mais il nous a semblé raisonnable. Des groupes possédant moins de 20% du nombre total des fragments de départ, ne sont pas représentatifs des conformations prédominantes de l'ensemble dont ils sont issus, ils sont donc éliminés.

Les deux premières catégories de l'histogramme de la figure 4.1 montre que plus de 10 000 groupes se situent en effet sous le seuil des 20%. L'augmentation du nombre de groupes pour la dernière catégorie par rapport à la précédente est due à la présence d'un certain nombre de structures n'ayant qu'un seul modèle, donc un seul fragment par groupe et se retrouvant alors dans cette dernière catégorie.

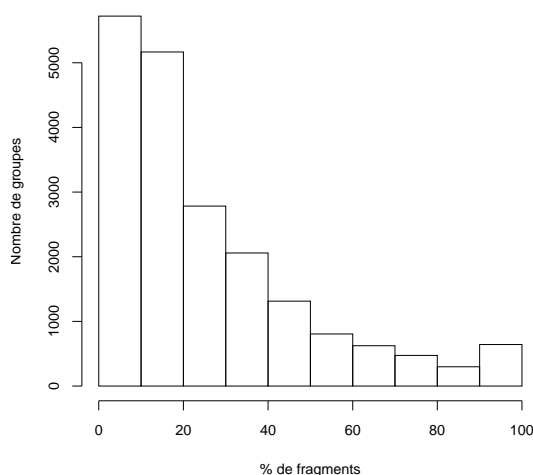


Fig. 4.1: Distribution de la population des groupes par rapport à la population totale des ensembles que constituent ces groupes, pour les fragments de 10 résidus après la classification faite pas NMRCLUST.

Après le critère numéraire, une autre sélection est faite sur un critère géométrique : celui des angles ϕ , ψ des résidus des fragments. Ce critère entraîne donc la suppression des

fragments ayant plus de 10% des résidus dans la zone la plus défavorable de la carte de Ramachandran. Pour des fragments de 10 résidus, sont éliminés ceux possédant plus d'un résidu dans la région extérieure, région nommée 'o' (voir figure 1.1 page 107, chapitre 1 de la partie III). Les structures sélectionnées doivent être de qualité et le choix de ce critère arbitraire, afin d'éliminer des fragments, nous a semblé adéquat. La figure 4.2 détaille la fréquence des résidus dans les différentes régions de la carte de Ramachandran. Le choix d'éliminer ces fragments, permet de n'en conserver pratiquement aucun ayant des résidus dans la région extérieure 'o' (figure 4.2 (a)). En moyenne seulement 0.6% des résidus constituant un fragment, se trouvent dans cette région après cette sélection. La figure 4.2 (b) montre que le groupe des fragments ayant plus de 90% de ses résidus dans les zones les plus favorables (g+a+c) possède deux fois moins de résidus (5.8% en moyenne, soit moins d'un résidu par fragment) dans la zone '*' (résidus hors zone), que le groupe complémentaire (11,3% en moyenne). Celui-ci possède moins de 90% de ses résidus dans les zones 'g', 'a' et 'c'. Le cumul des régions en hélices α , A et C, représente globalement plus de 50% des résidus d'un fragment.

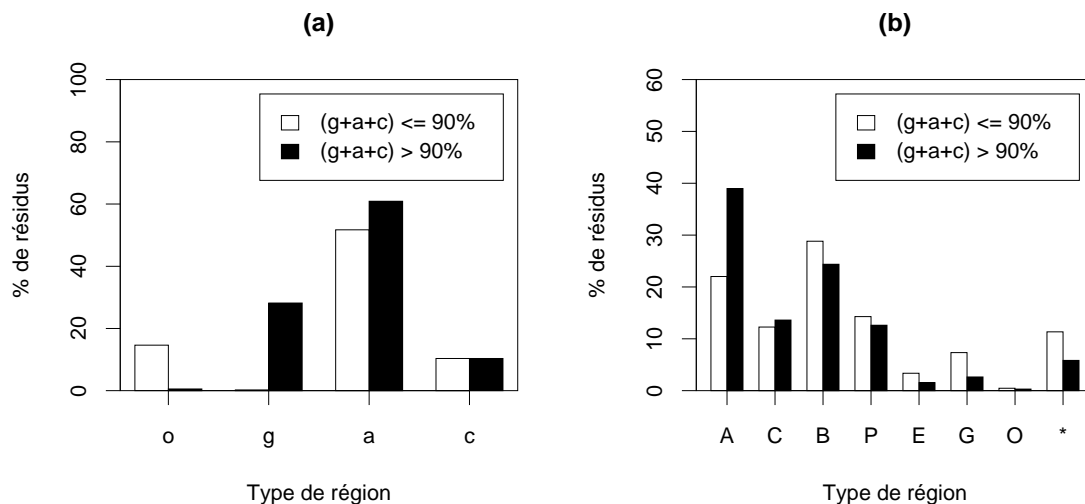


Fig. 4.2: Fréquence des résidus, dans les fragments de 10 résidus, répartis dans les différentes régions de la carte de Ramachandran. (a) représente les régions plus ou moins favorables (o : extérieur, g : acceptable, a : favorable et c : très favorable) tandis que (b) correspond aux régions liées aux structures définies dans le chapitre 2 (partie II). Les barres blanches correspondent aux fragments possédant moins de 90% des résidus dans les zones g, a et c, les noires correspondent aux fragments possédant plus de 90% des résidus dans ces 3 zones.

Après le passage de l'étape 2, les fragments ayant la même signature de Ramachandran sont regroupés en 785 familles, constituées chacune d'au moins 2 fragments. Moins de 1% des fragments du départ constitue les 71 clusters, représentant le résultat final de la classification.

Étape	Description	Nombre d'entités	Nombre de groupes
		97 protéines	
1a	Découpage	124 831 fragments	5 410 ensembles
1b	Classification NMRCLUST	19 886 fragments	-
1c	Sélection du fragment représentatif d'un groupe si le nombre de fragments dans le groupe représente au moins 20% du nombre total.	8 996 fragments	-
1d	Sélection des fragments ayant plus de 90% de résidus dans les régions favorables de la carte de Ramachandran.	7 484 fragments	-
2	Classification basée sur les angles ϕ , ψ , ω et sélection des groupes de plus d'un élément.	3 519 fragments	785 familles
3	Clustering hiérarchique et sélection des clusters ayant au moins 5 éléments.	887 fragments	71 clusters

Tab. 4.1: Tableau des résultats pour les fragments de 10 résidus.

Les résultats concernant les autres fragments de L résidus sont consignés dans le tableau 4.2 (page 70). Contrairement aux fragments de 10 résidus, ceux de 7 possèdent un nombre de clusters beaucoup plus important offrant la possibilité d'une étude des motifs. Plus la longueur des fragments est importante et plus le nombre de motifs obtenus est réduit. L'étude porte donc essentiellement sur les motifs de 7 et 10 résidus.

Étape	L = 7		L = 15		L = 20	
	Entités	Groupes	Entités	Groupes	Entités	Groupes
1b	131 831	6 277	113 391	5 350	102 207	4 784
1c	20 988	-	17 883	-	15 939	-
1d	9 609	-	8 021	-	7 168	-
2	8 059	-	7 460	-	6 404	-
3	6 577	1 017	1 663	575	806	347
	2 260	166	186	20	33	4

Tab. 4.2: Tableau des résultats pour les autres fragments de L résidus.

4.3 Les différents motifs obtenus

Les motifs obtenus à la suite de la classification peuvent être répartis en trois catégories : les fragments en hélice α (régions A et C), les fragments en feuillet β (régions B et P) et les fragments mixtes. Le tableau 4.3 (page 70) présente la répartition des types de motifs et montre que pour les fragments de 10 résidus, la plupart des 887 fragments sont des fragments ayant un motif en hélice α . Ceux-ci sont des éléments de structures secondaires prédominants des structures de départ (figure 4.2 page 68, (b)), que nous retrouvons logiquement ici. Les motifs mixtes connectent les éléments de structures secondaires, et représentent des régions variables en longueur et en conformation : les motifs en tournant. Les fragments de 7 résidus, plus courts, peuplent pour plus de la moitié d'entre eux (99/166) les catégories de motifs en tournant. Certains motifs structuraux particuliers sont davantage observés quand la longueur de la région entre deux éléments de structures secondaires est petite.

Catégorie	Nombre de clusters	
	L = 7	L = 10
Hélice α	33	39
Feuillet β	34	4
Mixte	99	28
Total	166	71

Tab. 4.3: Répartition des types de motifs.

La nomenclature des familles de tournants est basée sur les types de structures secondaires qui bordent le tournant, et sur les régions ϕ , ψ , ω des résidus constitutifs du celui-ci. La lettre α est utilisée pour une hélice α et la lettre β pour un brin β . Par exemple, la famille de motifs ayant pour signature 'ACGBB' est appelée $\alpha G\beta$. Elle représente une connexion entre une hélice α et un brin β dans laquelle le tournant est constitué par un résidu en conformation G. La nomenclature ainsi définie rend compte uniquement des groupes obtenus lors de l'étape basée sur les valeurs des angles dièdres de la chaîne principale (étape 2). Lorsqu'une famille est séparée en deux clusters suite à la dernière étape de la classification, la nomenclature doit être adaptée.

Les positions dans les fragments sont désignées quant à elles en utilisant la nomenclature décrite par Edwards et al. [14]. Les résidus dans les hélices α sont nommés par la lettre A, ceux dans les brins β par la lettre B, et ceux dans les tournants par la lettre L. Ces trois lettres sont suivies par un chiffre qui indique la position du résidu par rapport au tournant. Les résidus dans les structures secondaires sont numérotés par rapport aux extrémités de la région en tournant, par contre les résidus du tournant sont numérotés du N- vers le C-terminal. Par exemple, les résidus du motif $\alpha G\beta$ sont désignés par : ..., $-A2$, $-A1$, $L1$, $B1$, $B2$,

Le tableau 4.4 (page 72) met en évidence les clusters les plus représentatifs par catégories de motifs. Parmi les différents motifs obtenus, ceux n'ayant pas les résidus formant le tournant centrés ne sont pas pris en compte dans cette classification. Cela permet d'éviter la redondance due à la sélection contiguë des segments le long de la séquence. La diversité des motifs obtenus ici n'est pas très importante au regard de celle obtenue lors d'études de classification de motifs en tournant. L'ensemble des protéines de départ n'a pas été sélectionné dans ce but.

13 familles de motifs en tournant, contenant au moins 5 membres, sont identifiées, contenant parmi elles 4 types de connexions $\alpha\alpha$, 3 $\alpha\beta$, 3 $\beta\alpha$ et 3 $\beta\beta$.

Catégorie	Cluster
$\alpha\alpha$	
$\alpha B\alpha$	7.17, 7.128 , 10.34, 10.56
$\alpha GB\alpha$	10.64
$\alpha GBB\alpha$	7.147
$\alpha BBB\alpha$	7.9, 7.45
$\alpha\beta$	
$\alpha\beta$	7.31 , 7.63, 7.156, 10.12
$\alpha G\beta$	7.60, 7.92, 7.166, 10.19, 10.66
$\alpha BA\beta$	7.62, 7.160
$\beta\alpha$	
$\beta\alpha$	7.15, 7.47, 7.49, 7.71, 7.81, 7.101, 7.105, 7.106, 7.111, 7.113, 7.116, 7.125, 7.127, 7.145, 7.146, 7.157, 7.163, 10.10 , 10.15, 10.26, 10.27, 10.30
$\beta AB\alpha$	7.76, 7.107, 7.109, 7.149, 10.68
$\beta AAB\alpha$	7.19
$\beta\beta$	
$\beta A\beta$	7.27, 7.28, 7.108
$\beta G\beta$	7.36
$\beta AA\beta$	7.18, 7.85 , 10.41

Tab. 4.4: Répartition des motifs en tournant par catégorie. Le premier chiffre du cluster indique le nombre L de résidus dans les fragments et celui après le point est un identifiant. Les clusters en **gras** sont ceux étudiés.

4.4 Les patterns de contraintes nOe dans les structures régulières

Les patterns de contraintes nOe observés dans les structures régulières sont connus depuis longtemps. L'analyse faite par Wüthrich et al. [65] sur les distances inter protons observées dans des polypeptides polyalanines, offre une base à la validation de notre méthode.

4.4.1 L'hélice α

L'hélice α est caractérisée par des résidus i , $(i + 3)$ et i , $(i + 4)$ proches dans l'espace, entraînant la présence possible de distances inter protons de type $HN_i - HN_{i+1}$, $HA_i - HN_{i+3}$ et $HA_i - HN_{i+4}$. La figure 4.3 (page 74) met en évidence un réseau $HN_i - HN_{i+1}$ (graphe (b)) de contraintes nOe très fréquentes, puisqu'elles apparaissent dans plus de

50% des cas. La présence de contraintes $HN_i - HN_{i+2}$ est remarquable ainsi que dans une moindre mesure celles $HN_i - HN_{i+3}$ et $HN_i - HN_{i+4}$, plus disparates. Un nombre important de contraintes $HA_i - HN_j$ (graphe (a)) avec j compris entre i et $i + 4$ est présent, parmi elles les connexions $HA_i - HN_{i+3}$ ont une fréquence plus élevée. Nous retrouvons, dans le cas du motif présenté, le réseau dense de contraintes nOe des 4 résidus consécutifs. Il est intéressant de remarquer aussi la présence des contraintes $HA_i - HB_{i+3}$ sur le graphe (c). Aucune contrainte $HA - HA$ n'est intéressante à noter à partir du graphe (d).

La figure 4.4 (page 75) montre les contraintes nOe observées dans le cas des chaînes latérales. Les chaînes latérales sont caractérisées par tous les atomes, excepté les HN , les HA et les HB . Pour la représentation de ces contraintes, nous avons choisi de ne compter qu'une contrainte de même type par fragment. Par exemple, plusieurs contraintes peuvent être présentes dans un cas comme $HN_5 - SC_4$, mais une seule est comptabilisée par fragment. La fréquence observée sur le graphique est donc l'occurrence de cette contrainte observée au sein de la famille de fragments. Un réseau dense de contraintes est observé sur le graphe (a) de la figure 4.4 (page 75) mettant en jeu les contraintes $HB_i - HN_i$ et $HB_i - HN_{i+1}$. Les cinq graphiques suivants (de (b) à (f)) présentent des contraintes beaucoup moins fréquentes, impliquant dans la majorité des cas des résidus éloignés de quatre résidus au maximum. Le réseau constitué des contraintes impliquant les chaînes latérales met encore en évidence la caractéristique des hélices α : elles possèdent des résidus i , $(i + 3)$ et i , $(i + 4)$ proches dans l'espace.

Au regard des fréquences observées, l'hélice α est caractérisée par une présence importante de six types de contraintes $HA_i - HN_{i+1}$, $HA_i - HN_{i+3}$, $HN_i - HN_{i+1}$, $HA_i - HB_{i+3}$, $HB_i - HN_i$ et $HB_i - HN_{i+1}$.

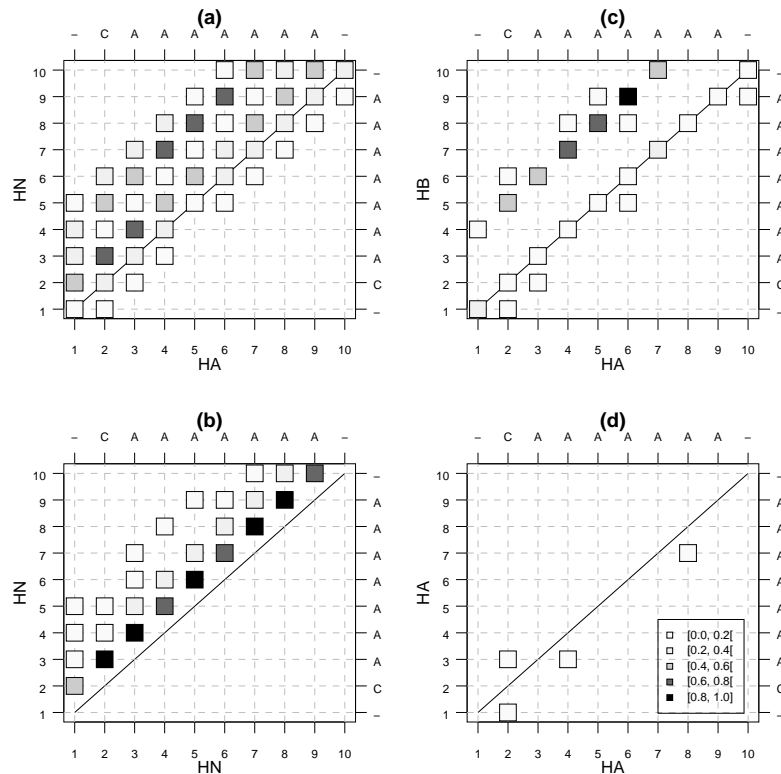


Fig. 4.3: Cartes des fréquences des contraintes nOe observées pour une hélice α dans un cluster de 40 fragments ayant pour signature de Ramachandran -CAAAAAA-. (a) représente les fréquences des contraintes nOe $HA_i - HN_j$, (b) celles pour $HN_i - HN_j$, (c) $HA_i - HB_j$ et (d) celles pour $HA_i - HA_j$.

4.4.2 Le brin β

Dans le cas des brins β , les distances intra-brins autres que celles entre les résidus voisins sont trop grandes pour être observées par RMN. La figure 4.5 (page 76) montre en effet que seules les contraintes $i, i + 1$ sont présentes dans ce cas. Le réseau de contraintes est beaucoup moins dense que dans le cas des hélices α . La présence des contraintes $HA_i - HN_{i+1}$ est une caractéristique connue pour ce type de structures secondaires.

La figure 4.6 (page 76) représentant les contraintes impliquant les chaînes latérales met en évidence la présence de contraintes $i, i + 2$. Les contraintes $SC_i - SC_{i+2}$ impliquant les chaînes latérales sont présentes, certaines contraintes $SC_i - HA_{i+2}$, $SC_i - HB_{i+2}$ et $SC_i - HN_{i+2}$ sont aussi présentes mais elles sont moins fréquentes. Elles sont la caractéristique d'un brin β où les résidus sont alternés, l'un au dessus puis l'autre en dessous du squelette. Nous pouvons noter la présence des contraintes $HB_i - HN_i$ et $HB_i - HN_{i+1}$ comme dans

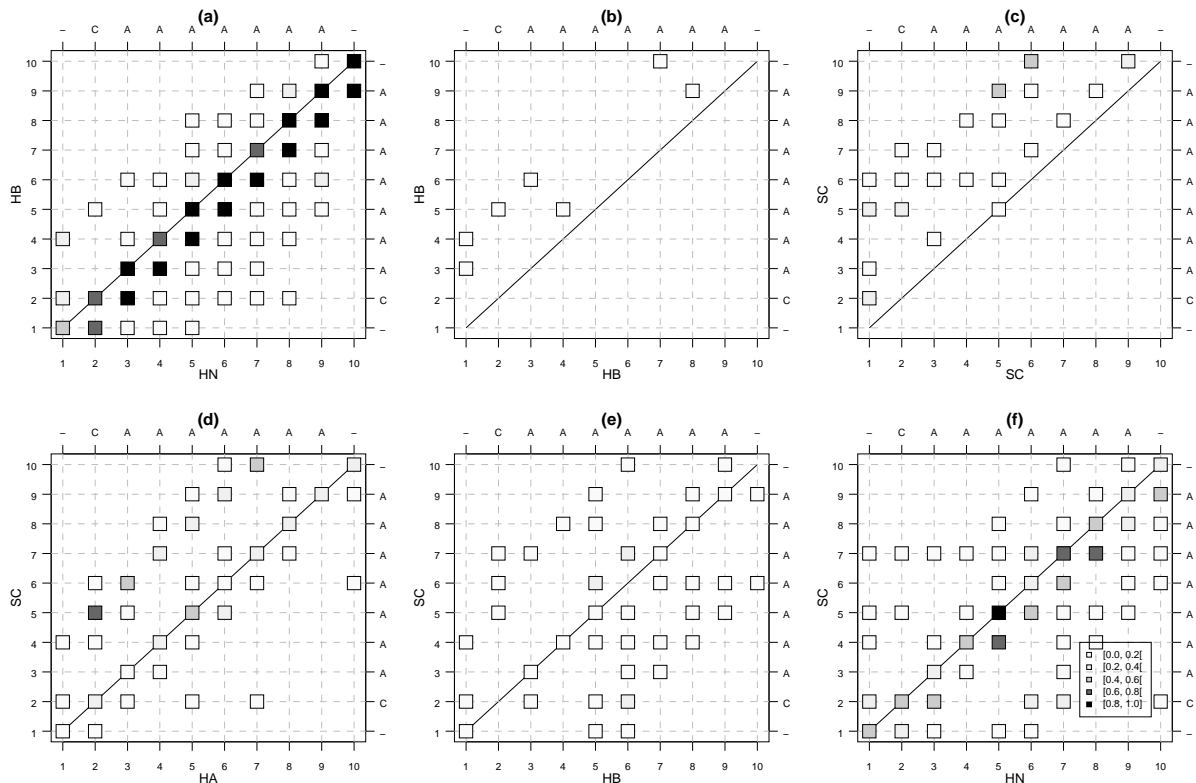


Fig. 4.4: Cartes des fréquences des contraintes nOe observées pour une hélice α dans un cluster de 40 fragments ayant pour signature de Ramachandran -CAAAAAAA-. (a) représente les fréquences des contraintes nOe $HN_i - HB_j$, (b) $HB_i - HB_j$, (c) $SC_i - SC_j$, (d) $HA_i - SC_j$, (e) $HB_i - SC_j$ et (f) $HN_i - SC_j$.

le cas des hélices α , mais l'absence des autres contraintes entourant ce réseau sur le graphe (a) est remarquable.

Le brin β est quant à lui caractérisé par un réseau de contraintes à plus courte portée. Il est représenté par un nombre importante de contraintes $HA_i - HN_{i+1}$ et par un réseau de contraintes $SC_i - SC_{i+2}$.

La présentation des résultats relatifs aux deux types de structures secondaires régulières que sont les hélices α et les brins β , a permis de mettre en évidence un réseau de contraintes nOe similaire à celui déjà connu. Les patterns de contraintes nOe que nous obtenons pour ces motifs ayant une structure secondaire régulière, nous ont permis de valider la méthode utilisée pour la classification des motifs structuraux. Une étude plus détaillée des motifs en tournant et de leurs caractéristiques expérimentales est donc envisageable. Elle est présentée dans la section suivante afin d'associer un réseau particulier de contraintes à un type de motif.

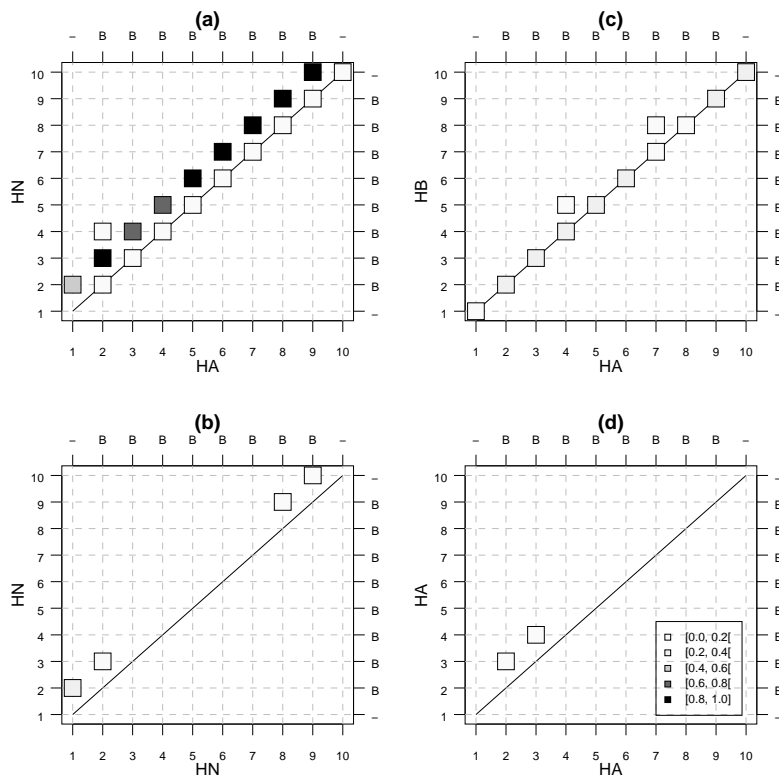


Fig. 4.5: Cartes des fréquences des contraintes nOe observées pour un brin β dans un cluster de 29 fragments ayant pour signature de Ramachandran -BBBBBBBB-.

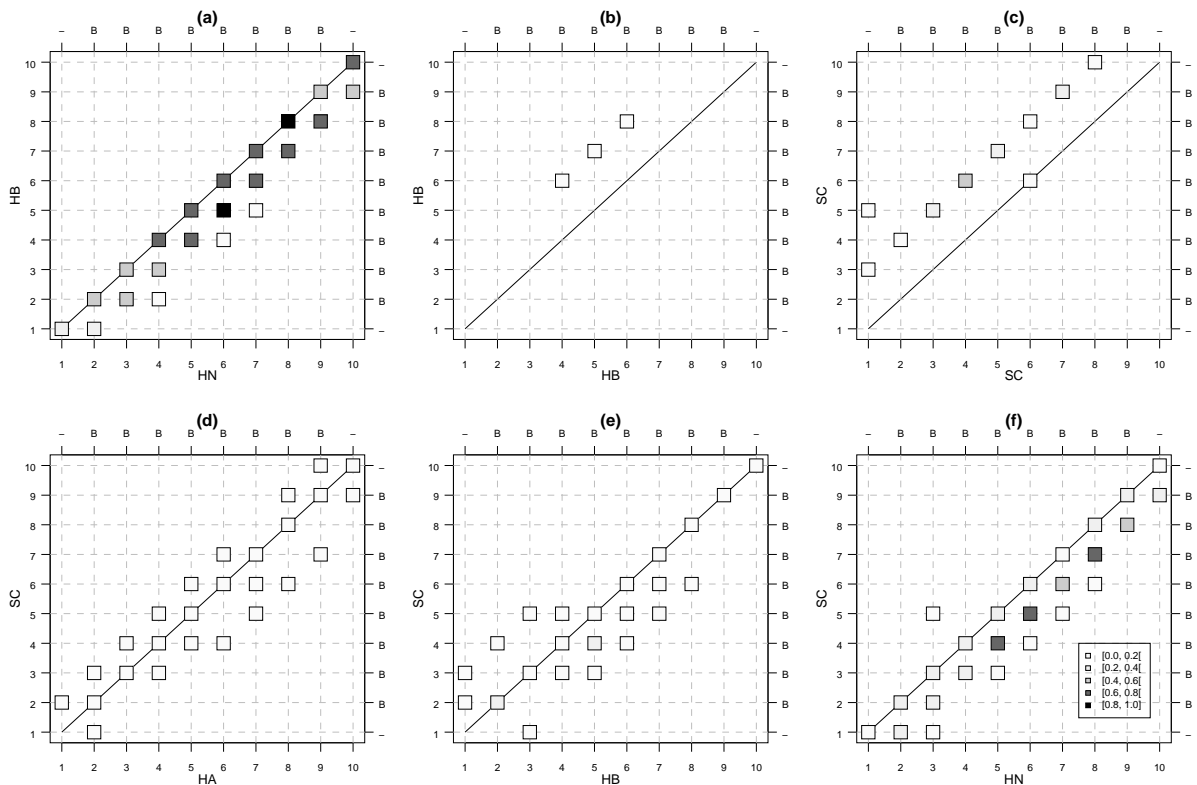


Fig. 4.6: Cartes des fréquences des contraintes nOe observées pour un brin β dans un cluster de 29 fragments ayant pour signature de Ramachandran -BBBBBBBB-. Cas des chaînes latérales.

4.5 Analyse détaillée des motifs en tournant

La sélection des motifs présentés est basée dans un premier temps sur la longueur des fragments puisqu'une même famille peut être identifiée avec des fragments de différentes longueurs. Les fragments les plus longs sont privilégiés car ils offrent l'opportunité d'une meilleure observation des contraintes. L'autre critère est celui de la diversité des fragments qui constituent un cluster. En effet, la méthode utilisée pour sélectionner les fragments issus des différents modèles représentatifs d'un segment de protéines (étape 1 de la classification des fragments, figure 2.1 page 55) permet de conserver plusieurs fragments représentatifs des différentes conformations possibles du segment. Ces fragments, ayant le même code PDB, peuvent se retrouver ensuite dans le même cluster à la fin de la classification. Si la redondance de ces fragments pour un même segment de protéine représentant différents modèles est trop importante, la famille n'est pas prise en compte car dans ce cas, la diversité des contraintes à analyser n'est pas importante.

4.5.1 motifs α - α

Parmi les quatre motifs α - α obtenus, un seul est présenté (figure 4.7 page 83 et figure 4.8 page 84) : le motif $\alpha B\alpha$ représenté par le cluster 7.128 composé de 14 fragments. Ce motif est assez fréquent dans les structures de protéines. La séquence consensus de ce motif met en évidence la présence d'un résidu polaire au niveau du résidu L1 impliqué dans le tournant. Le réseau de contraintes nOe représenté dans la figure 4.7 (page 83), montre la présence du résidu en conformation B parmi ceux en conformation d'hélice α . Le réseau caractéristique des hélices α n'est en effet observable qu'à partir du résidu L1 (résidu numéro 3). Le résidu -A1, du fait de son éloignement spatial dû à la présence du résidu L1, ne possède pas les contraintes $i, i + 3$ et $i, i + 4$. Le réseau des contraintes nOe impliquant les chaînes latérales, présenté dans la figure 4.8 (page 84) est assez disparate et la fréquence des observations n'est pas très élevée. Cependant, il est intéressant de noter la présence de huit contraintes particulières, impliquant le résidu L1, et se retrouvant aussi

sur les graphes d'autres motifs. Il s'agit des contraintes $\{HB_{L1}, SC_{L1}\} - \{HN_{L1,A1,A2,A3}\}$ impliquant les protons de la chaîne latérale avec les protons amides du résidu dans le tournant ainsi que les protons des trois résidus qui suivent.

4.5.2 motifs β - β

Les deux clusters sélectionnés représentant des motifs $\beta A \beta$ (figure 4.9 page 85) et $\beta A A \beta$ (figure 4.11 page 87) présentent chacun des caractéristiques particulières dues essentiellement au nombre différent de résidus impliqués dans le motif en tournant.

Les contraintes nOe observées pour le cluster 7.108 (figure 4.9 page 85) sont proches de celles représentant un brin β . La présence des contraintes $HA_{-B2} - HN_{-B1}$ et $HA_{B1} - HN_{B2}$ dans tous les fragments sélectionnés ainsi que la très faible fréquence de la contrainte $HA_{L1} - HN_{B1}$ permettent d'indiquer la localisation du tournant. La quasi absence des contraintes intra-résidus $HA - HN$ et $HA - HB$ est à noter, sachant qu'elles sont normalement présentes dans les structures en brin β (figure 4.5 page 76). Les contraintes $HN_i - HN_{i+1}$ (graphe (b)) n'apparaissent qu'à partir du résidu L1 (résidu 4) mais avec une faible fréquence, qui est sûrement davantage reliée à une particularité de cette famille de fragments. L'examen de la figure 4.10 (page 86) permet de remarquer l'absence de la contrainte $HN_{L1} - HB_{L1}$ (graphe (a)) et de mettre en évidence la présence des huit contraintes exposées lors du précédent motif. Dans ce cas ci, le résidu -B1 y est impliqué plutôt que le résidu L1 ($\{HB_{-B1}, SC_{-B1}\} - \{HN_{-B1,L1,B1,B2}\}$) et les contraintes observées sont peu fréquentes, il manque même la contrainte $HB_{-B1} - HN_{B2}$.

Le cluster 7.85 (figure 4.11 page 87) quant à lui est caractérisé par un motif en épingle à cheveux. La superposition des différentes structures n'est pas tellement satisfaisante puisque le r.m.s.d. moyen des structures deux à deux est de 1.70 Å. Une observation des contraintes $HA - HN$ (graphe (a)) met en évidence en plus du réseau caractéristique du brin β que sont les connexions $HA_i - HN_{i+1}$, cinq contraintes $HA_{L1} - HN_{L2,B1,B2}$ et $HA_{L2} - HN_{B1,B2}$. Deux de ces contraintes $HA_{L1} - HN_{L2}$ et $HA_{L1} - HN_{B1}$ ont été remarquées dans l'étude faite par Wagner et al. [58] comme étant représentatives d'un

motif en tournant. À celles-ci s'ajoutent les contraintes observées entre les protons amides (graphe (b)), la fréquence des contraintes $HN_i - HN_{i+1}$ est assez élevée et le réseau constitué par celles-ci est régulier. Deux contraintes particulières sont observées mais avec une fréquence moins importante, il s'agit de $HN_{-B1} - HN_{L2}$ et $HN_{L1} - HN_{B2}$, permettant probablement de contraindre le tournant et pouvant être une propriété de ce type de motif. Des contraintes entre résidus séparés dans la séquence d'au moins quatre résidus sont présentes sur les graphes (c) et (d), elles sont aussi présentes sur les graphes impliquant les chaînes latérales (figure 4.12 page 88). Cependant, dans cette figure peu de contraintes sont observées. Mais ici encore, le groupe des huit contraintes mettant en jeu la chaîne latérale du résidu L1 est présent ($\{HB_{L1}, SC_{L1}\} - \{HN_{L1,L2,B1,B2}\}$).

4.5.3 motifs α - β

Les motifs α - β sont représentés par deux clusters, un ayant comme motif $\alpha\beta$ et l'autre $\alpha G\beta$. Ce dernier type de connexion représente une part importante de l'ensemble des fragments répertoriés dans les familles α - β .

Le réseau des contraintes nOe observées pour les 8 fragments que composent le cluster 7.31 (figure 4.13 page 89) ayant pour motif $\alpha\beta$, se rapproche davantage de celui du brin β surtout en ce qui concerne les contraintes $HA - HN$. Les contraintes $HN - HN$ sont assez disparates mais une caractéristique particulière peut être mise en évidence : il s'agit de la contrainte $HN_{-X3} - HN_{B1}$. Pour l'observer par RMN sur pratiquement tous les fragments, le proton amide du résidu B1 doit être dirigé vers le tournant. L'observation de la figure 4.14 (page 90) des chaînes latérales montre que peu de contraintes sont observées. La présence du tournant en début de motif ne permet pas d'observer un réseau de contraintes important impliquant les chaînes latérales puisqu'une partie du brin β est éloignée du tournant. Des contraintes entre résidus proches le long de la séquence sont présentes, caractéristique du brin β .

La principale caractéristique du motif $\alpha G\beta$, représentée par le cluster 10.66 (figure 4.15 page 91), est que l'hélice α se termine par un résidu (position L1) en conformation hélice gauche (domaine G de la carte de Ramachandran). Des séquences consensus ont été construites pour identifier ce motif dit de 'Schellman' dont celle rapportée par Aurora et al. [2]. Selon la nomenclature utilisée, cette séquence consensus comporte une glycine en position L1, un acide aminé apolaire ou une arginine en B1, un acide aminé polaire ou une alanine en -A2, et au moins une des positions -A3, -A4 ou -A5 est apolaire ou une arginine. Le motif présenté dans la figure 4.15 (page 91) possède une séquence consensus qui s'accorde plutôt bien avec celle proposée par Aurora et al.. Associé à cette séquence consensus, ce motif est caractérisé aussi par trois ponts hydrogène entre les résidus -A4/B1 nommé 5-turn, -A3/L1 et -A4/L1 formant respectivement les ponts 3-turn et 4-turn [60, 14]. La présence en terme de fréquence observée des contraintes nOe pour ces types particuliers de résidus est remarquable sur les graphes (a) et (c) de la figure 4.15 (page 91) et sur les graphes de la figure 4.16 (page 92). Le contact le plus caractéristique de ce type de motif implique les résidus en position -A4 et B1. En effet, des contraintes $\{HA_{-A4,-A3}\} - HN_{B1}$ et $HB_{-A4} - \{HA_{B1,B2,X3}\}$ sont observées pour certains des fragments, mais aussi les contraintes $\{HA_{-A4,-A3}\} - HN_{L1}$. L'observation des contraintes $HA - HN$ nous renseigne aussi sur la présence du résidu L1 puisqu'aucune contrainte n'est présente entre ce résidu et les suivants. Les caractéristiques d'une hélice α en terme de contraintes nOe se retrouvent sur le graphe (b) et le passage au réseau de contraintes du brin β est remarquable. Cette observation est aussi applicable au graphe (a) des contraintes $HN - HB$ de la figure 4.16 (page 92). L'occurrence élevée des contraintes $HA_{-A4} - SC_{B1}$ et $HB_{-A4} - HN_{B1}$, ainsi que la présence moins fréquente des contraintes $\{HB_{-A4}, HN_{-A4}\} - SC_{B1}$, sont les caractéristiques d'observations expérimentales traduisant le fait que les résidus -A4 et B1 sont proches dans l'espace.

4.5.4 motifs β - α

La dernière famille de motifs présentée est celle des motifs β - α .

Les particularités du motif $\beta\alpha$ représentées dans la figure 4.17 (page 93) sont un réseau dense de contraintes nOe $HA - HN$ et $HN - HN$ mais peu fréquent, plutôt représentatif des hélices α . A partir du résidu A1, la présence des contraintes $HA_i - HN_{i+3}$ et $HA_i - HB_{i+3}$ est remarquable. La fréquence d'apparition des contraintes $HN_i - HN_{i+1}$ sur l'ensemble des 16 fragments peut aussi permettre de caractériser l'emplacement du tournant puisque à partir du résidu A1, la fréquence est plus élevée. En effet, le graphe (b) montre que la fréquence d'observation des contraintes est plus faible pour les deux résidus en conformation B, -B2 et -B1. Il met aussi en évidence la présence de nombreuses contraintes entre le résidu A4 et les cinq résidus le précédant. Les graphes des contraintes $HN - HB$ et $HN - SC$ (figure 4.18 page 94) sont assez similaires comme pour les autres motifs présentés. Ils se caractérisent ici par une plus grande ressemblance à ceux de l'hélice α qu'à ceux du brin β . Ils mettent en évidence de nombreuses contraintes entre le résidu -B2 et les cinq résidus le suivant. Ces contraintes impliquant des résidus éloignés dans la séquence traduisent la forme de la structure et le fait que le brin β et l'hélice α sont proches dans l'espace.

Le cas du cluster 10.68 ayant pour motif $\beta AB\alpha$ (figure 4.19 page 95) illustre davantage la présence de deux sortes de structures secondaires séparées par un motif en tournant constitué de deux résidus. Le réseau caractéristique de l'hélice α est observé à partir du résidu A1 avec notamment la présence des contraintes $HA_i - HN_{i+3}$ et $HA_i - HB_{i+3}$. Les contraintes $HA_{L1} - HN_{L2}$ et $HA_{L2} - HN_{A1,A2,A3}$ observées sur le graphe (a), ont aussi été remarquées dans l'étude du motif $\beta AA\beta$ précédent, où dans ce cas elles impliquaient les protons alpha des résidus L1 et L2 avec les protons amides des résidus L2, B1 et B2. Le graphe (b) montre des contraintes nOe entre les protons amides impliquant le résidu L2 et les quatre résidus qui suivent A1, A2, A3, A4. De plus, la figure 4.20 (page 96) indique la présence des huit contraintes déjà observées précédemment mais impliquant

ici le résidu L2 plutôt que le résidu L1. Elle met aussi en évidence quelques contraintes particulières plus fréquentes que d'autres impliquant les chaînes latérales. Les contraintes observées sur le graphe (a) entre HB_{A3} et $\{HN_{L1,L2}\}$ sont caractéristiques de ce motif. Le repliement de la structure tridimensionnelle montre que le brin β et l'hélice α sont proches dans l'espace, se traduisant par la présence ici de contraintes entre des résidus éloignés dans la séquence par exemple $HB_{-B1} - HB_{A3,A4}$, ou encore $SC_{-B1} - SC_{A4}$ et aussi $SC_{-B1} - HB_{A3}$.

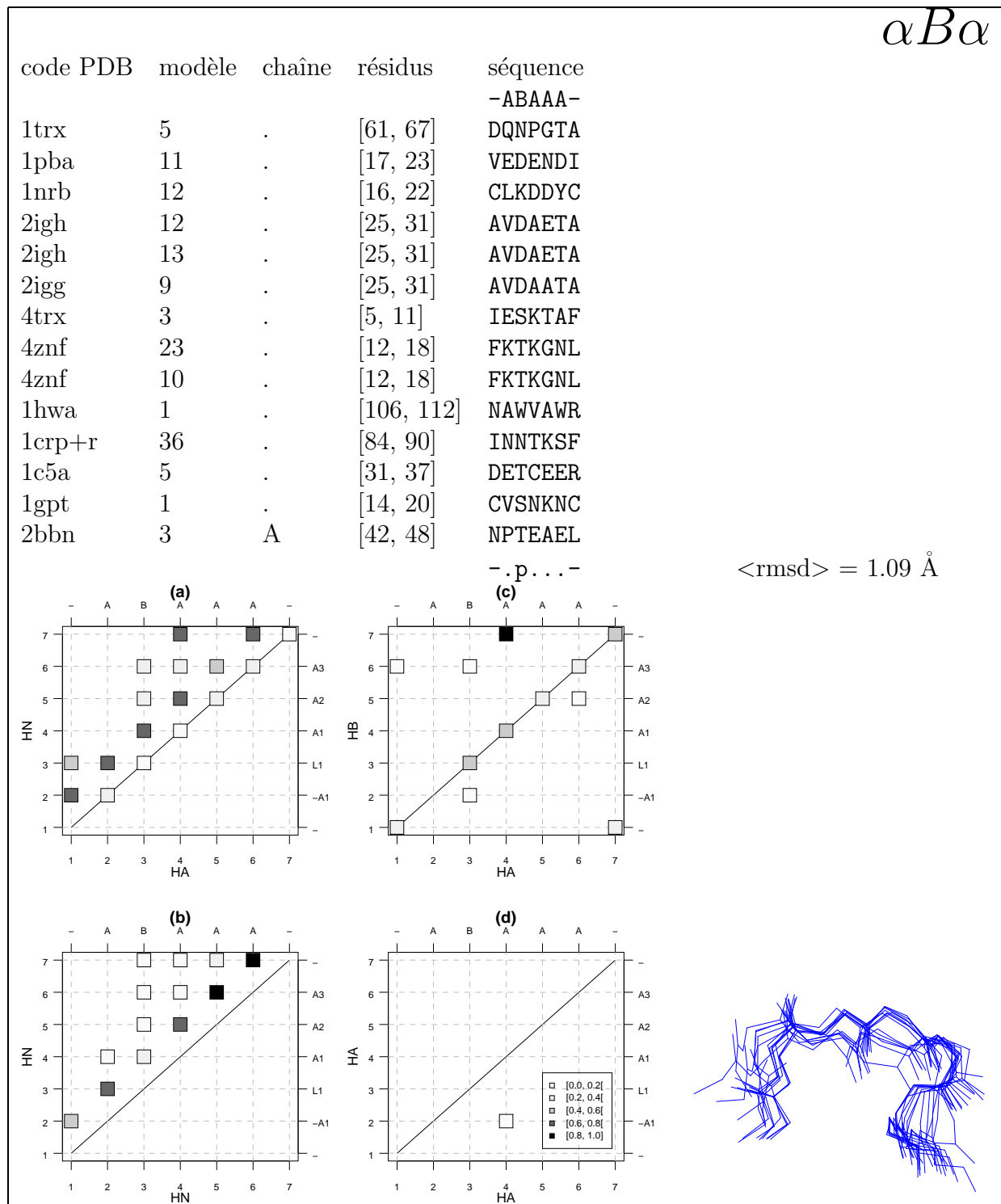


Fig. 4.7: Caractéristiques du motif $\alpha B\alpha$ (cluster 7.128) représentées par la liste des fragments, les réseaux des contraintes nOe et la superposition des structures.

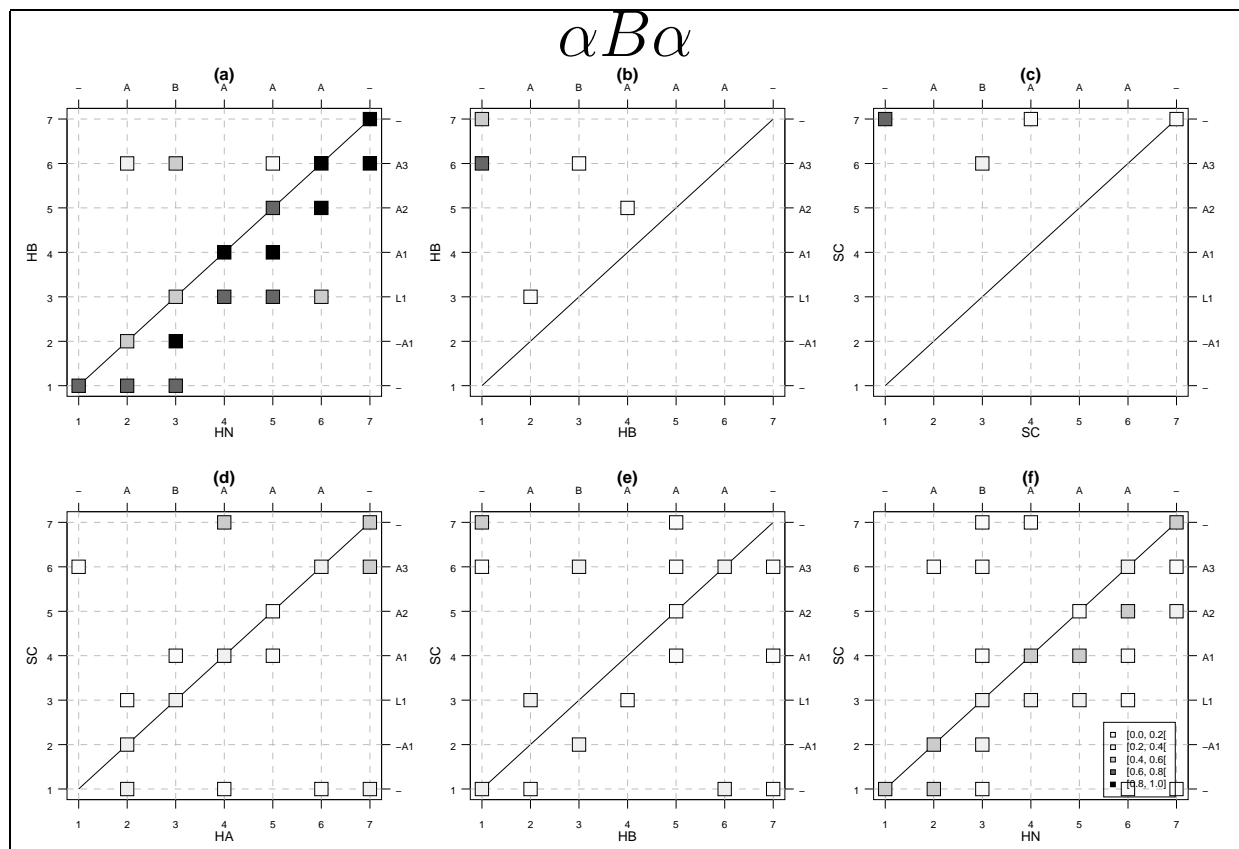


Fig. 4.8: Réseaux des contraintes nOe pour le motif $\alpha B\alpha$ (cluster 7.128) impliquant les chaînes latérales.

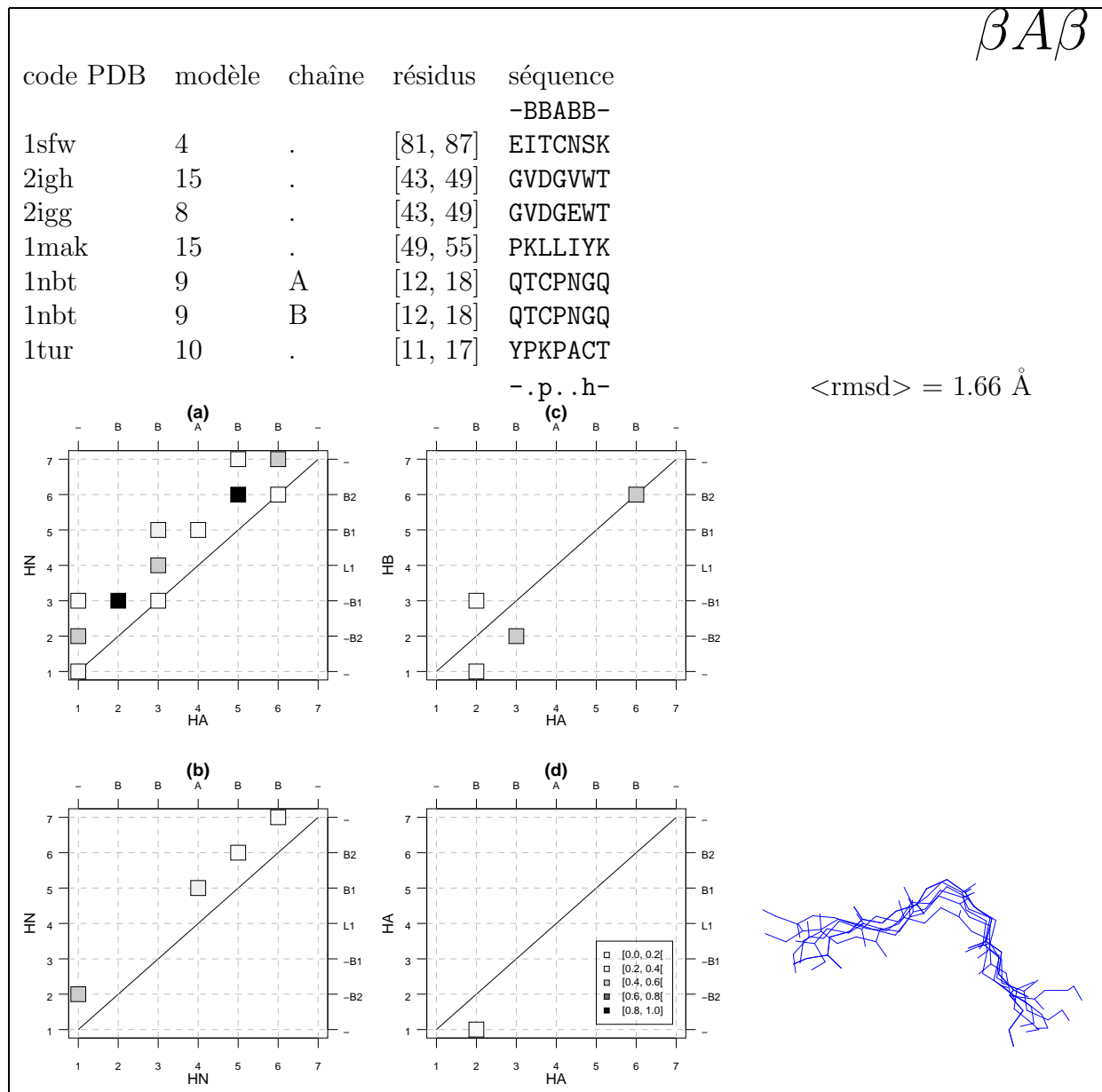


Fig. 4.9: Caractéristiques du motif $\beta A \beta$ (cluster 7.108).

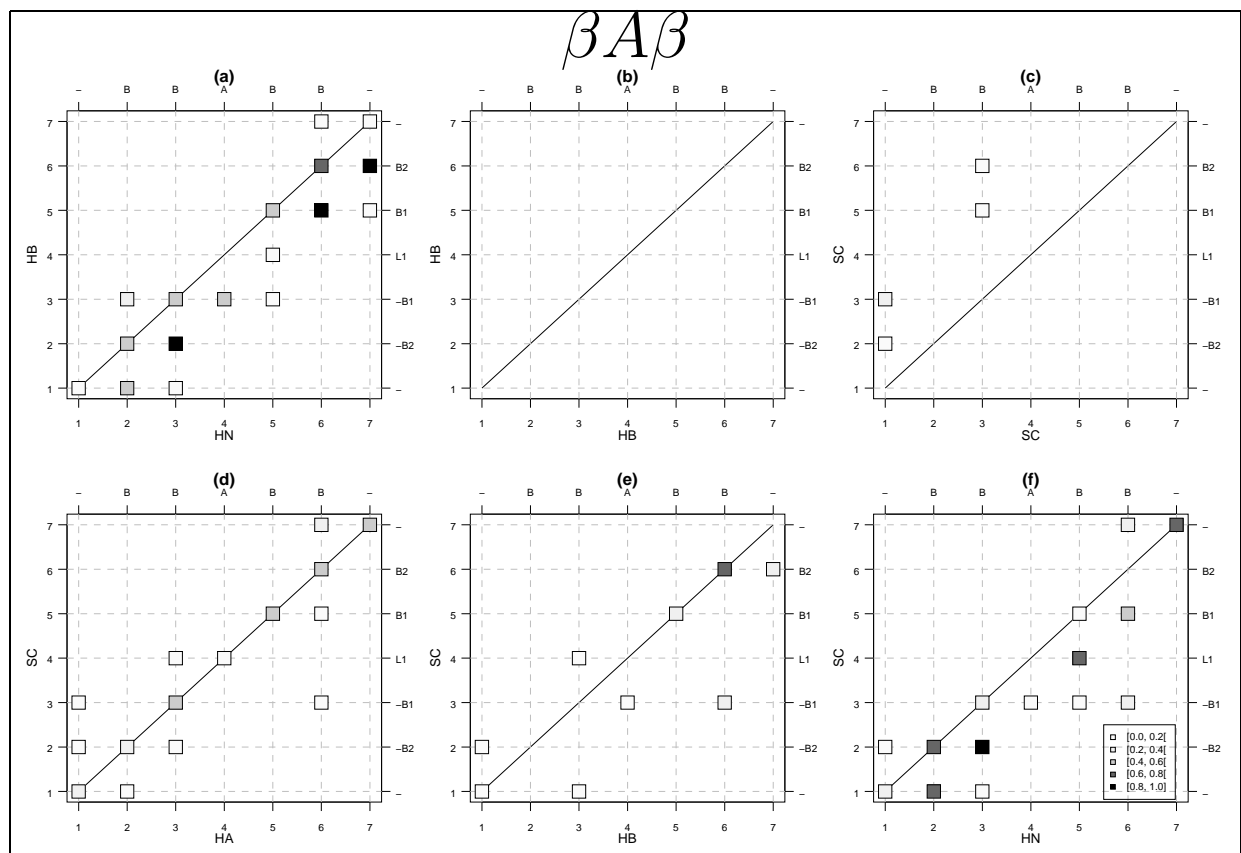


Fig. 4.10: Caractéristiques du motif $\beta A \beta$ (cluster 7.108). Cas des chaînes latérales.

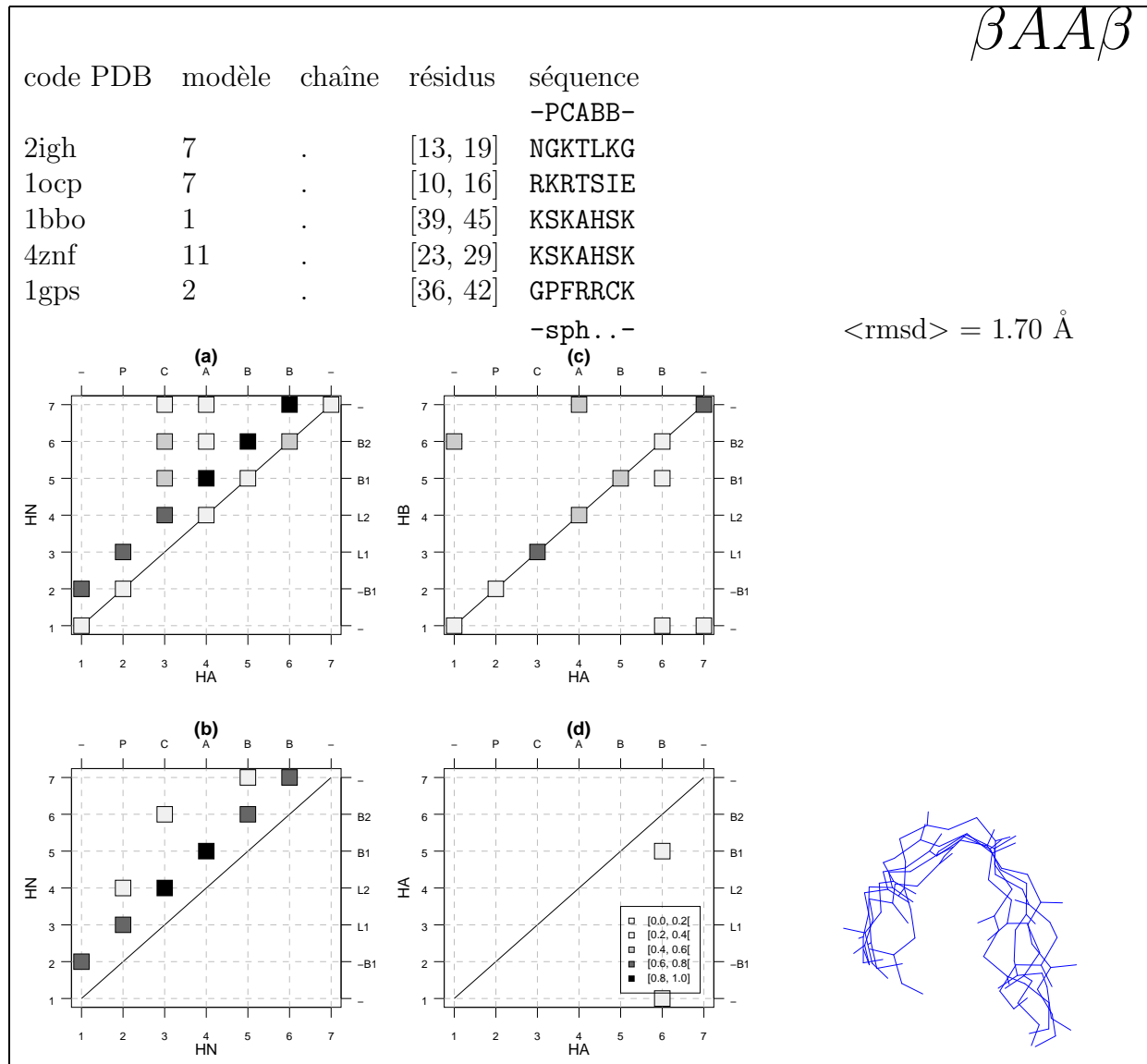


Fig. 4.11: Caractéristiques du motif $\beta A A \beta$ (cluster 7.85).

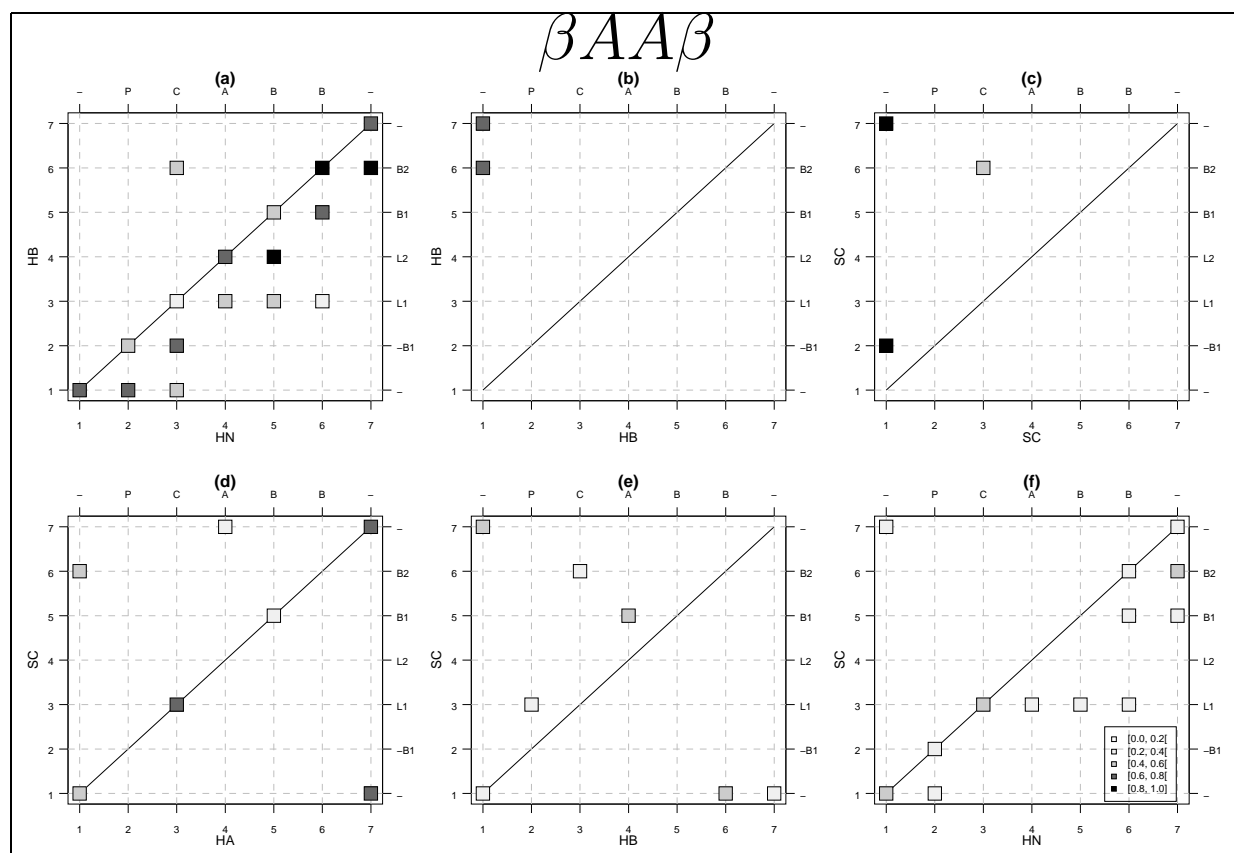


Fig. 4.12: Caractéristiques du motif $\beta A A \beta$ (cluster 7.85). Cas des chaînes latérales.

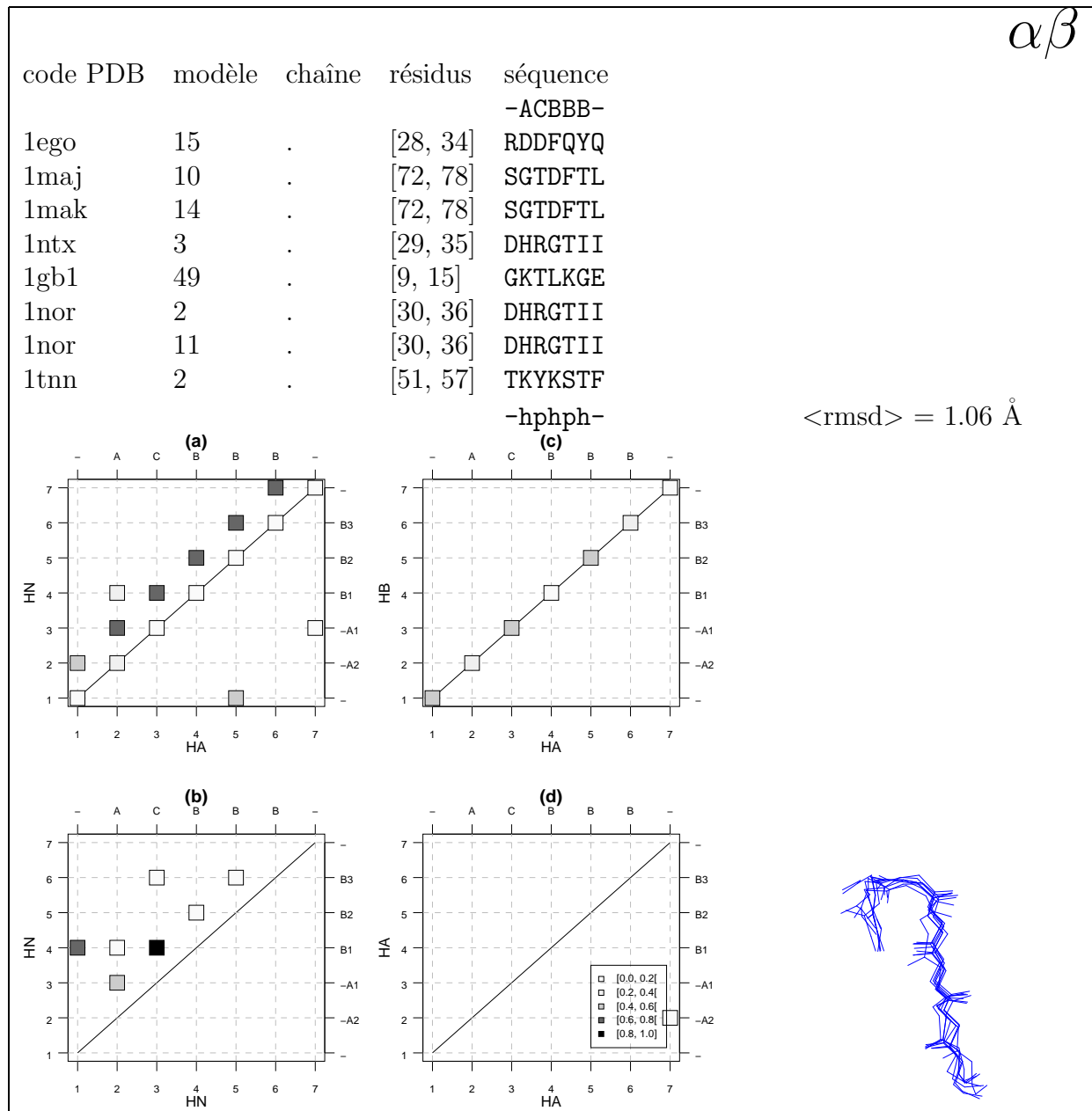


Fig. 4.13: Caractéristiques du motif $\alpha\beta$ (cluster 7.31).

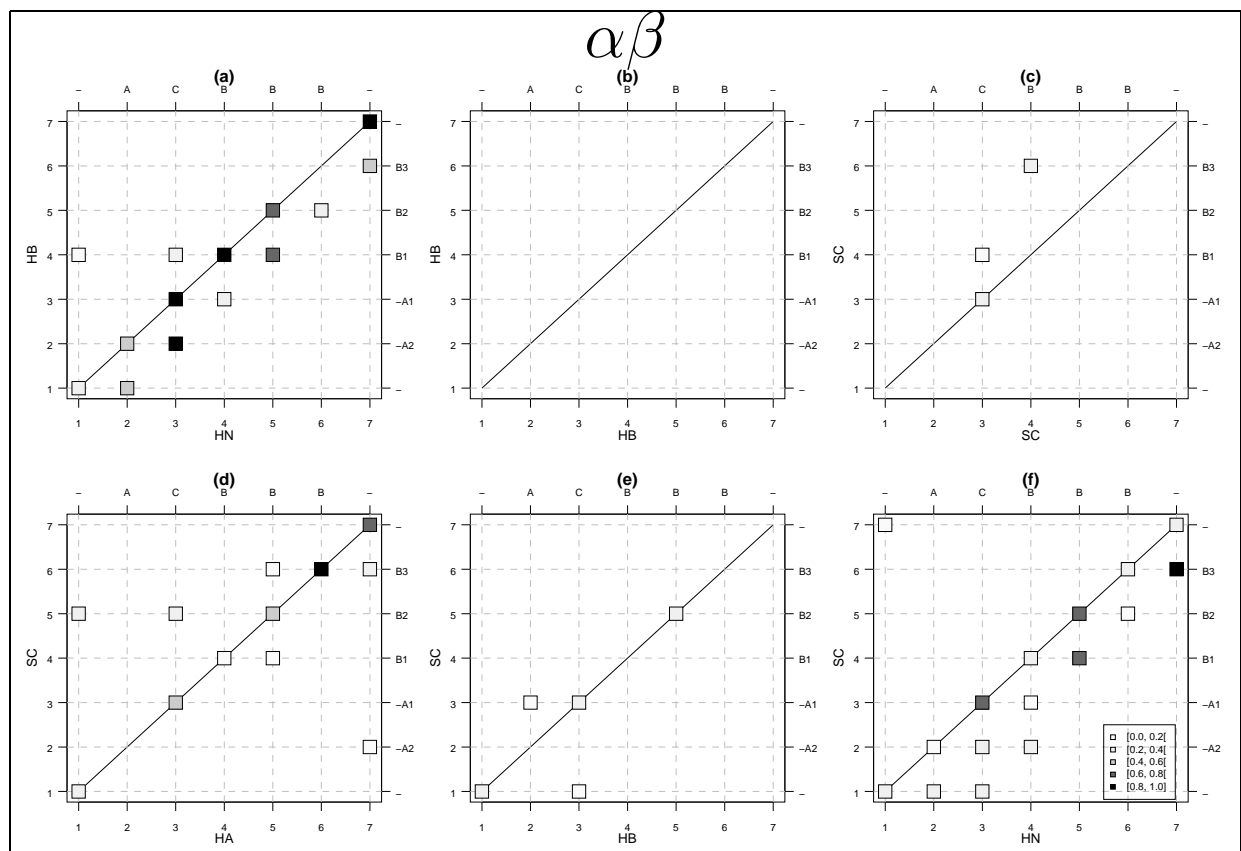


Fig. 4.14: Caractéristiques du motif $\alpha\beta$ (cluster 7.31). Cas des chaînes latérales.

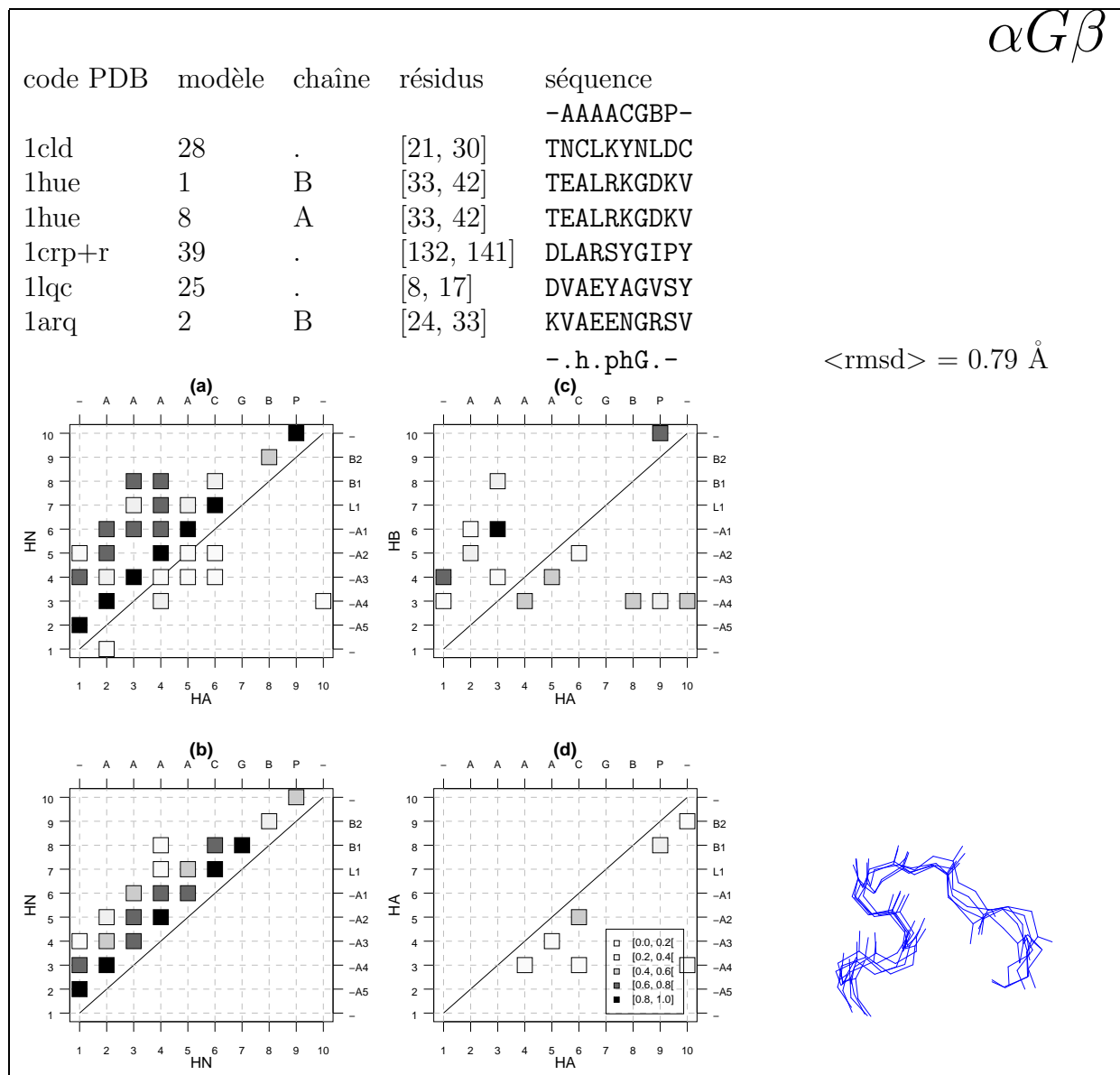


Fig. 4.15: Caractéristiques du motif $\alpha G\beta$ (cluster 10.66).

tel-00275947, version 1 - 25 Apr 2008

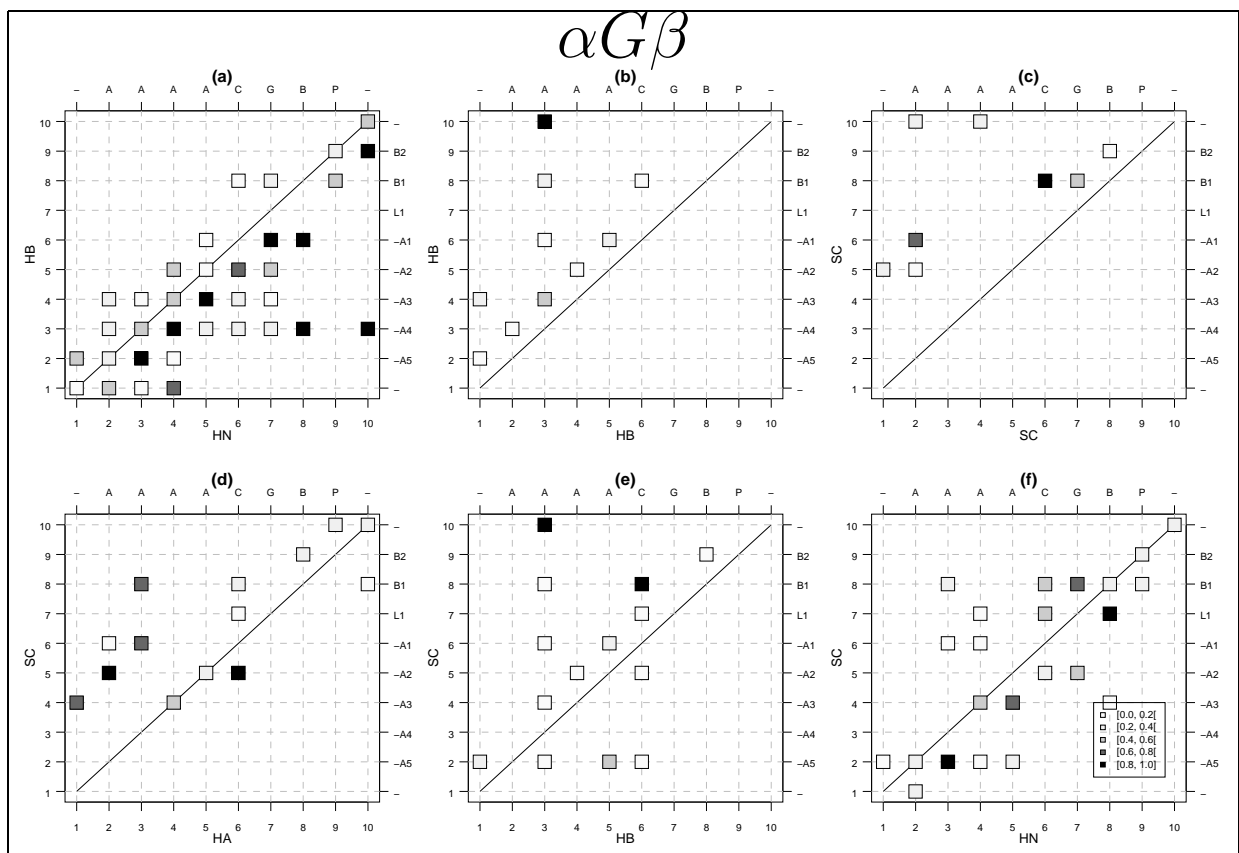
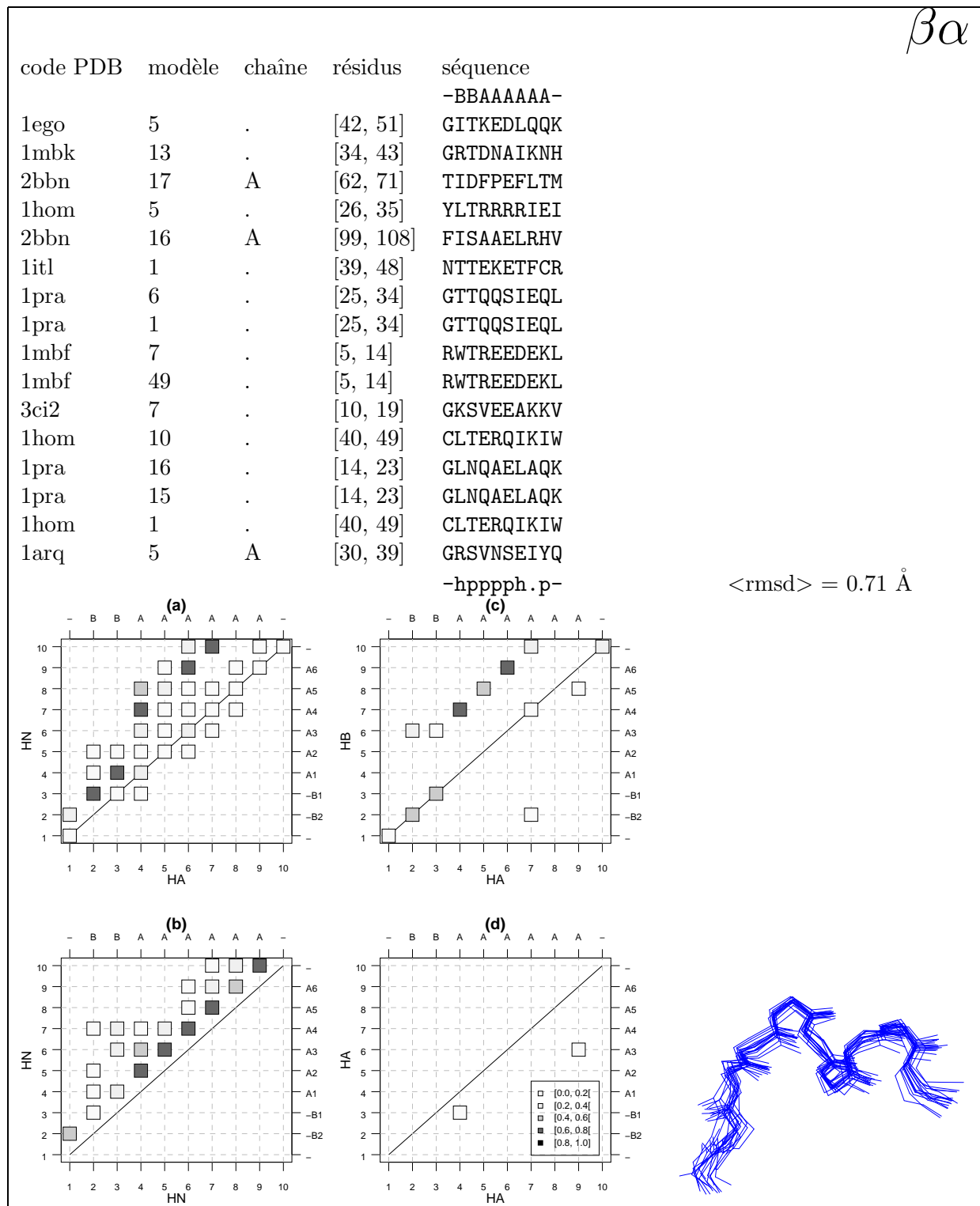


Fig. 4.16: Caractéristiques du motif $\alpha G\beta$ (cluster 10.66). Cas des chaînes latérales.

Fig. 4.17: Caractéristiques du motif $\beta\alpha$ (cluster 10.10).

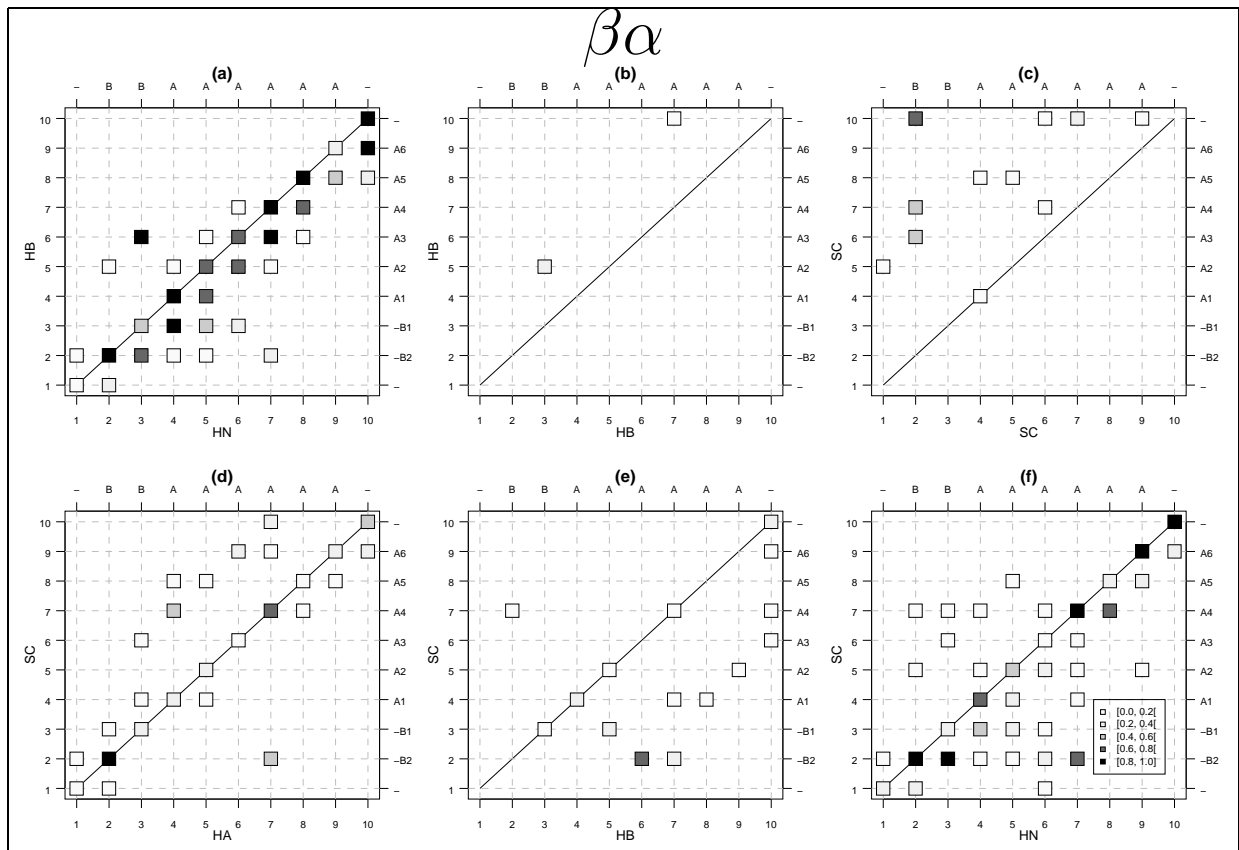
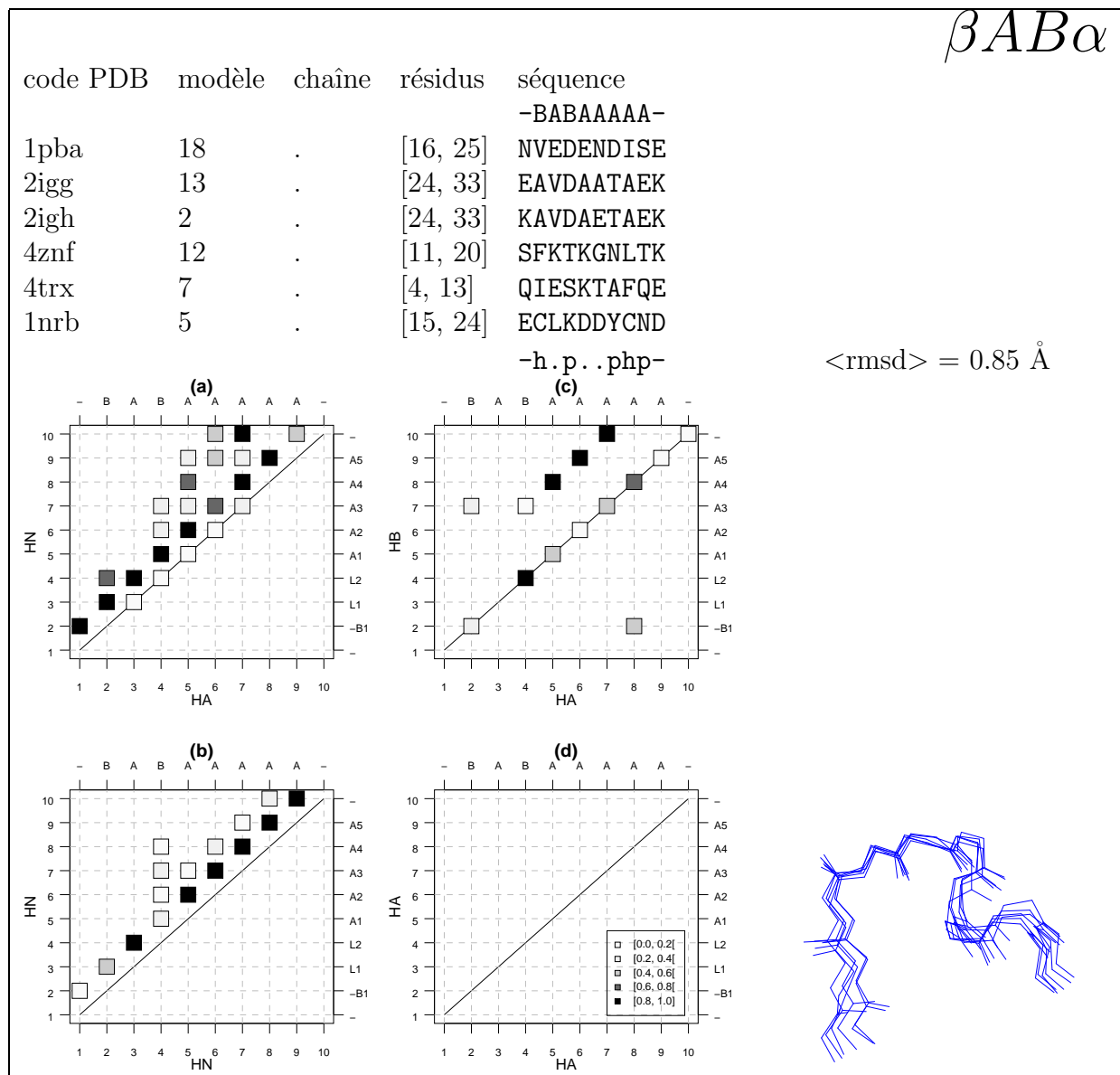


Fig. 4.18: Caractéristiques du motif $\beta\alpha$ (cluster 10.10). Cas des chaînes latérales.

Fig. 4.19: Caractéristiques du motif $\beta AB\alpha$ (cluster 10.68).

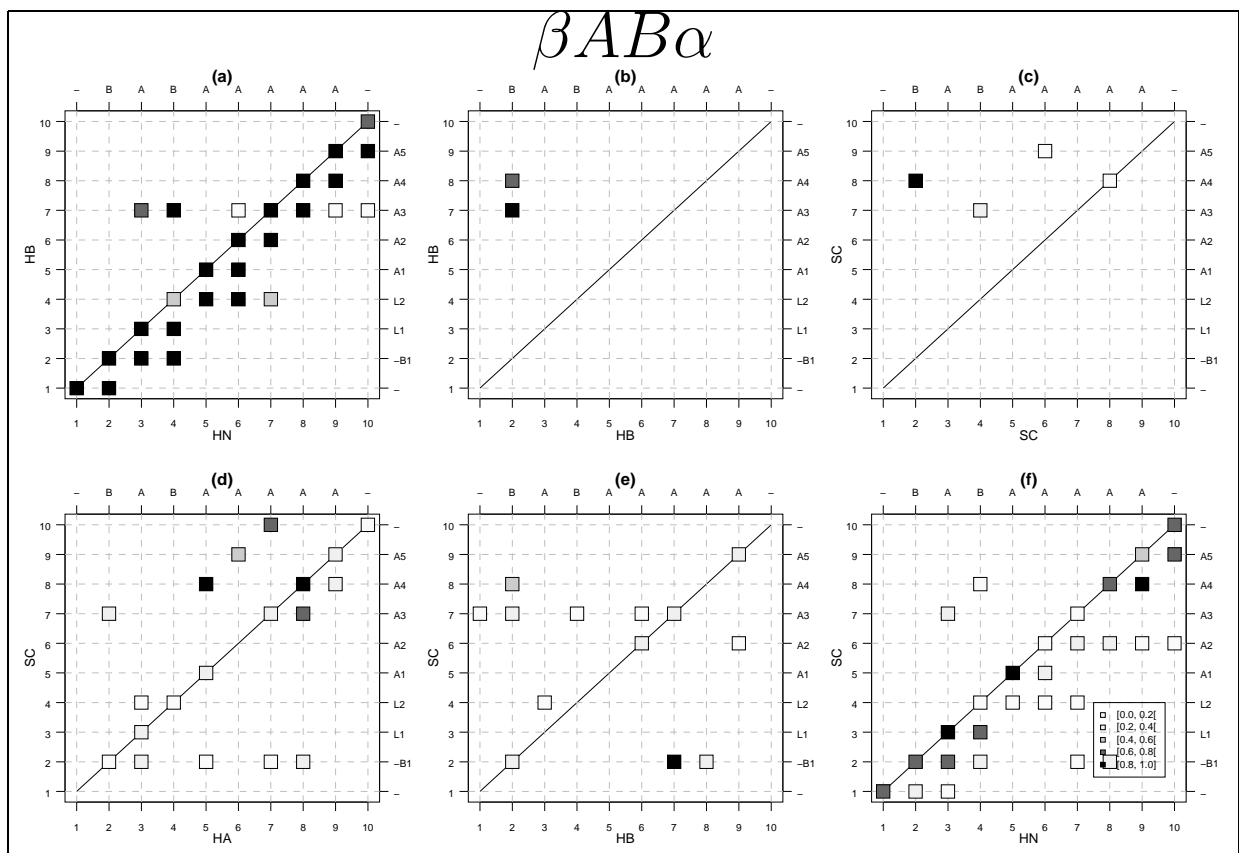


Fig. 4.20: Caractéristiques du motif $\beta AB\alpha$ (cluster 10.68). Cas des chaînes latérales.

4.6 Qualité locale des structures

Une application possible de cette méthode est l'analyse de la qualité locale des structures. En effet, la comparaison entre les distances calculées et les contraintes nOe observées met en évidence une forte corrélation. Les graphes (a) et (b) (figure 4.21) des distances calculées représentent les fréquences d'observation d'une distance inférieure à 5 Å entre les protons alpha et les protons amides. Il s'agit d'une observation faite sur les 40 fragments constituant le cluster de l'hélice α déjà présenté. Le réseau des distances est plus dense et plus régulier que celui des contraintes décrit précédemment.

L'observation des graphiques (figure 4.21) a permis de déterminer une nouvelle mesure de la qualité des structures. Celles que nous connaissons jusqu'à présent sont essentiellement basées sur des qualités géométriques [31]. La nouvelle mesure de qualité proposée est basée sur le nombre de contraintes nOe et est calculée comme suit :

$$\frac{\text{Nb de contraintes nOe observées}}{\text{Nb de distances} < 5\text{Å}} \quad (4.1)$$

Dans le cas du cluster de l'hélice α , une classification peut être faite sur base de cet indice de qualité (figure 4.22). Il varie de 10 à 90%. Les hélices ne sont donc pas toutes déterminées avec le même niveau de contraintes, mais les contraintes nOe ne sont pas toujours observables expérimentalement pour différentes raisons. Il pourrait être envisagé une étude des fragments ayant un faible indice (inférieur à 30%) afin de mettre en évidence des particularités au niveau de la séquence.

4.7 Améliorations de la méthode et perspectives d'analyses

Ce travail étant en cours, la présentation de l'état actuel des analyses permet d'envisager différents axes d'amélioration de cette méthode de classification de fragments de structures de protéines issues d'expériences RMN. La limitation du nombre des régions

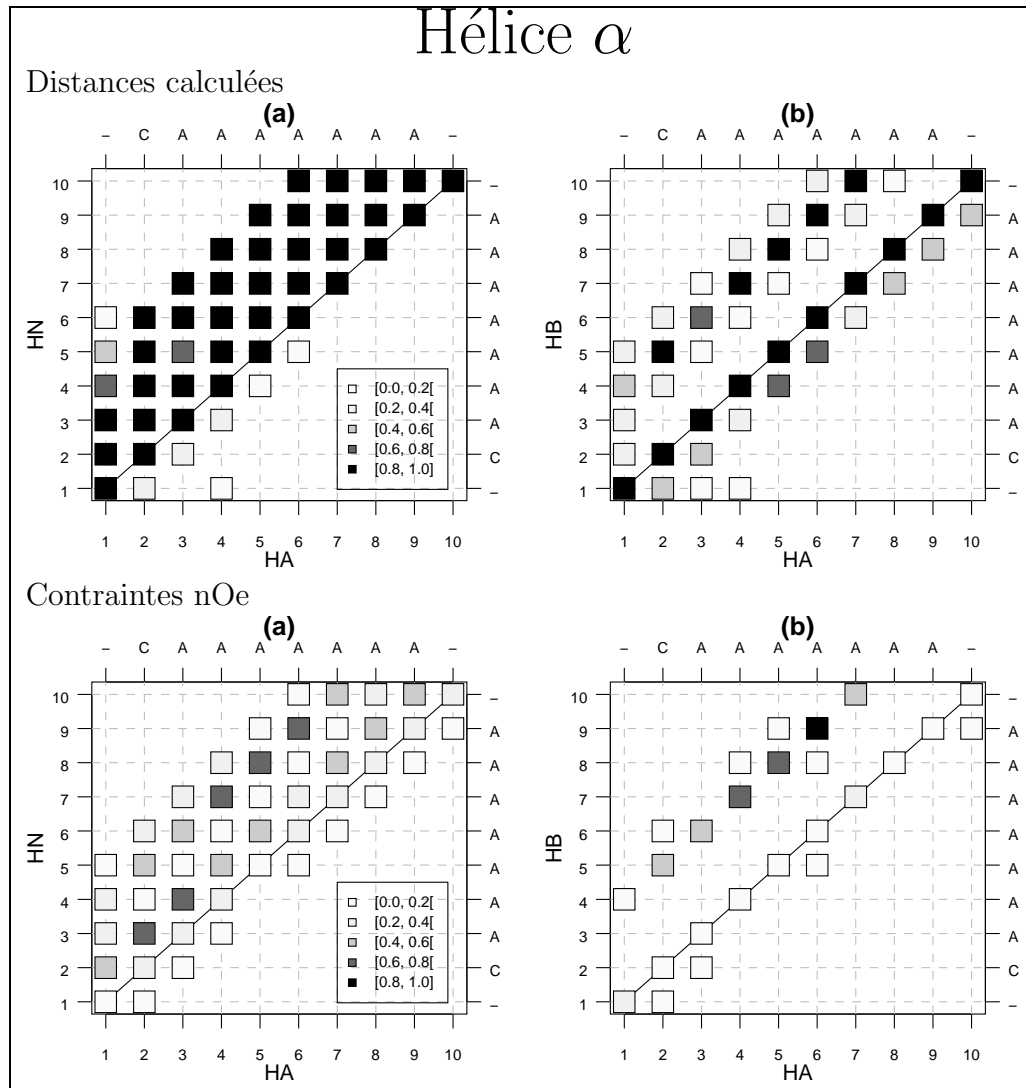


Fig. 4.21: Fréquences des distances calculées inférieures à 5 Å et des contraintes nOe observées dans le cluster de 40 fragments de l'hélice α . Le graphe (a) représente les fréquences des distances ou des contraintes entre les protons alpha et les protons amides, le (b) entre les protons alpha et les protons beta.

possibles de la carte de Ramachandran pourrait permettre de réduire le nombre de familles possédant la même signature de Ramachandran et ainsi limiter la dispersion des clusters. Par exemple, les résidus en conformation A et C pourraient être classés dans une même et unique région. Le choix de la longueur des fragments et des clusters représentatifs reste à améliorer. En effet, les fragments du motif $\beta AB\alpha$ (figure 4.19 page 95) de 10 résidus de longueur se retrouvent parmi les fragments de 7 résidus du motif $\alpha B\alpha$ (figure 4.7 page 83), mais ils ne sont cependant pas issus du même modèle de structure. L'élimination des fragments constituant les clusters représentatifs de plus grande longueur parmi

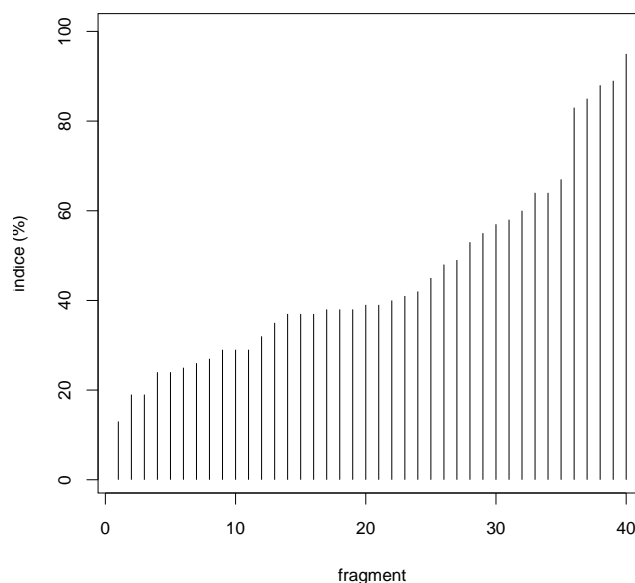


Fig. 4.22: Répartition du pourcentage de l'indice de qualité dans le cluster de l'hélice α constitué de 40 fragments.

le groupe de fragments de départ lors de la classification des fragments de plus petite longueur, permettrait de résoudre le problème du chevauchement des clusters.

Cette méthode offre malgré tout des perspectives d'analyses intéressantes. Notamment, l'analyse de l'indice de qualité pour d'autres fragments et l'analyse plus détaillée d'une liste de contraintes en fonction de cet indice sont envisageables. Une analyse de cet indice en fonction de la valeur de la distance observée ou de la valeur de la contrainte pourrait être proposée en classant les distances par catégorie courte, moyenne et longue distance. Le calcul du nombre moyen de violation des contraintes sur la longueur du fragment, comparé à cet indice, pourrait aussi être étudié.

L'augmentation du nombre de protéines étudiées permettrait d'apporter une plus grande diversité des motifs, d'analyser davantage de valeurs expérimentales, telles que les contraintes nOe, et d'envisager d'étendre cette analyse à celles des valeurs des déplacements chimiques. L'observation du nombre de protéines possédant des contraintes nOe dans la base de données BMRB est malheureusement faible (14 protéines pour 2 208 contraintes nOe) mais une liste d'environ 1 280 protéines ayant des contraintes nOe est disponible. Malheureusement, pour l'instant ces données sur les contraintes se trouvent stockées dans des fichiers ayant des formats différents (Amber, Discover, Dyana, Xplor,...).

La conversion des formats de fichiers des contraintes est un réel problème pour la prise en compte de toutes ces structures. Cependant, le nombre de protéines possédant des déplacements chimiques dans BMRB est important (2 141 protéines) et ces données sont bien structurées puisque stockées sous forme de fichier NMR-STAR et dans les tables de la base de données. La comparaison plus complète entre une liste de contraintes nOe et des valeurs de déplacements chimiques issus de protéines venant de BMRB et ayant des structures tridimensionnelle pourrait donc être envisagée.

Les applications d'une base de données de fragments ayant des contraintes nOe sont multiples. Elle peut servir notamment à la détermination des structures tridimensionnelles de protéines. La méthode décrite par Kraulis and Jones [28] permet en effet de générer une structure à partir d'un ensemble de contraintes de distance nOe, en recherchant dans une base de données de structures de protéines obtenues par radiocristallographie, des fragments de protéines satisfaisant aux contraintes observées. Cette méthode peut être appliquée à notre base de données de fragments issus de protéines obtenues par RMN et possédant déjà un réseau de contraintes nOe. Elle offrirait une comparaison plus réaliste et plus fiable des contraintes avec celles observées. Le logiciel TALOS [11] quant à lui fourni les valeurs des angles ϕ, ψ à partir des structures cristallographiques de 20 protéines ayant des déplacements chimiques, pouvant être utilisées comme contraintes lors de la modélisation de la structure recherchée. L'association de contraintes de distances à des contraintes angulaires pourrait permettre d'augmenter encore la qualité des structures modélisées.

Chapitre 5

Discussion

L'étude présentée sur les empreintes expérimentales d'un groupe de fragments représentant des motifs de différents types a permis de mettre en évidence des patterns de contraintes nOe caractéristiques. Dans cette première étude, les patterns nOe des structures secondaires régulières sont présents et observés. Nous avons pu remarquer un réseau de contraintes plus large dans le cas de l'hélice α que dans le cas du brin β étudié. L'hélice α est donc caractérisée par une présence prédominante de six types de contraintes $HA_i - HN_{i+1}$, $HA_i - HN_{i+3}$, $HN_i - HN_{i+1}$, $HA_i - HB_{i+3}$, $HB_i - HN_i$ et $HB_i - HN_{i+1}$. Le brin β est quant à lui représenté par un réseau de contraintes à plus courte portée. Il est caractérisé par une présence importante de contraintes $HA_i - HN_{i+1}$ rarement discutées dans la littérature, de contraintes $HA_i - HB_i$, $HB_i - HN_i$ et $HB_i - HN_{i+1}$, et d'un réseau remarquable de contraintes $SC_i - SC_{i+2}$. Les résultats obtenus dans le cas de l'hélice α et du brin β sont en accord avec les réseaux de contraintes nOe connus pour les structures secondaires régulières. Ces résultats nous ont donc permis de valider notre méthode de classification.

Des patterns communs ont été observés pour les différents types de motifs en tournant, mais des caractéristiques particulières, notamment dans le cas du motif $\alpha G \beta$, ont aussi été dégagées de cette analyse même s'il semble plus difficile de caractériser ces motifs uniquement sur base des contraintes nOe. Une particularité souvent observée dans le cas des motifs en tournant entre deux structures secondaires différentes, est la possibilité de

déterminer la présence du tournant par la distinction de la limite entre les deux types de réseaux de contraintes associés à ces deux structures secondaires régulières (motifs $\alpha G\beta$, $\beta\alpha$ et $\beta AB\alpha$). Les contraintes entre les protons alpha des résidus impliqués dans le tournant et les protons amides des trois résidus qui suivent, sont observées pour les motifs possédant deux résidus dans le tournant, comme $\beta AA\beta$ (figure 4.11 page 87) et $\beta AB\alpha$ (figure 4.19 page 95), mais aussi pour le motif $\alpha B\alpha$ (figure 4.7 page 83). Les autres motifs présentés ne possèdent pas cette caractéristique. Une autre propriété est associée à ces trois motifs, elle est caractérisée par un groupe de huit contraintes entre les protons de la chaîne latérale du résidu L1 et les protons amides des quatre résidus, dont L1, qui suivent (figures 4.12 page 88, 4.20 page 96 et 4.8 page 84). Le motif $\beta A\beta$ est quant à lui remarquable par une absence de contrainte impliquant le résidu L1. De même, le motif $\alpha\beta$ ne possède pas beaucoup de contraintes.

L'étude du motif $\alpha G\beta$ (figure 4.15 page 91) a permis de mettre en évidence la présence de contraintes nOe entre les résidus -A4/B1 impliqués dans le contact de Schellmann et dans le pont hydrogène 5-turn, mais aussi entre les résidus -A4/L1 et -A3/L1 impliqués dans les ponts hydrogènes 4-turn et 3-turn, caractéristiques de ce type de motif en tournant.

La confirmation des observations faites dans le cadre de cette analyse avec des informations issues de la littérature permet d'offrir des perspectives intéressantes à ce type d'étude. Malgré les difficultés à caractériser ces motifs seulement sur base des contraintes, une perspective d'étude des corrélations entre motifs structuraux, déplacements chimiques et contraintes nOe pourrait être envisagée dans le cadre de Pescador. Il est nécessaire pour cela d'obtenir un plus grand nombre de données. L'augmentation du nombre de données étudiées permettrait aussi d'obtenir une plus grande diversité des clusters et offrirait la possibilité d'une étude plus diversifiée et plus exhaustive des particularités de certaines contraintes nOe relatives aux motifs.

La mise en place de cette méthode et son développement ont pris une part importante dans ce travail. Cependant des perspectives intéressantes sont envisageables. La

comparaison des patterns de contraintes nOe observées avec celles pouvant être dérivées des données structurales actuelles, est une autre voie possible pour la validation de cette approche. Cette méthode peut en effet être utilisée pour l'analyse de fragments de protéines dans le but de déterminer la qualité des structures dans ces régions. Le nombre de contraintes nOe ayant une forte occurrence (les carrés noirs sur les graphes présentés précédemment) et permettant de distinguer telle ou telle conformation n'est souvent pas très important. Comment influencent-elles le calcul de la structure et la conformation finale observée? Est-il possible d'obtenir le même type de tournant sans certaines contraintes? La méthode de classification de fragments issus de protéines déterminées par RMN, présentée ici, peut permettre de répondre à ces interrogations ainsi qu'être utilisée dans l'analyse d'autres propriétés structurales. Cette méthode peut donc fournir une base de données de fragments permettant la détermination des structures tridimensionnelles en associant contraintes de distances et contraintes angulaires.

Conclusion générale

La première partie de ce travail a permis de décrire la base de données Pescador ainsi que son analyse. Pescador est une base de données centrée uniquement sur les données de conformations de peptides en solution. Le type de données permises par cette approche limite les données pouvant être traitées, mais d'un autre côté elle permet d'avoir un système de déposition rapide ainsi qu'un traitement simplifié des données. De plus, la base de données bien structurée qui est reliée aux dépositions permet une validation ainsi qu'une analyse facile et approfondie de ces données. L'accessibilité aux utilisateurs extérieurs en est par la même occasion simplifiée. Une importante quantité disponible de données sur les conformation de peptides et une rapidité à les regrouper afin de mieux comprendre les effets de la séquence sur les conformations, sont les deux principaux ingrédients pouvant nous garantir l'établissement de cette base de données spécialisée.

Pescador offre donc un nouveau moyen d'obtenir des valeurs de référence à partir de celles observées par RMN sur des peptides ou des segments de protéines, en fonction de la séquence ou de l'environnement. Un atout important réside dans le fait que le biais est réduit grâce à l'analyse d'un nombre important de peptides collectés sous différentes conditions et provenant de différents laboratoires. Le potentiel est aussi présent pour de nombreuses recherches sur les influences de la séquence, du pH ou de la température, sur des paramètres RMN spécifiques. Cependant, la clé pour améliorer les analyses reste la collecte d'un plus grand nombre des données. Nous souhaiterions donc que davantage de groupes impliqués dans la recherche sur les peptides déposent leur données dans Pescador.

La deuxième partie de ce travail a été consacrée au développement d'une méthode de classification de motifs structuraux récurrents de protéines, et à l'étude des relations

avec leurs empreintes expérimentales. Cette méthode permet de regrouper des fragments de protéines issus d'expérience RMN et possédant de multiples modèles, en clusters de fragments ayant une conformation similaire.

Une analyse des propriétés communes des clusters de fragments, en terme de contraintes nOe, a donné lieu dans un premier temps à la validation de notre méthode. Elle a permis de présenter une analyse intéressante des motifs en tournant et offre des perspectives prometteuses pour d'autres études.

Regrouper des données afin de les analyser rend possible en effet la mise en évidence de leurs propriétés globales. Ce travail a permis de caractériser deux de ces aspects : l'un au travers de Pescador sur les déplacements chimiques et l'autre au travers des motifs structuraux récurrents de protéines sur les contraintes nOe.

Troisième partie

Annexes

Chapitre 1

Matériels et Méthodes

1.1 Zones favorables de la carte de Ramachandran

La carte de Ramachandran définit les régions de l'espace conformationnel (ϕ, ψ) qui sont stériquement et énergétiquement favorables. Le nombre de résidus par structure ayant des couples (ϕ, ψ) dans les régions permises doit être maximum. Les contours des régions plus ou moins favorables ont été définis statistiquement lors d'une étude faite par Morris et al. [39]. La figure 1.1 met en évidence ces 4 régions. Les régions dites favorables sont dans ce travail les trois régions 'c', 'a' et 'g'.

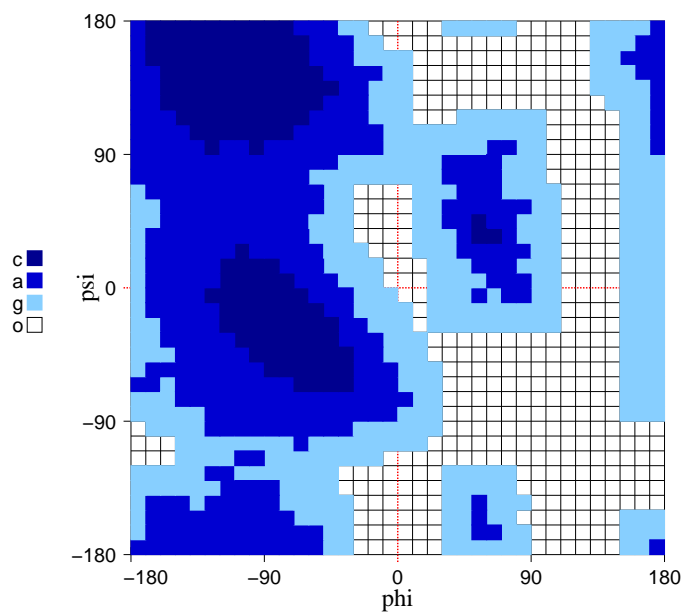


Fig. 1.1: Cartes de Ramachandran définissant les 4 régions [39] plus ou moins favorables. La région 'c' correspond à la région très favorable, 'a' à la région favorable, 'g' à la région acceptable et 'o' est la région extérieure.

1.2 Les valeurs des déplacements chimiques des trois peptides étudiés

Les valeurs des déplacements chimiques des protons alpha des peptides étudiés lors de l'application des facteurs de correction (chapitre 4 de la partie I) sont présentées dans les trois tableaux qui suivent :

- Le tableau 1.1 (page 109) correspond au peptide Carp Granulin 1-30 [57] ayant deux motifs en épingle à cheveux (code PDB : 1QGM).
- Le tableau 1.2 (page 110) correspond au fragment V3 en tournant avec un semblant de structure en hélice α [9]. Le motif en tournant β se situe dans la région 10-GPGR-13 du peptide.
- Le tableau 1.3 (page 111) correspond à la phosphatase acide lysosomale ayant un tournant β central [13]. Il se situe dans la région 5-PPGY-8.

Résidu	nr	$\delta_{H\alpha}$	$\Delta\delta_{Merutka}$	$\Delta\delta_{Merutka+Schw.}$	$\Delta\delta_{Pescador}$
VAL	1	3.890	-0.240	-0.240	-0.240
ILE	2	4.390	0.210	0.210	0.280
HIS	3	4.790	0.020	0.020	0.060
CYS	4	-	-	-	-
ASP	5	4.260	-0.370	-0.370	-0.310
ALA	6	4.060	-0.280	-0.280	-0.170
ALA	7	4.530	0.190	0.190	0.230
THR	8	4.320	-0.070	-0.070	0.020
ILE	9	4.630	0.450	0.450	0.420
CYS	10	5.250	0.670	0.560	0.340
PRO	11	4.450	0.010	0.010	0.040
ASP	12	4.430	-0.200	-0.200	-0.220
GLY	13	3.900	-0.110	-0.110	-0.190
THR	14	4.950	0.560	0.560	0.500
THR	15	4.561	0.170	0.170	0.170
CYS	16	4.851	0.270	0.270	0.200
SER	17	4.821	0.330	0.330	0.390
LEU	18	4.121	-0.230	-0.230	-0.310
SER	19	4.951	0.460	0.350	0.260
PRO	20	4.310	-0.130	-0.030	-0.050
TYR	21	4.650	0.090	0.090	0.140
GLY	22	3.735	-0.275	-0.115	-0.155
VAL	23	4.230	0.100	0.200	0.270
TRP	24	5.280	0.610	0.710	0.780
TYR	25	5.111	0.550	0.700	0.720
CYS	26	5.261	0.680	0.920	0.810
SER	27	5.011	0.520	0.410	0.310
PRO	28	4.611	0.170	0.260	0.210
PHE	29	4.661	0.040	0.040	0.030
SER	30	4.310	-0.180	-0.100	-0.130
DA			0.282	0.283	0.274
NEG			-0.072	-0.060	-0.061
POS			0.210	0.222	0.213

Tab. 1.1: Valeurs des déplacements chimiques des protons alpha observées pour chaque résidu du peptide Carp Granulin 1-30 (colonne $\delta_{H\alpha}$). Les déviations $\Delta\delta$ sont les différences entre les valeurs observées $\delta_{H\alpha}$ et les valeurs de référence corrigées ou non par les facteurs de correction. La colonne $\Delta\delta_{Merutka}$, représente les différences par rapport aux valeurs aléatoires de référence de Merutka et al. [35]. La colonne $\Delta\delta_{Merutka+Schw.}$ représente les différences par rapport aux valeurs aléatoires de référence de Merutka et al. [35] corrigées par les facteurs de Schwarzingger et al. [51]. La colonne $\Delta\delta_{Pescador}$ représente les différences par rapport aux valeurs de référence issues du sous-ensemble restreint et des facteurs de corrections de Pescador. Les déviations absolues (DA), les contributions négatives (NEG) et les positives (POS) sont la somme de ces différences.

Résidu	nr	$\delta_{H\alpha}$	$\Delta\delta_{Merutka}$	$\Delta\delta_{Merutka+Schw.}$	$\Delta\delta_{Pescador}$
TYR	1	4.210	-0.350	-0.350	-0.310
ASN	2	4.720	-0.040	0.040	0.100
LYS	3	4.230	-0.090	-0.090	-0.010
ARG	4	4.260	-0.080	-0.080	0.010
LYS	5	4.280	-0.040	-0.040	-0.010
ARG	6	4.350	0.010	0.010	0.050
ILE	7	4.150	-0.030	-0.030	0.060
HIS	8	4.710	-0.060	-0.060	-0.020
ILE	9	4.190	0.010	0.010	-0.070
GLY	10	4.105	0.095	-0.015	-0.175
PRO	11	4.460	0.020	0.020	-0.010
GLY	12	3.960	-0.050	-0.050	-0.010
ARG	13	4.260	-0.080	-0.080	-0.060
ALA	14	4.220	-0.120	-0.030	0.020
PHE	15	4.540	-0.080	0.020	0.040
TYR	16	4.620	0.060	0.140	0.040
THR	17	4.380	-0.010	0.070	0.110
THR	18	4.320	-0.070	-0.070	-0.040
LYS	19	4.290	-0.030	-0.030	-0.030
ASN	20	4.680	-0.080	-0.080	-0.030
ILE	21	4.180	0.000	0.000	0.010
ILE	22	4.500	0.320	0.320	0.320
GLY	23	3.990	-0.020	-0.020	0.000
CYS	24	4.530	-0.050	-0.050	-0.100
DA			0.075	0.071	0.068
NEG			-0.053	-0.045	-0.036
POS			0.021	0.026	0.031

Tab. 1.2: Valeurs des déplacements chimiques des protons alpha observées pour chaque résidu du fragment V3 en tournant. (Se reporter à la légende du tableau 1.1).

Résidu	nr	$\delta_{H\alpha}$	$\Delta\delta_{Merutka}$	$\Delta\delta_{Merutka+Schw.}$	$\Delta\delta_{Pescador}$
MET	1	4.16	-0.360	-0.360	-0.270
GLN	2	4.39	0.030	0.030	0.090
ALA	3	4.28	-0.060	-0.060	-0.020
GLN	4	4.58	0.220	0.110	-0.050
PRO	5	4.70	0.260	0.150	0.010
PRO	6	4.41	-0.030	-0.030	0.010
GLY	7	3.89	-0.120	-0.020	-0.040
TYR	8	4.46	-0.100	-0.100	-0.190
ARG	9	4.22	-0.120	-0.040	0.060
HIS	10	4.61	-0.160	-0.160	-0.100
VAL	11	4.07	-0.060	-0.060	-0.040
ALA	12	4.32	-0.020	-0.020	0.040
ASP	13	4.58	-0.050	-0.050	0.000
GLY	14	3.95	-0.060	-0.060	0.060
GLU	15	4.31	0.020	0.020	-0.010
ASP	16	4.57	-0.060	-0.060	0.060
HIS	17	4.69	-0.080	-0.080	0.030
ALA	18	4.15	-0.190	-0.190	-0.160
DA			0.111	0.089	0.069
NEG			-0.082	-0.072	-0.049
POS			0.029	0.017	0.020

Tab. 1.3: Valeurs des déplacements chimiques des protons alpha observées pour chaque résidu de la phosphatase acide lysosomale. (Se reporter à la légende du tableau 1.1).

1.3 Liste des 97 protéines

L'analyse des motifs structuraux est basée sur une liste de 97 protéines, provenant de la thèse de Doreleijers [12] sur la validation des structures RMN de biomolécules. Ces protéines issues d'expériences RMN, ont été déposées dans la PDB avec leurs contraintes nOe. Les fichiers de structures et de contraintes, provenant de la PDB ont été ensuite rassemblés, vérifiés et modifiés par Doreleijers avant d'avoir été convertis au format NMR-STAR.

PDB	Année	Fonction	Nb Modèles	Nb Rédius
1aps	91	HYDROLASE (ACTING ON ACID ANHYDRIDES)	5	98
1arq	93	GENE-REGULATING PROTEIN	16	106
1atx	90	SEA ANEMONE TOXIN	8	46
1bal	92	GLYCOLYSIS	56	51
1bbo	92	DNA-BINDING PROTEIN	60	56
1bcn	92	CYTOKINE	22	133
1bha	93	PHOTORECEPTOR	12	67
1bhb	93	PHOTORECEPTOR	12	67
1brv	96	GLYCOPROTEIN	48	19
1bus	90	PROTEINASE INHIBITOR	5	57
1c5a	90	COMPLEMENT FACTOR	41	66
1cb1	91	CALCIUM-BINDING PROTEIN	13	78
1ccm	93	PLANT SEED PROTEIN	8	46
1cey	94	SIGNAL TRANSDUCTION	46	128
1chl	94	NEUROTOXIN	7	36
1clb	95	CALCIUM-BINDING PROTEIN	33	75
1cld	95	TRANSCRIPTION REGULATION	29	33
1crp+r	93	ONCOGENE PROTEIN	20	166
1ctl	95	METAL-BINDING PROTEIN	19	85
1dec	94	BLOOD COAGULATION	25	39
1dmd	94	METALLOTHIONEIN	18	31
1dmf	94	METALLOTHIONEIN	18	28
1dtk	93	PRESYNAPTIC NEUROTOXIN	20	57
1edp	91	VASOCONSTRICTOR	1	17
1ego	91	ELECTRON TRANSPORT	20	85
1egr	91	ELECTRON TRANSPORT	20	85
1eph	92	GROWTH FACTOR	10	53
1epj	92	GROWTH FACTOR	5	53
1erc	94	PHEROMONE	20	40
1erd	94	PHEROMONE	20	40
1gb1	91	IMMUNOGLOBULIN BINDING PROTEIN	60	56
1gps	92	PLANT TOXIN	8	47
1gpt	92	PLANT TOXIN	8	47
1hic	92	HIRUDIN	20	51
1hiq	93	HORMONE	10	51
1his	92	HORMONE	15	46
1hit	92	HORMONE	9	51
1hiu	92	HORMONE	11	51
1hom	91	DNA-BINDING PROTEIN	19	68
1hue	95	DNA-BINDING	25	180
1hun	94	CYTOKINE(CHEMOTACTIC)	35	138

1hwa	92	HYDROLASE(O-GLYCOSYL)	1	129
ligl	94	GROWTH FACTOR	20	67
litl	92	CYTOKINE	1	130
lkal	95	PLANT PROTEIN	10	29
lkst	91	AGGREGATION INHIBITOR, GP ANTAGONIST	8	68
lleb	94	TRANSCRIPTION REGULATION	28	72
llqc	96	TRANSCRIPTION REGULATION	32	56
lmaj	93	IMMUNOGLOBULIN	15	113
lmak	93	IMMUNOGLOBULIN	15	113
lmbf	95	DNA BINDING PROTEIN	50	52
lmbk	95	DNA BINDING PROTEIN	50	52
lmdj+k	95	COMPLEX (ELECTRON TRANSPORT/PEPTIDE)	30	118
lmhu	90	METALLOTHIONEIN	1	31
lmrh	90	METALLOTHIONEIN	1	31
lmrt	90	METALLOTHIONEIN	1	31
lnbt	91	TOXIN	12	132
lnhn	94	DNA-BINDING	41	79
lnor	93	NEUROTOXIN	19	61
lnrb	95	NEUROTOXIN	20	63
lntx	92	NEUROTOXIN	20	60
locp	95	DNA-BINDING PROTEIN	20	67
lolh	94	ANTI-ONCOGENE PROTEIN	35	168
lpba	91	HYDROLASE(C-TERMINAL PEPTIDASE)	20	81
lpcp	94	LIPASE PROTEIN COFACTOR	25	93
lpdc	91	COLLAGEN-BINDING TYPE II DOMAIN	1	45
lpfl	94	REGULATORY PROTEIN	20	139
lpis	94	CARBOXYLIC ESTER HYDROLASE	20	124
lpk2	91	PLASMINOGEN ACTIVATOR	1	90
lpog	94	DNA BINDING PROTEIN	13	62
lpra	91	GENE REGULATING PROTEIN	20	69
lrgd	95	DNA-BINDING PROTEIN	11	71
lrpr	91	TRANSCRIPTION REGULATION	10	126
lrtn	95	CHEMOKINE	20	136
lsfw	96	HYDROLASE	18	124
lssp	95	DNA-BINDING PROTEIN	1	62
ltfs	95	TOXIN	20	60
ltmn	95	MUSCLE PROTEIN	16	91
ltrx	90	ELECTRON TRANSPORT	10	108
ltur	94	SERINE PROTEINASE INHIBITOR	12	56
ltus	94	SERINE PROTEINASE INHIBITOR	12	56
lznf	89	ZINC FINGER DNA BINDING DOMAIN	37	27
2aas	92	HYDROLASE(ENDORIBONUCLEASE)	32	124
2bbn	92	CALCIUM-BINDING PROTEIN	21	174
2gda	94	GLUCOCORTICOID RECEPTOR	24	72
2igg	92	IMMUNOGLOBULIN-BINDING PROTEIN	27	64
2igh	92	IMMUNOGLOBULIN-BINDING PROTEIN	24	61
2il8	90	CYTOKINE	30	142
2mhu	90	METALLOTHIONEIN	1	30
2mrh	90	METALLOTHIONEIN	1	30
2mrt	90	METALLOTHIONEIN	1	30
2sob	95	HYDROLASE (PHOSPHORIC DIESTER)	10	103
3ci2	91	SERINE PROTEASE INHIBITOR	20	64
3cti	91	PROTEINASE INHIBITOR (TRYPSIN)	6	29
4trx	90	ELECTRON TRANSPORT	33	105
4znf	90	ZINC FINGER / DNA BINDING DOMAIN	41	30
9pcy	91	ELECTRON TRANSPORT	16	99

Chapitre 2

Publication

– A. Pajon, W. F. Vranken, M. A. Jimenez, M. Rico and S. J. Wodak.

« PESCADOR : The PEptides in Solution ConformAtion Database : Online Resource. » *J Biomol NMR*, **23** : 85-102, 2002.

Accepté pour publication le 29 avril 2002 dans *Journal of Biomolecular NMR*.

Bibliographie

- [1] N. H. Andersen and H. Tong. Empirical parameterization of a model for predicting peptide helix/coil equilibrium populations. *Protein Sci*, 6(9) :1920–1936, Sep 1997.
- [2] R. Aurora, R. Srinivasan, and G. D. Rose. Rules for alpha-helix termination by glycine. *Science*, 264(5162) :1126–1130, May 1994.
- [3] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, 28(1) :45–48, Jan 2000.
- [4] R. L. Baldwin. Alpha-helix formation by peptides of defined sequence. *Biophys Chem*, 55(1-2) :127–135, Jun 1995.
- [5] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. Genbank. *Nucleic Acids Res*, 28(1) :15–18, Jan 2000. (eng).
- [6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1) : 235–242, Jan 2000.
- [7] A. Bundi and K. Wüthrich. ¹H-NMR parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH. *Biopolymers*, 18 :285–298, 1979.
- [8] A. Chakrabartty, J. A. Schellman, and R. L. Baldwin. Large differences in the helix propensities of alanine and glycine. *Nature*, 351(6327) :586–588, Jun 1991.
- [9] K. Chandrasekhar, A. T. Profy, and H. J. Dyson. Solution conformational preferences of immunogenic peptides derived from the principal neutralizing determinant of the hiv-1 envelope glycoprotein gp120. *Biochemistry*, 30(38) :9187–9194, Sep 1991.
- [10] W. J. Conover. *Practical nonparametric statistics*. New York : John Wiley and Sons, 1971.
- [11] G. Cornilescu, F. Delaglio, and A. Bax. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR*, 13 : 289–302, 1999.
- [12] J. F. Doreleijers. *Validation of Biomolecular NMR Structures*. PhD thesis, Utrecht University, 1999.
- [13] W. Eberle, C. Sander, W. Klaus, B. Schmidt, K. von Figura, and C. Peters. The essential tyrosine of the internalization signal in lysosomal acid phosphatase is part of a beta turn. *Cell*, 67(6) :1203–1209, Dec 1991.

- [14] M. S. Edwards, J. E. Sternberg, and J. M. Thornton. Structural and sequence patterns in the loops of beta alpha beta units. *Protein Eng*, 1(3) :173–181, Jun 1987.
- [15] A.V. Efimov. Patterns of loop regions in proteins. *Curr Opin Struc Biol*, 3 :379–384, 1993.
- [16] S. H. Gellman. Minimal model systems for beta sheet secondary structure in proteins. *Curr Opin Chem Biol*, 2(6) :717–725, Dec 1998.
- [17] S. R. Griffiths-Jones, A. J. Maynard, and M. S. Searle. Dissecting the stability of a beta-hairpin peptide that folds in water : Nmr and molecular dynamics analysis of the beta-turn and beta-strand contributions to folding. *J Mol Biol*, 292(5) :1051–1069, Oct 1999.
- [18] S. R. Griffiths-Jones and M. S. Searle. Structure, folding and energetics of cooperative interactions between the beta-strands of a de novo designed three-stranded antiparallel beta-sheet peptide. *J Am. Chem. Soc.*, 122 :8350–8356, 2000.
- [19] S. R. Griffiths-Jones, G. J. Sharman, A. J. Maynard, and M. S. Searle. Modulation of intrinsic phi,psi propensities of amino acids by neighbouring residues in the coil regions of protein structures : Nmr analysis and dissection of a beta-hairpin peptide. *J Mol Biol*, 284(5) :1597–1609, Dec 1998.
- [20] S. R. Hall. The star file : A new format for electronic data transfer and archiving. *J. Chem. Inf. Comput. Sci.*, 31 :326–333, 1990.
- [21] S. R. Hall and N. Spadaccini. The star file : Detailed specifications. *J. Chem. Inf. Comput. Sci.*, 34 :505–508, 1994.
- [22] R. Ihaka and R. Gentleman. R : A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3) :299–314, 1996.
- [23] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, A(32) :922–923, 1976.
- [24] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, A(34) :827–828, 1978.
- [25] L. A. Kelley, S. P. Gardner, and M. J. Sutcliffe. An automated approach for clustering an ensemble of nmr-derived protein structures into conformationally related subfamilies. *Protein Eng*, 9(11) :1063–1065, Nov 1996.
- [26] J. Kemmink and T. E. Creighton. The physical properties of local interactions of tyrosine residues in peptides and unfolded proteins. *J Mol Biol*, 245(3) :251–260, Jan 1995.
- [27] J. Kemmink, C. P. van Mierlo, R. M. Scheek, and T. E. Creighton. Local structure due to an aromatic-amide interaction observed by ¹H-nuclear magnetic resonance spectroscopy in peptides related to the N terminus of bovine pancreatic trypsin inhibitor. *J Mol Biol*, 230(1) :312–322, Mar 1993.
- [28] P. J. Kraulis and T. A. Jones. Determination of three-dimensional protein structures from nuclear magnetic resonance data using fragments of known structures. *Proteins*, 2(3) :188–201, 1987.

- [29] E. Lacroix, T. Kortemme, M. Lopez de la Paz, and L. Serrano. The design of linear peptides that fold as monomeric beta-sheet structures. *Curr Opin Struct Biol*, 9(4) : 487–493, Aug 1999.
- [30] E. Lacroix, A. R. Viguera, and L. Serrano. Elucidating the folding problem of α -helices : local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J Mol Biol*, 284(1) :173–191, Nov 1998.
- [31] R. A. Laskowski, J. A. Rullmann, M. W. MacArthur, R. Kaptein, and J. M. Thornton. AQUA and PROCHECK-NMR : programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, 8(4) :477–486, Dec 1996.
- [32] E. Lopez-Hernandez and L. Serrano. Structure of the transition state for folding of the 129 aa protein chey resembles that of a smaller protein, ci-2. *Fold Des*, 1(1) : 43–55, 1995.
- [33] E. Lopez-Hernandez and L. Serrano. Structure of the transition state for folding of the 129 aa protein chey resembles that of a smaller protein, ci-2. *Fold Des*, 1(1) : 43–55, 1996.
- [34] J. L. Markley, A. Bax, Y. Arata, C. W. Hilbers, R. Kaptein, B. D. Sykes, P. E. Wright, and K. Wüthrich. Recommendations for the presentation of NMR structures of proteins and nucleic acids. *J Mol Biol*, 280(5) :933–952, Jul 1998.
- [35] G. Merutka, H. J. Dyson, and P. E. Wright. 'random coil' ^1H chemical shifts obtained as a function of temperature and trifluoroethanol concentration for the peptide series GGXGG. *J Biomol NMR*, 5(1) :14–24, Jan 1995.
- [36] G. L. Millhauser, C. J. Stenland, K. A. Bolin, and F. J. van de Ven. Local helix content in an alanine-rich peptide as determined by the complete set of $^3\text{J}_{\text{HN}}$ alpha coupling constants. *J Biomol NMR*, 7(4) :331–334, Jun 1996.
- [37] E. J. Milner-White and R. Poet. Four classes of beta-hairpins in proteins. *Biochem J*, 240(1) :289–292, Nov 1986.
- [38] E. J. Milner-White and R. Poet. Four classes of beta-hairpins in proteins. *Trends Biochem Sci*, 12 :189–192, May 1987.
- [39] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton. Stereochemical quality of protein structure coordinates. *Proteins*, 12(4) :345–364, Apr 1992.
- [40] V. Munoz and L. Serrano. Elucidating the folding problem of helical peptides using empirical parameters. *Nat Struct Biol*, 1(6) :399–409, Jun 1994.
- [41] V. Munoz and L. Serrano. Elucidating the folding problem of helical peptides using empirical parameters. II. helix macrodipole effects and rational modification of the helical content of natural peptides. *J Mol Biol*, 245(3) :275–296, Jan 1995.
- [42] V. Munoz and L. Serrano. Elucidating the folding problem of helical peptides using empirical parameters. III. temperature and pH dependence. *J Mol Biol*, 245(3) : 297–308, Jan 1995.

- [43] V. Munoz, L. Serrano, M. A. Jimenez, and M. Rico. Structural analysis of peptides encompassing all alpha-helices of three alpha/beta parallel proteins : Che-y, flavodoxin and p21-ras : implications for alpha-helix stability and the folding of alpha/beta parallel proteins. *J Mol Biol*, 247(4) :648–669, Apr 1995.
- [44] B. Odaert, F. Jean, C. Boutillon, E. Buisine, O. Melnyk, A. Tartar, and G. Lipens. Synthesis, folding, and structure of the beta-turn mimic modified b1 domain of streptococcal protein g. *Protein Sci*, 8(12) :2773–2783, Dec 1999.
- [45] B. Oliva, P. A. Bates, E. Querol, F. X. Aviles, and M. J. Sternberg. An automated classification of the structure of protein loops. *J Mol Biol*, 266(4) :814–830, Mar 1997.
- [46] S. Padmanabhan and R. L. Baldwin. Tests for helix-stabilizing interactions between various nonpolar side chains in alanine-based peptides. *Protein Sci*, 3(11) :1992–1997, Nov 1994.
- [47] M. J. Rooman, J. Rodriguez, and S. J. Wodak. Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol*, 213(2) :327–336, May 1990.
- [48] P. Royston. Algorithm AS 181 : The W test for normality. *Applied Statistics*, 31 : 176–180, 1982.
- [49] C. M. Santiveri, M. Rico, and M. A. Jimenez. Position effect of cross-strand side-chain interactions on beta-hairpin formation. *Protein Sci*, 9(11) :2151–2160, Nov 2000.
- [50] C. M. Santiveri, M. Rico, and M. A. Jimenez. ¹³C(alpha) and ¹³C(beta) chemical shifts as a tool to delineate beta-hairpin structures in peptides. *J Biomol NMR*, 19 (4) :331–345, Apr 2001.
- [51] S. Schwarzingler, G. J. Kroon, T. R. Foss, J. Chung, P. E. Wright, and H. J. Dyson. Sequence-dependent correction of random coil NMR chemical shifts. *J Am Chem Soc*, 123(13) :2970–2978, Apr 2001.
- [52] B. R. Seavey, E. A. Farr, W. M. Westler, and J. L. Markley. A relational database for sequence-specific protein NMR data. *J Biomol NMR*, 1(3) :217–236, Sep 1991.
- [53] B. L. Sibanda and J. M. Thornton. Beta-hairpin families in globular proteins. *Nature*, 316(6024) :170–174, Jul 1985.
- [54] STAR Dictionary Definition Language : Initial Specification. S. r. hall and a. p. f. cook. *J. Chem. Inf. Comput. Sci.*, 35 :819–825, 1995.
- [55] G. Stoesser, W. Baker, A. van den Broek, E. Camon, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, N. Redaschi, P. Stoehr, M. A. Tuli, K. Tzouvara, and R. Vaughan. The embl nucleotide sequence database. *Nucleic Acids Res*, 30(1) :21–26, Jan 2002. (eng).
- [56] N. Taddei, F. Chiti, T. Fiaschi, M. Bucciantini, C. Capanni, M. Stefani, L. Serrano, C. M. Dobson, and G. Ramponi. Stabilisation of alpha-helices by site-directed mutagenesis reveals the importance of secondary structure in the transition state for acylphosphatase folding. *J Mol Biol*, 300(3) :633–647, Jul 2000.

- [57] W. F. Vranken, Z. G. Chen, P. Xu, S. James, H. P. Bennett, and F. Ni. A 30-residue fragment of the carp granulin-1 protein folds into a stack of two beta-hairpins similar to that found in the native protein. *J Pept Res*, 53(5) :590–597, May 1999.
- [58] G. Wagner, D. Neuhaus, E. Worgotter, M. Vasak, J. H. Kagi, and K. Wüthrich. Nuclear magnetic resonance identification of half-turn and 3(10)-helix secondary structure in rabbit liver metallothionein-2. *J Mol Biol*, 187(1) :131–135, Jan 1986.
- [59] C. M. Wilmot and J. M. Thornton. β -turns and their distortions : a proposed new nomenclature. *Protein Eng*, 3(6) :479–493, May 1990.
- [60] R. T. Wintjens, M. J. Rooman, and S. J. Wodak. Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J Mol Biol*, 255(1) :235–253, Jan 1996.
- [61] D. S. Wishart, C. G. Bigam, A. Holm, R. S. Hodges, and B. D. Sykes. ^1H , ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. i. investigations of nearest-neighbor effects. *J Biomol NMR*, 5(1) :67–81, Jan 1995.
- [62] D. S. Wishart and B. D. Sykes. The ^{13}C chemical-shift index : a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J Biomol NMR*, 4(2) :171–180, Mar 1994.
- [63] D. S. Wishart, B. D. Sykes, and F. M. Richards. The chemical shift index : a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, 31(6) :1647–1651, Feb 1992.
- [64] J. Wojcik, J. P. Mornon, and J. Chomilier. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol*, 289(5) :1469–1490, Jun 1999.
- [65] K. Wüthrich, M. Billeter, and W. Braun. Polypeptide secondary structure determination by nuclear magnetic resonance observation of short proton-proton distances. *J Mol Biol*, 180(3) :715–740, Dec 1984.
- [66] R. Zerella, P. Y. Chen, P. A. Evans, A. Raine, and D. H. Williams. Structural characterization of a mutant peptide derived from ubiquitin : implications for protein folding. *Protein Sci*, 9(11) :2142–2150, Nov 2000.