Thèse présentée pour obtenir le grade de
Docteur de l'Université de Strasbourg

par Matthieu Tanty

Développement de nouveaux outils et approches pour l'étude des protéines intrinsèquement désordonnées

Soutenue publiquement le 9 septembre 2011

Membres du jury

Pr. Christian Roumestand	Rapporteur externe
Dr. Martin Blackledge	Rapporteur externe
Pr. Waïs Hosseini	Examinateur interne
Dr. Véronique Receveur-Bréchot	Examinatrice externe
Dr. Jean-Philippe Starck	Invité
Dr. Marc-André Delsuc	Directeur de thèse

Table des matières

Remerciements

Je souhaite évidemment remercier mon directeur de thèse, le docteur Marc-André Delsuc. Avoir eu l'opportunité de travailler pendant 3 ans sous sa responsabilité a été une formidable expérience. Grâce à lui, l'un des buts essentiels que je m'étais fixés en commençant ma thèse a été largement rempli : apprendre encore et toujours. Il a su me faire profiter de ses connaissances encyclopédiques dans de nombreux domaines et me guider lorsque j'en avais besoin. J'espère avoir été à la hauteur de l'enjeu majeur que représente l'étude des IDP.

En deuxième lieu, je veux remercier le professeur Bruno Kieffer, toute l'équipe de RMN biomoléculaire de l'IGBMC passée et présente ainsi que les membres de la société NMRtec pour leur accueil et leur soutien tant technique que moral. Merci aussi à l'équipe de modélisation de l'IGBMC et à l'équipe oncoprotéines de l'IREBS sans qui le couloir serait bien calme ainsi qu'à Pascal Eberling sans qui rien n'aurait été possible.

Je remercie aussi les membres du jury, les professeurs Waïs Hosseini et Christian Roumestand ainsi que les docteurs Véronique Receveur-Bréchot, Martin Blackledge et Jean-Philippe Starck, qui ont répondu positivement à notre invitation et qui ont endossé la lourde charge de juger ce travail.

Un immense merci à ma fiancée, Julie, qui a accepté que nous soyons séparés de plus de 700 kilomètres pendant ces 3 ans. Merci aussi à toute ma famille qui m'a soutenu comme elle l'a toujours fait même si elle aurait évidemment préféré me voir plus souvent.

Enfin, merci aux personnes qui ont rendu ce séjour en Alsace riche en émotions : Katja, Emeline et Christian pour les inoubliables cours d'allemands, les magnifiques voyages et les mètres carrés de gâteaux ainsi que Marie-Aude, Justine et Eric pour nos épuisantes séances de badminton. Et merci à Doriane pour sa vue sur la cathédrale.

Matthieu Tanty

Abbréviations et accronymes

Voici la liste des abbréviations et accronymes rencontrés dans cette thèse, classés par ordre alphabétique.

BMRB	Biological Magnetic Resonance Bank
CD	Circular Dichroism
CSI	Chemical Shift Index
d_{f_e}	dimension fractale calculée par la méthode des distances bout-à-bout
d_{f_r}	dimension fractale calculée par la méthode des rayons de giration
DOSY	Diffusion-Ordered SpectroscopY
DLS	Dynamic Light Scattering
DSS	DiméthylSilapentaneSulfonate
DTT	dithiothréitol
FRET	Förster Resonance Energy Transfer
IDP	Intrinsically Disordered Protein ou protéine intrinsèquement désordonnée
ILT	Inverse Laplace Transform ou transformée de Laplace inverse
PDB	Protein DataBank
PDI	PolyDispersity Index ou indice de polydispersité
PEO	Poly(Ethylene Oxide) ou poly(oxyde d'éthylène)
PFGSE	Pulse Field Gradient Spin Echo ou écho de spin avec gradients de champ pulsés
PPI	PolyProline I
PPII	PolyProline II
RamaDA	Ramachandran Domain Analysis (voir chapitre 2)
RamaDP	Ramachandran Domain Prediction (voir chapitre 2)
RAR (γ)	Retinoic Acid Receptor ou récepteur de l'acide rétinoïque (γ)
RefDB	re-Referenced DataBase
RMN	Résonance Magnétique Nucléaire
RPE	Résonance Paramagnétique Électronique
SAXS	Small Angle X-ray Scattering
SH3	Src Homology-3
SLiM	Small Linear Motives ou petits motifs linéaires
SSP	Secondary Structure Propensity
TMS	TetraMéthylSilane
Tris	tris(hydroxyméthyl)aminométhane

Introduction

"L'incohérence n'existe pas, le désordre n'est qu'un ordre différent."

Robert Malaval

Les protéines intrinsèquement désordonnées (IDP : Intrinsically Disordered Proteins) occupent une place particulière au sein de leurs semblables. En effet, peu de protéines peuvent se targuer d'avoir mis à mal l'un des principaux paradigmes de la biologie : le paradigme structure-fonction.

Structure des protéines

La structure d'une protéine peut être décrite à différents niveaux :

- La structure primaire correspond à l'enchaînement des acides-aminés qui la composent. Elle constitue la carte d'identité de la protéine, l'empreinte digitale qui permet de l'associer à d'autres protéines de la même famille mais qui assure son unicité.
- La structure secondaire correspond à un agencement local d'une série d'acides-aminés. Cette structure a pour origine l'interaction entre les acides-aminés de la

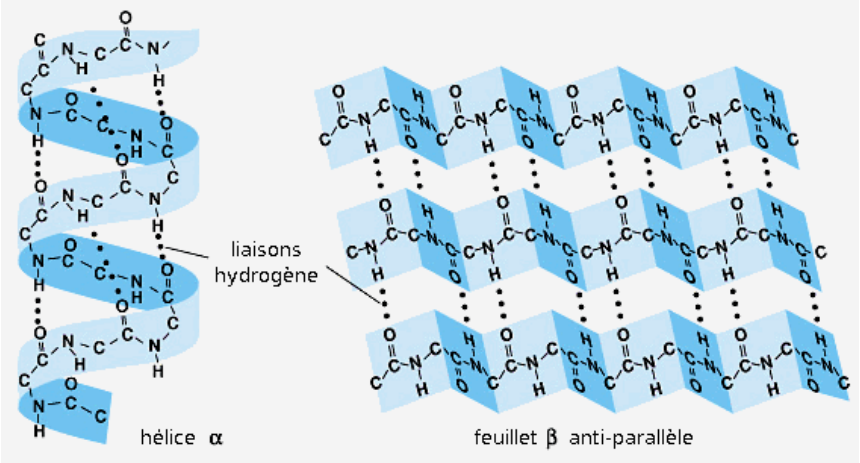

Figure 1 – *Exemples de réseaux de liaisons hydrogène au sein d'éléments de structure secondaire : l'hélice α (à gauche) et le feuillet β parallèle (à droite).*

protéine et leur environnement (solvant et autres molécules). La formation d'une telle structure permet alors d'abaisser l'énergie libre de la molécule et de la stabiliser. Les différents éléments de structure secondaires seront détaillés plus loin.

– La structure tertiaire d'une protéine correspond à la disposition de chacun des éléments de structure secondaire les uns par rapport aux autres.
– Enfin, la structure quaternaire correspond à l'assemblage de plusieurs protéines les unes avec les autres.

Les acides-aminés, de part leur formule chimique, sont à la fois des donneurs de liaisons hydrogène (*via* leur fonction amine) et des accepteurs d'hydrogène (*via* leur fonction carbonyle), la proline mise à part. La nature a su jouer de ces propriétés en stabilisant les éléments de structure secondaire par des liaisons hydrogène. La figure 1 montre deux exemples d'éléments de structure secondaire : l'hélice α et le feuillet β.

L'hélice α forme des liaisons hydrogène entre le carbonyle du résidu i et l'hydrogène de l'amine du résidu $i+4$. Il s'agit de la forme la plus courante d'hélice et l'une des plus stables car elle met en jeu des conformations facilement atteintes par les acides-aminés. Il existe cependant deux autres formes d'hélices, π et 3_{10}, qui forment des liaisons entre

les résidus i et $i + 5$ ou i et $i + 3$, respectivement.

Les feuillets β, quant à eux, forment des liaisons hydrogène entre plusieurs brins qui ne sont pas nécessairement à la suite les uns des autres dans la séquence. Les brins s'associent de façon parallèle, c'est-à-dire avec leurs extrémités N-ter et C-ter du même côté, ou anti-parallèle, c'est-à-dire avec les extrémités inversées.

Remise en cause du paradigme structure-fonction

Le paradigme structure-fonction postule que la fonction exacte d'une protéine émerge de sa structure tertiaire par rapprochement des fonctions chimiques présentes sur les chaînes latérales des acides-aminés. Ce paradigme a été formulé principalement après l'étude de nombreuses enzymes, liant ainsi la formation des sites actifs avec leur structure [1].

Cependant, à l'opposé des protéines repliées, les IDP ne possèdent pas de structure secondaire (et *a fortiori* pas de structure tertiaire) et se situent donc à mi-chemin entre les protéines globulaires très bien définies et les protéines dénaturées qui ont perdu la structure qu'elles possédaient. L'application du paradigme laisserait alors penser qu'elles n'ont pas de fonction. Pourtant, il s'avère que les IDP servent bel et bien à l'organisme pour, par exemple, la régulation de la transcription ou de la traduction de gènes, la transmission d'un signal dans la cellule ou encore la reconnaissance de molécules [2]. Elles sont aussi impliquées dans de nombreuses maladies neurodégénératives [3–5].

Le paradigme structure-fonction considère au départ une protéine comme un objet rigide, statique ou bien un ensemble de domaines rigides ayant quelques degrés de liberté les uns par rapport aux autres. La vision de la protéine a depuis changé afin d'intégrer un paramètre dynamique important : la flexibilité.

Dans un premier temps, cette flexibilité inclut les différentes formes allostériques de

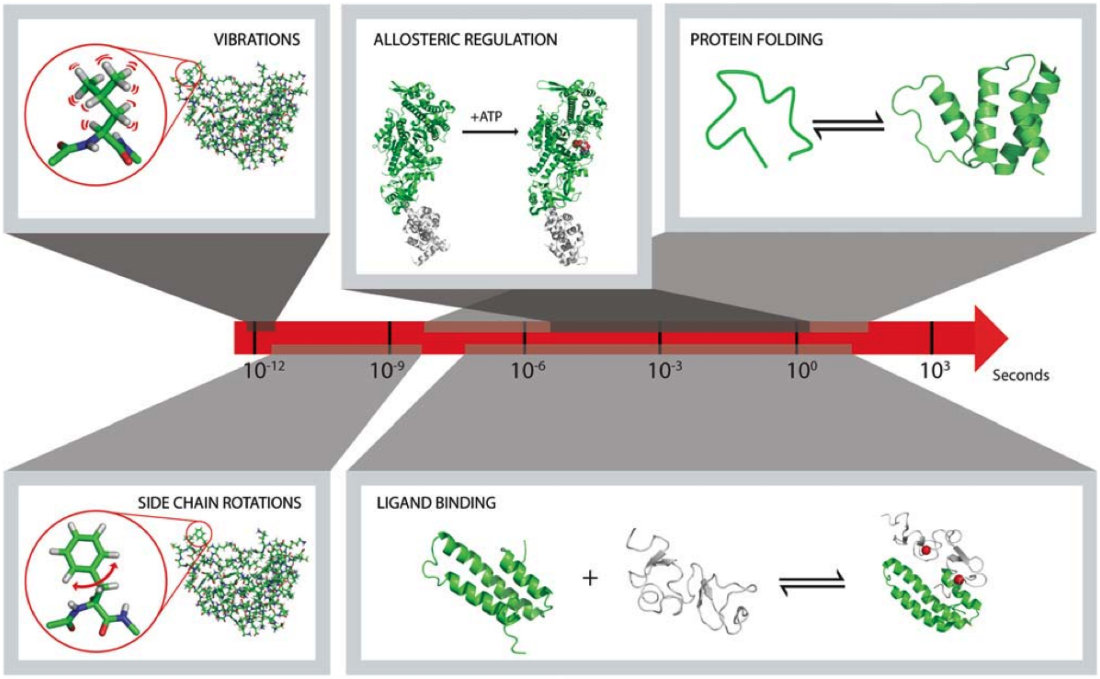

Figure 2 – *Différentes échelles de temps pour les phénomènes relevant de la flexibilité de la protéine. Figure extraite de Teilum et al. [6]*

la protéine, son adaptation à un ligand, la rotation des chaînes latérales de ses acides-aminés voire les vibrations de ses différentes liaisons chimiques. Ces phénomènes ont lieu à des échelles de temps différentes comme l'indique la figure 2. Pour chaque échelle de temps, une technique spectroscopique existe mais on ne peut pas les observer avec une seule technique bien que la RMN (Résonance Magnétique Nucléaire), par exemple, soit capable de couvrir un large intervalle d'échelles de temps. Les phénomènes décrits ne remettent pas en cause le paradigme.

La découverte des IDP, cependant, met le problème de la flexibilité des protéines au coeur du débat sur leur fonctionnement. Il s'agit bel et bien d'un problème car le repliement ou non des protéines actives malmène le paradigme structure-fonction. L'équilibre entre les formes désordonnée et structurée s'ajoute à la liste des phénomènes de flexibilité à prendre en compte dans une étude biologique.

Le désordre en biologie

Si pendant longtemps les IDP sont restées invisibles, c'est en grande partie parce que les outils développés en biologie sont basés sur le fait que les protéines que l'on souhaite observer sont repliées. Les IDP ont donc souvent été soit détruites par les procédés biochimiques utilisés, auxquels elles sont plus vulnérables, soit considérées comme de simples artefacts de manipulation. Lorsque les premières IDP ont été mises en évidence, on a alors pensé qu'elles n'avaient aucun rôle dans l'organisme ou qu'elles avaient été dénaturées dans les conditions de travail.

Cependant, les IDP sont des protéines actives. Elles servent notamment à la régulation de la transcription ou de la traduction de gènes, à la transmission de signaux dans la cellule ou encore à la reconnaissance de molécules [2]. Elles sont aussi impliquées dans de nombreuses maladies [7] comme les maladies neurodégénératives (maladies d'Alzheimer et d'Huntington par exemple [3–5]).

Plusieurs théories se sont alors suivies pour expliquer l'activité des IDP. Dans un premier temps, les IDP étudiées se liaient à leur partenaire en se structurant. Le paradigme structure-fonction était alors de nouveau le bienvenu. Il fut donc proposé que les formes repliée et dépliée d'une même protéine sont en équilibre en solution et que la forme repliée a une structure reconnaissable par le partenaire, qui peut alors interagir avec elle. La structure trouvée est appelée structure transitoire car elle n'apparaît que très rarement [8–10]. Dans ce modèle, la forme désordonnée est inactive.

Entre ces deux états extrêmes, un état intermédiaire déjà connu et défini par ailleurs [11], appelé *molten globule*, a été inséré dans le modèle. Il possède les mêmes éléments de structure secondaire que la forme repliée de la protéine mais n'en possède pas la structure tertiaire. Parfois cet état peut être observé en solution et ainsi donner des informations sur la forme structurée de la protéine [12].

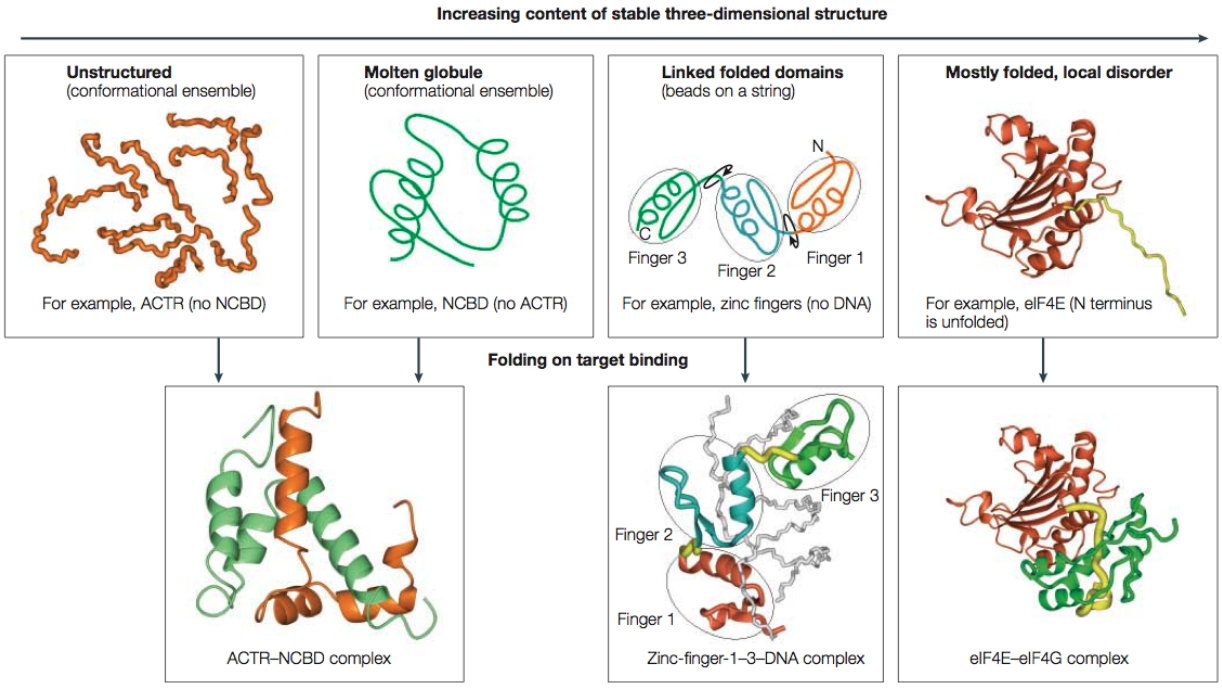

Figure 3 – *Classification. Figure extraite de Dyson et Wright [2]*

En dehors du molten globule, il existe de nombreuses façons de décrire une IDP ou une région désordonnée de protéine. Pour y voir clair, Dyson et Wright ont recueilli les plus communes et ont établi une classification en fonction de leur degré de structure [2]. Cette classification est reprise par la figure 3.

On remarque dans cette classification que seules les IDP qui se replient en liant leur partenaire (à gauche de la figure) sont considérées. Or, on sait depuis peu que ce modèle d'équilibre replié/déplié n'est pas applicable à toutes les IDP [13, 14]. Il existe en effet des IDP qui lient leurs partenaires sans pour autant gagner en structure.

En effet, certaines IDP sont reconnues par leur partenaire uniquement grâce à leur séquence primaire. De petits motifs linéaires (SLiM : Small Linear Motives) dans l'enchaînement d'acides-aminés permettent une interaction de faible intensité car peu spécifique [15, 16]. De plus, concentrer ainsi les interactions possibles permet aux IDP d'interagir avec de multiples partenaires tout en gardant une faible taille. C'est vrai-

semblablement le moyen utilisé par les protéines virales afin de déstabiliser tout un interactome [17].

Notre étude

L'étude des IDP devient un enjeu majeur afin de comprendre cette partie de la biologie, méconnue jusqu'à il y a peu. Malheureusement, une grande majorité des outils développés en biologie pour les protéines repliées ne sont pas applicables aux IDP puisqu'elle suppose la présence d'une structure tertiaire. Bien que de plus en plus de groupes de recherches s'intéressent à ces protéines et aux façons de les étudier [18–21], il est nécessaire de créer ou de découvrir de nouveaux moyens d'analyser les IDP.

Pour apporter notre pierre à l'édifice, nous avons abordé les IDP sous différents angles en utilisant une approche multi-disciplinaire. Parmi toutes les techniques spectroscopiques disponibles, nous nous sommes principalement intéressés à la RMN car elle permet d'avoir une vue dynamique de la protéine étudiée. La bioinformatique tient aussi une place importante au sein de cette étude puisqu'elle nous a permis de développer des modèles du comportement des IDP. Néanmoins, le SAXS (Small Angle X-ray Scattering) vient aussi compléter les résultats obtenus.

Trois grandes méthodes d'analyse des IDP ont été développées durant cette thèse. Dans un premier temps, nous parlerons de la détermination de la dimension fractale des protéines afin de connaître leur comportement hydrodynamique. Nous avons pour cela utilisé une méthode propre à la chimie des polymères que nous avons appliqué à des peptides polyproline ainsi qu'à une IDP salivaire riche en proline. Dans un deuxième temps, nous décrirons quelles informations peuvent être tirées des conformations des acides-aminés d'une protéine, structurée ou non, et comment prédire ces conformations à partir des déplacements chimiques. Enfin, le dernier point donnera un aperçu de

l'étude de différents peptides à partir de données expérimentales, des outils précédents et d'un générateur aléatoire de conformations.

Chapitre 1

Diffusion et dimension fractale

Les protéines, quelles qu'elles soient, sont composées d'une chaîne d'acides-aminés. Elles peuvent donc être considérées comme des hétéropolymères.

Il est alors tentant d'aller chercher chez les polymères ce qui manque chez les IDP, à savoir des paramètres pertinents pour pouvoir les décrire facilement. Le paramètre sur lequel nous nous sommes arrêtés est la dimension fractale.

Les protéines ne sont pas des objets fractaux à proprement parler. En effet, en mathématiques, une fractale est infinie et ne souffre pas d'un changement d'échelle, elle présente toujours le même schéma quelque soit le niveau de détails observé. Évidemment, aucun objet physique n'est infini et donc ne peut pas pas être considéré à proprement parler comme fractal. Dans le cas des protéines par exemple, un feuillet n'est pas composé de petits feuillets eux-mêmes composés de plus petits feuillets, etc...

Par contre, si les objets ne sont pas fractals, certaines de leurs propriétés peuvent l'être. Pour être considérées comme telles, ces propriétés doivent avoir des valeurs différentes en fonction de l'échelle à laquelle elles sont mesurées et répondre à la loi de puissance suivante :

$$P \propto E^D \tag{1.1}$$

où P est la valeur de la propriété, E le facteur d'échelle et D la dimension fractale de la propriété. Par exemple, la carte de la Grande-Bretagne n'est pas un objet fractal mais

Figure 1.1 – *Mesures de longueur de la côte de Grande-Bretagne à différentes échelles (de gauche à droite : 200 km, 100 km et 50 km).*

on peut montrer que la longueur de sa côte l'est. La figure 1.1 montre comment, selon l'échelle à laquelle on mesure la longueur de la côte, on obtient des valeurs différentes. En variant les échelles, on peut extraire la dimension fractale de la longueur de la côte de Grande-Bretagne qui est de 1.25.

Tout comme pour la carte de la Grande-Bretagne, certains objets finis et en particulier les polymères et les protéines, ne sont pas fractals mais possèdent des propriétés qui le sont [22]. Il existe de nombreuses sortes de dimensions fractales qui ne décrivent pas les mêmes comportements selon la propriété observée et l'échelle choisie. La dimension fractale que nous présentons ici est une dimension fractale en masse, qui décrit la façon dont la masse est repartie autour du centre de masse de la protéine. La propriété observée est le rayon de giration et le facteur d'échelle est le nombre de résidus des protéines.

Nous verrons dans ce chapitre que la dimension fractale définie ici se révèle assez simple à mesurer et offre un intervalle de valeurs confortable. Afin de déterminer les avantages et les limitations de ce paramètre, nous avons étudié un motif structural par-

ticulier : l'hélice polyproline-II (PPII). Cette hélice est d'autant plus importante qu'elle peut interagir avec de nombreux partenaires qui la reconnaissent spécifiquement et qu'on la retrouve dans certaines protéines structurées. Nous nous sommes ensuite logiquement tournés vers une IDP contenant plus de 40% de prolines pour comparer sa dimension fractale à celle d'une hélice polyproline.

1.1 Introduction

1.1.1 Dimension fractale

La théorie de Flory [23], bien que basée sur de nombreuses approximations, est l'un des piliers de la physico-chimie des polymères. Flory y definit notamment un paramètre d'interaction χ relié à l'enthalpie de mélange d'un polymère dans un solvant ΔH_m par l'équation 1.2, où k est la constante de Boltzmann, T la température de la solution, n le nombre d'unités monomériques en solution au total et ϕ la probabilité qu'une molécule de solvant soit proche d'une unité monomérique.

$$\Delta H_m = kTn\phi\chi \tag{1.2}$$

χ décrit donc les interactions entre le solvant et le polymère et permet de connaître le comportement d'un polymère à une température donnée dans un solvant connu.

En écrivant le potentiel chimique dû aux interactions entre une unité monomérique et le solvant, Flory obtient alors l'équation suivante :

$$\mu - \mu_0 = RT(\tfrac{1}{2} - \chi)v^2 \tag{1.3}$$

R est la constante des gaz parfaits et v la fraction volumique de polymère en solution. De plus, voyant expérimentalement que χ dépend de la température, il définit une

température Θ telle que l'équation 1.4 soit remplie.

$$\frac{1}{2} - \chi \propto 1 - \frac{\Theta}{T} \tag{1.4}$$

Selon le solvant utilisé, la température Θ change. Il existe alors des conditions particulières de solvant et de température pour lesquelles la différence de potentiel chimique donnée à l'équation 1.3 est nulle. Il s'agit du cas $\chi = \frac{1}{2}$ $(T = \Theta)$, qu'on appellera conditions Θ. Dans ces conditions-là, une unité monomérique interagira autant avec les autres unités qu'avec le solvant.

Les conditions Θ ont longuement été étudiées par Flory qui en a fait un état de référence. Il démontre qu'en théorie, pour $\chi = \frac{1}{2}$, le rayon de giration d'un polymère croît comme son nombre de monomères N élevé à la puissance $\frac{1}{2}$. Il définit ensuite α comme étant le rapport entre le rayon de giration d'un polymère pour un χ quelconque, R_g, et son rayon de giration à $\chi = \frac{1}{2}$, R_{g_0} :

$$R_g = \alpha R_{g_0} \propto \alpha N^{1/2} \tag{1.5}$$

Il démontre par ailleurs que α peut s'exprimer comme une fonction de $(\frac{1}{2} - \chi)$ et de N. On peut alors décrire R_g par l'équation 1.6 où ν, appelé exposant de Flory, dépend de χ et permet donc lui aussi de connaître le comportement d'un polymère en solution.

$$R_g \propto N^{\nu} \tag{1.6}$$

Pour $\alpha = 1$, l'équation 1.5 nous donne $\nu = \frac{1}{2}$, le polymère se trouve en conditions Θ ($\chi = \frac{1}{2}$). Une unité monomérique aura autant intérêt à interagir avec une autre unité qu'avec une molécule de solvant. Dans cet état, le polymère parcourt l'espace par une 'marche aléatoire' (*random walk*). Partant d'une extrémité du polymère, chaque pas représente l'unité monomérique suivante et peut se faire dans toutes les directions de

l'espace sans distinction, en respectant les contraintes stériques.

Pour $\alpha << 1$, on a $\alpha \propto -(\frac{1}{2} - \chi)N^{-\frac{1}{6}}$ d'où $\nu = \frac{1}{3}$ et $\chi > \frac{1}{2}$, on parle de 'polymère replié' (*collapsed polymer*). Le polymère préfère se replier sur lui-même pour minimiser le contact avec le solvant, les chaînes peuvent librement s'interpénétrer. C'est le cas d'un 'mauvais solvant' dans lequel le polymère se dissolvera mal voire précipitera.

Pour $\alpha >> 1$, on a $\alpha \propto (\frac{1}{2} - \chi)N^{\frac{1}{10}}$ d'où $\nu = \frac{3}{5}$ et $\chi < \frac{1}{2}$, le polymère est appelé 'à volume exclu'. Les unités monomériques préfèrent le contact avec le solvant plutôt qu'avec les autres unités, le polymère adopte alors une conformation très étendue pour maximiser le contact avec le solvant, alors appelé 'bon solvant'. Les chaînes de polymères ne peuvent pas s'interpénétrer, il y a donc un volume minimal de solvant autour de chaque molécule, appelé volume exclu.

Le changement de température et/ou de solvant modifie la valeur de χ et de ν et provoque le passage du polymère d'un état à un autre.

Comme précisé précédemment, on peut étudier les protéines de la même façon qu'un polymère. On peut donc leur appliquer la théorie de Flory en gardant en tête que cette théorie, avec ses hypothèses fortes et ses approximations, ne prend pas en compte les interactions fortes comme les interactions de Van der Walls, les interactions électrostatiques ou bien les liaisons hydrogène.

Dans le cas des valeurs de ν, on retrouve pour les protéines les mêmes propriétés qu'un polymère ainsi qu'un changement d'état possible par modification du solvant ou de température. Un cas bien connu est le passage d'une protéine globulaire (ν proche de $\frac{1}{3}$) à une protéine dénaturée (ν proche de $\frac{3}{5}$) par chauffage, refroidissement extrême ou ajout d'urée ou de chlorure de guanidinium. Ces phénomènes, bien que de nature différente, peuvent être décrits par l'étude de la dimension fractale.

$$d_f = \frac{1}{\nu} \tag{1.7}$$

molécule (solvant)	d_f
cylindre infini	1
polymère à volume exclu (bon solvant)	$\frac{5}{3} \approx 1.67$
polymère à marche aléatoire (conditions Θ)	2
sphère	3
protéines dénaturées	1.71
poly(oxide d'éthylène) ($CDCl_3$)	1.73
poly(oxide d'éthylène) (D_2O)	1.86
peptides β-amyloïdes	2.27
protéines globulaires	2.56

Tableau 1.1 – *Valeurs de dimensions fractales théoriques (partie supérieure) et recueillies dans la littérature (partie inférieure).*

La dimension fractale d_f dont nous nous servirons tout au long de notre étude est définie comme l'inverse de l'exposant de Flory (voir équation 1.7). Etudier la dimension fractale plutôt que l'exposant de Flory permet une appréhension plus aisée de l'objet observé. En effet, un cylindre infini a une dimension fractale de 1 alors qu'une sphère a une dimension fractale de 3. Pour ces raisons d'abstraction, manipuler une dimension fractale est souvent considéré comme plus clair.

On remarque par ailleurs que l'intervalle de dimensions fractales compris entre 1 et 1.67 ne fait pas partie de la théorie de Flory et n'est décrit que par la théorie des dimensions fractales.

Dans le tableau 1.1 et la figure 1.2 sont compilées des valeurs théoriques et expérimentales de d_f trouvées dans la littérature à température ambiante [24,25]. On voit à travers l'exemple du poly(oxyde d'éthylène) (PEO : poly(ethylene oxide)) que la valeur de d_f est bien dépendante du solvant. On peut aussi remarquer que les dimensions fractales des protéines ont déjà été étudiées. Les protéines dénaturées et les protéines globu-

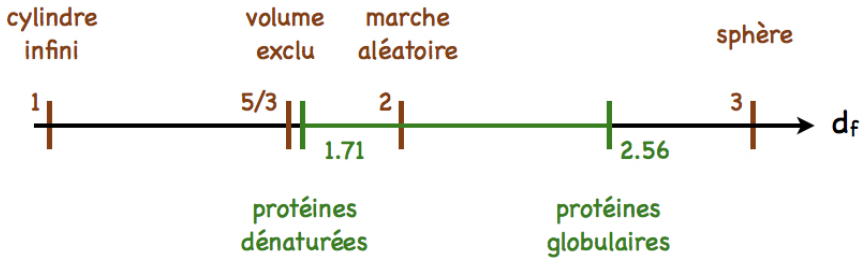

Figure 1.2 – *Axe des dimensions fractales avec, en marron, les valeurs théoriques extraites de la théorie de Flory ou des propriétés des dimensions fractales et, en vert, des valeurs trouvées dans la littérature pour les protéines.*

laires possèdent les valeurs extrêmes de d_f. On remarque d'ailleurs que les protéines dénaturées se comportent comme des polymères à volume exclu ($d_f = \frac{5}{3}$), avec une dimension fractale de 1.71, et non comme des polymères à marche aléatoire. Cela signifie qu'ils adoptent une conformation plus étendue qu'en conditions Θ afin de maximiser la surface d'interaction avec le solvant.

En partant d'un intervalle théorique de valeurs compris entre 1 (cylindre infini) et 3 (sphère), on se retrouve avec un intervalle expérimental compris entre 1.71 (protéines dénaturées) et 2.56 (protéines globulaires). Il paraît alors bien faible au vu de tous les phénomènes qu'il peut nous être amené d'observer. Cependant, la valeur de d_f pour les peptides β-amyloïdes semble déjà indiquer une certaine pertinence du paramètre. En effet, les peptides β-amyloïdes sont des IDP connus pour s'agréger sous forme de fibres très compactes en créant des feuillets β entre eux et ensuite précipiter. Cette prédominance à la précipitation ainsi que leur propension à adopter une conformation étendue se traduisent par une forte dimension fractale ($d_f = 2.27$).

Ainsi pour confirmer ou infirmer le choix de d_f comme paramètre pertinent des IDP, tout va résider dans la précision de notre technique de mesure ainsi que de l'interprétation possible des phénomènes étudiés.

1.1.2 Diffusion-Ordered SpectroscopY (DOSY)

La DOSY est une méthode de RMN permettant d'extraire le coefficient de diffusion translationnel de chaque composé présent dans un mélange. Le coefficient de diffusion translationnel est une mesure de l'ampleur du mouvement Brownien pour une molécule à l'intérieur d'une solution, il dépend de nombreux paramètres comme la température et la viscosité du milieu mais aussi de la masse et de la forme de la molécule.

Une expérience de DOSY est basée sur l'emploi de gradients de champ magnétique. Ces gradients s'ajoutent au champ magnétique constant B_0 toujours présent dans le spectromètre. Leur intensité dépend de la hauteur z sur l'axe colinéaire à B_0. L'équation 1.8 donne la valeur du champ magnétique total en z, $B(z)$, en fonction de z et de g, la force du gradient.

$$B(z) = B_0 + gz \tag{1.8}$$

Dans un plan perpendiculaire au champ magnétique, on sait que les spins tournent autour de l'axe de ce champ avec une certaine fréquence. Cette fréquence dépend de l'intensité du champ magnétique appliqué. Or dans notre cas cela signifie que la fréquence des spins dépend de la hauteur z de la molécule dans le tube.

Les séquences de type écho de spin avec gradients de champ magnétique pulsés (PFGSE : Pulse Field Gradient Spin Echo), dont la figure 1.3 est l'exemple le plus simple, utilisent dans un premier temps cette propriété pour encoder la hauteur des spins dans le tube en envoyant un gradient pendant un laps de temps δ. Le système est ensuite laissé libre d'évoluer pendant un temps $\frac{\Delta}{2}$. Une impulsion de 180° renverse le spin pour le placer à sa position symétrique puis le même temps d'évolution $\frac{\Delta}{2}$ est laissé. Un gradient est alors appliqué avec la même puissance et pendant le même temps δ que

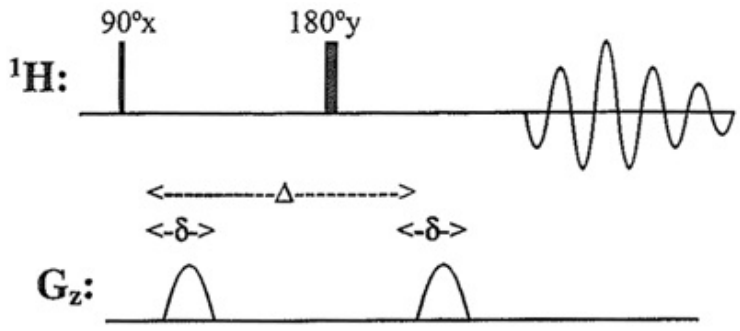

Figure 1.3 – *Séquence d'impulsion simple pour une expérience de DOSY. Les impulsions concernant les spins des protons sont indiquées en haut et les impulsions de gradient sont indiquées en bas.*

précédemment puis l'acquisition a lieu sur un des axes du repère tournant.

La séquence étant complètement symétrique, si le spin n'a pas changé de hauteur au cours de l'expérience, il sera également affecté par les deux gradients et le signal obtenu sera en théorie le même que si la séquence n'était composée que d'une impulsion de 90°. Par contre, si le spin change de position pendant l'expérience, il ne sera pas affecté de la même façon par les deux gradients, il ne reviendra pas à sa position initiale et on ne récupèrera donc pas la pleine intensité du signal.

$$S \propto e^{-D(\gamma g \delta)^2(\Delta - \delta/3))} \tag{1.9}$$

Cette décroissance du signal S, décrite par l'équation 1.9, dépend de paramètres expérimentaux (l'intensité du gradient g, le temps entre deux impulsions de gradient Δ et la durée du gradient δ) mais aussi du rapport gyromagnétique de l'atome observé γ et du coefficient de diffusion D.

Afin d'extraire D de façon la plus précise possible, nous avons suivi la méthode décrite par Stejskal et Tanner [26], à savoir faire varier l'un des paramètres expérimentaux dont dépend l'intensité du signal. Le paramètre que nous avons choisi est l'intensité du gradient.

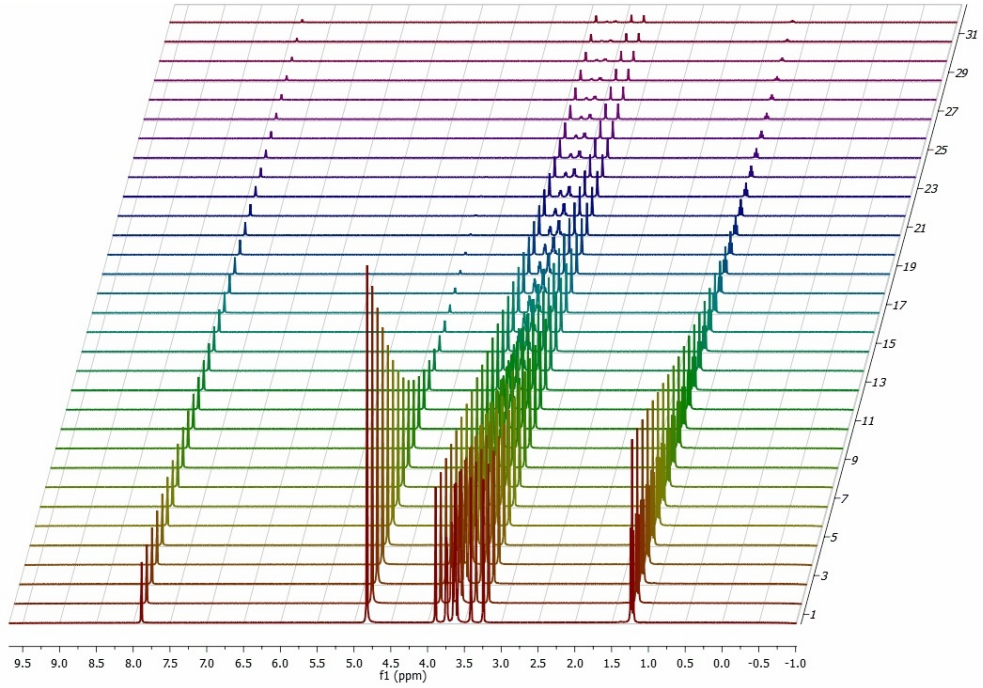

Figure 1.4 – *Exemple de résultat d'une expérience de DOSY. L'ensemble des spectres montre plusieurs décroissances différentes correspondant à toutes les molécules de l'échantillon pour lesquelles on peut observer le signal de RMN du proton.*

L'ensemble des spectres obtenu montre la décroissance des signaux pour tous les déplacements chimiques comme on peut le voir sur l'exemple de la figure 1.4. Grâce à la courbe obtenue à chaque déplacement chimique, il est possible de trouver une courbe gaussienne qui décrit au mieux la décroissance du signal et ainsi obtenir D. Malheureusement, cette méthode n'est pas assez précise car elle ne peut prendre en compte la décroissance que d'une seule molécule. Si au moins deux molécules ont un déplacement chimique en commun, il faut utiliser une autre méthode.

Pour cela, la transformée de Laplace inverse (ILT : Inverse Laplace Transform) résout notre problème. Elle permet d'extraire, pour chaque déplacement chimique, une distribution de coefficients de diffusion. Ainsi, lorsque plusieurs molécules ont des déplacements chimiques identiques, il est possible de discerner les différentes distributions qui leur correspondent.

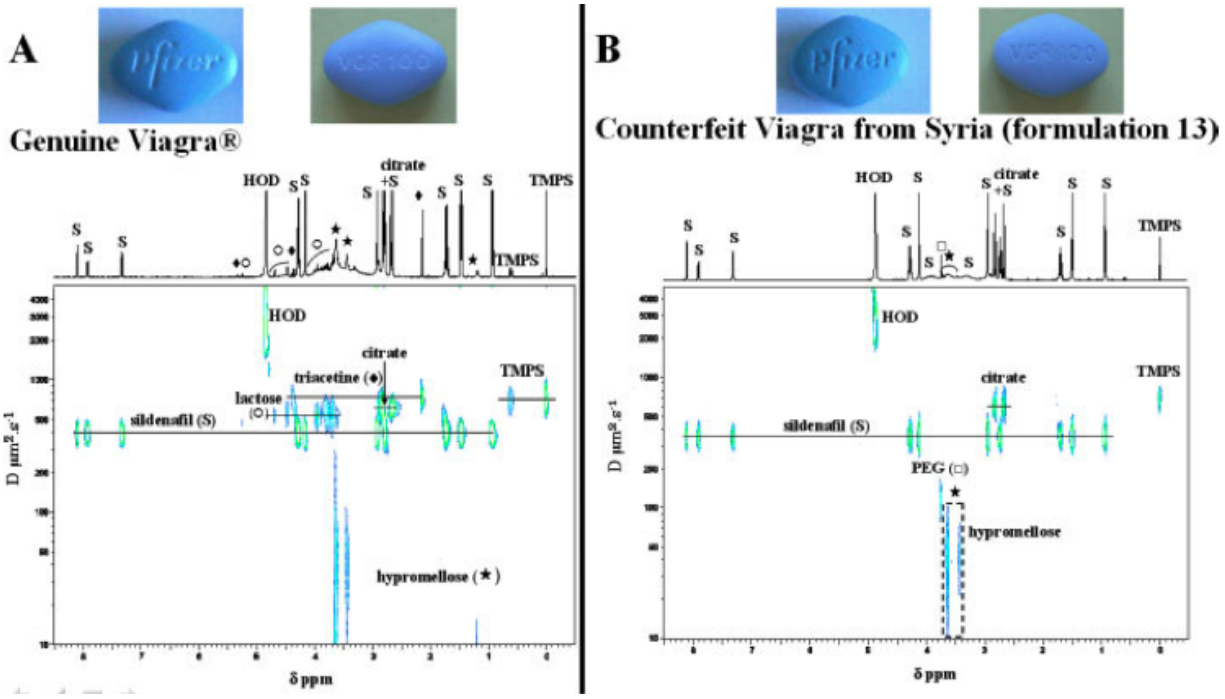

Figure 1.5 – *Comparaison de la composition de deux médicaments à travers leur spectres DOSY respectifs. Figure extraite de l'article de Trefi et al. [27].*

La figure 1.5 donne un exemple d'utilisation de spectres traités par l'ILT. Il s'agit ici de comparer la composition de deux pilules. Chaque molécule possède un seul coefficient de diffusion ce qui permet de la différencier des autres. Les auteurs de l'étude dont a été extraite la figure [27] montrent que la DOSY, par une seule expérience, permet de savoir si un médicament a été contrefait ou non. Par ailleurs, le traitement avec l'ILT permet de déterminer la polydispersité d'un échantillon en évaluant la largeur des distributions observées (voir Annexe A). On voit bien par ces exemples que l'un des intérêts majeurs de la DOSY est d'avoir une sorte de chromatographie de l'échantillon étudié sans pour autant séparer réellement ses composés et ce dans des temps relativement courts (une expérience de DOSY pouvant prendre près d'une heure au minimum).

1.1.3 Détermination de d_f par DOSY

Augé et al. [24] ont montré que la dimension fractale d_f d'une famille homogène de polymères ou de protéines pouvait être déterminée grâce à leurs coefficients de diffusion.

En effet, nous avons vu dans la partie 1.1.1 l'expression du rayon de giration d'un polymère R_g en fonction de son nombre d'unités monomèriques. En combinant les équations 1.5 et 1.7, on obtient l'équation suivante, qui relie donc R_g et la dimension fractale d_f :

$$R_g \propto N^{1/d_f} \tag{1.10}$$

D'un autre côté, on a remarqué empiriquement que le coefficient de diffusion D d'une molécule est relié à sa masse M (équation 1.11). Un exemple de cette relation est donné par la figure 1.6 dans le cas du PEO dans D_2O à température ambiante. De plus, D est aussi relié au rayon hydrodynamique de la molécule par l'équation de Stokes-Einstein (équation 1.12).

$$D \propto M^{-\mu} \tag{1.11}$$

$$D = \frac{kT}{6\pi\eta R_h} \tag{1.12}$$

Le rayon hydrodynamique R_h représente la valeur du rayon qu'aurait une sphère de même masse diffusant à la même vitesse que la molécule, η est la viscosité du solvant, k la constante de Boltzmann et T la température de l'échantillon.

R_h et R_g sont des paramètres différents mais très corrélés et les auteurs de l'étude précédemment citée avancent que le paramètre μ reliant D et M est le même paramètre

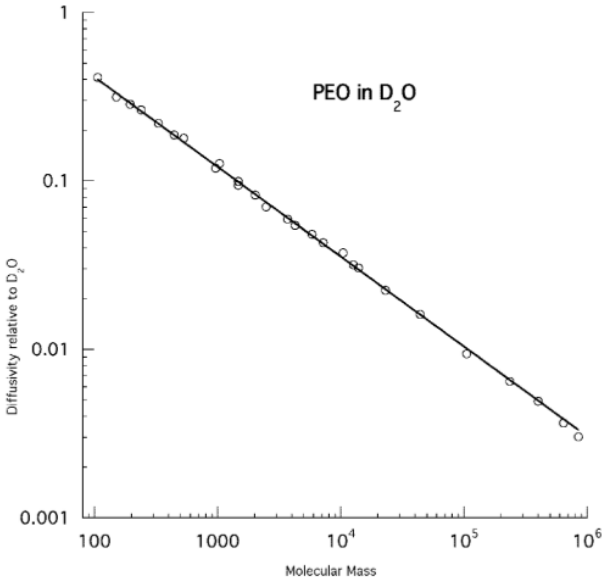

Figure 1.6 – *Graphique représentant la diffusivité mesurée par rapport à l'eau pour une molécule par DOSY en fonction de sa masse molaire. Cas d'une famille de PEO de différentes tailles dans D_2O. Figure extraite de l'article de Augé et al. [24]. Ici, la pente de la droite est de -0.54.*

que celui reliant R_g et N. On a alors l'équation suivante :

$$D \propto M^{-1/d_f} \tag{1.13}$$

Cette relation est valable au coeur de l'intervalle de valeurs de dimensions fractales mais se révèle incorrecte dans le cas de dimensions fractales faibles c'est-à-dire pour le cas des cylindres finis ou infinis. Cette possibilité n'est d'ailleurs pas prévue par la théorie de Flory. En effet, une conformation de type cylindrique imposerait que les liaisons chimiques du polymère soient parallèles les unes aux autres pour qu'il grandisse de façon linéaire, ce qui va à l'encontre-même de la disposition tétragonale des atomes de carbone.

Pour déterminer d_f, il suffit donc maintenant de prendre la pente de la courbe $ln(D) = f(ln(M))$, qui vaut $-\frac{1}{d_f}$. Grâce à la DOSY, nous pouvons déterminer le coef-

ficient de diffusion de chaque membre d'une famille de molécules. La figure 1.6 montre un exemple de détermination de la dimension fractale pour le PEO dans D_2O. Le paramètre appelé diffusivité, D_r, est le rapport du coefficient de diffusion de la molécule appartenant à la famille sur le coefficient de diffusion d'une molécule de référence, ici l'eau. En effet, on comprend bien par l'équation 1.12 que le coefficient de diffusion est très sensible à la température et à la viscosité du milieu. Cette dépendance n'existe pas pour le rapport de deux coefficients de diffusion mesurés lors de la même expérience. En gardant la même molécule de référence sur toute une famille et en utilisant la diffusivité plutôt que le coefficient de diffusion, on s'affranchit des problèmes d'inhomogénéité de température ou de viscosité dans l'échantillon. La pente de la courbe $ln(D_r) = f(ln(M))$ est toujours $-\frac{1}{d_f}$ mais la précision de la mesure de d_f augmente.

La méthode développée par Augé et al. pour la détermination de d_f consiste donc en une mesure de la diffusivité des membres d'une famille homogène de molécules (polymères de différentes tailles, protéines ayant les mêmes propriétés et/ou des séquences comparables...) par DOSY et prenant soin de les mesurer avec la même molécule de référence et à la même température. La pente de la courbe $ln(D) = f(ln(M))$ permet alors d'extraire d_f avec une grande précision. C'est cette méthode que nous utiliserons tout au long de cette étude.

1.2 La proline, un acide aminé à part

Des particularités...

Si tous les acides-aminés peuvent potentiellement être retrouvés dans les IDP, la proline fait partie des acides-aminés qu'on retrouve un peu plus souvent dans les IDP que dans les protéines structurées. Elle cumule à elle seule un nombre conséquent de singularités qui intriguent.

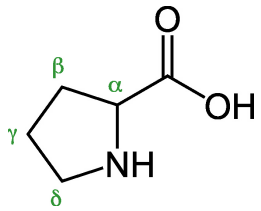

Figure 1.7 – *Représentation de la L-proline avec la nomenclature des carbones du cycle (en vert).*

Elle est par exemple le seul acide-aminé naturel dont la chaîne latérale est reliée à son groupement amine (voir figure 1.7). La première conséquence de cette cyclisation est la disparition de tout hydrogène amide lors de la synthèse peptidique. La proline, dans une protéine, ne peut donc pas jouer le rôle de donneur de proton dans le cadre d'une liaison hydrogène. Elle apparaît donc assez peu dans les éléments de structure secondaire car elle ne peut pas pleinement participer à leur stabilisation. La deuxième conséquence est la création d'une forte contrainte stérique. En effet, contrairement aux acides aminés possédant des cycles (l'histidine, la phenylalanine, la tyrosine et le tryptophane), le cycle de la proline est directement sur la chaîne principale d'une protéine. Ceci réduit l'espace conformationnel de la proline elle-même mais aussi sur l'acide-aminé qui la précède sur la chaîne peptidique. L'ensemble des conformations adoptées par ces acides-aminés est décrit en détails au chapitre 2.

Une autre particularité de la proline est sa capacité à adopter une conformation cis. En effet, l'angle dihèdre ω formé autour de la liaison peptidique ne peut prendre que deux valeurs : $0°$ (conformation cis) ou $180°$ (conformation trans). La grande majorité des acides-aminés gardent constamment une conformation trans, seule la proline peut être trouvée à l'état naturel dans les deux conformations.

... et une structure caractéristique.

Malgré ces extravagances, la proline peut s'insérer dans un motif structural répétitif, l'hélice polyproline. Il existe deux formes d'hélices polyproline, elle sont représentées

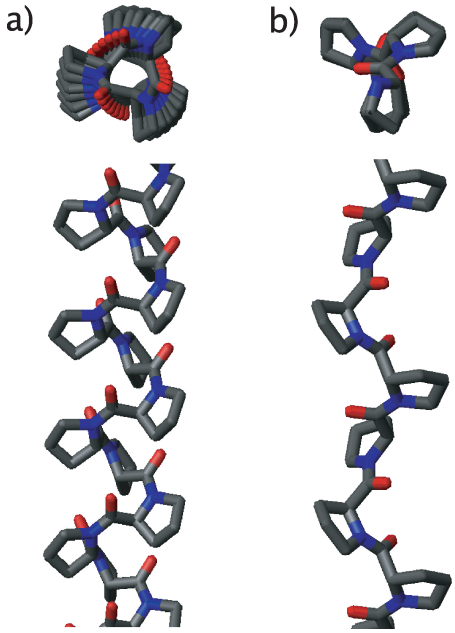

Figure 1.8 – *Vue de dessus (en haut) et vue de côté (en bas) des hélices polyprolines. a) type I, b) type II.*

en figure 1.8. L'hélice polyproline-I (PPI) est composée exclusivement de prolines en conformation cis. Il s'agit d'une hélice de symétrie droite comptant 3.3 résidus par tour et grandissant de 1.90Å par proline. Cette forme est beaucoup plus stable dans un alcool que dans l'eau [28], voilà pourquoi elle n'est pas présente dans les protéines.

L'hélice polyproline-II (PPII), quant à elle, est une hélice de symétrie gauche comptant 3 résidus par tour et grandissant de 3.2Å par acide-aminé. On se rend bien compte de la nature très étendue de cette hélice sur la figure 1.9 où sont représentées des structures idéales de décaglycines. L'hélice PPII, malgré son nom, peut aussi être composée d'acides-aminés différents d'une proline. Quelques études se sont d'ailleurs penchées sur l'effet de tels acides aminés sur la stabilité de cette hélice en les insérant dans des peptides riches en proline [30–32]. Dans le chapitre 2, nous déterminerons la propension de chaque acide-aminé à se trouver dans la conformation majoritaire des hélices PPII.

Evidemment, aucune de ces deux hélices n'est stabilisée par des liaisons hydrogène.

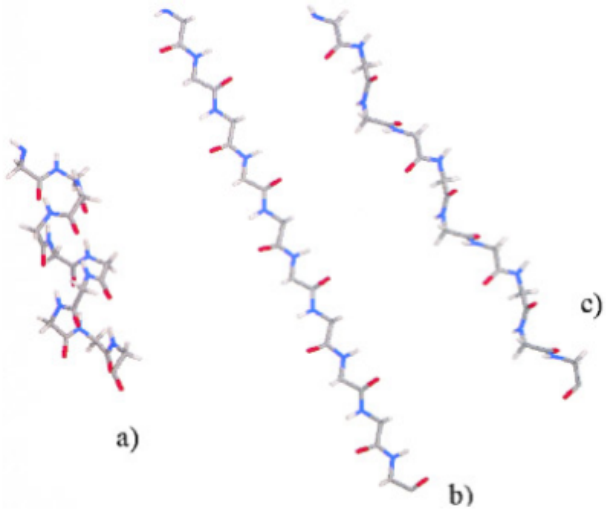

Figure 1.9 – *Exemples de polyglycines idéales : a) en hélice α, b) en brin β, c) en PPII. Figure extraite de l'article de Bochicchio et Tamburo [29].*

Si aucune explication complète sur l'existence de ces hélices n'est disponible dans la littérature, plusieurs conjectures ont été formulées dans lesquelles l'eau joue un rôle important [29]. Cela expliquerait par ailleurs pourquoi les peptides polyprolines sont solubles en toutes proportions dans l'eau alors qu'il s'agit majoritairement d'une chaîne carbonée et que, par conséquent, ces peptides devraient être fortement hydrophobes.

Quelles que soient les causes de l'existence de l'hélice PPII, elle se révèle être un élément essentiel de la transmission de signaux à l'intérieur de la cellule. En effet, un grand nombre de protéines reconnaissent ce motif et interagissent avec [32, 33]. C'est le cas notamment des domaines SH3 (Src Homology-3) [34]. Nous traiterons au chapitre 3 l'exemple de l'interaction entre le domaine A/B, riche en prolines, du récepteur nucléaire à l'acide rétinoïque γ (RARγ : Retinoic Acid Receptor γ) et l'un des domaines SH3 de la vinexine β. Cette interaction provoque la répression de la transcription de certains gènes [35].

Figure 1.10 – *Hélices PPII composées de 20 prolines chacune. A gauche, toutes les liaisons peptidiques sont trans, la distance entre les carbones α des résidus 1 et 20 est de 60.8Å. A droite, la liaison entre P9 et P10 est cis, la distance entre les carbones α des résidus 1 et 20 est maintenant de 35.9Å.*

Affiner le modèle de l'hélice PPII

En ce qui concerne la taille de l'hélice PPII, deux écoles divisent les utilisateurs de FRET (Förster Resonance Energy Transfer). Le FRET permet de savoir si deux molécules fluorescentes appelées chromophores, sont proches l'une de l'autre. La molécule A doit impérativement pouvoir absorber l'énergie libérée par la fluorescence de la molécule B pour que cette méthode fonctionne. Si les chromophores sont assez proches l'un de l'autre, B transfère son énergie à A qui peut alors fluorescer. Pour calibrer les chromophores et connaître la distance minimale nécessaire à l'obtention de transfert de fluorescence, les fluorospectroscopistes ont utilisé des polyprolines de différentes tailles. Connaissant la longueur d'une hélice par le modèle décrit précédemment et se basant sur sa grande rigidité supposée, ils l'ont pendant longtemps considérée comme une règle [36].

Cependant, des mesures complémentaires tendraient à montrer que la distance bout-à-bout attendue n'est pas la valeur réelle. Une explication de cette différence est la remise en cause de la rigidité de l'hélice. Elle aurait en réalité une tendance à se courber légèrement [37]. L'autre explication, la plus courante, est la prise en compte de l'isomérisation cis-trans. Entre 5 et 10% de prolines se trouveraient en conformation cis

SRSARSPPGKPQGPPQQEGNKPQGPPPPGKPQGPPPAGGNPQQPQ
PPAGKPQGPPPPPQGGRPPRPAQGQQPPQ

Figure 1.11 – *Séquence de la protéine IB-5. En rouge, le motif répétitif.*

naturellement dans les hélices PPII [38,39]. La figure 1.10 montre l'effet d'une liaison cis sur la distance bout-à-bout d'un peptide polyproline. On comprend alors que si toutes les liaisons sont potentiellement en cis, la distribution de distances bout-à-bout est plus large que prévue et la distance moyenne observée est plus faible que celle attendue pour un modèle d'hélice rigide.

Un modèle biologique : IB-5

L'étude que nous avons menée sur la dimension fractale cherche à confirmer ou infirmer certains modèles avancés pour l'hélice PPII mais plusieurs questions découlent de ces modèles : les protéines riches en proline forment-elle aussi des hélices PPII et, si non, peut-on discerner les différents comportements en solution ?

Pour répondre à ces questions, nous avons utilisé comme modèle biologique la protéine IB-5 dont la séquence est donnée par la figure 1.11. Elle possède 40% de prolines et sa complexité de séquence est très faible. IB-5 est une protéine salivaire complètement désordonnée qui se lie aux polyphénols. Elle est l'une des causes de l'astringence des boissons ou des fruits fortement chargés en tannins. En effet, elle se lie aux tannins et précipite, laissant une sensation de sécheresse en bouche.

Une étude d'IB-5 a été réalisée par CD (Circular Dichroism) et RMN [40]. Cette étude montre que IB-5 se structure peu voire pas du tout lorsqu'elle lie ses partenaires. Cependant, son spectre CD ressemble beaucoup au spectre d'une hélice PPII tout en suggérant un niveau de désordre supérieur. Malheureusement, en CD, le spectre d'une hélice PPII et celui d'une protéine *random-coil* sont très similaires, il donc difficile de les discriminer.

Il est intéressant pour nous de voir si la dimension fractale permet de repérer ces deux comportements plus précisément que le CD.

1.3 Dimension fractale de l'hélice PPII

La méthode décrite dans la partie 1.1.3 a été appliquée à une famille de peptides polyprolines de différentes tailles.

Dans un premier temps, on a mesuré la diffusivité des peptides polyprolines par rapport au Tris (trishydroxyméthylaminométhane) qui tient lieu de molécule de référence. Les peptides choisis pour l'étude sont P_3, P_6, P_6C, $P_{11}C$, $P_{14}C$ et $P_{20}C$. Ces peptides ont été synthétisés dans un service commun de l'institut et la cystéine terminale a du être ajoutée pour augmenter le rendement des synthèses peptidiques.

Les expériences de DOSY ont été réalisées à 300K sur un spectromètre RMN Bruker Avance 600MHz équipé d'une cryo-sonde. La concentration de peptide dans chacun des cas est de 1mM ainsi que la concentration de Tris. Deux équivalents de DTT (dithiothréitol) ont systématiquement été ajoutés dans les solutions afin d'éviter la formation de ponts disulfure entre les cystéines et les mesures ont été réalisées sur des solutions fraîches (de quelques minutes à deux heures). La séquence d'impulsions choisie est basée sur des gradients bipolaires. La suppression du solvant est réalisée par *excitation sculpting*. La durée de chaque gradient est de 2.8ms, le temps entre deux gradients est de 150ms et le temps de *led* est de 5ms. 40 points ont été enregistrés en 32 scans soit au total 45 minutes d'expérience. Les spectres ont été traités grâce au logiciel NMRNotebook.

La figure 1.12 donne la courbe $ln(D_r) = f(ln(M))$ pour ces six peptides. Pour calculer la dimension fractale, nous nous sommes servis des quatre derniers points de la courbe et on trouve $d_f = 1.2$. Il s'agit d'une valeur très faible qui se rapproche

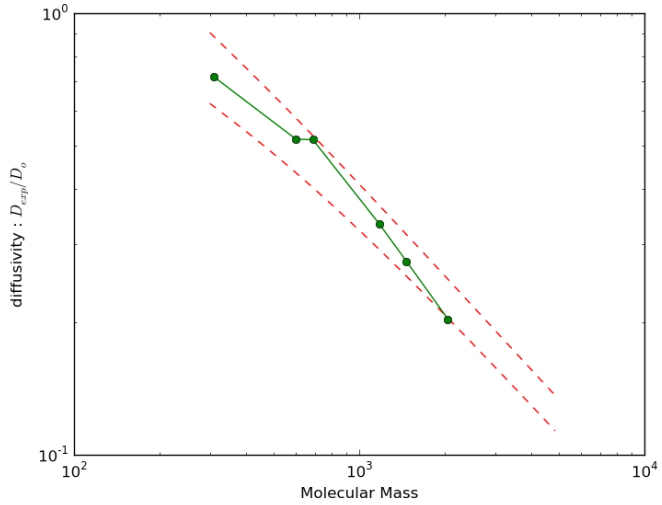

Figure 1.12 – *En vert, la courbe de la diffusivité des six peptides polyproline décrits dans le texte en fonction de leur masse moléculaire. En rouge, les courbes théoriques d'après Ortega et García de la Torre [41] pour une extension de 3.2Å par résidu et un diamètre de 4.6Å (en haut) ou 8.0Å (en bas).*

énormément de la dimension fractale d'un cylindre infini ($d_f = 1$). Le modèle de cylindre rigide semble donc convenir.

Modèle du cylindre et oligomérisation

Pour confirmer la longueur de l'hélice, nous nous sommes intéressés au modèle de Ortega et García de la Torre qui permet de calculer le coefficient de diffusion d'un cylindre à partir de son diamètre d et de sa longueur l [41] (voir équation 1.14).

$$D = \frac{kT}{6\pi\eta(\frac{3l}{16d^2})^{\frac{1}{3}}(1.009 + 0.01395\ ln(\frac{l}{d}) + 0.0788\ ln(\frac{l}{d})^2)} \tag{1.14}$$

La température T choisie pour les calculs est celle des expériences, 300K, et la viscosité de l'eau η a été placée à 1 centipoise (soit 10^{-3} Pa.s). Nous avons voulu savoir quel diamètre aurait un cylindre diffusant à la même vitesse que les peptides étudiés pour une extension de l'hélice e égale à la valeur théorique de 3.2Å. Pour calculer la diffusi-

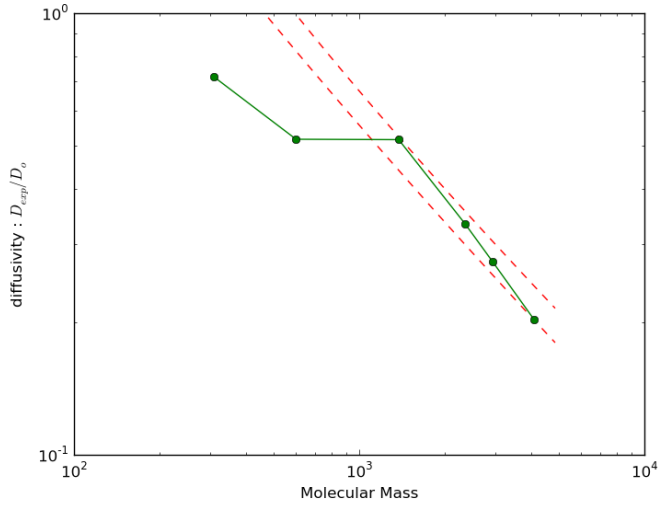

Figure 1.13 – *En vert, la courbe de la diffusivité des six peptides polyproline dimérisés (sauf P_3 et P_6) en fonction de leur masse moléculaire. En rouge, les courbes théoriques d'après Ortega et García de la Torre [41] pour une extension de 3.2Å par résidu et un diamètre de 1.1Å (en haut) ou 1.9Å (en bas).*

vité de chaque peptide, le résultat de l'équation a été divisée par les références de Tris mesurées par DOSY.

Selon le peptide observé, le diamètre trouvé est différent, il varie dans un intervalle entre 4.6Å et 8.0Å (voir figure 1.12). La plus grande distance mesurable dans le cycle de la proline est proche de 4Å, ces valeurs sont donc compatible avec un cylindre équivalent à une hélice PPII. En faisant varier e, on ne peut toujours pas retrouver l'ensemble des coefficients de diffusion car la pente de la courbe reste identique.

On ne peut alors pas écarter l'hypothèse que, malgré l'ajout de DTT, une dimérisation a eu lieu. Si on multiplie la masse des peptides possédant une cystéine terminale par deux tout en gardant les autres paramètres intacts, on obtient la même dimension fractale. Cela est logique dans la mesure où les propriétés de croissance et de symétrie restent les mêmes dans les deux cas. On peut donc quoi qu'il arrive faire confiance à la dimension fractale. Par contre, si l'ensemble des valeurs de diffusivité ne peut toujours

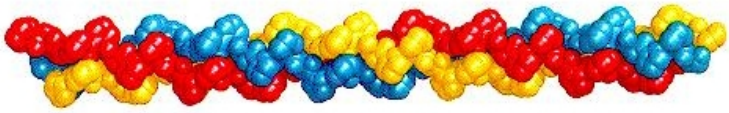

Figure 1.14 – *Structure d'une hélice de collagène. Trois chaines en conformation majoritairement PPII s'associent en une super-hélice.*

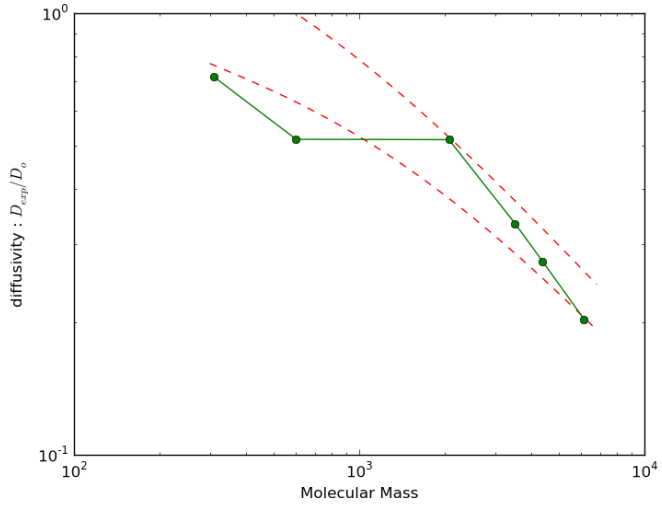

Figure 1.15 – *En vert, la courbe de la diffusivité des six peptides polyproline trimérisés (sauf P_3 et P_6, jugés trop petits) en fonction de leur masse moléculaire. En rouge, les courbes théoriques d'après Ortega et García de la Torre [41] pour une extension de 2.9Å par résidu et par chaîne et un diamètre de 4.5Å (en haut) ou 8.8Å (en bas)*

pas être prédit en une seule fois, les valeurs possibles pour le rayon d'un cylindre d'extension 3.2Å par résidu deviennent extrêmement faible : entre 1.1 et 1.9Å (voir figure 1.13).

Une autre hypothèse possible est le regroupement de trois peptides en hélice PPII pour former une super-hélice comme dans le collagène (voir figure 1.14). L'association en super-hélice diminue l'extension de chaque peptide à environ 2.9Å par résidu [42]. La figure 1.15 montre le résultat pour des objets observés trois fois plus lourds (sauf pour P_3 et P_6 jugés trop petits), les valeurs trouvées pour le diamètre sont comprises entre 4.5Å et 8.8Å qui paraissent faibles pour un tel assemblage. L'ensemble des valeurs

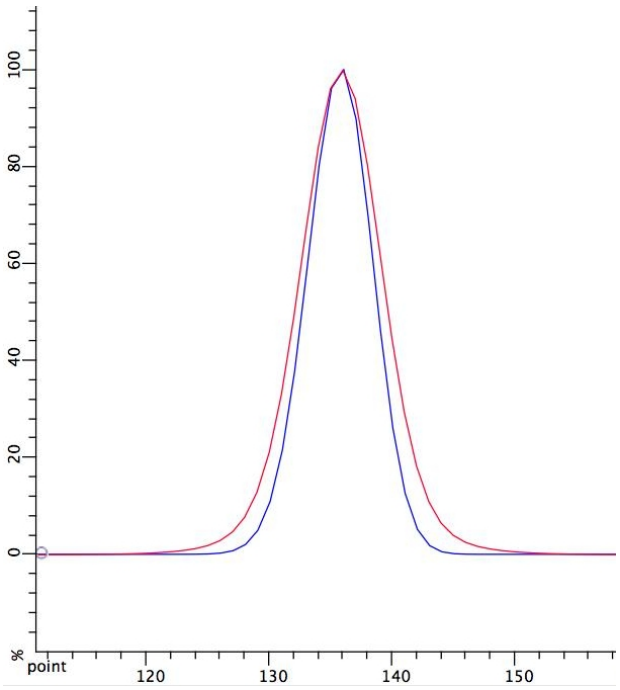

Figure 1.16 – *Distributions de coefficients de diffusion extraites de la DOSY de P_3 (en bleu) et de $P_{20}C$ (en rouge) au même déplacement chimique. Les distributions ont été artificiellement translatées pour que leurs maxima coïncident.*

ne peut toujours pas être prédit par une seule valeur de rayon mais la dimension fractale reste la même.

Avec ou sans oligomerisation (dimérisation, trimérisation ou plus), le modèle cylindrique ne permet pas d'expliquer le comportement de toute la famille de peptides. Il ne semble donc pas adapté pour la polyproline. Cependant, on voit bien que la dimension fractale, même très proche de la valeur de celle du cylindre infini, décrit un autre comportement que celui du cylindre.

De plus, l'oligomérisation a pour conséquence un élargissement de la distribution des coefficients de diffusion. La figure 1.16 donne ces distributions extraites au même déplacement chimique pour P_3 et pour $P_{20}C$ puis artificiellement translatées pour que leurs maxima coïncident. On voit bien que l'élargissement des distributions n'est pas significatif et ne permet pas de justifier l'hypothèse d'oligomérisation. Pour la suite

du chapitre, nous considérerons donc uniquement le cas dans lequel il n'y a pas d'oligomérisation.

À la recherche d'un nouveau modèle

On cherche maintenant à prévoir les diffusivités en dehors de tout modèle géométrique. Dans la littérature récente, on trouve l'équation 1.15 qui donne le rayon hydrodynamique d'une IDP R_h en fonction de son nombre d'acides-aminés N, de la proportion de prolines P, de sa charge nette Q [43]. Dans notre cas, l'équation peut être réduite car Q vaut toujours 0.

$$\begin{aligned} R_h &= 2.49(1.24P + 0.904)(0.00759Q + 0.963)N^{0.509} \\ &= 2.40(1.24P + 0.904)N^{0.509} \end{aligned} \tag{1.15}$$

Pour comparer les valeurs expérimentales avec les résultats de l'équation précédente, nous avons calculé la diffusivité en faisant le rapport du rayon hydrodynamique du Tris sur le rayon hydrodynamique du peptide (voir équation 1.16).

$$D_r = \frac{D}{D_{Tris}} = \frac{kT}{6\pi\eta R_h}\frac{6\pi\eta R_{h_{Tris}}}{kT} = \frac{R_{h_{Tris}}}{R_h} \tag{1.16}$$

La figure 1.17 montre la comparaison entre l'expérience et les valeurs trouvées avec l'équation précédente. Sur le graphique de gauche, l'équation a été utilisée directement avec une forte correction pour les prolines, P allant de 0.86 à 1.0. On remarque que les valeurs de diffusivité sont quasiment deux fois trop petites mais reproduisent le même schéma que les valeurs expérimentales avec une cassure entre P_6 et P_6C prouvant qu'il y a bien un effet de composition du peptide sur la diffusivité.

En enlevant dans l'équation 1.15 la correction due aux prolines, c'est-à-dire en met-

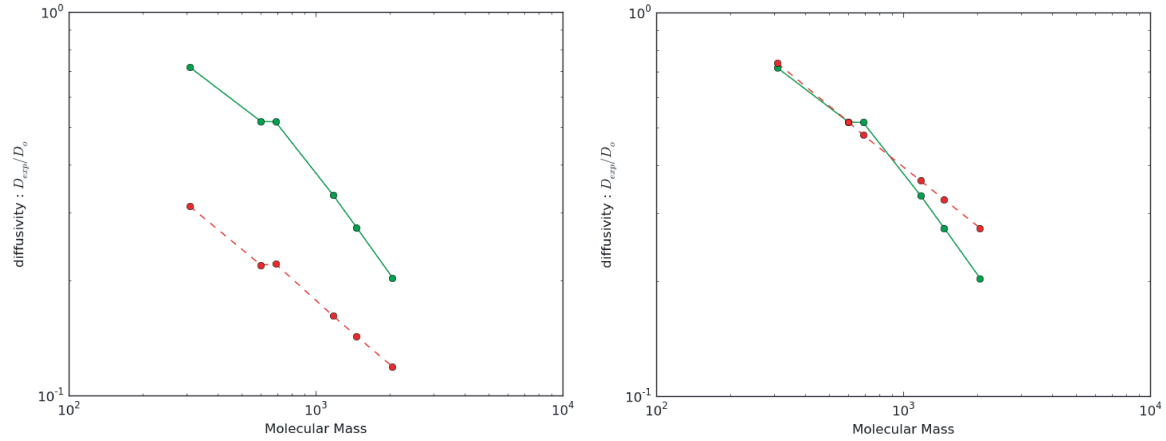

Figure 1.17 – *Comparaison entre les valeurs expérimentales de diffusivité pour les peptides polyproline (en vert) et les valeurs calculées à partir de l'équation 1.15 (en rouge, en pointillés). À gauche, avec la correction pour les prolines ; à droite, sans la correction ($P = 0$).*

tant P à 0 constamment, on obtient le graphique de droite sur la figure 1.17. Les valeurs obtenues pour la diffusivité sont cette fois-ci du même ordre de grandeur, celles de P_3 et P_6 sont même en bon accord avec l'expérience. Cependant, les peptides les plus longs sont mal prédits car la pente de la courbe est toujours de -0.509 d'après l'équation 1.15. Les auteurs considèrent donc que toutes les IDP ont une même dimension fractale à environ 1.96, ce qui est incorrect d'après des études précédentes (voir tableau 1.1). De plus, l'ensemble des protéines utilisé par les auteurs de l'équation ne semble pas contenir d'hélices polyproline donc ce motif n'a pas été pris en compte dans leur étude.

Là encore, et malgré les nombreux paramètres considérés par l'équation 1.15, la dimension fractale semble être l'un des rares paramètres pertinents à prendre en compte pour décrire les peptides.

Relation entre rayon de giration et distance bout-à-bout

Au lieu de remettre en cause le modèle utilisé pour la polyproline, on peut remettre en cause la théorie de la dimension fractale mise en place. La valeur de dimension frac-

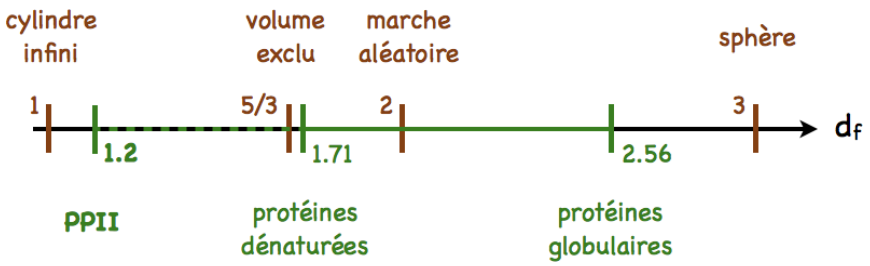

Figure 1.18 – *Axe des dimensions fractales avec, en marron, les valeurs théoriques extraites de la théorie de Flory ou des propriétés des dimensions fractales et, en vert, la valeur mesurée pour l'hélice PPII et des valeurs trouvées dans la littérature pour les protéines.*

tale trouvée pour la polyproline se trouve en dehors de l'intervalle défini pour la théorie de Flory et la loi de puissance reliant R_g et N (voir équation 1.6) ne peut pas s'étendre à des valeurs de ν supérieures à $\frac{3}{5}$. On a donc besoin d'un nouveau modèle mathématique expliquant qu'il existe des dimensions fractales proches de 1.

Avant de relier R_g et N, Flory a tout d'abord constaté le même type de relation entre la racine carrée de la moyenne des carrés des distances bout-à-bout d'un polymère $\sqrt{\overline{l^2}}$ et son nombre de résidus. Par exemple, il semble évident que dans le cas d'un polymère qui grandit comme un cylindre, la distance bout-à-bout moyenne $\overline{l}$ est directement proportionnelle à N. Il en est de même pour $\sqrt{\overline{l^2}}$. De plus, pour un polymère dans les conditions Θ, Flory démontre que $\sqrt{\overline{l^2}}$ est proportionnel à $N^{\frac{1}{2}}$. Il est donc possible d'établir une loi de puissance entre $\sqrt{\overline{l^2}}$ et N du même type que la relation entre R_g et N. Cette loi est applicable sur l'intervalle manquant à la théorie présentée plus haut et on peut en déduire une dimension fractale sur l'intervalle 1.0-2.0.

A priori, la dimension fractale extraite de R_g ou $\sqrt{\overline{l^2}}$ devrait être la même. En effet, il est possible relier R_g et $\sqrt{\overline{l^2}}$ sur l'intervalle de dimension fractale qui nous intéresse. Pour un cylindre de diamètre d,

$$R_g = \sqrt{\frac{\overline{l^2}}{12} + \frac{d^2}{8}}$$
$$= \sqrt{\frac{\overline{l^2}}{12}} \propto \sqrt{\overline{l^2}} \quad si\ le\ cylindre\ est\ infini \tag{1.17}$$

De même, pour un polymère en conditions Θ, Flory démontra que $R_g \propto \sqrt{\overline{l^2}}$.

Deux méthodes de calcul théorique de dimensions fractales ont été mises en place pour comparer les résultats à partir d'un fichier PDB décrivant une polyproline de 20 résidus (P_20) placée dans une conformation d'hélice PPII parfaite ((φ,ψ) = (-75°, 150°)). Ce fichier a été créé avec la bibliothèque python du logiciel de modélisation moléculaire MMTK [44]. La première méthode consiste à mesurer itérativement la distance bout-à-bout d'une partie du peptide. Le processus suit l'ordre des acides-aminés dans la séquence et un acide-aminé est ajouté à chaque itération. La deuxième méthode consiste à calculer le rayon de giration d'une partie du peptide. Le processus commence à l'acide aminé le plus proche du centre de masse et suit l'ordre des acides-aminés dans la séquence, un acide-aminé est ajouté à chaque itération, alternativement à un bout puis à l'autre.

Dans les deux cas, en traçant le graphique des valeurs obtenues en fonction du nombre de résidus dans chaque sous-partie, on peut extraire la dimension fractale de la régression linéaire en log-log. La figure 1.19 montre le résultat dans les deux cas et la dimension fractale extraite diffère. Elle est de 1.02 pour les distances bout-à-bout (d_{f_e}) et de 1.24 pour les rayons de giration (d_{f_r}). On observe par ailleurs que la méthode des R_g n'a pas encore convergé et que donc la régression linéaire ne passe pas par tous les points du graphique.

La valeur trouvée par $\sqrt{\overline{l^2}}$ est celle correspondant le mieux à la théorie des dimensions fractales car, P_20 n'est pas infini mais sa longueur est très grande par rapport à son

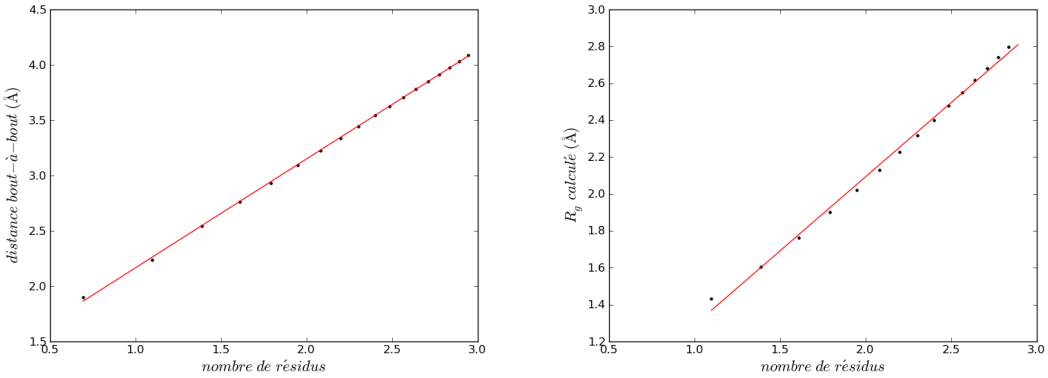

Figure 1.19 – *À gauche, la méthode de calcul par les distances bout-à-bout ($d_{f_e} = 1.02$). À droite, la méthode de calcul par les rayons de giration ($d_{f_r} = 1.24$)*

diamètre, on peut donc penser qu'un nombre restreint d'acides-aminés suffit pour que la méthode converge. Cependant, la méthode des R_g nécessiterait certainement de très grandes protéines pour atteindre la même valeur. La relation entre R_g et $\sqrt{\overline{l^2}}$ démontrée par Flory n'est pas remise en cause par cette expérience mais on peut penser que d_{f_e} décrit mieux les petites molécules que d_{f_r}. Pour de très grandes protéines, les valeurs trouvées par les deux méthodes devraient être les mêmes.

En ce qui concerne l'équation donnée par Ortega, le problème de convergence reste le même. On voit bien sur les figures précédentes (1.12, 1.13 et 1.15) et sur l'équation 1.14 que la pente de la courbe tend vers une valeur proche de 1.0 (soit une dimension fractale de 1.0). À grande échelle, ce modèle est donc bien capable *a priori* de décrire les polyprolines, ce qui pose problème ce sont donc les petites molécules qui gardent la même symétrie cylindrique rigide.

La valeur trouvée pour d_f, bien que très proche du modèle du cylindre, surprend. Outre le fait qu'elle décrit un comportement inexpliqué pour le moment, elle sort complètement de l'intervalle de dimension fractale délimité auparavant par les protéines

dénaturées et les protéines globulaires (comme on peut le voir clairement sur la figure 1.18). De plus, elle sort aussi de toutes les considérations théoriques de Flory. Cependant, Flory ne prend pas en compte les interactions fortes. Or, dans notre cas, il est possible que des interactions de Van der Walls ait lieu entre les cycles de prolines pour stabiliser l'hélice. De plus, l'isomérisation cis-trans des prolines ne peut être prise en compte dans aucun modèle développé jusqu'à présent or il s'agit d'un paramètre important qu'il faut étudier plus en détails.

Grâce aux hélices PPII, on élargit l'éventail des possibilités en termes de dimension fractale. On sait maintenant qu'elles peuvent s'étaler de 1.2 à 2.56 soit plus des deux tiers de l'intervalle théorique total. Ce surplus de valeurs nous conforte dans le choix de la dimension fractale comme paramètre pertinent. Reste à savoir si des comportements différents mais très proches ont des dimensions fractales différentes. Pour cela, nous avons étudié la protéine IB-5 qui est très riche en prolines pour savoir si elle avait un comportement d'hélice PPII.

1.3.1 Dimension fractale de IB-5

On souhaite utiliser la même technique que précédemment pour déterminer la dimension fractale de IB-5. L'inconvénient de notre méthode de mesure de la dimension fractale est qu'il nous faut une famille homogène de plusieurs peptides ou protéines de tailles différentes mais l'avantage de IB-5 est que sa séquence est très répétitive. Nous avons donc pu extraire plusieurs peptides dont les séquences sont données en figure 1.20.

Les expériences de DOSY ont été menées dans les mêmes conditions que pour la série des polyprolines à savoir à 300K sur un spectromètre Bruker Avance 600MHz avec cryo-sonde, pour une concentration de peptides de 1mM et une concentration de Tris de 1mM. Pour plus de détails, se référer au paragraphe 1.3.

```
PPP
KPQGPP
KPQGPPQQEGN
QQEGNKPQGPPPPGKPQG
KPQGPPQQEGNKPQGPPP
```

Figure 1.20 – *Séquences des peptides utilisés en plus d'IB-5 pour l'étude de la dimension fractale.*

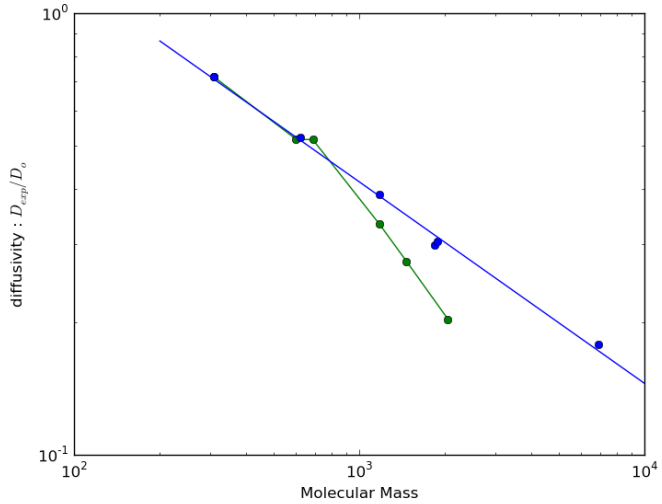

Figure 1.21 – *En vert, la courbe de la diffusivité des six peptides polyproline décrits dans le texte en fonction de leur masse moléculaire. En bleu, les valeurs de diffusivité de la famille composée d'IB-5 et de peptides issus de sa séquence (points) et la droite qui a servi à extraire la dimension fractale d'IB-5.*

La mesure de leur diffusivité et de celle d'IB-5 nous permet de tracer le graphique de la figure 1.21. On remarque que le comportement de l'hélice polyproline est différent du comportement des peptides issus d'IB-5. La dimension fractale d'IB-5 est d'environ 2.2.

Même si elle possède près de 40% de prolines, IB-5 ne se comporte pas comme une hélice PPII en solution. Sa dimension fractale se rapproche plus des polymères à marche aléatoire ($d_f = 2.0$) qui sont très solubles ou même des peptides β-amyloïdes ($d_f = 2.27$) qui ont une forte tendance à l'agrégation et à la précipitation. Or, pour IB-5, on retrouve, par DOSY et DLS (Dynamic Light Scattering), un rayon de giration plus petit que pour

une marche aléatoire et une forte propension à la précipitation à partir d'une certaine concentration ou en présence de polyphénols [40].

Dans ce cas précis, on peut retrouver d'importantes caractéristiques d'IB-5 à partir des connaissances accumulées dans la littérature, en la rapprochant de protéines ayant la même dimension fractale. On remarque aussi qu'un écart de 0.2 en dimension fractale (ici entre un polymère à marche alétoire et IB-5) peut suffire à décrire deux comportements différents. La dimension fractale confirme ainsi sa pertinence.

Il faudrait maintenant être capable de mesurer la dimension fractale d'une grande bibliothèque de peptides et de protéines pour être capable d'extraire des intervalles dans lesquels les protéines ont toutes les mêmes propriétés. L'étude des IDP est complexe donc pouvoir relier les protéines entre elles et en déduire leurs propriétés permettrait d'ouvrir de nouvelles pistes de recherche.

Chapitre 2

RamaDA et RamaDP

La biologie structurale est basée, comme son nom l'indique, sur l'étude de la structure des protéines et des acides nucléiques. Elle ne semble donc pas adaptée à l'étude des IDP. Cependant, il est nécessaire de dissocier les notions de conformation et d'élément de structure secondaire pour comprendre que la biologie structurale a sa place dans leur analyse. C'est ce que nous montrerons dans ce chapitre en décrivant la conception de deux outils utilisables tant pour les IDP que pour les protéines structurées.

Ces outils seront utilisés tout au long de cette étude et servirons dans un premier temps à avoir une meilleure compréhension du comportement de la protéine IB-5 en solution d'après ses déplacements chimiques.

2.1 RamaDA (Ramachandran Domain Analysis)

2.1.1 Introduction

Les IDP et les protéines repliées sont composées des mêmes acides aminés. Ces acides aminés ont donc les mêmes propriétés et sont amenés à subir les mêmes contraintes stériques. Les conformations qu'ils adoptent seront donc les mêmes. Partant de ce constat, un ensemble représentatif des conformations adoptées par les protéines repliées conviendra également aux IDP. Pour notre étude, nous choisi d'utiliser un en-

semble composé de 500 des meilleures structures de protéines présentes dans la PDB (Protein DataBank). Lovell et ses collaborateurs [45] ont sélectionné ces dernières pour représenter au mieux l'espace conformationnel de tous les acides aminés. Certains fichiers obsolètes ont été remplacés par les fichiers PDB récents (1XFF, 1GOK, 1E70 and 1IG5). 1A1Y n'ayant pas de pendant récent, il a été gardé. L'ensemble ainsi décrit sera nommé top500 par la suite. Il contient 110 018 acides aminés.

Il existe sur le squelette d'une protéine trois angles dihèdres par acide-aminé : φ, ψ et ω. Ces angles dihèdres sont très variables à l'exception de ω qui ne peut prendre que deux valeurs : environ 0° pour un acide-aminé cis et environ 180° pour un trans. Ces angles définissent la conformation spatiale de l'acide-aminé.

Ramachandran et Ramakrishnan [46] mettent en avant en 1965 que seuls quelques couples (φ, ψ) sont possibles et les représentent dans le graphe de ψ en fonction de φ. Ce graphe sera appelé diagramme de Ramachandran dans la suite de cette thèse. Les régions vides, ou non-autorisées, du diagramme ont été justifiées en grande partie par des exclusions stériques alors que les autres régions, appelées autorisées ou favorisées, ont été associées à des éléments de structure secondaire.

Depuis le travail de Ramachandran, sept domaines conformationnels ont été identifiés dans les régions autorisées du diagramme (voir figure 2.1) [45, 47–49]. β et hélice-droite décrivent les conformations majoritairement présentes dans les éléments de structure secondaire et PPII celles des hélices polyproline. Hélice-gauche correspond aux conformations qu'on retrouverait majoritairement dans des hélices de symétrie gauche, γ correspond aux conformations spécifiques observées dans les coudes γ [50], ζ est exclusivement composé d'acides aminés précédant une proline [51] et PPII_R (appelé aussi beta_{PR} [51]) correspond à une conformation qu'on pourrait observer dans des hélices PPII-droite. Chaque acide-aminé de chaque protéine se trouve dans l'un de ces sept domaines conformationnels. Le diagramme de Ramachandran est donc largement

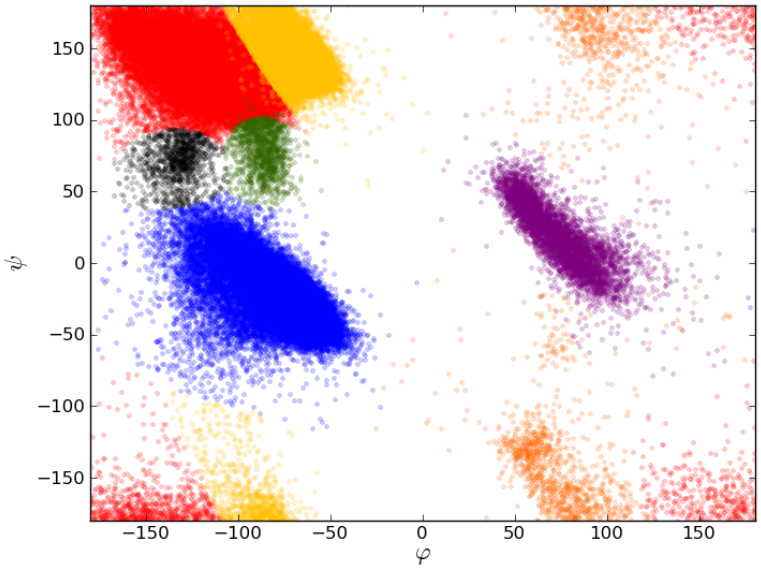

Figure 2.1 – *Diagramme de Ramachandran de top500, chaque point désigne un acide-aminé. Sept domaines sont représentés : β en rouge, PPII en jaune, hélices-droite en bleu, hélices-gauche en violet, γ en vert, ζ en noir et PPII$_R$ en orange.*

utilisé pour vérifier la validité de la structure globale d'une protéine, le pourcentage d'acide-aminés présents dans les régions non-autorisées servant le plus souvent de paramètre discriminant [52, 53].

Concernant les domaines conformationnels, les associer à des éléments de structure secondaire est un abus de langage. En effet, les acides aminés d'un élément de structure secondaire donné adoptent quasiment tous la même conformation du fait de la nature-même de l'élément mais, comme on peut le voir dans le cas des hélices sur la figure 2.2, toutes sortes de conformations peuvent être retrouvées.

Cette différence de notion entre élément de structure secondaire et conformations est la clé de voûte du raisonnement mené ici. Les conformations peuvent permettre d'étudier les protéines repliées aussi bien de les IDP alors que les éléments de structure secondaires sont réservés aux seules protéines repliées. Deux notions différentes mènent à deux types d'information. Nous tenterons dans cette étude d'extraire toute

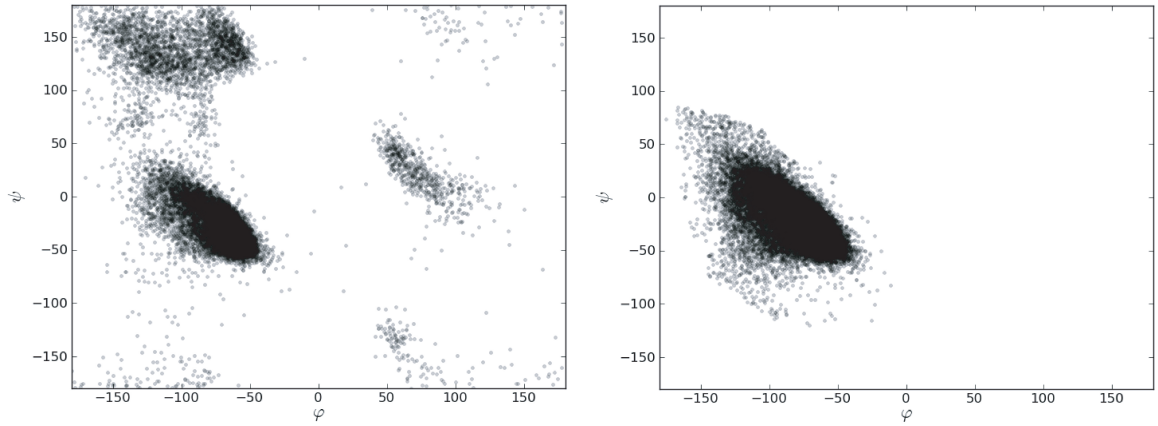

Figure 2.2 – *Diagrammes de Ramachandran sur top500. A gauche, les acides aminés appartenant à une hélice d'après le prédicteur de structure secondaire DSSP. A droite, les acides aminés dont la conformation se situe dans le domaine des hélices d'après Ramachandran.*

l'information contenue dans les conformations des acides aminés.

Dans le cas des protéines repliées, les acides aminés gardent le plus souvent leur conformation au cours du temps. Seules leurs parties flexibles peuvent changer de conformation comme le font les IDP. Dans ce dernier cas, un acide-aminé peut soit naviguer entre plusieurs domaines conformationnels, soit rester dans un seul. Une multitude de possibilités s'offrent alors à nous et connaître la ou les conformations des acides aminés d'une protéine au cours du temps pourrait donc nous informer sur le comportement de cette dernière en solution. Pour cela, nous avons créé un modèle du diagramme de Ramachandran et développé un outil permettant d'attribuer à chaque acide-aminé son ou ses domaines conformationnels.

2.1.2 Mise en place d'un modèle gaussien

Pour attribuer une conformation à un acide-aminé, il nous faut créer un modèle du diagramme de Ramachandran. Nous devons notamment définir clairement les limites de chaque domaine conformationnel.

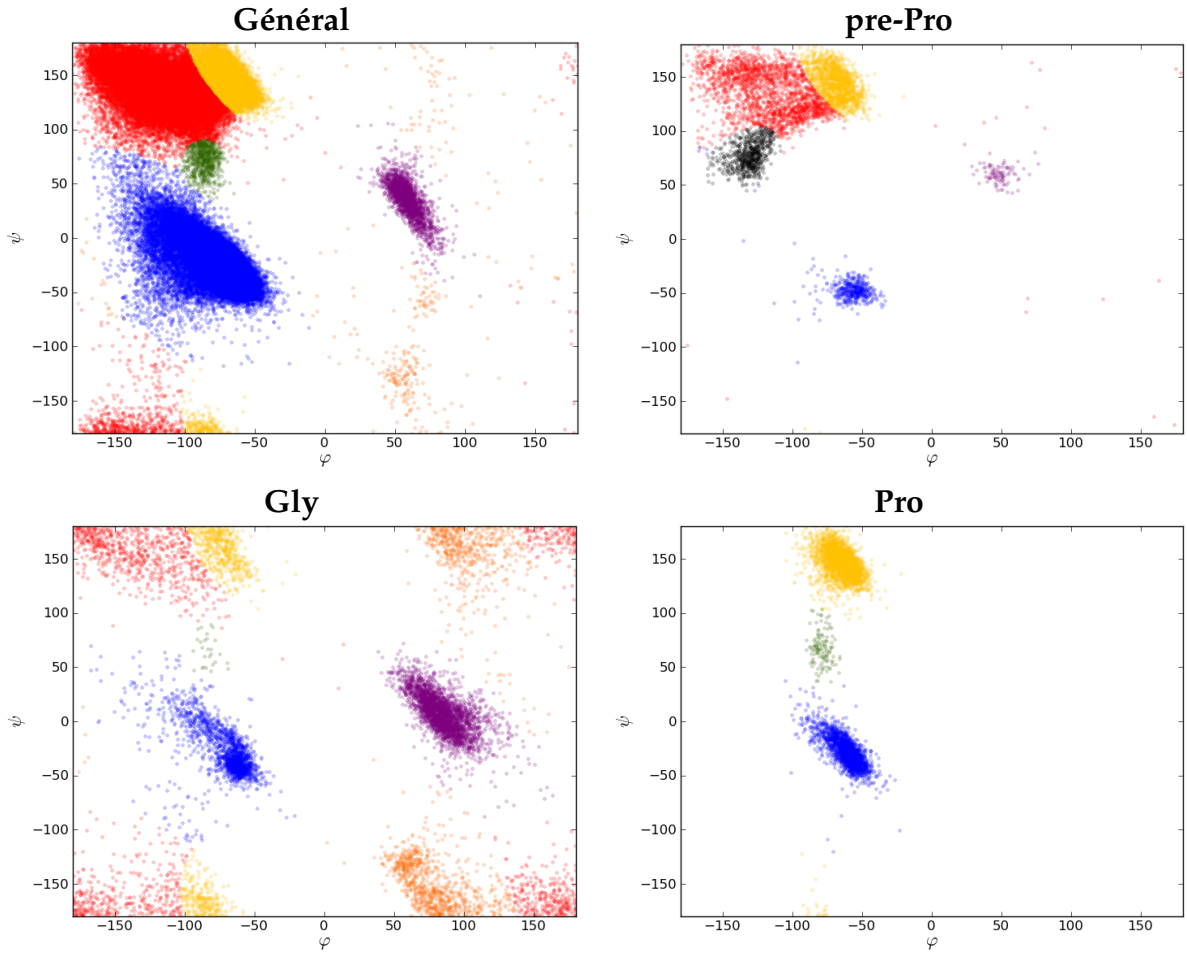

Figure 2.3 – *Diagrammes de Ramachandran de top500, pour les quatre sous-ensembles définis. Domaines conformationnels colorés comme suit : β en rouge, PPII en jaune, hélices-droite en bleu, hélices-gauche en violet, γ en vert, ζ en noir et $PPII_R$ en orange.*

L'étude de top500 nous a amené à définir quatre sous-ensembles d'acides aminés dont la distribution des couples (φ, ψ) est différente : les acides aminés précédant une proline (appelé pre-Pro), les glycines restantes (Gly), les prolines restantes (Pro) et tous les autres acides aminés (General). Ces sous-ensembles de couples (φ, ψ) sont représentés en figure 2.3.

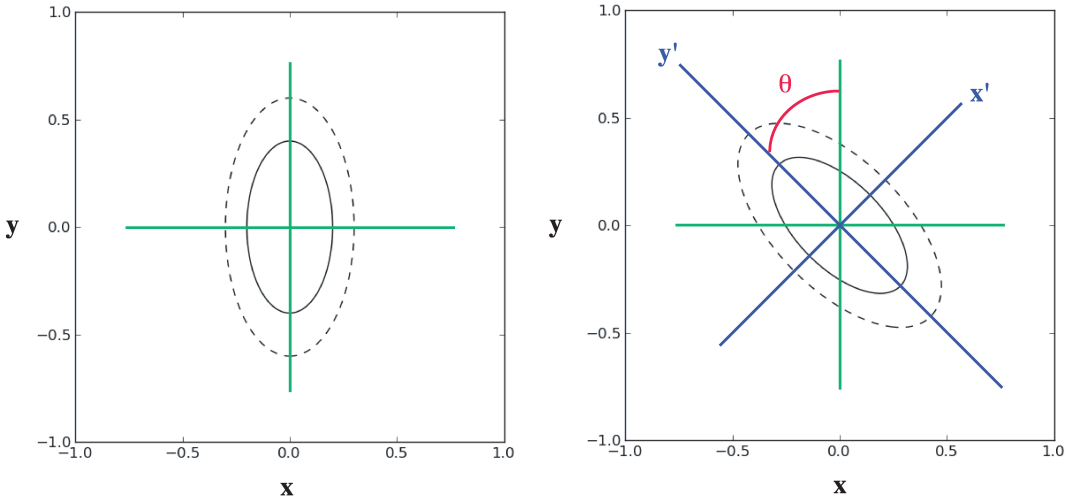

Figure 2.4 – *À gauche, une gaussienne centrée en (0.0,0.0) d'écart-type 0.1 sur x et 0.2 sur y. À droite, la même gaussienne, tournée d'un angle $\theta = -45°$. L'équation de la gaussienne de gauche dans le repère (x,y) est valable pour celle de droite dans le repère (x',y').*

Fonctions utilisées

Pour modéliser les quatre sous-ensembles d'acides aminés, le diagramme de Ramachandran a été découpé en carrés de 1° sur 1°. Un histogramme du nombre d'acides aminés présents par carré a été recueilli pour top500. Afin de représenter au mieux cet histogramme, nous avons utilisé une somme de fonctions gaussiennes-2D dont l'équation générale de chacune de ces fonctions est donnée par l'équation 2.1.

$$G(x,y) = Ae^{-\left(\frac{(x-x_c)^2}{2{\sigma_x}^2} + \frac{(y-y_c)^2}{2{\sigma_y}^2}\right)} \tag{2.1}$$

A désigne une constante d'échelle, (x_c, y_c) sont les coordonnées du centre de la gaussienne et σ_x et σ_y sont les écarts-types selon les axes de la gaussienne. Dans notre cas, les axes principaux des gaussiennes définies pour chaque domaine ne correspondent pas aux axes φ et ψ. Ces axes sont tournés d'un angle θ, présenté dans la figure 2.4. On appelle alors (φ_c, ψ_c) les coordonnées du centre de la gaussienne dans le repère du dia-

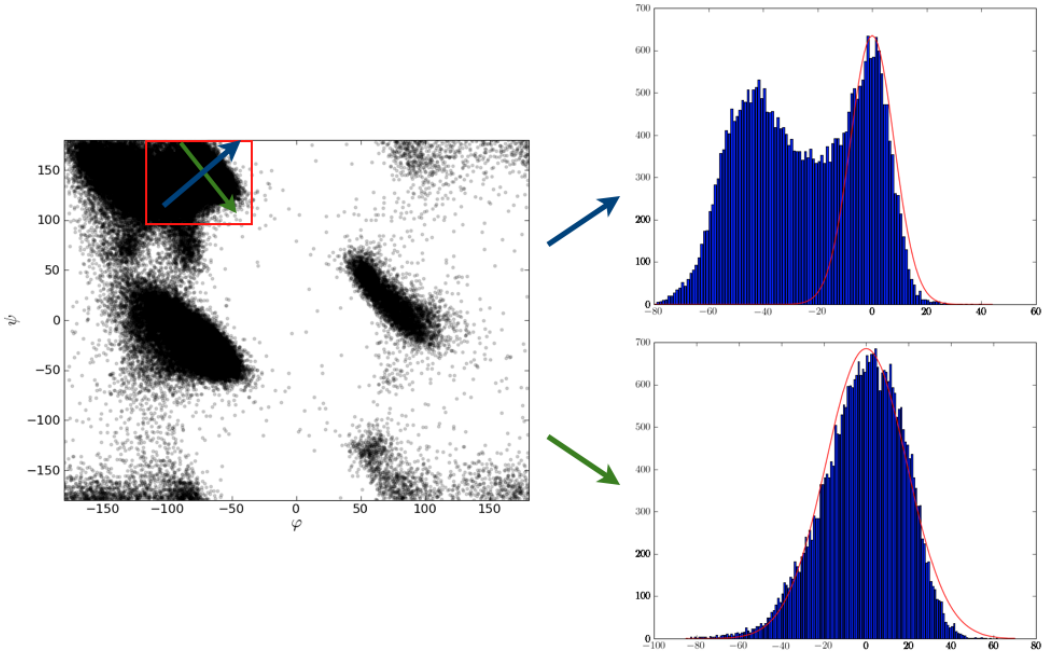

Figure 2.5 – *Le diagramme de Ramachandran de top500 (à gauche) est transformé en histogramme de présence d'un acide-aminé (à droite). Les histogrammes de droite sont ceux recueillis selon les axes principaux de la gaussienne définie empiriquement pour le domaine PPII. Ces axes sont représentés par des fléches sur le diagramme de Ramachandran. La courbe en rouge représente la coupe de la gaussienne-2D selon ses axes principaux. Le deuxième domaine, visible dans l'histogramme en haut à droite, est le domaine β voisin.*

gramme de Ramachandran et $\sigma_{\varphi'}$ et $\sigma_{\psi'}$ les écarts-types de la gaussienne selon les axes du référentiel tourné.

Dans le cadre d'un espace de topologie torique comme celui dans lequel nous nous trouvons, il faudrait utiliser une définition particulière des gaussiennes afin qu'elles soient définies correctement dans tout l'espace. En effet, φ et ψ sont définis sur l'intervalle $]-180°, 180°]$ modulo 360°, une gaussienne ayant pour centre [180°,180°] va donc se retrouver dans les quatre coins du diagramme. Cependant, en première approximation, nous pouvons utiliser la définition donnée précédemment en prenant garde aux

Domaine	(φ_c, ψ_c)	$\sigma_{\varphi'}$	$\sigma_{\psi'}$	**Angle**
hélice-droite (1)	(-63.07°, -42.23°)	3.54°	5.77°	-36.31°
(2)	(-62.15°, -28.74°)	12.00°	4.69°	-61.76°
(3)	(-83.72°, -16.01°)	30.22°	10.95°	-56.15°
hélice-gauche (Gly)	(82.38°, 6.89°)	7.55°	20.63°	-32.81°
(pre-Pro)	(47.06°, 5.93°)	5.58°	6.79°	-27.38°
(General)	(56.35°, 39.04°)	4.86°	15.18°	-25.00°
β	(-119.12°, 136.48°)	15.77°	29.98°	-52.51°
PPII	(-68.03°, 144.89°)	9.65°	16.67°	-30.37°
γ	(-84.90°, 69.28°)	5.85°	10.82°	-6.51°
ζ	(-130.46°, 76.31°)	5.90°	12.80°	12.25°
PPII_R	(76.13°, -162.12°)	11.75°	41.02°	-29.27°

Tableau 2.1 – *Paramètres des fonctions gaussiennes utilisés pour modéliser le diagramme de Ramachandran.*

repliements de l'espace sur lui-même.

Le nombre de gaussiennes par domaine conformationnel ainsi que leurs paramètres (centre, écarts-types, angle θ) ont été évalués dans un premier temps de façon empirique. La figure 2.5 montre l'exemple du domaine PPII. Dans ce cas, on a choisi d'utiliser une seule gaussienne et on peut voir que les paramètres définis sont déjà en bon accord avec l'histogramme de présence de top500.

Pour affiner les valeurs trouvées, un programme informatique a été créé. Il permet de faire une minimisation au moindre carré de la somme des gaussiennes définies. Il calcule donc en chaque point de l'histogramme l'écart entre la somme des gaussiennes et la valeur de l'histogramme et fait la somme de leurs carrés. Les paramètres des gaussiennes sont alors modifiés de façon à minimiser cette valeur jusqu'à atteindre une valeur stable. Les paramètres trouvés à la fin de la minimisation sont regroupés dans le tableau 2.1.

Analyse des résultats

On peut remarquer que les sept domaines conformationnels cités en introduction ne sont pas présents dans les quatre sous-ensembles. Les prolines sont restreintes aux domaines PPII, γ et hélices-droite tandis que le domaine ζ apparaît dans pre-Pro alors que les domaines γ et PPII$_R$ en sont absents.

Il est intéressant de noter que le domaine des hélices-gauche n'est pas décrit par la même gaussienne selon les sous-domaines. De plus, la plupart des domaines sont définis par une seule gaussienne. Seul le domaine des hélices droite est décrit par 3 gaussiennes, que nous noterons par la suite (1), (2) et (3). Après analyse, ces gaussiennes ne peuvent pas être associées à un certain type d'acide-aminé chacune, nous avons donc cherché à savoir si elles pouvaient alors être spécifiques à un type d'hélice.

Répondre à cette question nécessite tout d'abord d'associer un acide-aminé et sa conformation. Pour cela, chaque fonction gaussienne définie précédemment a servi de loi de probabilité de présence d'un couple (φ, ψ) dans le domaine concerné. Pour chaque acide-aminé, la probabilité de présence de son couple d'angles dihèdres est déterminée indépendamment pour chaque domaine. La plus grande probabilité désigne alors la conformation adoptée par l'acide-aminé. Cette approche statistique est la base du programme RamaDA (pour **Rama**chandran **D**omain **A**nalysis) que nous avons développé.

Les gaussiennes définies se chevauchent toutes sur le diagramme. Une grande majorité des zones de chevauchement pose peu de problème dans la mesure où elles se trouvent dans les parties "non-autorisées" du diagramme. Cependant il peut subsister un doute sur l'attribution des domaines β et PPII. Ces deux domaines conformationnels étant associés à des structures régulières, nous avons ajouté à RamaDA la possibilité de changer l'attribution d'un acide-aminé dont la conformation β (respectivement, PPII) se trouve entouré d'au moins quatre autres acides aminés en conformation PPII (respectivement, β).

Domaine RamaDA	Analyse DSSP associée	pourcentage de présence
(1)	hélice α	95.5 %
	hélice π	<0.1 %
	hélice 3_{10}	0.6 %
	autres	3.9 %
(2)	hélice α	46.1 %
	hélice π	<0.1 %
	hélice 3_{10}	21.1 %
	autres	32.8 %
(3)	hélice α	25.9 %
	hélice π	0.2 %
	hélice 3_{10}	10.5 %
	autres	63.4 %

Analyse DSSP	Domaine RamaDA associé	pourcentage de présence
hélice α	(1)	73.5 %
	(2)	17.8 %
	(3)	8.5 %
	autres	0.2 %
hélice π	(1)	16.1 %
	(2)	6.5 %
	(3)	74.2 %
	autres	3.2 %
hélice 3_{10}	(1)	3.7 %
	(2)	63.0 %
	(3)	26.7 %
	autres	6.6 %

Tableau 2.2 – *Comparaison des analyses RamaDA et DSSP pour les hélices.*

RamaDA a été utilisé pour attribuer la conformation de tous les acides aminés de top500. Dans un premier temps, les trois fonctions gaussiennes décrivant le domaine des hélices-droite ont été traitées séparément. Celles-ci se chevauchant fortement, on utilise le même principe que précédemment pour limiter les erreurs d'attribution. Les résultats ont été comparés à l'analyse des mêmes fichiers PDB par DSSP. DSSP [54] est un programme permettant de détecter les éléments de structures secondaires à partir des motifs de liaisons hydrogène présents dans une protéine. Ces mêmes motifs lui permettent de discriminer différents types d'hélices : α, 3_{10} et π. L'accès aux résultats de l'analyse DSSP des fichiers de top500 a été possible *via* la base de données en ligne contenant les résultats de l'analyse de tous les fichiers présents dans la PDB.

Le tableau 2.2 montre en quelles proportions un acide-aminé dont la conformation est hélice-droite pour RamaDA est effectivement dans une hélice pour DSSP, et inversement. Les acides aminés sans attribution DSSP sont laissés de côté. Il apparaît que 93% des acides aminés en conformation (1) pour RamaDA sont dans une hélice α d'après DSSP. Cependant, les conformations (2) et (3), bien qu'elles décrivent la quasi-totalité

des hélices 3_{10} et π respectivement, se retrouvent très fréquemment dans une hélice alpha. On décide donc de renommer la conformation (1) en α et de considérer les conformations (2) et (3) ensemble sous le nom générique de hélice-droite.

Le modèle du diagramme de Ramachandran décrit ici possède donc maintenant huit domaines conformationnels. La figure 2.6 présente le diagramme de Ramachandran des quatre sous-ensembles définis dans top500 ainsi que les contours des fonctions gaussiennes utilisées par RamaDA pour chaque domaine conformationnel. On voit bien sur cette figure les différents chevauchements de gaussiennes problématiques discutés précédemment.

Cependant, il existe, pour les IDP, un état structural qui n'est pas visible directement sur le diagramme de Ramachandran : l'état *random-coil*. En effet, un acide-aminé appartenant à une région désordonnée d'une protéine n'a pas, par définition, de conformation stable dans le temps. Par contre, à un instant donné, cet acide-aminé a obligatoirement une conformation puisqu'il a une position dans l'espace. Or, un fichier PDB issu de RMN et donnant plusieurs modèles d'une même protéine peut être considéré comme un ensemble de clichés de la protéine en différents instants.

Partant de ces deux constats, il est apparu que l'état *random-coil* pouvait être déterminé par l'analyse de tous les modèles d'un fichier PDB issu de la RMN. Si pour tous les modèles, un acide-aminé a la même conformation alors il s'agit d'une partie peu ou pas flexible de la protéine alors que si les modèles ne donnnent pas les mêmes résultats, il s'agit d'une partie très flexible. On considère par la suite que la conformation d'un acide-aminé est *random-coil* si moins de 65% des modèles concordent. De plus, il existe des acides aminés qui oscillent entre les domaines conformationnels β et PPII, leur conformation sera alors appelée étendue. Bien sûr, ces attributions n'auront de sens que si l'ensemble des modèles est représentatif de la façon dont se comporte la protéine en solution.

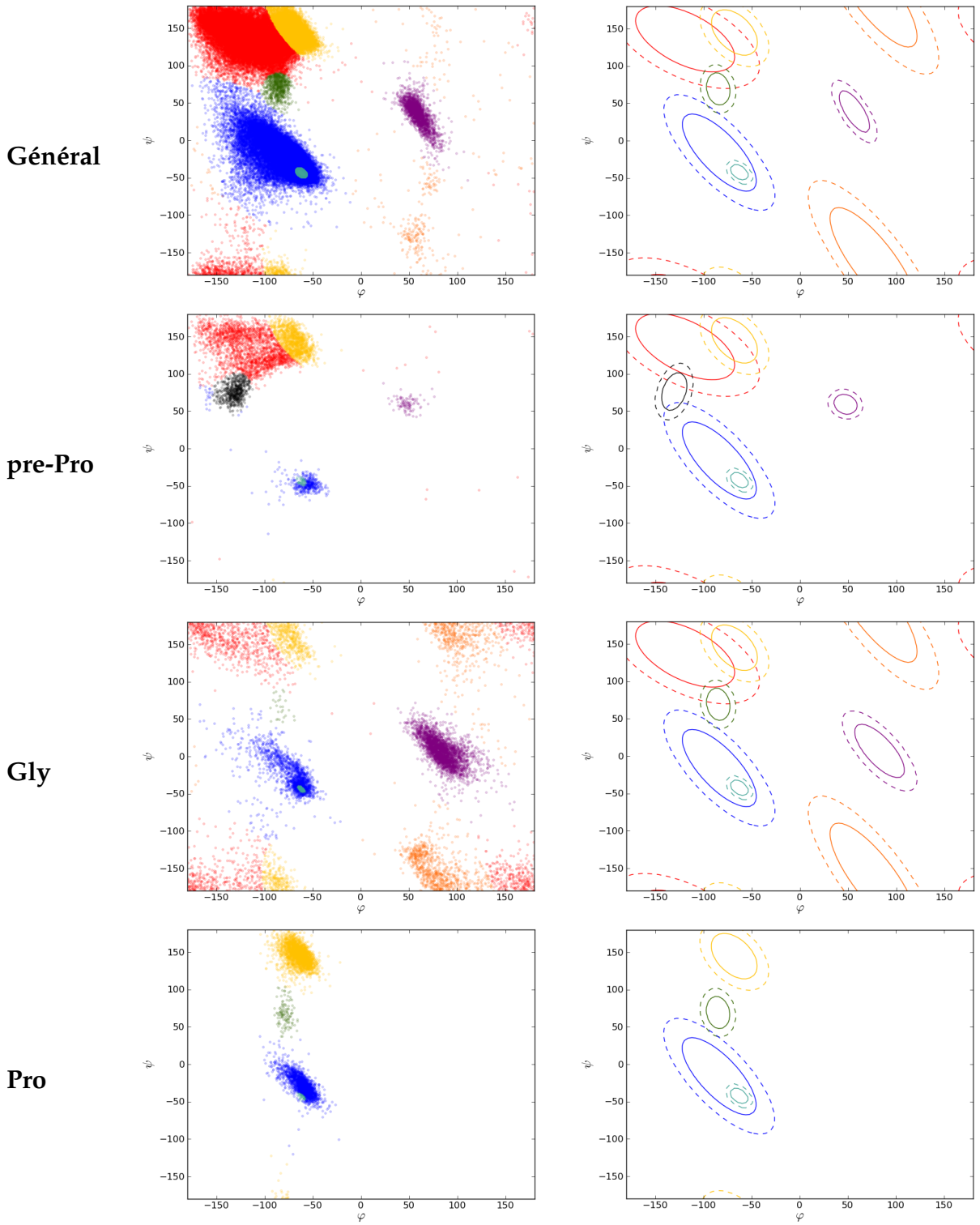

Figure 2.6 – *Diagramme de Ramachandran de top500, pour les quatre sous-ensembles définis. Domaines conformationnels colorés comme suit : β en rouge, PPII en jaune, hélices-droite en bleu, α en turquoise, hélices-gauche en violet, γ en vert, ζ en noir et $PPII_R$ en orange. Les courbes en pointillés donnent le contour des gaussiennes à une distance de trois fois l'écart-type par rapport au centre, les courbes en trait plein donnent le contour à une distance de deux fois l'écart-type.*

Nom	α	hélices-droite	β	PPII	hélices-gauche	γ	ζ	PPII_R	*random-coil*	*étendu*
Lettre	A	H	B	P	L	G	Z	Q	R	e

Tableau 2.3 – *Equivalence entre lettre et nom de domaine ou d'état structural.*

Grâce à l'attribution conformationnelle de RamaDA et à son extension aux problèmes de flexibilité et de *random-coil*, il est possible d'avoir la description complète d'une protéine. On peut donc largement ouvrir l'éventail d'applications du diagramme de Ramachandran pour la biologie structurale.

2.1.3 Applications

Dans cette partie, les domaines conformationnels seront nommés par une lettre, comme c'est le cas en sortie du programme RamaDA. Le tableau 2.3 donne les correspondances entre les lettres et les noms des domaines conformationnels ou des états structuraux qu'ils représentent.

Toutes les applications présentées dans ce paragraphe, exceptée la dernière, sont disponibles dans la version en ligne de RamaDA à l'adresse http ://ramada.u-strasbg.fr. Une version régulièrement mise à jour du programme est téléchargeable depuis cette adresse pour une utilisation hors-ligne.

RamaDA a été écrit en Python 2.5 avec l'aide de la bibliothèque Biopython [55]. L'affichage de la partie graphique est optimisé pour Firefox 3.0.

Validation de structure de protéine

Valider une structure de protéine consiste en l'évaluation de la probabilité de présence de chaque acide-aminé en sa position (φ, ψ) sur le diagramme de Ramachandran. C'est probablement l'utilisation la plus commune du diagramme de Ramachandran en

	top500	**PDB**
z-score (minimum)	0.688	0.521
(maximum)	1.900	4.215
(moyen)	1.190	1.386

Tableau 2.4 – *Statistiques sur les z-scores de deux ensembles de structures de protéine.*

biologie structurale. Si tous les acides aminés d'une structure ont de faibles probabilités d'existence, la structure doit être remise en question.

Pour concentrer cette information en un seul paramètre, nous avons fait appel au z-score. En effet, grâce à notre approche statistique et l'équation 2.2, reliant la probabilité de présence d'un acide-aminé dans son domaine conformationnel P et son z-score Z, il est très facile de le déterminer.

$$P = e^{-\frac{Z^2}{2}} \tag{2.2}$$

D'après cette équation, Z peut être positif ou négatif. Comme seul l'écart à la position de plus grande probabilité nous intéresse, nous ne considérerons par la suite que la solution positive.

On peut alors calculer le z-score d'une structure de protéine comme la moyenne des z-scores de tous les acides aminés qui la composent. Une protéine dont les acides aminés sont répartis dans les domaines conformationnels selon une loi gaussienne a un z-score de 1.253. Un z-score trop éloigné de cette valeur remet en cause la validité de la structure proposée.

Quelques statistiques issues de top500 et de la PDB sont rassemblées dans le tableau 2.4. Les z-scores moyens correspondent bien à ce qui était attendu. Ils sont tous les deux proches de la valeur théorique et légèrement meilleur pour top500 vu qu'il est composé de 500 des meilleures structures de la PDB. Ce dernier point est d'ailleurs corroboré par

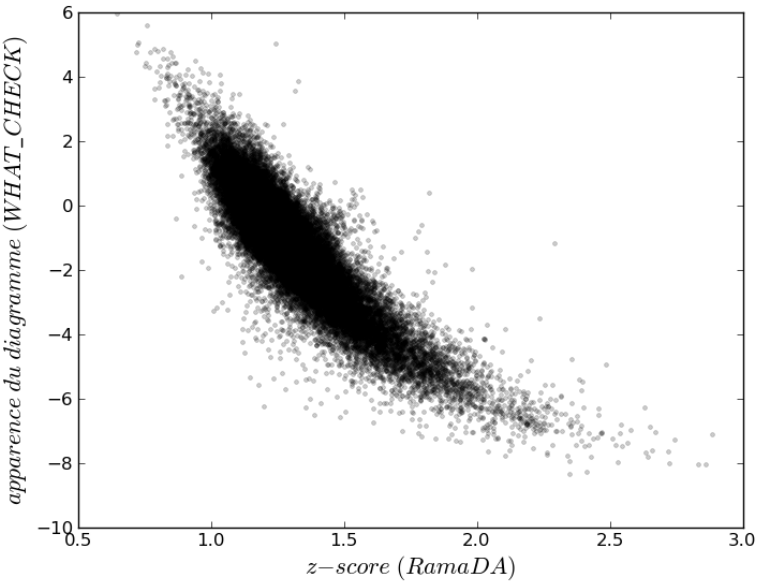

Figure 2.7 – *Corrélation entre le paramètre d'*apparence du diagramme *donné par WHAT_CHECK et le z-score donné par RamaDA. Le coefficient de Pearson est de -0.88.*

les z-scores minimum et maximum de top500 qui montrent une faible dispersion du paramètre. Dans la PDB, un fichier ayant un faible z-score peut correspondre à une structure artificiellement construite dans les zones autorisées du Ramachandran. De même, un fort z-score maximum confirme qu'il existe dans la base de données des structures discutables.

Afin de montrer l'efficacité des z-scores à valider des structures, ils ont été comparés au paramètre d'*apparence du diagramme de Ramachandran* utilisé par le logiciel de validation de structures WHAT_CHECK [53]. La figure 2.7 montre, pour top500, l'excellente corrélation entre ces deux paramètres dont le coefficient de Pearson est -0.88.

La validation de structure peut paraître inutile lorsque l'on parle de IDP puisqu'aucun fichier de structure ne décrit ces protéines. On verra par ailleurs au chapitre suivant que la génération de conformères aléatoires d'IDP pourra combler ce manque. Cependant, les boucles et autres régions désordonnées de protéines sont présentes dans la

PDB. Si toute la région possède un mauvais z-score alors on peut supposer qu'elle a mal été décrite, une telle région devra alors être considérée avec précaution dans l'étude des parties désordonnées de la protéine.

Indications sur les éléments de structure secondaire

La comparaison réalisée précédemment entre RamaDA et DSSP pour les hélices (voir paragraphe 2.1.2) a été élargie aux domaines conformationnels β et PPII afin d'avoir des données sur toutes les motifs réguliers. Cette fois-ci, les acides aminés ne possédant pas d'attribution DSSP sont pris en compte. Le tableau 2.5 regroupe les résultats de cette comparaison.

On remarque que les hélices et les brins étendus de DSSP sont retrouvés à environ 90% par RamaDA. Rien d'étonnant à cela puisque, comme cela a été évoqué dans l'introduction, domaines conformationnels et éléments de structure secondaires sont fortement liés. Cependant, si l'adéquation n'est pas parfaite, c'est parce que domaines conformationnels et éléments de structure secondaires ne sont en aucun cas équivalents (voir figure 2.2 pour rappel).

Si les différents types d'hélices et les brins β sont reconnus par DSSP, l'hélice PPII n'est, quant à elle, pas détectée par ce logiciel. La cause en est que DSSP se base en priorité sur les réseaux de liaisons hydrogène et que les hélices PPII n'en ont pas. Et c'est le cas pour de nombreux autres logiciels de détection de structures secondaires comme P-SEA [56] par exemple. RamaDA, quant à lui, est capable de montrer des enchaînements d'acides aminés en conformation `P`, susceptibles de former des hélices PPII.

De tels enchaînements, de plus de 5 résidus, ont été repérés 39 fois dans top500. Après observation des structures correspondantes, on peut affirmer que la plupart des enchaînements sont effectivement des hélices PPII. Seuls 12 n'ont pu être confirmés. Dans la PDB, on trouve 27 902 enchaînements de plus de 5 résidus en conformation `P` et

Domaine RamaDA	Analyse DSSP associée	pourcentage de présence
A	hélice α	95.2 %
	hélice π	<0.1%
	hélice 3_{10}	0.5 %
	brin étendu	0.2 %
	non attribué	0.4 %
	autres	3.6 %
H	hélice α	33.7 %
	hélice π	0.1 %
	hélice 3_{10}	14.8 %
	brin étendu	2.5 %
	non attribué	8.9 %
	autres	40.0 %
B	hélice α	<0.1 %
	hélice π	-
	hélice 3_{10}	<0.1 %
	brin étendu	64.9 %
	non attribué	23.1 %
	autres	12.0 %
P	hélice α	<0.1 %
	hélice π	-
	hélice 3_{10}	0.4 %
	brin étendu	15.6 %
	non attribué	61.3 %
	autres	22.7 %

Analyse DSSP	Domaine RamaDA associé	pourcentage de présence
hélice α	A	73.1 %
	H	26.6 %
	B	<0.1 %
	P	<0.1 %
	autres	0.3 %
hélice π	A	12.9 %
	H	83.9 %
	B	-
	P	-
	autres	3.2 %
hélice 3_{10}	A	3.0 %
	H	90.4 %
	B	0.1 %
	P	1.2 %
	autres	5.3 %
brin étendu	A	0.2 %
	H	2.9 %
	B	85.1 %
	P	9.5 %
	autres	2.3 %
non attribué	A	0.6 %
	H	11.8 %
	B	34.0 %
	P	42.0 %
	autres	11.6 %

Tableau 2.5 – *Comparaison des analyses RamaDA et DSSP.*

575 de plus de 10 résidus. Par exemple, le fichier 1BFD, dont l'attribution complète est donnée en figure 2.8, possède une hélice PPII de 14 résidus. Seuls 3 acides aminés de cette hélice sont des prolines, confirmant le caractère universel de l'hélice PPII.

Afin de mettre en évidence les éléments de structure secondaire repérables par RamaDA, une interface graphique a été développée. À partir de trois acides aminés consécutifs de même conformation, une image indique la possible présence d'une hélice α, d'une hélice-droite, d'un brin β ou d'une hélice PPII (voir un exemple en figure 2.8).

La détection d'hélices PPII peut se révéler intéressante du point de vue de l'étude de la fonction d'une protéine, notamment pour l'implication de telles hélices sur la fonction et le nombre de partenaires envisageables pour la protéine. Cependant, il existe de nombreuses régions ou protéines désordonnées n'ayant pas d'éléments de structure secondaire. Nous allons voir grâce à l'application suivante que l'attribution conformationnelle peut nous aider à mieux les décrire.

Détermination de signatures de domaines - Applications aux mains-EF

Les hélices PPII ne sont pas les seuls éléments caractéristiques que l'on peut retrouver par RamaDA. En effet, les boucles, par exemple, peuvent avoir une signature conformationnelle particulière que l'on peut retrouver dans plusieurs protéines différentes. Nous traiterons ici, sur l'exemple des mains-EF, le fait que l'attribution conformationnelle de RamaDA permet de repérer ces signatures et d'associer les protéines présentant les mêmes.

Les mains-EF sont des domaines protéiques composés de deux hélices α séparées par une boucle de 9 acides aminés capable de lier un ion calcium [57].

Pour déterminer si cette boucle a une signature conformationnelle propre, 4 protéines connues pour avoir au moins une main-EF et dont la structure a été caractérisée par RMN ont été analysées avec RamaDA. L'attribution conformationnelle des mains-EF de la calerythrine (PDB 1NYA), de la calmoduline (PDB 2K0E), de la parvalbumine (PDB 1RJV) et d'une protéine de canal sodium du muscle cardiaque (PDB 2KBI) sont regroupées dans le tableau 2.6 avec leur séquence et leur analyse DSSP.

On voit clairement que là où l'enchaînement d'acides aminés et l'analyse DSSP donnent un consensus large, RamaDA est capable de livrer une signature précise de la boucle d'une main-EF. Grâce à cette signature et à l'attribution conformationnelle complète de la PDB, nous avons pu trouver 482 fichiers contenant une ou plusieurs protéines

Protéine	Séquence	RamaDA	DSSP
calerythrine (PDB 1NYA)	DFDGNGALE	LAHLHLBBB	-SS--SSB-
	DKNADGQIN	BHHLHLBBP	-SS--SEEE
	DTNGNGELS	BHHLHLBBP	-TT-SSEEE
calmoduline (PDB 2K0E)	DKDGDGTIT	eHHLHLBBP	-SS-SS-B-
	DADGNGTID	RHHLHLBBP	-SSSSSSB-
	DKDGNGYIS	eHHLHLBBP	SSS-SS-EE
	DIDGDGQVN	eHHLHLBBP	-SS-SSEE-
parvalbumine (PDB 1RJV)	DKDKSGFIE	RHHLHLBBB	-TT-SS-B-
	DKDGDGKIG	GHHLHLBPB	-SSSSSSB-
canal cardiaque sodium (PDB 2KBI)	DPEATQFIE	ZHHLHLBBP	-TT--SEEE
consensus	DX[DNE][GAK]X[GQ]X[ILV]X	X[HA]HLHLB[BPe][BPe]	[-S][ST][ST][-S][-S][-SE][BE][-E]

Tableau 2.6 – *Comparaison entre 10 mains-EF : séquence, attribution RamaDA et analyse DSSP. X correspond à n'importe quelle valeur et les valeurs entre crochet montrent les différentes possibilités pour un même acide-aminé.*

1BFD

```
Chain A
z-score : 1.21
```

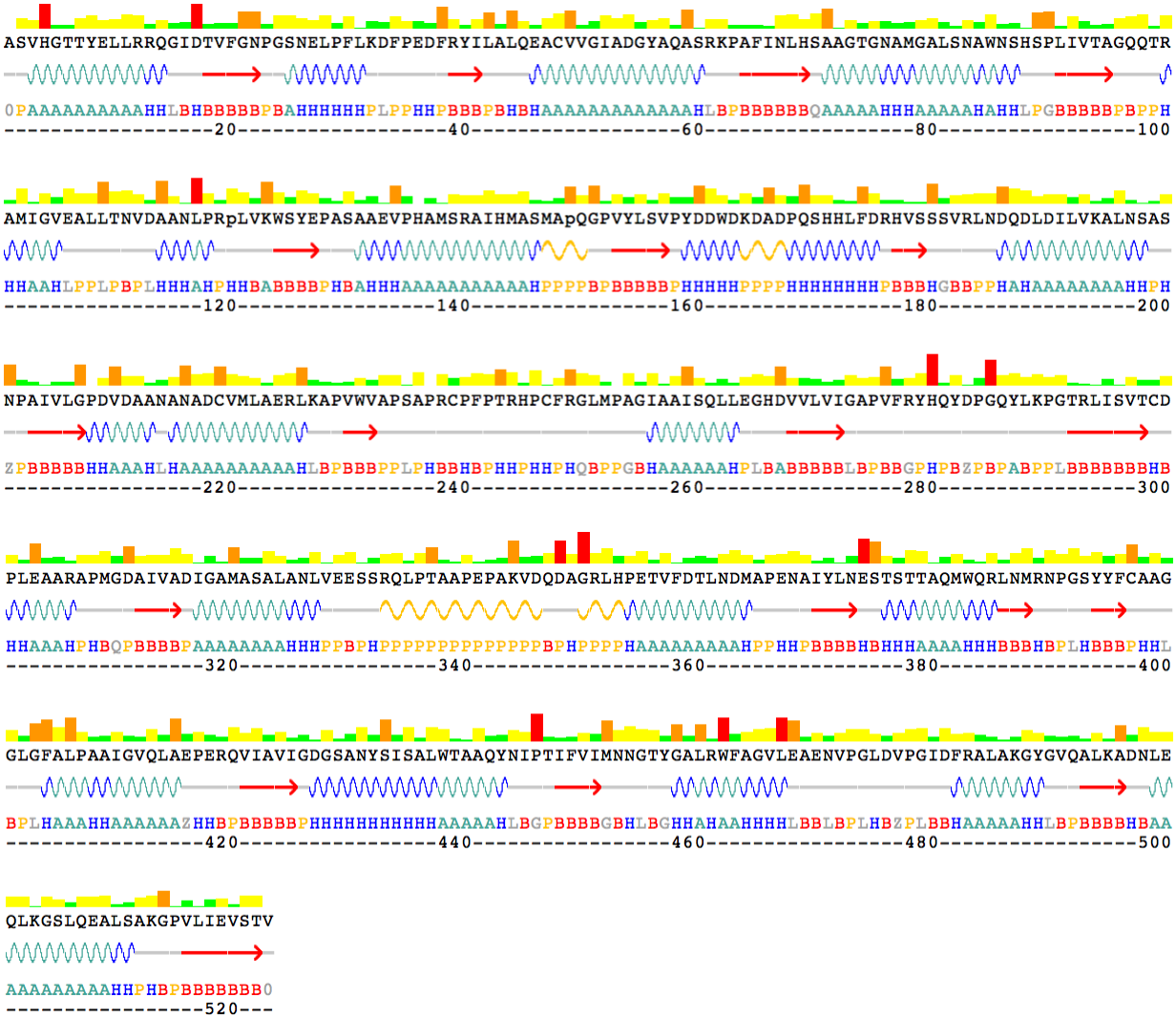

Figure 2.8 – *Résultat obtenu en sortie de RamaDA. Exemple de la benzoylformate de* pseudomonas putida *(fichier PDB 1BFD). Le première ligne montre le z-score sous forme d'un histogramme. Un z-score inférieur à 1 sera donnée en vert, un z-score compris entre 1 et 2 sera jaune, entre 2 et 3 orange et supérieur à 3 rouge. La deuxième ligne correspond à la séquence de la protéine. Les cis-prolines sont détectées par leur angle ω et indiquée par p. La troisième ligne donne les indications de structure (vagues bleue pour les hélices, turquoise pour les hélices α, jaune pour les hélices PPII et flèche rouge pour les brins β). La dernière ligne donne le résultat de l'attribution des domaines conformationnels pour chaque acide-aminé de la protéine.*

qui possèdent deux hélices séparées d'une boucle similaire à une main-EF. En regardant la littérature, il s'avère que 465 de ces fichiers sont décrits comme contenant une main-EF ou des boucles connues pour lier le calcium, que de telles propriétés n'ont pas été décrites pour les 17 restants. Le taux de faux positifs éventuels de notre recherche s'élève donc seulement à 3.5 %.

La recherche de signatures peut donc permettre d'associer facilement des protéines par la fonction d'une de leur boucle. Elle semble donc être une application prometteuse de RamaDA pour les domaines désordonnés, il faudra néanmoins la tester sur de nombreux domaines pour prouver toute son efficacité.

Population des domaines conformationnels

Afin de mieux comprendre le comportement d'un ensemble d'acides aminés tel qu'une boucle ou toute une partie désordonnée de protéine, et pour compléter ou remplacer l'analyse des signatures de domaines décrite précédemment, il est intéressant de regarder quels domaines conformationnels sont préférentiellement adoptés par les acides aminés.

Pour cela, des statistiques ont été réalisées sur top500 et sont présentées dans les tableaux 2.7 et 2.8. Comme remarqué sur le diagramme de Ramachandran (voir figure 2.3), le sous-ensemble pre-Pro présente une distribution des acides aminés très différente des autres sous-ensembles.

Acide-aminé	A	H	B	P	L	G	Q	Z
A	39.03 %	26.78 %	19.01 %	12.44 %	1.29 %	1.16 %	0.29 %	-
C	19.19 %	24.55 %	39.24 %	12.71 %	3.15 %	0.92 %	0.19 %	-
D	21.79 %	35.59 %	20.62 %	12.71 %	5.03 %	3.60 %	0.67 %	-
E	36.71 %	30.37 %	19.87 %	10.12 %	1.98 %	0.69 %	0.25 %	-
F	25.16 %	22.58 %	40.52 %	8.66 %	1.85 %	1.08 %	0.14 %	-
G	11.33 %	12.48 %	16.60 %	9.18 %	34.93 %	0.44 %	15.04 %	-
H	22.15 %	28.73 %	33.04 %	8.73 %	4.94 %	1.98 %	0.42 %	-
I	29.35 %	15.73 %	47.36 %	6.67 %	0.11 %	0.61 %	0.18 %	-
K	29.44 %	30.48 %	24.40 %	10.98 %	3.35 %	0.88 %	0.47 %	-
L	33.45 %	26.11 %	28.59 %	10.03 %	0.69 %	0.95 %	0.18 %	-
M	34.57 %	24.24 %	28.80 %	8.77 %	1.67 %	1.76 %	0.20 %	-
N	17.19 %	33.47 %	23.49 %	9.35 %	12.50 %	3.24 %	0.76 %	-
P	6.03 %	38.41 %	-	52.89 %	-	2.67 %	-	-
Q	34.50 %	29.07 %	23.05 %	8.97 %	2.98 %	0.91 %	0.50 %	-
R	31.86 %	28.50 %	25.40 %	10.45 %	2.70 %	0.84 %	0.26 %	-
S	17.64 %	33.91 %	29.05 %	16.13 %	2.02 %	0.50 %	0.76 %	-
T	19.66 %	28.82 %	38.71 %	11.86 %	0.37 %	0.34 %	0.25 %	-
V	25.62 %	15.83 %	50.71 %	7.05 %	0.17 %	0.47 %	0.15 %	-
W	27.10 %	26.21 %	33.29 %	10.40 %	1.40 %	1.40 %	0.19 %	-
Y	22.26 %	24.16 %	40.25 %	9.78 %	2.14 %	1.18 %	0.24 %	-

Tableau 2.7 – *Statistiques sur la présence dans chaque domaine conformationnel des acides aminés ne précédant pas une proline.*

Acide-aminé	A	H	B	P	L	G	Q	Z
A	4.96 %	8.62 %	24.28 %	46.48 %	2.35 %	-	-	13.32 %
C	0.88 %	0.88 %	35.96 %	33.33 %	5.26 %	-	-	23.68 %
E	2.87 %	6.56 %	36.89 %	33.20 %	4.51 %	-	-	15.98 %
D	0.91 %	1.82 %	44.38 %	34.04 %	3.65 %	-	-	15.20 %
G	3.55 %	10.00 %	60.00 %	25.81 %	0.32 %	-	-	0.32 %
F	4.89 %	4.44 %	40.00 %	22.67 %	0.89 %	-	-	27.11 %
I	7.87 %	6.56 %	54.10 %	20.00 %	0.33 %	-	-	11.15 %
H	-	4.40 %	33.96 %	39.62 %	-	-	-	22.01 %
K	3.63 %	5.28 %	41.25 %	33.00 %	4.29 %	-	-	12.54 %
L	7.06 %	8.83 %	34.22 %	38.63 %	0.44 %	-	-	10.82 %
M	5.66 %	8.49 %	37.74 %	35.85 %	-	-	-	12.26 %
N	1.89 %	1.26 %	36.79 %	22.96 %	5.35 %	-	-	31.76 %
P	0.64 %	5.10 %	-	94.26 %	-	-	-	-
Q	2.75 %	7.14 %	42.31 %	30.77 %	3.30 %	-	-	13.74 %
R	1.94 %	4.85 %	44.17 %	31.07 %	2.43 %	-	-	15.53 %
S	2.37 %	3.05 %	39.32 %	43.39 %	2.37 %	-	-	9.49 %
T	4.22 %	4.22 %	49.40 %	31.63 %	1.51 %	-	-	9.04 %
V	7.61 %	4.89 %	59.51 %	17.12 %	-	-	-	10.87 %
W	5.17 %	5.17 %	43.10 %	25.86 %	-	-	-	20.69 %
Y	5.64 %	4.10 %	44.62 %	25.13 %	0.51 %	-	-	20.00 %

Tableau 2.8 – *Statistiques sur la présence dans chaque domaine conformationnel des acides aminés précédant une proline.*

L'approche développée pour le programme RamaDA montre donc là tout son potentiel en donnant l'opportunité de caractériser simplement les parties désordonnées d'une protéine. L'attribution conformationnelle prend en compte tous les acides aminés d'une protéine, on a donc une information locale sur toute la séquence. Les applications présentées ici transforment cette information locale en information globale sur une protéine ou une région d'une protéine : la validation de structure permet de prendre ou non en considération une partie désordonnée dans une étude et la détection d'hélices polyproline ou de signatures de domaines permet de mieux comprendre la ou les fonction(s) d'une protéine.

Cependant, on voit bien ici que le talon d'Achille de RamaDA réside dans l'obtention d'un fichier décrivant la structure de la protéine d'intérêt afin de l'analyser. Pour pallier à ce manque, ce qui sera souvent le cas dans le domaine des IDP, les tableaux de proportions de domaines conformationnels par acide-aminé peuvent, dans un premier temps, aider à comprendre le comportement de certaines régions de la protéine. Mais cela ne peut pas suffire pour l'étude complète d'une protéine. Voilà pourquoi nous avons développé un moyen d'obtenir l'attribution conformationnelle d'une protéine à partir de ses déplacements chimiques observés par RMN.

2.2 RamaDP (Ramachandran Domain Prediction)

Les déplacements chimiques sont les données les plus basiques recueillies en RMN et pourtant ils contiennent une information sur la structure de la protéine étudiée. Ils sont donc la base de nombreux logiciels comme Talos+ [58] ou Rosetta [59] et de nombreuses techniques comme le Chemical Shift Index (CSI) [60] qui, dans le cas des protéines structurées, extraient cette information pour proposer une structure. Dans le cas des IDP, les déplacements chimiques sont moins dispersés et une seule structure ne

suffit pas à décrire le comportement de la protéine en solution. La notion de propension à être dans une certaine conformation a été introduite notamment via le logiciel SSP (Secondary Structure Propensities) [18]. Ce logiciel, sur la base d'un CSI adapté aux IDP, permet de déterminer pour chaque acide-aminé d'une protéine sa propension à se trouver dans une hélice ou un feuillet.

Dans ce paragraphe, nous présenterons les limites de telles techniques du point de vue des IDP ainsi que l'approche globale découlant de RamaDA que nous avons mise en place pour obtenir une description plus complète d'une protéine ou de régions désordonnées à partir de leurs seuls déplacements chimiques.

2.2.1 À la convergence de SSP et Talos+

Il existe deux grandes tendances dans l'étude les protéines structurées à partir de leurs déplacements chimiques. Soit on cherche à localiser les éléments de structure secondaire, soit on veut réaliser une structure complète de la protéine, les deux possibilités n'étant pas incompatibles.

Pour localiser les éléments de structure secondaire, le CSI reste une référence. Il permet d'extraire l'information structurale contenue dans la dispersion des déplacements chimiques des atomes du squelette protéique (H_N, H_α, C_α, C_O, N) et de C_β. En plaçant un seuil dépendant de l'atome, il permet de récupérer une information ternaire hélice/feuillet/autre sur chaque acide-aminé de la protéine observée. Ce seuil correspond à la valeur appelée *random-coil* d'un acide-aminé non-inclus dans un élément de structure secondaire.

D'un autre côté, pour réaliser une structure complète d'une protéine à partir de ses déplacements chimiques, Talos+ [58] est un exemple de logiciel capable de donner pour chaque acide-aminé une valeur de ses angles dihèdres (φ, ψ). Grâce à un réseau neuronal, il repère dans sa base de données la valeur la plus probable pour le cas observé.

Toutes ces méthodes sont calibrées pour les protéines structurées et ne sont pas applicables directement aux IDP. Cependant, il existe un pendant IDP à CSI, le SSP [18]. Ce logiciel donne pour chaque acide-aminé sa propension à être dans une hélice ou dans un feuillet. Cela permet notamment de trouver des structures secondaires transitoires dans les sites de liaison de la protéine, lorsqu'il y en a. L'avantage de ce logiciel est qu'il redéfinit précisément les seuils *random-coil* appliqués à chaque atome.

En effet, l'ensemble des déplacements chimiques d'un acide-aminé dépend de sa conformation, qui modifie son environnement chimique, mais aussi de sa nature-même et de celle de ses successeurs et prédécesseurs dans la séquence. Dans ce cadre, pour n'observer que la dispersion des déplacements chimiques dûe à la conformation, il faut lever la dépendance liée à la nature des acides aminés. C'est ce que Schwarzinger et al. [61] proposent à travers un tableau de valeurs de *random-coil* et de corrections basées sur la séquence de la protéine. Comme la dispersion des déplacements chimiques pour les IDP est moindre que celle des protéines structurées, il est intéressant d'utiliser ces corrections pour augmenter la précision de l'étude de déplacements chimiques.

On voit bien ici que le référencement des déplacements chimiques par rapport au 0 ppm est très important. Le DSS (diméthylsilapentanesulfonate) ou le TMS (tetraméthylsilane) sont les références les plus précises utilisées. Le référencement permet d'affiner les connaissances statistiques sur la dispersion des déplacements chimiques et donc d'améliorer les prédictions des conformations. SSP s'appuie d'ailleurs largement sur une base de données extraite de RefDB (Re-referenced protein chemical shift DataBase) [62], une base de données de déplacemements chimiques re-référencés par rapport au DSS, afin d'affiner les valeurs de seuil hélice/feuillet du CSI. Cela rend SSP légitime du point de vue de l'étude des IDP.

De la même façon, le concept de Talos+ a été élargi aux IDP grâce à une approche de statistique bayesienne par Wang et al. [63]. La précision de Talos+ pour les angles

dihèdres étant de $\pm$ 10 à 15°, soit l'équivalent d'un peu moins que la surface couverte par le domaine γ que nous avons défini précédemment, le logiciel de Wang et al. prédit ici plus modestement un domaine conformationnel du Ramachandran parmi trois plutôt que sur les angles eux-mêmes. On peut ainsi obtenir la probabilité de présence d'un acide-aminé dans ces trois domaines d'après ses déplacements chimiques.

On peut regretter dans les approches citées ici qu'aucune description totale des IDP soit disponible. En effet, pourquoi ne pas prendre en compte tous les domaines conformationnels connus dans le Ramachandran pour avoir une vue d'ensemble ainsi qu'une vision dynamique de la protéine ? De plus, les bases de données de déplacements chimiques utilisées pour les études statistiques ne sont pas nécessairement référencées par rapport au 0 ppm. Cela fausse bien évidemment les résultats obtenus : les bases de données risquent de ne pas être correctes et les déplacements chimiques de la protéine étudiée peuvent de pas refléter exactement les conformations adoptées.

En utilisant certains concepts de ces logiciels, nous avons développé une nouvelle approche de la prédiction de structure à partir des déplacements chimiques, basée sur tous les domaines conformationnels définis pour RamaDA ainsi que leur statistique de présence. Cet outil, appelé RamaDP (pour **Rama**chandran **D**omain **P**rediction), a pour but de nous donner une vue d'ensemble des conformations adoptées par chaque acide-aminé.

2.2.2 Utilisation des probabilités bayesiennes

Pour la détermination la plus précise possible des propensions d'un acide-aminé à adopter les conformations du diagramme de Ramachandran, on souhaite utiliser toutes les données qu'on a à disposition. Ces données sont les déplacements chimiques du plus grand nombre d'atomes pertinents (H_N, H_α, C_α, C_O,N et C_β) possible, la nature de l'acide-aminé observé et la nature de l'acide-aminé suivant dans la séquence, sans

oublier les tableaux 2.7 et 2.8 qui donnent la probabilité de présence d'un acide-aminé dans les différents domaines conformationnels.

Théorie des probabilités conditionnelles

La méthode statistique la plus appropriée dans ce genre de cas est l'utilisation de probabilités conditionnelles. Par exemple, vous vous demandez qu'elle est la probabilité que votre boîte aux lettres soit pleine (évènement A). A priori, vous avez une chance sur deux qu'elle le soit. Mais si vous savez si le facteur est passé ou non (évènement B), alors la probabilité que votre boîte aux lettres soit pleine est sensiblement modifiée ! La connaissance d'un évènement B non-indépendant de A permet donc de modifier la probabilité de A, $P(A)$, pour coller au maximum avec la réalité. La probabilité $P(A \mid B)$, probabilité de A sachant B, est appelée probabilité conditionnelle et possède de nombreuses propriétés.

Le théorème de Bayes (équation 2.3) permet relier $P(A)$ et $P(A \mid B)$.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2.3}$$

De plus, la probabilité que A et B aient lieu ensemble, $P(A \cap B)$, peut être exprimée en fonction de probabilités conditionnelles grâce à l'équation suivante :

$$P(A \cap B) = P(B \mid A)P(A) = P(A \mid B)P(B) \tag{2.4}$$

La dernière propriété de ces probabilités qui sera utilisée ici est le théorème des probabilités totales. Si B peut être entièrement décrit par n sous-ensembles B_i qui ne se recoupent pas, alors :

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i) \tag{2.5}$$

Application à notre étude

Dans notre cas, la probabilité qu'on souhaite déterminer est la probabilité d'adopter la conformation C_i sachant qu'on a un acide-aminé A (précédant ou non une proline) ayant pour déplacements chimiques mesurés l'ensemble σ. Cette probabilité est notée $P(C_i \mid (A \cap \sigma))$ et peut être calculée grâce à l'application du théorème de Bayes qui la relie à la probabilité d'adopter cette conformation, $P(C_i)$, à celle d'être l'acide-aminé A et d'avoir les déplacements chimiques σ, sachant ou non si la conformation C_i est adoptée, $P(A \cap \sigma)$ et $P((A \cap \sigma) \mid C_i)$ (voir équation 2.6).

$$P(C_i \mid (A \cap \sigma)) = \frac{P((A \cap \sigma) \mid C_i)P(C_i)}{P(A \cap \sigma)} \tag{2.6}$$

De plus, la réunion des conformations C_i est une partition exacte de l'ensemble des conformations que peut adopter un acide-aminé. On peut donc passer de l'équation 2.6 à l'équation 2.7 grâce au théorème des probabilités totales.

$$P(C_i \mid (A \cap \sigma)) = \frac{P((A \cap \sigma) \mid C_i)P(C_i)}{\sum_j P((A \cap \sigma) \mid C_j)P(C_j)} \tag{2.7}$$

Pour chaque conformation, on peut aussi écrire

$$P((A \cap \sigma) \mid C_i)P(C_i) = P(\sigma \mid (A \cap C_i))P(A \mid C_i)P(C_i) \tag{2.8}$$

Or, on a déjà, par les tableaux 2.7 et 2.8, la probabilité de présence d'un acide-aminé dans les différents domaines conformationnels c'est-à-dire $P(C_i \mid A)$, on utilise donc une nouvelle fois le théorème de Bayes pour arriver à l'équation suivante :

$$P((A \cap \sigma) \mid C_i)P(C_i) = P(\sigma \mid (A \cap C_i))P(C_i \mid A)P(A) \tag{2.9}$$

De plus, l'ensemble des déplacements chimiques σ est composé de l'intersection de

sous-ensembles σ_k représentant chacun un atome. On verra au prochain paragraphe que ces ensembles n'ont pu être corrélés par manque de données, on considèrera donc ces ensembles comme indépendants. On obtient alors l'équation suivante :

$$P((A \cap \sigma) \mid C_i)P(C_i) = [\prod_k P(\sigma_k \mid (A \cap C_i))]P(C_i \mid A)P(A) \tag{2.10}$$

Lorsqu'on reporte l'équation 2.10 dans l'équation 2.7, on obtient alors :

$$\begin{aligned} P(C_i \mid (A \cap \sigma)) &= \frac{[\prod_k P(\sigma_k \mid (A \cap C_i))]P(C_i \mid A)P(A)}{\sum_j [\prod_k P(\sigma_k \mid (A \cap C_j))]P(C_j \mid A)P(A)} \\ &= \frac{[\prod_k P(\sigma_k \mid (A \cap C_i))]P(C_i \mid A)}{\sum_j [\prod_k P(\sigma_k \mid (A \cap C_j))]P(C_j \mid A)} \end{aligned} \tag{2.11}$$

À ce stade, nous avons d'ores et déjà $P(C_i \mid A)$. Il faut maintenant déterminer la loi de probabilité suivie par les déplacements chimiques en connaissant la nature de l'acide-aminé et sa conformation $P(\sigma_k \mid (A \cap C_i))$.

2.2.3 Distribution des déplacements chimiques

La BMRB (Biological Magnetic Resonance Bank) [64] est une banque de données regroupant les attributions RMN de milliers de protéines. Malheureusement, ces attributions ne sont pas toutes référencées par rapport au DSS ou au TMS, les déplacements chimiques ne peuvent donc pas être pris en compte dans une étude statistique comme celle-ci. Les fichiers re-référencés de la base de données RefDB leur ont donc été préférés.

Afin de constituer un ensemble statistiquement représentatif des déplacements chimiques existants, seules les protéines de la RefDB dont l'identité de séquence est inférieure à 30% ont été conservées. Le tri des séquences a été effectué par l'algorithme de Smith-Waterman. L'ensemble ainsi formé contient 257 fichiers soit 34 864 acides

aminés.

Pour réaliser l'analyse RamaDA, les fichiers PDB correspondant aux protéines étudiées ont été récupérés. La RefDB donne une équivalence avec les fichiers PDB mais il s'agit la plupart du temps de fichiers PDB nés de l'analyse par cristallographie ou non-déposés par les auteurs du fichiers BMRB. On a donc recherché les fichiers PDB les plus proches des fichiers RefDB, en préférant les fichiers déposés par les mêmes auteurs que le fichier BMRB initial et ceux issus d'une analyse par RMN car porteurs de plus d'informations pour RamaDA.

L'analyse RamaDA de tous ces fichiers nous a permis de classer les acides aminés selon leur conformation. La population de chaque acide-aminé ayant les six déplacements chimiques qui nous intéressent n'est malheureusement pas assez importante pour utiliser des matrices de covariance, les atomes ont donc été traités séparément. Ces matrices nous auraient permis de corréler les distributions de déplacements chimiques entre elles et ainsi gagner en précision pour la détermination des propensions.

Il est possible, à partir des distributions de déplacements chimiques par acide-aminé et par conformation, d'affiner ces distributions en utilisant les corrections de Schwarzinger et al. [61]. Ces corrections portent sur le type d'acide-aminé observés ainsi que sur les 4 acides aminés les plus proches en séquence. Elles permettent de calculer précisément la valeur des déplacements chimiques en s'affranchissant de l'effet de la séquence protéique. La distribution des déplacements chimiques est alors uniquement due à la conformation adoptée.

La différence de déplacements chimiques $\Delta\delta$ entre la valeur donnée par Schwarzinger et al. et la valeur expérimentale a été calculée pour chaque acide-aminé de top500. Un histogramme par sous-ensemble de top500, par acide-aminé, par conformation et par atome a été réalisé. L'allure de chacun de ses histogrammes a été approchée par des sommes de gaussiennes par la même méthode que RamaDA c'est-à-dire en mini-

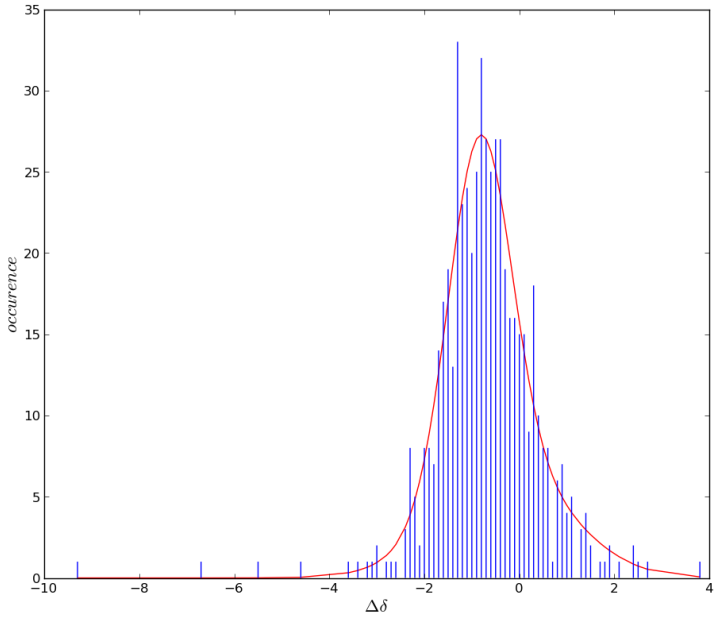

Figure 2.9 – *Distribution des écarts de déplacements chimiques $\Delta\delta$ des prolines du sous-ensemble Pro de top500, pour la conformation PPII et pour le carbone α. En bleu, l'histogramme extrait de top500. En rouge, la somme de deux gaussiennes qui représente au mieux l'histogramme.*

misant la somme des carrés des écarts entre les gaussiennes et l'histogramme. La figure 2.9 montre la distribution du domaine conformationnel PPII pour le carbone α de la proline. Dans ce cas-là, deux gaussiennes ont été nécessaires.

Pour les hélices, il est difficile voire impossible de différencier la conformation α de la conformation hélices-droite. De plus, dans les rares cas où cela est possible, le nombre de points recueilli n'est pas suffisant pour mener une étude statistique. Nous avons donc décidé de ne pas différencier les deux conformations. RamaDP regroupera donc toutes les conformations d'hélices sous le nom d'hélices-droite.

Le programme développé est téléchargeable à la même adresse que RamaDA : http ://ramada.u-strasbg.fr. Il est fourni avec la base de données regroupant les valeurs des différentes gaussiennes représentant les distributions de déplacements chimiques.

Il peut lire en entrée les fichiers de déplacements chimiques de type Sparky, Cara ou BMRB/RefDB et donne en sortie un histogramme cumulatif des propensions trouvées pour chaque acide-aminé dont les déplacements chimiques ont été fournis.

2.2.4 Application aux protéines structurées

Dans le cadre de RamaDP, on peut bien évidemment choisir la conformation dont la probabilité est la plus élevée afin de décrire un acide-aminé. Pour des protéines structurées, cette approche suffit *a priori* à décrire l'acide-aminé correctement. Afin de connaître la fiabilité de RamaDP quant à l'attribution d'un domaine conformationnel pour les protéines structurées, tous les fichiers de RefDB utilisés précédemment pour décrire les distributions de déplacements chimiques ont été prédits. L'attribution de RamaDP a été ensuite comparée à l'analyse RamaDA des fichiers PDB correspondants. Le tableau 2.9 contient les pourcentages d'acides aminés attribués au même domaine par RamaDA et RamaDP pour les différents sous-domaines et les différentes conformations.

On remarque que certains domaines sont très mal voire pas du tout retrouvés par RamaDP, c'est notamment le cas des domaines `Q` et `G`. Ceci est certainement dû aux statistiques bayesiennes qui privilègient les domaines conformationnels pour lesquels la propension d'occurence (ici, $P(C \mid A)$) est grande. Il n'est donc pas étonnant de bien prédire la grande majorité des conformations `H` et `B`.

Comme c'est le cas pour RamaDA, une correction à cette attribution a été introduite afin de prendre en compte les structures secondaires. Tout acide-aminé trouvé au milieu d'une séquence de quatre acides aminés ayant la même attribution `H`, `B` ou `P` est automatiquement modifié pour avoir la même attribution que ces voisins. Cette correction n'a pas été appliquée aux prolines puisqu'elle baissait significativement l'accord entre RamaDA et RamaDP. Le tableau 2.10 montre les nouveaux pourcentages d'acides aminés prédits dans les mêmes domaines par RamaDA et RamaDP.

Sous-domaine	Domaine	Accord RamaDA/RamaDP (correct / total)	
Tous	B	80.77 %	(7218 / 8937)
	P	31.82 %	(876 / 2753)
	H	73.04 %	(9273 / 12696)
	L	33.73 %	(364 / 1079)
	Q	16.94 %	(83 / 490)
	Z	9.62 %	(10 / 104)
	G	21.22 %	(66 / 311)
	tous	*68.58 %*	*(17890 / 26085)*
Général	B	82.53 %	(6605 / 8003)
	G	18.02 %	(40 / 222)
	H	74.43 %	(8803 / 11828)
	L	37.85 %	(190 / 502)
	P	28.56 %	(508 / 1779)
	Q	- - -	(0 / 274)
	tous	72.29 %	(16146 / 22334)
pre-Pro	B	75.99 %	(440 / 579)
	H	12.96 %	(14 / 108)
	L	5.88 %	(1 / 17)
	P	33.88 %	(104 / 307)
	Z	9.62 %	(10 / 104)
	tous	51.03 %	(569 / 1115)
Pro	G	33.33 %	(26 / 78)
	H	69.10 %	(293 / 424)
	P	43.76 %	(263 / 601)
	tous	52.76 %	(582 / 1103)
Gly	Q	38.43 %	(83 / 216)
	B	48.73 %	(173 / 355)
	H	48.51 %	(163 / 336)
	L	30.89 %	(173 / 560)
	P	1.52 %	(1 / 66)
	G	- - -	(0 / 11)
	tous	38.68 %	(593 / 1533)

Tableau 2.9 – *Pourcentage d'acides aminés attribués au même domaine par RamaDA et RamaDP dans l'ensemble de fichiers de RefDB utilisés pour la mise au point de RamaDP.*

Sous-domaine	Domaine	Accord RamaDA/RamaDP (correct / total)	
Tous	B	81.76 %	(7307 / 8937)
	P	30.95 %	(852 / 2753)
	H	75.14 %	(9540 / 12696)
	L	33.09 %	(357 / 1079)
	Q	14.90 %	(73 / 490)
	Z	7.69 %	(8 / 104)
	G	20.90 %	(65 / 311)
	tous	*69.78 %*	*(18202 / 26085)*
Général	B	83.34 %	(6670 / 8003)
	G	17.57 %	(39 / 222)
	H	76.45 %	(9043 / 11828)
	L	37.25 %	(187 / 502)
	P	27.15 %	(483 / 1779)
	tous	73.53 %	(16422 / 22334)
pre-Pro	B	74.96 %	(434 / 579)
	H	24.07 %	(26 / 108)
	L	5.88 %	(1 / 17)
	P	34.20 %	(105 / 307)
	Z	7.69 %	(8 / 104)
	tous	51.48 %	(574 / 1115)
Pro	G	33.33 %	(26 / 78)
	H	69.34 %	(294 / 424)
	P	43.76 %	(263 / 601)
	tous	52.86 %	(583 / 1103)
Gly	Q	33.80 %	(73 / 216)
	B	57.18 %	(203 / 355)
	H	52.68 %	(177 / 336)
	L	30.18 %	(169 / 560)
	P	1.52 %	(1 / 66)
	tous	40.64 %	(623 / 1533)

Tableau 2.10 – *Pourcentage d'acides aminés attribués au même domaine par RamaDA et RamaDP dans l'ensemble de fichiers de RefDB utilisés pour la mise au point de RamaDP avec correction consistant à changer l'attribution d'un acide-aminé trouvé entre 4 acides aminés décrivant une structure régulière.*

Avec ces résultats, on se rapproche des caractéristiques du logiciel Talos+. En effet, Talos+ est capable de re-prédire correctement 86.35% des acides aminés de sa base de données (là où la version précédente, Talos, n'en prédisait correctement que 72.31%) [58] mais la définition de bonne prédiction n'est pas la même que celle de RamaDP. Talos+ divise le Ramachandran en 3 régions : *alpha*, qui correspond à `H`, *positive-*φ, qui correspond à `L` et *beta*, qui englobe tous les autres domaines conformationnels. En prenant cette définition des régions du Ramachandran, RamaDP est, lui, capable de re-prédire correctement sa base de données à 78.81%.

Afin de comparer Talos+ et RamaDP sur un même ensemble de fichiers, nous avons choisi les 9 protéines utilisées par Shen et al. pour démontrer l'efficacité de Talos+. Les fichiers PDB (3E0E, 3E0H, 3EVX, 2HZ5, 1O5U, 1SRR, 1TTZ, 1VC1 et 2VCZ) ont été analysés par RamaDA et les fichiers BMRB (15849, 16097, 5357 et 15760), ou RefDB lorsqu'ils existaient (6546, 6210, 5899, 6263 et 5921), ont été utilisés par RamaDP pour prédire la conformation majoritaire de chaque acide-aminé. Les résultats de ces deux analyses ont été comparés.

Le tableau 2.11 regroupe les pourcentages de bonnes prédictions pour Talos+ et RamaDP. On remarque qu'au total avec les mêmes critères que Talos+, RamaDA prédit à peu près aussi bien cet ensemble de protéines (87.07% pour Talos+ contre 82.53% RamaDP). Au cas par cas cependant, on voit qu'un écart de plus de 10% peut exister en faveur de Talos+. En prenant des fichiers de déplacements chimiques référencés par rapport au DSS, on peut réduire cet écart. Le fichier BMRB 15760 associé à 2VZC peut être échangé avec le fichier BMRB 15899 qui, lui, est référencé. Le taux de bonnes prédictions pour cette protéine passe alors à 82.95% au lieu 79.53% avec les régions *alpha, beta et positive-*φ de Talos+ et de 76.38% à 80.31% quand on regarde tous les domaines conformationnels. De même, en échangeant le fichier BMRB 5357 avec le fichier RefDB 16006, on obtient 75.00% (respectivement, 82.95%) de bonnes prédictions au lieu

fichier PDB / fichier BMRB associé	3E0E / 15849	3E0H / 16097	3EVX / 6546	2HZ5 / 6210	1O5U / 5357	1SRR / 5899	1TTZ / 6363	1VC1 / 5921	2VZC / 15760	tous
Talos+	84.38 %	81.94%	87.26%	91.58%	93.10%	85.47%	90.28%	89.72%	91.41%	87.07%
RamaDP [a]	83.51%	86.27%	77.22%	84.44%	84.09%	84.30%	81.33%	83.64%	79.53%	82.53%
RamaDP [b]	74.23%	78.43%	69.62%	80.00%	71.59%	80.99%	73.33%	78.18%	76.38%	75.86%

fichier PDB / fichier BMRB associé	3E0E / 15849	3E0H / 16097	3EVX / 6546	2HZ5 / 6210	1O5U / 16006	1SRR / 5899	1TTZ / 6363	1VC1 / 5921	2VZC / 15899	tous
Talos+	84.38 %	81.94%	87.26%	91.58%	93.10%	85.47%	90.28%	89.72%	91.41%	87.07%
RamaDP [a]	83.51%	86.27%	77.22%	84.44%	82.95%	84.30%	81.33%	83.64%	82.95%	83.02%
RamaDP [b]	74.23%	78.43%	69.62%	80.00%	75.00%	80.99%	73.33%	78.18%	80.31%	76.64%

Tableau 2.11 – *Pourcentage de prédictions correctes pour un ensemble de 9 protéines, par Talos+ [58] et RamaDP ([a] en prenant les régions définies par Talos+, [b] en prenant les domaines conformationnels). Le deuxième tableau met en avant les nouveaux couples formés qui améliorent la prédiction de RamaDP (en rouge).*

de 71.59% (respectivement, 84.09%) en prenant en compte tous les domaines conformationnels (respectivement, avec les critères de Talos+). Les autres fichiers sont soit déjà correctement référencés soit ne possèdent d'*alter ego* référencé. Ces deux modifications montrent bien la nécessité d'avoir des déplacements chimiques correctement référencés par rapport au DSS pour RamaDP. Avec ces nouveaux fichiers, la proportion totale de bonnes prédictions augmente légèrement pour l'ensemble : 83.02% avec les critères de Talos+ et 76.64% pour tous les domaines conformationnels.

On peut donc considérer que Talos+ et RamaDP sont capables de prédire de façon similaire les conformations majoritaires des protéines à partir de leurs déplacements chimiques. Grâce notamment au référencement par rapport au DSS, RamaDP peut même être plus précis que Talos+ en donnant un domaine conformationnel plutôt qu'une large région du diagramme de Ramachandran.

2.2.5 Application aux IDP

Pour les protéines structurées, la conformation majoritaire des acides aminés suffit à décrire le comportement global. On vient de voir que RamaDP était capable de prédire cette conformation majoritaire avec une bonne précision. Cependant, dans le cas des IDP, l'ensemble des conformations présentes est nécessaire à la compréhension des protéines. La méthode de calcul utilisée par RamaDP permet aussi d'avoir un jeu complet de propensions (dont la somme est égale à 1) qui donne une vue d'ensemble de l'espace conformationnel parcouru par l'acide-aminé dans le cas d'une IDP. La figure 2.10 montre le résultat de RamaDP pour la protéine IB-5, vue précédemment, dont le spectre RMN a été attribué récemment (par Franck Paté, données non publiées) et le compare au résultat de la méthode SSP pour la même protéine. SSP donne un nombre compris entre -1 et 1, à l'instar du CSI, qui indique une propension à former une hélice (valeur positive) ou un feuillet (valeur négative).

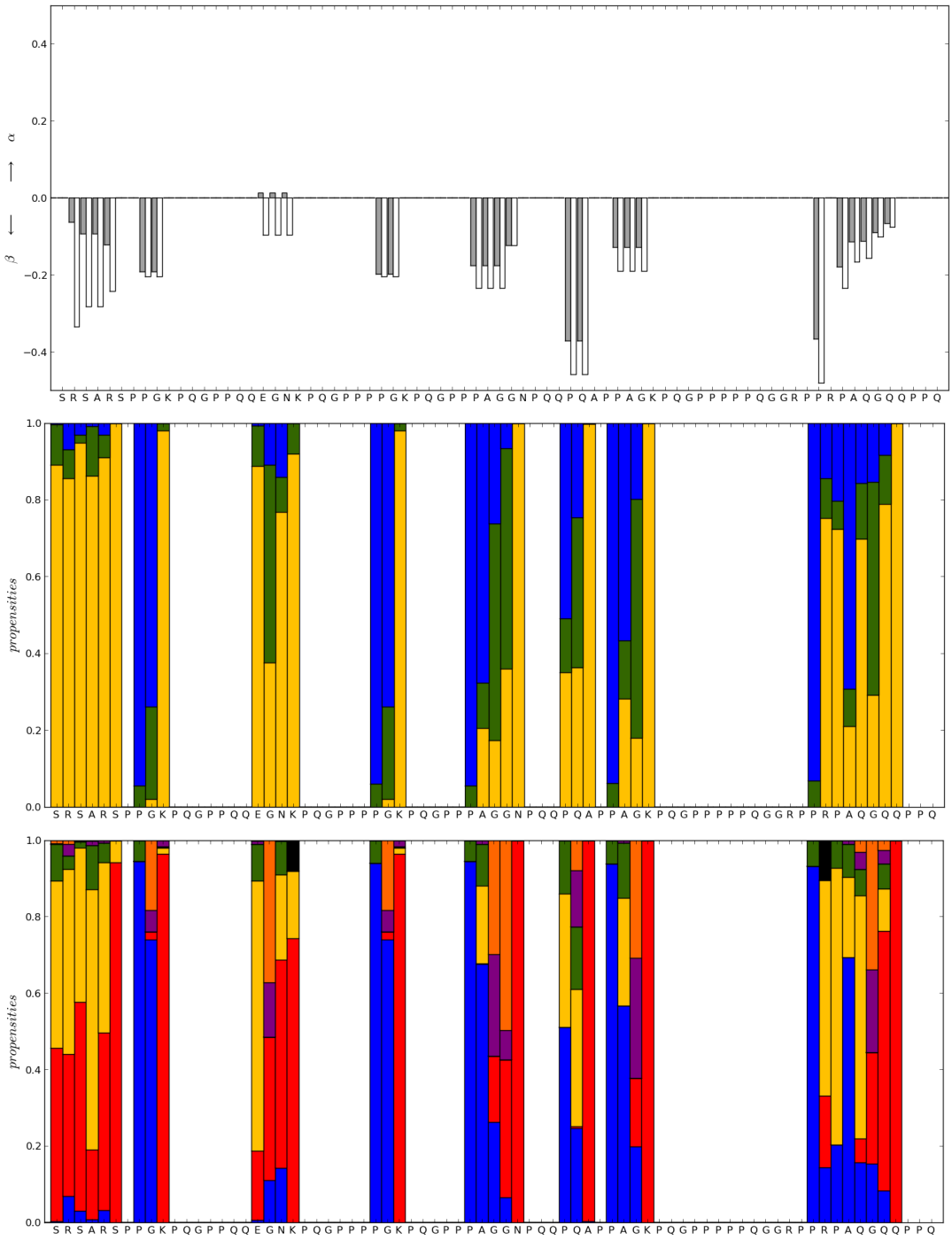

Figure 2.10 – *En haut, histogramme des résultats de SSP avec la correction de Schwarzinger et al. (en gris) et sans (en blanc). Une valeur positive montre une propension à être en hélice α alors qu'une valeur négative montre une propension à être en brin β. En bas, histogramme cumulatif des propensions données par RamaDP pour chaque acide-aminé et chaque domaine conformationnel (*`H`* en bleu,* `B` *en rouge,* `P` *en jaune,* `L` *en violet,* `G` *en vert,* `Z` *en noir et* `Q` *en orange). Au milieu, histogramme cumulatif des propensions données par RamaDP regroupées en trois ensembles : jaune pour conformations étendues (*`B` *et* `P`*), vert pour les conformations créant des coudes (*`L`*,* `G`*,* `Z` *et* `Q`*) et bleu pour* `H`*.*

Dans le cas de SSP, on a comparé le résultat de l'analyse en prenant en compte ou non la correction de Schwarzinger et al. Rappelons que cette correction permet de s'affranchir de la dépendance des déplacements chimiques à la nature-même des acides aminés et de leur environnement proche dans la séquence. Les valeurs de référence (ou *random-coil*) utilisées dans SSP ne prennent pas en compte cette dépendance c'est pourquoi on peut voir une différence significative entre les deux résultats de la figure 2.10. Dans notre cas, la correction se contente de diminuer les valeurs de propensions tout en gardant les mêmes tendances. Cependant, l'effet observé prouve bien que la séquence peptidique joue un rôle dans la distribution des déplacements chimiques.

D'après SSP, IB-5 aurait une certaine propension à former des feuillets β de façon transitoire ou, plus généralement, à adopter des conformations étendues (`H` et B). On sait effectivement que la protéine a tendance à être étendue mais la dimension fractale calculée précédemment (voir chapitre 1) donne à penser que le comportement d'IB-5 est beaucoup plus complexe que celui d'un brin β. Malheureusement, SSP est incapable de quantifier le temps passé dans la structure transitoire qu'il prédit car le score donné n'est pas borné. De plus, SSP ne peut pas traiter l'information provenant d'acides aminés précédant une proline, ce qui est le cas de près de 40% des acides aminés d'IB-5.

Avec RamaDP, l'information donnée est plus complète, elle n'est pas regroupée en un seul descripteur. On est en mesure d'avoir une vision dynamique totale de chaque acide-aminé. En effet, la propension donnée pour chaque domaine conformationnel correspond au temps passé par l'acide-aminé dans ce domaine. Pour les 6 premiers acides aminés d'IB-5, il y a accord avec SSP, les conformations étendues (`B` en rouge et `P` en jaune) dominent. Cependant, pour les autres acides aminés, il est difficile voire impossible de comparer les deux résultats car SSP offre une vision binaire là où RamaDP propose 7 possibilités.

En regroupant les conformations étendues (`B` et `P`) et les conformations créant des

coudes (`L`, `G`, `Z` et `Q`), on obtient l'histogramme au milieu de la figure 2.10. Grâce à cet histogramme, il est plus facile de comparer SSP et RamaDP. L'accord entre les résultats de ces deux logiciels ne se fait qu'aux extrémités de la protéine. En effet, à part dans quelques cas (N40 et A46 par exemple), SSP donne une propension à former des feuillets là où RamaDP trouve une propension de conformations étendues inférieure ou égale à 50%.

On sait, d'après les expériences précédentes, que la dimension fractale d'IB-5 est de 2.2. Or, si tous ses acides aminés étaient en conformation étendue, la dimension fractale de la protéine serait plus proche de 1.67 (protéine dénaturée) voire inférieure, on peut donc penser que RamaDP donne une description plus réaliste que SSP avec des acides aminés en conformation `H` ou en coude.

Le tableau 2.12 regroupe les valeurs de propensions trouvées par RamaDP pour IB-5.

Acide-aminé	H	B	P	L	G	Z	Q
S1	0.33%	45.33%	43.74%	0.08%	9.61%	–	0.91%
R2	6.93%	37.13%	48.36%	3.12%	3.48%	–	0.97%
S3	3.09%	54.54%	40.32%	0.18%	1.69%	–	0.17%
A4	0.86%	18.15%	68.05%	1.17%	11.55%	–	0.22%
R5	3.19%	46.45%	44.60%	0.59%	5.04%	–	0.13%
S6	0.00%	94.12%	5.80%	0.03%	–	0.05%	–
P8	94.51%	–	0.00%	–	5.49%	–	–
G9	73.95%	2.07%	–	5.71%	0.00%	–	18.27%
K10	0.00%	96.44%	1.51%	1.73%	–	0.32%	–
E18	0.70%	17.94%	70.74%	0.80%	9.55%	–	0.27%
G19	10.97%	37.53%	–	14.31%	0.00%	–	37.19%
N20	14.16%	54.54%	22.32%	0.13%	8.69%	–	0.15%
K21	0.01%	74.24%	17.63%	0.00%	–	8.11%	–
P28	94.04%	–	0.00%	–	5.96%	–	–
G29	73.95%	2.07%	–	5.71%	0.00%	–	18.27%
K30	0.00%	96.44%	1.51%	1.73%	–	0.32%	–
P36	94.42%	–	0.00%	–	5.58%	–	–
A37	67.65%	0.02%	20.44%	0.86%	10.88%	–	0.14%
G38	26.26%	17.31%	–	26.57%	0.00%	–	29.86%
G39	6.60%	36.04%	–	7.64%	0.00%	–	49.72%
N40	0.00%	99.99%	0.00%	0.00%	–	0.01%	–
P44	51.00%	–	34.97%	–	14.03%	–	–
Q45	24.67%	0.46%	35.83%	14.78%	16.39%	–	7.87%
A46	0.31%	99.69%	0.00%	0.00%	–	0.00%	–
P48	93.89%	–	0.00%	–	6.11%	–	–
A49	56.63%	0.02%	28.17%	0.52%	14.53%	–	0.13%
G50	19.87%	17.97%	–	31.40%	0.00%	–	30.76%
K51	0.00%	100.00%	0.00%	0.00%	–	0.00%	–
P65	93.27%	–	0.00%	–	6.73%	–	–
R66	14.36%	18.82%	56.41%	0.05%	–	10.37%	–
P67	20.34%	–	72.39%	–	7.27%	–	–
A68	69.33%	0.04%	20.91%	0.86%	8.71%	–	0.15%
Q69	15.70%	6.15%	63.59%	4.43%	6.93%	–	3.20%
G70	15.39%	29.08%	–	21.64%	0.00%	–	33.88%
Q71	8.38%	67.82%	11.02%	3.55%	6.60%	–	2.63%
Q72	0.02%	99.98%	0.00%	0.00%	–	0.00%	–

Tableau 2.12 – *Propensions trouvées par RamaDP pour chaque acide-aminé d'IB-5 dont l'attribution RMN a été faite, par domaine conformationnel.*

Certaines valeurs retiennent notre attention. Par exemple, les prolines 8, 28, 36, 48 et 65 ont une infime propension à se trouver en conformation `P`. Contrairement à ce qu'on pourrait penser de prime abord, l'abondante proportion de prolines au sein d'IB-5 ne mène donc pas à la formation de multiples petites hélices PPII. On a d'ailleurs l'impression en regardant l'histogramme simplifié de RamaDP que les prolines citées sont là pour empêcher la protéine d'être trop étendue en imposant majoritairement des conformations `H`.

L'analyse de RamaDP et la valeur de dimension fractale amènent bien aux mêmes conclusions c'est-à-dire à une protéine globalement étendue mais moins étendue qu'en conditions Θ. Si d_f permet d'avoir une vision globale du comportement de la protéine, RamaDP permet de comprendre la dynamique d'IB-5 acide-aminé par acide-aminé.

Chapitre 3
Détermination d'ensembles de conformations

Afin de décrire complètement une IDP, il paraît inévitable de connaître l'ensemble des conformations qu'elle adopte. Pour cela, nous avons développé un générateur de conformations capable de faire décrire à la protéine tout son espace conformationnel. Mais créer un ensemble de conformations représentatif du comportement de la protéine en solution n'est pas chose aisée. Il doit être créé à partir d'un groupe de conformations aléatoire puis trié sur plusieurs paramètres. Se posent alors deux problèmes : comment avoir un groupe de départ aussi aléatoire que possible et quels paramètres utiliser pour le filtrer ?

Dans ce chapitre, nous montrerons l'efficacité et la fiabilité de notre générateur aléatoire ainsi que l'étude de la polyproline et d'un modèle biologique. Les données expérimentales que l'on essaiera de reproduire proviennent de techniques spectroscopiques différentes.

3.1 Générateur de conformations aléatoires

3.1.1 Tirage aléatoire des angles dihèdres

Le diagramme de Ramachandran indique toutes les conformations possibles pour un dipeptide ainsi que la propension de ce dipeptide à les adopter. Nous avons donc choisi de nous baser sur ce diagramme pour générer les conformations des protéines

étudiées. Le modèle du diagramme de Ramachandran créé pour RamaDA a été repris afin de tirer des couples d'angles dihèdres (φ,ψ). Le générateur utilise les gaussiennes définies précédemment (voir chapitre 2) comme loi de probabilité de présence d'un couple (φ,ψ). Pour chaque domaine conformationnel, on est donc capable de tirer des angles sur la surface correspondante, en respectant leur propension à exister. Pour que chaque acide-aminé soit parfaitement représenté, les statistiques de présence dans chaque domaine conformationnel (voir tableaux 2.7 et 2.8) ont été utilisées. Ainsi, chaque acide-aminé n'ayant pas la même propension à se trouver dans chaque domaine conformationnel, on respecte leurs particularités.

La figure 3.1 montre un exemple de diagramme de Ramachandran pour un tirage de 10 000 couples (φ, ψ) pour une alanine selon qu'elle est précédée ou non d'une proline. On voit bien que la présence de l'alanine dans chaque domaine conformationnel est variable et que la présence ou non d'une proline a bien une influence sur le résultat du tirage. Le domaine des hélices-droite étant décrit par une gaussienne large, certains couples (φ, ψ) tirés pour l'alanine n'apparaissent pas dans le diagramme de l'ensemble pre-Pro de top500. Cependant, on peut penser qu'en instaurant une vérification des conflits stériques au sein d'une protéine, ces couples auront tendance à disparaître. L'accord avec les diagrammes de Ramachandran de top500 est néanmoins important et valide à la fois notre modèle gaussien et le caractère aléatoire du tirage.

3.1.2 Génération des conformères

Pour appliquer les angles dihèdres tirés à une chaîne peptidique, la bibliothèque python du logiciel MMTK [44] a été utilisée. Ce logiciel de modélisation moléculaire est capable de lire un fichier PDB puis de travailler sur la protéine et enfin de sortir le fichier PDB correspondant aux modifications appliquées. Dans notre cas, les angles dihèdres sont modifiés d'après le tirage aléatoire décrit précédemment. On vérifie ensuite que

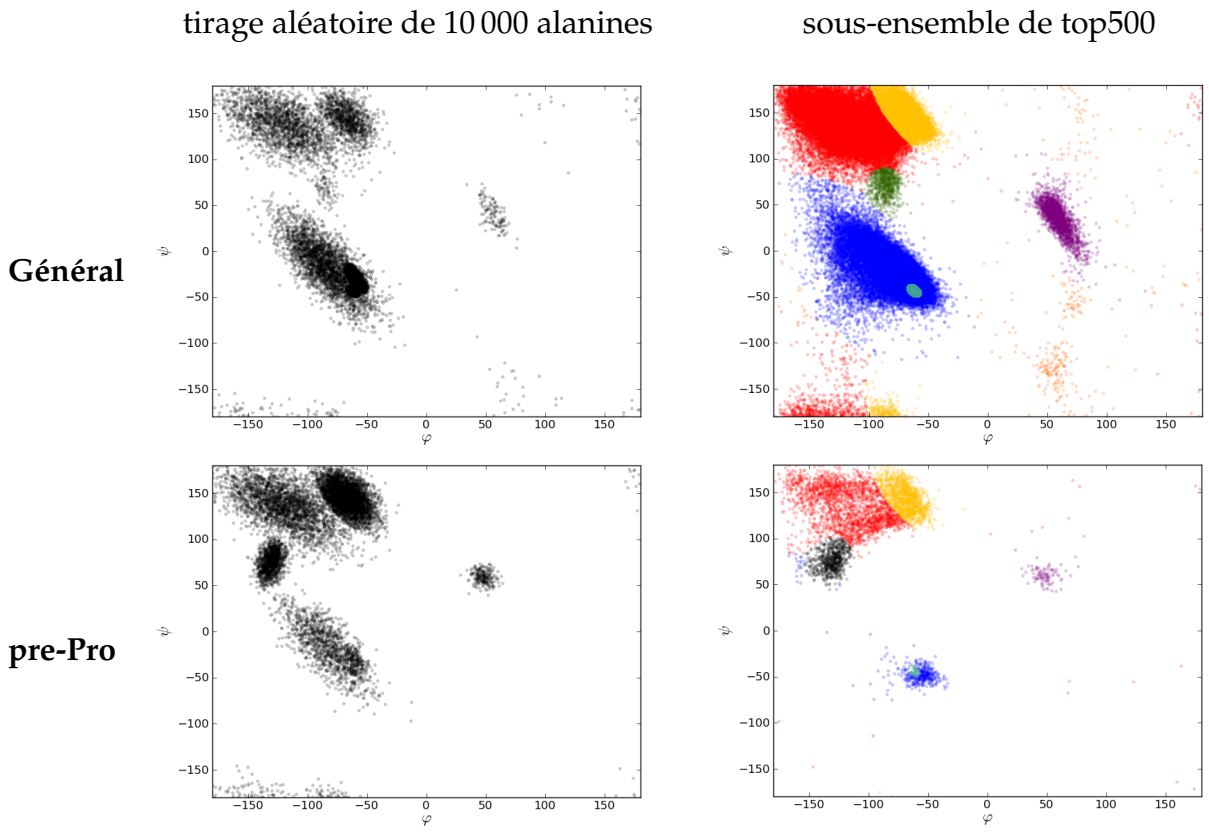

Figure 3.1 – *À gauche, les diagrammes de Ramachandran pour un tirage de 10 000 alanines, à droite, les diagrammes de Ramachandran de top500, tous acides-aminés confondus. Domaines conformationnels colorés comme suit : β en rouge, PPII en jaune, hélices-droite en bleu, α en turquoise, hélices-gauche en violet, γ en vert, ζ en noir et $PPII_R$ en orange. En haut, dans le cas du sous-ensemble Général, en bas, dans celui du sous-ensemble pre-Pro.*

le squelette de la protéine ne se replie pas sur lui-même. Si c'est le cas, on continue à travailler sur la molécule, sinon on recommence l'opération du début. Les atomes du squelette sont alors considérés comme figés dans l'espace et une minimisation d'énergie libre a lieu sur les chaînes latérales. Les atomes figés permettent de garder l'information du tirage aléatoire d'angle et les atomes des chaînes latérales se placent de façon à limiter les conflits stériques. Le champ de force choisi pour minimiser l'énergie libre du système est Amber99.

Une fois la minimisation réalisée on mesure les distances entre tous les atomes de

la protéine pour vérifier qu'il n'existe plus aucun conflit stérique. On décide arbitrairement de considérer que tout conformère ayant deux atomes séparés dans l'espace d'une distance inférieure à 0.9Å sera refusé. Les conformères gardés le sont sous la forme d'un fichier qui adopte la nomenclature des fichiers PDB. Le générateur répète l'opération jusqu'à ce que le nombre de conformères demandé soit atteint. De plus, la parallèlisation du programme permet de gagner du temps en séparant la tâche et en la lançant de front sur plusieurs processeurs du même ordinateur.

L'éviction des conformères ne répondant pas aux contraintes stériques imposées va d'ores et déjà engendrer un tri des conformères et créer un biais en favorisant les angles dihèdres qui évitent les conflits stériques. On peut donc se demander ce que représente l'ensemble généré.

Le diagramme de Ramachandran peut être interprété comme une carte de potentiels d'énergie. En effet, l'énergie libre due à la conformation dépend des angles dihèdres. Dans les régions *non-autorisées* du diagramme, le potentiel est quasiment nul alors qu'au coeur des domaines conformationnels, le potentiel est important.

Dans un modèle thermodynamique basé uniquement sur les contraintes stériques, le diagramme de Ramachandran ainsi que la sélection des conformères exempts de conflits stériques placent donc un ensemble généré infini à l'équilibre thermodynamique.

Un grand ensemble de conformères représentera donc un état à l'équilibre qui ne correspond pas obligatoirement à l'état de marche aléatoire de la protéine ni aux conditions Θ ou à un quelconque état défini jusqu'à présent. La sélection *a posteriori* permettra, elle, de se rapprocher le plus possible de l'état de la protéine en solution, c'est-à-dire à l'équilibre thermodynamique d'un système prenant en compte les contraintes stériques ainsi que les interactions intra- et intermoléculaires potentielles.

3.2 Outils d'analyse

Le but de notre générateur étant de créer à terme un ensemble de conformères reproduisant les propriétés de la protéine en solution, il est important de développer de nombreux outils pour analyser les ensembles créés ainsi que pour trier les conformères.

Pour vérifier que les angles tirés correspondent bien aux statistiques voulues, on peut tracer le diagramme de Ramachandran de chaque acide-aminé et récupérer les proportions de domaines conformationnels adoptés grâce à RamaDA. L'intérêt d'une telle analyse est bien évidemment de comparer les valeurs obtenues avec les valeurs des tableaux 2.7 et 2.8 qui regroupent les valeurs de propensions de chaque acide-aminé à se trouver dans un domaine conformationnel. On aura alors une idée du biais engendré par l'environnement et/ou l'opportunité de choisir la proportion de domaines conformationnels. Les matrices de distances entre les carbones α peuvent aussi montrer la structuration d'une protéine. Si des acides-aminés éloignés en séquence s'avèrent proches dans l'espace, alors le peptide montre une certaine structure.

De plus, certains paramètres mesurables par différentes techniques spectroscopiques peuvent être facilement calculés. C'est le cas du rayon de giration et de l'ensemble de distances à l'intérieur d'une protéine, accessibles par SAXS, ainsi que l'ensemble des distances bout-à-bout d'un échantillon, donné par des expériences de RPE (Résonance Paramagnétique Électronique).

La dimension fractale d'un ensemble de conformations aléatoires peut elle aussi être calculée à partir des deux méthodes décrites au chapitre 1. Ces deux méthodes sont complémentaires puisqu'elles convergent à des vitesses différentes (voir chapitre 1). d_{f_e} et d_{f_r} peuvent donc avoir des valeurs différentes pour un même ensemble.

Les paramètres présentés ici sont les plus simples que l'on peut extraire d'un ensemble. Il existe cependant des programmes capables de calculer des paramètres bien

plus complexes. Dans le cas de la RMN, il est possible, par exemple, de calculer les RDC (Residual Dipolar Coupling) [65] ou bien encore le résultat d'une expérience de Het-Sofast [66]. Bien sûr, il est intéressant de développer ces outils et de les coupler au générateur uniquement si on a recueilli les données expérimentales et que les paramètres sont pertinents pour le tri des conformères.

3.3 Cas de $\mathbf{A}_{50}$ et $\mathbf{P}_{20}$

Afin de tester notre générateur, deux exemples simples ont été traités, une polyalanine de 50 résidus et une polyproline de 20 résidus.

3.3.1 $\mathbf{A}_{50}$

30 000 conformères d'une polyalanine de 50 résidus (A_{50}) ont été générés sur 4 processeurs en 8 jours environ. Le rayon de giration R_g moyen de cet ensemble est de 14.75Å et sa dimension fractale d_{f_e} s'élève à 1.9. Comme prévu, l'ensemble ne correspond pas à un état précis de la protéine mais à un état d'équilibre thermodynamique d'un modèle basé uniquement sur les contraintes stériques.

De plus, la matrice des distances Cα-Cα, présentée en figure 3.2, confirme qu'aucune structuration n'est visible. L'analyse RamaDA des 30 000 fichiers générés montre que les statistiques de présence des différents domaines conformationnels sont bel et bien reproduites dans ce cas (voir figure 3.3). Le générateur est donc fiable tant au niveau de son caractère aléatoire qu'au niveau des statistiques du diagramme de Ramachandran.

Bien évidemment, dans notre cas l'alanine est un acide-aminé dont la chaîne latérale n'est pas très encombrante et qui peut adopter de nombreuses conformations, les statistiques de RamaDA correspondent donc bien aux statistiques de départ. Cependant, il est certain qu'une séquence contenant des acides-aminés encombrants et/ou dont

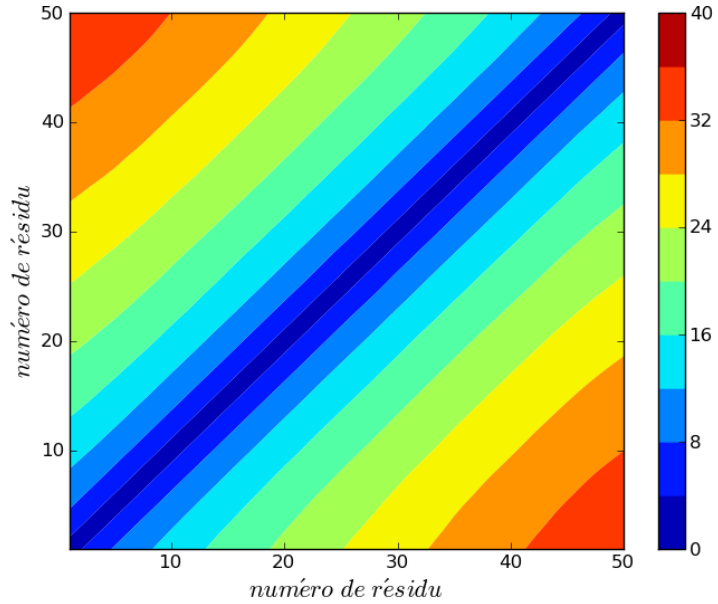

Figure 3.2 – *Matrice des distances Cα-Cα pour un ensemble de 30 000 conformères de poly-alanine. La distance moyenne entre deux Cα se lie sur l'échelle colorée donnée à droite. Chaque couleur représente un intervalle de 4Å. Sur la matrice, les zones sont parallèles les unes aux autres, montrant une absence de structuration de l'ensemble des peptides.*

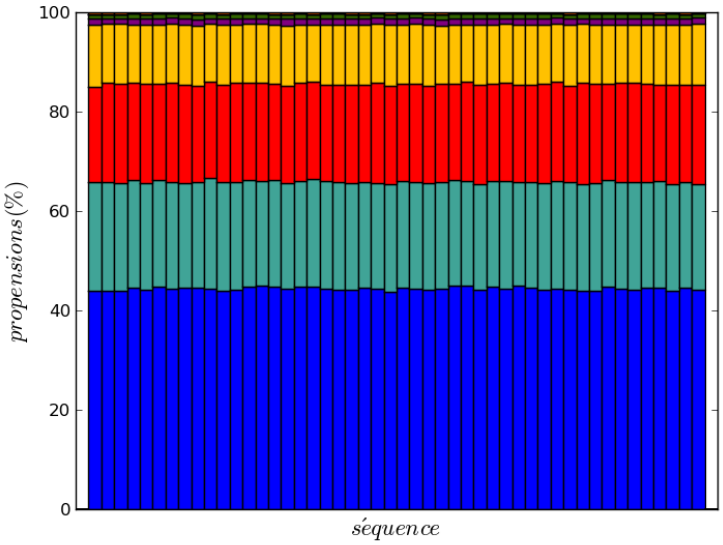

Figure 3.3 – *Analyse RamaDA pour un ensemble de 30 000 conformères de poly-alanine. Le code couleur est le même que précédemment : `A` en turquoise, `H` en bleu, `B` en rouge, `P` en jaune, `G` en vert, `L` en violet et `Q` en orange.*

l'éventail de conformations est réduit mènera à un ensemble de conformations contraint dont les statistiques RamaDA seront différentes des statistiques de départ à cause de la nature-même de l'environnement. Le cas de la polyalanine nous permet de nous assurer qu'aucun biais statistique n'existe, ce qui se passera avec les autres peptides générés sera alors uniquement la conséquence de leurs propriétés intrinsèques. Nous verrons par ailleurs le cas de la polyproline dans le paragraphe suivant.

La valeur 30 000 a été choisie arbitrairement pour que l'on soit certain que toutes les valeurs calculées auront convergé vers leur valeur à l'infini. Il est intéressant de connaître à partir de combien de conformères un ensemble reproduit déjà les paramètres d'un ensemble infini. Pour cela, des ensembles contenant de 100 à 30 000 conformères ont été créés. La figure 3.4 présente l'évolution du rayon de giration et de la dimension fractale calculés pour ces ensembles en fonction du nombre de conformères de l'ensemble. L'intervalle de confiance à 95% autour de la mesure de R_g, I, ne correspond pas à la distribution des valeurs observées dans l'ensemble des conformères mais à la précision de la mesure. Il est calculée à partir de l'écart-type de la distribution σ_d et du fractile à 0.975 de la loi de Student calculée avec $n-1$ degrés de liberté, $t_{0.975}(n-1)$ où n est le nombre de conformères dans l'ensemble (voir équation 3.1).

$$I = [R_g - \frac{t_{0.975}(n-1)\sigma_d}{\sqrt{n}}, R_g + \frac{t_{0.975}(n-1)\sigma_d}{\sqrt{n}}] \tag{3.1}$$

Dans le cas de A_{50}, quelle que soit la taille de l'ensemble, on observe que les valeurs de dimension fractale d_{f_e} et d_{f_r} sont différentes, une protéine de 50 acides-aminés n'est donc pas assez grande pour que les deux méthodes aient convergé. On remarque que toutes les grandeurs sont assez stables sur ce large intervalle et qu'elles atteignent leur valeur finale aux alentours de 1 000 conformères par ensemble.

La figure 3.5 montre cependant que 1 000 semble trop peu et que 2 000 conformères

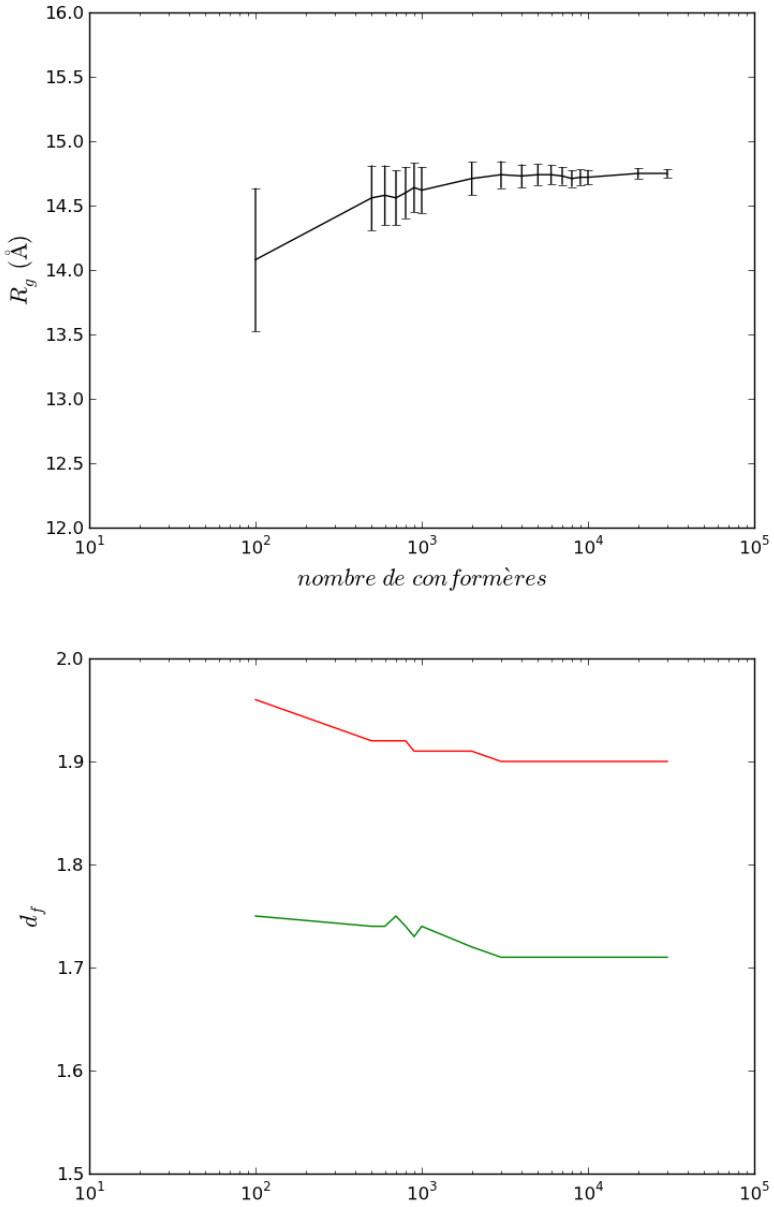

Figure 3.4 – *En haut, évolution du rayon de giration moyen (avec intervalle de confiance à 95%) selon le nombre de conformères présents dans un ensemble de poly-alanines. En bas, évolution de la dimension fractale calculée à partir des R_g (en rouge) et celle calculée à partir des distances bout-à-bout (en vert).*

seraient assez pour atteindre la stabilisation de la matrice des distances Cα-Cα et des statistiques RamaDA. La valeur minimale du nombre de conformères par ensemble

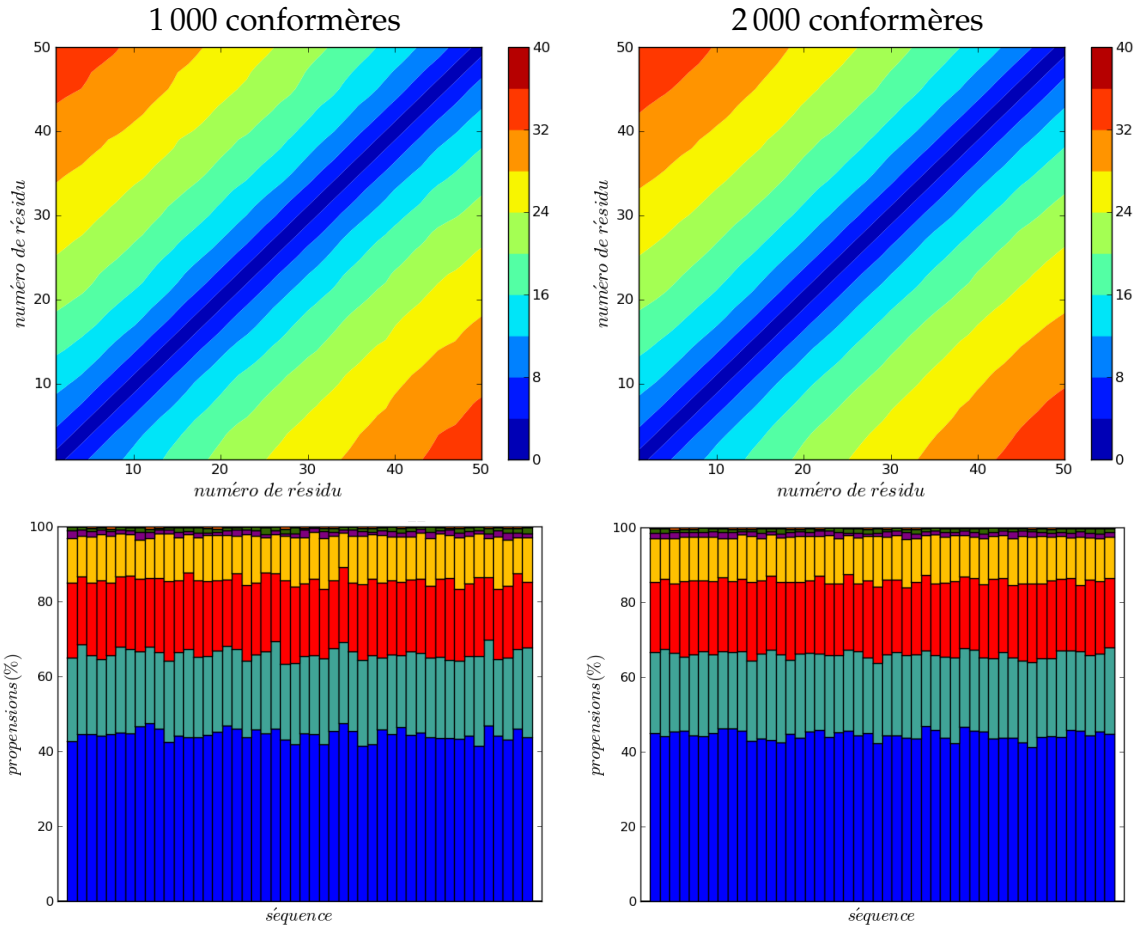

Figure 3.5 – *À gauche la matrice des distances Cα-Cα (en haut) et les statistiques RamaDA (en bas) sur un ensemble de 1 000 conformères. À droite les mêmes statistiques sur un ensemble de 2 000 conformères.*

donnée ici dépend évidemment du système sur lequel on travaille et une séquence plus contrainte nécessitera certainement moins de conformères pour être entièrement représentative du comportement de la molécule. Néanmoins, la polyalanine étant très peu contrainte, on a ici un ordre de grandeur acceptable pour tout type de peptides de cette taille.

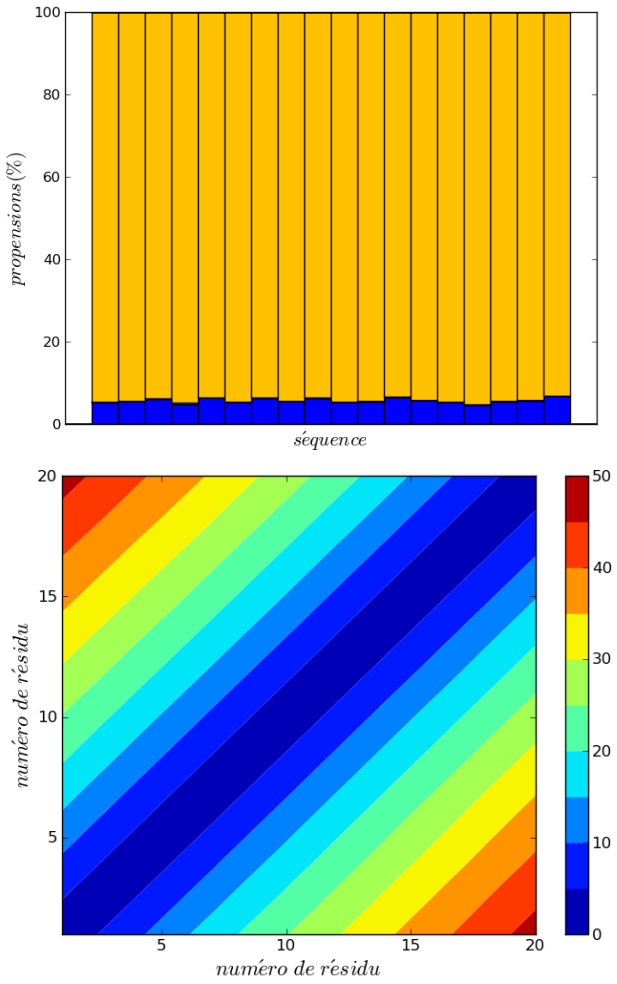

Figure 3.6 – *En haut, l'analyse RamaDA des 2 000 conformères de P20. En bas, leur matrice de distances Cα-Cα.*

3.3.2 $\mathbf{P_{20}}$

2 000 conformères d'une polyproline de 20 résidus (P_{20}) ont été générés sur 4 processeurs en 2 heures environ.

Le R_g moyen est de 15.78 $\pm$ 1.60Å sur l'ensemble. On remarque donc que la dispersion de l'ensemble est très faible. La matrice des distances Cα-Cα semble montrer qu'il n'y a aucune structuration de l'ensemble mais l'analyse RamaDA indique que, comme prévu par les statistiques de top500, plusieurs prolines consécutives mènent majoritairement à la formation d'hélices PPII (voir figure 3.6). Ces deux analyses sont donc

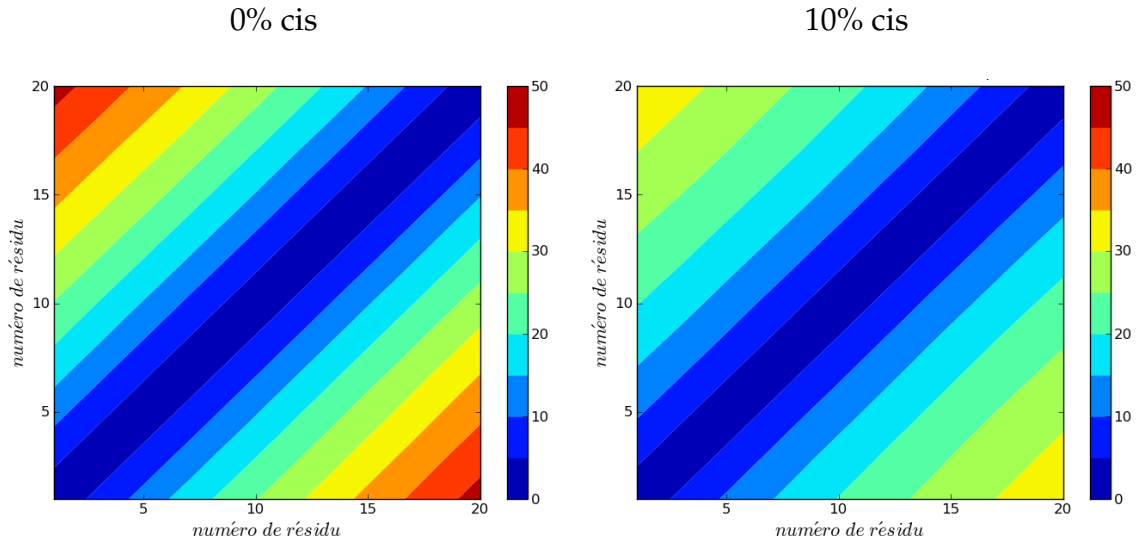

Figure 3.7 – *Matrice des distances Cα-Cα de l'ensemble avec 0% de cis-prolines (à gauche) et 10% de cis-prolines (à droite).*

complémentaires dans l'étude des hélices PPII.

La dimension fractale calculée n'est pas la même selon les deux méthodes car le peptide choisi est petit. df_r vaut 1.37 et df_e 1.12. On retrouve bien une dimension fractale faible et d_{f_e} est même très proche de la valeur mesurée par DOSY (d_f = 1.2). L'ensemble généré semble donc être d'ores et déjà relativement représentatif du comportement de la polyproline en solution.

Cependant, dans l'ensemble présenté ci-dessus, toutes les prolines sont en conformation trans. Il est intéressant de connaître l'effet de l'isomérisation cis-trans sur les paramètres mesurés par notre analyse. Pour cela, des ensembles de 2 000 conformères ont été générés avec une probabilité de 2%, 5% ou 10% d'être en conformation cis pour chaque proline.

Si l'analyse RamaDA des ensembles reste inchangée pour ces ensembles, la matrice des distances Cα-Cα indique le rapprochement des prolines entre elles (voir figure 3.7). Le tableau 3.1 regroupe les valeurs de dimension fractale et de rayon de giration calculées sur les ensembles et on peut noter que d_f augmente avec le pourcentage de cis-

Pourcentage de cis-prolines	d_{f_e}	d_{f_r}	R_g **(Å)**
0%	1.12	1.37	15.78 ± 1.60
2%	1.16	1.42	15.06 ± 2.01
5%	1.22	1.51	14.04 ± 2.36
10%	1.31	1.65	12.79 ± 2.45

Tableau 3.1 – *Valeurs de dimension fractale et de rayon de giration moyen en fonction de la proportion de cis-prolines dans l'ensemble étudié.*

Nombre de conformères	d_{f_e}	d_{f_r}	R_g **(Å)**
1 000	1.12	1.37	15.78 ± 1.60
2 000	1.12	1.37	15.76 ± 1.58

Tableau 3.2 – *Valeurs de dimension fractale et de rayon de giration en fonction du nombre de conformères dans l'ensemble étudié.*

prolines alors que le R_g moyen diminue et que l'ensemble des valeurs est plus large.

Ces résultats correspondent à ce que l'on peut effectivement attendre de l'isomérisation cis-trans. En effet, une cis-proline a tendance à casser les hélices PPII. Plus il y a de cis-prolines et moins le peptide ressemble à un cylindre. De plus, le peptide étant plus compact, son rayon de giration diminue.

On constate qu'en considérant que l'analyse RamaDA correspond aux conformations du peptide en solution, un taux de cis-prolines compris entre 2 et 5% permet de retrouver une dimension fractale de 1.2. Cette valeur est en accord avec des études menées précédemment par différentes techniques [31,38].

Les conformations des prolines étant à plus de 90% en `P`, il n'est pas nécessaire d'avoir 2 000 conformères par ensemble pour atteindre la convergence des paramètres calculés. Un ensemble de 1 000 conformères peut suffire à représenter l'ensemble des conformères de la polyproline. On peut d'ailleurs voir en figure 3.8 et dans le tableau 3.2 la comparaison des résultats obtenus pour des ensembles de 1 000 et 2 000 conformères dans le cas de 0% de cis-prolines. Mise à part une différence de l'ordre de 0.1% pour le rayon de giration, les autres analyses donnent les mêmes résultats.

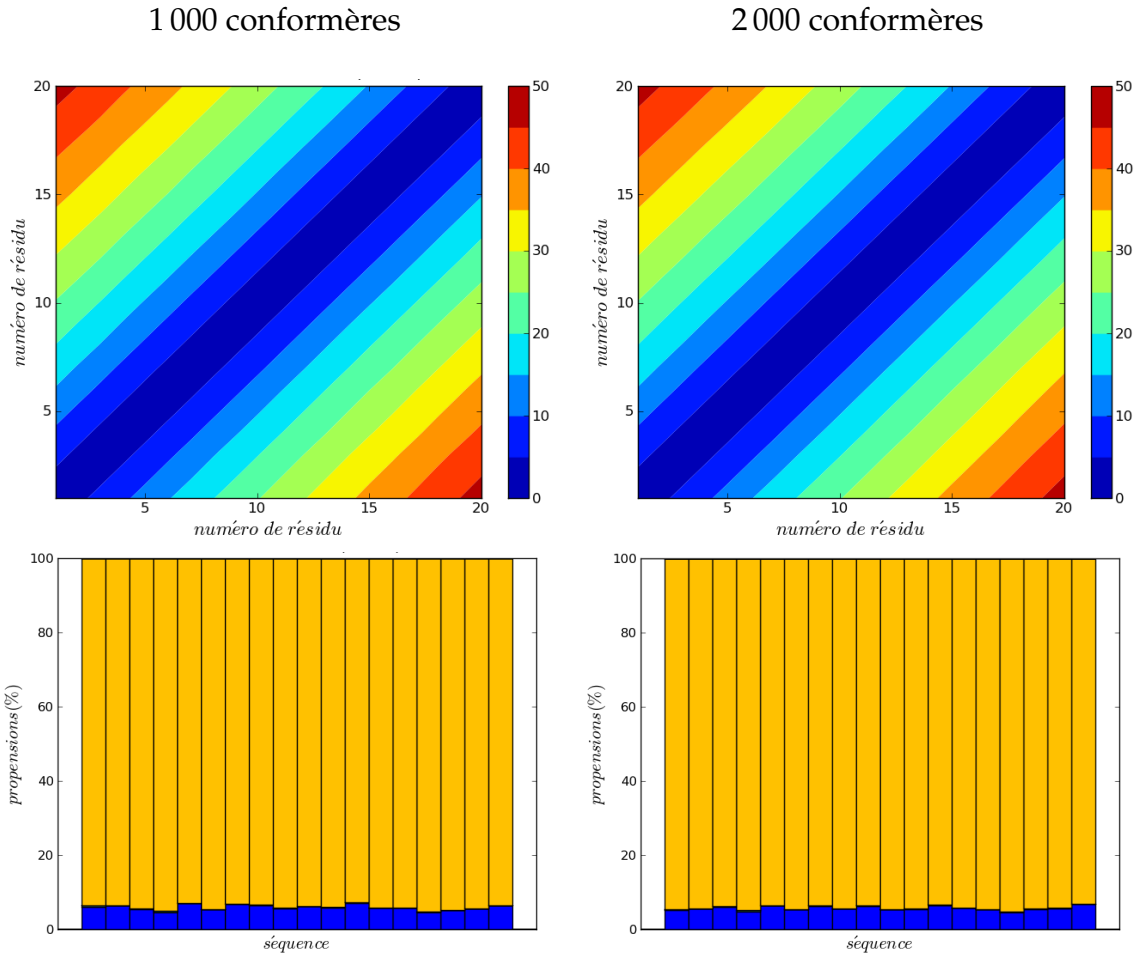

Figure 3.8 – *Comparaison des résultats obtenus en fonction du nombre de conformères. En haut, la matrice des distances Cα-Cα, en bas, l'analyse RamaDA.*

3.4 RARγ

3.4.1 Présentation du système

Les récepteurs nucléaires à l'acide rétinoïque (RAR : Retinoic Acid Receptors) sont des régulateurs de la transcription des gènes dans l'organisme. Ils sont composés de 3 grands domaines : le domaine de liaison à l'ADN (DBD : DNA Binding Domain), le domaine de liaison à l'acide rétinoïque (LBD : Ligand Binding Domain) et à tous les autres ligands et le domaine A/B désordonné. Si le mécanisme d'action du domaine

Figure 3.9 – *Positions relatives des domaines principaux de RARγ sur la séquence. Les sérines phosphorylables du domaine A/B sont indiquées. La séquence présentée ici est la séquence exacte des utilisés pour mesurer les K_d avec la vinexine β*

peptide	K_d (μM)
PI121 (non phosphorylé)	37 ± 9
PI120 (phosphorylé en S7)	134 ± 13
PI119 (phosphorylé en S9)	280 ± 17
PI118 (doublement phosphorylé)	544 ± 22

Tableau 3.3 – *K_d mesurés par RMN pour l'interaction entre un peptide issu de RARγ et la vinexine β (valeurs extraites de Lalevée et al. [67]).*

A/B n'est pas encore tout à fait établi, il est clair cependant que sa phosphorylation est essentielle pour déclencher la liaison de l'acide rétinoïque et l'activité de la protéine.

Plusieurs isotypes de ces récepteurs existent (α, β, γ) et de légers changements dans leurs séquence permet de les caractériser. L'isotype γ est particulier car il nécessite une double phosphorylation par deux molécules différentes pour son activation. La figure 3.9 montre les sérines du domaine A/B qui doivent être phosphorylées pour RARγ. En fait, l'activité de RARγ non-phosphorylé est bloquée par la vinexine β qui vient se fixer sur la région à phosphoryler avec un coefficient de dissociation (K_d) faible [35].

L'interaction entre la vinexine β et le domaine A/B de RARγ se fait par reconnaissance d'une hélice PPII par l'un des domaines SH3 de la vinexine β. Les domaines SH3 (Src Homology-3) sont connus en général pour se lier spécifiquement à ces hélices [34]. Des peptides synthétiques phosphorylés ou non, dont la séquence est présentée en figure 3.9, ont été utilisés pour mesurer précisement leur K_d avec la vinexine β par

peptide	nom utilisé
PPPPPSPPPPPRVYK non-phosphorylé	S
PPPPPSPPPPPRVYK phosphorylé	S*
PPPPPPPPPPPRVYK	P
PPPPPAPPPPPRVYK	A
PPPPPVPPPPPRVYK	V

Tableau 3.4 – *Séquence et nom des peptides-modèles.*

RMN [67]. Les valeurs de K_d sont recueillies dans le tableau 3.3. Selon le degré de phosphorylation, les peptides se lient avec plus ou moins d'affinité sur la vinexine β. Plus le peptide est phosphorylé et moins il se lie à la vinexine.

Afin d'étudier cette interaction plus précisément et comprendre le rôle de la phosphorylation des sérines, une modélisation de l'interaction des peptides avec la vinexine β a été réalisée (par Marc Quinternet, données non publiées). Il semblerait que l'extrémité C-terminale des peptides servent à l'ancrage sur la vinexine β, ce qui laisse les sérines dans le solvant et sans contact avec le partenaire. Il est difficile à ce stade de comprendre exactement la base moléculaire de la modulation de l'affinité observée.

Nous avons donc développé un modèle simplifié de ces peptides où nous nous sommes intéressés dans un premier temps à la seule sérine 79 ainsi qu'aux potentielles hélices PPII formées sur les peptides naturels. La phosphorylation de la sérine 79 semble avoir un plus grand effet que celle de la sérine 77, au vu des K_d, c'est pour cela qu'elle lui a été préférée. La séquence des peptides-modèles est présentée dans le tableau 3.4. Pour étudier l'impact de l'hélice polyproline nous avons aussi remplacé la sérine par une proline, une alanine ou une valine. En effet, l'alanine et la valine sont les acides-aminés qui adoptent respectivement le plus souvent et le moins souvent la conformation PPII lorsqu'ils se trouvent devant une proline (voir tableau 2.8, serine exclue). Les acides-aminés en C-terminal (RVYK) ont été gardés car ils sont *a priori* un point d'ancrage important pour l'interaction avec la vinexine β. Notre modèle se base donc sur deux hypothèses :

peptide	K_d (μM)	K_d des peptides naturels équivalents (μM)
P	33 ± 2	—
V	53 ± 1	—
A	45 ± 2	—
S	120 ± 8	37 ± 9 (PI121) et 134 ± 13 (PI120)
S*	408 ± 14	280 ± 17 (PI119) et 544 ± 22 (PI118)

Tableau 3.5 – *K_d recueillis par RMN pour les peptides synthétiques avec la vinexine β, comparés, si possible aux peptides naturels équivalents.*

l'extrémité C-terminale du peptide lui permet de s'accrocher à la vinexine β et la propension d'hélices PPII dans le peptide module l'affinité.

3.4.2 Mesures réalisées

Les peptides correspondants ont été synthétisés. Afin de connaître exactement le contenu des échantillons et pour se débarrasser des sels laissés par la synthèse des peptides et leur purification par HPLC, les solutions de peptides ont été dialysés contre de l'eau milliQ ou contre de l'eau milliQ avec 2% de glycérol pour les échantillons réservés au SAXS. Malgré la perte d'une partie non négligeable de produit, il apparaît que cette méthode est la seule méthode fiable. Les solutions récupérées sont nos solutions-mères de peptides.

Les K_d de l'interaction entre chacun des peptides et la vinexine β ont été mesurés par RMN (par Christian Koehler, données non publiées). Ces résultats sont réunis dans le tableau 3.5. Les K_d recueillis sont proches de ceux des peptides ayant exactement la séquence de RARγ. Il le sont d'autant plus si l'on ne considère que les peptides naturels phosphorylés en sérine 77. Cela valide donc notre modèle et montre que seuls l'ancrage RVYK et la conformation en PPII sont nécessaires à l'interaction entre les peptides et la vinexine β.

Des expériences de DOSY ont été réalisées sur les peptides dans l'eau milliQ à une concentration de 1mM à 300K avec 1mM de Tris comme référence. Le spectromètre uti-

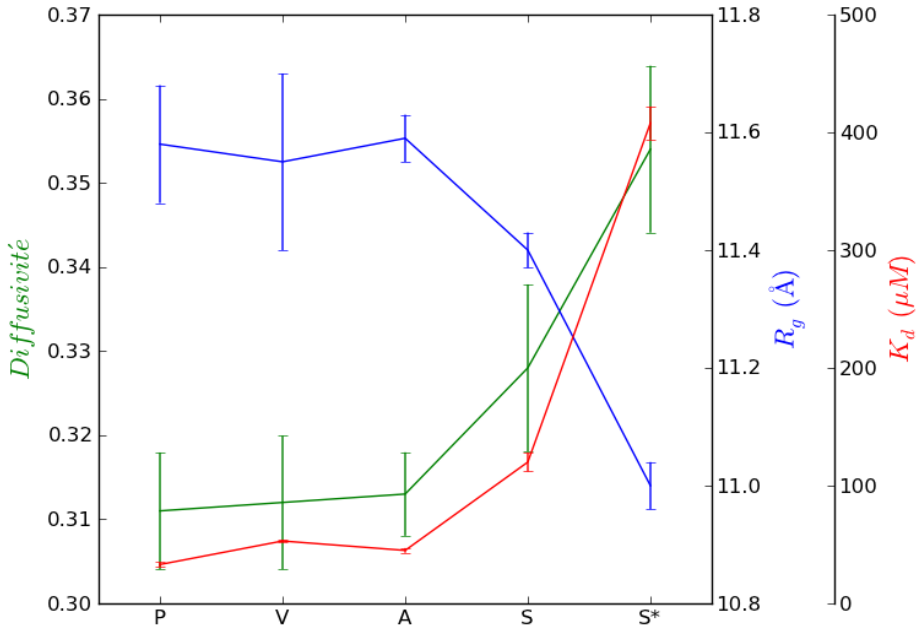

Figure 3.10 – *Graphique des valeurs de diffusivité mesurées par DOSY en référence au Tris (en vert) et valeurs de R_g mesurées par SAXS (en bleu). Comparaison avec l'évolution des K_d mesurés pour l'interaction entre les peptides et la vinexine β (en rouge).*

lisé est un spectromètre Bruker Avance 600MHz équipé d'une cryo-sonde. Les conditions des expériences sont les mêmes que précédemment (voir chapitre 1). Les spectres ont été traités par NMRNotebook.

Les expériences de SAXS ont été réalisées sur la ligne SWING du centre de rayonnement synchrotron Soleil de Gif-sur-Yvette. Les peptides ont été étudiés sur une gamme de concentration allant de 0.25mM à 3mM dans l'eau milliQ avec 2% de glycérol, nécessaire à la protection contre la radiolyse. Le blanc utilisé pour ces expériences est le bain de dialyse des peptides.

La figure 3.10 regroupe les coefficients de diffusion mesurés par DOSY et les rayons de giration obtenus par l'application de l'approximation de Guinier sur les courbes de SAXS grâce à l'outil Autorg (version 2.4.1) de la suite ATSAS [68]. On voit bien que

les résultats convergent, le peptide S* a le plus petit R_g et donc le plus grand D alors que P a le grand R_g et donc le plus petit D. Toutes les valeurs concordent. Mais le plus intéressant est que l'évolution des données de diffusion correspond à l'évolution des K_d mesurés. Cela confirme donc que la conformation du peptide est vraiment la clé de la reconnaissance par la vinexine β et que sa séquence n'a pas un grand impact.

Cependant, les prédictions de conformations faites sur A et V semblent être incorrectes puisque l'affinité de la vinexine β pour ces peptides est la même et que leurs conformations semblent aussi identiques au vu des diffusivités et des R_g. L'environnement fortement dominé par les prolines tend donc a priori à contraindre A et V à adopter les mêmes conformations, à gommer leurs différences. Créer un ensemble de conformations ayant les mêmes propriétés que les peptides nous en apprendra sûrement plus sur leur comportement.

3.4.3 Analyse des conformations

Les figures 3.11 et 3.12 et le tableau 3.6 contiennent les résultats obtenus pour des ensembles de 1 000 conformères de chaque peptide. Ces ensembles ont été créés en 1h environ chacun, sur 4 processeurs en parallèle. Comme indiqué précédemment, la dimension fractale calculée par la méthode des distances bout-à-bout, pour le peptide P, est très proche de la valeur de dimension fractale mesurée au chapitre 1 pour l'hélice PPII. L'analyse RamaDA nous confirme par ailleurs que toutes les prolines ont une forte propension à se trouver en conformation `P`. De plus, l'accord entre les valeurs calculées et les valeurs réelles de R_g est très grand. Les ensembles de conformations générées sont donc très proches du comportement réel des peptides. Encore une fois la fiabilité du générateur est confortée.

Les matrices de distance Cα-Cα sont toutes quasiment identiques. On y repère la partie C-terminale du peptide qui est purement aléatoire ainsi que les deux hélices

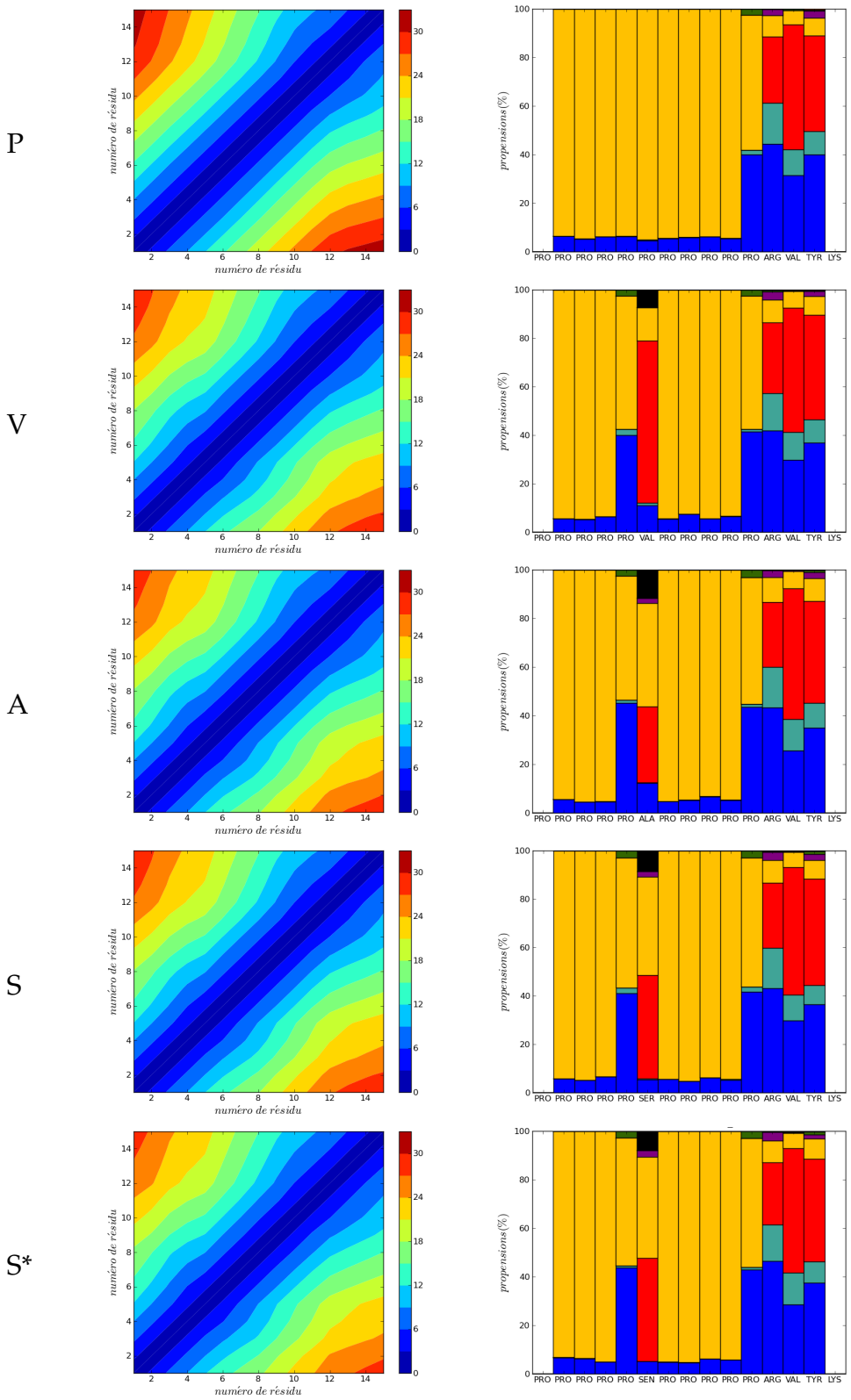

Figure 3.11 – *Résultats obtenus sur des ensembles de 1 000 conformères. À gauche, les matrices de distances Cα-Cα et à droite, l'analyse RamaDA.*

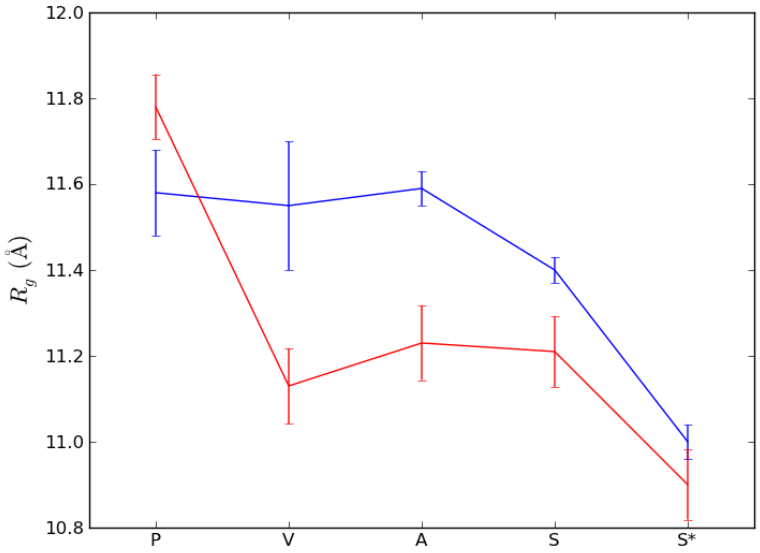

Figure 3.12 – *Comparaison entre les valeurs de R_g calculées (en rouge), avec l'intervalle de confiance à 95%, et les valeurs expérimentales (en bleu).*

peptide	R_g **calculé (Å)**	R_g **expérimental (Å)**	d_{f_e}
P	11.77 ± 1.20	11.58 ± 0.10	1.17
V	11.13 ± 1.40	11.55 ± 0.15	1.24
A	11.23 ± 1.38	11.59 ± 0.04	1.25
S	11.21 ± 1.31	11.40 ± 0.03	1.24
S*	10.90 ± 1.30	11.00 ± 0.04	1.25

Tableau 3.6 – *R_g et d_{f_e} calculés pour des ensembles de 1 000 conformères de chaque peptide. Comparaison avec les R_g issus des expériences de SAXS.*

PPII séparée par un acide-aminé (sauf dans le cas de P). Cependant, en prenant comme référence le peptide P, on peut représenter le logarithme du rapport entre les matrices de distances des peptides et de la référence choisie. Le figure 3.13 donne ces matrices pour V, A, S et S*. Pour ces quatre peptides, on remarque aisément la mobilité apportée au peptide par l'acide-aminé central (résidu 6) par rapport à la proline. Cet effet est le plus important pour S*. De plus, on voit que la seconde partie du peptide, à partir du résidu 7, n'est quasiment pas altérée par la substitution de la proline centrale en une

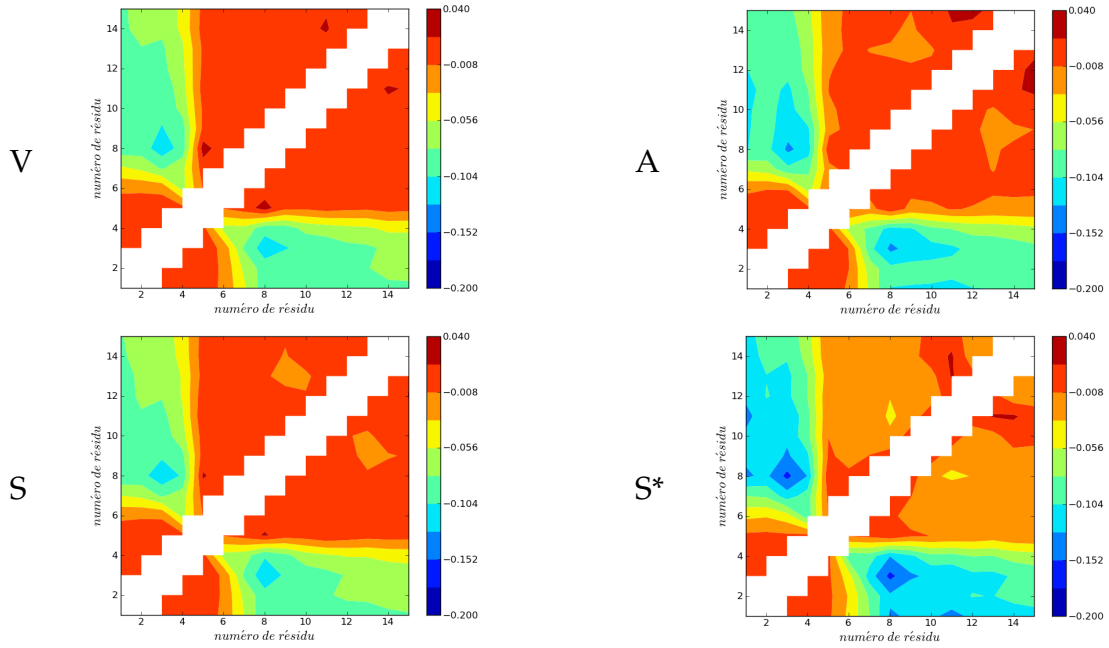

Figure 3.13 – *Matrices des logarithmes des distances Cα-Cα des peptides rapportée à P.*

valine mais semble plus sensible à la sustitution en une sérine phosphorylée. Les distances entre ces acides-aminés sont sensiblement modifiées et s'écartent des distances mesurées dans le cas du peptide P c'est-à-dire qu'elles semblent s'écarter d'une hélice PPII.

Pourtant, peu de différences sont notables au niveau des analyses RamaDA et les propensions à être en conformation `P` sont les mêmes pour tous les peptides (voir figure 3.11). De même, la séquence d'ancrage RVYK des peptides adopte toujours les mêmes conformations et ne semble pas affecté par le changement de l'acide-aminé placé entre les deux hélices PPII. Cet acide-aminé adopte les conformations qui lui sont propres, la valine ayant moins de propension à se trouver en conformation `P` que les autres. On peut toutefois noter que la proline placée avant l'acide-aminé central et celle avant la séquence d'ancrage ont une forte tendance à adopter une conformation `H`, qu'elles n'auraient pas si elles étaient placées ailleurs.

Les différences les plus marquées se situent au niveau des R_g. En effet, si les conformations semblent très proches, le moindre changement de conformation d'un acide-aminé se traduit au niveau de son rayon de giration. Les valeurs trouvées sont assez proches des mesures réalisées en SAXS et au vu de la précision des mesures et du calcul on peut donc considérer que ces ensembles sont représentatifs du comportement des peptides en solution. En sélectionnant les conformères pour faire en sorte que la proline placée avant l'acide-aminé central (la valine, l'alanine ou la sérine) ait les mêmes statistiques RamaDA que ces voisines, on peut augmenter les R_g de A et V et les rapprocher encore plus des valeurs expérimentales. Malheureusement, l'ensemble sélectionné devient alors trop petit pour qu'on ait pleinement confiance dans les résultats.

Cependant, les peptides générés ici ne comportent que des prolines en conformation trans. Or, les études menées précédemment sur l'isomérisation cis-trans des polyprolines montrent un effet important. On peut donc étudier l'impact de cette isomérisation sur les peptides étudiés. Pour cela, des ensembles de 1 000 conformères de peptides ont été générés, en 1h chacun sur 4 processeurs en parallèles, avec différentes probabilités de trouver une proline en conformation cis (2%, 5% et 10% pour chaque proline).

L'influence de l'isomérisation sur les R_g des peptides est mise en évidence par la figure 3.14. Comme précédemment, l'augmentation du nombre de prolines-cis fait grandement décroître les rayons de giration de chaque peptide. Même 2% de prolines-cis suffisent à observer une chute des R_g. Ce résultat est prévisible puisque les prolines-cis ont tendance à casser les hélices PPII et à rendre le peptide plus compact.

Il paraît maintenant évident que l'ajout de l'isomérisation cis-trans à notre étude la rend plus complexe. Néanmoins, les ensembles de départ semblent déjà correspondre en grande partie à des ensembles représentatifs et la sélection de conformères particuliers simplifie la tâche. Il faudrait maintenant générer un plus grand nombre de conformères pour chaque peptide afin de pouvoir sélectionner sans craindre de perdre

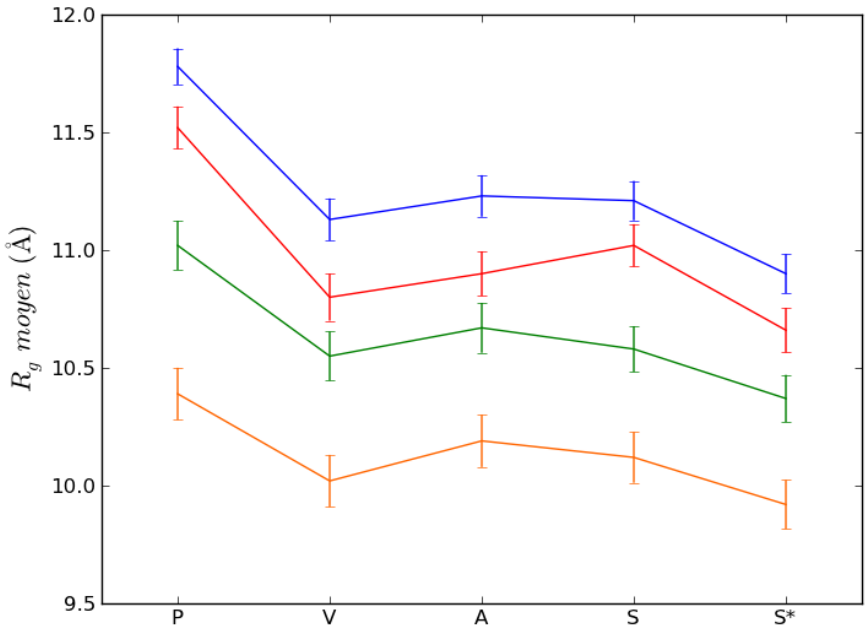

Figure 3.14 – *Graphique des valeurs de R_g pour des probabilités différentes d'avoir une cis-proline : 0% (en bleu), 2% (en rouge), 5% (en vert) et 10% (en orange), avec intervalle de confiance à 95%.*

de l'information et étudier les différentes conformations correspondant aux ensembles finaux pour mieux aborder l'interaction entre ces peptides et la vinexine β. Acquérir de nouvelles données de différents types serait aussi appréciable afin de sélectionner les conformères sur plusieurs critères, des expériences de CD sont d'ailleurs en cours sur ces peptides.

Avec cet exemple, le générateur montre sa fiabilité en reproduisant la dimension fractale d'une hélice PPII mais aussi sa sensibilité à tout changement conformationnel comme l'apparition de cis-prolines. Les peptides jugés au départ désordonnés s'avèrent particulièrement structurés avec une très forte propension à former des hélices polyprolines.

En accumulant des données expérimentales variées et en développant les outils pour les prendre en compte dans le tri des conformères, on a bon espoir de pouvoir former des ensembles de conformères représentatifs du comportement de la protéine étudiée en solution.

Conclusions et Perspectives

L'étude présentée ici a pour but d'amener de nouveaux outils à l'analyse des IDP, de générer de nouvelles façons d'aborder ces protéines afin d'en comprendre toutes les facettes.

Dans une première partie, nous avons vu que, bien qu'une protéine ne soit pas un objet fractal, il était possible de définir et de mesurer une dimension fractale. La détermination de coefficients de diffusion par DOSY sur une famille de molécule homogène suffit à l'obtention de d_f de manière précise. Ainsi nous avons pu mesurer la dimension fractale de l'hélice PPII. Le modèle utilisé jusuqu'à présent pour ce motif est une hélice étendue et rigide. La valeur de d_f obtenue, 1.2, surprend. En effet les hypothèses contraignantes de la théorie de Flory ne permettent pas d'expliquer cette valeur alors qu'elle correspond bien à un cylindre dans la théorie des dimensions fractales. Aucun des modèles proposés (dimérisation, super-hélice de collagène, calcul de R_h en fonction de N) ne permet pour le moment d'expliquer cette valeur. Établir un corollaire des équations de Flory pourrait être une piste intéressante.

Quoi qu'il en soit, avoir une famille homogène de peptides peut être considéré comme l'un des points faibles de la méthode de mesure de d_f pour les protéines. Or, les IDP ont une complexité de séquence moindre que celle des protéines structurées, on peut ainsi extraire des peptides de la séquence de la protéine pour créer cette fa-

mille. La dimension fractale est donc autant adapté à l'hélice PPII, structurée, qu'aux IDP. Grâce à cette approche, nous avons déterminé que la dimension fractale de l'IDP salivaire riche en proline IB-5, dont la séquence est très répétitive, était de 2.2. En comparant cette valeur à celle de la littérature, il est possible de rattacher la protéine aux peptides β-amyloïdes, qui, eux aussi, sont très étendus et ont une forte propension à la précipitation.

Dans une deuxième partie, nous avons développé un modèle du diagramme de Ramachandran afin de s'en servir pour attribuer à chaque acide-aminé son domaine conformationnel. 8 domaines conformationnels ont été définis dans les régions *autorisées* du diagramme. Certains d'entre eux n'apparaissent que pour une catégorie précise d'acides-aminés comme par exemple le domaine `Z` qui ne contient que des acides-aminés précédant une proline dans la séquence peptidique. Le domaine `L` quant à lui, n'existe pas chez les prolines et n'est pas le même pour les glycines, les acides-aminés précédant une proline ou les autres acides-aminés.

Le modèle gaussien mis en place pour représenter le diagramme de Ramachandran, appelé RamaDA, a de nombreuses applications pour les protéines structurées. Il permet de valider les structures de protéines aussi précisément que WHAT_CHECK, de repérer la grande majorité des structures secondaires ou bien de déterminer les signatures de domaines comme celle des mains EF, qu'il repère à 96.5%. Cependant, les conformations des acides-aminés étant les mêmes pour les protéines structurées et pour les IDP, RamaDA s'applique tout aussi bien aux IDP. Le talon d'Achille de RamaDA réside dans l'obtention de fichiers de type PDB pour les IDP susceptibles d'être analysés.

Grâce aux statistiques réalisées par RamaDA sur l'ensemble de protéines top500, il est possible de connaître la population d'acides-aminés présents dans chaque domaine conformationnel. Avec la théorie des probabilités conditionnelles et une analyse fine des distributions de déplacements chimiques par acide-aminé et par domaine confor-

mationnel, un logiciel de prédiction des domaines conformationnels, appelé RamaDP, a été écrit.

Pour des déplacements chimiques correctement référencés par rapport au DSS, RamaDP est capable de prédire correctement près de 70% des conformations majoritaires de chaque acide-aminé d'une protéine. La fiabilité du logiciel est donc faite pour les protéines structurées, pour qui la conformation majoritaire suffit à les décrire. Mais le véritable potentiel de RamaDP réside dans sa description dynamique globale de chaque acide-aminé. En effet, chaque acide-aminé à une propension à se trouver dans chacun des 8 domaines conformationnels et RamaDP est capable de les prédire ensemble. IB-5 a fait l'objet d'une prédiction de conformations par RamaDP. Après comparaison avec SSP, il s'avère que RamaDP fournit une information plus détaillée et plus en accord avec la valeur de dimension fractale mesurée précédemment.

Enfin, dans une dernière partie, un générateur de conformations aléatoires a été créé à partir du modèle du diagramme de Ramachandran développé pour RamaDA. Après avoir prouvé que le tirage des angles dihèdres réalisé est bien aléatoire et respecte les statistiques de présence des acides-aminés dans les domaines conformationnels, nous avons démontré sur l'exemple de la poly-alanine qu'il n'y avait aucun biais statistique sur l'ensemble des conformères créés. Un nombre minimum de conformères par ensemble a été donné à titre indicatif pour s'assurer de la fiabilité des résultats.

De nombreux outils ont été développés pour engranger autant d'information que possible sur les ensembles générés : calculs de R_g, de d_f, des distances Cα-Cα... Toutes ces informations nous ont servi pour l'étude préliminaire de l'interaction entre des peptides issus du récepteur nucléaire RARγ et la vinexine β. À partir de 1 000 conformères de chacun des 5 peptides-modèles choisis et de données expérimentales de RMN et de SAXS, nous avons pu valider notre modèle et commencer à analyser les différences et les similitudes observées entre les peptides. Malheureusement, les ensembles générés sont

pour l'instant trop petits pour pouvoir sélectionner certaines conformations et étudier l'impact de telles sélections sur R_g et d_f par exemple.

L'un des avantages de ce générateur de conformations aléatoires pour l'étude des IDP est le fait qu'il résout le problème posé par l'obtention de fichiers analysables par RamaDA. On peut ainsi avoir une vision complète de la protéine, acide-aminé par acide-aminé. De plus, avec ce générateur, on a accès aux conformères rares aussi facilement qu'aux conformères majoritaires, ce qui est difficile à atteindre par modélisation moléculaire. Avec des ensembles de conformères plus grands, il sera possible, avec tous les outils développés, d'étudier l'interaction entre RARγ et la vinexine β plus finement et de comprendre le rôle des sérines phosphorylées dans cette interaction.

De même, il serait intéressant d'étudier IB-5 par le biais de ce générateur maintenant que sa dimension fractale a été déterminée et qu'une grande partie des conformations de ces acides-aminés a été prédite par RamaDP. Pour cette protéine aussi, il est important de connaître son comportement en solution afin de comprendre comment elle lie les polyphénols.

Plus généralement, toute IDP pour laquelle un nombre de données expérimentales important est disponible peut être étudiée par ce biais, même s'il faut tout d'abord créer les outils susceptibles de filtrer les ensembles de conformères pour faire correspondre les paramètres calculés avec la réalité. De plus, la création d'une banque de données de dimensions fractales permettrait de savoir s'il est possible de relier des intervalles de valeurs avec certaines propriétés comme cela semble être le cas.

L'étude des IDP n'en est qu'à ses balbutiements, notre étude aura ouvert de nouvelles pistes de recherche. En s'appuyant sur les outils développés ici, le vaste monde des IDP nous ouvre encore un peu plus ces portes.

Bibliographie

[1] Roca M, Messer B, Hilvert D, Warshel A : **On the relationship between folding and chemical landscapes in enzyme catalysis**. *Proc Natl Acad Sci USA* 2008, **105**(37) :13877–82.

[2] Dyson HJ, Wright PE : **Intrinsically unstructured proteins and their functions**. *Nat Rev Mol Cell Biol* 2005, **6**(3) :197–208.

[3] Gusella JF, MacDonald ME : **Molecular genetics : unmasking polyglutamine triggers in neurodegenerative disease**. *Nat Rev Neurosci* 2000, **1**(2) :109–15.

[4] Goedert M : **Alpha-synuclein and neurodegenerative diseases**. *Nat Rev Neurosci* 2001, **2**(7) :492–501.

[5] Goedert M : **Tau protein and the neurofibrillary pathology of Alzheimer's disease**. *Trends Neurosci* 1993, **16**(11) :460–5.

[6] Teilum K, Olsen JG, Kragelund BB : **Functional aspects of protein flexibility**. *Cell. Mol. Life Sci.* 2009, **66**(14) :2231–47.

[7] Midic U, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN : **Protein disorder in the human diseasome : unfoldomics of human genetic diseases**. *BMC Genomics* 2009, **10 Suppl 1** :S12.

[8] Smith LJ, Fiebig KM, Schwalbe H, Dobson CM : **The concept of a random coil. Residual structure in peptides and denatured proteins**. *Fold Des* 1996, **1**(5) :R95–106.

[9] Morozova LA, Haynie DT, Arico-Muendel C, Dael HV, Dobson CM : **Structural basis of the stability of a lysozyme molten globule**. *Nat Struct Biol* 1995, **2**(10) :871–5.

[10] Song J, Guo LW, Muradov H, Artemyev NO, Ruoho AE, Markley JL : **Intrinsically disordered gamma-subunit of cGMP phosphodiesterase encodes functionally relevant transient secondary and tertiary structure**. *Proc Natl Acad Sci USA* 2008, **105**(5) :1505–10.

[11] Hughson FM, Wright PE, Baldwin RL : **Structural characterization of a partly folded apomyoglobin intermediate**. *Science* 1990, **249**(4976) :1544–8.

[12] Dunker AK, Obradovic Z : **The protein trinity–linking function and disorder**. *Nat Biotechnol* 2001, **19**(9) :805–6.

[13] Dunker AK, Silman I, Uversky VN, Sussman JL : **Function and structure of inherently disordered proteins**. *Curr Opin Struct Biol* 2008, **18**(6) :756–64.

[14] Hazy E, Tompa P : **Limitations of Induced Folding in Molecular Recognition by Intrinsically Disordered Proteins**. *ChemPhysChem* 2009, **10**(9-10) :1415–1419.

[15] Edwards RJ, Davey NE, Shields DC : **SLiMFinder : a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins**. *PLoS ONE* 2007, **2**(10) :e967.

[16] Davey NE, Shields DC, Edwards RJ : **Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery**. *Bioinformatics* 2009, **25**(4) :443–50.

[17] Davey NE, Travé G, Gibson TJ : **How viruses hijack cell regulation**. *Trends in Biochemical Sciences* 2011, **36**(3) :159–69.

[18] Marsh JA, Singh VK, Jia Z, Forman-Kay JD : **Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein : implications for fibrillation**. *Protein Sci* 2006, **15**(12) :2795–804.

[19] He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK : **Predicting intrinsic disorder in proteins : an overview**. *Cell Res* 2009, **19**(8) :929–49.

[20] Konrat R : **The protein meta-structure : a novel concept for chemical and molecular biology**. *Cell. Mol. Life Sci.* 2009, **66**(22) :3625–39.

[21] Jensen MR, Salmon L, Nodet G, Blackledge M : **Defining Conformational Ensembles of Intrinsically Disordered and Partially Folded Proteins Directly from Chemical Shifts**. *J Am Chem Soc* 2010.

[22] Dewey TG : *Fractals in molecular biophysics*. Oxford University press 1997.

[23] Flory PJ : *Principles of polymer chemistry*. Ithaca, New York 1953.

[24] Augé S, Schmit PO, Crutchfield CA, Islam MT, Harris DJ, Durand E, Clemancey M, Quoineaud AA, Lancelin JM, Prigent Y, Taulelle F, Delsuc MA : **NMR measure of translational diffusion and fractal dimension. Application to molecular mass measurement**. *The journal of physical chemistry B* 2009, **113**(7) :1914–8.

[25] Wilkins DK, Grimshaw SB, Receveur V, Dobson CM, Jones JA, Smith LJ : **Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques**. *Biochemistry* 1999, **38**(50) :16424–31.

[26] Stejskal EO, Tanner JE : **Spin diffusion measurements : spin echoes in the presence of a time-dependant field gradient**. *The Journal of chemical physics* 1965, **42** :5.

[27] Trefi S, Gilard V, Balayssac S, Malet-Martino M, Martino R : **The usefulness of 2D DOSY and 3D DOSY-COSY (1)H NMR for mixture analysis : application to genuine and fake formulations of sildenafil (Viagra)**. *Magnetic resonance in chemistry : MRC* 2009.

[28] Kakinoki S, Hirano Y, Oka M : **On the stability of polyproline-I and II structures of proline oligopeptides**. *Polym Bull* 2005, **53**(2) :109–115.

[29] Bochicchio B, Tamburro AM : **Polyproline II structure in proteins : identification by chiroptical spectroscopies, stability, and functions**. *Chirality* 2002, **14**(10) :782–92.

[30] Kelly MA, Chellgren BW, Rucker AL, Troutman JM, Fried MG, Miller AF, Creamer TP : **Host-guest study of left-handed polyproline II helix formation**. *Biochemistry* 2001, **40**(48) :14376–83.

[31] Vila J, Baldoni H, Ripoll D, Ghosh A, Scheraga H : **Polyproline II helix conformation in a proline-rich environment : A theoretical study**. *Biophys J* 2004, **86**(2) :731–742.

[32] Cubellis MV, Caillez F, Blundell TL, Lovell SC : **Properties of polyproline II, a secondary structure element implicated in protein-protein interactions**. *Proteins* 2005, **58**(4) :880–92.

[33] Rath A, Davidson AR, Deber CM : **The structure of "unstructured" regions in peptides and proteins : role of the polyproline II helix in protein folding and recognition**. *Biopolymers* 2005, **80**(2-3) :179–85.

[34] Mayer BJ : **SH3 domains : complexity in moderation**. *J Cell Sci* 2001, **114**(Pt 7) :1253–63.

[35] Bour G, Gaillard E, Bruck N, Lalevée S, Plassat JL, Busso D, Samama JP, Rochette-Egly C : **Cyclin H binding to the RARalpha activation function (AF)-2 domain directs phosphorylation of the AF-1 domain by cyclin-dependent kinase 7**. *Proc Natl Acad Sci USA* 2005, **102**(46) :16608–13.

[36] Stryer L, Haugland RP : **Energy transfer : a spectroscopic ruler**. *Proc Natl Acad Sci USA* 1967, **58**(2) :719–26.

[37] Doose S, Neuweiler H, Barsch H, Sauer M : **Probing polyproline structure and dynamics by photoinduced electron transfer provides evidence for deviations from a regular polyproline type II helix**. *Proc Natl Acad Sci USA* 2007, **104**(44) :17400–5.

[38] Best RB, Merchant KA, Gopich IV, Schuler B, Bax A, Eaton WA : **Effect of flexibility and cis residues in single-molecule FRET studies of polyproline**. *Proc Natl Acad Sci USA* 2007, **104**(48) :18964–9.

[39] Dolghih E, Ortiz W, Kim S, Krueger BP, Krause JL, Roitberg AE : **Theoretical Studies of Short Polyproline Systems : Recalibration of a Molecular Ruler**. *J Phys Chem A* 2009, **113**(16) :4639–4646.

[40] Pascal C, Paté F, Cheynier V, Delsuc MA : **Study of the interactions between a proline-rich protein and a flavan-3-ol by NMR : residual structures in the natively unfolded protein provides anchorage points for the ligands**. *Biopolymers* 2009, **91**(9) :745–56.

[41] Ortega A, de la Torre J : **Hydrodynamic properties of rodlike and disklike particles in dilute solution**. *J Chem Phys* 2003, **119**(18) :9914–9919.

[42] Berisio R, Vitagliano L, Mazzarella L, Zagari A : **Crystal structure of the collagen triple helix model [(Pro-Pro-Gly)(10)](3)**. *Protein Sci* 2002, **11**(2) :262–70.

[43] Marsh JA, Forman-Kay JD : **Sequence Determinants of Compaction in Intrinsically Disordered Proteins**. *Biophys J* 2010, **98**(10) :2383–2390.

[44] Hinsen K : **The molecular modeling toolkit : A new approach to molecular simulations**. *J Comput Chem* 2000, **21**(2) :79–85.

[45] Lovell SC, Davis IW, Arendall WB, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC : **Structure validation by Calpha geometry : phi,psi and Cbeta deviation**. *Proteins* 2003, **50**(3) :437–50.

[46] Ramakrishnan C, Ramachandran GN : **Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units**. *Biophys J* 1965, **5**(6) :909–33.

[47] Muñoz V, Serrano L : **Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices : comparison with experimental scales**. *Proteins* 1994, **20**(4) :301–11.

[48] Kleywegt GJ, Jones TA : **Phi/psi-chology : Ramachandran revisited**. *Structure* 1996, **4**(12) :1395–400.

[49] Hovmöller S, Zhou T, Ohlson T : **Conformations of amino acids in proteins**. *Acta Crystallogr D Biol Crystallogr* 2002, **58**(Pt 5) :768–76.

[50] Milner-White EJ : **Situations of gamma-turns in proteins. Their relation to alpha-helices, beta-sheets and ligand binding sites**. *J Mol Biol* 1990, **216**(2) :386–97.

[51] Ho BK, Brasseur R : **The Ramachandran plots of glycine and pre-proline**. *BMC Struct Biol* 2005, **5** :14.

[52] Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC : **MolProbity : all-atom structure validation for macromolecular crystallography**. *Acta Crystallogr D Biol Crystallogr* 2010, **66**(Pt 1) :12–21.

[53] Hooft RW, Vriend G, Sander C, Abola EE : **Errors in protein structures**. *Nature* 1996, **381**(6580) :272.

[54] Kabsch W, Sander C : **Dictionary of protein secondary structure : pattern recognition of hydrogen-bonded and geometrical features**. *Biopolymers* 1983, **22**(12) :2577–637.

[55] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL : **Biopython : freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics* 2009, **25**(11) :1422–3.

[56] Labesse G, Colloc'h N, Pothier J, Mornon JP : **P-SEA : a new efficient assignment of secondary structure from C alpha trace of proteins**. *Comput Appl Biosci* 1997, **13**(3) :291–5.

[57] Lewit-Bentley A, Réty S : **EF-hand calcium-binding proteins**. *Curr Opin Struct Biol* 2000, **10**(6) :637–43.

[58] Shen Y, Delaglio F, Cornilescu G, Bax A : **TALOS+ : a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts**. *J Biomol NMR* 2009, **44**(4) :213–23.

[59] Shen Y, Vernon R, Baker D, Bax A : **De novo protein structure generation from incomplete chemical shift assignments**. *J Biomol NMR* 2009, **43**(2) :63–78.

[60] Wishart D, Sykes B, Richards F : **The Chemical-Shift Index - A fast and simple method for the assignment of protein secondary structure through NMR-spectroscopy**. *Biochemistry* 1992, **31**(6) :1647–1651.

[61] Schwarzinger S, Kroon GJ, Foss TR, Chung J, Wright PE, Dyson HJ : **Sequence-dependent correction of random coil NMR chemical shifts**. *J Am Chem Soc* 2001, **123**(13) :2970–8.

[62] Zhang H, Neal S, Wishart DS : **RefDB : a database of uniformly referenced protein chemical shifts**. *J Biomol NMR* 2003, **25**(3) :173–95.

[63] Wang J, Liu H : **A Bayesian-probability-based method for assigning protein backbone dihedral angles based on chemical shifts and local sequences**. *J Biomol NMR* 2007, **37** :31–41.

[64] Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao H, Markley JL : **BioMagResBank**. *Nucleic Acids Res* 2008, **36**(Database issue) :D402–8.

[65] Bernadó P, Blanchard L, Timmins P, Marion D, Ruigrok RWH, Blackledge M : **A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering**. *Proc Natl Acad Sci USA* 2005, **102**(47) :17002–7.

[66] Schanda P, Forge V, Brutscher B : **HET-SOFAST NMR for fast detection of structural compactness and heterogeneity along polypeptide chains**. *Magn Reson Chem* 2006, **44** :S177–S184.

[67] Lalevée S, Bour G, Quinternet M, Samarut E, Kessler P, Vitorino M, Bruck N, Delsuc MA, Vonesch JL, Kieffer B, Rochette-Egly C : **Vinexinß, an atypical "sensor" of retinoic acid**

receptor gamma signaling : union and sequestration, separation, and phosphorylation. *FASEB J* 2010, **24**(11) :4523–34.

[68] Petoukhov MV, Konarev PV, Kikhney AG, Svergun DI : **ATSAS 2.1 - towards automated and web-supported small-angle scattering data analysis**. *Journal of applied crystallography* 2007, **40** :S223–S228.

Annexe A

Polydispersité du PEO par DOSY

Il est très difficile voire impossible d'avoir uniquement une seule taille d'un même polymère dans un échantillon. La distribution de masses est donc un paramètre important à mesurer pour caractériser l'échantillon. De plus, cette distribution module les propriétés du polymère. Un indicateur couramment utilisé pour décrire la distribution des masses est l'indice de polydispersité (PDI : PolyDispersity Index). Il peut être calculé comme le rapport de la masse moyenne en masse M_w sur la masse moyenne en nombre M_n.

Par DOSY, il était déjà possible de connaître la distribution des masses dans un échantillon grâce au traitement du spectre par ILT. Nous avons donc utilisé cette technique pour déterminer le PDI de 15 mélanges de PEO connus. La différence de signal observée entre les monomères terminaux de la chaîne et la chaîne entière permet de retrouver les valeurs de coefficients de diffusion relatives à M_w et M_n et donc d'en déduire le PDI.

L'article ci-après présente les expériences menées et montre que la méthode décrite est efficace et fiable. On peut donc à présent déterminer le PDI d'un échantillon précisément par une seule expérience de DOSY dans la mesure où les monomères terminaux ont leur propre signal de RMN. Certains polymères entrent dans cette catégorie, les autres peuvent potentiellement être modifiés chimiquement aux extrémités afin de réaliser l'expérience.

Polydispersity index of polymers revealed by DOSY-NMR

Justine Viéville[a,b], Matthieu Tanty[a,b], Marc-André Delsuc[a,b,*]

[a] *Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), UMR 7104, 1 rue Laurent Fries, BP 10142, 67404 Illkirch cedex, France*
[b] *NMRTEC, Bioparc B, boulevard Sébastien Brant, 67400 Illkirch, France*

Abstract

The polydispersity of a polymer chain is usually measured by its polydispersity index (PDI). In this study we present a method which allows to estimate the PDI of linear polymers from a simple diffusion experiment.

The approach is based on the differential diffusion profile observed for the main polymer chain signal vs the extremity signal. From this difference, a statistical analysis of the DOSY spectrum allows the PDI to be estimated accurately, to the condition that the Flory coefficient of the polymer chain is known. Alternatively, the mass average molar mass M_w and the number average molar mass M_n can be extracted separately from the same spectrum.

Results on PEO mixes reveal that, using this new method, PDI can be estimated with a very good accuracy. This method can easily be applied to almost any kind of linear polymers.

Keywords: polymer, polydispersity, DOSY, fractal dimension, Inverse Laplace Transform

1. Introduction

Polymers are characterized by a distribution of molecular masses. For linear polymers this distribution can be expressed in terms of average chain length and polydispersity of the chain length. The average chain length

*corresponding author. tel : (+33) (0) 3 68 85 47 29 fax : (+33) (0) 3 68 85 47 18
Email addresses: vieville@igbmc.fr, justine.vieville@nmrtec.com (Justine Viéville), tanty@igbmc.fr (Matthieu Tanty), delsuc@igbmc.fr (Marc-André Delsuc)

is routinely measured by NMR spectroscopy by measuring the ratio of the integrals of the main chain signal to the extremity signals.

Self-diffusion, measured by pulsed field gradient NMR (PFGNMR) is sensitive to molecular size and provides an approach to the determination of the distribution of molecular mass. In the case of linear polymer chains, the diffusion coefficient is linked to the molecular mass through the following equation:

$$M \propto D^{-d_f} \tag{1}$$

where D is the diffusion coefficient of the molecule, M its mass and d_f its fractal dimension [1, 2, 3]. The fractal dimension is a measure of the way the chain extends into the solvent, and is equal to the inverse of the Flory coefficient ν : $d_f = 1/\nu$. It is comprised between 5/3 and 3 [4].

The effect of polydispersity on PFGNMR measurements has already been well studied [5, 6, 7] and is known to lead to non exponential decays, even for weak polydispersity [8]. The analysis of non exponential decays requires the use of Inverse Laplace Transform (ILT) in order to estimate the molecular mass distribution. This is an ill-posed mathematical problem, to which an approximate solution can only be constructed. This was investigated by Chen et al using the CONTIN algorithm [1], but due to the approximate reconstruction, it is not possible to extract a useful value of the polydispersity index from this approach.

Polydispersity is commonly measured by the polydispersity index (PDI). For a given polymer sample, it is defined as the ratio of the mass average molar mass (M_w) to its number averaged molar mass (M_n).

$$PDI = \frac{M_w}{M_n} \tag{2}$$

For a homopolymer linear chain, assuming that the mass of the chain is equal to the product of its length by the mass of the monomeric unit M, the average molecular masses M_n and M_w expressions are given in equations 3 and 4,

$$M_n = \frac{\sum n_i M_i}{\sum n_i} = NM \tag{3}$$

$$M_w = \frac{\sum m_i M_i}{\sum m_i} = \frac{\sum n_i M_i^2}{\sum n_i M_i} \tag{4}$$

where n_i is the number of molecules of mass M_i, m_i the mass of molecules of mass M_i, and N the averaged chain length.

M_n and M_w are statistical features of the same polymer distribution but with different weightings. Because of this difference, M_w is always greater than or equal to M_n, and PDI is always greater than or equal to 1.

NMR parameters are also obtained as statistical average on the whole sample. Signals measured from the extremity of the chain are weighted by the number of molecules, while signals measured on the whole polymer chain, extremities included, are weighted by mass of the molecules. Thus, the two different weightings used for defining M_n and M_w can be observed in NMR, depending on the measure being performed either on the extremities or on the whole polymer chain.

This property applies also to PFGNMR measurement of diffusion coefficients. From equations 1 and 2, it is thus possible to express PDI as follows

$$PDI = \left(\frac{\langle D_w \rangle}{\langle D_n \rangle} \right)^{-d_f} \tag{5}$$

where $\langle D_n \rangle$ is the mean diffusion coefficient measured from the ILT analysis of the PFGNMR signal of the extremity units, and $\langle D_w \rangle$ the mean diffusion coefficient measured for the whole polymer.

From this theoretical presentation, it appears that the polydispersity index can be determined from a simple PFGNMR measurement, by comparing the signals originated from the main chain to the signal of the extremities, and applying equation 5, given the preliminary knowledge of fractal dimension of the chain.

To confirm this hypothesis, 2D-DOSY spectra were registered for different mixes of poly-ethyleneoxide (PEO) in water with calibrated chain lengths and PDIs. Experimental values were confronted to theoretical ones.

2. Results and Discussion

2.1. 1D NMR

The 1D-^{1}H NMR spectrum of mix L is shown in Figure 1. Very few peaks are observed and are easily assigned. Besides the resonance at 3.7 ppm of the principal chain, other resonances are observed. Four different spin systems can be identified. The signal A corresponds to the chain methylene group, and the signal B to the penultimate methylene. The signal C corresponds to the signal assigned to the terminal methylene group bearing the hydroxy function, and last signals D and E show the ^{13}C-satellites of the chain protons.

As one can see on this example, protons on the PEO's extremities are clearly identified. It is the case for all the mixes which makes the calculation of the average chain length N always possible.

2.2. The DOSY experiment

Figure 2 shows the NMR signal decay for signals A and C from PEO mix L versus the square of the gradient strength. Both curves are columns extracted respectively at 3.69 ppm and 3.63 ppm, from the diffusion experiment performed on mix L after Fourier transform and baseline correction. As expected, being obtained from a quite polydisperse sample (here PDI=2.51) both curves present a strong non-exponential decay, as can be seen from the non-linearity of the log-plot. The ILT analysis of the decays produces a DOSY spectra with peaks broaden along the diffusion axis. However, despite this broadening, this experiment reveals two different diffusion regimes. Due to marked difference in the diffusion coefficients, the 2D-DOSY spectrum displayed in Figure 3 unequivocally confirms that two different diffusion profiles can be extracted. In this example, the DOSY peak summit measured for the extremity was found to be D=1.83 $10^3\,\mu m^2 s^{-1}$, whereas the DOSY peak summit measured for the chain was found to be D=1.11 $10^3\,\mu m^2 s^{-1}$.

It should be noted that the measure relies on the complete measurement of all the different polymers present in the sample. In consequence, the PFG experiment should be designed to allow a signal attenuation from the longest chains, sufficient for a correct measurement of their diffusion coefficient. In the present case, all experiments have been performed in the same conditions, optimized on the largest monodisperse polymer studied. However, when the composition of the mix is unknown, one should use large enough PFG intensities to ensure the signal attenuation of the heaviest polymers. As a rule of thumb, a final attenuation around 10% on a monodisperse species is usually required to permit a precise determination of the diffusion coefficient. Thus, when studying an unknown polydisperse polymer, one should try to reach at least a 1% attenuation for the main signal.

Mega-dalton polymers have already been precisely measured on standard spectrometers[3]. So, the main size limitation for the application of this technique is the possibility to reliably detect signals from the chain extremities. Of course, this is more difficult to achieve on large polymers, as the extremity signals might be too faint to be observed. This was done here for PEO polymers up to 10 kDa, despite the fact that this signal only integrates as a CH_2.

The method requires that the extremity of the polymer presents an isolated signal in the NMR spectrum. This condition is not really stringent. On the PEO samples, the small shift of 0.08 ppm observed between the chain and the extremity signals, is sufficient for the study. Given this shift difference, two different diffusion coefficient distributions can be extracted by ILT from one 2D-DOSY spectrum. By integrating over the regions displayed in Figure 3, the barycenters of these distributions are calculated to estimate the PDI.

2.3. Comparison to theoretical values

Table 1 gathers the expected and measured values of N and PDI for all the different analyzed PEO mixes and Figure 4 shows the comparison between the theoretical and measured values of N and PDI. The very good correlation between theory and measurement indicates that the quality of the method.

The value of the fractal dimension d_f is the only free parameter which is needed to extract the PDI for the 2D-DOSY spectrum. From the Flory theory, it is predicted to be 5/3 and 3 for fully solvated and collapsed polymer chains, respectively[4]. In a θ solvent, where polymer-polymer interactions are equal to polymer-solvent and solvent-solvent interactions, the polymer behaves as a Gaussian chain and the exponent d_f is predicted to be 2.

Results presented here have been obtained with a d_f value of 1.86, as was determined by several studies [3, 9]. Are also added in Figure 4, points showing the impact of varying the d_f values. It can be observed that while an error on this value may have an impact on the PDI accuracy, this impact is not very important.

3. Conclusion

We have shown that DOSY NMR can bring valuable information on polydisperse polymers. With the proposed approach the polydispersity index as well as average chain length can readily be determined for linear polymers. To assess the polydispersity index from the 2D DOSY spectra, the barycenter of diffusion peak is calculated, this is made possible here thanks to the ILT analysis of the DOSY signal, which conserves the properties of the polymer distribution [7]. With this approach, results are independent of the average chain length and only the fractal dimension of the polymer chain d_f must be known. Experimental average chain length and polydispersity index did not

indicate significative difference when compared to supplier data. This technique was shown to equally reliable and accurate for both high (5.23) and low (1.04) polydispersity indexes. This method requires a separate NMR proton signal for the extremity of the studied polymer to be observed. However this condition is not stringent, as it was easily fulfilled here in the case of PEO, where only 0.08 ppm separates both signals. It will be easily extended to polymers with different extremity chemical patterns (for example, a methyl- or amide- group). Moreover, in the case of very large polymers with low extremities signals, a chemical modification of the extremities will allow the use of the method presented here.

Compared to other PDI determination technique such as Mass Spectrometry or Size Exclusion Chromatography, this approach presents the unique advantage of a direct measure which does not require any interaction with a static phase, separation, ionization or dilution of the polymer. It does not require any special calibration, equipment, or preparation and is a rapidly obtained with a diluted polymer sample. NMR as always been a powerful spectroscopy for the study of polymers, with DOSY NMR and the proposed procedure, the range of the physico-chemical parameters which can be accessed by NMR is further extended.

4. Experimental

4.1. Sample preparation

A set of 17 PEO standards, with masses ranging from 106 Da to 10730 Da were purchased from American Polymer Standards Corporation (Mentor, Ohio, USA). Each standard has been dissolved in Milli-Q water to 10% (w/v) solutions. These solutions were used to create 15 mixes with controlled PDI from 1.04 to 5.23 and a concentration range from 0.05 to 1% (w/v). Each mix contains 10% D_2O (v/v) and 1% (v/v) of a 1 mM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) aqueous solution. The details of the 17 standard PEO as well as the 15 mixes are given in the supplementary materials.

4.2. NMR spectroscopy

1D-^{1}H and 2D-DOSY experiments were carried out on each PEO mix at 298 K on a 500 MHz Bruker Avance I NMR spectrometer employing a 5 mm TXI probe equipped with z-gradients delivering up to 53 G/cm.

1D-^{1}H experiments were obtained with presaturation of water, 128 scans of 16k data points, and recycle time of 6.1 seconds. Fourier transform was

applied with zerofilling and 0.5 Hz exponential broadening. Careful spline polynomial correction was applied to each 1D-^{1}H spectra before integration.

2D-DOSY experiments were acquired using a LED experiment with bipolar pulses [10] and water presaturation. Gradients were linearly sampled from 0.5 G/cm to 45.7 G/cm in 40 points. 32 scans were acquired on 16k data points, for a total acquisition time of 1 hour and 7 minutes.

The gradient pulse length was $\delta/2 = 1.5$ ms and the Δ diffusion delay was adapted to the sample for values in the 100 ms - 180 ms range. The DOSY spectra were obtained by applying an Inverse Laplace Transform (ILT) along the diffusion axis, using the *Gifa* algorithm [11, 12] embedded into the commercial software NMRnotebook (NMRTEC Illkirch). Careful spline polynomial correction was applied along the F2 dimension before the ILT processing, which was computed on 256 points, using the highest quality available in the algorithm.

4.3. Average chain length determination

For each 1D-^{1}H spectrum, peaks corresponding to extremity CH_2 protons and chain CH_2 protons were integrated. There are 4 protons at the extremities and $4N - 4$ protons inside the chain, where N is the number of monomeric units in the polymer chain. The ratio of both integrals allows the average chain length N to be extracted.

4.4. PDI determination

For a given sample, the diffusion coefficient distributions given by the 2D-DOSY spectrum were studied over ranges of chemical shift intervals. Mean diffusion coefficients were computed as a barycenter along the diffusion axis by integrated over determined spectral regions of the 2D spectrum. $\langle D_n \rangle$ was computed as the mean diffusion coefficients measured over the chemical shifts corresponding to the extremity of the chain. $\langle D_w \rangle$ should be evaluated over the whole polymer chain, thus the averaging was performed over a range of chemical shifts, encompassing all polymer signals, main chain and extremities included.

PDI was then obtained using equation 5, using a value of d_f equal to 1.86 [3, 9]. From the average chain length N, the number average molecular mass M_n was computed using equation 3, from the values of PDI and M_n, the mass average molecular mass was determined from $M_w = PDI\ M_n$.

All the experimental values are given in Table 1. The complete procedure has been programmed into a python script, which can be embedded as a

macro into the NMRnotebook software. The python script of this macro is available in the supplementary materials.

5. Acknowledgement

The authors want to thank Marie-Aude Coutouly for her help in developing the NNB macro. JV and MT also acknowledge NMRTEC for financial support.

6. References

[1] A. Chen, D. Wu, C. Johnson, Determination of molecular-weight distributions for polymers by diffusion-ordered NMR, J Am Chem Soc 117 (1995) 7965–7970.

[2] D. K. Wilkins, S. B. Grimshaw, V. Receveur, C. M. Dobson, J. A. Jones, L. J. Smith, Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques, Biochemistry 38 (1999) 16424–31.

[3] S. Augé, P.-O. Schmit, C. A. Crutchfield, M. T. Islam, D. J. Harris, E. Durand, M. Clemancey, A.-A. Quoineaud, J.-M. Lancelin, Y. Prigent, F. Taulelle, M.-A. Delsuc, NMR measure of translational diffusion and fractal dimension. application to molecular mass measurement, J Phys Chem B 113 (2009) 1914–8.

[4] P. Flory, Principles of polymer chemistry, Cornell University Press, Ithaca, NY. (1953).

[5] E. V. Meerwall, Interpreting pulsed-gradient spin-echo diffusion experiments in polydisperse specimens, J Magn Reson 50 (1982) 409–416.

[6] P. T. Callaghan, D. N. Pinder, A pulsed field gradient NMR study of self-diffusion in a polydisperse polymer system: Dextran in water, Macromolecules 16 (1983) 968–973.

[7] P. Callaghan, D. Pinder, Influence of polydispersity on polymer self-diffusion measurements by pulsed field gradient nuclear magnetic-resonance, Macromolecules 18 (1985) 373–379.

[8] G. Fleischer, The effect of polydispersity on measuring polymer self-diffusion with the NMR pulsed field gradient technique, Polymer 26 (1985) 1677–1682.

[9] K. Chari, B. Antalek, J. Minter, Diffusion and scaling behavior of polymer-surfactant aggregates, Phys Rev Lett 74 (1995) 3624–3627.

[10] D. Wu, A. Chen, C. Johnson, An improved diffusion-ordered spectroscopy experiment incorporating bipolar-gradient pulses, J Magn Reson Ser A 115 (1995) 260–264.

[11] M.-A. Delsuc, T. Malliavin, Maximum entropy processing of DOSY NMR spectra, Anal Chem 70 (1998) 2146–2148.

[12] D. Tramesel, V. Catherinot, M.-A. Delsuc, Modeling of NMR processing, toward efficient unattended processing of NMR experiments, J Magn Reson 188 (2007) 56–67.

7. tables

	N		PDI		M_n (g.mol^{-1})		M_w (g.mol^{-1})	
PEO mix	theo	exp	theo	exp	theo	exp	theo	exp
mix A	36.3	36.1	1.04	1.06	1615	1588.4	1679.6	1683.7
mix B	49.4	53.7	1.07	1.08	2190	2362.8	2343.3	2551.8
mix C	107.5	120	1.11	1.12	4750	5280	5272.5	5913.6
mix D	8.7	8.7	1.12	1.12	400.6	382.8	447.5	428.7
mix E	8.8	9.5	1.14	1.17	403.4	418	459.7	489.1
mix F	23	23.8	1.26	1.31	1021	1047.2	1268.3	1371.8
mix G	71.5	85.3	1.28	1.28	3165	3753.2	4051.2	4804.1
mix H	44.8	47.5	1.34	1.15	1989.3	2090	2518	2403.5
mix I	34.7	37	1.54	1.5	1544.3	1628	2382.7	2442
mix J	38.1	39.8	2.01	2.2	1696.9	1751.2	3238.5	3852.6
mix K	15.7	15.9	2.12	1.89	710.3	699.6	1332.1	1322.2
mix L	19.1	19.4	2.51	2.24	858.5	853.6	2008.2	1912.1
mix M	21.4	24.8	3.3	2.41	961.7	1091.2	3062.8	2629.8
mix N	22.4	23.8	3.41	2.75	1001.7	1047.2	3328	2879.8
mix O	13.5	13.9	5.23	4.48	611.3	611.6	3099.6	2740

Table 1: Experimental results compared to theoretical values.

8. figure legends

Figure 1

1D-^{1}H NMR spectrum of a PolyEthyleneOxide in D_2O, mix L. Assignment is given in insert, D and E are the ^{13}C satellites of A.

Figure 2

log-plot of the observed decays for varying gradient values squared for mix L. Diamonds are from the signal of the main chain (signal A Figure 1), dots are from the signal of the extremity (signal C).

Figure 3

2D DOSY spectrum of mix L. The black rectangle on the right is the region over which the integration is made to determine $\langle D_n \rangle$. $\langle D_w \rangle$ is determined by integration over the whole spectral range shown by outer red rectangle.

Figure 4

Correlations curves for the average chain length N (left) and the PDI (right). Red lines correspond to $theo = exp$. The PDI was computed with the fractal dimension value $d_f = 1.86$ (blue square); $d_f = 1.96$ (upper bar); and $d_f = 1.76$ (lower bar).

9. figures

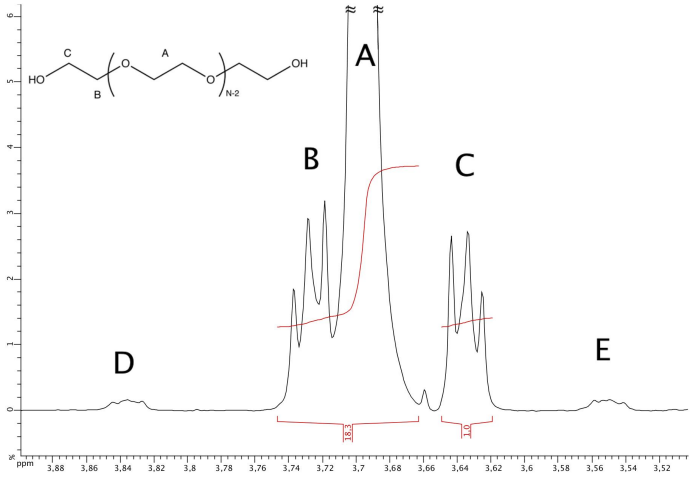

Figure 1: 1D-^{1}H NMR spectrum of a PolyEthyleneOxide in D_2O, mix L. Assignment is given in insert, D and E are the ^{13}C satellites of A.

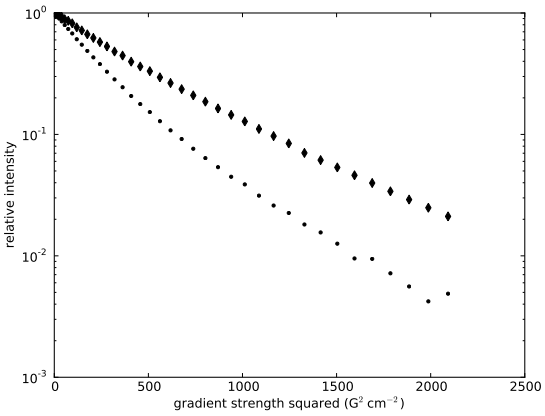

Figure 2: log-plot of the observed decays for varying gradient values squared for mix L. Diamonds are from the signal of the main chain (signal A Figure 1), dots are from the signal of the extremity (signal C).

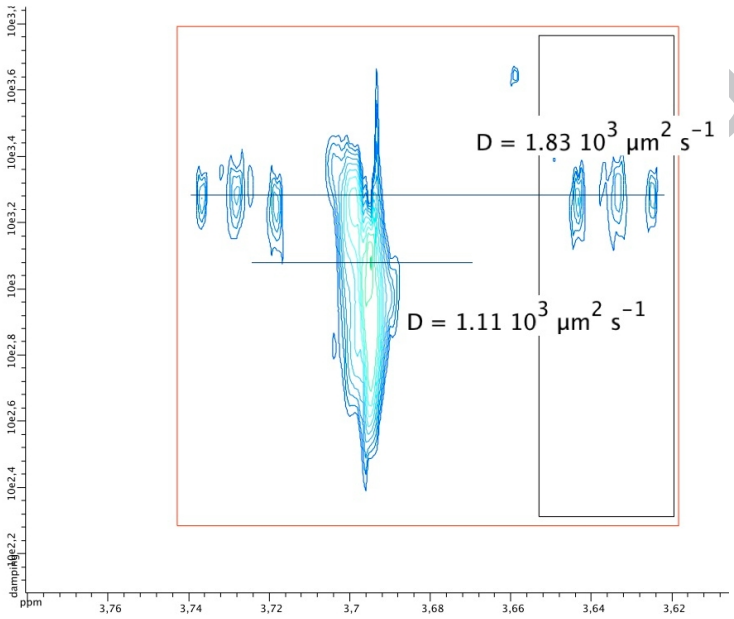

Figure 3: 2D DOSY spectrum of mix L. The black rectangle on the right is the region over which the integration is made to determine $\langle D_n \rangle$. $\langle D_w \rangle$ is determined by integration over the whole spectral range shown by outer red rectangle.

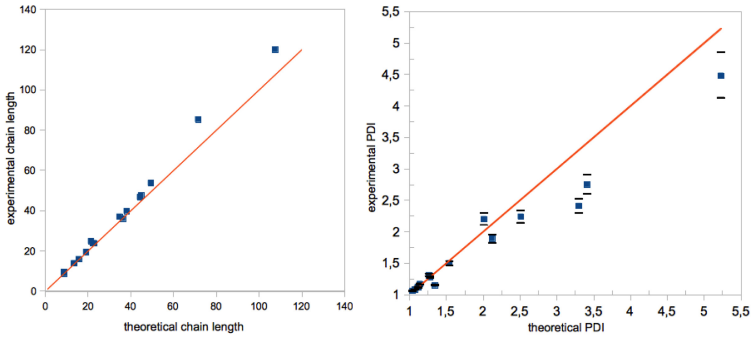

Figure 4: Correlations curves for the average chain length N (left) and the PDI (right). Red lines correspond to $theo = exp$. The PDI was computed with the fractal dimension value $d_f = 1.86$ (blue square); $d_f = 1.96$ (upper bar); and $d_f = 1.76$ (lower bar).