Etude de la dynamique conformationnelle des protéines intrinsèquement désordonnées par résonance magnétique nucléaire

Valéry Ozenne

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Physique pour les Sciences du Vivant**

Arrêté ministériel : 7 août 2006

Présentée par

Valéry Ozenne

Thèse dirigée par **Martin Blackledge**

préparée au sein de l'équipe **Flexibilité et Dynamique des Protéines de l'Institut de Biologie Structurale**
et de l'**Ecole Doctorale de Physique de Grenoble**

Caractérisation des protéines intrinsèquement désordonnées par résonance magnétique nucléaire

Thèse soutenue publiquement le **28 novembre 2012**,
devant le jury composé de :

Dr. Marc-André Delsuc
Docteur à l'Institut de Génétique et de Biologie Moléculaire et Cellulaire, Strasbourg, Rapporteur
Dr. Guido Pintacuda
Docteur au Centre de RMN à Très Hauts Champs, Lyon, Rapporteur
Dr. Pau Bernado
Docteur au Centre de Biologie Structurale, Montpellier, Examinateur
Prof. Michael Nilges
Professeur à l'Institut Pasteur, Paris, Examinateur
Prof. Franz Bruckert
Professeur à l'Institut National Polytechnique de Grenoble, Examinateur
Dr. Martin Blackledge
Docteur à l'Institut de Biologie Structurale, Grenoble, Directeur de thèse

Remerciements

Je tiens à remercier les membres du jury d'avoir accepté de juger ce travail.

Je remercie le Docteur Marc-André Delsuc et le Docteur Guido Pintacuda rapporteurs de ce travail, le Professeur Franz Bruckert qui m'a fait l'honneur d'être président du jury du thèse et qui n'est point étranger à ma découverte 3 ans plus tôt de cette équipe de recherche où j'ai eu la chance de travailler. Enfin je souhaite vivement remercier le Docteur Pau Bernado et le Professeur Michael Nilges pour la discussion enrichissante qu'ils ont su mener lors de la soutenance.

Je remercie Martin Blackledge, mon directeur de thèse, qui m'a acceuilli en stage de Master puis en thèse et m'a fait découvrir la Biologie, la RMN. J'ai beaucoup appris durant ces années en Science évidemment mais aussi sur le fonctionnement d'un projet de recherche et sur la recherche en général.

Je remercie aussi toutes personnes qui m'ont encadré et enseigné les connaissances nécessaires à l'étude de ces protéines désordonnées et plus encore : Malène Ringkjøbing Jensen (Hi Hi chef), Loïc Salmon, Gabrielle Nodet, Guillaume Communie, Antoine Licinio, Luca Mollica, Robert Schneider, Jie-Rong Huang, Jaka Kragelj, Damien Maurin, Eric Condamine, Paul Guerry, Elise Delaforge et Anton Abyzov.

*Doubles tournées de remerciements à ceux qui m'ont aidé à écrire et corriger cette thèse.
**Triples tournées de remerciements à ceux qui m'ont côtoyé en salle 5239.
***Je remercie beaucoup Guillaume mais un peu moins que prévu pour ne pas avoir vérifié l'attribution de la Phosphoprotéine, et ceci malgré les prolongations du délai.

Je remercie techniciens, thésards, post-doc's que j'ai rencontré durant ma thèse Mickael, Mélanie, Laurianne, Zsofia, Pavel, Rime, Axel, Maria, Paul, Mingxi, Morgane, Sarah, Sandy, Lionel, Michael, Cristina, Nicolas.

Je remercie l'ensemble des institutions qui m'ont permis à bien de mener ce travail et notamment le Département des Sciences du Vivant du CEA pour ma bourse de thèse, l'Institut de Biologie Structurale, le PSB et l'EMBL.

Je tiens à remercier le personnel administratif de l'IBS, en particulier Dominique Ribiero et Linda Ponnet, l'équipe informatique : Fabrice Leger, Frederic Metoz, Didier Depoisier et Jean-Luc Parouty et toutes les équipes avec qui j'ai pu échanger.

Je remercie mes différents collaborateurs : Frank Gabel, Martin Weik, Pau Benardo, Markus Zweckstetter, Stephan Bibow, Kim, Magnus Kjaergaard, Jens Danielsson, le Groupe informatique pour les Scientifiques du Sud Est avec qui j'ai travaillé pour développer l'application flexible-meccano : Merci à Céline Charavay, Stéphane Ségard, Frédéric Bauer et Madalina Ghita.

Je remercie quelques autres institutions tel que au Hasard, la station des Sept Laux, le Tholonet, le O'callaghan, le Ptit Vélo, le Taraf Tukur qui m'ont permis à bien de

mener ce travail.

Une autre tournée de remerciements incluant ma famille et mes amis proches : Papa, Maman et Brice ainsi que PA, Laurie, Bruno, Bouss, Elo, Bouss, Antoine, Léo, Hélène, Nico, Oriane, Pauline, Maud, Thibaud, Pauline, Clément, Jessica, Romain, Christophe, Cerise, Matias, Caro, Charlène, Tom, Raph, Brice (la fine équipe), Matth, Loic (le bien nommé), Grand, Remz, Perrine, Munch, Bou, Poual, Hadrien et la team de Chateau Jobis, Loris, Mathilde, Emma, Piwi, Laurent, Alex.

En cas d'oubli de ma part, n'hésitez pas à contribuer :

Je remercie ....
....

Je remercie ....
....

Je remercie ....
....

Je remercie ....
....

Table des matières

Liste des figures

Liste des tableaux

Introduction

Au cours des siècles, les progrès dans le domaine des mathématiques et de la physique ont bouleversé l'approche scientifique. Le processus de découverte ou la compréhension du vivant a été initialement guidé par l'expérience. Cette curiosité a amené le scientifique à formaliser ses observations de manière abstraite par des schémas ou par des équations. L'apparition de modèle[1] est devenue inévitable pour étudier un phénomène physique. Par la suite, la théorie commença à devancer occasionnellement l'expérience, les exemples et domaines d'applications concernés sont nombreux au début du XXe siècle, nous pouvons citer la prédiction de la planète Pluton en astronomie observée 30 ans plus tard, les éléments du tableau périodique en chimie[2] ou dans le domaine de la physique des particules avec les résultats concernant l'existence du boson de Higgs presque 50 ans après sa formulation est un des derniers exemples les plus médiatiques.

Les progrès théoriques et techniques ont toujours bouleversé l'approche scientifique, dernièrement l'arrivée de la simulation numérique et l'augmentation des moyens de calcul a fondamentalement changé la place de l'expérimentateur dans le processus de découverte. Celui-ci reste l'élément sans qui rien ne peut être prouvé mais son rôle au sein du monde scientifique s'est restreint. La simulation numérique a pris une place fondamentale en recherche, il est devenu usuel qu'elle prenne les devants sur l'expérimentation. Les raisons sont diverses, notamment économiques, mais cet argument serait réducteur à lui seul, la maitrise des outils de modélisation a permis de faire émerger de nombreux domaines et champs applications scientifiques : la météorologie, les contraintes des matériaux pour l'industrie, l'aérospatiale, l'aéronautique. Un des domaines d'expertise française est historiquement la modélisation des essais nucléaires et plus généralement de l'atome. La recherche médicale et notamment la biologie structurale n'échappa à cette règle, la compréhension des mécanismes du vivant a énormément évolué au cours des dernières années, l'apparition des moyens informatiques a permis à de nouveaux champs de recherche d'émerger : la dynamique moléculaire, la génomique structurale et fonctionnelle[3] et l'étude des protéines intrinsèquement désordonnées (PIDs) ou protéines fonctionnelles ne possédant pas de structure tridimensionnelle. Ce champ de recherche est en effet très récent et son importance s'est accrue depuis le début des années 2000 grâce aux outils de la bio-informatique. Elle a joué, et continue à le faire, un rôle considérable à la fois dans la découverte et la compréhension des protéines intrinsèquement désordonnées.

Les protéines possèdent une place très importante dans le génome, elles occupent des rôles extrêmement vastes et sont associés à presque tous les processus biologiques. Leur étude est un enjeu majeur à la fois pour la compréhension des mécanismes du vivant, de l'origine de la vie et plus concrètement pour la recherche et l'industrie

1. La modélisation consiste à prendre en compte les éléments fondamentaux d'un système et à décrire les grandeurs physiques de ce dernier. La relation entre ces grandeurs étant exprimées par les équations traduisant les lois de la physique régissant le comportement de l'objet.

2. Le tableau périodique de Mendeleïev date précisément de 1869

3. Techniquement la génomique n'est pas de la simulation numérique mais plutôt de l'analyse de donnée par informatique ou *data mining*.

médicale. La biologie structurale dont l'objet est l'étude du vivant est difficilement dissociable de son champ d'application : la conception de médicament [1] (*Drug Design*). La compréhension des mécanismes biologiques à ouvert les portes à de nouveaux champs de recherches associés à la conception et synthèse de médicaments. Le rôle des protéines pour expliquer les mécanismes biologiques s'appuie sur le paradigme suivant :

Le paradigme *structure-fonction* associe la structure tridimensionnelle d'une protéine à sa fonction. Le concept de *"clé-serrure"* a été introduit par Emil Fisher en 1894[4] à l'issu de ces travaux sur les réactions enzymatiques. Il constate l'incroyable spécificité et sélectivité des enzymes assurant ainsi leur fonctionnement. La structure tridimensionnelle d'une protéine ou d'un ligand dicte sa fonction, l'association ou l'interaction de deux protéines étant guidée par leur compatibilité structurelle et chimique. Cette hypothèse perdura jusqu'à la fin du XX^e siècle et guida l'ensemble des programmes de recherche telle que les projets de génomique structurale.

Dans les années 1970 des chercheurs mettent en évidence l'existence de protéines fonctionnelles partiellement ou complètement dépliées [2] contredisant le paradigme *structure-fonction*. D'autres découvertes occasionnelles suivent mais sont laissées à l'écart vu la place grandissante de la résolution de structure par cristallographie aux rayons X ou spectroscopie RMN [3]. À partir des années 2000, la mise en évidence des caractéristiques de ces protéines non repliées mais fonctionnelles alla de pair avec la reconnaissance de ce champ de recherche et la remise en question du dogme *structure-fonction*. L'orientation des recherches se divisa avec :

- la mise en évidence de nombreux systèmes fonctionnels incluant des protéines ou régions désordonnées. Ces systèmes très variés révèlent le rôle primordial des protéines désordonnées. L'incapacité à adopter une structure 3D précise confère à ces protéines une grande flexibilité structurale, ce qui entraine des impacts biologiques multiples. La nature transitoire de ces interactions est particulièrement bien adaptée à des phénomènes de régulation qui nécessitent des réponses rapides à des changements environnementaux.

- l'avènement de la bio-informatique confirma l'importance notable de la place des protéines intrinsèquement désordonnées dans le génome.La bio-informatique a ainsi prédit l'existence de 45-50% de régions[5] désordonnées au sein du génome eucaryote[4]. L'application de cette même analyse computationnelle au génome procaryote indique 30% de régions désordonnées. Nous dénombrons à ce jour (2012-07-01) 1467 régions désordonnées réparties dans 667 protéines désordonnées recensées dans une base de données en ligne : DISPROT [5]. Cette base de données est l'analogue de la Protein Data Bank des protéines repliées [6].

- l'étude des caractéristiques biophysiques par différentes méthodes biophysiques et la mise en place de modèles pour interpréter les données expérimentales des protéines non repliées ou dénaturées. L'introduction du concept l'état *random-coil* ou absence de structure tridimensionnelle précise a permis d'améliorer nettement la compréhension des protéines désordonnées.

Toutes ces études ont permis d'une part la reconnaissance de rôle fonctionnel des protéines désordonnées mais aussi d'étendre le paradigme *structure-fonction* de la biologie structurale à une classe de protéine jusqu'ici peu connue.

4. Ce concept est donc bien antérieur à la Biologie Structurale.
5. C'est-à-dire constitué d'au moins 30 résidus contigus

Ce travail de thèse intervient donc quelques années après la reconnaissance des protéines désordonnées par la communauté scientifique. Cette thèse multidisciplinaire met en jeu des connaissances de physique, de spectroscopie par résonance magnétique nucléaire, de biologie, de chimie et d'informatique. Elle s'oriente sur la simulation numérique des protéines désordonnées à partir des connaissances actuelles de l'état déplié. La formulation théorique des propriétés des protéines étant à l'heure actuelle limitée, la description la plus adéquate a recours à la modélisation numérique. Le modèle adopté est une description par ensemble explicite de structures en échange traduisant le caractère flexible du système étudié. Ce modèle abandonne toute description biologique et adopte une vision physico-chimique, les protéines sont considérées comme des hétéro-polymères constitués d'éléments, les acides aminés, possédant des propriétés physiques et chimiques particulières. Nous pouvons distinguer l'information locale décrite par les angles dièdres (ϕ,ψ) traduisant l'orientation des plans peptidiques les uns par rapport aux autres de l'information à moyenne et longue portée traduisant les distances entre résidus. L'objet de cette thèse est donc de proposer un modèle viable permettant de caractériser quantitativement l'échantillonnage local et les interactions à longue portée présentes au sein des protéines désordonnées.

Pour cela, l'un des aspects cruciaux de ce travail de thèse repose sur la validation expérimentale du modèle. Nous avons recours à la résonance magnétique nucléaire (RMN), une méthode spectroscopique extrêmement sensible permettant d'obtenir de l'information variée, à l'échelle atomique et particulièrement adaptée à l'étude des mouvements sur différentes échelles de temps. Le caractère dynamique d'un système est compliqué à traiter, il est relatif à la méthode expérimentale utilisée pour l'observer. La RMN offre une diversité de mesures impressionnantes permettant d'accéder à l'ensemble des échelles de temps et ainsi à l'information caractéristique des mouvements des protéines. L'interprétation des paramètres RMN appliqués aux protéines désordonnées est un des éléments essentiels de ce manuscrit.

Le second atout essentiel de la RMN est sa résolution à l'échelle atomique, cette méthode exploite une caractéristique intrinsèque des atomes : le spin. La RMN étudie les transitions entre les différents états de spin, la seule condition pour l'atome considéré est l'utilisation d'un spin non nul. Dans le cas des protéines constituées principalement de noyaux Hydrogène, Carbone, Azote, et d'Oxygène. L'hydrogène $^{1}H^{N}$ présent abondamment possède un spin 1/2, il est en de même pour les isotopes ^{13}C, ^{15}N qui bien que de faibles abondances peuvent être exploités en enrichissant artificiellement la protéine.

La thèse est divisée en huit chapitres :

* L'objet des trois premiers chapitres est une introduction des outils nécessaires pour effectuer une description moléculaire de l'état déplié. Au premier chapitre, après une présentation des propriétés et caractéristiques connues ou supposées des protéines intrinsèquement désordonnées, nous mettrons en perspective l'importance de ces dernières au sein du génome et leur rôle dans de nombreux mécanismes clés. Nous continuerons en présentant la méthode expérimentale utilisée pendant cette thèse pour étudier ces protéines : la résonance magnétique nucléaire, une méthode à résolution atomique et adaptée à l'étude des systèmes dynamiques. Nous présenterons les observables mesurés et analysés permettant la caractérisation des protéines désordonnées : les déplacements chimiques caractérisant l'environnement électronique du noyau, les couplages dipolaires sensibles à la fois à l'information locale et à grande distance et la relaxation paramagnétique utile pour caractériser les interactions à grande portée. Nous terminerons cette

partie par un chapitre introduisant les modèles d'analyse des protéines désordonnées. Ces protéines hautement flexibles nécessitent le recours à une description statistique nommé la description par ensemble.

* Les deux chapitres suivants s'articulent autour de la caractérisation de l'échantillonnage local des protéines désordonnées, le premier après une présentation des méthodes développées au sein du groupe démontre précisément dans quelle mesure il est possible de combiner la résonance magnétique nucléaire et la description par ensemble pour caractériser l'état déplié de la protéine Ubiquitine dénaturée dans l'Urée. Le deuxième propose une étude précise de la relation liant les valeurs des paramètres RMN : les couplages dipolaires résiduels et les déplacements chimiques avec leur échantillonnage conformationnel. Prenant en compte ses considérations, nous mettons en place un protocole utilisant le minimum de données expérimentales pour caractériser le paysage énergétique des protéines désordonnées. Cette méthode est appliquée sur la partie C-terminale N_{tail} de la nucléoprotéine du virus de la Rougeole et sur la construction K18 de la protéine Tau.

* Les deux chapitres suivants concernent la caractérisation des interactions à longue portée, les protéines désordonnées sont impliquées dans de nombreuses maladies neurodégénératives caractérisées par la présence d'agrégats. L'existence d'ordre à longue portée pourrait être importante dans le mécanisme de formation et d'agrégation des fibrilles. Nous combinons des mesures de relaxation paramagnétique à une description par ensemble pour caractériser l'ordre présent dans la protéine α-Synucléine et la protéine Tau. La protéine Tau fait l'objet d'un chapitre complet où nous effectuons une description moléculaire de cette dernière en combinant l'ensemble des paramètres RMN disponibles.

* Un chapitre additionnel présente le projet de développement logiciel réalisé en parallèle pendant la thèse : il s'agit du développement de l'application Java Flexible-Meccano en collaboration avec le Groupe Informatique Pour les Scientifiques du sud-est (GIPSE) du Commissariat à l'Energie Atomique (CEA) de Grenoble. Cette application est maintenant disponible sur Internet. Nous présenterons les grandes lignes de ce projet.

* Nous terminerons ce manuscrit par un chapitre conclusif sur les travaux effectués et présentons les perspectives.

Réexaminer le paradigme structure-fonction des protéines

1

Les protéines intrinsèquement désordonnées (PIDs) ou "protéines non repliées à l'état natif" sont des protéines fonctionnelles dépourvues de structures secondaires et tertiaires stables en conditions physiologiques en l'absence d'un partenaire, et dont les fonctions nécessitent l'état désordonné [7, 8]. L'incapacité à adopter une structure 3D précise confère à ces protéines une grande flexibilité structurale, ce qui entraine des impacts biologiques multiples.

Qu'entendons-nous par protéines intrinsèquement désordonnées ? Ces protéines appelées aussi protéines flexibles ou dépliées sont souvent classées par opposition aux protéines repliées dites aussi ordonnées. La structure 3D d'une protéine repliée est relativement stable, les angles dièdres définissant l'orientation des plans peptidiques varient mais autour d'une position d'équilibre et effectue seulement occasionnellement des changements coopératifs plus importants. Dans le cas des protéines dépliées, il n'existe pas de structure précise mais un ensemble dynamique dont la position des atomes ou la valeur des angles dièdres change de manière constante au cours du temps sans transition coopérative spécifique.

Quelles protéines sont concernées ? D'après les prédictions de la bio-informatique une large partie du génome est concernée. Une façon alternative de répondre à cette question est de s'intéresser à l'absence de densité d'électron lors des mesures par cristallographie indiquant potentiellement la présence de structures flexibles. Seulement 32% des structures cristallines de la Protein Data Bank sont exempts de désordre [9]. Ainsi, les protéines complètement repliées ne sont pas si fréquentes, il existe souvent des fragments (de toutes tailles de quelques résidus à l'ordre de la centaine ou plus) présentant une tendance au désordre.

Comment définir le désordre ? La différenciation entre protéine repliée et protéine dépliée a été énoncée pour insister sur l'existence de protéines fonctionnelles n'ayant pas les mêmes propriétés physiques. Pourtant, cette distinction est pour ainsi dire sujette à de nombreuses critiques, il faut garder en tête que ces termes ne sont que des représentations usuelles d'une problématique et ne correspondent pas forcément à la réalité physique : comme précédemment énoncé toute protéine quelle qu'elle soit est un système en interaction avec son environnement et n'est donc pas strictement statique, elle possède une tendance à l'ordre ou au désordre mais avec des degrés variables. C'est pour cela qu'il est préférable de parler de continuum ordre-désordre.

Quel est le rôle des protéines intrinsèquement désordonnées ? La nature flexible des PIDs est particulièrement bien adaptée à des phénomènes de régulation qui nécessitent des réponses rapides à des changements environnementaux. Les PIDs sont ainsi impliqués dans de nombreux mécanismes biologiques où elles tiennent un rôle clé. Liées

à leur forte flexibilité, elles possèdent de nombreux atouts fonctionnels que sont la plasticité structurale et la grande surface d'association à des partenaires. Ils s'avèrent que ces protéines soient finement régulées et impliquées dans de nombreux mécanismes pathologiques.

Dans ce chapitre nous allons approfondir ces questions. Nous serons amenés à revisiter les caractéristiques clés de la biologie structurale. Il s'agit en effet d'étendre les concepts connus à une classe de protéines ne possédant pas les mêmes propriétés physiques. Nous présenterons alors plus spécifiquement les propriétés des protéines désordonnées en s'appuyant sur des systèmes biologiques connus ou en relation avec le travail de thèse présenté ultérieurement dans le manuscrit.

1.1 Le rôle des protéines désordonnées

Les protéines intrinsèquement désordonnées sont impliquées dans de nombreux mécanismes biologiques : la régulation de la transcription [10], le contrôle du cycle cellulaire, la reconnaissance moléculaire [11, 12], la signalisation cellulaire [13, 14] et dans de nombreux processus pathogènes. Ces protéines sont associées à un nombre impressionnant de maladies telles que le cancer [15], les maladies neurodégénératives comme Alzheimer [16, 17] et le Parkinson [18], les maladies cardiovasculaires ou le diabète [19, 20].

À l'heure actuelle, il est encore très difficile d'évaluer précisément le rôle des protéines désordonnées mais on ne peut ignorer leur implication dans des cascades d'événements menant à l'état pathogène.. L'un des exemples les plus frappants est la protéine α-Synucléine dont nous ignorons la fonction mais qui est clairement associée à la maladie de Parkinson [21] : cette protéine est le constituant principal des fibrilles trouvées dans les corps de Lewy, un des marqueurs de cette maladie. Un autre exemple bien connu est la protéine Tau impliquée dans la régulation des microtubules qui est associée à la dégénérescence neuronale [22, 23] : la protéine Tau fait partie de la famille des protéines *microtubule-associated proteins* (MAPs) participant à la formation du cytosquelette [24] cependant à l'état pathogène cette protéine s'agrège en fibrilles et implique directement ou indirectement la mort des cellules neuronales, nous présenterons ce mécanisme plus en détail au chapitre 7.

Afin d'identifier leur place au sein du génome, différentes pistes sont exploitées et de nombreux concepts sont présentés [19, 25, 26, 27, 28], de manière générale il semblerait que les protéines désordonnées soient situées à des rôles clés des mécanismes biologiques, elles possèdent la main mise sur de nombreux mécanismes de contrôle et leur fonctionnement serait finement régulé [10]. Ainsi, une altération de la régulation des protéines désordonnées pourrait aisément être associée à un mauvais fonctionnement et se traduire par des processus pathogènes [29].

1.2 Le continuum ordre-désordre

Les protéines ou régions désordonnées sont parts intégrantes du génome existant. Il est usuel de distinguer 3 familles de protéines : les protéines désordonnées, les protéines globules fondues et les protéines repliées. Pourtant, la frontière entre ces familles n'étant pas aussi nette que nous pourrions l'entendre, de nombreuses protéines sont composées à la fois de régions ordonnées et de régions désordonnées et il n'y a pas de définition précise délimitant le degré de structure résiduel nécessaire pour considérer une protéine comme repliée ou au contraire désordonnée. Il serait donc préférable de ne pas dissocier les protéines désordonnées des protéines ordonnées et d'utiliser un continuum désordre-ordre où le degré d'ordre ou de désordre présent au sein de la protéine varie suivant les conditions biologiques (pH, température, partenaire). Au premier abord, ce continuum peut paraître surprenant ou inadapté, pourtant la diversité des systèmes biologiques est telle qu'il soit difficile de parler distinctement de protéines uniquement désordonnées ou de protéines uniquement ordonnées.

La figure 1.1 présente quatre systèmes différents. Ce continuum désordre-ordre s'illustre très bien lors de changements conformationnels liés de l'appariement avec un partenaire. Pour les cas mentionnés, nous constatons un repliement partiel ou total des régions flexibles ou *random-coil*. La caractérisation des protéines désordonnées dans leur état fondamental est cruciale pour la compréhension de leurs propriétés biophysiques

mais reste une approche insuffisante pour obtenir un aperçu des aspects fonctionnels de ces protéines. Il est pour cela nécessaire d'étudier les voies d'interaction et la cinétique de ces protéines.

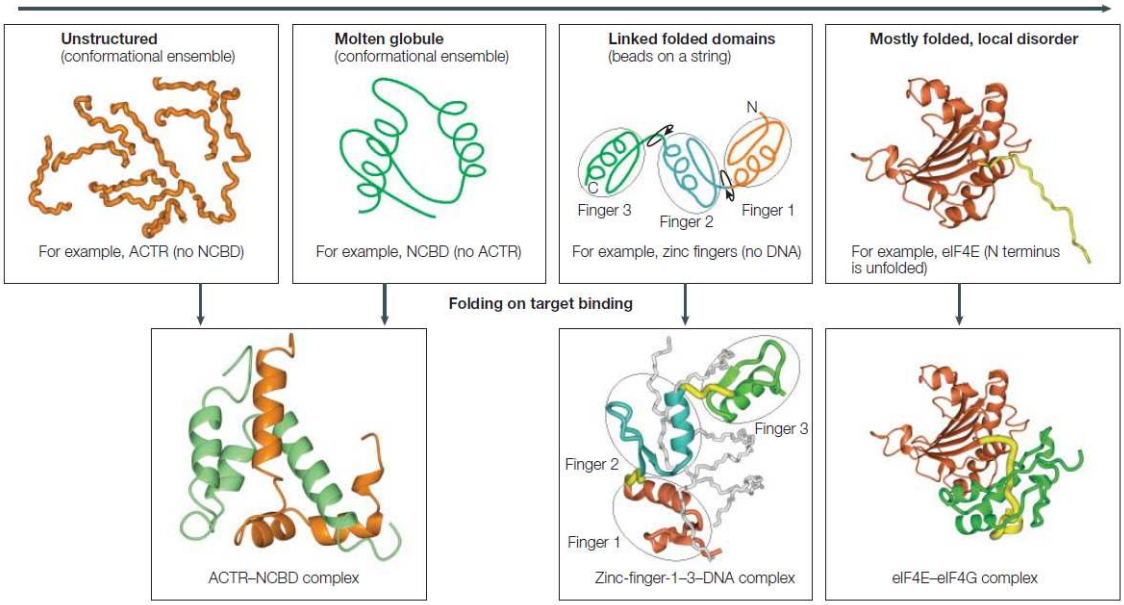


FIGURE 1.1 – ***Le continuum désordre-ordre.*** *Nous présentons ici plusieurs systèmes plus ou moins désordonnés. À gauche, la protéine ACTR est principalement désordonnée. Un état plus structuré dit globule fondu est ensuite présenté avec le nuclear-receptor co-activator-binding domain (NCBD). En troisième est présentée un protéine constituée de trois domaines repliés séparés par des motifs flexibles. Une protéine repliée possédant sa partie N-terminale dépliée représente l'état le plus structuré. L'état désordonné n'est pas permanent, une modification des conditions biologiques, i.e. l'ajout d'un partenaire, peut induire un repliement partiel ou total des régions désordonnées. Est affiché à gauche le complexe ACTR-NCBD (en orange et en vert). Le second cadre montre une structure bien définie des trois zinc-fingers of TFIIIA liés à un oligonucleotide. Le cadre à droite montre le complexe entre IF4E (orange) et eIF4G (vert) impliquant un repliement de la partie N-terminale de eIF4E (jaune). Figure extraite de Dyson and Wright [30].*

La concept de continuum ordre-désordre est tout aussi valable en prenant comme référence les protéines repliées. Il est d'usage de représenter ces protéines comme une structure 3D fixe, cette vision est cependant simplificatrice de la réalité, il existe de nombreux mouvements et vibrations existants même pour ces protéines. Ces mouvements peuvent par exemple être liés à la présence d'un état excité [31]. Pour illustrer nos propos, nous pouvons nous référer tout d'abord aux travaux portant sur la dynamique des protéines structuréesKay et al. [32], Salmon et al. [33]. La présence de désordre ou de flexibilité est inévitable même au sein des protéines repliées et a probablement un rôle capital dans le fonctionnement de ces protéines.

1.3 Flexibilité et plasticité structurale des protéines désordonnées

Initialement, le désordre conformationnel a été considéré comme un facteur peu enclin à la réalisation des fonctions biologiques. Cette idée a fortement été révisée depuis. Le désordre se traduit concrètement par une flexibilité accrue de la chaine principale des protéines désordonnées et peut aboutir à deux effets majeurs : l'un est une surface d'association nettement plus importante que pour une protéine repliée, l'autre est la plasticité structurale, ou conformationnelle, qui permet l'association avec

plusieurs partenaires [34]. La protéine adopterait alors différentes structures suivant le partenaire rencontré. Ainsi, d'un point de vue fonctionnel, la présence de désordre est potentiellement bénéfique lors de l'appariement avec un partenaire [35].

À titre d'exemple, ces protéines peuvent être impliquées dans ce que l'on appelle : *one-to-many binding* et *many-to-one binding*. La partie C-terminale de la protéine p53, intrinsèquement désordonnée, peut se lier avec quatre partenaires différents. Suivant le partenaire considéré, la protéine adopte différentes structures transitoires localisées sur différents résidus : p53 forme une hélice α liée à S100$\beta\beta$, un feuillet β liée à la sirtuin et à une conformation étendue lorsqu'elle se lie à CBP et la cycline A2 [36]. Nous parlons alors de plasticité structurale.

1.4 Les transitions désordre-ordre

Dans l'exemple de la section précédente, nous avons mentionné qu'en plus de leur flexibilité, certaines protéines désordonnées sont caractérisées par des transitions désordre-ordre lors de l'appariement leur partenaire. En effet, l'association avec une autre protéine peut induire un repliement local et transitoire en structure secondaire. Ce repliement potentiellement nécessaire à la reconnaissance moléculaire est un sujet primordial pour la compréhension des mécanismes biologiques. Pour expliquer ces arguments, nous allons brièvement présenter un système étudié au laboratoire concernant les bases moléculaires de la réplication du virus de la rougeole.

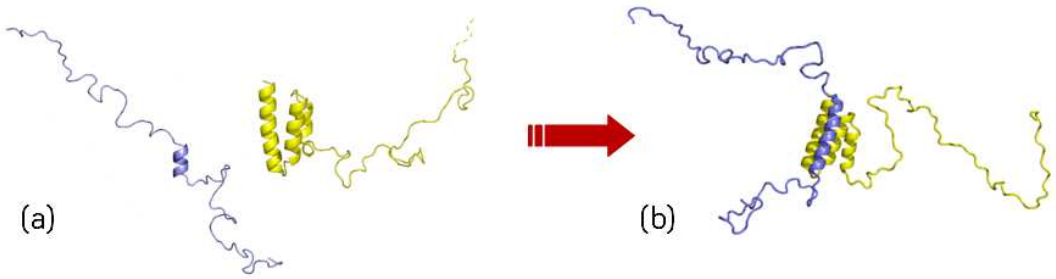


Figure 1.2 – ***Repliement partiel de N_{tail} avant appariement avec la protéine P.*** *(a) : La forme N_{tail} isolée possède une propension à former des hélices α et cette structuration transitoire est exactement située sur la zone d'appariement avec son partenaire physiologique P. (b) : Lors de l'appariement la protéine N_{tail} effectue un repliement plus important.*

Le génome du virus de la rougeole est enfermé à l'intérieur d'une capside assemblée à partir de multiples copies de la nucléoprotéine N. La partie intrinsèquement désordonnée C-terminale de la nucléoprotéine N nommée N_{tail} s'associe avec la phosphoprotéine P pour permettre la transcription et la réplication du virus (1.2). Il a été démontré par résonance magnétique nucléaire que la forme N_{tail} isolée possède une propension à former des hélices α et que cette structuration transitoire est exactement située sur la zone d'appariement avec son partenaire physiologique P. Le repliement transitoire joue un rôle primordial dans la spécificité de la signalisation de cette interaction [37], il facilite potentiellement la reconnaissance moléculaire de la protéine N_{tail} par la protéine P.

A l'issu de ces travaux, une question demeure, elle concerne la cinétique de l'interaction. L'interaction avec un partenaire est caractérisée par un repliement transitoire des protéines désordonnées. La nature et l'instant où ce repliement est effectué sont aussi des points clés de la reconnaissance moléculaire des protéines désordonnées. Deux

voies possibles d'interaction, présentées en figure sont 1.3 :

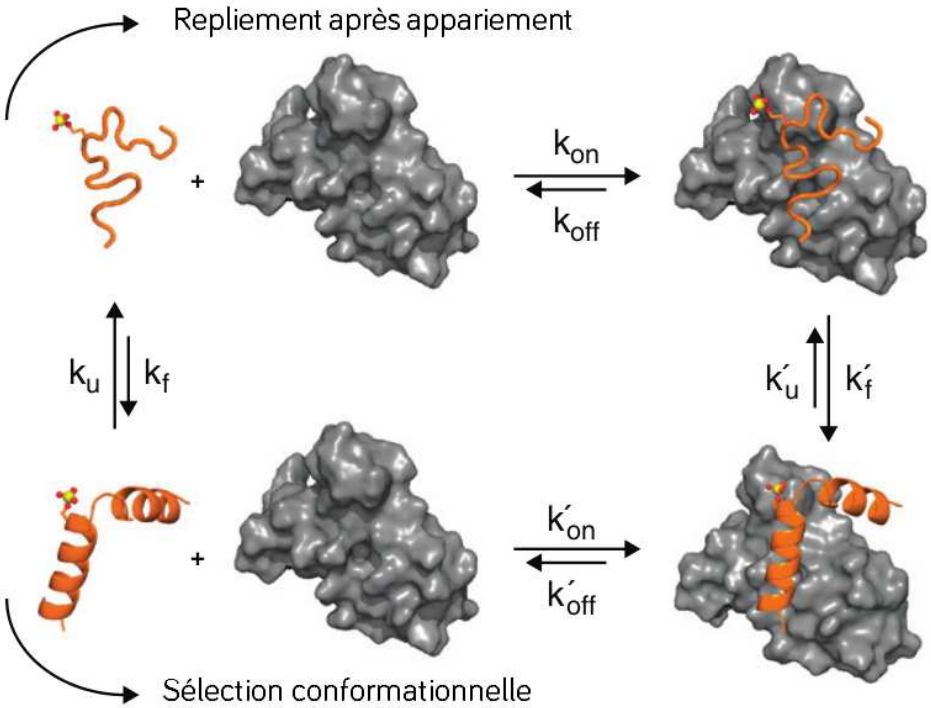


FIGURE 1.3 – *Présentation des deux voies d'interaction possibles pour des protéines dépliées. Il y a sélection conformationnelle lorsque l'état replié de la protéine est choisi par le ligand. De manière opposée nous avons le mécanisme de repliement après appariement. Les valeurs k_{off}, k_{on}, k_u et k_f sont les constantes d'échanges associées. Figure extraite de Kiefhaber et al. [38].*

- La sélection conformationnelle (*conformational selection*). Le repliement partiel en éléments structurés facilite la reconnaissance moléculaire en proposant avant l'appariement un site d'association au partenaire.

- Le repliement après appariement (*induced fit*). La protéine dépliée se lie aux ligands puis se replie.

Les deux situations présentées sont difficiles à discriminer, les deux mécanismes peuvent coexister même si l'un prime sur l'autre. Le mécanisme de sélection conformationnel est supposé largement indépendant de la concentration en ligand, pour autant la présence de ligand supplémentaire pourrait aussi favoriser la présence de la bonne conformation. La compréhension de la cinétique et des échelles de temps décrivant ces mécanismes sont nécessaires pour comprendre les bases de la reconnaissance moléculaire des protéines désordonnées.

1.5 La séquence : une spécificité des protéines intrinsèquement désordonnées

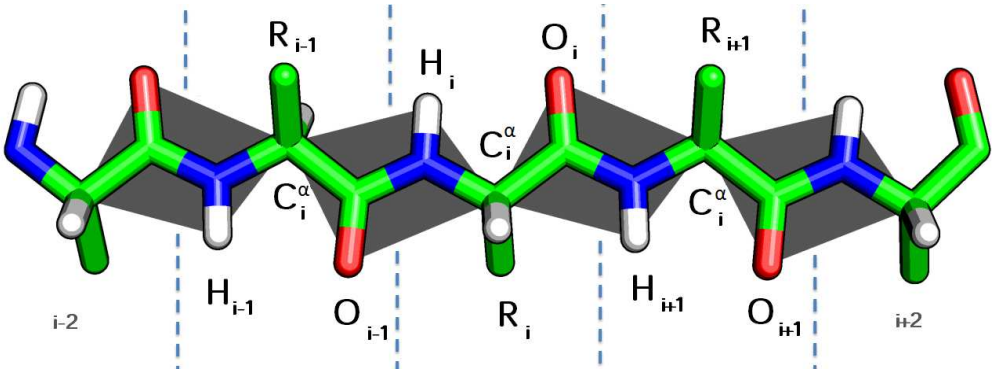


Figure 1.4 – ***Représentation schématique de 5 résidus et des plans peptidiques associés.*** *Les résidus sont tous identiques en l'absence de chaine latérale spécifique. Les pointillés délimitent les acides aminés.*

D'un point de vue chimique ou physique, les protéines peuvent être considérées comme des hétéro-polymères, elles sont composées d'acides aminés reliés entre eux par des liaisons peptidiques (figure 1.4). Ces acides aminés, au nombre de 20, possèdent un groupe amine (NH_2) et une fonction acide carboxylique (COOH). Ils se distinguent par leur chaine latérale R . La composition en acides aminés d'une protéine constitue la structure primaire. La majeure partie de l'information définissant la forme d'une protéine ordonnée, et par conséquent sa fonction, est directement codée dans la structure primaire, c'est un des postulats fondamentaux de la biologie structurale. Le parallèle avec les protéines désordonnées reste vrai.

La composition en acides aminés

Une des premières découvertes associées aux protéines désordonnées fut leur composition en acide aminé. De manière générale, les protéines désordonnées contiennent peu de résidus hydrophobes qui habituellement constituent le coeur des protéines repliées mais une forte propension de résidus chargés ou polaires. La présence de domaines similaires ou répétitions est aussi une caractéristique de ces protéines.

Il a alors été possible de classer les acides aminés en trois classes : les résidus favorisant l'ordre : C, W, Y, I, F, V, L, H, T, N, les résidus favorisant le désordre E, P, Q, S, R, K, M, D et les résidus neutres A et G. Pour simplifier, les résidus hydrophobes ont tendance à favoriser l'ordre tandis que les résidus chargés ou polaires ont tendance à favoriser le désordre [7, 39, 40] (figure 1.5).

1.6 Les logiciels de prédiction du désordre

Les protéines désordonnées se caractérisent par une flexibilité ou dynamique inhérente bien plus élevée à la plupart des protéines jusqu'alors étudiées. Ce champ de recherche en biologie structurale est en effet très récent et son importance a énormément accru depuis le début des années 2000. La place des protéines désordonnées au sein du

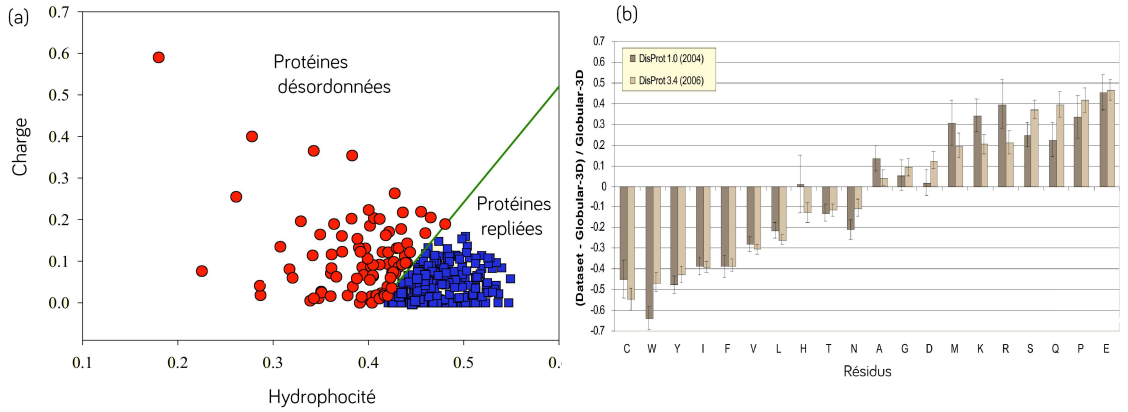


FIGURE 1.5 – ***Caractéristique de la séquence d'acides aminés des protéines désordonnées.*** *(a) : Le cadre en haut présente la charge en fonction de l'hydrophobicité d'un jeu de 275 protéines repliées (en bleu) et 91 protéines dépliées (en rouge). (b) : L'histogramme présente la composition en acide aminé de protéines ou régions (supérieures à 10 résidus) intrinsèquement désordonnées. Le gris foncé correspond à 152 protéines extraites de la base de données de DisProt 1.0, le gris clair correspond à 460 protéines issues de DisProt 3.4. Figure extraite de Dunker et al. [41].*

génome humain a été mise en évidence grâce aux outils de la bio-informatique. Elle a joué un rôle considérable dans le développement du champ de recherche consacré aux protéines intrinsèquement désordonnées. À ce jour, seule l'analyse computationnelle permet d'estimer rapidement le degré de désordre d'un jeu de protéines en s'appuyant sur la spécificité de leur séquence. La prédiction du désordre des protéines a notamment permis aux chercheurs d'estimer l'importance du désordre au sein du génome et d'identifier les potentielles fonctions associées aux protéines désordonnées en caractérisant les sites d'interactions tels que les sites d'associations [42, 11] ou les sites de modifications post-traductionnelles [43, 44].

Un nombre important de prédicteurs de désordre sont apparus au cours des dernières années [45, 46, 47]. Les algorithmes de ces programmes diffèrent cependant : certains s'appuyant sur la charge et l'hydrophobicité des résidus, d'autre sur le poids relatif des acides aminés au sein de la séquence ou encore sur l'énergie d'interaction entre résidus définis au sein d'un motif de la protéine. Il existe maintenant un tel nombre de prédicteurs qu'il a été développé en conséquence des métaprédicteurs combinant les différentes prédictions de ces algorithmes et permettant de peser plus raisonnable les propensions de ces algorithmes pour décrire les régions ou domaines désordonnées des protéines. Nous citerons à titre d'exemple deux algorithmes, la liste est évidemment non exhaustive :

- IUPred [48] estime l'énergie d'interaction entre résidus sur une fenêtre d'acides aminés. En d'autres termes, l'algorithme étudie la capacité d'une chaine polypeptidique à former des contacts. Les paramètres pris en compte sont le type d'acide aminé, la propension de l'acide aminé au désordre.

- DisEMBL [49] utilise un réseau de neurones artificiels pour prédire la propension au désordre d'une protéine. Il s'appuie sur trois critères : le degré de mobilité du résidu, l'absence de coordonnées dans les structures par diffraction aux rayons X et la propension à échantillonner une structure secondaire précise.

Bien que la valeur exacte de chiffre soit soumise à de nombreuses incertitudes, ces logiciels prédisent l'existence de 45-50% de régions[1] désordonnées au sein du génome

1. C'est-à-dire constitué d'au moins 30 résidus contigus

eucaryote. L'application de cette même analyse computationnelle au génome procaryote indique 30% de régions désordonnées [4].

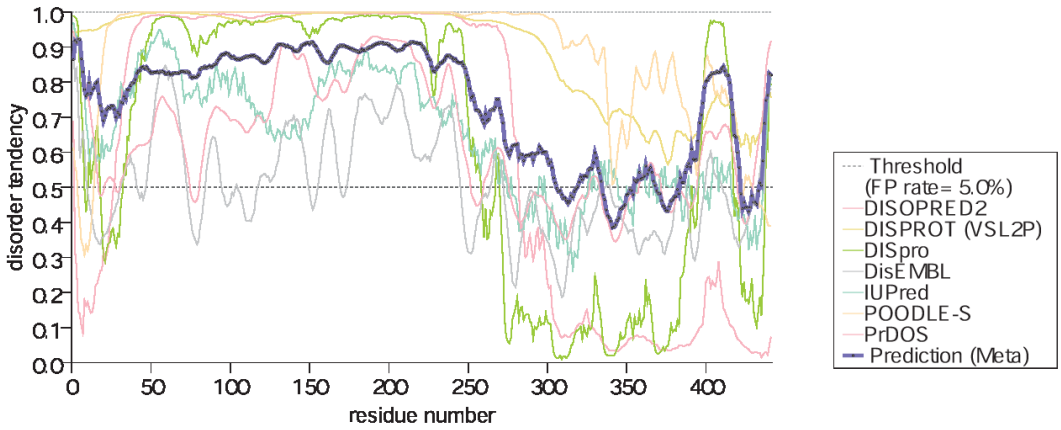


FIGURE 1.6 – ***Application du metaprédicteurs metaPrDOS à Tau.*** *Application du metaprédicteurs metaPrDOS [50] sur la protéine Tau htau40 de 441 résidus. La valeur 1 indique une protéine à fort désordre, la valeur 0 indique une protéine à faible désordre. La protéine est globalement désordonnée, à l'exception de la région allant des résidus 250 à 380 et la partie N-terminale qui présente potentiellement des structures transitoires.*

Ces approches ont grandement contribué à la reconnaissance de l'état déplié, elles présentent cependant leur limite pour étudier le degré de désordre présent dans une protéine à l'échelle atomique. À titre d'exemple, nous présentons en figure 1.6 la prédiction du degré de désordre de la protéine Tau humaine complète htau40. Les prédicteurs indiquent une protéine désordonnée, exception faite de la partie N-terminale (résidu 420 à 440) et de la région des répétitions, domaine d'appariement de la protéine aux microtubules (résidu 250 à 350). Le degré de désordre indiqué pour cette région varie de 0.1 à 0.9 suivant le prédicteur considéré, ce qui indique clairement la nécessité de recourir à des outils plus précis pour étudier l'état déplié. Une caractérisation structurale de la protéine Tau sera effectuée aux chapitres 5 et 7.

1.7 La région Polyproline : une seconde spécificité des protéines désordonnées

En dépit des progrès constants réalisés en biologie structurale, les connaissances concernant la région Polyproline (PPII) sont relativement en marge, cette région reste la plus méconnue de l'espace Ramachandran. Il a été estimé à 10% la part des résidus de structures connues échantillonnant cette région et à seulement 2% la part des résidus formant des hélices de type PPII d'une longueur supérieure à 4 résidus Rucker et al. [51].

Hélice Polyproline II

Une hélice Polyproline II est un type de structure secondaire. Nous désignons par hélice PPII ou hélice Polyproline II une suite de résidus conjoints adoptant les angles dièdres suivants $(\phi, \psi) = (-75, +145)$ et comptant une périodicité de 3 résidus exactement.

Il a précédemment été suggéré que l'inclusion du solvant à proximité de la chaine principale est la force principale guidant la formation des hélices PPII. Cette conformation étant spatialement étendue, l'orientation des atomes de carbone et d'azote favorise les interactions avec les molécules d'eau du solvant.

Les hélices PPII sont notamment courantes au sein des séquences riches en Proline, caractéristique des protéines désordonnées. Les protéines désordonnées sont très riches en résidu polaire ou chargé tel que les Glutamines, les Asparagines et en petit résidu tel que les Alanines et Glycines. Il a été montré qu'un nombre important de ces résidus ont tendance à échantillonner fortement la région Polyproline de l'espace Ramachandran. En effet, courant des années 2000, de nombreuses études se sont focalisées sur la place de la région Polyproline au sein des protéines non structurées. Des mesures de dichroïsme circulaire couplées à des mesures par résonance magnétique nucléaire ont permis de progresser dans la compréhension des conformations PPII. Différentes hypothèses concernant le rôle des ces conformations sont apparues, Blanch et al. [52] soumettent l'idée que les hélices PPII sont des précurseurs de la formation des plaques amyloïdes, ces travaux seront notamment soutenus par Eker et al. [53]. Différentes études concordent sur plusieurs points importants relatifs aux propriétés conformationnelles des protéines non-structurées : le niveau de structuration en hélice PPII serait dépendant en température [54, 51], une diminution de la température favorisant cet effet.

Plusieurs méthodes biophysiques indiquent que la région Polyproline est une des spécificités des protéines désordonnées. Nous confirmons cette tendance au chapitre 5 en étudiant l'échantillonnage conformationnel de deux protéines désordonnées par résonance magnétique nucléaire et insisterons sur la place de ces conformations dans l'état déplié.

1.8 Les échelles de temps : une troisième spécificité des protéines désordonnées

Il est d'usage de les représenter par un ensemble de structures en interconversion : les protéines intrinsèquement désordonnées sont présumées fluctuer rapidement entre différentes conformations ou sous-états. Le temps définissant l'interconversion des conformations est un élément crucial pour comprendre les propriétés des PIDs, pour autant sa définition est relative au modèle proposé : considérant un continuum de conformations et une transition symbolisée par le passage de la conformation i à la conformation $i+n$, le temps d'interconversion entre ces deux états sera fonction des différences structurelles existantes entre ces dernières. Il n'y a donc pas définition absolue des temps caractéristiques des protéines désordonnées.

L'hypothèse standard consiste à considérer un temps d'interconversion de l'ordre de la microseconde à la nanoseconde. Ce phénomène est donc plus lent que le temps de

repliement en structure secondaire estimé à quelques nanosecondes par des mesures de cinétiques ultra-rapides [55, 56] ou les vibrations locales de la chaine principale existant sur une échelle allant de la nanoseconde à la picoseconde.

Conclusion du chapitre

L'étude des protéines intrinsèquement désordonnées est un domaine de recherche récent. L'introduction du désordre dans le génome a initialement fortement perturbé la communauté scientifique et l'a amené à réviser ou étendre les modèles en cours. À l'issu de ce parcours, la notion de désordre ou de flexibilité se révèle très utile pour expliquer de nombreux mécanismes biologiques jusque-là inconnus ou ignorés. La flexibilité est en effet un atout fonctionnel : la plasticité structurale et la grande surface d'association à des partenaires offrent à ces protéines un rôle clé dans de nombreux mécanismes de régulation, de signalisation. Un autre point important concerne l'implication de ces protéines dans plusieurs processus pathogènes.

La compréhension de ces mécanismes nécessite préalablement une étude structurale de ces protéines, c'est-à-dire une meilleure compréhension des conformations adoptées et des échelles de temps définissant les caractéristiques de ces dernières. La description moléculaire d'un tel système est un sujet compliqué à traiter, notamment de par son caractère dynamique. Une des méthodes expérimentales les plus appropriées est la résonance magnétique nucléaire que nous allons introduire au chapitre suivant.

La résonance magnétique nucléaire appliquée aux protéines intrinsèquement désordonnées

2

La résonance magnétique nucléaire est une méthode physique utilisant les propriétés magnétiques des noyaux possédant un spin nucléaire pour obtenir sur des échantillons en solution ou en solide de l'information structurelle et dynamique.

Principe physique de la RMN

La RMN est une méthode de perturbation : considérant comme système à étudier un spin nucléaire unique soumis à un champ magnétique constant B_0, ce spin va s'aligner avec le champ magnétique. Le système étant à son équilibre, nous le perturbons par l'application d'ondes radio-fréquence, le système évolue alors dans un état excité déterminé à la fois par l'environnement physico-chimique et par les caractéristiques des ondes envoyées. Nous mesurons alors l'évolution et le retour à l'équilibre du spin, c'est-à-dire la réponse induite du système à cette perturbation.

L'étude de macromolécules biologiques telles que de protéines complexifie évidemment la problématique, cependant le principe physique est identique à celui mentionné. La RMN permet d'obtenir de l'information à l'échelle atomique, spécifique à l'environnement chimique de chaque spin de la protéine.

La résonance magnétique nucléaire est ainsi un outil de choix pour étudier les protéines désordonnées à l'échelle atomique, ce qui explique l'intérêt grandissant pour développer ou adapter les méthodes de RMN aux protéines désordonnées. Nous aborderons uniquement dans cette thèse la RMN en liquide, la RMN en solide offre un panel d'applications impressionnantes avec notamment l'étude des fibrilles, conséquence directe de l'agrégation de nombreuses protéines désordonnées impliquées dans des maladies neurodégénératives, qui est actuellement un champ d'application très en vogue. Nous citons à titre d'exemple quelques systèmes récemment étudiés : les fibrilles de la protéine Tau [57, 58], les fibrilles amyloïdes β [59, 60], les fibrilles amyloïdes de la protéine α-synucléine [61].

Un second atout que confère la RMN est la grande diversité des interactions mesurables. Nous présenterons les interactions les couplages spin-spin, spin-lattice et spin-électron. Ces interactions sont principalement intramoléculaires mais peuvent sous certaines conditions être intermoléculaires. Les paragraphes suivants présentent successivement les concepts clés utilisés : le déplacement chimique, les couplages scalaires, les couplages dipolaires résiduels et la relaxation paramagnétiques. Les bases théoriques seront donc expliquées de manière à mettre en relief les applications potentielles de ces

interactions dans l'étude des protéines désordonnées. Ainsi, nous tacherons de garder un formalisme épuré, les détails calculatoires seront si nécessaire reportés en annexe pour plus de clarté.

2.1 Description quantique pour un spin 1/2

Plus formellement, la résonance magnétique nucléaire permet de caractériser les transitions énergétiques entre spins nucléaires. Les propriétés des spins nucléaires sont définies par le nombre quantique de spin s. L'opérateur moment cinétique intrinsèque (ou spin) $\hat{S}$ est associé au nombre quantique de spin s et peut être caractérisé à partir des deux opérateurs suivants :

- $\hat{S}^2$ décrit la norme du moment cinétique de spin et est quantifié par s(s+1).

- $\hat{S}_z$ représente la projection du moment cinétique le long de l'axe z. Il est quantifié à partir du moment magnétique de spin m qui varie entre -s et +s par pas de 1.

Le moment magnétique de spin $\hat{\boldsymbol{\mu}}_s$ est associé au moment cinétique de spin $\hat{S}$ par l'équation :

$$\hat{\boldsymbol{\mu}}_s = \gamma\hbar\hat{\boldsymbol{S}} \tag{2.1}$$

où γ est le ratio gyromagnétique du noyau en question, c'est une propriété intrinsèque du noyau.

En présence d'un champ magnétique statique $\boldsymbol{B}^0$, l'Hamiltonien[1] d'interaction entre le champ magnétique et le moment magnétique de spin s'exprime :

$$\hbar\mathcal{H}^Z = -\hat{\boldsymbol{\mu}}_s\boldsymbol{B}^0 = -\gamma\hbar\hat{\boldsymbol{S}}\boldsymbol{B}^0 \tag{2.2}$$

Le cas le plus simple à traiter correspond à un spin 1/2 isolé, l'Hamiltonien a alors deux valeurs propres notées α et β correspondant à m=+1/2 et m=-1/2. L'application du champ B^0 induit une levée de dégénérescence ou une séparation des niveaux d'énergies. Les énergies respectives sont alors :

$$E_\alpha = -\frac{1}{2}\hbar\gamma B^0 \qquad \text{et} \qquad E_\beta = +\frac{1}{2}\hbar\gamma B^0 \tag{2.3}$$

Ainsi la différence d'énergie entre les deux états s'écrit :

$$\Delta E = \hbar\gamma B^0 = h\nu_0 \qquad \Longrightarrow \qquad \nu_0 = \frac{\gamma B^0}{2\pi} \tag{2.4}$$

La séparation des niveaux d'énergies sera d'autant plus nette si le champ est intense.

Le moment magnétique de spin n'est pas statique mais tourne autour du champ B^0 à la vitesse de la fréquence de larmor ν_0 défini dans l'équation 2.4. Les transitions entre l'état α et β sont observables à cette fréquence et constitue la base de la spectroscopie RMN : l'intensité du signal RMN dépend de la différence de population entre les deux états d'énergie du noyau. Cette différence peut être quantifiée en utilisant la distribution de Boltzmann. Considérant N_α et N_β les populations respectives des deux niveaux d'énergie précédents, le rapport de populations s'exprime :

$$\frac{N_\alpha}{N_\beta} = \exp\left(\frac{\Delta E}{kT}\right) = \exp\left(\frac{\gamma\hbar B^0}{kT}\right) \approx 1 + \frac{\gamma\hbar B^0}{kT} \tag{2.5}$$

Pour un proton plongé dans un champ magnétique de 18.8 Tesla (800 MHz) à 25°C , la rapport de population est :

$$\frac{N_\alpha}{N_\beta} = 1.00013 \tag{2.6}$$

1. On parle d'Hamiltonien Zeeman ou d'interaction Zeeman

Bien que très faible, l'excès de population dans l'état α induit une modification de l'aimantation mesurable.

Une dernier point à mentionner concerne le champ magnétique observé, ce dernier n'est pas exactement le champ appliqué, la présence des électrons induits une constante d'écran modifiant légèrement la valeur de B^0 observé. Cette manifestation appelée le déplacement chimique est particulièrement utilisée en RMN des protéines, nous allons la présenter dans la section suivante.

2.2 Le déplacement chimique

Le déplacement chimique est la signature chimique de l'environnement électronique d'un noyau observé. Appliquant un champ externe statique sur un noyau, un champ induit est créé par l'intermédiaire des électrons, le champ magnétique observé est la somme du champ induit et du champ statique. Considérant les champs magnétiques observés sur deux noyaux dans deux sites de la même molécule, si l'environnement électronique diffère entre les deux sites, il en résultera un champ observé différent.

Nous allons maintenant présenter le formalisme associé aux déplacements chimiques. Le spin nucléaire est sensible à la somme du champ statique $\boldsymbol{B}^0$ et du champ induit $\boldsymbol{B}^{induit}$ généré par le nuage électronique :

$$\boldsymbol{B}^{loc} = \boldsymbol{B}^0 + \boldsymbol{B}^{induit} \qquad \text{avec} \qquad \boldsymbol{B}^{induit} = \boldsymbol{\delta}\boldsymbol{B}^0 \tag{2.7}$$

où $\boldsymbol{\delta}$ est une matrice 3*3 appelée le tenseur du déplacement chimique du spin $\boldsymbol{S}$. Ainsi :

$$\boldsymbol{B}^{loc} = \boldsymbol{B}^0(1+\boldsymbol{\delta}) \tag{2.8}$$

En un milieu isotrope, seules les valeurs principales du tenseur sont conservées qui équivaut alors à :

$$\delta^{iso} = 1/3\left(\delta^{XX} + \delta^{YY} + \delta^{ZZ}\right) \tag{2.9}$$

où les termes δ^{XX}, δ^{YY} et δ^{ZZ} dépendent chacun de l'orientation moléculaire de la protéine mais leur somme y est invariante.

Combinant l'équation 2.4 et l'équation 2.8 et considérant que le champ ressenti n'est pas vraiment B^0 mais le champ local B^{loc}, nous avons les expressions respectives :

$$\nu^0 = -\gamma B^0\left(1+\delta^{iso}\right)/(2\pi) \tag{2.10}$$

Selon cette expression, la fréquence de Larmor serait dépendante du champ magnétique. Pour pouvoir comparer les spectres entre eux a été introduit la notion de déplacement chimique δ tel que :

$$\delta_{ppm} = 10^{-6} * \frac{\nu^0 - \nu^{ref}}{\nu^{ref}} \tag{2.11}$$

où ν^0 est la fréquence de Larmor en Hertz et ν^{ref} une fréquence de référence. Le tétraméthylisilane (TMS) et le 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) sont les composés courants utilisés comme référence : $\delta(^1H^N$ TMS)=0 ppm.

2.3 Échelle de temps et échange chimique

2.3.1 Échelle de temps et échange chimique

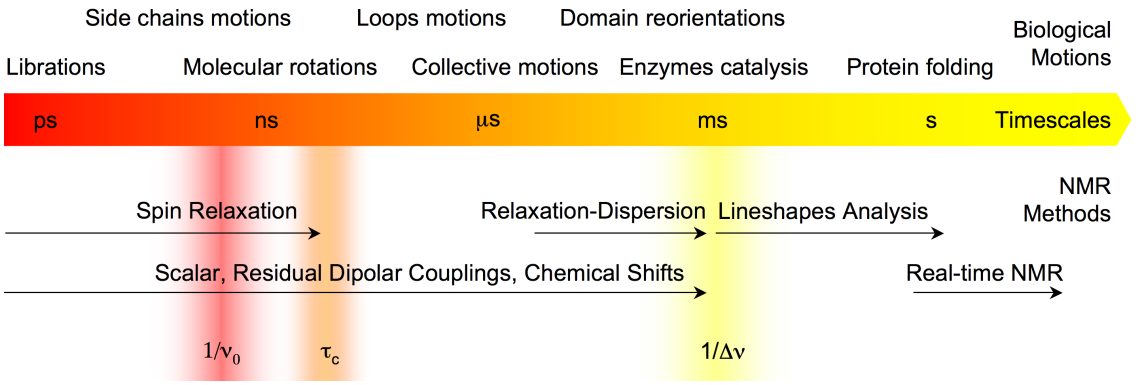


FIGURE 2.1 – ***Les échelles de temps et mouvements associés aux protéines.*** *ν_0 est la fréquence de Larmor du noyau étudié, τ_C est le temps de corrélation globale de la protéine, $\Delta\nu$ est la différence de fréquence entre les deux conformations. Les mouvements caractéristiques des protéines et les expériences RMN associées à leur étude sont affichés.*

La spectroscopie RMN est une méthode extrêmement sensible aux échelles de temps et particulièrement adaptée à l'étude de la dynamique des protéines [62, 63]. La figure 2.1 présente les différents mouvements existant au sein des protéines. Pour chacun d'entre eux il est possible d'associer une ou plusieurs expériences de RMN permettant de caractériser les phénomènes physiques en jeu. Nous allons les présenter brièvement.

Les phénomènes physiques s'effectuant à une échelle de temps plus lente que le temps de mesure du signal RMN peuvent être observés en réalisant des mesures en temps réel [64]. A l'opposé, la RMN offre une forte sensibilité à des mouvements s'effectuant sur des gammes nettement plus rapide.

Les mouvements s'effectuant à des échelles de temps proche de la fréquence de Larmor influenceront la relaxation du système de spin et modifieront donc les propriétés du signal RMN. La dynamique de ces mouvements est ainsi détectable par des expériences de relaxation [65]. La valeur limite de l'échelle de temps étant le temps de corrélation globale traduisant la rotation de la molécule dans le milieu qui est de l'ordre de 5 à 20 ns pour une protéine repliée de taille moyenne à température ambiante [66].

Une seconde échelle de temps importante est la limite définissant l'échange chimique. L'échange chimique a lieu lorsqu'un même noyau observe deux environnements chimiques différents. L'exemple le plus concret est l'adoption par le noyau de deux conformations A et et B en échange, de fréquence respective ν_A et ν_B. La mesure du déplacement chimique sera fonction des gammes de temps du mouvement relativement à la différence de fréquence entre les deux conformations et impactera ou non le spectre RMN obtenu.

$$A \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} B \tag{2.12}$$

La constante d'échange ou d'interconversion est définie par $k_{ex} = k_{on} + k_{off}$, cette valeur est primordiale et aboutit en fonction du rapport avec $\Delta\nu$ où $\Delta\nu = \nu_B - \nu_A$ aux différents régimes suivants (présentés aussi en figure 2.2) :

- Échange lent : $k_{ex} \ll \Delta\nu$, la mesure RMN donnera alors deux pics positionnés en

ν_A et en ν_B

- Échange moyen : $k_{ex} \sim \Delta\nu$, nous obtenons dans ce cas un élargissement de la raie spectrale

- Échange rapide : $k_{ex} \gg \Delta\nu$, l'échelle de temps d'interconversion des conformations étant plus rapide que l'écart entre les deux déplacements chimiques, nous obtenons un unique pic situé entre les fréquences ν_A et ν_B des deux conformations respectives.

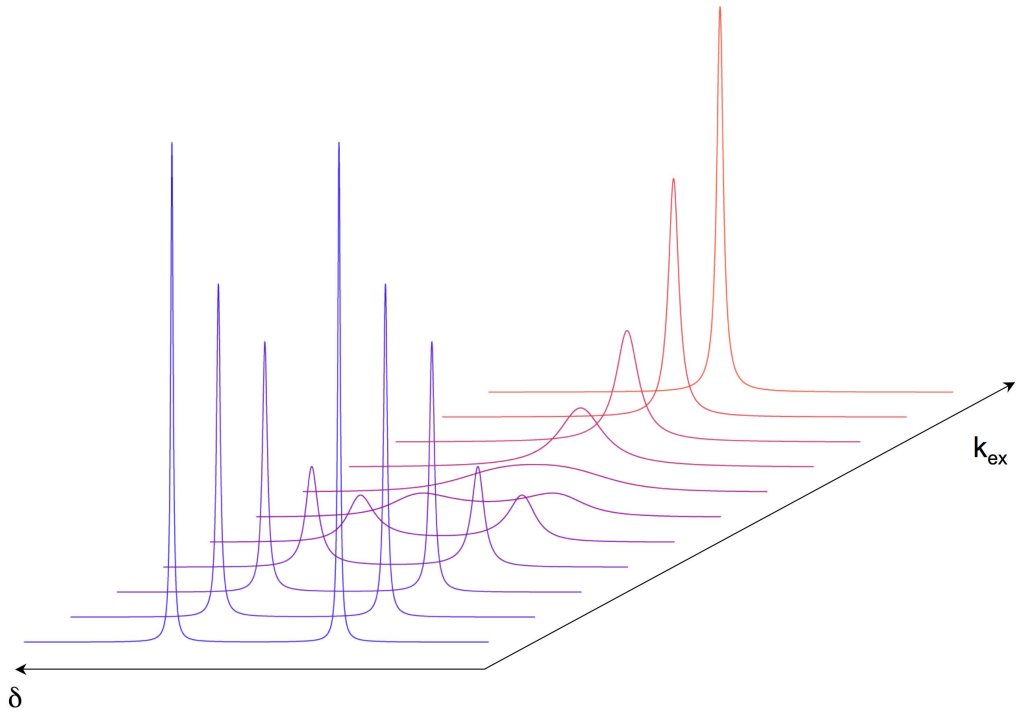


FIGURE 2.2 – ***Représentation de l'échange chimique sur un spectre 1D.*** *Nous sommes en présence de deux sites de populations identiques en échange. Le taux d'échange ou d'interconversion augmente en partant du bleu vers le rouge.*

Ainsi, si le mouvement s'effectue sur une échelle plus lente que la constante d'échange, il peut être observé en analysant directement l'évolution des spectres, si au contraire le mouvement s'effectue à une échelle de plus rapide que la constante d'échange, le mouvement peut être caractérisé par des expériences de relaxation dispersion [67].

Les couplages scalaires, les couplages dipolaires résiduels et les déplacements chimiques sont sensibles à une gamme très large de mouvements. Notamment, les couplages dipolaires se sont révélés des outils extrêmement puissants pour étudier la dynamique des protéines repliées ou dépliées. Ils fournissent de l'information structurale et dynamique en traduisant l'orientation moyenne du vecteur internucléaire relative au champ magnétique statique $\boldsymbol{B}^0$.

Un dernier point à aborder concerne les échelles de temps des protéines intrinsèquement désordonnées, nous avons précédemment souligné que l'interconversion des structures pourrait s'effectuer sur une gamme allant de la microseconde à la nanoseconde, quelles sont les conséquences sur les déplacements chimiques des protéines désordonnées ?

Échange rapide entre conformations

Les protéines intrinsèquement désordonnées sont en échange rapide entre différentes conformations. Par conséquent, le déplacement chimique obtenu est la moyenne dans le temps des signaux de chaque conformation. Nous mesurons donc normalement un spectre possédant un pic unique par résidu indiquant un échange rapide entre les conformations. Par conséquent, l'observable mesurée nous renseigne sur une conformation moyenne adoptée par la protéine représentant l'ensemble des états accessibles de la protéine. Toute la difficulté consiste alors à proposer une méthode quantitative pouvant modéliser ou représenter l'état désordonné à partir d'une information moyenne.

Considérant un ensemble de conformations explicites représentant l'état déplié, la valeur du déplacement chimique CS_i du résidu i sera donc la moyenne des déplacements chimiques du résidu i de chaque conformation k :

$$CS_i = \frac{1}{N} \sum_{k=1}^{N} CS_{i,k} \tag{2.13}$$

Nous considérons un ensemble infiniment grand de conformations représentant l'ensemble des états accessibles de la protéine.

2.3.2 Le déplacement chimique secondaire

Les déplacements chimiques sont les mesures les plus précises qu'ils soient possible de faire par RMN cependant leur utilisation n'en reste pas complexe. Ils sont premièrement dépendants en température, en pH [68], en salinité mais surtout l'information fournie par les déplacements chimiques est relative. Suivant l'acide aminé concerné, la valeur des déplacements chimiques absolue est nettement différente, traduisant l'influence des chaines latérales. A titre d'exemple les déplacements chimiques $^{13}C^{\alpha}$ des Glycines se situent autour de 45 ppm, des Prolines autour de 63 ppm et des Alanines autour de 52 ppm. Cependant la dispersion des valeurs correspondant à chaque acide aminé est relativement similaire et inférieure à ± 4 ppm pour le déplacement chimique $^{13}C^{\alpha}$, ainsi une manière usuelle de représenter les déplacements chimiques d'un hétéropolymère est de soustraire la valeur absolue mesurée à une valeur de référence, nous parlons alors de déplacements chimiques secondaires notés (SCS) pour *Secondary Chemical Shifts*

Le référencement des déplacements chimiques est important pour améliorer la compréhension et l'interprétation des données issues des déplacements chimiques. Considérant un ensemble de protéines dont les déplacements chimiques et la structure sont connus, il a été attribué une valeur médiane pour chaque acide aminé qui sert référence pour calculer les déplacements chimiques secondaires [69]. Il existe différentes tables de référence, nous utiliserons dans ce manuscrit les travaux de Zhang et al. [70] comme référence.

2.4 Manifestation spectrale des protéines désordonnées

L'attribution séquentielle est un pré-requis pour étudier toute protéine par RMN, cela s'applique évidemment aussi aux protéines désordonnées. Un des premiers points constatés est la réduction de l'étendu spectrale des spectres HSQC, notamment selon la dimension proton $^{1}H^{N}$. La nature flexible des protéines désordonnées diminue l'existence de liaisons hydrogène au sein des structures et tend à une uniformisation de

l'environnement électronique des protons qui sont alors en échange avec le solvant. Ainsi, nous observons une diminution de la dispersion spectrale, majoritairement pour la dimension proton $^1H^N$ mais légèrement aussi pour la dimension azote ^{15}N (figure 2.3).

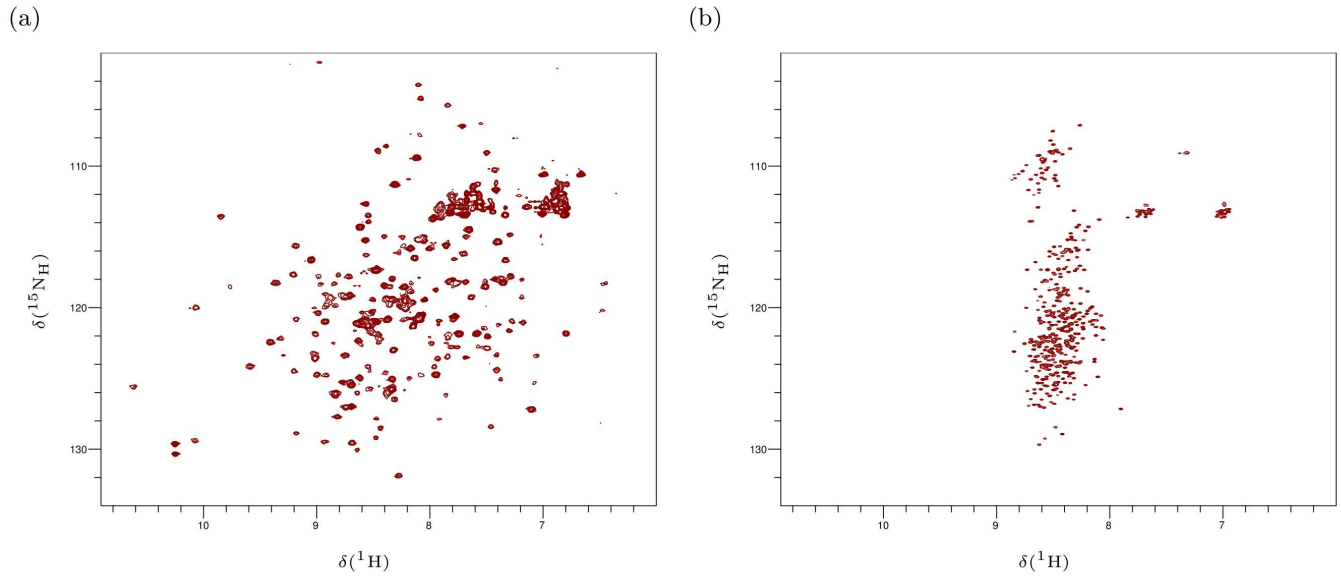


FIGURE 2.3 – ***Comparaison des spectres HSQC d'une protéine repliée et d'une protéine dépliée.*** *(a) : Spectre HSQC d'une protéine homodimérique de* 2*244 *résidus. (b) : Spectre HSQC d'une protéine dépliée de* 441 *résidus (Tau). Les largeurs spectrales azote et proton sont identiques pour les deux cas. Nous observons une diminution de la dispersion spectrale, majoritairement pour la dimension proton* $^1H^N$ *mais légèrement aussi pour la dimension azote* ^{15}N*.*

Une deuxième conséquence des propriétés des protéines désordonnées est la diminution de la largeur de raie. Les vitesses de relaxation des protéines désordonnées sont principalement régies par des mouvements internes liés à la flexibilité de la chaine principale s'effectuant sur des échelles de temps plus rapide que la rotation globale.

Démonstration de ce résultat : la largeur de raie $\nu_{1/2}$ est fonction de la vitesse de relaxation R_2 suivant la formule suivante :

$$\nu_{1/2} = \frac{R_2}{\pi} = \frac{1}{T_2 \pi} \approx A\tau_c \tag{2.14}$$

où T_2 est le temps de relaxation, τ_c le temps de corrélation de la protéine, A une constante.

2.5 Les couplages scalaires

Le couplage scalaire est une interaction spin-spin, il nécessite une liaison covalente entre les noyaux considérés. Le couplage J est une manifestation spectrale de la liaison chimique. Elle traduit l'influence mutuelle des nuages électroniques des noyaux. L'Hamiltonien décrivant l'interaction dipôle dipôle indirecte entre le spin $\mathcal{S}_j$ et $\mathcal{S}_k$ s'écrit :

$$\hat{\mathcal{H}}^J = 2\pi \hat{\mathcal{S}}_j \boldsymbol{J}_{jk} \hat{\mathcal{S}}_k \tag{2.15}$$

où $\boldsymbol{J}_{jk}$ est le tenseur du couplage J. En milieu isotrope, nous obtenons :

$$\hat{\mathcal{H}}_{jk}^{iso} = 2\pi J_{jk}(S_{jx}S_{kx} + S_{jy}S_{ky} + S_{jz}S_{kz}) \tag{2.16}$$

Bien qu'il existe plusieurs couplages J, nous nous focalisons sur le couplage ${}^3J(H^N H^\alpha)$ nommé ${}^3J$. La dépendance du couplage scalaire ${}^3J$ en fonction de l'angle dièdre ϕ peut être paramétrée selon la relation de Karplus [71] (figure 2.4) :

$$ {}^3J(H^N H^\alpha) = A\cos^2(\phi - 60^\circ) - B\cos(\phi - 60^\circ) + C \tag{2.17} $$

où les paramètres A,B et C ont été optimisé en mesurant la constante de couplage scalaire sur des protéines dont la structure est connue par cristallographie aux rayons X [71, 72]. Connaissant l'équation 3.2.2, les couplages scalaires peuvent être utilisés comme contrainte pour caractériser l'échantillonnage conformationnel des protéines.

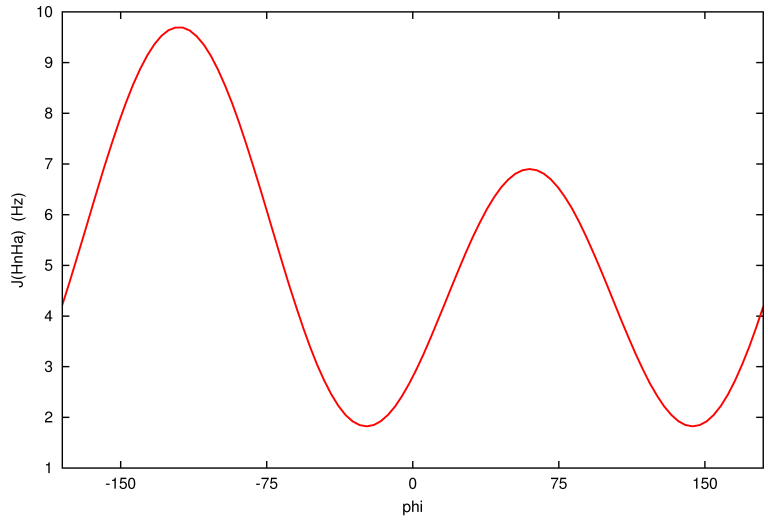


FIGURE 2.4 – ***Dépendance angulaire du couplage scalaire*** ${}^3J$ ***selon la relation de Karplus.*** *La relation de Karplus est présentée en équation 3.2.2. Les constantes utilisées sont A=6.4, B=-1.4 et C=1.9 [71]*

Dans le cas des protéines désordonnées, considérant un ensemble de conformation k en échange rapide, le couplage scalaire ${}^3J_i$ du résidu i correspond à la conformation moyenne et traduit donc l'orientation moyenne ou la distribution de l'angle ϕ au sein de l'ensemble [73, 74] :

$$ {}^3J_i = \frac{1}{N}\sum_{k=1}^{N} {}^3J_{i,k} \tag{2.18} $$

Nous considérons un ensemble infiniment grand de conformations représentant l'ensemble des états accessibles de la protéine.

2.6 Les couplages dipolaires résiduels

Les couplages dipolaires résiduels (CDRs) sont des paramètres très précis permettant de caractériser l'échantillonnage conformationnel des protéines ordonnées ou désordonnées. Nous allons dans cette partie présenter le formalisme de l'interaction dipôle-dipôle, nous calculerons après application de l'approximation séculaire la valeur du couplage en milieu isotrope puis anisotrope. Nous conclurons cette section en expliquant comment interpréter la valeur des couplages dipolaires pour les protéines désordonnées.

2.6.1 Origine de l'interaction dipolaire

Chaque spin nucléaire est associé à un moment magnétique de spin et génère un champ magnétique dans l'espace environnant. Deux spins nucléaires $\boldsymbol{I}$ et $\boldsymbol{S}$ se trouvant dans un même voisinage interagiront donc chacun avec l'autre : le premier "voyant" le champ magnétique créé par le second et inversement. L'interaction est mutuelle (figure 2.5) :

$$E_{\mathrm{IS}} = -\boldsymbol{\mu}_{\mathrm{I}}.\boldsymbol{B}_{\mathrm{S}} = -\boldsymbol{\mu}_{\mathrm{S}}.\boldsymbol{B}_{\mathrm{I}} \tag{2.19}$$

avec

$$\boldsymbol{B}_I(r_{\mathrm{IS}}) = \frac{\mu_0}{4\pi r_{\mathrm{IS}}^3}\left[3(\boldsymbol{\mu}_I.\boldsymbol{e}_{\mathrm{IS}})\boldsymbol{e}_{\mathrm{IS}} - \boldsymbol{\mu}_I\right] \tag{2.20}$$

où $\boldsymbol{r}_{\mathrm{IS}}$ est le vecteur internucléaire joignant le noyau I au noyau S et $\boldsymbol{e}_{\mathrm{IS}} = \frac{\boldsymbol{r}}{r_{\mathrm{IS}}}$ le vecteur associé normalisé.

Contrairement aux couplages scalaires, l'interaction dipolaire ne nécessite pas de liaison électronique entre les deux atomes mis en jeu pour exister. Elle s'effectue sans intermédiaire entre les deux noyaux, c'est pourquoi elle s'appelle "interaction dipôle-dipôle directe". Les couplages dipolaires peuvent donc être aussi bien intramoléculaires qu'intermoléculaires.

Les moments magnétiques associés s'écrivent :

$$\hat{\boldsymbol{\mu}}_{\mathrm{I}} = \gamma_I \hbar \hat{\boldsymbol{I}} \qquad \text{et} \qquad \hat{\boldsymbol{\mu}}_{\mathrm{S}} = \gamma_S \hbar \hat{\boldsymbol{S}} \tag{2.21}$$

L'expression complète de l'interaction dipôle-dipôle entre les spins I et S est donnée par l'Hamiltonien de spin en unité de fréquence :

$$H_{\mathrm{IS}}^D = d_{\mathrm{IS}}(3(\hat{\boldsymbol{I}}.\boldsymbol{e}_{\mathrm{IS}})(\hat{\boldsymbol{S}}.\boldsymbol{e}_{\mathrm{IS}}) - (\hat{\boldsymbol{I}}.\hat{\boldsymbol{S}})) \qquad \text{avec} \qquad d_{\mathrm{IS}} = -\frac{\gamma_{\mathrm{I}}\gamma_{\mathrm{S}} h \mu_0}{16\pi^3 r_{\mathrm{IS}}^3} \tag{2.22}$$

ou γ_{I} et γ_{S} sont les rapports gyromagnétiques des deux spins (en $rad.s^{-1}.T^{-1}$), μ_0 la permittivité du vide, avec h la constante de Planck et r_{IS} la distance entre les deux spins (en m).

2.6.2 L'approximation des hauts champs

Lorsque les spins sont placés dans un champ $\boldsymbol{B}^0$ intense, seule la composante des spins selon la direction du champ est significative dans l'Hamiltonien de l'interaction dipôle-dipôle direct. Ainsi, si l'on suppose que le champ $\boldsymbol{B}^0$ est colinéaire à l'axe z alors l'Hamiltonien de l'équation (2.22) se réécrit de la façon suivante :

$$H_{\mathrm{IS}}^D(\theta_{\mathrm{IS}}) = d_{\mathrm{IS}}\frac{3cos^2\theta_{\mathrm{IS}} - 1}{2} 2\hat{I}_z\hat{S}_z \tag{2.23}$$

où θ_{IS} est l'angle que fait le vecteur internucléaire avec la direction du champ magnétique B^0. Par ailleurs, nous notons :

$$D_{\mathrm{IS}} = d_{\mathrm{IS}}P_2(cos(\theta_{\mathrm{IS}})) \qquad \text{avec} \qquad P_2(cos(\theta_{\mathrm{IS}})) = \frac{3cos^2\theta_{\mathrm{IS}} - 1}{2} \tag{2.24}$$

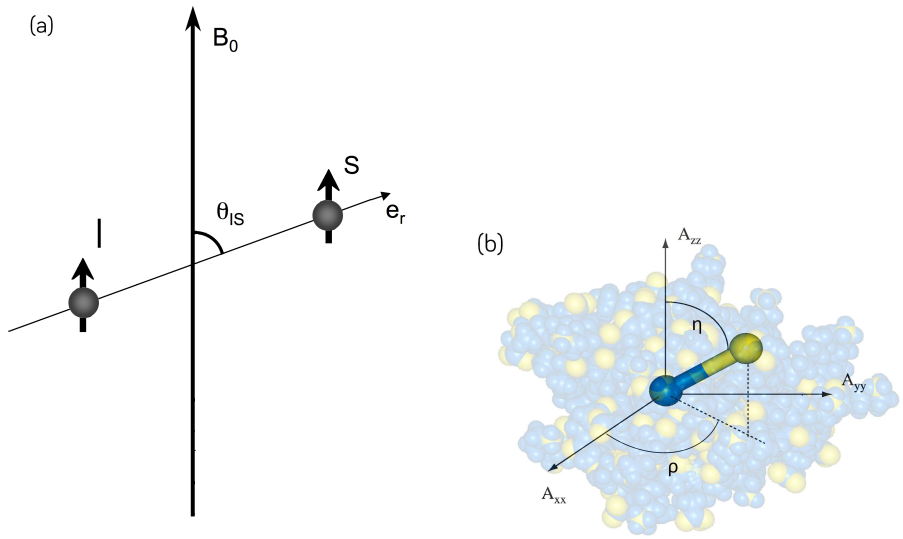


FIGURE 2.5 – ***Le couplage dipolaire résiduel est fonction de l'orientation du vecteur internucléaire par rapport au champ magnétique.*** *(a) : Le couplage dipolaire : le champ induit du spin I modifie le champ effectif pour le spin S (b) : L'orientation du vecteur internucléaire par rapport aux axes principaux x, y et z du tenseur d'orientation A défini les angles η et ρ*

2.6.3 Couplages dipolaires en milieux isotropes

Le cas particulier des liquides isotropes est important car les mesures en RMN des biomolécules s'effectuent couramment dans ce type de milieu. Dans un liquide isotrope, les molécules se réorientent uniformément (θ_{IS} prend toutes les valeurs possibles avec une densité de probabilité uniforme sur l'intervalle de temps de la mesure) et avec une vitesse suffisamment grande pour que seule la moyenne des couplages dipolaires intramoléculaires soit mesurable. La mesure du couplage dipolaire en RMN liquide est décrite par la moyenne dans le temps et sur l'ensemble (I,S) de l'Hamiltonien dipolaire :

$$D_{IS} = \langle d_{IS} P_2(cos(\theta_{IS}(t))) \rangle \approx d_{IS} \langle P_2(cos\theta_{IS}(t)) \rangle \tag{2.25}$$

Une première approximation, le terme $r^3_{IS}(t)$ est constant au cours du temps. Cette approximation est valable pour les couplages opérant entre deux noyaux liés de manière covalente. Les vibrations de la chaine polypeptidique étant nettement plus rapides que l'échelle de temps de la mesure, cela équivaut à remplacer la distance instantanée par une distance effective entre noyaux.

L'isotropie de la réorientation rend la moyenne $\langle P_2(cos(\theta_{IS}(t))) \rangle$ nulle (figure 2.6). Aucun couplage dipolaire intramoléculaire n'est donc mesurable dans ces milieux. D'où :

$$\langle H^D_{IS}(\theta_{IS}(t)) \rangle = 0 \tag{2.26}$$

Ainsi, dans un liquide isotrope, en solution aqueuse par exemple, les couplages dipolaires ne sont pas mesurables. Il n'est aussi pas possible d'avoir accès à l'information de l'interaction dipôle-dipôle sans l'introduction de milieu orientant.

2.6.4 Couplages dipolaires résiduels en milieux orientant

Une des récentes avancées dans le monde de la RMN haute résolution est venue de l'introduction des milieux orientants pour faciliter la mesure des couplages dipolaires résiduels [75, 76, 77]. Dans de tels milieux, les protéines sont partiellement orientées, leur espace orientationnel devient alors suffisamment anisotrope pour permettre la mesure précise de couplages dipolaires résiduels. L'anisotropie induite par ces milieux

reste suffisamment faible pour ne pas perturber le comportement des protéines et préserver au spectre RMN sa lisibilité.

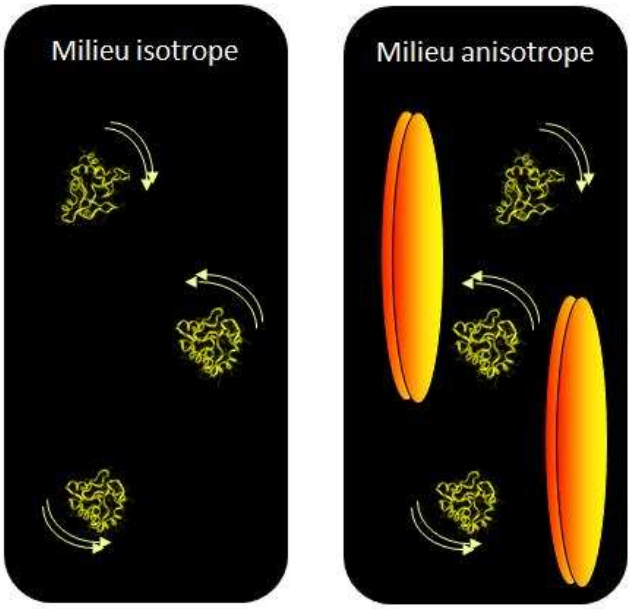


FIGURE 2.6 – ***Milieu isotrope et anisotrope.*** *(a) : En milieu isotrope, le couplage est moyenné à zéro. (b) : En milieu anisotrope, sont représentées des bicelles, la restriction du mouvement moléculaire fait apparaitre le couplage dipolaire résiduel.*

Il existe maintenant beaucoup de milieux orientant qu'on a coutume de classer selon leur forme, leur mode d'interaction et leur comportement physique. Pour qu'il soit possible d'étudier une macromolécule dans ces milieux, l'ordre imposé à la molécule doit être très petit, typiquement moins de 0.002. L'ordre orientationnel des particules du cristal liquide est, quant à lui, grand ; il se situe entre 0.5 et 0.85. Il est clair que les interactions des macromolécules avec le cristal liquide doivent être faibles. Par ailleurs, les milieux doivent être aqueux pour que les protéines soient dans un environnement proche de leur milieu naturel. Ces conditions sont satisfaites en utilisant des solutions aqueuses de cristaux liquides fortement diluées ou les milieux avec les cavités anisotropes. Deux milieux orientants courants sont les gels polyacrylamide étiré [78, 79] et les bicelles [80, 81] qui s'alignent parallèlement au champ magnétique (figure 2.6).

2.6.5 Expression du couplage dipolaire dans le repère lié à la molécule

Ainsi, dans une solution partiellement orientée, la densité de probabilité des valeurs prises par θ (noté précédemment $\theta_{IS}(t)$) est modifiée : il existe des orientations favorisées. La moyenne $\langle P_2(cos(\theta_{IS}))\rangle$ est non nulle.

Afin de simplifier l'interprétation des résultats, il est nécessaire de réécrire la valeur du couplage dipolaire dans une base liée au repère propre de la molécule que l'on supposera en première approximation invariant pour un état conformationnel donné.

Les axes privilégiés d'orientation de la molécule sont définis par un tenseur (matrice 3x3) et le degré d'orientation selon chaque axe par un scalaire. La valeur de l'interaction dipolaire est divisée en deux composantes : axial et rhombique.

L'alignement d'une molécule par rapport au champ magnétique $\boldsymbol{B}^0$ peut être décrit par une matrice réelle et symétrique A'. Cette matrice est diagonalisable et les trois

composantes de la matrice diagonale A'_{xx}, A'_{yy} et A'_{zz} représentent les probabilités des axes x, y et z d'être parallèles à $\boldsymbol{B}^0$. Comme seules les différences entre ces probabilités ont un effet, on peut définir une nouvelle matrice A telle que $\mathrm{tr}(A) = 0$ avec $|A_{zz}| > |A_{yy}| > |A_{xx}|$. L'orientation du vecteur internucléaire par rapport aux axes x, y et z est repérée par les angles η et ρ tels qu'ils sont définis dans la figure 2.5b.

Le couplage dipolaire entre deux spins peut alors être écrit selon l'équation 2.27. Cette équation peut elle-même être simplifiée en (2.28) où $A_a = A_{zz}/2$ est la composante axiale du tenseur d'alignement et $A_r = \frac{1}{3}(A_{xx} - A_{yy})$ est la composant rhombique.

$$D_{\mathrm{IS}}(\eta, \rho) = d_{\mathrm{IS}}[A_{zz}\cos^2\eta + A_{xx}\sin^2\eta\cos^2\rho + A_{yy}\sin^2\eta\cos^2\rho] \tag{2.27}$$

$$D_{\mathrm{IS}}(\theta, \rho) = d_{\mathrm{IS}}[A_a(3\cos^2\eta - 1) + \frac{3}{2}A_r\sin^2\eta\cos 2\rho] \tag{2.28}$$

Pour être capable de mesurer des couplages dipolaires résiduels utilisables, il faut une amplitude du tenseur d'orientation de l'ordre de 10^{-3}. En effet, si cette amplitude est trop petite, les valeurs de couplage dipolaire seront trop petites pour être mesurées précisément. Si par contre elle est trop grande, la complexité des couplages entre spins va rendre le spectre plus difficile à interpréter.

2.6.6 Interprétation du couplage dipolaire pour des protéines repliées

L'interprétation des couplages dipolaires résiduels mesurés sur des protéines repliées est relativement directe. La moyenne des couplages dipolaires résiduels issue de l'équation 2.25 peut, en supposant que l'alignement global n'est pas couplé aux fluctuations rapide de la chaine principale, s'interpréter en terme d'orientation relative du vecteur internucléaire par rapport au tenseur d'alignement global de la molécule, où le tenseur d'alignement décrit l'alignement relatif de la protéine en utilisant un tenseur d'ordre 2 (équation 2.28).

Cette hypothèse a été testée à plusieurs reprises sur des protéines repliées[82] : à partir d'une structure PDB il est possible d'approximer le tenseur d'alignement en estimant l'alignement stérique de cette dernière et ainsi de calculer empiriquement la valeur des couplages dipolaires. Les couplages dipolaires résiduels mesurés dans des molécules repliées présentant une structure stable fournissent de l'information pour l'affinement de structure, l'étude des contacts longues portés sur les chaines étendues et sur les complexes et enfin l'étude de la dynamique locale de la protéine à une échelle de temps allant de la microseconde à la picoseconde [83, 84].

2.6.7 Interprétation du couplage dipolaire pour des protéines dépliées

Dans le cas de protéines intrinsèquement désordonnées, la moyenne de l'équation 2.25 ne s'exprime pas si simplement. L'alignement de chaque molécule par rapport à la moyenne temporelle et à la moyenne sur toutes les conformations de l'ensemble varie de manière significative en fonction de la forme de la conformation.

Considérant un ensemble infiniment grand de conformations, la valeur des couplages dipolaires résiduels d'un ensemble doit donc être exprimée comme la moyenne sur toutes les conformations de la moyenne dans le temps :

$$D_{\mathrm{IS}} = d_{\mathrm{IS}}\frac{1}{N}\sum_{k=1}^{N}\frac{1}{t_{max}}\int_{t=0}^{t_{max}} P_2(cos\theta_{k,\mathrm{IS}}(t))dt \tag{2.29}$$

où N est le nombre de conformations. En supposant que pour un temps suffisamment grand, une protéine échantillonne toutes les conformations de l'ensemble, l'expression se réduit à une moyenne dans le temps des conformations d'une protéine. (hypothèse d'ergodicité)

$$D_{\mathrm{IS}} = d_{\mathrm{IS}} \frac{1}{t_{max}} \int_{t=0}^{t_{max}} P_2(cos\theta_{\mathrm{IS}}(t))dt \tag{2.30}$$

Dans le cas d'une description explicite de structures en échange rapide, les couplages dipolaires peuvent donc être calculés sur chaque conformation et moyennés sur l'ensemble :

$$D_{\mathrm{IS}} = \frac{1}{N} \sum_{k=1}^{N} D_{k,\mathrm{IS}} \tag{2.31}$$

Les couplages dipolaires résiduels sont sensibles à la nature structurale et dynamique du système étudié. Pourtant, en raison de la haute flexibilité de la chaine latérale des protéines désordonnées les CDRs pourraient être moyennés à zéro. Il n'en n'est rien, les CDRs ont été mesurés sur plusieurs systèmes et adoptent une courbe en forme de cloche moyennée à 0 aux extrémités de la protéine et non moyennée à 0 au centre de la protéine (se référer à la figure 2.7 à gauche). Ces valeurs traduisent l'orientation moyenne du vecteur internucléaire D_{NH} relative au champ magnétique au centre de la chaine et à l'extrémité de la chaine.

L'existence de ces couplages a été démontrée théoriquement en considérant une chaine peptidique libre, effectuant une marche aléatoire [85, 86]. Cette approche a permis de reproduire la forme en cloche des CDRs pour des protéines de différentes longueurs mais a ses limites. Le principal problème réside dans la non-distinction des spécificités des acides aminés. Considérés comme tous identique, il est alors impossible d'obtenir une description quantitative de l'ordre local à l'échelle atomique.

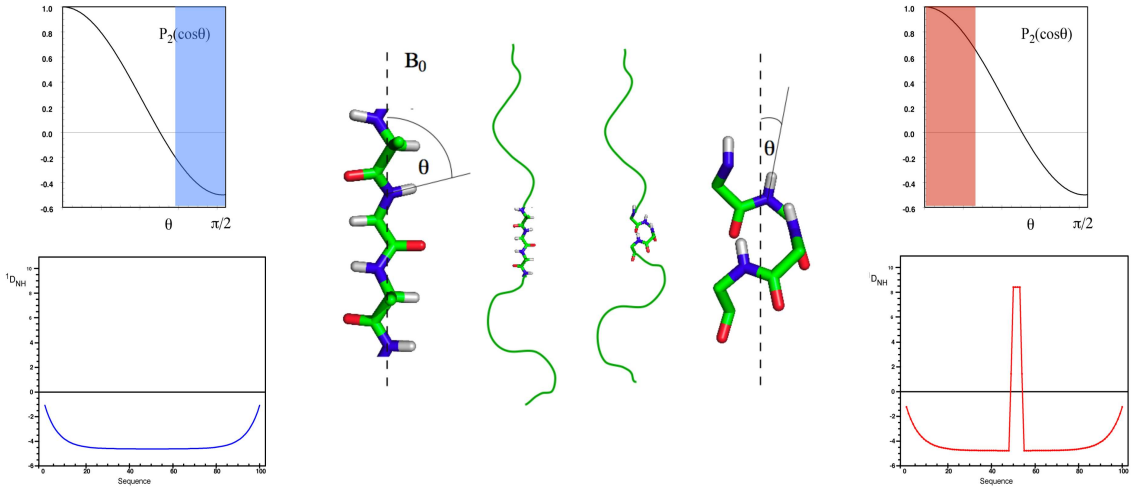


FIGURE 2.7 – ***Schéma du vecteur internucléaire par rapport au champ magnétique pour deux conformations.*** *(a) : Conformation "dépliée" (b) Conformation possédant un repliement un hélice α. L'orientation du vecteur internucléaire relative au champ magnétique est ainsi modifiée. SOnt affichés respectivement les valeurs prise par la fonction* $P_2(cos\theta_{\text{N-H}}$ *et* $D_{\text{N-H}}$

Sur la 2.7 à , les déviations des couplages D_{NH} par rapport à la courbe en forme de cloche nous informe de la présence de structures résiduels au sein de la protéine [78, 87]. Ces variations s'expliquent en considérant la valeur moyenne $\langle d_{\mathrm{NH}} P_2(cos(\theta_{\mathrm{NH}}))\rangle$ où θ_{NH} est l'angle entre le vecteur $N_i - H_i^{\mathrm{N}}$ et le champ magnétique $\boldsymbol{B}^0$ et P_2 le polynôme de

Legendre du second ordre. L'orientation relative du vecteur internucléaire dépend des conformations adoptées par la protéine.

2.7 La relaxation paramagnétique

Les protéines possèdent de l'information à moyenne et longue distance, dans le cas des protéines repliées, il est d'usage de mesurer l'effet NOE pour *Nuclear Overhauser Effect*. Cet effet existant à moins de 6Å n'est cependant pas très adapté à l'étude des protéines désordonnées : d'une part les contacts existants sont très faiblement peuplés et d'autre part le temps de corrélation n'est pas favorable aussi à l'obtention d'un signal élevé. Pour cela une méthode alternative existe : elle consiste en présence d'une Cystéine au sein de la séquence à attacher à un tag ou radical libre, c'est-à-dire à une molécule nommée *thiol-reactive nitroxide label* comprenant un électron non-apparié, l'électron célibataire étant souvent porté par un atome d'azote. Il est alors possible d'utiliser le rapport gyromagnétique de l'électron qui est 650 fois plus important que celui du proton $^1H^N$. Cette méthode nommée la relaxation paramagnétique (PRE) est particulièrement utile pour caractériser le présence d'ordre à longue portée (jusqu'à 25 Å) au sein des protéines désordonnées.

De nombreuses techniques permettent d'ajouter des radicaux paramagnétiques au sein d'une protéine, nous nous focalisons ici sur l'utilisation d'un radical spécifique nommé MTSL ((1-oxyl-2,2,5,5-tetramethyl-Δ3-pyrroline-3-methyl)-methanethiosulfonate spin label). En l'absence de cystéine, l'expérimentateur mute un acide aminé, de préférence une Alanine, en Cystéine pour ajouter la chaine latérale sur la position désirée de la séquence. Il faut évidemment éviter de perturber les structures coopératives existantes au sein de la séquence pour obtenir la description la plus proche des conditions biologiques.

2.7.1 Présentation du formalisme de la relaxation

Les mouvements des protéines peuvent être étudiés en spectroscopie RMN à partir de la relaxation. Pour traiter cette information, il est nécessaire de recourir à la théorie de Redfield-Abragam [88]. Cette dernière montre comment les expressions des constantes de relaxation sont calculées à partir de l'évolution du système de spins exprimé par la matrice densité σ. Nous traiterons dans cette section uniquement le cas de la relaxation paramagnétique de proton $^1H^N$ impliquant un couplage dipolaire entre le spin de l'électron et les spins nucléaires $^1H^N$ de la protéine mais nous ne développerons pas ici le calcul.

Les vitesses de relaxation s'expriment en fonction des valeurs adoptées par la densité spectrale $J(\omega)$ aux fréquences de transitions. La contribution de la relaxation paramagnétique causée par l'interaction dipolaire entre le spin nucléaire et le spin de l'électron s'exprime :

$$\Gamma_2^{para} = K\left[4J(0) + 3J(\omega_H)\right] \tag{2.32}$$

$$\Gamma_1^{para} = 2K\left[3J(\omega_H)\right] \tag{2.33}$$

avec K une constante égale à $1.23 * 10^{-32} cm^6 s^{-2}$ et $J(\omega_H)$ la fonction de densité spectrale échantillonnée à la fréquence de Larmor du proton.

La fonction de densité spectrale traduit la contribution de l'ensemble des mouvements à une fréquence donnée. C'est la transformée de Fourier réelle de la fonction

d'autocorrélation $C(t)$:

$$j(\omega) = 2\int_0^\infty \cos(\omega t) C(t) dt \tag{2.34}$$

où la fonction d'autocorrélation traduisant "la mémoire" des spins ou du vecteur internucléaire $\boldsymbol{e}_{\mathrm{IS}}$ du couplage considéré dans le repère de la molécule.

$$C(t) = \frac{1}{5}\langle P_2(\boldsymbol{e}_{\mathrm{IS}}(0)\boldsymbol{e}_{\mathrm{IS}}(t))\rangle \tag{2.35}$$

où $P_2(x)$ du polynôme de Lagrange d'ordre 2.

Un des problèmes majeurs de cette approche est la modélisation des mouvements à priori inconnus avant l'analyse des données. Une des plus grandes avancés dans ce domaine fut l'introduction du modèle libre : au lieu de modéliser physiquement le mouvement des protéines, Lipari and Szabo [89] suggère l'utilisation de deux paramètres : un paramètre d'ordre généralisé S^2 qui traduit la restriction spatiale du mouvement du vecteur en question et un temps de corrélation ou temps caractéristique du mouvement.

Dans notre cas d'étude, nous considérons deux mouvements s'effectuant sur des échelles de temps différentes, le mouvement de rotation global de la protéine et le mouvement interne de la chaine latérale où est attaché le tag paramagnétique. La fonction de corrélation est le produit de la fonction de corrélation de rotation globale $C_0(t)$ par celles des mouvements internes $C_{\mathrm{I}}(t)$:

$$C(t) = C_0(t)C_{\mathrm{I}}(t) \tag{2.36}$$

où les fonctions de corrélation respective s'écrivent :

$$C_0(t) = \exp(-t/\tau_R) \qquad \text{et} \qquad C_{\mathrm{I}}(t) = S^2 + (1-S^2)\exp(-t/\tau_{\mathrm{I}}) \tag{2.37}$$

avec τ_R le temps de corrélation global ou rotationnel et τ_{I} le temps de corrélation interne. Nous avons alors la fonction de corrélation suivante :

$$C(t) = S^2\exp(-t/\tau_{\mathrm{R}}) + (1-S^2)\exp(-t/\tau_{\mathrm{E}}) \qquad \text{avec} \qquad \frac{1}{\tau_{\mathrm{E}}} = \frac{1}{\tau_{\mathrm{I}}} + \frac{1}{\tau_{\mathrm{R}}} \tag{2.38}$$

La densité spectrale s'écrit alors :

$$j(\omega) = \frac{2}{5}\left(\frac{S^2\tau_{\mathrm{R}}}{1+\omega^2\tau_{\mathrm{R}}^2} + \frac{(1-S^2)\tau_{\mathrm{E}}}{1+\omega^2\tau_{\mathrm{E}}^2}\right) \tag{2.39}$$

Considérant que les fluctuations locales de la chaine principale s'effectuent sur une échelle de temps plus lente que les phénomènes de relaxation, nous pouvons écrire :

$$J(\omega) = \left\langle \frac{1}{r_{\mathrm{IS}}^6} \right\rangle j(\omega) \tag{2.40}$$

2.7.2 Le modèle de Gillespie et Shortle

Pour expliquer les origines de la relaxation paramagnétique dans les protéines désordonnées, il est commode de différencier alors deux mouvements principaux intervenant dans la relaxation. Nous devons considérer plusieurs phénomènes physiques. L'interaction dipolaire entre l'électron non-apparié de la chaine latérale (I) et le spin du proton $^1H^N$ de la chaine peptidique (S) est modulée par :

(i) la réorientation du vecteur $\boldsymbol{e}_{\mathrm{IS}}$ dans le référentiel du laboratoire

(ii) la variation de distance r_{IS}

où $\boldsymbol{e}_{\mathrm{IS}}=\boldsymbol{r}_{\mathrm{IS}}/r_{\mathrm{IS}}$

Le traitement théorique appliqué aux protéines désordonnées a été initialement formulé par Gillespie et Shortle [90, 91]. Le modèle s'appuie sur deux hypothèses :

- D'une part, l'existence d'un ensemble de protéines dépliées en échange conformationnel. Ces molécules sont cependant considérées comme rigide sur une échelle de temps et soumis à la rotation globale avec un temps de corrélation global τ_c. Il n'existe pas dans ce cas de mouvements internes, $S^2 = 1$

- D'autre part, l'existence de fluctuation conformationnelle, c'est-à-dire de variation des distances r_{IS} au cours du temps sur une échelle plus lente que la rotation globale de la molécule.

La densité spectrale utilisée est donc la suivante :

$$J(\omega) = \left\langle \frac{1}{r_{\mathrm{IS}}^6} \right\rangle \frac{\tau_c}{1+\omega^2\tau_c^2} \tag{2.41}$$

Nous obtenons en conséquence la formulation suivante :

$$\Gamma_2^{para} = K \left\langle \frac{1}{r_{\mathrm{IS}}^6} \right\rangle \left[4\tau_c + \frac{\tau_c}{1+\omega^2\tau_c^2} \right] \tag{2.42}$$

Dans ce modèle la relaxation paramagnétique est dépendante d'une part du temps de corrélation global de la protéine qui est uniforme pour toute la protéine et d'autre part de la distribution de distance $\langle r_{IS}^{-6} \rangle$. La plupart des travaux se sont concentrés sur la détermination ensemble de conformations reproduisant les données expérimentales. De nombreux systèmes ont ainsi été étudiés en traduisant en distance effective ou distribution de distance l'interaction électron-spin. Répétant cette expérience en utilisant une série de radicaux ajoutés singulièrement le long de la chaine, il est alors possible d'obtenir une distribution de distances traduisant la présence d'ordre résiduel à longue portée.

Le traitement quantitatif des données PRE reste délicat, ce modèle a été validé sur de nombreux systèmes et semble bien adapté aux protéines dépliées possédant une mobilité interne proche d'une protéine repliée. Pour autant considérant une protéine fortement désordonnée, les hypothèses précédentes suggérant une protéine rigide sur un intervalle de temps proche de la nanoseconde peuvent poser problème. Une alternative consiste alors à privilégier les phénomènes de diffusion translationnelle pour représenter le mouvement des protéines désordonnées [92, 93].

2.7.3 Le modèle prenant en compte la dynamique de la chaine latérale MTSL

L'ajout d'une chaine latérale MTSL nécessite une réflexion sur la dynamique du système, cette dernière est probablement hautement flexible, son orientation est seulement contrainte par la chaine principale et les chaines latérales environnantes. Démontré par Iwahara et al. [94], l'incorporation d'un modèle dynamique améliore nettement la reproduction de la distance dite effective entre l'électron et le spin. Notre description, prenant en compte la flexibilité de la chaine latérale MTSL, s'appuiera sur deux modèles dynamiques différents. Le premier modèle considère un ensemble de structures en échange

représentant l'état d'équilibre d'un système hautement flexible. Le second traduit la mobilité de la chaine latérale MTSL de chaque structure. Nous ferons l'hypothèse suivante :

Hypothèse du modèle de chaine MTSL

La dynamique du squelette de la chaine principale de la protéine s'effectue sur une échelle de temps différente (plus lente) que celle de la chaine latérale MTSL. Elles peuvent donc être traitées statistiquement de manière indépendante.

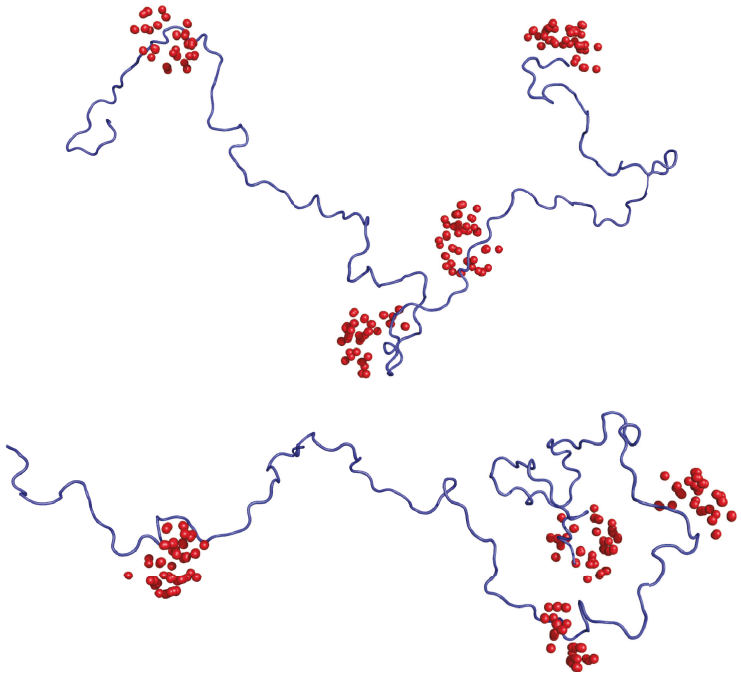


FIGURE 2.8 – ***Positions échantillonnées par l'électron non apparié de chaines latérales MTSL.*** *Deux structures arbitraires de la protéine α-Synucléine sont générées avec* FLEXIBLE-MECCANO. *Le squelette en bleu et le radical libre en rouge. Les positions des tags paramagnétiques sont celles utilisées expérimentalement, c'est à dire A18C, A76C, A90C et A140C.*

La réorientation de la chaine latérale MTSL s'effectuerait sur une échelle de temps de l'ordre de la picoseconde (d'une dizaine à quelques centaines de picosecondes) tandis que l'échelle de temps définissant l'échange ou l'interconversion des conformations serait proche de la microseconde [95]. Soulignons que les échelles de temps définissant la réorganisation du squelette des protéines désordonnées, le repliement de la chaine en structures secondaires transitoires sont des problématiques primordiales du domaine. La mise en place d'expériences et de modèles permettant de déterminer les temps caractéristiques des PIDs permettrait d'affiner les modèles existants [96, 93]. Nous décrivons le mouvement de la chaine latérale MTSL à partir d'une base de données de rotamères déterminé par Sezer et al. [97]. Nous échantillonnons l'ensemble des positions et retenons les conformations n'ayant pas d'encombrement stérique avec les atomes de la chaine principale. La flexibilité de la chaine latérale est décrite figure 2.8.

Considérant un ensemble de conformation en échange rapide représentant l'état déplié, pour chaque conformation k, la contribution à la relaxation transversale en raison de la présence d'un électron non apparié à proximité du résidu m, notée $\Gamma_{2,k,i,m}$, du résidu i s'exprime comme [98, 66] :

$$\Gamma_{2,k,i,m} = \frac{2}{5}\left(\frac{\mu_0 \gamma_H g_e \mu_B}{4\pi}\right)^2 S_e(S_e+1)\left[4J(0)+3J(\omega_H)\right] \tag{2.43}$$

où g_e est la facteur de Landé de l'électron, μ_B le magnéton de Bohr et S_e le spin de l'élection.

Utilisant une expression initialement développée pour la relaxation croisée homonucléaire, la fonction de densité spectrale J traduisant l'interaction électron spin nucléaire peut être exprimée selon le modèle libre, un paramètre d'ordre dépendant à la fois de la longueur et de l'orientation du vecteur électron spin est invoqué.

$$J(\omega) = \left\langle r_{H-e}^{-6} \right\rangle \left[\frac{S_{H-e}^2 \tau_R}{1 + \omega^2 {\tau_R}^2} + \frac{S_{H-e}^2 \tau_E}{1 + \omega^2 {\tau_E}^2} \right] \tag{2.44}$$

où S_{H-e}^2 est le paramètre d'ordre du vecteur électron spin, r_{H-e} est la distance instantanée électron-spin et les temps de corrélation sont données par les formules suivantes :

$$\frac{1}{\tau_R} = \frac{1}{\tau_C} + \frac{1}{\tau_S} \qquad \text{et} \qquad \frac{1}{\tau_E} = \frac{1}{\tau_C} + \frac{1}{\tau_S} + \frac{1}{\tau_I} \tag{2.45}$$

où τ_C est le temps de corrélation global de la molécule, τ_S est le temps de relaxation électron spin et τ_I est le temps de corrélation du mouvement de l'électron par rapport au spin nucléaire.

Le paramètre d'ordre S_{H-e}^2 peut être décomposé suivant sa composante radiale et sa composante angulaire.

$$S_{H-e}^2 = \frac{4\pi}{5} \left\langle r_{H-e}^{-6} \right\rangle^{-1} \sum_{m=-2}^{2} \left| \frac{Y_{2,m}(\theta, \phi)}{r_{H-e}} \right|^2 \simeq S_{ang}^2 S_{rad}^2 \tag{2.46}$$

où

$$S_{rad}^2 = \left\langle r_{H-e}^{-3} \right\rangle^2 \left\langle r_{H-e}^{-6} \right\rangle^{-1} \qquad \text{et} \qquad S_{ang}^2 = \frac{4\pi}{5} \sum_{m=-2}^{2} \left| \left\langle Y_{2,m}(\theta, \phi) \right\rangle \right|^2 \tag{2.47}$$

où le couple (θ, ϕ) représente l'orientation du vecteur électron spin dans le repère moléculaire. Ces équations permettent d'estimer la contribution de l'électron non apparié à la relaxation transverse pour une structure donnée. La contribution effective d'un ensemble de N structures est donnée par la formule :

$$\Gamma_{2,i,m} = \frac{1}{N} \sum_{k=1}^{N} \Gamma_{2,k,i,m} \tag{2.48}$$

Le ratio d'intensité du résidu i où la chaine latérale MTSL est attachée au résidu m est estimé par la formule :

$$\left[\frac{I_{para}}{I_{dia}} \right]_{i,m}^{calc} = \frac{R_{2,i} e^{-\Gamma_{2,i,m} \tau_{mix}}}{R_{2,i} + \Gamma_{2,i,m}} \tag{2.49}$$

où $R_{2,i}$ est la relaxation intrinsèque du proton amide de l'acide aminé i, τ_{mix} est le temps de mélange de 10 ms pendant lequel la relaxation a lieu et $\Gamma_{2,i,m}$ est la contribution du centre paramagnétique à la relaxation.

Conclusion du chapitre

Les observables RMN se révèlent adaptées à l'étude des protéines désordonnées, un des points essentiels à retenir concerne l'échelle de temps définissant l'interconversion des structures[2]. Ce phénomène est plus rapide que la microseconde ou plus lent

2. Comme en mentionnée en section 1.8, les temps caractéristiques restent relatifs aux transitions structurales, il est cependant possible d'émettre une hypothèse sur l'ordre de grandeur de ces phénomènes.

que la nanoseconde, par conséquent les protéines intrinsèquement désordonnées sont en échange rapide et nous observons un unique pic par résidu contenant l'information moyennée sur l'ensemble des conformations. Cette supposition est valide pour les déplacements chimiques, les couplages dipolaires résiduels et scalaires. Dès lors, l'étape suivante consiste à interpréter cette information, la solution retenue est un modèle statistique de conformations explicite en échange traduisant le dynamique du système. Nous allons donc au chapitre suivant présenter les prémices de l'approche.

La description par ensemble appliquée aux protéines désordonnées

3

L'interprétation des données expérimentales des protéines intrinsèquement désordonnées est un des problèmes clés associés à ce champ de recherche. Ne pouvant pas associer la fonction d'une protéine désordonnée à sa structure tridimensionnelle, le paradigme de la biologie structurale à défaut d'être obsolète a dû être révisé. C'est sur cet aspect que repose la majeure partie de mon travail de thèse, c'est à dire développer une description moléculaire de l'état désordonné compatible avec l'ensemble des mesures expérimentales disponibles. Nous présenterons dans un premier temps le modèle utilisé pour le traitement de nos données, ce modèle statistique couramment appelé description par ensemble. Nous verrons quels ont été les premiers travaux effectués sur les protéines désordonnées et en particulier le potentiel des générateurs statistique de structures à l'état *random-coil*, nous présenterons ensuite d'autres modèles alternatifs. Dans un dernier temps, nous insisterons sur les points essentiels abordés durant la thèse.

3.1 L'état déplié défini par un ensemble de structures en échange rapide

Une description dite structurale des protéines désordonnées consiste à déterminer les règles qui définissent le comportement de ces protéines en termes de probabilité d'échantillonner telle ou telle conformation puis de corréler ces probabilités avec la fonction ou les fonctions de ces protéines. La description peut être communément réalisée en utilisant une description par ensemble explicite de structures en échange et à l'équilibre.

La description par ensemble

La description par ensemble est une formulation d'un système physique où l'ensemble représente toutes les différentes configurations (ou états) accessibles et les probabilités correspondantes d'échantillonner ces configurations. Le principal avantage repose sur la possibilité d'associer un grand nombre d'états microscopiques avec une grandeur macroscopique mesurable.

Dans notre cas, il s'agit de construire un nombre fini mais grand de structures qui caractérisent la plage d'états accessibles de ces protéines. Il est alors possible de comparer la valeur moyenne des propriétés de chaque conformation à des mesures expérimentales et d'en déduire des caractéristiques biophysiques. La formulation mathématique d'un tel modèle s'appuie sur une hypothèse essentielle, l'hypothèse d'ergodicité. Sur un temps suffisamment long, chaque structure échantillonne l'ensemble des états accessibles *i.e.* des conformations possibles. Ainsi à l'équilibre, la mesure de l'observable moyenné sur chacune des conformations à un instant donné est égale à la mesure de l'observable d'une unique conformation moyennée pendant un temps important.

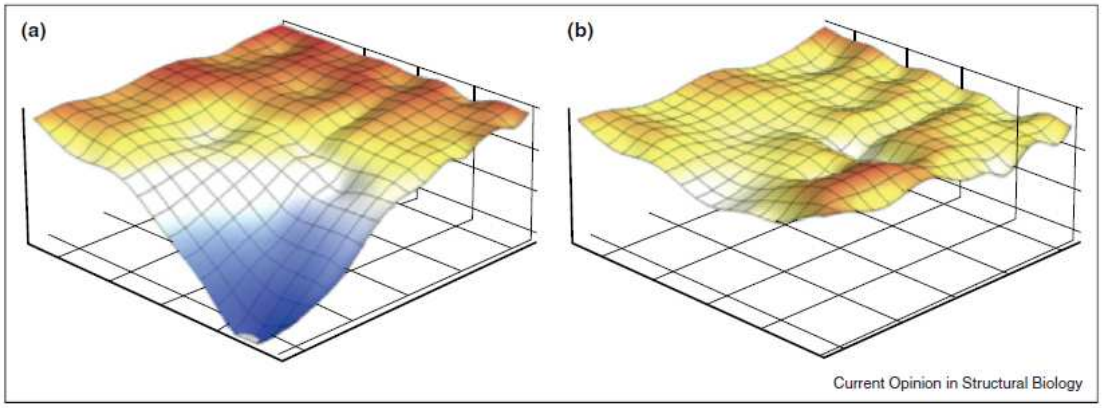


FIGURE 3.1 – ***Paysage énergétique d'une protéine repliée et dépliée.*** *Le paysage énergétique d'une protéine repliée (a) est clairement défini par un minimum comparé à celui d'une protéine désordonnée (b). Figure extraite de [99].*

Une façon similaire de représenter ces états consiste à considérer le diagramme énergétique des protéines. Les protéines échantillonnent un nombre varié de conformations durant leur temps de vie biologique à cause des fluctuations thermales. La probabilité d'échantillonner telle ou telle conformation est déterminée par la topographie du paysage énergétique associé. Les protéines repliées possèdent un paysage énergétique où il existe un minimum bien défini (figure 3.1 (a)) tandis que les protéines désordonnées possèdent un paysage énergétique plat et par conséquent échantillonnent un grand nombre d'états à température ambiante (figure 3.1 (b)).

3.2 La description par ensemble de l'état *random-coil*

3.2.1 La définition historique

Le terme *random-coil* fait directement référence à la physique des polymères : c'est une description de l'état d'un polymère dans laquelle, en l'absence d'interaction spécifique, chaque sous unité du polymère, reliée entre elles, échantillonne aléatoirement l'ensemble des directions de l'espace sans tenir compte des interactions stériques. Une des premières descriptions statistiques basées sur les propriétés physiques des polymères est le modèle *random-coil* de Flory, calculant une fonction de distribution des conformations. En dépit de son apparente simplicité, ce modèle permet d'estimer la distance moyenne entre les extrémités[1] d'un polymère de longueur N , le rayon de giration moyen ou la distribution des distances moyennes entre résidus. A titre d'exemple, le rayon de giration d'un polymère décrit par la formule Flory s'écrit :

$$R_g = R_0 N^{\nu} \tag{3.1}$$

où N est le nombre la longueur de la chaine peptidique, R_0 est une constante et ν dépend du comportement structurel du polymère.

Considérant un polymère *random-coil*, ν vaut théoriquement 0.6. La mesure d'un ensemble de protéines dénaturées par diffraction aux petits angles par rayons X (SAXS) a aboutit une valeur de ν de 0.598 [100]. Les protéines désordonnées n'obéissent pas forcément à cette loi, la dénaturation des protéines perturbe le comportement conformationnel de nature à les rendre plus étendues. L'étude *in-silico* de 23 protéines désordonnée combinée à des données SAXS prédit une valeur de ν de 0.522 [101]. Il est important de prendre en compte les spécificités structurales de la protéine pour ce type d'analyse. En effet, la présence de contact transitoire entre deux domaines de la protéine réduira fortement la rayon de giration, à l'opposé la présence de structures étendues de type feuillets β ou motifs Polyproline contribuera à augmenter le rayon de giration de l'ensemble. Les protéines, bien que régies par des propriétés hautement plus complexes qu'une chaine polymérique, peuvent être étudiées en incluant des propriétés biophysiques au sein du modèle. Dans ce cas, les sous-unités du polymère sont alors des acides aminés ayant des caractéristiques spécifiques : charge, polarité, densité électronique. Nous présenterons dans la section suivante d'autres modèles plus élaborés décrivant l'état *random-coil*.

En biologie, le terme *random-coil* est historiquement utilisé pour désigner les protéines n'ayant pas de structuration précise, c'est un terme antérieur à la formulation *protéines intrinsèquement désordonnées*. La structure d'une protéine repliée est un juste équilibre entre les interactions locales ou à courte portée qui déterminent l'échantillonnage conformationnel des acides aminés, les interactions à moyenne portée traduisant le repliement de certaines régions en structures secondaires et les interactions à longue portée définissant la structure tertiaire *i.e.* la forme globale de la protéine. Pour comprendre le degré de persistance de chacune de ces interactions, il est crucial d'étudier les interactions à courte portée majoritairement présentes dans les protéines dites *random-coil*. À l'issu de ce chapitre introductif, nous l'utiliserons alors davantage pour désigner la description statistique des protéines désordonnées développée au sein du groupe et la base de données d'angles dièdres associées.

1. end-to end-distance

3.2.2 Le modèle statistique *random-coil*

Étant donné le très grand nombre de degrés de liberté existant, définir l'espace conformationnel de ces protéines est particulièrement ardu et nécessite le recours à des jeux de données expérimentaux afin de délimiter les propriétés structurales d'une protéine donnée. La RMN est particulièrement adaptée à la mesure des protéines *random-coil*. Comme exposée précédemment, les paramètres RMN sont moyennés dans le temps et peuvent transcrire l'environnement chimique moyen ou la distance effective entre deux atomes. L'interprétation des résultats, plus délicate, nécessite l'introduction d'une description par ensemble de structures. Smith et al. [73] présente les premiers travaux permettant de reproduire des données expérimentales de RMN de peptides ou protéines *random-coil*.

La description par ensemble utilise une distribution d'angles dièdres (ϕ,ψ) réalisée à partir de structures cristallographiques en extrayant les angles spécifiquement à chaque acide aminé. Les premières bases de données considèrent l'ensemble des angles dièdres présents dans les structures cristallographiques. Ils permettent d'identifier trois régions de l'espace Ramachandran : la région correspond au hélice α nommée αR, la région correspond au feuillet β nommée β la région correspondant au hélice 3_{10} nommée αL, la région αR étant largement majoritaire en de la présence des nombreuses hélices présentes dans les protéines repliées [73]. À partir d'une base de donnée de 2020 structures PDB présentée en figure 3.2, nous pouvons évaluer l'évolution de la distribution des angles (ϕ,ψ) en fonction de la présence ou non des structures secondaires. Les bases de données les plus en accord avec les propriétés de l'état déplié ne prennent pas en compte des structures secondaires hélices α et feuillets β, les boucles et les résidus adjacents de ces structures [103, 102]. La distribution des angles dièdres est ensuite analysée par acide aminé, ces derniers échantillonnent telle ou telle région de l'espace Ramachandran avec une probabilité spécifique. Pour la plupart des acides aminés, les régions de l'espace occupées sont relativement les mêmes exception faite des Prolines et des Glycines.

Afin de vérifier la validité des modèles, il est possible de comparer les paramètres RMN prédits et expérimentaux. S'appuyant sur l'équation traduisant la dépendance du couplage ${}^3J$ en fonction de la distribution de l'angle dièdre ϕ, Penkett et al. [106] calcul la valeur du couplage ${}^3J$ issu de leur modèle statistique. Les auteurs mettent non seulement en évidence la dépendance de la valeur des couplages ${}^3J$ en fonction des acides aminés (figure 3.3) mais aussi celui des voisins : la distribution d'angles (ϕ,ψ) est influencée par les résidus précédents ou suivant ce dernier.

Ces études pionnières du domaine mettent en perspective l'importance de ce champ d'investigation. Ces protéines dites *random-coil* sont non seulement importantes pour la compréhension de la physique des protéines mais se révèlent fonctionnelles et associées à de nombreux mécanismes biologiques importants : nous citerons la protéine Tau, la kinase p21, la protéine Aβ précurseur des plaques amyloïdes de la maladie d'Alzheimer.

3.3 Inventaire des méthodes existantes pour décrire l'état déplié.

La description par ensemble est maintenant communément admise comme un méthode efficace pour décrire les protéines intrinsèquement désordonnées. Un large panel d'algorithmes et de logiciels sont apparus mais diffèrent suivant le modèle physique envisagé. Nous distinguons majoritairement deux approches : les descriptions par

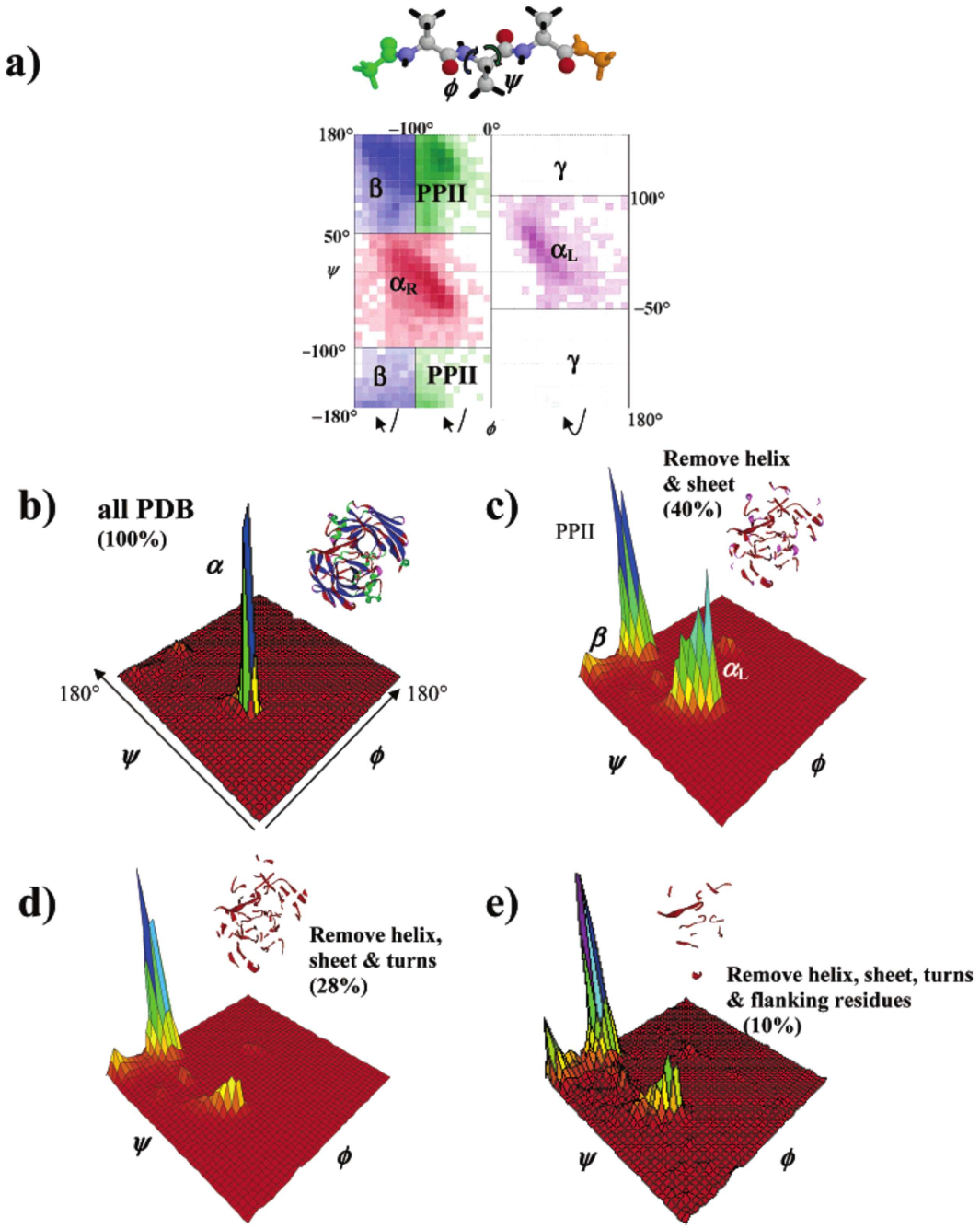


FIGURE 3.2 – ***Influence des structures secondaires sur les bases de données random-coil.*** *(a) :Schéma d'un tripeptide AAA (Ace-$(Ala)_3$-Nme), la position des angles (ϕ,ψ) est indiquée, ci-dessous la carte de Ramachandran comprenant la distribution d'angles dièdres de 2020 structures. Les couleurs et rectangles délimitent quatre régions : la région αR, la région αL, la région β, la région PPII. Nous affichons ensuite la distribution en terme de probabilité des angles dièdres dans l'espace de Ramachandran (b) : Les structures entières sont incluses. (c) : Les structures secondaires hélices α et feuillets β sont retirés. (d) : Les structures secondaires hélices α et feuillets β et les boucles sont retirés. (e) : Les structures secondaires hélices α et feuillets β, les boucles et les résidus adjacents sont retirés. La figure est extraite de Jha et al. [102].*

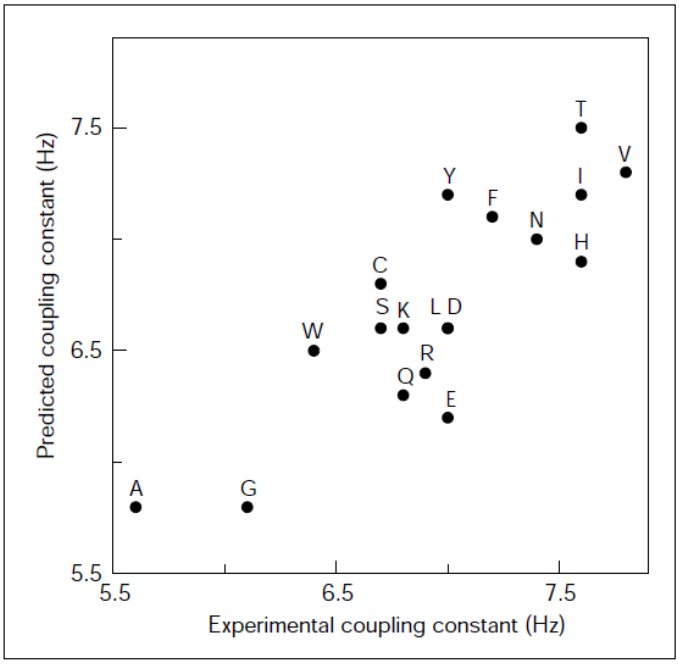


FIGURE 3.3 – ***Comparaison des couplages*** ^3J ***expérimentaux avec ceux calculés pour une distribution random-coil.*** *Les données présentées sont une compilation des couplages extraits du lyzozyme dénaturé et de la protéine Barnase [104]. Pour calculer théoriquement les valeurs des couplages* ^3J*, l'équation de Karplus (équation 3.2.2) est appliquée à la base de données d'angles* (ϕ,ψ) *issue de Smith et al. [105].*

ensemble sans contraintes et les descriptions par ensemble avec contraintes. Nous présentons maintenant une liste non exhaustive des méthodes développées pour chacune de ces descriptions.

3.3.1 La description statistique *random-coil*

La description statistique *random-coil* présentée précédemment peut être implémentée avec logiciels exploitant une base de données d'angles dièdres (ϕ,ψ) pour générer des ensembles de structures représentant l'état désordonnée. La difficulté essentielle de cette étape consiste à obtenir le paysage énergétique le plus étendu possible, les conformations générées doivent échantillonner le maximum d'états possibles. Ces générateurs de structures *random-coil* se sont révélés bien adaptés à cette problématique et peuvent créer très rapidement des ensembles de structures sur lesquels nous calculons les paramètres RMN afin de les comparer aux données expérimentales. Nous citons à titre d'exemple les logiciels TraDES [107] ou FLEXIBLE-MECCANO [108, 109] qui génèrent des ensembles de structures à partir d'une base de donnée d'angles dièdres en tenant compte des contraintes stériques entre atomes. Aucune autre hypothèse ou biais n'est inclus dans le calcul de structures.

3.3.2 La dynamique moléculaire sans contrainte.

Une approche standard impliquant une description explicite de la structure des protéines est bien évidemment la dynamique moléculaire. Cette méthode utilisée pour étudier les protéines repliées peut s'appliquer sous certaines conditions aux protéines désordonnées. Les protéines désordonnées étant particulièrement en contact avec le solvant, ainsi il est crucial de simuler correctement l'interactions avec les molécules d'eau. Le second point essentiel concerne le choix du champ de force, ce dernier doit

être en adéquation avec les caractéristiques biophysiques des protéines désordonnées et permettre des repliements transitoires et des contacts à longue-portée tout en incluant un régime *random-coil*. Lindorff-Larsen et al. [110] ont dernièrement appliqué un nouveau protocole à la protéine *Acyl-CoA-binding protein* (ACBP) dénaturée en considérant la protéine dans une boite de 23569 molécules d'eau en équilibre avec 19 ions chlorure. Le champ de force CHARMM22 et le modèle TIP3P ont été respectivement appliqués à la protéine ACBP dénaturée et au solvant.

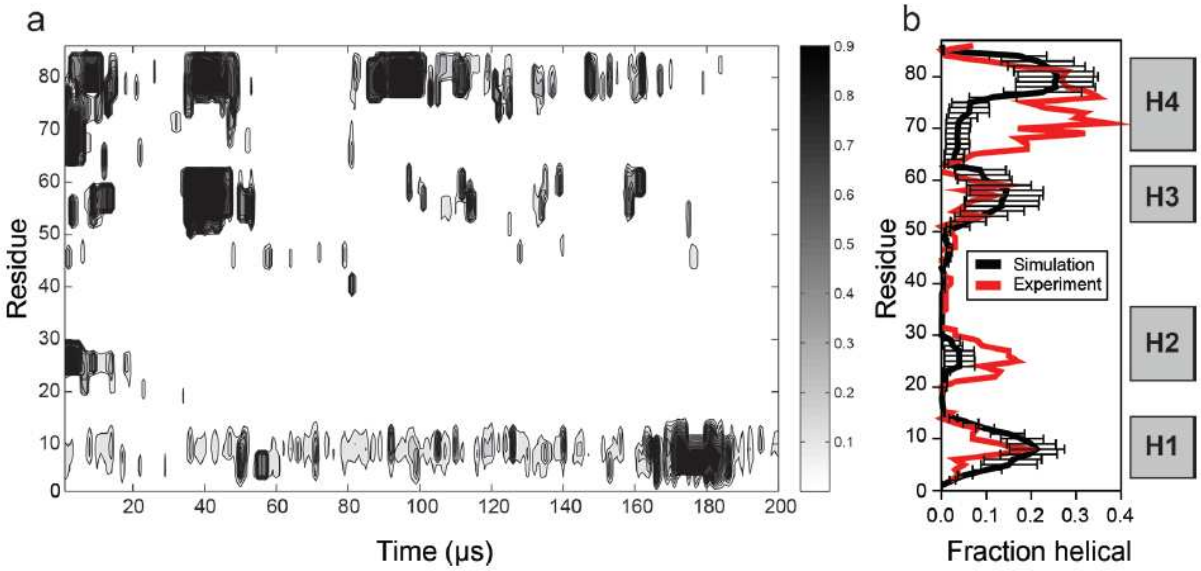


FIGURE 3.4 – ***Repliements transitoires des hélices de la protéine ACBP dénaturée.*** *(a) Evolution au cours de la simulation de la propension au repliement en hélice en fonction de la position dans le séquence. Un gradient de couleur (niveau de gris) permet d'estimer la fraction de structuration. (b) Moyenne sur la durée de la simulation de la structuration en hélice (en noir), estimation, issue de données expérimentales, de la propension en structures hélicoïdales de la protéine (en rouge). Les cadres en gris représentent la position des hélices lorsque la protéine est repliée. Figure extraite de Lindorff-Larsen et al. [110].*

Les auteurs ont choisi délibérément de ne pas inclure de contraintes, l'objectif étant de valider le champ de force utilisé. Ils ont pu observer le repliement et dépliement des hélices présentes dans la forme native (figure 3.4) et calculer des valeurs de relaxation ^{15}N en accord avec des données expérimentales.

Deux mécanismes associés au champ de force sont difficiles à maitriser en dynamique moléculaire et sont souvent formulés en sa défaveur. D'une part, l'utilisation d'un champ de force restreint considérablement le paysage énergétique accessible par la protéine à l'opposé d'une description par ensemble classique où les conformations accèdent à l'ensemble des états possibles, il est ainsi difficile d'obtenir des transitions structurales importantes entre conformations. De l'autre, la comparaison des structures simulées avec celles issues d'un modèle *random-coil* aboutissent généralement à une surestimation de l'échantillonnage de la région hélicoïdale.

3.3.3 Le modèle Meta-Structure

L'approche Meta-Structure est un modèle topologique considérant la protéine comme un réseau artificiel décrit par des noeuds et de chemins, les noeuds étant les résidus, les chemins traduisant les interactions entre noeuds. Ce modèle présume l'existence de propriétés définissant le voisinage de chaque noeud [111]. L'approche Meta-Structure prend le contre-pied des descriptions par ensembles statistiques : ces

dernières supposent l'existence d'ordre local au niveau du résidu et utilisent donc la distribution d'angles dièdres de protéines repliées pour prédire le comportement des protéines dépliées tandis que l'approche Méta-Structure suppose l'existence d'ordre à moyenne ou longue portée entre les résidus et traduit plutôt l'hydrophobicité ou la charge des résidus. Le modèle est alors défini en cartographiant les interactions présentes dans un jeu de protéines repliées. Ainsi deux résidus sont voisins si la distance les séparant est inférieure à 8Å, la distance étant arbitraire. La topologie θ entre deux noyaux A et B est défini par deux points : le plus court chemin entre ces deux noyaux et l'espacement l_{AB} en résidu dans la séquence. La topologique complète est alors statistiquement évaluée en appliquant ce modèle à un ensemble de structures cristallographiques repliées. En résulte des fonctions de distribution définit par les paramètres suivant : $\rho(\theta, A, B, l_{AB})$.

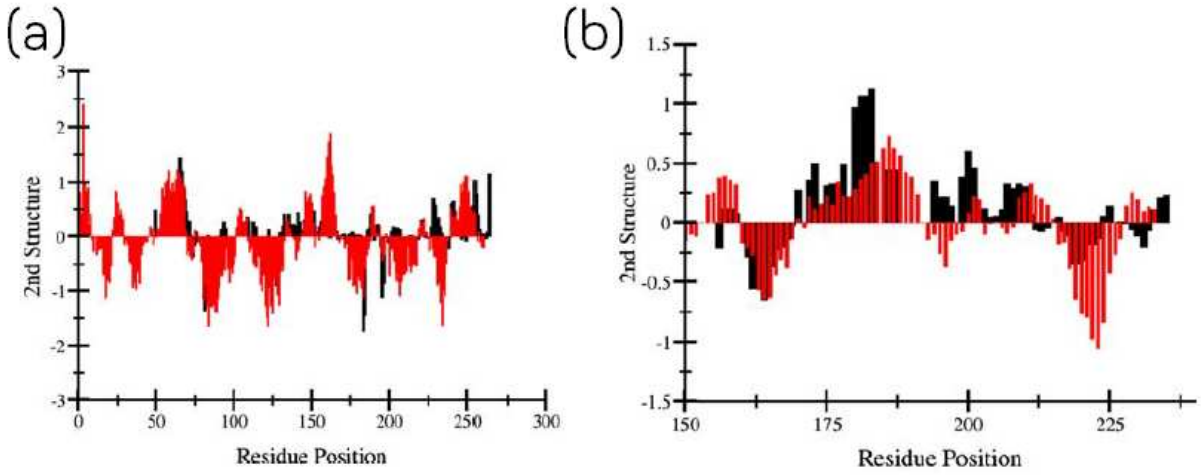


FIGURE 3.5 – ***Ordre résiduel prédit par le modèle Meta-Structure.*** *Comparaison de la prédiction de structures secondaires issue du modèle Meta-Structure avec les valeurs expérimentales correspondent pour deux protéines intrinsèquement désordonnées (a) : CTD-ICln et (b) : Osteopontin . Les valeurs positives et négatives indiquent respectivement de la structuration en hélice α et en feuillet β (figure extraite de Konrat [111]).*

Il est alors possible en utilisant uniquement la séquence primaire d'une protéine et ces fonctions de distribution de prédire les paramètres suivants : la distance effective entre deux atomes, la compacité , la présence de structures secondaires (figure 3.5) [111].

3.3.4 Les descriptions par ensembles sous contraintes.

Il existe différentes méthodes de description par ensemble sous contraintes, elles se situent dans la continuité des approches présentées précédemment. L'objet est de caractériser au mieux l'état déplié des protéines désordonnées : après avoir généré un vaste ensemble de structures par dynamique moléculaire ou avec un générateur statistique *random-coil*, il s'agit ensuite de sélectionner un sous-ensemble de structures minimisant la différence entre les données simulées et les données expérimentales permettant ainsi de capturer les caractéristiques biophysiques de l'ensemble telles que les dimensions globales de la protéines [112], la présence de contact à longue portée [113], [114], la présence de structures secondaires transitoires [115] ou même encore l'échantillonnage local de chaque résidu ([109]). Ce problème d'analyse multivariée est délicat, nous sommes en présence d'un système possédant un nombre de degrés de liberté très élevés qui est contraint par des données RMN ou de diffraction aux petits angles par rayons X (SAXS) peu nombreuses. Ces contraintes sont appliquées soit par l'intermédiaire

d'un champ de force [116], soit en utilisant des algorithmes génétiques sélectionnant les meilleures structures en accord ou soit encore en pesant le poids des structures utilisées dans une approche bayésienne [117]. Ces méthodes ont été appliquées avec succès avec d'autres algorithmes combinant données RMN et SAXS [118, 119]. Des consensus semblent maintenant se dessiner concernant le traitement et la validation des approches comme en témoignent ces revues [99, 120, 121].

L'interprétation des données RMN ou SAXS des protéines désordonnées est un défi scientifique nécessitant des connaissances en mathématiques, physiques. Cette rapide présentation des méthodes n'est évidemment pas exhaustive et ne permet pas d'assimiler la complexité des problèmes rencontrés lors du traitement des données. C'est pourquoi nous présenterons au paragraphe suivant les grandes lignes à approfondir concernant le fonctionnement de la description par ensemble sous contraintes.

3.4 Avant-propos, réflexion

La construction d'ensemble de structures représentant l'état désordonné est sujette à de nombreuses incompréhensions. Le premier point concerne le modèle, nous devons abandonner la vision classique de la biologie structurale consistant à associer l'état d'une protéine à une structure unique. Nous passons ainsi d'un modèle comprenant une représentation visuellement accessible à un modèle statistique.

La plupart des logiciels créant des ensembles s'appuient sur la génération de structures en format PDB, l'ensemble de ces conformations représentant l'ensemble des états accessibles de la protéine désordonnée. Pour autant l'utilisation de structures PDB n'est pas une condition nécessaire à la description par ensemble. L'utilisation de format PDB facilite tout de même le travail computationnel, notamment pour calculer les paramètres RMN les plus courants comme les CDRs, les déplacements chimiques, les PREs et les couplages ^3J. Ils servent aussi à vérifier visuellement les configurations des conformations échantillonnées.

Nous sommes confrontés au paradoxe suivant : d'un côté, la détermination d'un ensemble de conformations en accord avec les valeurs expérimentales ne garantit pas l'existence physique des conformations, la vision considérant ces structures en échange reste réductrice de la problématique. Les échelles de temps régissant l'interconversion des structures étant probablement plus rapide que la mesure des paramètres RMN, nous ne pouvons trancher sur l'existe réelle de conformation fixe représentant l'état désordonné. Il est relativement désuet de représenter l'état déplié en traçant un ensemble des structures dans l'espace réel (x,y,z), une représentation montrant la distribution des angles dièdres dans l'espace Ramachandran est plus appropriée. En d'autres termes, un ensemble de structures PDB est juste une représentation discrète de l'état déplié, les probabilités de distributions des paramètres physiques sont le meilleur moyen de présenter leurs caractéristiques.

La description par ensemble explicite est une représentation statistique de l'état déplié, il est conçu d'un grand nombre de conformations supposées en échange à l'équilibre. La sélection de sous-ensembles requiert la définition ou du moins une réflexion sur le nombre de structures nécessaires pour décrire l'état déplié. A priori, le degré de désordre est corrélé avec le nombre de conformations nécessaires pour atteindre l'équilibre statistique et représenter correctement l'état déplié. Une protéine complètement dépliée nécessite plus de conformations qu'une protéine transitoirement repliée, de même une protéine de 400 résidus nécessite plus de conformations qu'un

peptide de 10 résidus. Il est aussi possible de peser le poids des conformations présentes pendant la sélection. Ce choix est principalement méthodologique et n'engage pas l'approche considérée, qui nécessite juste alors N paramètres de plus.

La détermination d'un sous-ensemble de structures est un problème d'analyse quantitative multivariée. Comme précédemment exposé nous sommes en présence d'un système présentant un nombre de degrés de liberté impressionnant et nous ne disposons que de mesures expérimentales restreintes. Il est ainsi particulièrement difficile de déterminer une seule solution à ce problème. L'utilisation même de données RMN ou SAXS ne permet pas de déterminer sans ambiguïté des ensembles en accord avec les données expérimentales. Il faut donc évoluer prudemment lors des protocoles de sélections, la dégénérescence des solutions est à priori inévitable sans la combinaison des différents paramètres.

Les questions fondamentales

Nous mettons maintenant en avant différentes questions ou problèmes que nous nous efforcerons de mentionner ou résoudre dans les chapitres suivants :

- Point 1 : l'accord avec les données expérimentales ne garantit pas le sens physique de l'ensemble en sens ensemble respectant une distribution de Boltzmann
- Point 2 : le nombre de degrés de liberté d'un tel système est très important comparativement aux nombres de données disponibles pour caractériser de tel système
- Point 3 : il n'existe pas de solution unique en accord avec les données expérimentales
- Point 4 : quel nombre de structures devons-nous utiliser dans une description par ensemble ?

Conclusion du chapitre

Durant ces trois chapitres, nous avons introduit les concepts et outils fondamentaux pour comprendre et étudier les protéines désordonnées. Ce manuscrit propose ainsi d'expliquer comment en combinant la résonance magnétique nucléaire avec des outils numériques il est possible d'obtenir une description moléculaire de l'état déplié. La solution retenue pour décrire la dynamique des protéines désordonnées est de recourir à un modèle statistique appelé la description par ensemble explicite de structures, décrivant l'ensemble des états accessibles. En continuité, le chapitre suivant présentera les réalisations du groupe Flexibilité et Dynamique des Protéines et notamment le modèle choisi pour les travaux de thèse.

La description par ensemble : présentation et application à la protéine Ubiquitine dénaturée dans l'urée

4

Ce chapitre englobe les travaux réalisés au sein de l'équipe en amont de la thèse. Il se divise en deux parties, l'une concerne la présentation de la description par ensemble, la seconde concerne l'application de cette méthode pour caractériser la protéine Ubiquitine dénaturée dans l'urée à 8M en combinant la description par ensemble sous contraintes avec les couplages dipolaires résiduels.

Courant des années 2000, de nombreuses questions sont adressées à la communauté scientifique pointant du doigt la nécessité de développer des méthodes et outils physiques pour caractériser les protéines intrinsèquement désordonnées [122]. Deux aspects doivent être mis en place : de nouvelles méthodes expérimentales adaptées à la dynamique des protéines intrinsèquement désordonnées sont à développer mais aussi des outils d'analyse et de traitement des données. En effet, l'abandon de la représentation des protéines par une unique structure a nettement complexifié la compréhension des mécanismes biologiques associées aux protéines intrinsèquement désordonnées et constitue un défi à relever. La description par ensemble explicite de structures présentées au chapitre précédent (en section 3) se trouva un modèle viable pour analyser le comportement des protéines désordonnées. C'est dans cette optique que le logiciel Flexible-Meccano a été développé par le groupe Flexibilité et Dynamique des protéines. Nous aborderons donc le fonctionnement de cet algorithme et les applications et perspectives qui en ont découlé. Ces travaux se sont concrétisés par plusieurs publications combinant le modèle développé et les paramètres RMN et SAXS [123, 108, 124, 112].

Une deuxième étape fut la mise en place d'une approche plus quantitative, il s'agissait d'exploiter la diversité des structures composant les ensembles générés par le logiciel Flexible-Meccano pour sélectionner des sous-ensembles de structures les plus en adéquation possible avec les données expérimentales et ainsi caractériser au mieux les règles et spécificités régissant le comportement des protéines désordonnées. La fonction guidant la sélection cherche alors uniquement à minimiser la différence entre les données expérimentales et les données simulées en choisissant les structures les plus appropriées. Ce problème d'optimisation, à priori trivial, nécessita l'implémentation d'un algorithme génétique : Asteroids pour *A Selection Tool for Ensemble Representations of Intrinsically Disordered States*, la mise en place d'outils spécifiques pour calculer des couplages dipolaires résiduels et de nombreux tests pour valider la consistance de l'approche. Nous avons finalement mis au point une méthode combinant

Flexible-Meccano et Asteroids permettant de sélectionner des ensembles de structures représentatifs de l'état déplié en utilisant les CDRs. Nous allons donc présenter les étapes fondamentales de la description par ensemble sous contraintes et répondre aux questions précédemment évoquées concernant le nombre de structures à utiliser ou la dégénérescence des solutions. Les tests présentés serviront de gages de qualité de la description et seront continuellement repris et appliqués dans les chapitres suivants sur d'autres systèmes.

4.1 La description de l'état *random-coil*

4.1.1 Présentation du logiciel Flexible-Meccano

La description par ensemble explicite consiste à créer un ensemble de conformations[1] dites en échange rapide à l'équilibre qui reproduisent en moyenne les données expérimentales. Nous cherchons à échantillonner l'ensemble du paysage énergétique d'une protéine donnée puis nous comparons les prédictions aux données expérimentales afin de valider ou non le modèle proposé.

L'algorithme Flexible-Meccano développé en 2005 par Bernadó et al. [123] est un générateur de structures respectant les conditions dites *random-coil*, c'est à dire n'ayant pas d'ordre résiduel. Ce modèle statistique ne cherche donc pas à déterminer le minimum de l'espace conformationnel qui serait en accord avec les données expérimentales mais propose simplement une description de l'état déplié des protéines désordonnées. La présence d'ordre, transitoire ou non, au sein de ces protéines n'est donc pas initialement pris en compte dans la description. Les structures générées constituant l'ensemble sont créées selon deux règles :

1 Une base de données d'angles dièdres (ϕ,ψ), spécifique à chaque acide aminé, a été conçue à partir de structures cristallographiques à haute résolution obtenues par diffraction X en retirant préalablement toutes les régions repliées de ces protéines : hélices α, feuillets β (figure 4.1). L'algorithme construit la protéine de la région C-terminale à la région N-terminale. Pour cela il oriente successivement les plans peptidiques de chaque résidu en sélectionnant aléatoirement un couple d'angles dièdres (ϕ,ψ) dans la base de données. Il réitère cette opération indépendamment pour chaque structure de manière à générer l'ensemble.

2 Un test d'exclusion stérique est introduit à chaque ajout d'un plan peptidique afin de prendre en compte l'influence de la chaine principale et des chaines latérales, le modèle d'exclusion stérique est réalisé avec une sphère localisée sur le carbone $^{13}C^{\beta}$ de chaque résidu. Le rayon est différent suivant l'acide aminé considéré pour tenir compte du volume de la chaine latérale.

random-coil ou la base de données standard de Flexible-Meccano

Par définition, nous nommerons par la suite les valeurs ou les ensembles conçus à partir de la base de données standard de Flexible-Meccano les valeurs ou les ensembles *random-coil*.

Nous calculons alors pour chaque structure les paramètres RMN ou SAXS associés et nous moyennons ensuite linéairement ces paramètres sur l'ensemble des structures et ceci spécifiquement pour chaque acide aminé :

- La méthode de calcul des couplages dipolaires résiduels est implémentée à partir des travaux de Berlin et al. [125] en estimant l'alignement stérique de la protéine [82].

- Le calcul de la relaxation paramagnétique est présenté en section 6.1. Le temps de corrélation global vaut 5 ns et le temps de corrélation interne vaut 500 ps. Le

1. Nous utiliserons indifféremment le terme structure ou conformation pour désigner les individus constituant les ensembles.

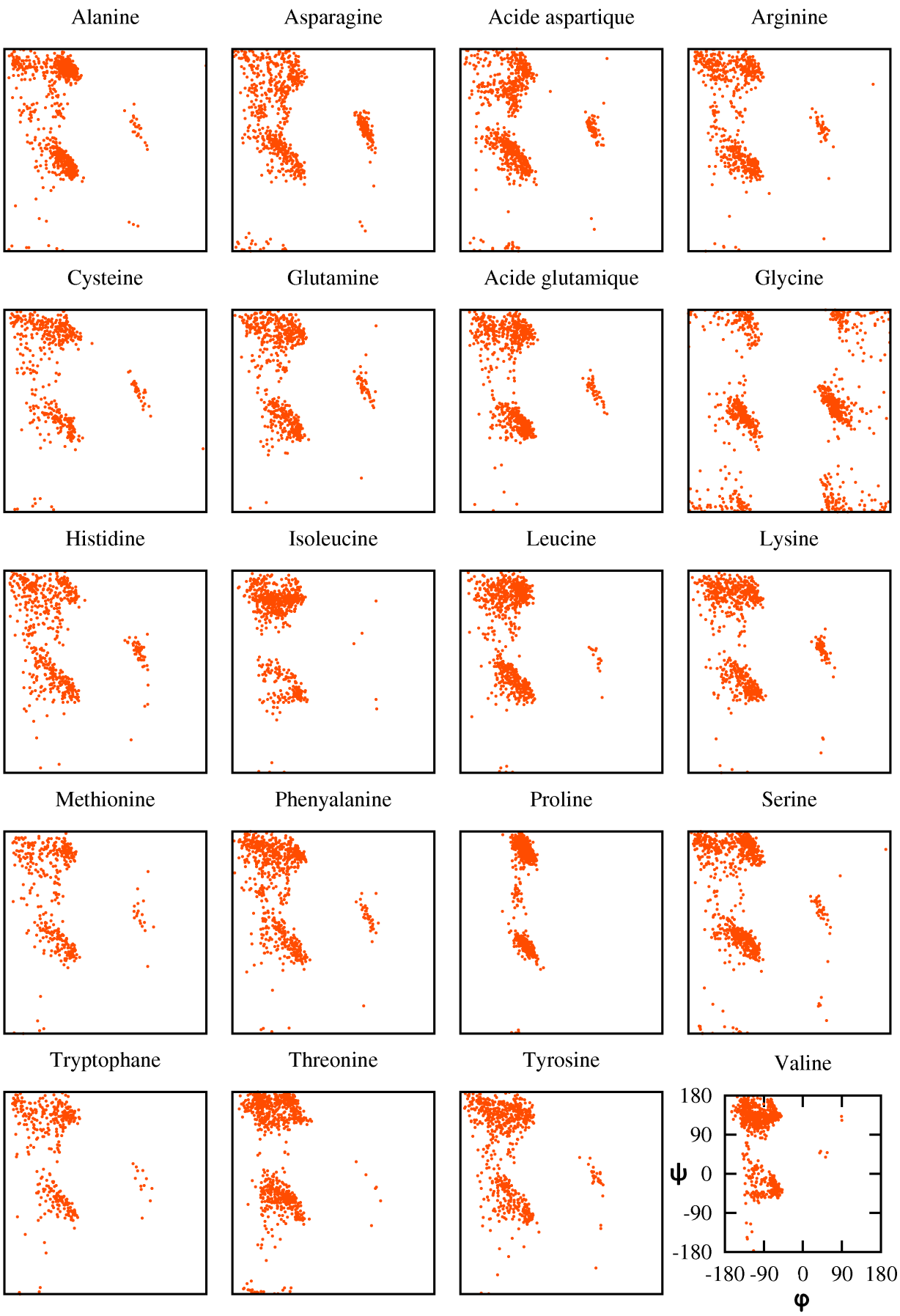


FIGURE 4.1 – ***Distribution par acide aminé des angles dièdres de la base de données standard de* Flexible-Meccano.** *Les 20 acides aminés sont affichés dans l'espace de Ramachandran* $(\phi,\psi)\in[-180^\circ,180^\circ]$.

profil d'intensité I/I_0 est extrait de la formule 2.49.

- Le calcul des déplacements chimiques sera effectué avec le logiciel SPARTA [126] ou SPARTA+ [127] qui seront présentés au chapitre 5. Ils nécessitent l'ajout préalable des chaines latérales avec le logiciel SCOMP [128]. Les déplacements chimiques secondaires sont calculés à partir de la base de donnée refDB [70].

- Le couplage ^3J est calculé à partir de la relation de Karplus [126].

- Le profil d'intensité SAXS est calculé avec le logiciel CRYSOL [129].

Afin de valider la description statistique présentée, nous comparons les valeurs prédites aux valeurs expérimentales. Ce modèle a été testé sur différents systèmes et permet de reproduire relativement bien l'ensemble des données RMN ou SAXS disponibles. Les publications de Bernadó et al. [108], [112] montre la concordance du modèle avec les données RMN et les données SAXS.

De nouveaux jeux expérimentaux sont comparés avec les paramètres RMN prédits sur les pages suivantes avec dans l'ordre : la partie N-terminale N_{tail} de la nucléoprotéine N du virus de la rougeole (en figure 4.2), la région d'appariement de la protéine Tau au microtubulee nommée K18 (en figure 4.3), la partie centrale de la protéine Tau nommée K32 (en figure 4.4), la protéine ACBP pour *Acyl-CoA-binding protein* (en figure 4.5) et la protéine ACTR pour *activator for thyroid hormone and retinoid receptor* (en figure 4.6). Pour chacune, la présence de structures transitoires potentielles ou provenant de la forme native est indiquée dans la partie supérieure de la figure. Les ensembles sont générés avec FLEXIBLE-MECCANO, les paramètres RMN sont calculés sur chaque conformation puis moyennés sur l'ensemble. Le calcul des paramètres RMN est effectué suivant le paragraphe précèdent.

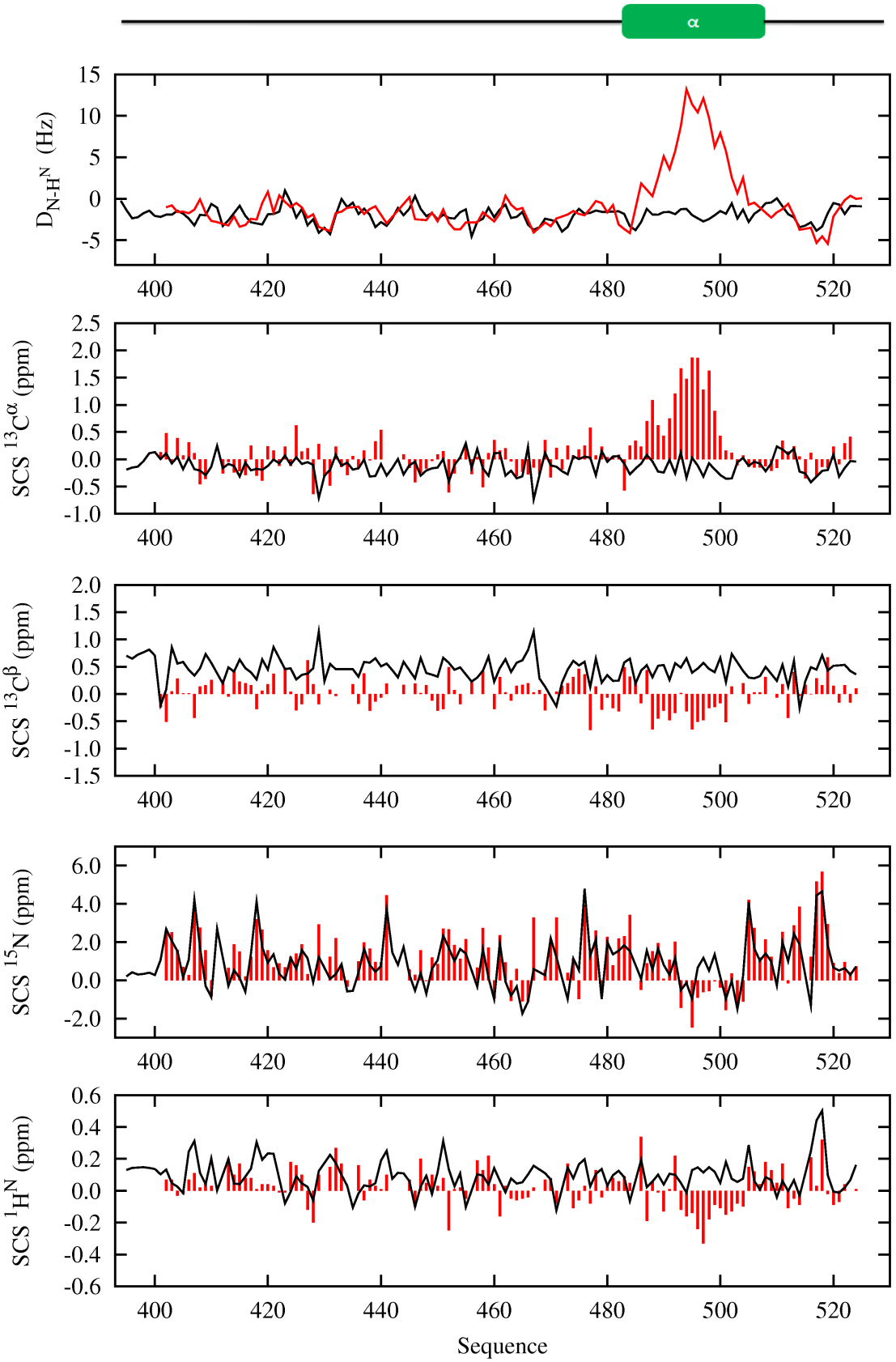


FIGURE 4.2 – ***Comparaison d'une simulation random-coil aux données expérimentales de*** N_{tail}. *Nous générons avec* FLEXIBLE-MECCANO *un ensemble de* 100000 *structures sur lequel est calculé et moyenné les CDRs* D_{NH} *et les déplacements chimiques secondaires* $^{13}C^{\alpha}$, $^{13}C'$, $^{13}C^{\beta}$, ^{15}N *et* $^{1}H^{N}$ *(noir). Les données expérimentales sont en rouge.*

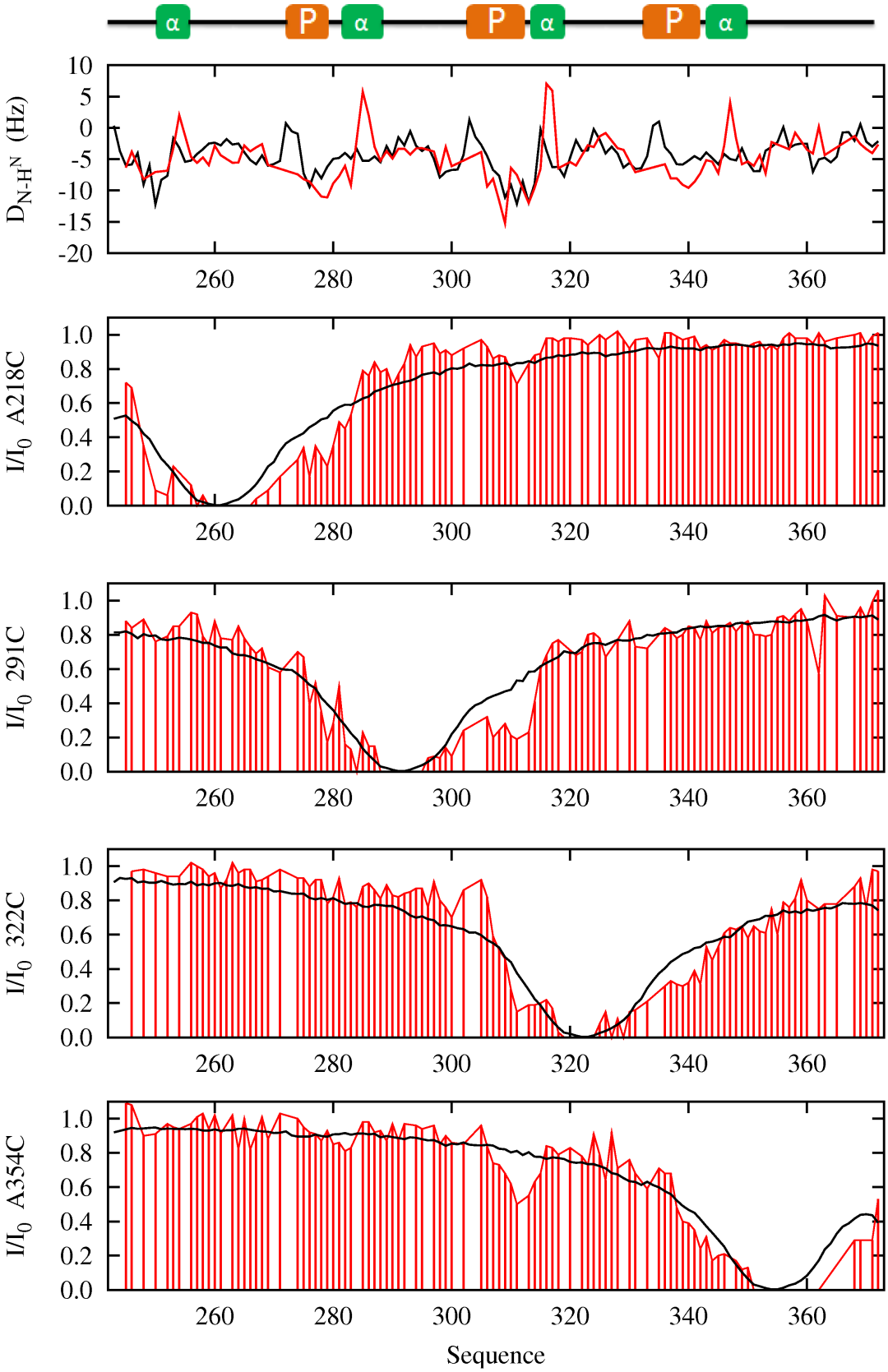


FIGURE 4.3 – ***Comparaison d'une simulation random-coil aux données expérimentales de K18.*** *Nous générons avec* FLEXIBLE-MECCANO *un ensemble de* 100000 *structures sur lequel est calculé et moyenné les CDRs D_{NH} et les valeurs de relaxation paramagnétique des cystéines : A*260*C*, 291*C*, 322*C*, *A*354*C (noir). Les données expérimentales sont en rouge.*

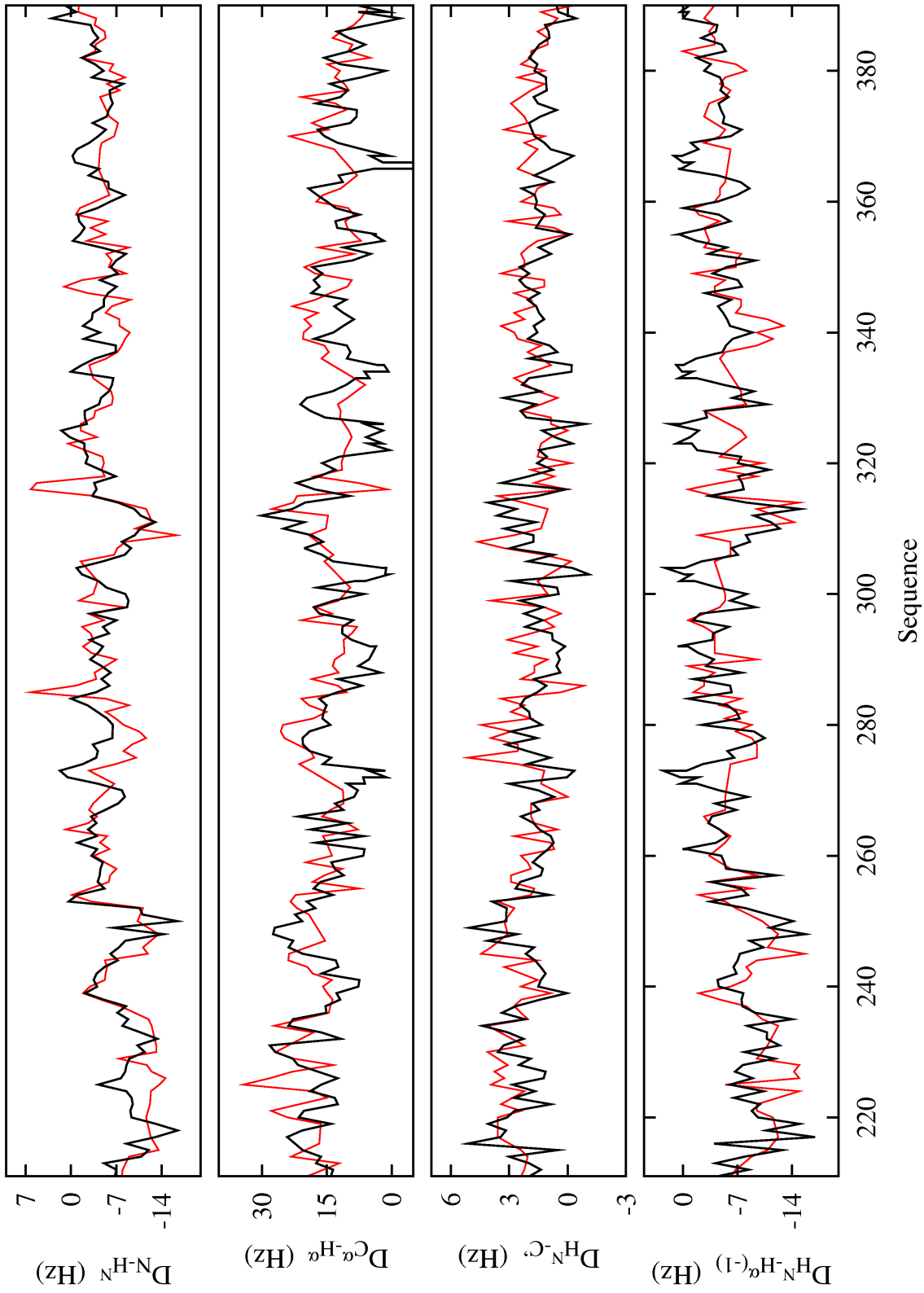


FIGURE 4.4 – ***Comparaison d'une simulation random-coil aux données expérimentales de K32.*** *Nous générons avec* FLEXIBLE-MECCANO *un ensemble de* 150000 *structures sur lequel est calculé et moyenné les CDRs D_{NH}, $D_{C^\alpha H^\alpha}$, $D_{C'H^N}$ et $D_{H^N H^\alpha}$ (noir). Les données expérimentales sont en rouge.*

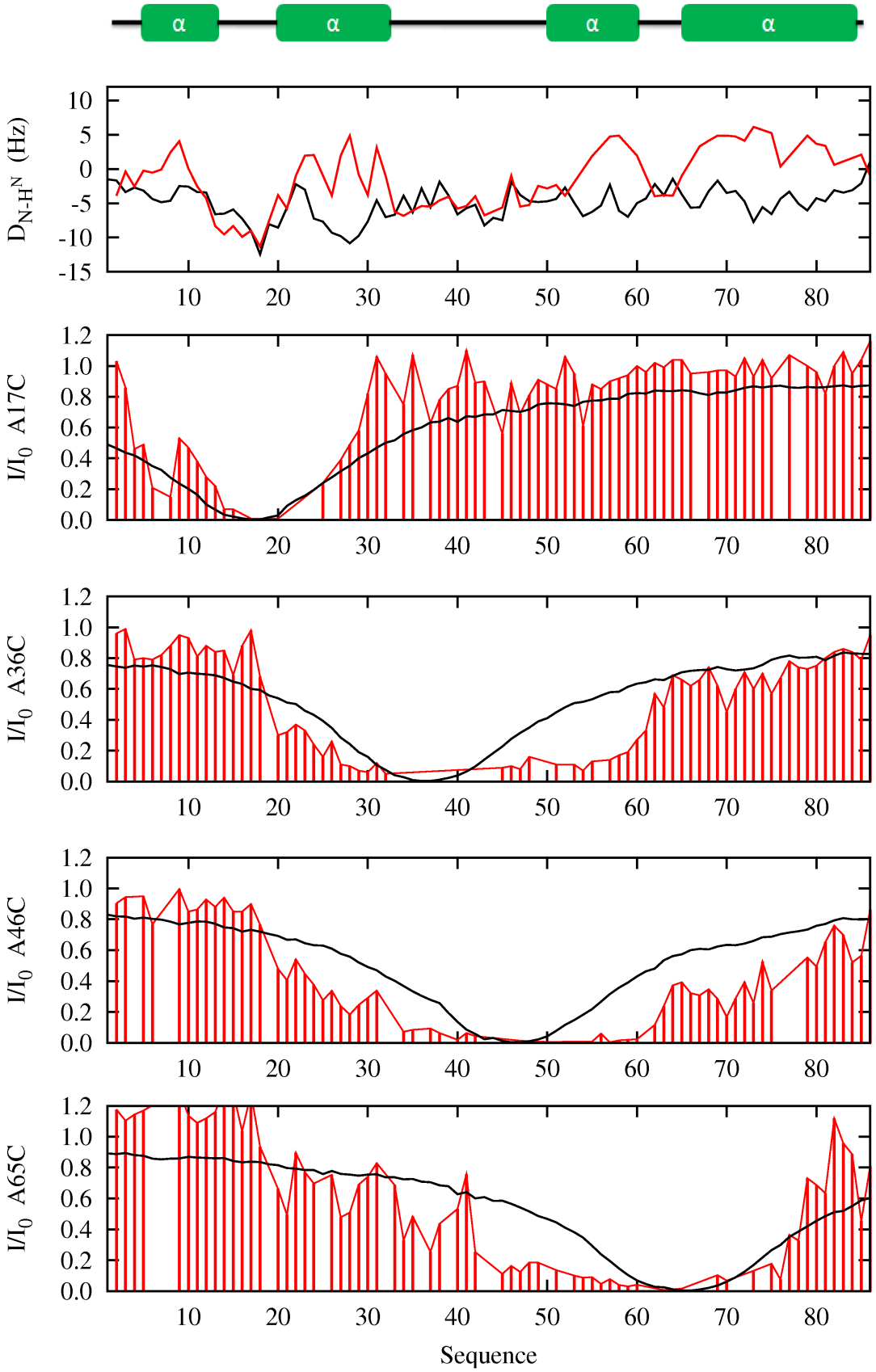


FIGURE 4.5 – ***Comparaison d'une simulation random-coil aux données expérimentales de ACBP.*** *Nous générons avec* FLEXIBLE-MECCANO *un ensemble de* 100000 *structures sur lequel est calculé et moyenné les CDRs D_{NH} et les valeurs de relaxation paramagnétique des cystéines : A17C, A36C, A46C, A65C (noir). Les données expérimentales sont en rouge.*

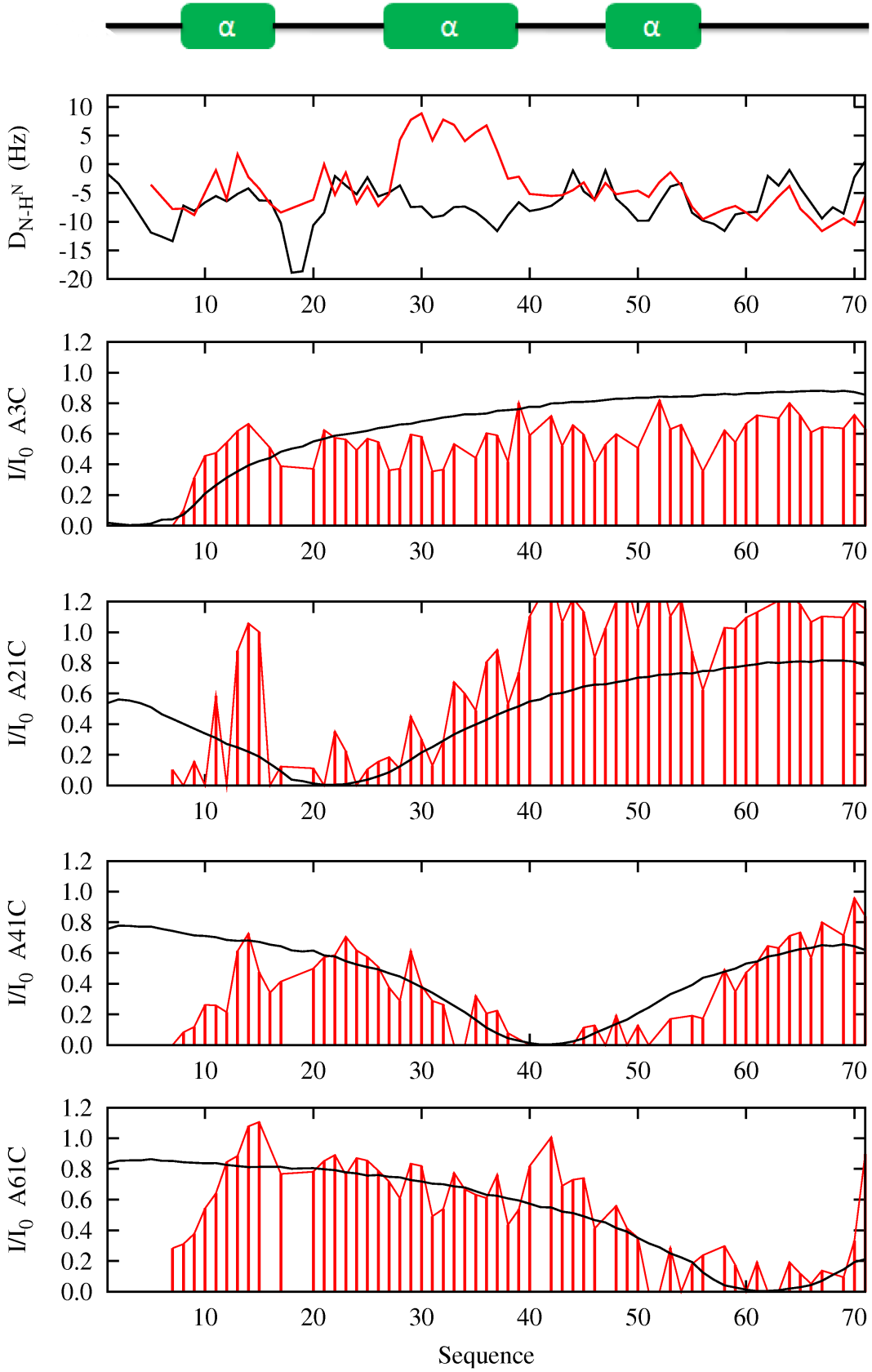


FIGURE 4.6 – ***Comparaison d'une simulation random-coil aux données expérimentales de ACTR.*** *Nous générons avec* FLEXIBLE-MECCANO *un ensemble de* 100000 *structures sur lequel est calculé et moyenné les CDRs D_{NH} et les valeurs de relaxation paramagnétique des cystéines : A*4*C, A*21*C, A*41*C, A*61*C (noir). Les données expérimentales sont en rouge.*

4.1.2 Convergence et nombre de structures

Une des questions clés de la description par ensemble est le nombre de structures nécessaires pour décrire l'état déplié. En pratique, nous n'abordons pas distinctement cette question sous cet angle mais nous observons le nombre de structures assurant la convergence des paramètres biophysiques et des paramètres RMN issus de ce modèle statistique.

Du point de vue de l'échantillonnage, le nombre de degrés de liberté est tel que toutes les combinaisons d'angles (ϕ,ψ) ne peuvent être calculées pour une protéine donnée. Par exemple, considérant 3 couples d'angles dièdres différents par acide aminé pour une protéine de 100 acides aminés, nous obtenons à la limite supérieure 3^{100} possibilités de structures différentes. Si l'on considère la taille de la base de données de Flexible-Meccano dénombrant en moyenne 1000 couples (ϕ,ψ) par acide aminé, le temps de calcul de l'ensemble des conformations serait astronomique. Dans ces conditions, comment pouvons-nous définir le nombre de structures nécessaires et suffisantes pouvant représenter les caractéristiques biophysiques et les différents paramètres RMN de l'ensemble ? Nous nous réfèrerons à la définition suivante :

Le nombre de structures nécessaires

Un ensemble de structures en échange rapide possède un nombre nécessaire et suffisant de structures lorsque l'ajout d'une nouvelle structure ne modifie pas significativement la valeur du paramètre calculé sur l'ensemble des structures. Ainsi, le nombre de structures est fixé par la convergence du paramètre calculé. Ce résultat est explicité pour les CDRs en figure 4.11.

En pratique, nous calculons des ensembles de différentes tailles, et nous définissons une plage de structures à partir de laquelle la valeur du paramètre calculé a convergé. Au-delà de cette plage, l'ensemble reste statistiquement valable mais requiert un coût computationnel plus important.

En d'autres termes, le nombre de structures est fixé à la fois par les caractéristiques physiques de la protéine et par la convergence du paramètre calculé, il n'existe donc pas de règle précise définissant le nombre de structures nécessaires pour représenter l'état déplié. Pour un paramètre RMN donné, ce nombre dépend fortement de la longueur de la protéine, de la présence ou non d'un échantillonnage local spécifique ou d'ordre résiduel. Plus le degré de liberté de la protéine est important, plus il faudra de structures dans l'ensemble pour atteindre l'équilibre statistique.

4.1.3 Les avantages de la description *random-coil*

Un modèle simple et prédictif

Le principal avantage du modèle *random-coil* est sa relative simplicité, il n'inclut qu'un nombre limité d'hypothèses par opposition aux calculs de dynamique moléculaire. Nous distinguons de plus deux atouts majeurs :

- La comparaison des données simulées avec les valeurs expérimentales permet dans un premier temps de valider le modèle mais aussi de détecter la présence

d'ordre local. Les différences entre données simulées et données expérimentales traduisent la présence d'ordre résiduel n'ayant pas été pris en compte dans la description statistique[2]. Cet ordre peut être une structure secondaire transitoire de type hélice α, feuillet β, boucle, un contact transitoire entre deux régions de la protéine.

- En spécifiant un échantillonnage spécifique, nous pouvons alors étudier la dépendance et la sensibilité des paramètres RMN en fonction de l'échantillonnage local. En étudiant le profil des CDRs, il a été possible d'identifier et de quantifier le niveau de structure secondaire au sein de plusieurs protéines [74, 130]. Nous avons aussi par ailleurs mis en évidence la dépendance des couplages dipolaires résiduels vis-à-vis des interactions transitoires à longue portée [131, 113].

Biais et ordre résiduel

Afin d'étudier plus quantitativement l'échantillonnage des protéines, il est possible d'introduire un biais dans l'échantillonnage statistique utilisé par FLEXIBLE-MECCANO. A titre d'exemple, un biais consiste à échantillonner spécifiquement une région de l'espace de Ramachandran sur un ou plusieurs acides aminés afin d'introduire de l'ordre résiduel.

Une application concrète

Considérons le problème suivant extrait des publications de Jensen et al. [130] et Jensen et al. [132] : la protéine N_{tail} présenté en figure 4.2 possèdent une structuration en hélice entre les résidus 480 et 510, afin de caractériser quantitativement la propension de la protéine à échantillonner la région hélicoïdale nous devons déterminer un ensemble en accord avec les données expérimentales. Pour cela une solution consiste à créer non pas un mais plusieurs ensembles possédant des caractéristiques physiques variées (présence d'une hélice de longueur n, de propension 100%, en position x). Calculant pour chaque cas les paramètres RMN correspondants, il suffit de choisir et peser les ensembles reproduisant au mieux les données pour déterminer les caractéristiques du système.

La résolution de ce problème nécessite la création d'un nombre exhaustif d'ensembles représentant l'étendue des solutions. Il a ainsi été possible de quantifier le niveau de structures secondaires présent entre les résidus 480 et 510 de la région N-terminale de la nucléoprotéine N du virus Sendaï [130] et du virus de la Rougeole [132]. Ce dernier cas, présenté en figure 4.7, sera étudié de nouveau au chapitre suivant (en section 5.3.2).

L'introduction de biais pour améliorer la reproduction des données n'est cependant pas généralisable d'un point de vue pratique à l'étude d'une protéine complète en raison du nombre de cas à envisager. En effet, une telle description bien que physiquement réaliste serait d'une part peu convéniente et d'autre part risquerait de mener dans certains cas l'utilisateur à des solutions préconçues ou mal appropriées par méconnaissance de la protéine. Il est donc nécessaire d'envisager une approche complémentaire à la description *random-coil*.

2. Se référer à titre d'exemple aux figures 4.2, 4.3, 4.4, 4.5 et 4.6

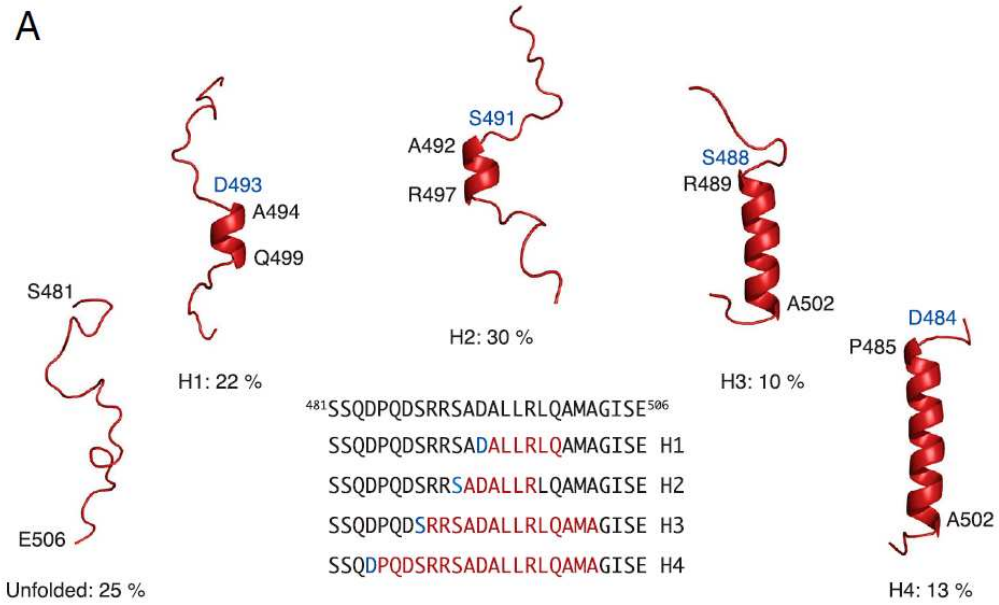


FIGURE 4.7 – ***Caractérisation des hélices α de la protéine*** N_{tail}. *Quatre hélices différentes sont nécessaires pour reproduire correctement les données. Elles sont précédées d'une Sérine ou d'un Acide aspartique servant à stabiliser l'hélice par des interactions N-capping. La figure est extraite de Jensen et al. [132].*

4.2 La sélection de sous-ensembles

4.2.1 Un problème d'optimisation

La comparaison des données expérimentales avec les données prédites issues de Flexible-Meccano n'est parfois pas suffisante pour définir correctement le paysage énergétique des PIDs. Une solution appropriée consiste à contraindre l'ensemble de structures à partir des données RMN. Par contraindre, nous entendons sélectionner le ou les meilleurs sous-ensembles de structures en accord avec les données expérimentales. Nous ne modifions cependant pas la valeur des angles dièdres présents dans l'ensemble mais souhaitons obtenir la distribution reproduisant au mieux les données expérimentales. Cette nuance est de toute importance, la distribution résultante sera donc toujours issue de la base de données standard de Flexible-Meccano.

Nous disposons d'un nombre de données expérimentales limitées qui sera toujours nettement inférieur aux degrés de liberté des ensembles, la solution est donc fortement dégénérée c'est-à-dire qu'il n'existera pas de solution unique ou parfaite à ce problème. Nous sommes confrontés à un problème d'optimisation[3], nous devons déterminer la ou les meilleures solutions en adéquation avec nos données. Une réponse adéquate repose sur l'utilisation d'algorithmes génétiques qui sont adaptés au problème d'optimisation à N variables. Leur efficacité réside d'ailleurs lorsque l'espace de recherche est grand.

4.2.2 L'algorithme génétique ASTEROIDS

Asteroids est un algorithme génétique développé au laboratoire[4] spécialement conçu pour sélectionner des sous-ensembles de structures en accord avec des jeux de données RMN ou SAXS. La fonction d'évaluation des sous-ensembles sélectionnés est le χ^2 traduisant la reproductivité des données définit par :

3. Nous pouvons aussi parler de problème d'analyse quantitative multivariée.
4. Asteroids a initialement été écrit par Gabrielle Nodet puis complété par Loïc Salmon de manière à intégrer l'ensemble des paramètres RMN.

$$\chi^2 = \sum_i \left(\frac{X_i^{simu} - X_i^{exp}}{\sigma_i} \right)^2 \tag{4.1}$$

où X_i^{exp} et X_i^{calc} sont respectivement la valeur expérimentale et simulée du paramètre X de l'acide aminé i avec σ_i l'erreur sur la mesure.

Cette fonction peut être complexifiée en incorporant différents paramètres, la prise en compte du poids relatif des données est alors un point crucial.

$$\chi^2 = \sum_i \left(\frac{X_i^{simu} - X_i^{exp}}{\sigma_{X,i}} \right)^2 + \sum_i \left(\frac{Y_i^{simu} - Y_i^{exp}}{\sigma_{Y,i}} \right)^2 + ... \tag{4.2}$$

où Y_i^{exp} et Y_i^{simu} sont respectivement la valeur expérimentale et simulée du paramètre Y de l'acide aminé i et $\sigma_{Y,i}$ l'erreur respective .

L'algorithme dispose de I solutions (nommé individu) générées aléatoirement, chaque individu est constitué de n structures de l'ensemble *pool*. L'algorithme doit trier les structures de manière à déterminer l'individu I qui reproduit aux mieux les données cibles. L'algorithme tend itérativement vers la meilleure solution en suivant trois règles d'évolution :

0 Nous générons préalablement avec Flexible-Meccano[5] un ensemble de structures *random-coil* ou *pool*.

1 La génération aléatoire de solution : I individus sont aléatoirement créés en sélectionnant à chaque fois n structures du *pool*.

2 Le croisement : les individus de l'étape précédente sont associés par couples, il y a donc I/2 couples. Chaque couple donne alors deux solutions filles par croisement des populations.

3 Les mutations : cela consiste à substituer 1% des structures des individus (ou du moins une structure). Nous différencions les mutations internes et externes suivant l'utilisation où non des structures de l'étape précédente. Le taux de mutation est résolument faible pour conserver le principe de sélection et d'évolution de l'algorithme.

Durant la procédure d'évolution, il est impossible d'obtenir deux fois le même individu, d'autre part chaque individu ne peut contenir deux fois la même structure.

Après chaque évolution, nous avons alors 5I individus qui sont triés par tournoi, la fonction d'évaluation étant le χ^2 définit précédemment. Le nombre de survivants au processus de sélection est réduit progressivement pour éviter toute convergence dans un minimum local.

4.2.3 Les deux tests de validation fondamentaux

Pour chaque jeu de donnée ou protéine concernée, il est nécessaire de tester la sensibilité de l'algorithme et la consistance de l'approche envisagée. Pour cela nous

5. L'exemple n'est pas restrictif à Flexible-Meccano, des structures provenant de dynamique moléculaire peuvent évidemment être utilisées à condition de calculer pour chacune les paramètres RMN correspondants.

effectuerons pour chaque cas les deux tests suivants :

Le test in-silico

Le test in-silico, présenté en figure 4.8, a pour objectif d'évaluer les capacités et la sensibilité de l'algorithme aux paramètres RMN et biophysiques considérées. Il est réalisé avec des données *in-silico* nommées aussi données synthétiques. Il précèdera toujours la validation croisée.

1 Nous créons avec FLEXIBLE-MECCANO un ensemble nommé cible ayant des caractéristiques biophysiques spécifiques (contact, structure résiduelle)

2 Nous calculons les jeux de données correspondants dites données cibles ou données de références.

3 Nous créons un *pool random-coil*.

4 Nous sélections avec ASTEROIDS un sous-ensemble de structures issu du *pool* qui reproduisent en moyenne les données cibles.

5 Nous comparons les données et les caractéristiques biophysiques du sous-ensemble obtenu avec celles de l'ensemble cible. La reproduction des données est assurée tandis que la reproduction des caractéristiques biophysiques est nettement plus délicate.

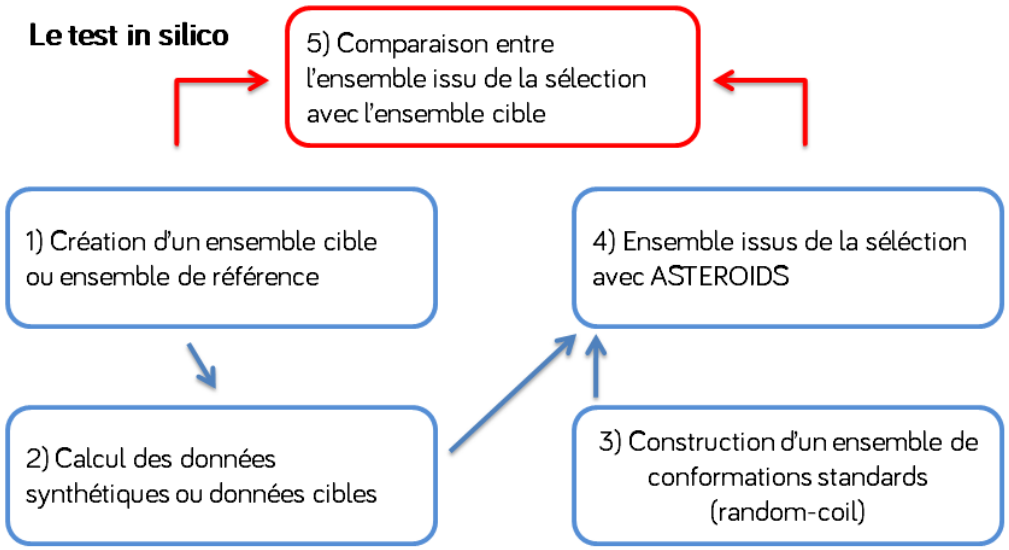


FIGURE 4.8 – ***Protocole du test in-silico.*** *Le protocole se déroule en cinq étapes : 1) Création d'un ensemble de référence 2) Calcul des paramètres correspondants qui serviront de cible, 3) Création d'un ensemble random-coil, 4) Sélection avec* ASTEROIDS *d'un sous-ensemble en accord avec les données, 5) Comparaison des données et des caractéristiques de l'ensemble sélectionné avec l'ensemble de référence.*

La validation croisée

La validation croisée présentée en figure 4.9 est uniquement réalisée avec des données expérimentales. Ce test a pour objectif d'évaluer la consistance de l'approche envisagée et est d'une manière générale un gage de qualité.

1 Partants d'un ou plusieurs jeux de données expérimentales, nous excluons un pourcentage des données. Il est possible d'exclure une partie des données d'un même paramètre RMN ou au contraire d'un autre paramètre RMN pour étudier l'interdépendance de ces derniers.

2 Nous créons un *pool random-coil*.

3 Nous sélections avec ASTEROIDS un sous-ensemble de structures issu du *pool* qui reproduit les données expérimentales.

4 Nous comparons les données expérimentales exclues avec les valeurs prédites issues de la sélection. De plus, nous comparons les paramètres biophysiques obtenus à l'issu de la validation croisée à ceux obtenus pour une sélection où toutes les données sont incluses. Si les caractéristiques sont similaires, nous disposons alors d'un nombre suffisant de données pour caractériser le système.

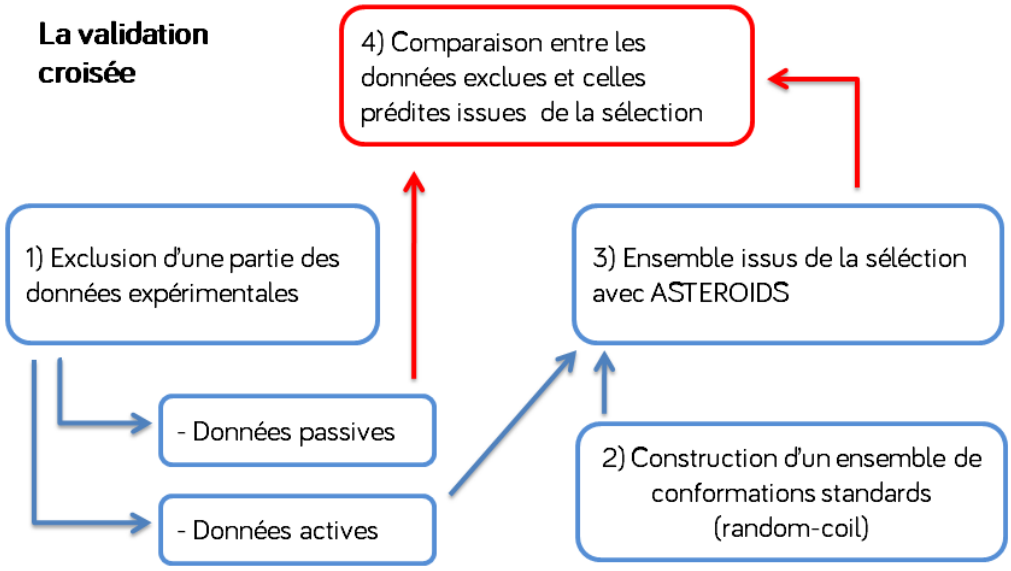


FIGURE 4.9 – ***Protocole de la validation croisée.*** *Le protocole se déroule en quatre étapes : 1) Exclusion d'une partie des données expérimentales 2) Création d'un ensemble random-coil, 3) Sélection avec* ASTEROIDS *d'un sous-ensemble en accord avec les données actives 5) Comparaison des données passives avec celles prédites issues de la sélection.*

4.3 APPLICATION AUX COUPLAGES DIPOLAIRES RÉSIDUELS DE L'UBIQUITINE DÉNATURÉE DANS L'URÉE

4.3.1 Introduction

La description *random-coil* est un outil pratique pour caractériser qualitativement les protéines désordonnées et estimer la présence d'ordre résiduel au sein d'une protéine.

Dans le cadre d'une étude plus générale, nous souhaitons exploiter la sensibilité des couplages dipolaires résiduels pour obtenir une description quantitative de l'échantillonnage conformationnel d'une protéine résolue à l'échelle de l'acide aminé.

Ce chapitre présente les fondamentaux de la description par ensemble sous contraintes. La sélection de sous-ensembles est un problème d'optimisation complexe qui nécessite à la fois une bonne compréhension des couplages dipolaires résiduels et une étude approfondie de la sensibilité de l'algorithme génétique aux données RMN. Les problèmes rencontrés sont à la fois d'ordre computationnel, et théorique. Aussi évident que cela puisse paraitre il s'agit d'une part de calculer et reproduire les données expérimentales correctement et d'autre part de déterminer ou reproduire les caractéristiques biophysiques associées à ces données.

- Dans un premier temps, nous introduirons de nombreux concepts et définirons la plage de fonctionnement de l'algorithme sur des données synthétiques. Différents points seront abordés, le nombre de structures d'un sous-ensemble, l'introduction de la fenêtre glissante, la sensibilité des couplages dipolaires résiduels à l'information à longue portée, la dégénérescence des solutions, l'utilisation du χ^2, la reproduction des caractéristiques biophysiques. Cette partie offrira un aperçu des problèmes de sélection pouvant intervenir lors de la caractérisation de protéines intrinsèquement désordonnées. Nous présenterons ainsi les problèmes courants et associés à la description par ensemble sous contraintes. Notamment, nous souhaitons avertir tout utilisateur de la finesse de cette approche sans évidemment remettre en question son utilisation. Nous montrerons ainsi qu'il est possible moyennant l'utilisation de 200 structures de déterminer un échantillonnage conformationnel précis avec l'algorithme Asteroids.

- Dans un second temps, nous appliquerons le protocole retenu et déterminerons le paysage énergétique de l'Ubiquitine dénaturée dans l'urée. Nous verrons ainsi que l'Ubiquitine adopte des conformations plus étendues que celle initialement présente dans un ensemble *random-coil*. Cela nous permettra par ailleurs de présenter les outils et concepts utilisés pour comparer les propriétés biophysiques des protéines désordonnées.

Cette partie est bien évidemment focalisée sur les couplages dipolaires résiduels de la protéine Ubiquitine mais a pour vocation de mettre en évidence les problèmes liés à l'utilisation d'un algorithme génétique. Les concepts rencontrés et méthodes utilisées s'appliqueront de manière similaire dans les chapitres suivants.

4.3.2 Matériel et méthodes

Données expérimentales

Les données expérimentales de l'Ubiquitine dans l'urée à 8M ont été mesurées par Meier et al. [133], la protéine a été alignée avec un gel polyacrylamide étiré à $pH = 2.5$. Nous disposons de 8 couplages dipolaires : D_{NH}, $D_{C'C^\alpha}$, $D_{C^\alpha H^\alpha}$, $D_{H^N H^\alpha}$, $D_{H_i^N H_{i-1}^\alpha}$, $D_{H_i^N H_{i+1}^N}$, et $D_{H_i^N H_{i+2}^N}$.

Calcul des couplages dipolaires résiduels

Nous distinguons deux méthodes de calcul des couplages dipolaires résiduels :

1 La méthode du tenseur d'alignement global est réalisée en estimant le tenseur d'alignement de chaque structure de l'ensemble. Nous supposons la structure

rigide sur un intervalle de temps suffisant long. Cette méthode développée par Zweckstetter [82] est la plus courante. Les CDRs seront alors notés D_{IS}^{GT}.

2 La méthode du tenseur d'alignement local ou de la fenêtre glissante consiste à ne considérer qu'une section de la protéine de longueur n (n étant impair) pour estimer le tenseur d'alignement stérique puis à calculer uniquement la valeur des couplages dipolaires du résidu central $(n+1)/2$ de ce domaine [134]. Cette opération est alors répétée en décalant séquentiellement la fenêtre afin de définir un nouveau domaine juxtaposé et ainsi calculer la valeur des couplages du résidu central suivant. Les CDRs seront alors notés D_{IS}^{LAW}. LAW signifiant *Local Alignement Window*.

La fenêtre glissante

Par définition, nous nommerons fenêtre glissante de longueur n, l'utilisation d'un tenseur local d'alignement pour calculer la valeur des CDRs sur une région de la séquence de longueur n.

Ligne de base

La ligne de base est obtenue en l'absence d'interaction locale, c'est dire sans échantillonnage conformationnel spécifique. Pour cela, nous créons à partir d'une séquence polyvaline de 100 acides aminés un ensemble de 100000 structures et calculons le profil des CDRs avec le tenseur global. La ligne de base d'une protéine s'extrait aisément de la valeur des couplages dipolaires D_{NH} et $D_{C^\alpha H^\alpha}$ en utilisant la paramétrisation suivante :

$$B(x) = 2b\cosh(a(x-d)) - c \tag{4.3}$$

où x est numéro d'acide aminé, a, b, c et d des constantes dépendant uniquement de la longueur de la chaine.

La paramétrisation est réalisée par minimisation dans le logiciel GNUPLOT des paramètres a, b et c. d étant égal à $\frac{n+1}{2}$.

Les régions de l'espace Ramachandran

Afin de quantifier la reproduction de l'échantillonnage conformationnel de nos ensembles lors de test *in-silico*, nous divisons l'espace de Ramachandran en 4 régions (figure 4.10) :

- α_L : $\phi > 0^\circ$
- α_R : $\phi < 0^\circ$ et -120° < ψ < 50°
- β_S : -180° < ϕ < -90° et $\psi < 120^\circ$ ou $\psi > 50^\circ$
- β_P : -90° < ϕ < -0° et $\psi < 120^\circ$ ou $\psi > 50^\circ$

Les angles dièdres de chaque résidu sont alors comptabilisés par quadrant puis normalisés de manière à obtenir la probabilité d'échantillonner telle ou telle région de l'espace Ramachandran. Nous introduisons la métrique suivante :

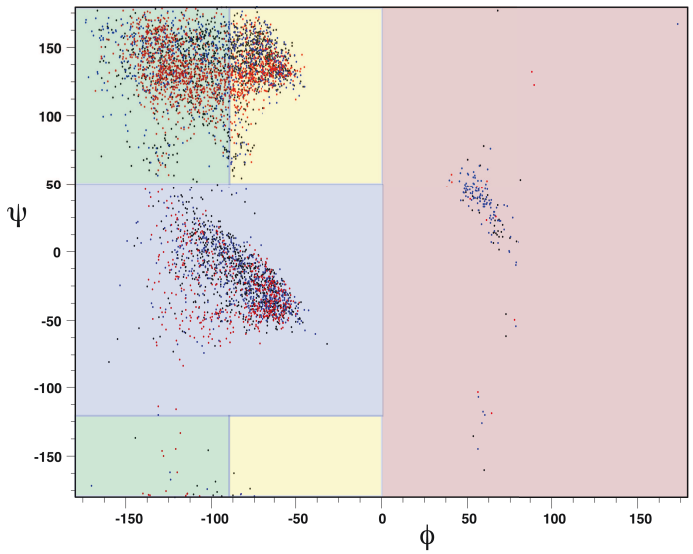


FIGURE 4.10 – ***Définition des* 4 *régions de l'espace Ramachandran.*** *Les régions sont α_L (en rouge), α_R (en bleu), β_P (en jaune) et β_s (en vert). Les points correspondent à l'échantillonnage standard, ou random-coil, des acides aminés suivant : Valine (en rouge), Lysine (en bleu) et Leucine (en noir).*

$$\chi^2_{rama} = \sum_{i,q} (p^{select}_{i,q} - p^{cible}_{i,q})^2 \tag{4.4}$$

où i est le numéro d'acide aminé, q le cadrant associé à la population p_q.

Ces régions permettent de plus d'estimer les déviations de l'échantillonnage conformationnel de sélections avec des données expérimentales en prenant pour référence les populations d'un ensemble *random-coil*.

Sélection et jeux de données simulées

Nous définissons deux régimes : le premier utilise l'échantillonnage standard de FLEXIBLE-MECCANO *i.e. random-coil*, nous le nommerons ici (S), le second échantillonne plus souvent les régions étendues de l'espace Ramachandran ($50^\circ < \psi < 180^\circ$), nous le nommerons (E). Les probabilités d'être dans la région étendue sont 78% pour le régime (E) et 59% pour le régime (S). Nous calculons les CDRs correspondants et utilisons ces données comme cible de l'algorithme ASTEROIDS. Les données synthétiques sont moyennées sur 50000 structures pour atteindre la convergence.

L'ensemble *pool*, c'est-à-dire à partir duquel les structures sont sélectionnées, est constitué des deux ensembles (S) et (E) et contenant chacun 6000 structures. Nous effectuons parallèlement la sélection d'ensembles ayant pour cible l'échantillonnage (S) ou (E).

La sélection avec ASTEROIDS est réalisée avec 100 individus, le nombre de structures par individus sera précisé dans la partie Résultat suivant le cas considéré. 2000 itérations d'évolution sont nécessaires pour assurer la convergence des sous-ensembles obtenus.

Le poids des couplages dipolaires est optimisé en fonction de leur gamme respective, soit 1.0 pour D_{NH} et $D_{H^N_i H^\alpha_{i-1}}$, 2.0 pour $D_{C^\alpha H^\alpha}$, 0.5 pour $D_{C'C^\alpha}$, $D_{H^N H^\alpha}$ et $D_{H^N_i H^N_{i+1}}$ et 0.33 pour $D_{H^N_i H^N_{i+2}}$. Les CDRs sont évidemment moyennés sur l'ensemble puis multiplié à chaque itération part un facteur (*scaling factor*) pour tenir compte de l'alignement relatif

du milieu dans lequel ils sont mesurés. Nous utiliserons dans un deuxième temps deux facteurs multiplicatifs K_1 et K_2 appliqués respectivement au couplage D_{NH}, $D_{C^\alpha H^\alpha}$, $D_{C'C^\alpha}$ et $D_{H^N H^\alpha}$, $D_{H_i^N H_{i-1}^\alpha}$, $D_{H_i^N H_{i+1}^N}$, $D_{H_i^N H_{i+2}^N}$.

4.3.3 Résultats

Convergence et fenêtre glissante

La description par ensemble nécessite l'étude préalable de la convergence des paramètres calculés. La convergence d'un paramètre est validée si en calculant m simulations différentes de n structures, la valeur moyenne prédite du paramètre est similaire. Les couplages dipolaires étant extrêmement sensibles à l'échantillonnage conformationnel des protéines, nous avons dans un premier temps cherché à déterminer le nombre de structures nécessaires pour atteindre l'équilibre statistique.

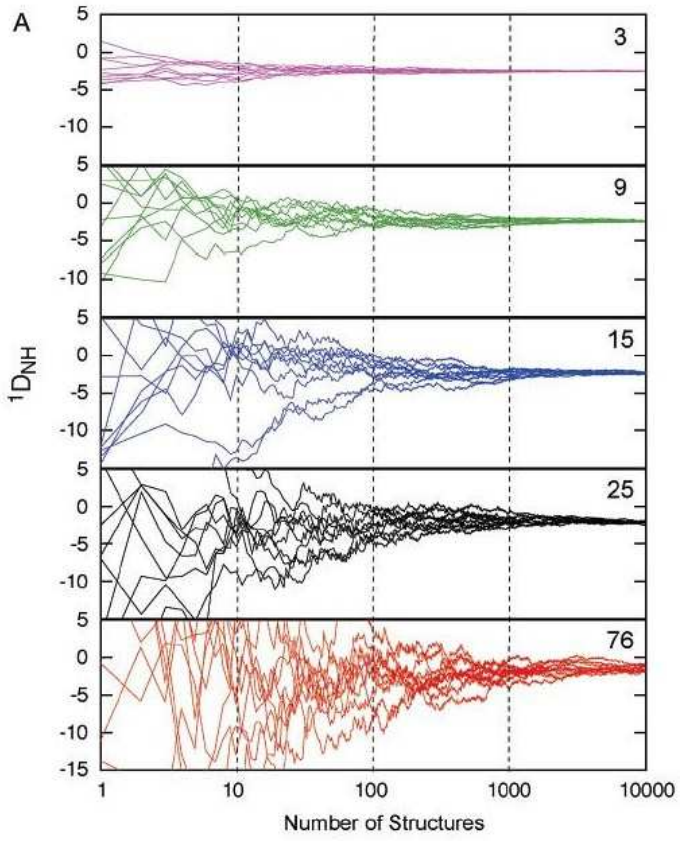


FIGURE 4.11 – ***Convergence du couplage dipolaire résiduel D_{NH}.*** *Convergence du couplage dipolaire résiduel D_{NH} du résidu* 41 *de l'Ubiquitine en fonction du nombre de structures et pour différentes longueurs de fenêtre glissante :* 3 *(en violet),* 9 *(en vert),* 15 *(en bleu),* 25 *(en noir) ou d'un tenseur d'alignement global de 76 résidus (en rouge). L'usage de la fenêtre glissante améliore concrètement la convergence des CDRs.*

La réponse, présentée en figure 4.11, diffère nettement suivant la méthode de calcul utilisée. Un ensemble structure de 10000 est nécessaire pour atteindre la convergence si l'on considère la méthode du tenseur d'alignement global. L'application de la fenêtre glissante, c'est-à-dire d'un tenseur d'alignement de plus petite dimension, permet de réduire fortement le nombre de structures, l'ordre de grandeur retenu est une gamme proche de la centaine. Cette solution développée se révèle nettement plus adaptée à nos problématiques de sélection.

En effet, la sélection d'un sous-ensemble de structures en accord avec les données expérimentales nécessite la convergence des paramètres calculés sur les dimensions de cet ensemble sous peine de retranscrire une information erronée. Si cette condition n'est pas respectée, l'échantillonnage conformationnel déterminé ne sera pas spécifiquement relié aux données mesurées.

La méthode de la fenêtre glissante influence nettement la gamme des CDRs obtenus qui deviennent environ 5 à 10 fois inférieurs aux valeurs calculées avec le tenseur d'alignement global. L'application d'un facteur multiplicatif permet de comparer ces deux jeux de données, cependant la reproduction se révèle imparfaite aux extrémités de la séquence (figure 4.13). Cette différence est due aux effets de ligne de base prédite théoriquement par Louhivuori et al. [85] et repris par Obolensky et al. [86]. Ces équipes modélisent la protéine par polymère effectuant une marche aléatoire dans un milieu orienté, ils ont mis en évidence l'influence de la longueur de la chaine peptidique sur son degré de flexibilité et par conséquent sur la valeur des couplages dipolaires résiduels.

Ligne de base et reproduction des données

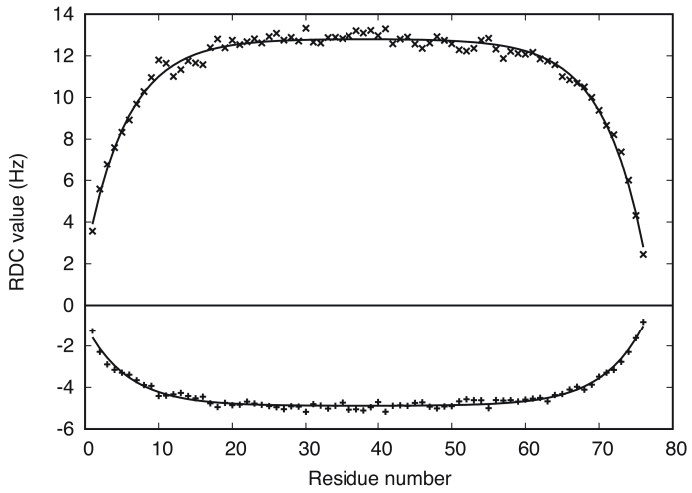


FIGURE 4.12 – ***Paramétrisation de la ligne de base.*** *Valeurs des CDRs D_{NH} et $D_{C^\alpha H^\alpha}$ d'un polyvaline de 76 résidus généré avec* FLEXIBLE-MECCANO *(en pointillé), courbe issue de la paramétrisation dont l'équation est présenté en 4.3 (en trait continu). La paramétrisation est identique pour les deux couplages.*

Nous observons une courbe en forme de cloche traduisant le degré de flexibilité de la chaine (figure 4.12). La flexibilité étant plus importante aux extrémités, la valeur de la ligne de base tend vers 0. La ligne de base est une caractéristique de la chaine peptidique est donc unique pour tous les couplages résiduels dipolaires.

Nous pouvons dès lors multiplier la ligne de base par la valeur des CDRs calculés avec la fenêtre glissante pour retrouver la valeur des CDR calculés avec le tenseur global affiché en figure 4.13 :

$$\mathrm{D}_{\mathrm{IS}}^{\mathrm{GT}}(x) = |B(x)|\mathrm{D}_{\mathrm{IS}}^{\mathrm{LAW}}(x) \tag{4.5}$$

où x est numéro d'acide aminé, $B(x)$ est la paramétrisation de la ligne de base, $\mathrm{D}_{\mathrm{IS}}^{\mathrm{LAW}}(x)$ les CDRs calculés avec la fenêtre glissante et $\mathrm{D}_{\mathrm{IS}}^{\mathrm{GT}}(x)$ les CDRs calculés avec le tenseur d'alignement global.

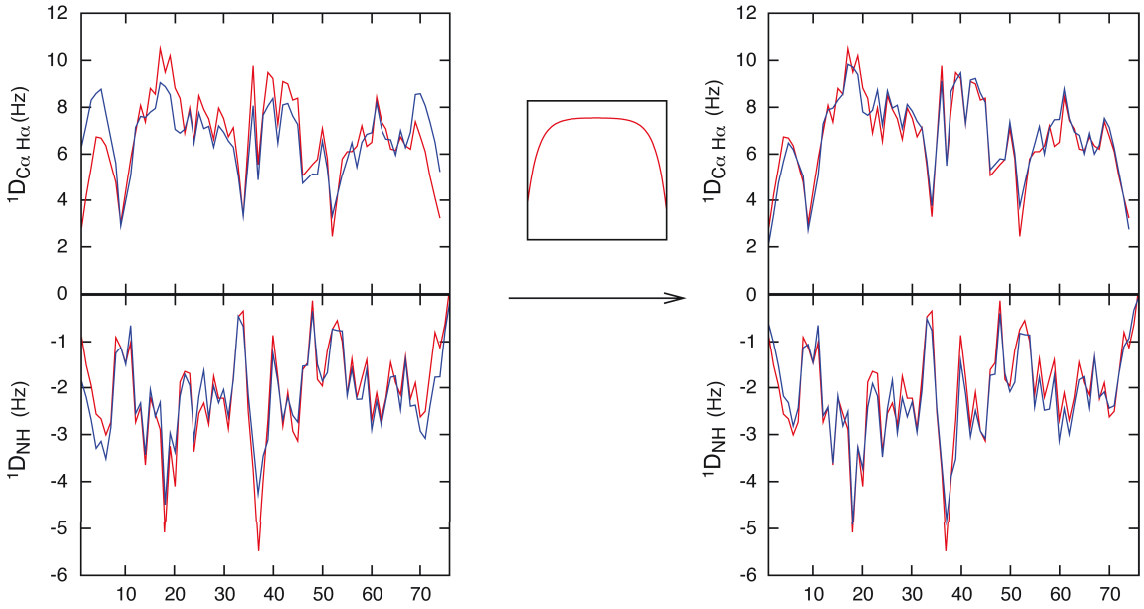


FIGURE 4.13 – ***Influence de la ligne de base sur les CDRs.*** *Le calcul des CDRs avec la fenêtre glissante sur 200 structures (en bleu) ne prend pas en compte les interactions à longue portée, la comparaison de ces derniers avec les CDRs calculés avec le tenseur d'alignement global sur 100000 structures (en rouge) est présentée à gauche. Les extrémités de la séquence ne sont pas correctement reproduites. L'incorporation de la ligne de base avec l'équation 4.5 module correctement la courbe des CDRs (à droite).*

La combinaison des CDR calculés en utilisant la fenêtre glissante avec l'information à longue distance caractérisée par une ligne de base permet de reproduire correctement le profil des CDRs calculés avec le tenseur d'alignement global (figure 4.13) et ceci avec un nombre de structure de l'ordre de la centaine.

Information locale et à longue portée

Les CDRs sont à la fois sensibles à l'échantillonnage local et à l'information à longue portée. L'utilisation de la fenêtre glissante est nécessaire pour accélérer la convergence des CDRs mais entraine la perte de l'information à longue distance de la chaine dépliée. Pour pallier ce problème, nous réintroduisons ces effets à longue portée par l'intermédiaire de la ligne de base.

Afin de fixer la longueur de la fenêtre glissante, nous comparons la reproduction des CDRs calculés de part et d'autre de l'équation 4.5. La métrique considérée étant le χ^2 suivant :

$$\chi^2_{CDR} = \sum_i (D^2_{i,LAW} - D^2_{i,GT}) \tag{4.6}$$

Les résidus voisins, c'est-à-dire une fenêtre glissante de 3 acides aminés, ne permettent pas d'obtenir le χ^2 optimal. La persistance de l'information locale des CDRs est donc supérieure aux uniques résidus voisins[6].

Afin de reproduire correctement les valeurs des couplages dipolaires (figure 4.14), nous choisissons une longueur de fenêtre glissante de 15 résidus permettant d'après la figure 4.11 d'obtenir l'équilibre statistique pour une centaine de structures.

6. Pour information, nous déterminerons précisément la longueur de persistance des CDRs vis-à-vis de l'échantillonnage conformationnel au chapitre 5 (en section 5.2.5) et nous comparons ce résultat avec celui des déplacements chimiques.

Caractérisation de l'échantillonnage conformationnel *in silico*

Nous allons maintenant tester à la fois la capacité de l'algorithme à reproduire un échantillonnage spécifique et déterminer le nombre de structures nécessaires à la description. Nous effectuons donc des tests *in-silico* avec pour cible soit l'ensemble étendu (E), soit l'ensemble standard (S), le *pool* étant constitué de la somme de l'ensemble (E) et l'ensemble (S). Les résultats étant très similaires, les figures concernant l'échantillonnage (E) en tant que cible ne seront pas affichées.

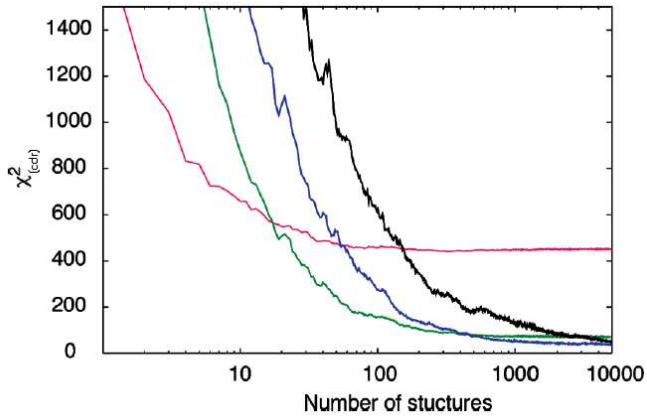


FIGURE 4.14 – ***Détermination de la longueur de la fenêtre glissante.*** *Evolution du χ^2_{CDR} en fonction du nombre de structures et pour différentes longueurs de fenêtre glissante :* 3 *(en violet),* 9 *(en vert),* 15 *(en bleu), et* 25 *(en noir).*

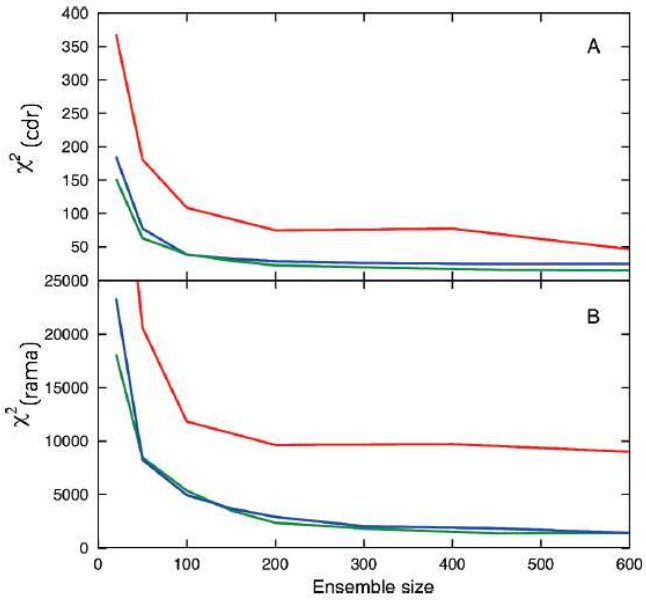


FIGURE 4.15 – ***Détermination de la taille du sous-ensemble sélectionné.*** *Evolution du χ^2_{CDR} et χ^2_{rama} en fonction du nombre de structures et pour différentes longueurs de fenêtre glissante :* 9 *(en vert),* 15 *(en bleu) et avec un tenseur d'alignement global (en rouge).*

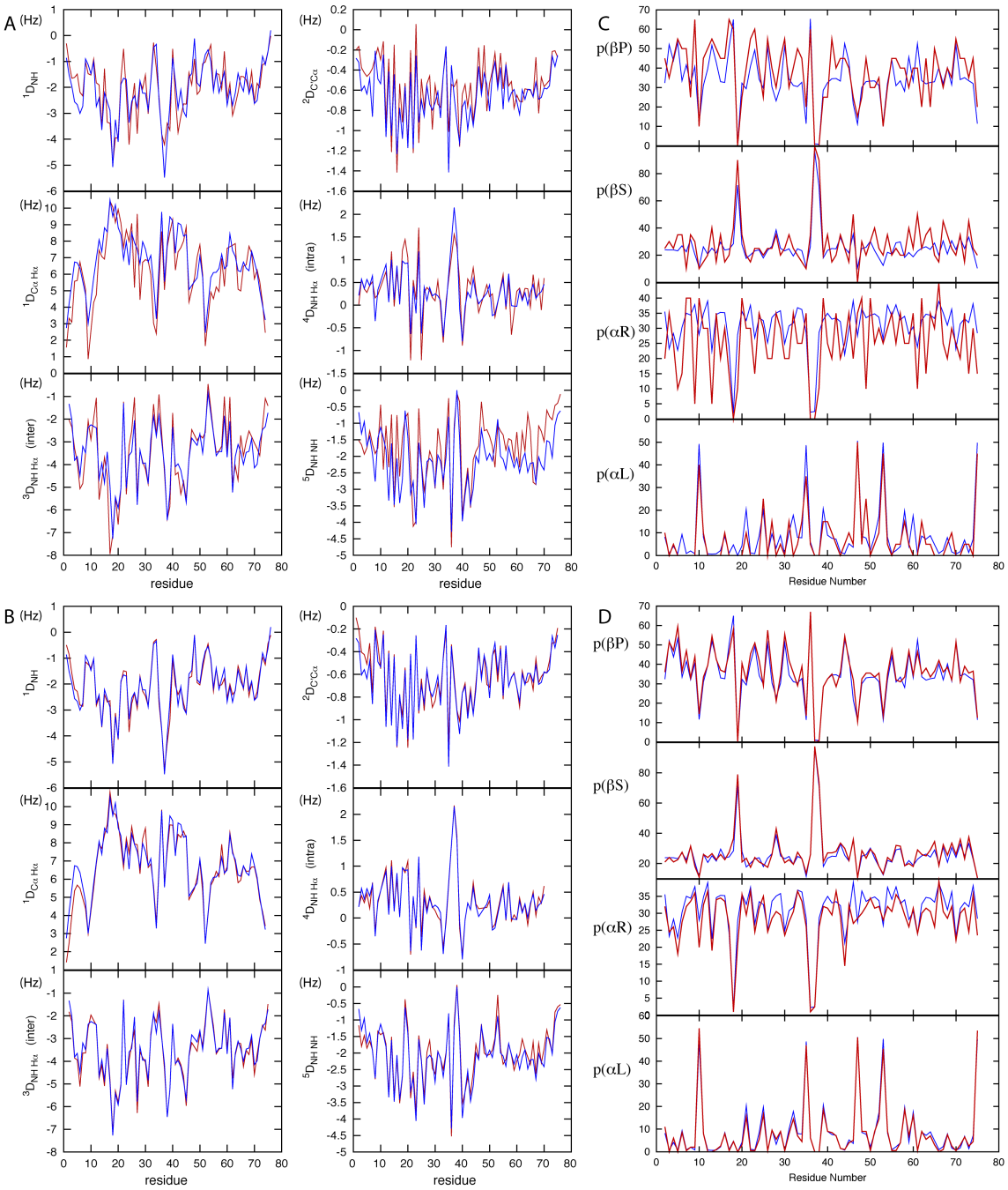


FIGURE 4.16 – ***Reproduction des CDRs et de l'échantillonnage conformationnel pour un ensemble de 20 et 200 structures.*** *Reproduction des CDRs (A-B) et de l'échantillonnage conformationnel (C-D) issus de la sélection avec* ASTEROIDS *sur des données synthétiques : (A) et (C) sous-ensembles de 20 structures, (B) et (D) sous-ensembles de* 200 *structures. Pour (A) et (B), sont présents les couplages D_{NH}, $D_{C^\alpha H^\alpha}$, $D_{C'C^\alpha}$, $D_{H^N H^\alpha}$, $D_{H_i^N H_{i-1}^\alpha}$ et $D_{H_i^N H_{i+1}^N}$ avec les données cibles (en bleu), les données issues de la sélection avec* ASTEROIDS *(en rouge). Pour (C) et (D), population des quatre régions de l'espace Ramachandran, sont affichés l'échantillonnage cible (en bleu) et l'échantillonnage issu de la sélection (en rouge).*

Résumé

- Le calcul des CDRs avec le tenseur d'alignement global n'est pas adapté aux problématiques de sélection de sous-ensembles de petites tailles (*i.e.* de l'ordre de 10^2 structures).

- La reproduction des valeurs cibles est une condition nécessaire mais insuffisante pour conclure sur la qualité de la sélection.

- L'utilisation d'un nombre trop faible de structures ne garantit pas la reproduction du bon échantillonnage conformationnel de l'ensemble cible. Nous utiliserons par la suite sauf mention contraire des ensembles de 200 structures et une fenêtre glissante de 15 résidus.

Quel que soit le nombre de structures présentes dans ces simulations, le tenseur global d'alignement ne permet ni d'obtenir la meilleure reproduction des données, ni le meilleur échantillonnage conformationnel (figure 4.15). Cette solution pourtant couramment utilisée avec un faible nombre de structures (*i.e.* inférieur à 10000) n'est absolument pas appropriée pour caractériser les protéines désordonnées. L'utilisation d'une fenêtre glissante de 9 ou 15 acides aminés semble quand à elle prometteuse. La figure 4.15 indique qu'un ensemble de l'ordre de 200 structures offre une bonne reproduction des caractéristiques de l'ensemble cible.

Choisissant alors une fenêtre glissante de 15 résidus, nous effectuons une étude comparative similaire entre un sous-ensemble de 20 structures et un sous-ensemble de 200 structures. Au vu de la figure 4.16, bien que la reproduction des CDRs soit dans les deux cas au dessus de la gamme de bruit de mesures expérimentales, l'échantillonnage conformationnel de l'ensemble de 20 structures n'est absolument pas satisfaisant avec une marge d'erreur atteignant parfois 30% de la valeur souhaitée.

Application aux données de l'Ubiquitine

Nous générons un ensemble pool contenant 12000 structures (se référer à la section 4.3.2) à partir duquel l'algorithme génétique Asteroids sélectionne des sous-ensembles de 200 structures en utilisant une fenêtre glissante de 15 acides aminés. Afin de reproduire correctement les données, deux facteurs multiplicatifs furent utilisés : $K_1 = 0.58$ pour les couplages D_{NH}, $D_{C^\alpha H^\alpha}$, $D_{C'C^\alpha}$ et $K_2 = 0.96$ pour les couplages $D_{H^N H^\alpha}$, $D_{H_i^N H_{i-1}^\alpha}$, $D_{H_i^N H_{i+1}^N}$ et $D_{H_i^N H_{i+2}^N}$. Les résultats de la sélection sont présentés en figure 4.17.

La reproduction des données est nettement meilleure en utilisant deux facteurs multiplicatifs. Ce résultat proviendrait de la dynamique des mouvements relatifs des plans peptidiques qui modifie la gamme du couplages proton-proton entre résidus. La distance n'est plus fixe, nous avons dans ce cas une distance effective à prendre en compte. N'incorporant pas cet effet dans notre modèle, nous utilisant alors un deuxième facteur d'ajustement.

Pour analyser en détail l'échantillonnage obtenu, nous pouvons comparer en figure 4.17 les populations d'angles dièdres résultantes de la sélection à celles d'un ensemble *random-coil*. Nous constatons une diminution générale des populations p(αR) au profits des populations étendues p(βS) et p(βP). Une autre façon d'observer ces modifications consiste à affiche la densité d'angle (ϕ,ψ) par acide aminé dans l'espace

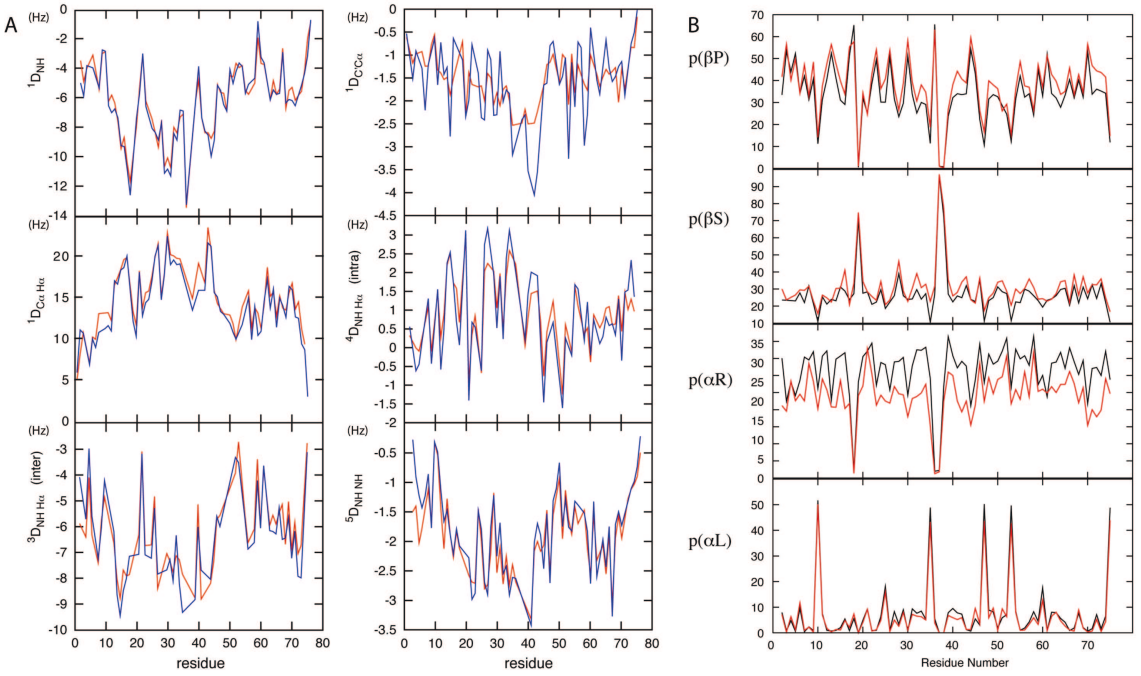


FIGURE 4.17 – ***Reproduction des CDRs et de l'échantillonnage conformationnel de l'Ubiquitin dénaturée dans l'urée.*** *Pour (A), sont présent les couplages D_{NH}, $D_{C^\alpha H^\alpha}$, $D_{C'C^\alpha}$, $D_{H^N H^\alpha}$, $D_{H_i^N H_{i-1}^\alpha}$ et $D_{H_i^N H_{i+1}^N}$, la sélection (en rouge) a été effectuée avec deux facteurs multiplicatifs, les données expérimentales sont juxtaposées (en bleu). Pour (B), échantillonnage conformationnel de l'Ubiquitine (en rouge) comparé à celui du random-coil (en noir).*

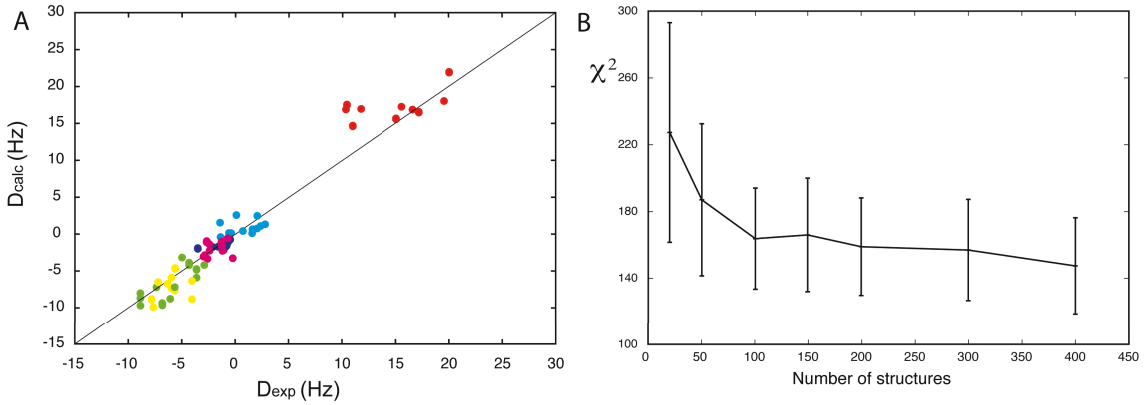


FIGURE 4.18 – ***Validation croisée des données de l'Ubiquitine dénaturée dans l'urée.*** *(A) Validation croisée des CDRs issus de la sélection avec* ASTEROIDS*, les données passives sont comparées aux données expérimentales : D_{NH} (en vert), $D_{C^\alpha H^\alpha}$ (en rouge), $D_{C'C^\alpha}$ (en bleu foncé), $D_{H^N H^\alpha}$ (en cyan), $D_{H_i^N H_{i-1}^\alpha}$ (en jaune) and $D_{H_i^N H_{i+1}^N}$ (en violet). (B) Reproduction des données passives en fonction du nombre de structures dans le sous-ensemble. Pour chaque dimension,* 10 *répétitions ont été effectuées afin de moyenner le χ^2.*

de Ramachandran. Pour cela, nous récupérons pour chaque acide aminé l'ensemble des angles dièdres, ces points sont ensuite interpolés et tracés avec un gradient de couleur. Nous pouvons ainsi visualiser à l'échelle atomique l'échantillonnage conformationnel de la protéine (figure 4.19A). Une dernière représentation graphique qui sera souvent utilisée ultérieurement consiste à normaliser la densité d'angle (ϕ,ψ) résultant de la sélection avec la densité d'angle (ϕ,ψ) d'un ensemble *random-coil* de manière à localiser les modifications induites par la sélection (figure 4.19B). Nous identifions un régime plus étendu pour les acides aminés Thréonine, Arginine et Acide Glutamique à l'opposé des acides aminés hydrophobes qui présentent des modifications conformationnelles mineures. L'inclusion des molécules d'urée autour de la chaine principale influencerait l'échantillonnage conformationnel adopté, ce point a été confirmé par Huang et al. [135] en combinant des données RMN et SAXS.

Pour tester la robustesse du protocole, nous effectuons des validations croisées, c'est à dire, nous retirons préalablement 10% des données expérimentales sur l'ensemble des couplages avant d'effecter de nouvelles sélections pour des sous-ensembles allant de 25 à 400 structures (figure 4.18 A). Nous comparons alors la reproduction des données passives, c'est-à-dire non incluses dans la sélection, avec les données expérimentales. Deux conclusions importantes sont à souligner : d'une part, tous les couplages sont correctement reproduits, nous disposons donc d'un nombre suffisant de données pour caractériser notre système. Considérant le nombre de degrés de liberté d'un ensemble de protéines désordonnées, ce résultat valable pour ce jeux de couplages dipolaires n'est cependant pas généralisable. Il devra être vérifié au cas par cas pour l'ensemble des simulations futures. D'autre part, la reproduction des données passives dépend fortement de la dimension des sous-ensembles, l'évolution du χ^2 en figure 4.18 B indique un plateau à partir de la centaine de structures, ce qui confirme nos choix précédents d'utiliser 200 structures. La validation croisée est aussi un test efficace pour quantifier le nombre de structures nécessaires lors de description par ensemble sous contraintes.

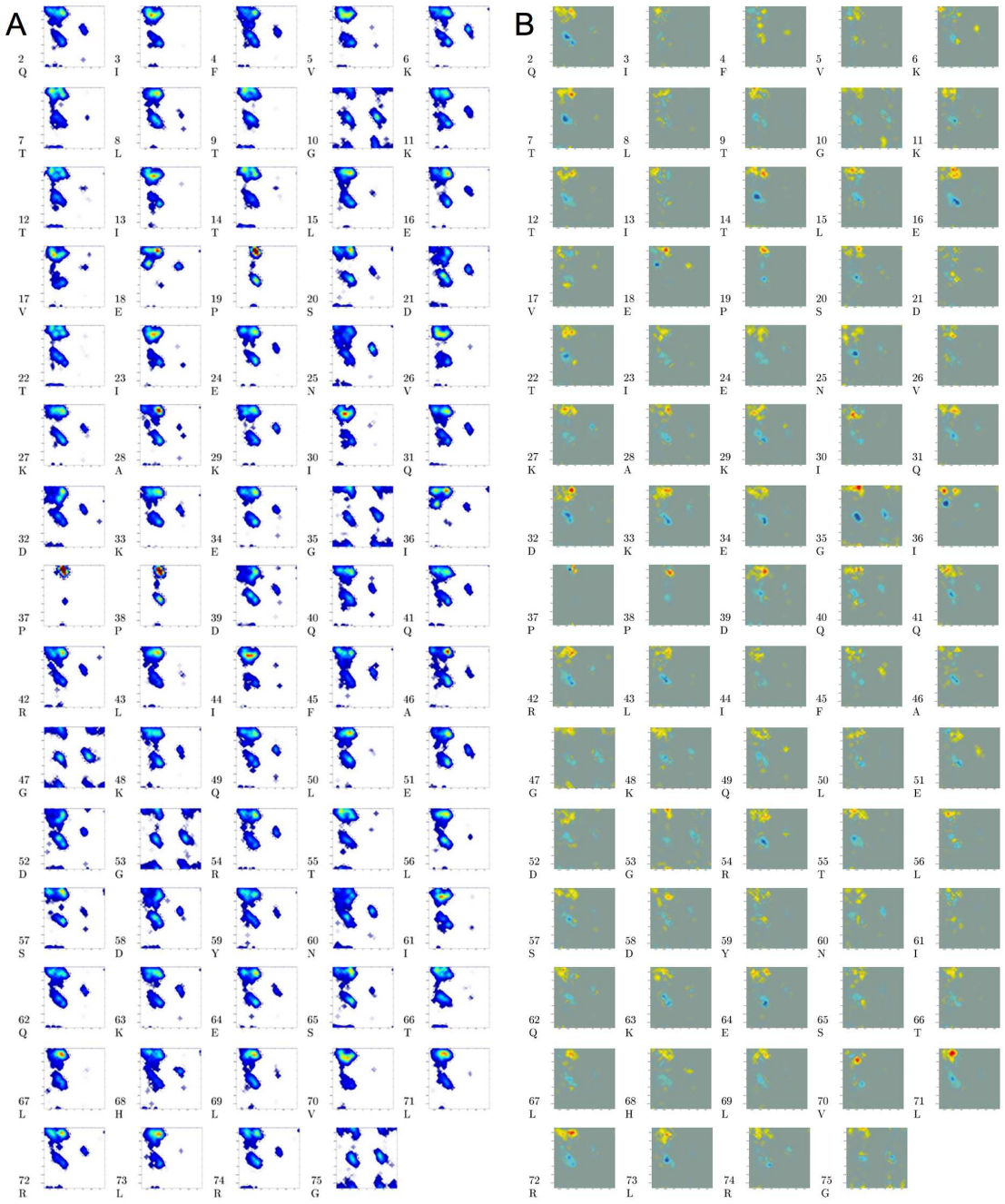


FIGURE 4.19 – ***Distribution des angles dièdres de l'Ubiquitin dénaturée dans l'urée résolue à l'échelle de l'acide aminé.*** *(A) Echantillonnage conformationnel issu de la sélection avec* ASTEROIDS. 10 *répétitions sont effectuées afin de créer une base de données de* 2000 *structures. Le code de couleur traduit la densité d'angles, par ordre croissant nous avons blanc (population inexistante), bleu foncé, cyan, vert, jaune et rouge (population très dense). (B) Idem mais normalisé avec la distribution d'un ensemble random-coil. Le code de couleur est alors le suivant, augmentation de la population en jaune puis rouge, diminution de la population en cyan puis bleu.*

Conclusion du chapitre

Ce chapitre présente les fondamentaux de la description par ensemble. Ce projet qui débuta courant des années 2000 commença par la mise en place d'un modèle statistique permet de caractériser l'état déplié. Utilisant comme modèle la description par ensemble explicite de structure, l'équipe a développé l'algorithme Flexible-Meccano générant des structures PDB sans contrainte structurale, c'est à dire *random-coil*. À partir de cette description, il est possible de calculer les paramètres RMN et de les moyenner sur l'ensemble pour reproduire les mesures expérimentales de nombreux systèmes désordonnées. Ainsi après avoir présenté le fonctionnement de l'algorithme, nous avons exposé les avantages et qualités de cette approche et nous l'avons illustré sur cinq protéines désordonnées sur lesquelles nous disposions de différentes mesures RMN : déplacements chimiques, couplages dipolaires résiduels et relaxation paramagnétique.

La seconde partie de ce projet fut la mise en place d'un protocole à la fois quantitatif et généralisable à l'ensemble des systèmes désordonnées. Il s'agissait d'introduire les paramètres RMN comme contraintes dans la description. Nous nous sommes heurtés à un problème bien plus complexe que prévu. Le système étudié est fortement dégénéré, nous avons constaté avec des tests in-silico que la reproduction des données ne garantissait pas la reproduction de l'échantillonnage conformationnel. Pour résoudre ce problème, nous avons répondu à plusieurs questions essentielles, tel que le nombre de structures nécessaires dans un sous-ensemble pour décrire l'état déplié. Pour une protéine donnée de longueur N, le nombre de structure est fixé par le paramètre RMN servant de cible. Après introduction d'une méthode de calcul des couplages dipolaires combinant l'information locale : la fenêtre glissante et l'information à longue portée : la ligne de base, nous avons choisi pour caractériser l'Ubiquitine des sous-ensembles de 200 structures. Cette taille permet de reproduire correctement les données et les caractéristiques biophysiques de l'ensemble. Nous avons appliqué le protocole aux données expérimentales et identifié l'échantillonnage conformationnel de l'Ubiquitine dénaturée dans l'urée, ce dernier est plus étendu que celui d'un ensemble *random-coil*. Pour statuer sur la qualité de la description, nous avons réalisé des validations croisées en retirant 10% des données qui ne contredisait pas la reproduction des données. Cela nous a permis d'évaluer une nouvelle fois dans quelle mesure notre protocole permet de décrire l'état déplié d'une protéine sans ambiguïté.

5 Une méthode pour quantifier précisément l'ordre local des protéines intrinsèquement désordonnées

Ce chapitre présente le travail de thèse de 3e année, il se situe dans la continuité des projets passés cherchant à caractériser l'espace conformationnel des protéines intrinsèquement désordonnées. Deux publications ont été réalisées par le groupe dans ce sens durant mes deux premières années de thèses, l'une présentée au chapitre précédent (en section 4.3) montre comment les couplages dipolaires résiduels peuvent être utilisés comme contraintes dans une description par ensemble pour caractériser la flexibilité de la protéine Ubiquitine dénaturée dans l'urée. La seconde réalisée environ un an plus tard par Jensen et al. [136] combine les déplacements chimiques dans une description par ensemble sous contraintes pour caractériser la partie C-terminale N_{tail} de la nucléoprotéine N du virus Sendaï.

La question posée est alors la suivante : pouvons-nous définir un protocole utilisant le minimum de données RMN permettant de caractériser n'importe quel échantillonnage conformationnel d'une protéine intrinsèquement désordonnée quelconque ? Nous souhaitons évidemment utiliser les déplacements chimiques et/ou les couplages dipolaires résiduels les plus courants. La mise en place d'un tel protocole serait d'un intérêt majeur pour la communauté RMN et faciliterait à la fois le travail d'interprétation et pourrait améliorer la compréhension générale de l'état déplié. La détermination de structures transitoires de type hélices α ou feuillet β est un sujet déjà bien traité dans la littérature. Il existe plusieurs algorithmes permettant à partir des déplacements chimiques de prédire le type et la propension de structures secondaires α, β et PPII dans les protéines désordonnées [137, 138]. Nous proposons ici une méthode généralisable à l'ensemble de l'espace Ramachandran, c'est à dire pouvant aussi quantifier précisément le degré d'échantillonnage de la région Polyproline que nous étudierons tout particulièrement.

Nous allons dans un premier temps examiner la signature des paramètres RMN pour trois motifs structurés : hélices α, feuillets β, hélices Polyproline. Nous nous intéresserons particulièrement à la relation entre les valeurs des paramètres et l'échantillonnage correspondant. En effet, une des questions sous-jacentes à examiner préalablement à toute description par ensemble sous contraintes est si nous pouvons à partir des valeurs discriminer chaque région de l'espace Ramachandran. Nous illustrerons ces propos sur un exemple très simple en premier temps, un tripeptide AAA et montrerons que la réponse concernant la région Polyproline est particulièrement ambigüe. Nous identifierons plusieurs dégénérescences intrinsèques aux paramètres RMN, ce qui nous amènera à proposer un protocole performant combinant les déplacements chimiques

$^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N et $^{1}H^{N}$ et les couplages dipolaires résiduels D_{NH}. Cette méthode sera premièrement testée sur un système de référence *in-silico* puis sera appliquée sur deux protéines désordonnées : la partie C-terminale N_{tail} de la nucléoprotéine N du virus de la rougeole et sur la protéine K18, région où la protéine Tau s'apparie aux microtubules.

Nous caractériserons ainsi les spécificités structurales de ces deux protéines désordonnées. Nous identifierons plusieurs régions transitoirement structurées préalablement connues et analyserons la distribution de l'échantillonnage conformationnel correspondant aux régions dites *random-coil*. Nous montrerons la place importante de la région Polyproline dans l'échantillonnage des angles dièdres des protéines désordonnées : nous noterons une diminution globale de l'échantillonnage de la région βS au profit de la région βP sur deux systèmes expérimentaux. Concernant la caractérisation structurale de la protéine K18, nous déterminerons la nature des motifs situés de la région d'appariement de la protéine Tau qui ne sont pas des feuillets β mais des motifs PPII. Notre méthode semble la première à souligner et quantifier à l'échelle atomique une telle tendance à échantillonner si distinctement la région Polyproline.

5.1 Matériel et méthodes

5.1.1 Le calcul des déplacements chimiques

Présentation des logiciels de prédiction

Les déplacements chimiques (présentés en section 2.2) sont particulièrement ambigus à interpréter, ils contiennent à la fois de l'information structurale et dynamique et disposent d'une sensibilité unique mais restent difficiles à prédire théoriquement. Ce paradoxe a amené certains groupes à développer des approches empiriques permettant de prédire la valeur des déplacements chimiques d'une structure donnée. L'enjeu est important, cela permet de prédire les spectres d'une protéine pour faciliter l'attribution, d'améliorer la compréhension globale des déplacements chimiques et ainsi d'envisager la détermination de structures uniquement à partir de ces derniers [139, 140, 141]. Nous allons donc brièvement présenter quelques-uns de ces logiciels :

- Sparta : Le programme Sparta [126] prédit les déplacements chimiques d'un tripeptide donné. La méthode de calcul est une approche statistique mettant en jeu l'analogie entre les paramètres de la structure voulue et ceux d'une base de données de 200 structures cristallographiques à haute résolution, l'attribution de chacune de ces protéines étant complète. Les paramètres pris en considération sont notamment le type d'acide aminé, les angles (ϕ,ψ) de la chaine principale et l'orientation des chaines latérales.

- Sparta+ : Le programme Sparta+ est une mise à jour de Sparta. La structuration de l'algorithme est profondément modifiée, Shen and Bax [127] utilisent comme modèle de calcul un réseau de neurones artificiels[1] qui améliore concrètement la vitesse de l'algorithme. Néanmoins, le modèle physique repose essentiellement sur les mêmes paramètres.

- ShiftX2 : Le programme ShiftX2 est analogue au programme Sparta+, il combine d'une part, une approche statistique et d'autre part une approche perceptive pour calculer les valeurs des déplacements. Le logiciel prend notamment en compte la température et le pH [142, 143].

- CamShift : Le programme CamShift établit une relation entre l'environnement conformationnel et la distribution de distances interatomiques. Kohlhoff et al. [144] modélisent cette relation par une fonction polynomiale, les termes de cette série traduisant les paramètres physiques à considérer : le type d'acide aminé, d'atome, les distances covalentes.

Initialement conçus pour des protéines repliées, ces programmes prédisent la valeur des déplacements chimiques d'une structure. Il est aisé d'appliquer la prédiction à la description par ensemble en calculant la valeur des déplacements chimiques pour chaque structure puis en les moyennant. Une inconnue subsiste, pour chaque logiciel, les auteurs valident la consistance de leur logiciel en comparant les données prédites sur des structures connues et attribuées[2], ils peuvent ainsi estimer l'erreur sur la prédiction. Dans le cas d'un ensemble de structures, qu'advient l'erreur à l'issue de la moyenne des déplacements chimiques ? A priori, deux cas existent, nous sommes soit en présence d'une erreur systématique qui restera similaire quel que soit le nombre de structures, soit en présence d'une erreur aléatoire pouvant tendre vers zéro à l'issue de la moyenne. La réponse à cette question est évidemment difficile à déterminer mais

1. C'est un algorithme à la fois statistique et perceptif.
2. Les structures choisies ne sont donc pas celles incluses dans la base de données

nous mentionnons le problème pour souligner la dépendance de l'approche vis-à-vis des logiciels de prédiction.

Dépendance structurale des déplacements chimiques

Nous créons un ensemble de 50000 pentadecapeptide polyalanine, les valeurs de déplacement chimique de chaque conformation sont calculées avec SPARTA. Nous disposons alors d'une base de donnée de 50000 angles dièdres avec les valeurs des déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{1}H^{\alpha}$, ^{15}N et $^{1}H^{N}$ correspondant. Nous moyennons ces derniers localement dans l'espace de Ramachandran afin d'assurer la convergence des valeurs. : tous les déplacements chimiques situés dans un cercle de rayon 1° centré en position (ϕ_0,ψ_0) sont moyennés.

5.1.2 Le calcul des couplages dipolaires résiduels

Dépendance structurales des couplages dipolaires résiduels

Nous créons un ensemble de 500000 pentadeca-peptide polyalanines avec FLEXIBLE-MECCANO. Les CDRs sont calculés sur chaque structure avec le logiciel PALES [82] puis moyennés sur un rayon de 1° dans l'espace de Ramachandran afin d'assurer la convergence des valeurs.

Mise en évidence de la longueur de persistance des CDRs

Nous créons un ensemble de 500000 pentadeca-peptides polyalanines avec FLEXIBLE-MECCANO dépourvus de contraintes stériques entre atomes et pour lequel on ne restreint pas l'échantillonnage, les angles (ϕ,ψ) peuvent donc échantillonner toutes les valeurs comprises entre [-180 :180][-180 :180]. Les couplages dipolaires résiduels sont calculés sur chacun des peptides puis après les avoir tracés dans l'espace Ramachandran en fonction de la valeur des angles (ϕ,ψ) respectifs ils sont moyennés sur une fenêtre de 2 degrés par 2 degrés afin d'assurer la convergence des valeurs.

Nous répétons ensuite ce calcul en utilisant cette fois-ci l'échantillonnage conformationnel de l'acide aminé Alanine et calculons de la même manière les CDRs.

5.1.3 Données cibles

Nous considérons une protéine de 60 acides aminés, dont la séquence est arbitraire, nous avons cependant retiré les résidus Glycines et Prolines car leur échantillonnage conformationnel est particulier et nécessite une étude séparée. Nous définissons trois fragments partiellement structurés long chacun de 5 acides aminés, espacés de plus de 15 résidus avec une propension de 50%, ces éléments seront générés de manière non coopérative, nous distinguons : un élément échantillonnant la région αR (en position 10-14), un élément échantillonnant la région βS (en position 27-31) et un élément échantillonnant la région βP (en position 45-49).

Nous générons cet ensemble de 100000 structures sur lequel nous calculons les CDRs suivants D_{NH}, $D_{C^{\alpha}H^{\alpha}}$, $D_{C'C^{\alpha}}$ et $D_{C'H^{N}}$ avec un tenseur d'alignement global. Ces couplages moyennés sur l'ensemble serviront de données cibles dans les tests *in-silico*. Parallèlement pour chaque structure, nous ajoutons les chaines latérales avec le logiciel SCOMP puis les protons avec le logiciel REDUCE, nous calculons ensuite avec le logiciel

SPARTA ou SPARTA+ les valeurs de déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N et $^{1}H^{N}$ qui sont ensuite moyennées sur l'ensemble.

5.1.4 Sélection avec Asteroids

Le protocole est largement inspiré de celui utilisé au chapitre 4 :

0 Nous générons un *pool* de 20000 conformations *random-coil* avec FLEXIBLE-MECCANO.

1 Pour chaque conformation, nous ajoutons les chaines latérales avec le logiciel SCOMP.

2 Nous ajoutons ensuite les protons avec le logiciel REDUCE.

3 Nous calculons les déplacements chimiques avec logiciel le SPARTA (ou avec le logiciel SPARTA+) et les CDRs avec une fenêtre glissante de 15 acides aminés.

4 Nous sélectionnons avec ASTEROIDS cinq sous-ensembles de 200 conformations reproduisant les données cibles et extrayons alors une base de données de 1000 angles dièdres par résidus (5*200).

5 Nous générons alors un nouvel ensemble contenant 18000 conformations issues de la nouvelle base de données et 2000 issues du *random-coil*.

6 Nous réitérons alors les étapes 4 et 5 jusqu'à convergence du χ^2.

Lors de la sélection avec ASTEROIDS, l'erreur des déplacements chimiques vaut 2.0 pour ^{15}N, 0.04 pour $^{1}H^{N}$, 1.0 pour $^{13}C^{\alpha}$, $^{13}C^{\beta}$ et $^{13}C'$ et l'erreur des CDRs vaut 0.8 pour D_{NH} et $D_{C'H^N}$, 1.6 pour $D_{C^{\alpha}H^{\alpha}}$, 0.4 pour $D_{C'C^{\alpha}}$. 20000 itérations sont effectuées et 100 individus utilisés.

Le protocole de sélection reprend en grande partie celui présenté dans l'article de Jensen et al. [136] sur la partie C-terminale de la nucléoprotéine N du virus Sendaï. Un point essentiel diffère, afin de prendre en comptent les effets des résidus voisins, nous sélectionnons toujours des structures entières et non uniquement des résidus unique pendant plusieurs itérations puis des structures entières. La sélection de structures entières est computationnellement plus couteuse mais offre une meilleure reproduction de l'échantillonnage conformationnel.

5.1.5 Ajout d'un bruit gaussien

La difficulté essentielle consiste à caractériser le bruit de chaque observable :

- Pour les CDRs, il provient principalement de la mesure, nous choisissons par défaut un bruit gaussien dont la largeur à mi-hauteur vaut 0.5 Hz pour le couplage D_{NH}, l'amplitude du bruit est alors adaptée à la gamme des autres couplages c'est-à-dire 1.0 Hz pour $D_{C^{\alpha}H^{\alpha}}$, 0.25 Hz pour $D_{C'C^{\alpha}}$ et 0.5 Hz pour $D_{C'H^N}$.

- Pour les déplacements chimiques, l'erreur provient principalement des prédicteurs et non de la mesure expérimentale, nous utilisons 25% de l'erreur estimée sur la prédiction par les logiciels, cette erreur est spécifique à chaque déplacement

chimique : 0.22 ppm pour $^{13}C^{\alpha}$, 0.24 ppm pour $^{13}C^{\beta}$, 0.25 ppm pour $^{13}C'$, 0.6 ppm pour ^{15}N, 0.12 ppm pour $^{1}H^{N}$.

5.1.6 Données expérimentales

Les données de la protéine N_{tail} comprennent les déplacements chimiques : $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{N}$ mesurés à 25°C et les CDRs D_{NH} mesurés dans un cristal liquide composé de polyéthylène glycol et d'hexan-1-ol [132].

Les données de la protéine K18 comprennent les déplacements chimiques : $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{N}$ mesurés à 5°C [145] et les CDRs D_{NH} mesurés dans un gel polyacrylamide étiré [74]. La base de données de SPARTA étant essentiellement constituée de protéines attribuées entre 20° et 30°C, nous avons décalé avec un *offset* les déplacements chimiques de la construction K18 de manière à reproduire les déplacements de la protéine Tau mesurée à 25°C pour éviter toute inconsistance avec les logiciels de prédictions. Cet ajustement est crucial et n'implique pas de biais notable dans la description structurale de la protéine. Pour étayer ces propos nous pouvons nous référer à la publication de Kjaergaard et al. [146]. Les auteurs ont étudié la dépendance en température des déplacements chimiques afin d'identifier son origine. En effet lors de l'augmentation en température, la modification de la valeur des déplacements chimiques peut être due aux changements de l'échantillonnage conformationnel ou à une dépendance intrinsèque des déplacements chimiques vis-à-vis de la température. La réponse est vraisemblablement une combinaison des deux phénomènes.

Les auteurs ont donc cherché à quantifier ces effets en évaluant les changements structuraux à l'échelle atomique. Pour cela ils mesurent sur la protéine ACTR les déplacements chimiques $^{13}C^{\alpha}$ et $^{13}C'$ à 5°C et 45°C. Le référencement des déplacements chimiques est alors crucial pour quantifier la présence de structures secondaires. Les auteurs utilisent non pas les bases de référence standards mais mesurent en conséquence les déplacements chimiques de la même protéine à 5°C et 45°C dans l'urée. Au vu de ces résultats, les changements structuraux liés à l'augmentation de la température sont mineurs, la différence entre les valeurs des déplacements chimiques serait donc du à une dépendance intrinsèque en température. Ce résultat souligne l'importance de référencer les spectres.

La dépendance en température des déplacements chimiques est évidemment difficile à évaluer mais doit être prise en compte en référençant les spectres à température ambiante (20°C-25°C) pour offrir le maximum de précision dans la caractérisation de l'échantillonnage conformationnel des PIDs. Nous avons donc aligné notre spectre à 5°C sur ceux de Tau à 25°C, les modifications intervenant principalement sur les valeurs de $^{13}C^{\beta}$, ^{15}N, $^{1}H^{N}$.

5.2 RÉSULTATS

5.2.1 Dépendance structurale des déplacements chimiques

Un objectif de la RMN appliquée aux protéines désordonnées est la détermination d'un ensemble de structures représentant l'échantillonnage conformationnel du système [136, 138]. Bien que les déplacements chimiques soient aisément accessibles expérimentalement pour un bon nombre de protéines, une description par ensemble s'appuyant sur la reproduction de ces valeurs nécessite la prédiction des déplacements chimiques sur des structures créées *in-silico*. D'énormes progrès ont été réalisés dans ce domaine

au cours des 10 dernières années, de nombreux programmes combinant une approche empirique et s'appuyant sur de larges bases de données ont émergé tels que SPARTA, SPARTA+, SHIFTX2, CAMSHIFT (se référer à la section 5.1.1).

Afin de répondre a notre problématique, il nous est apparu essentiel de présenter les prédictions du logiciel SPARTA effectuées sur des structures désordonnées créées par le logiciel FLEXIBLE-MECCANO et d'évaluer le lien entre les valeurs des déplacements chimiques prédits et l'échantillonnage conformationnel correspondant. La méthode de calcul des déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{1}H^{N}$, ^{15}N et $^{1}H^{\alpha}$ est présentée en section 5.1.1.

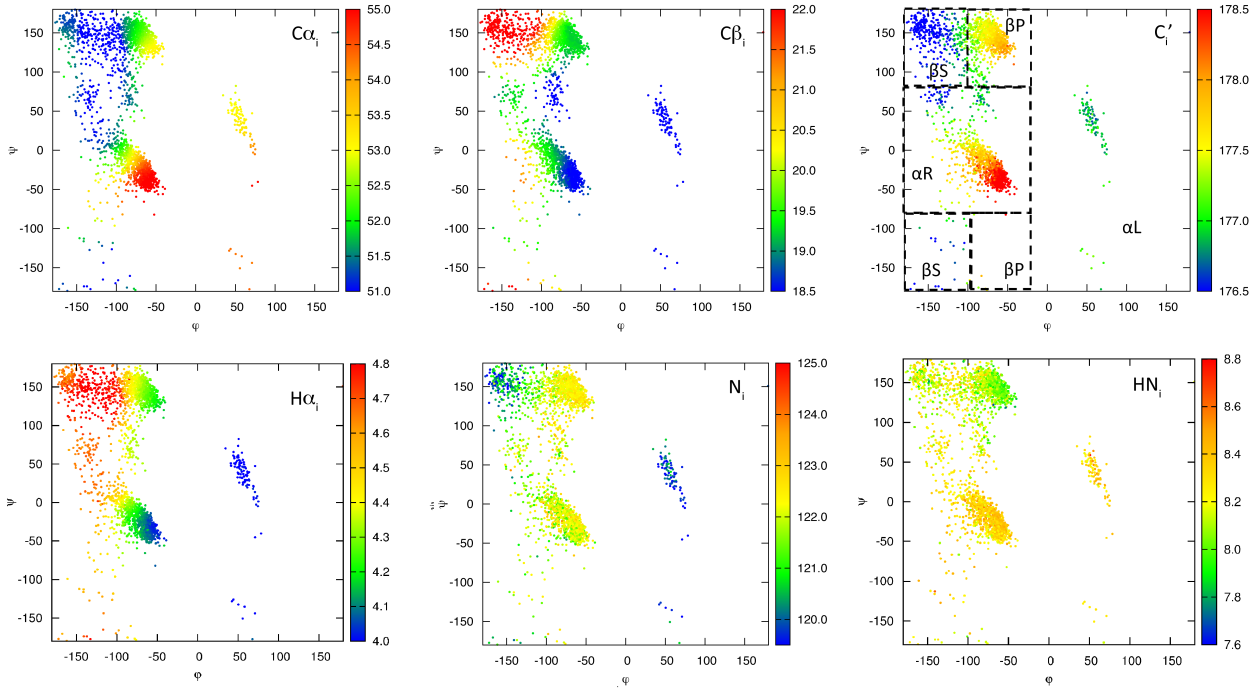


FIGURE 5.1 – ***Prédictions des déplacements chimiques du résidu central i en fonction de l'échantillonnage (ϕ,ψ) du résidu i.*** *La prédiction est réalisée avec le logiciel* SPARTA. *De gauche à droite, en haut :* $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, *en bas :* $^{1}H^{\alpha}$, ^{15}N, $^{1}H^{N}$. *Pour chaque cadre, nous avons en abscisse la valeur de ϕ, en ordonnée la valeur de ψ, le code de couleur est un gradient croissant des couleurs froides aux couleurs chaudes.*

La distinction entre les valeurs des déplacements chimiques est faite par un gradient de couleur, les valeurs maximales correspondant aux couleurs chaudes, les valeurs minimales aux couleurs froides. La problématique peut se résumer à la question suivante : pouvons-nous visuellement discriminer les différentes régions de l'espace Ramachandran en fonction des valeurs prédites par le logiciel SPARTA. Si la réponse est non, il en sera de même pour toute approche cherchant à reproduire les déplacements chimiques expérimentaux telle que la description par ensemble sous contraintes utilisant le χ^2 comme fonction pour guider la sélection. Nous commençons en figure 5.1 par étudier la valeur des déplacements chimiques du résidu central i en fonction des angles (ϕ,ψ) de ce même résidu :

- Le déplacement chimique $^{13}C^{\alpha}$ est bien différencié avec ces maximales dans la région αR et ces minimales dans la région βS. Nous notons une continuité entre la région βP et la région αR supérieure.

- Le déplacement chimique $^{13}C^{\beta}$ est le miroir du déplacement chimique $^{13}C^{\alpha}$ avec ces minimales dans la région αR, ces maximales dans la région βS. Nous obser-

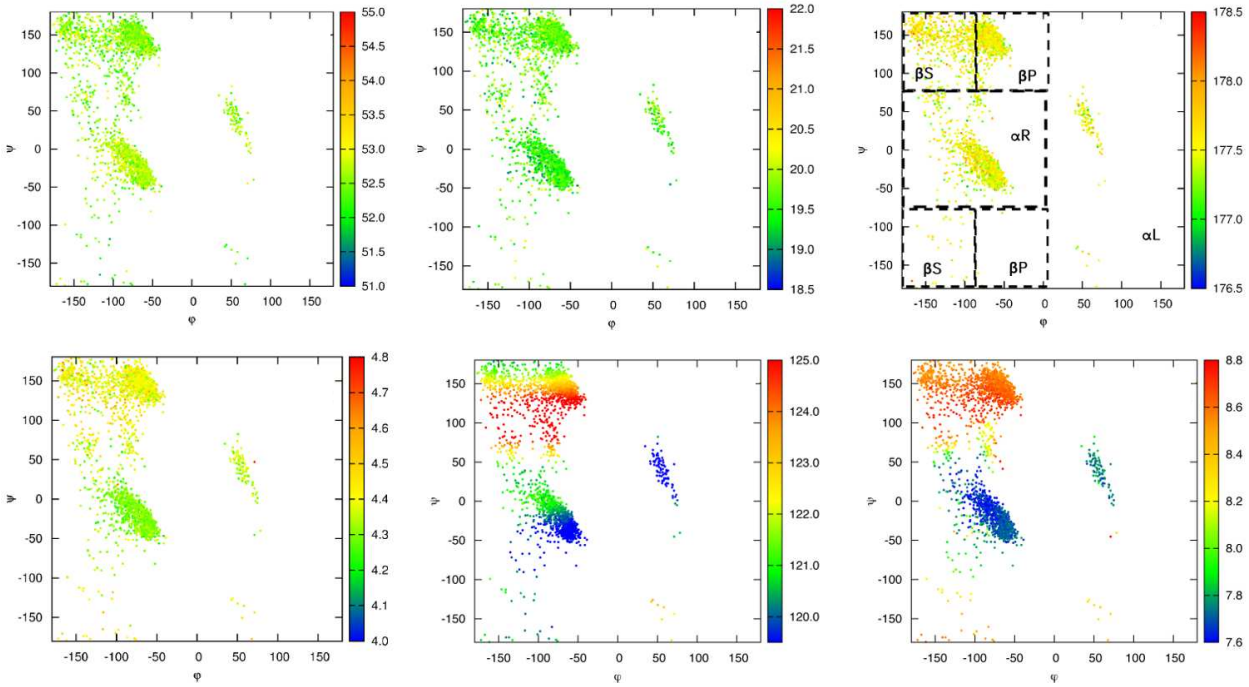


FIGURE 5.2 – ***Prédiction des déplacements chimiques du résidu voisin i+1 en fonction de l'échantillonnage (ϕ,ψ) du résidu i.*** *La prédiction est réalisée avec le logiciel* SPARTA. *De gauche à droite, en haut :* $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, *en bas :* $^{1}H^{\alpha}$, ^{15}N, $^{1}H^{N}$. *Pour chaque cadre, nous avons en abscisse la valeur de ϕ, en ordonnée la valeur de ψ, le code de couleur est un gradient croissant des couleurs froides aux couleurs chaudes.*

vons à nouveau une continuité entre la région βP et la région αR supérieure.

- Le déplacement chimique $^{1}H^{\alpha}$ à la même distribution de valeurs dans l'espace Ramachandran que les $^{13}C^{\beta}$.

- La distribution des déplacements chimiques $^{13}C'$ est similaire à celle de $^{13}C^{\alpha}$, cependant la région αL n'est pas dégénérée avec la région même région de l'espace Ramachandran.

- Les azotes ^{15}N et protons $^{1}H^{N}$ ne permettent pas très bien de différencier les régions de l'espace Ramachandran .

La figure 5.2 est la transposée du cas précédent avec une faible sensibilité des déplacements chimiques voisins (i+1) $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ et $^{1}H^{\alpha}$ et une forte sensibilité des ^{15}N et $^{1}H^{N}$ vis-à-vis des angles (ϕ,ψ) du résidu i. Nous distinguons une différence d'environ 1 ppm pour le proton $^{1}H^{N}$ et de 3 ppm pour l'azote ^{15}N entre les régions αR supérieure et βP ou βS.

5.2.2 Les déplacements chimiques cibles des simulations *in-silico*

L'analyse effectuée met en relief le lien entre la valeur des déplacements chimiques à l'échantillonnage conformationnel d'un tripeptide AAA. Une étude approfondie nécessiterait l'analyse de toutes les combinaisons d'acides aminés au sein d'un tripeptide. Cette solution bien que réalisable d'un point de vu calculatoire n'en reste pas moins peu pragmatique. Une description par ensemble semble alors un outil de choix pour étudier la complexité d'une chaine polypeptidique comme une protéine.

Afin de tester les capacités de notre approche à détecter un échantillonnage conformationnel spécifique, nous créons un système de référence appelé cible avec logiciel FLEXIBLE-MECCANO et calculons les paramètres RMN comme mentionné précédemment. La séquence en question, longue de 60 résidus, possède trois motifs transitoirement structurés et non coopératifs : une hélice α, un feuillet β et une hélice PPII.

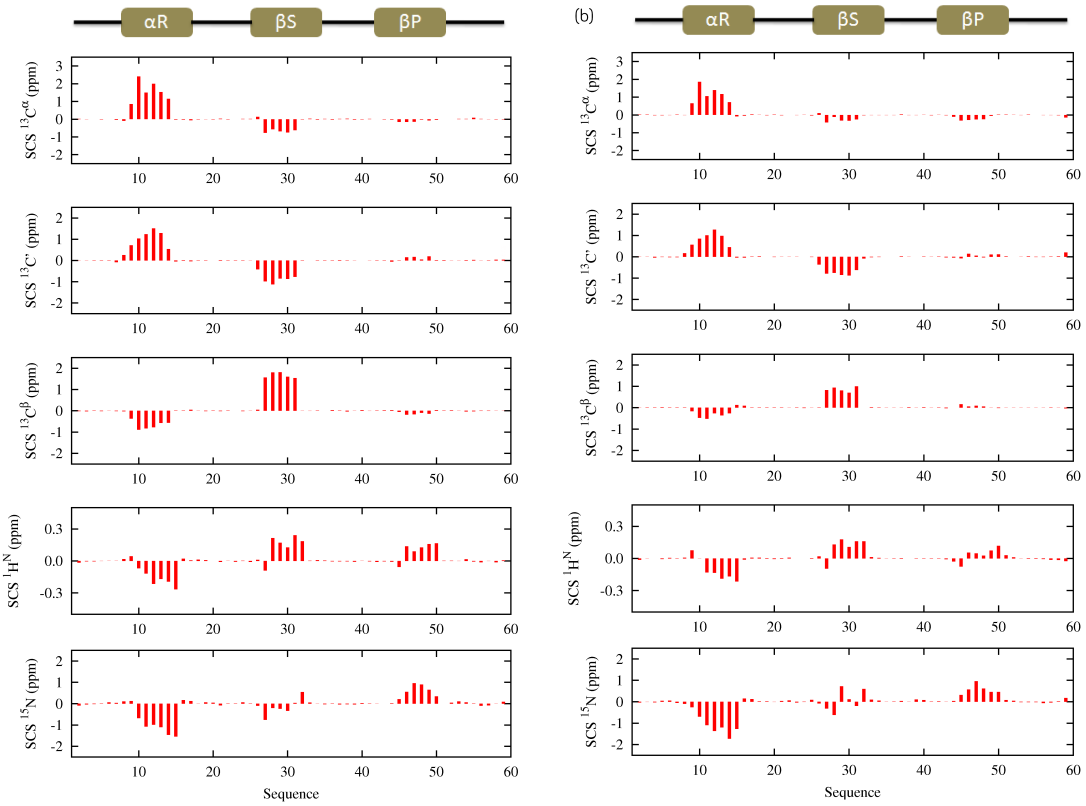


FIGURE 5.3 – ***Différence entre les déplacements chimiques secondaires cibles et ceux du random-coil.*** *Nous soustrayons les valeurs random-coil aux valeurs cibles afin d'identifier la signature des déplacements chimiques dans les régions transitoirement structurées. De haut en bas, nous avons :* $^{13}C^{\alpha}$, $^{13}C'$, $^{13}C^{\beta}$, ^{15}N *et* $^{1}H^{N}$. *À gauche, les prédictions sont calculées avec le logiciel* SPARTA, *et à droite avec le logiciel* SPARTA+.

La figure 5.3 montre la différence entre les valeurs issues de l'ensemble cible et celles issues d'un ensemble *random-coil* pour les déplacements chimiques suivant $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N et $^{1}H^{N}$. Nous retrouvons les résultats présentés précédemment c'est -à-dire une forte sensibilité des carbones pour les éléments de type hélice ou feuillet et une sensibilité quasi nulle pour les éléments échantillonnant la région Polyproline. Nous constatons malgré tout une sensibilité des protons et azotes pour cette dernière région.

Résumé

Concernant la région PPII, nous observons à partir des déplacements chimiques carbones deux dégénérescences lourdes de conséquences : d'une part nous pouvons difficilement différencier un motif PPII d'un motif *random-coil*, d'autre part, il existe une continuité des valeurs entre la région βP et la région αR supérieure. Toute méthode utilisant ces prédictions ne pourra pas caractériser correctement l'échantillonnage conformationnel de protéines échantillonnant la région PPII. L'ajout des ^{15}N et $^{1}H^{N}$ permet probablement de pallier ce problème mais nécessite de plus amples investigations.

Il apparait essentiel au vu des progrès réalisés dans la prédiction des déplacements chimiques de considérer l'ensemble des prédicteurs : Sparta, Sparta+, SHIFTX[3].. La figure 5.3b a été réalisée en utilisant le logiciel de prédiction Sparta+. Nous observons toujours une forte sensibilité des déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ pour l'hélice α et le feuillet β mais la sensibilité devient peu notable pour le motif PPII. Seuls les déplacements chimiques ^{15}N et $^{1}H^{N}$ sont distinctement marqués par cet échantillonnage.

Nous notons de subtiles différences entre Sparta et Sparta+. Concernant l'hélice α, la différence entre valeurs cibles et *random-coil* donne pour le logiciel Sparta+ des valeurs inférieures au logiciel Sparta, nous avons respectivement 0.5 ppm, 0.2 ppm, 0.2 ppm de moins pour les déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$. Il est en de même pour l'élément de type feuillet. Ces différences peuvent jouer un rôle non négligeable lors de la caractérisation des structures secondaires en influençant la propension des hélices obtenues. Les sélections selon des déplacements chimiques $^{13}C^{\alpha}$ à 1.5 ppm ou 1 ppm donneront des sous-ensembles avec une population d'angles dièdres différentes dans la région αR. Une manière de valider le résultat est d'utilisé les couplages dipolaires résiduels qui sont aussi sensibles à la propension des hélices. Cela permettrait de déterminer le logiciel le plus en accord avec les données expérimentales. D'autre part, nous constatons aussi un offset pour les valeurs du $^{13}C^{\beta}$ ainsi que des variations dans la prédiction des protons $^{1}H^{N}$.

5.2.3 Sélection d'ensembles avec Asteroids

Afin de tester les capacités de l'algorithme à reproduire les données et l'échantillonnage conformationnel, nous réalisons plusieurs tests *in-silico* en incluant un nombre variable de données, nous définissons donc les simulations suivantes :

- Simulation 1 : la sélection inclut les déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$
- Simulation 2 : la sélection inclut les déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{N}$

Pour chaque itération du protocole définit en section 5.1.4, nous regardons d'un part la reproduction des données incluses dans la sélection (en figure 5.4), des données exclues de la sélection et nous analysons la convergence du χ^2 et l'évolution des ensembles *pool* (données non montrées). Nous utiliserons deux représentations graphiques différentes pour vérifier la reproduction de l'échantillonnage conformationnel : la première montre quantitativement la reproduction des populations d'angles

3. Le programme ShiftX ne sera pas utilisé car la gamme des déplacements chimiques ^{15}N et $^{1}H^{N}$ n'est absolument pas en accord avec les données expérimentales.

dièdres des 4 régions de l'espace Ramachandran (βS, βP, αR, αL) pour chaque acide aminé (figure 5.5), la seconde montre "géographiquement" dans l'espace de Ramachandran les modifications de distribution d'angles (ϕ,ψ) induites par la sélection (figure 5.6).

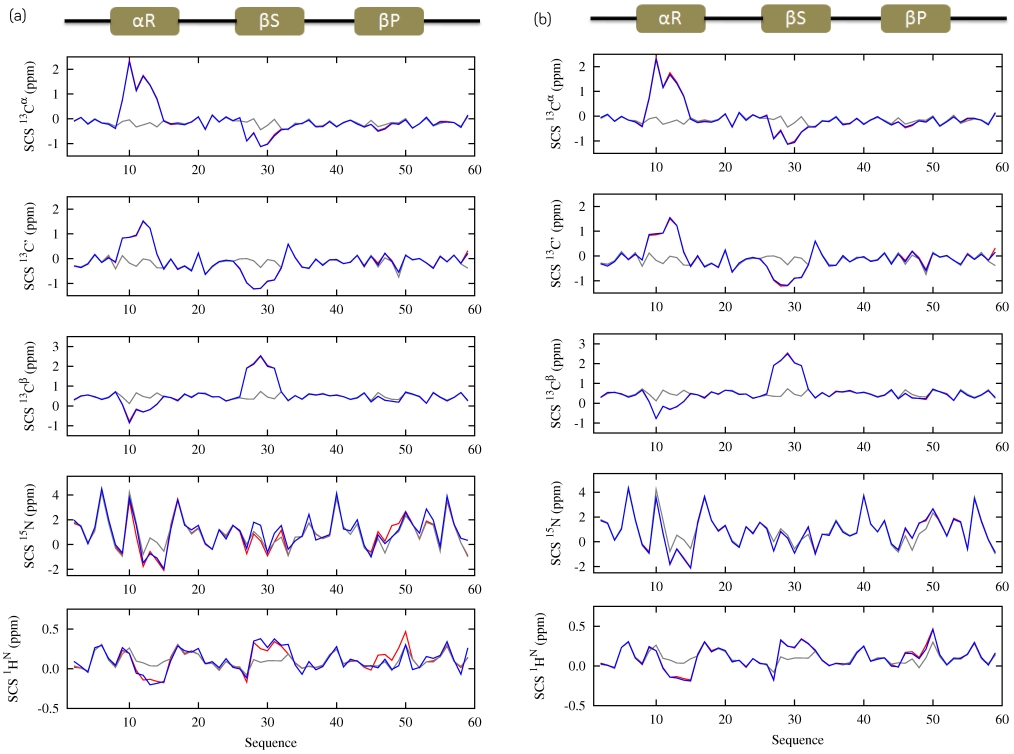


FIGURE 5.4 – ***Reproduction des déplacements chimiques cibles des simulations*** 1 ***et*** 2. *(a) : La simulation* 1 *affiche une parfaite reproduction des déplacements chimiques* $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ *inclus dans la sélection. (b) Idem pour la simulation* 2 *avec* $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ *et* ^{15}N, $^{1}H^{N}$. *Nous avons la cible en rouge, les données issues de la sélection avec* ASTEROIDS *en bleu et le random-coil en noir.*

Présentée en figure 5.4, la reproduction des données incluses dans la sélection avec ASTEROIDS est excellente pour les deux simulations. 5 itérations furent effectuées pour assurer la convergence et la bonne reproduction des données. Par ailleurs, nous souhaitons quantifier la reproduction de l'échantillonnage conformationnel, pour cela nous comparons pour chaque résidu les populations d'angles dièdres par régions de l'espace Ramachandran présentées en figure 4.10[4]. Les régions structurées en hélice α et feuillet β sont quantitativement échantillonnées (figure 5.5), l'erreur est inférieure à 5% sur l'ensemble des motifs. Cependant, la région structurée en hélice PPII est inégalement détectée : l'utilisation des déplacements chimiques proton $^{1}H^{N}$ et azote ^{15}N semble un pré-requis pour caractériser quantitativement la région PPII. Encore une fois, même sur des données *in-silico* une très bonne reproduction des données cibles ne permet pas de valider l'ensemble obtenu, l'observation des paramètres biophysiques est nécessaire pour statuer sur la qualité du protocole.

Nous obtenons un résultat quantitatif pour les fragments comprenant une hélice α et un feuillet β. La valeur moyenne du sous-ensemble sélectionné devant correspondre aux données cibles, elle ne permet pas d'ambigüité dans le choix des structures à l'algorithme pour ces deux motifs. En effet, elle doit correspondre aux valeurs des régions αR ou βS qui sont des extrémums comparé à la gamme des valeurs de chaque déplacement chimique (se référer à la figure 5.1). A l'opposé, les valeurs correspondantes à la région

4. La limite séparant la région βS de la région βP était auparavant ϕ=-90°, nous utiliserons maintenant ϕ=-100°.

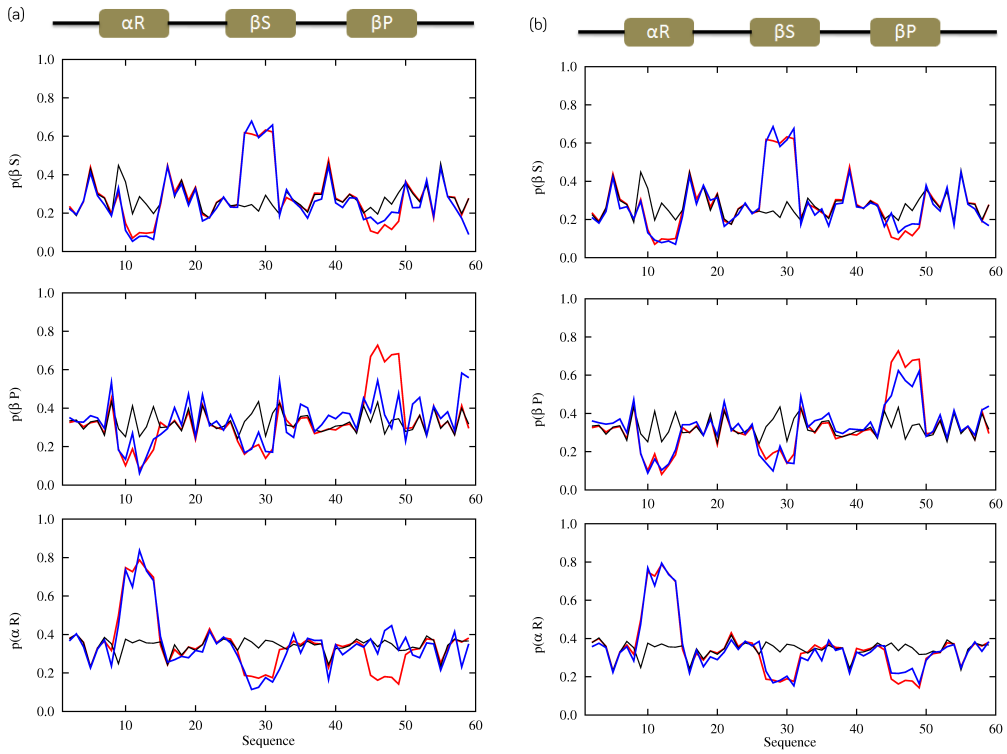


FIGURE 5.5 – ***Reproduction de l'échantillonnage conformationnel cible des simulations*** 1 ***et*** 2. *Les populations des régions de l'espace Ramachandran, (a) : simulation* 1. *(b) : simulation* 2. *Les cadres correspondent aux régions βS (en haut), βP (au milieu), αR (en bas) avec en rouge l'échantillonnage cible, en bleu l'échantillonnage issu de la sélection avec* ASTEROIDS, *en noir l'échantillonnage random-coil.*

βP sont médianes, il existe alors un nombre de combinaisons de structures bien plus important pour obtenir la valeur moyenne voulue. L'utilisation d'un nombre restreint de contraintes ne permet donc pas d'identifier correctement la distribution des angles dièdres de cette région.

Une façon alternative de valider la reproduction de l'échantillonnage conformationnel consiste à comparer dans l'espace de Ramachandran la distribution des angles dièdres pour différents ensembles. Nous allons détailler comment ont été tracés les cadres de la ligne du haut de la figure 5.6.

Nous extrayons les angles dièdres correspondant à chaque motifs structurés des structures de l'ensemble cible et de l'ensemble *random-coil*. Nous disposons alors de deux bases de données d'angles (ϕ,ψ). La solution la plus simple consiste à tracer ces points sur un graphique 2D avec ϕ en abscisse et ψ en ordonné. Pour traduire la distribution local d'angle dièdres, nous utilisons une fonction d'interpolation et passons d'une distribution discrète à une distribution continue. Un gradient de couleur permet alors de distinguer les différences de distribution d'angles dièdres. Souhaitant discriminer au mieux les modifications de l'échantillonnage conformationnel, nous normalisons la distribution de l'ensemble cible avec celle de l'ensemble *random-coil*. Nous obtenons alors les graphiques suivants (figure 5.6) :

- Cadre haut 1er à gauche : Le motif n'a pas de modification structurale, l'ensemble cible et l'ensemble *random-coil* possède la même distribution d'angle (ϕ,ψ) (en vert).

- Cadre haut 2e à gauche : Le motif αR échantillonne spécifiquement la région (-60°, -30°), l'ensemble cible n'a donc pas la même distribution d'angle (ϕ,ψ) que

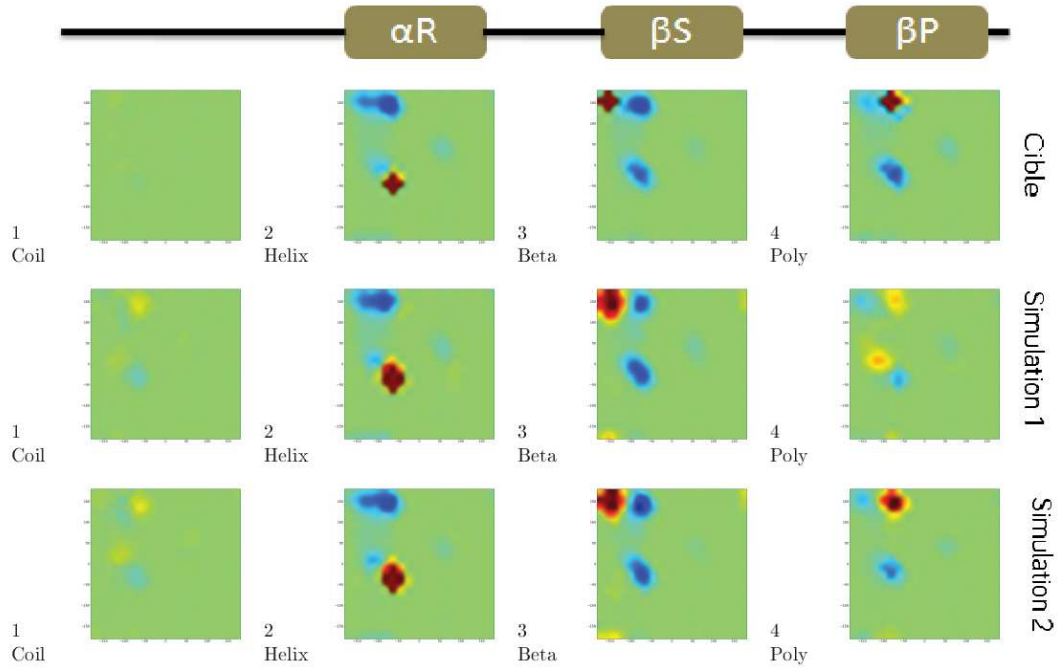


FIGURE 5.6 – ***Reproduction de l'échantillonnage conformationnel cible des simulations*** 1 ***et*** 2***.*** *De gauche à droite, nous distinguons la région non structurée, la région en hélice α, la région en feuillet β, la région en hélice PPII. (en haut) : échantillonnage cible. (au milieu) : échantillonnage après sélection des données* $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$. *(en bas) : échantillonnage après sélection des données* $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{N}$.

l'ensemble *random-coil* dans cette région. Par conséquent, la distribution d'angles dièdres augmente dans la région (-60˚, -30˚) (en rouge) et diminue ailleurs (en bleu).

- Cadre haut 2e à droite : Le motif βS échantillonne spécifiquement la région (-150˚, 150˚), l'ensemble cible n'a donc pas la même distribution d'angle (ϕ,ψ) que l'ensemble *random-coil* dans cette région. Par conséquent, la distribution d'angles dièdres augmente dans la région (-150˚, 150˚) (en rouge) et diminue ailleurs (en bleu).

- Cadre haut 1er à droite : Le motif βP échantillonne spécifiquement la région (-75˚, 150˚), l'ensemble cible n'a donc pas la même distribution d'angle (ϕ,ψ) que l'ensemble *random-coil* dans cette région. Par conséquent, la distribution d'angles dièdres augmente dans la région (-75˚, 150˚) (en rouge) et diminue ailleurs (en bleu).

Comparons maintenant les deux lignes suivantes, nous constatons comme attendu que les angles dièdres de la région *random-coil* ne sont pas modifiés, les angles dièdres du motif αR ont bien une propension plus importante dans cette région pour les deux simulations, il en est de même pour le feuillet β qui est correctement échantillonné. Pour le motif PPII, nous constatons une réponse différente suivant la simulation considérée, la simulation 1 indique deux augmentations de l'échantillonnage, une dans la région désirée, la région βP mais aussi une autre dans la région αR supérieure. La région αR inférieure correspondant à une hélice standard. Par ailleurs, la simulation 2 donne le résultat attendu. Ces tests *in-silico* correspondent aux prévisions effectuées en étudiant le lien entre les valeurs des déplacements chimiques et la distribution des angles dièdres d'un tripeptide AAA. Ils nous permettent de généraliser la portée de ces observations

et nous incitent à prendre en compte les valeurs des déplacements chimiques azote ^{15}N et proton $^{1}H^{N}$ pour lever la dégénérescence et caractériser correctement les PIDs.

Nous devons cependant apporter un bémol à cette conclusion, en effet les déplacements chimiques ^{15}N et $^{1}H^{N}$ sont particulièrement sensibles aux variations de pH et de température, de plus, en dépit des progrès importants réalisés ces dernières années, la prédiction de ces déplacements chimiques est soumise à de larges incertitudes, la plupart des logiciels affichent une erreur supérieure à 1.5 ppm pour l'azote et 0.5 ppm pour le proton. Dans ces conditions, il serait judicieux de ne pas les incorporer, pour cela nous allons répéter cette étude sur les CDRs.

5.2.4 La dépendance structurale des couplages dipolaires résiduels

Les couplages dipolaires appliqués aux PIDs ont fait l'objet de nombreuses études [78, 15, 147, 148, 87]. Nous allons dans cette section poursuivre notre travail de compréhension des couplages dipolaires résiduels et étudier la relation intrinsèque reliant leurs valeurs à l'échantillonnage conformationnel. Nous commençons par un cas simple, un pentadecapeptpide constitué d'Alanine.

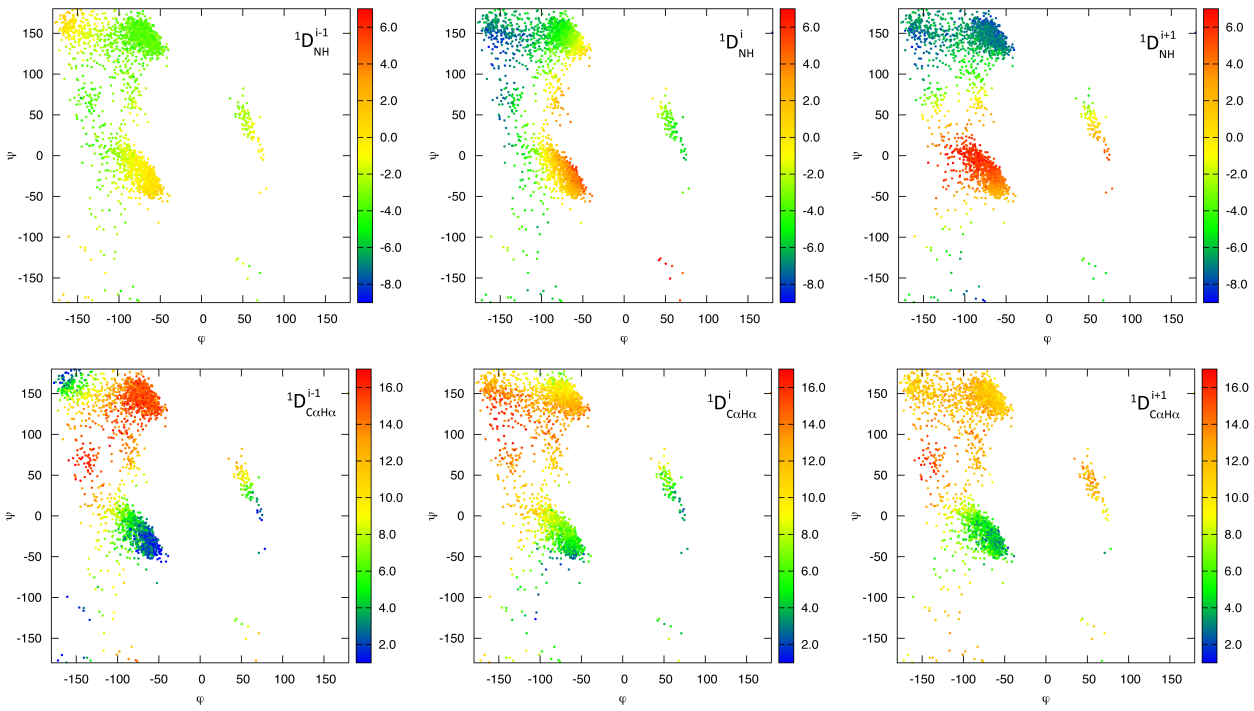


FIGURE 5.7 – ***Prédiction des couplages D_{NH} et $D_{C^{\alpha}H^{\alpha}}$ des résidus i-1, i, i+1, en fonction de l'échantillonnage conformationnel du résidu central i.*** *Les couplages sont moyennés dans une sphère d'un degré pour assurer la convergence des valeurs.*

La figure 5.7 présente les valeurs des couplages D_{NH} et $D_{C^{\alpha}H^{\alpha}}$ du résidu i en fonction des angles dièdres du résidu central i. Comme attendu, les CDRs D_{NH} sont positifs dans la région αR et négatifs dans la région étendue. Les valeurs sont similaires dans les régions βS et βP. De manière similaire, les CDRs $D_{C^{\alpha}H^{\alpha}}$ permettent de distinguer la région étendue de la région hélicoïdale αR mais ne permettent pas une distinction entre la région βS et la région βP. Un second point important est la sensibilité des couplages voisins (i+1 et i-1) à l'échantillonnage du résidu central i, le cas du couplage $D_{C^{\alpha}H^{\alpha}}$ du résidu i-1 est particulièrement frappant.

5.2.5 Mise en évidence de la longueur de persistance des CDRs

L'intérêt d'étudier la longueur de persistance des CDRs a été introduit en calculant des fenêtres glissantes de plusieurs tailles au chapitre 4, nous avons constaté qu'une fenêtre glissante de 3 résidus n'était pas suffisante pour retranscrire toute l'information des CDRs, *i.e.* la valeur du couplage du résidu i est sensible à l'échantillonnage conformationnel de voisins plus éloigné que le voisin direct. Pour déterminer cette longueur en résidu, nous créons un nouvel ensemble de structures n'ayant pas de contraintes stériques et dont les angles (ϕ,ψ) peuvent échantillonner toutes les valeurs comprises entre [-180 :180] [-180 :180] et calculons les CDRs. Les valeurs des couplages des 7 précédents et 7 suivants résidus du résidu central i, sont extraits en fonction de l'échantillonnage de ce dernier et présentés en figure 5.8A. Le résidu central est le numéro 8.

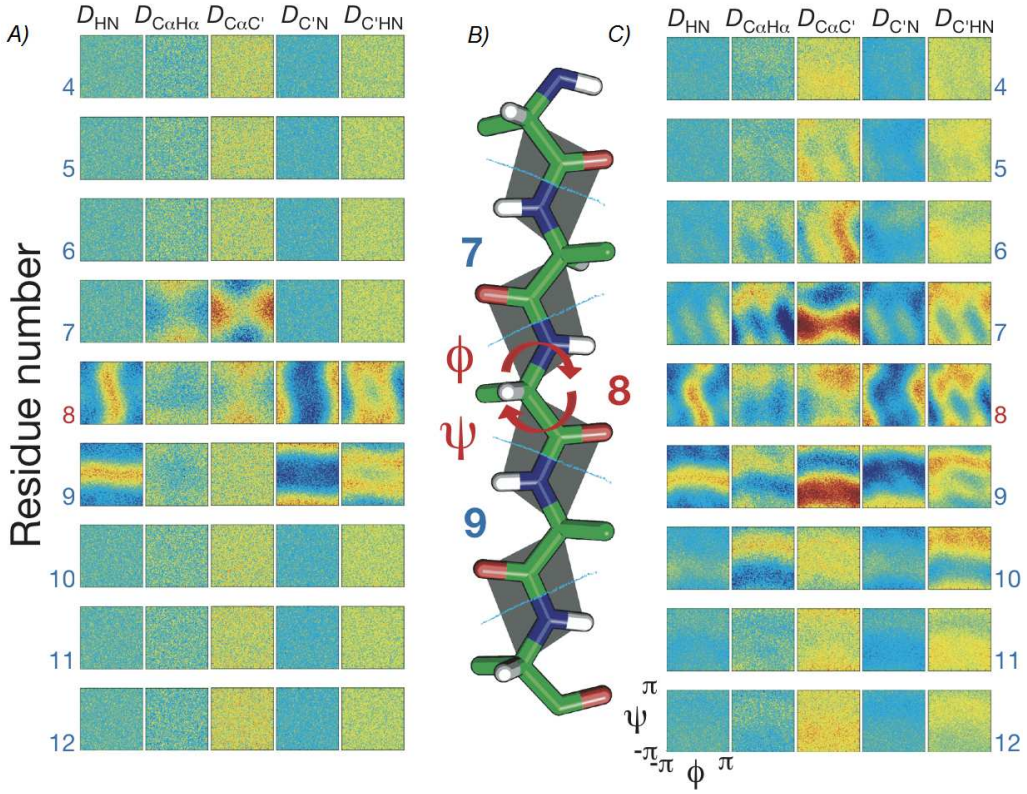


FIGURE 5.8 – ***Valeurs des couplages D_{NH}, $D_{C^\alpha H^\alpha}$, $D_{C'C^\alpha}$ pour les résidus i-7 à i+7, en fonction de l'échantillonnage conformationnel du résidu central i.*** *Les couplages sont moyennés pour assurer la convergence des valeurs. En A) Pas de restriction dans l'échantillonnage des résidus, en B) représentation schématique d'un plan peptidique, en C) Echantillonnage spécifique aux alanines pour tous les résidus excepté le résidu central.*

Au vu des profits obtenus, seuls le résidu central et ces voisins directs possèdent une dépendance vis-à-vis de l'échantillonnage du résidu central. Les différences entre les voisins i-1 (résidu 7) et i+1 (résidu 9) s'expliquent par les couplages considérés et leurs positions dans le plan peptidique par rapport aux angles dièdres du résidu central. Ainsi, le couplage résiduel D_{NH} du résidu central est particulièrement sensible aux variations de l'angle dièdre ϕ du résidu central i, le couplage résiduel D_{NH} du résident i-1 est cependant trop loin pour subir l'influence de l'échantillonnage du résidu central i tandis que le couplage résiduel D_{NH} du résidu i+1 est particulièrement sensible aux variations de l'angle dièdre ψ. Ces observations s'appliquent de manière analogue aux autres couplages. Pour autant, on observe un net désaccord avec la figure 5.7, notamment les valeurs du couplage $D_{C^\alpha H^\alpha}$ présentent une distribution forte surprenante : il semblerait que ce couplage soit indépendant de l'échantillonnage du résidu central i. Nous réitérons alors ce calcul en considérant cette fois-ci l'échantillonnage d'une ala-

nine pour chaque résidu, excepté le résidu central, le résultat est présenté en figure 5.8C.

Longueur de persistance des CDRs

La distribution angulaire de la chaine principale influence la valeur des CDRs et ceci même à longue distance. Nous pouvons affirmer en première approximation que l'échantillonnage des voisins de 1er et 2nd ordre contribuent majoritairement aux valeurs des CDRs. La longueur de persistance de l'information dipolaire varie suivant le couplage concerné.

Ce calcul met en évidence la relation complexe qui lie la valeur des couplages dipolaires résiduels à l'échantillonnage conformationnel des PIDS et nous incite à recourir à une description par ensemble afin de traiter la problématique posée.

5.2.6 Les CDRs cibles des simulations *in-silico*

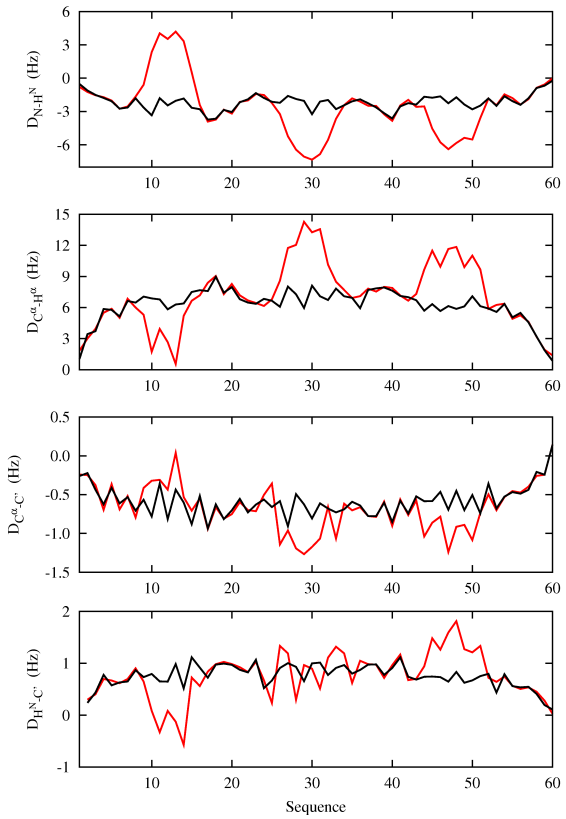


FIGURE 5.9 – ***Comparaison des CDRs cibles avec ceux du random-coil.*** *Les valeurs de l'ensemble cible sont en rouges et celles issues de l'ensemble random-coil en noir. Nous avons de haut en bas les CDRs suivant, : D_{NH}, $D_{C^\alpha H^\alpha}$, $D_{C'C^\alpha}$ et $D_{C'H^N}$.*

Calculant les CDRs sur notre ensemble cible, nous comparons en figure 5.9 le résultat obtenu aux CDRs d'un ensemble *random-coil*, l'identification des motifs structurés est immédiate. Il est cependant difficile de différencier un élément échantillonnant la région βS d'un élément échantillonnant la région βP en utilisant les couplages D_{NH}, $D_{C^\alpha H^\alpha}$, $D_{C'C^\alpha}$, seul le couplage $D_{C'H^N}$ semble en mesure de discriminer ces deux régions.

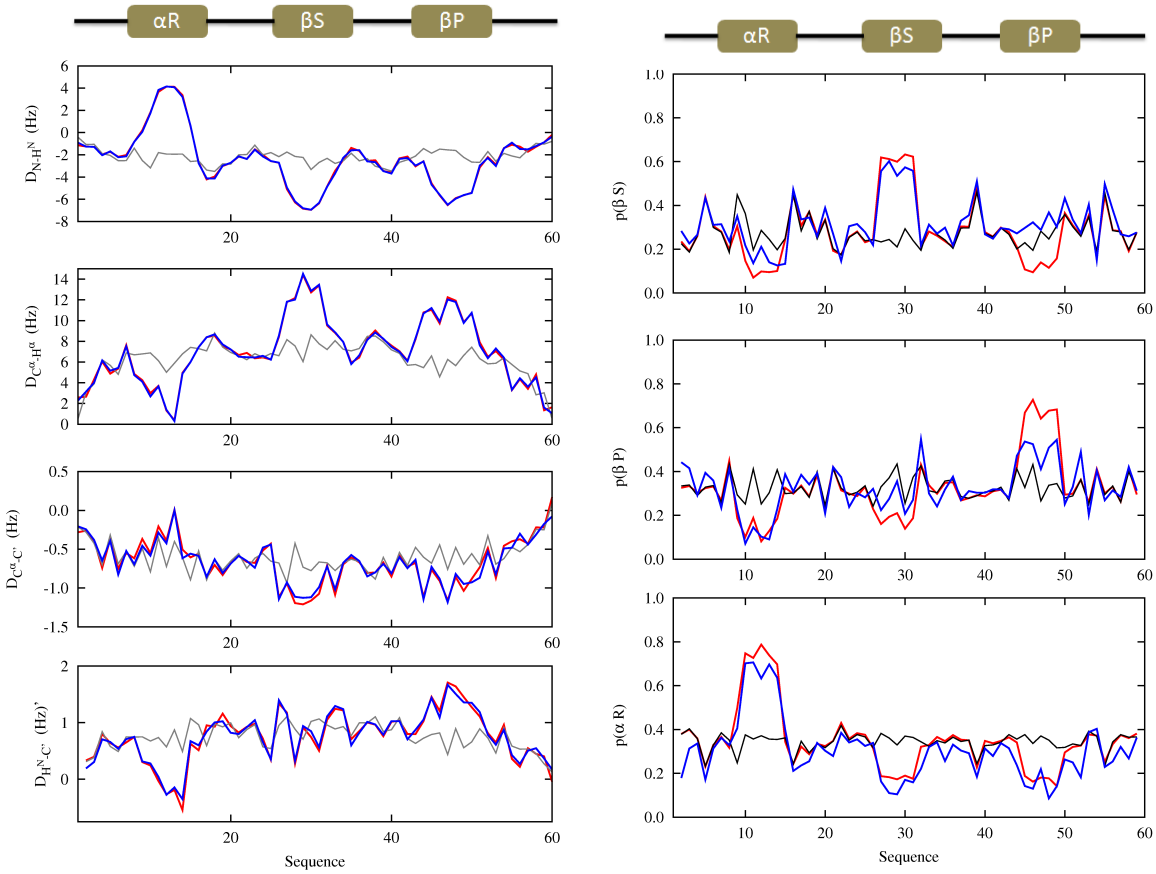


FIGURE 5.10 – ***Reproduction des 4 CDRs et de l'échantillonnage cible*** *À gauche, nous avons les CDRs suivant de haut en bas : D_{NH}, $D_{C^\alpha H^\alpha}$, $D_{C'C^\alpha}$ et $D_{C'H^N}$. À droite, les populations d'angles dièdres correspondant aux trois régions p(βS),p(βP),p(αR). Les données cibles sont en rouge, les données issues de la sélection avec* ASTEROIDS *en bleu et le random-coil en noir.*

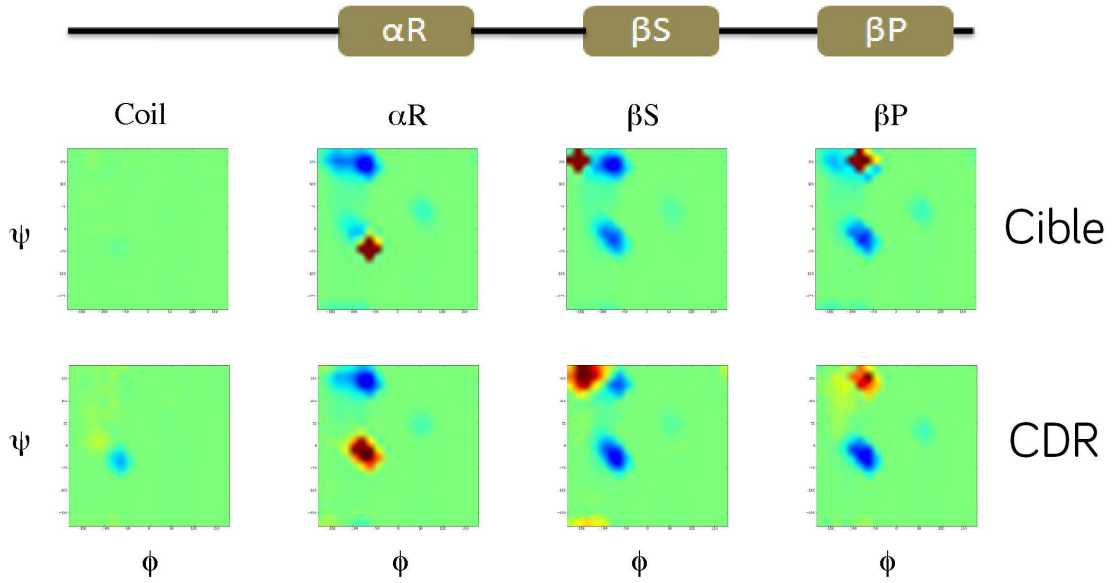


FIGURE 5.11 – ***Reproduction de l'échantillonnage conformationnel cible après sélections selon les CDRs suivants : D_{NH}, $D_{C^\alpha H^\alpha}$, $D_{C'C^\alpha}$ et $D_{C'H^N}$.*** *Les cadres sont tracés comme précédemment, ils comparent la distribution des angles (ϕ,ψ) dans l'espace Ramachandran de l'ensemble cible (en haut) et de l'ensemble à l'issu de la sélection (en bas). L'abscisse correspond à ϕ, l'ordonnée correspond à ψ. Dans chaque cas, la distribution d'angles dièdres est normalisée par rapport à celle équivalente d'un ensemble random-coil.*

5.2.7 Sélection d'ensembles avec Asteroids

Nous réalisons donc des *tests-in silico* afin de généraliser les observations mentionnées. Nous appliquons le protocole présenté en section 5.1.4 avec comme cible les 4 CDRs suivant : D_{NH}, $D_{C^\alpha H^\alpha}$, $D_{C'C^\alpha}$ et $D_{C'H^N}$. La reproduction des données en figure 5.10 est très bonne, cependant nous notons évidemment une dégénérescence entre la région βS et la région βP mais aussi une dégénérescence entre la région αR inférieure et la région αR supérieure (se référer à la figure 5.11 sur le cadre du motif αR en bas, la tache rouge est nettement plus étendue que la cible). Nous notons aussi une diminution de l'échantillonnage de la région αR en faveur des autres régions. Cette information se visualise sur la figure 5.10B sur cadre du bas. L'échantillonnage de la région Polyproline se révèle nettement plus étendu que voulu, nous constatons sur le cadre du motif βP une augmentation de l'échantillonnage dans toute la région étendue.

5.2.8 Conclusion partielle

Les travaux présentés mettant en relation la distribution des angles dièdres des structures et les valeurs des paramètres RMN ont permis de mettre en évidence la présence d'une dégénérescence intrinsèque au sein des déplacements chimiques et des CDRs. L'utilisation d'une description par ensemble couplée à une sélection avec Asteroids confirme clairement ces résultats. Nous avons une bonne reproduction des valeurs cibles mais ne s'accompagnant pas forcément d'une reproduction de l'échantillonnage cible au vu des dégénérescences exposées : les déplacements chimiques possèdent une dégénérescence entre la région βP et la région αR supérieure qui peut cependant être levée par l'utilisation des déplacements chimiques ^{15}N et $^{1}H^N$. Les CDRs possèdent une dégénérescence à la fois au sein de la région étendue et au sein de la région hélicoïdale.

D'autres points plus proches des contraintes expérimentales méritent d'être soulevés. Concernant les CDRs, les expériences les plus couramment effectuées aboutissent à la mesure des couplages suivants : D_{NH}, $D_{C^\alpha H^\alpha}$, $D_{C'C^\alpha}$ et $D_{C'H^N}$. L'incorporation de ces données reste tout de même délicate notamment pour les couplages $D_{C'C^\alpha}$ dont la gamme de mesure est très faible. Concernant les déplacements chimiques, l'erreur expérimentale est négligible. Malgré tout, la dépendance en température ou en pH et les potentiels erreurs de référencement des spectres doivent être attentivement pris en compte avant toute sélection. De plus, l'incertitude des logiciels de prédictions est un point critique intervenant dans toute procédure de calcul impliquant des déplacements chimiques. Une question essentielle subsiste : l'erreur de prédiction est-elle moyennée vers zéro ou au contraire amplifiée du fait du nombre de structures présentes au sein d'un ensemble. Il convient alors de proposer une méthode combinant à la fois les CDRs et les déplacements chimiques permettant la reproduction des caractéristiques biophysiques de l'ensemble et prenant en compte l'influence potentielle des contraintes expérimentales exposées sur l'échantillonnage conformationnel obtenu.

5.2.9 Combinaison des déplacements chimiques et couplages dipolaires résiduels

Il apparait essentiel de comprendre dans quelle mesure, la combinaison des déplacements chimiques et des couplages dipolaires résiduels permet ou non de caractériser le paysage énergétique des protéines désordonnées. Il s'agit donc dans un premier temps de déterminer quelle combinaison de données permet de retrouver le bon échantillonnage conformationnel puis dans un deuxième temps de déterminer le nombre minimal de jeux de données pouvant reproduire ce résultat.

Nous présenterons en parallèle les tests *in-silico* effectués avec les deux prédicteurs

suivants : SPARTA et SPARTA+. De nouveau, différentes simulations sont envisagées, nous nous contenterons de présenter les suivantes :

- Simulation 1 : la sélection inclut les déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$
- Simulation 2 : la sélection inclut les déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{N}$
- Simulation 3 : la sélection inclut les couplages dipolaires D_{NH}, $D_{C^{\alpha}H^{\alpha}}$, $D_{C'C^{\alpha}}$ et $D_{C'H^{N}}$
- Simulation 4 : la sélection inclut $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ + D_{NH}, $D_{C^{\alpha}H^{\alpha}}$, $D_{C'C^{\alpha}}$ et $D_{C'H^{N}}$
- Simulation 5 : la sélection inclut $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ + D_{NH}, $D_{C'H^{N}}$
- Simulation 6 : la sélection inclut $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ + D_{NH}
- Simulation 7 : la sélection inclut $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{N}$ + D_{NH}, $D_{C^{\alpha}H^{\alpha}}$, $D_{C'C^{\alpha}}$ et $D_{C'H^{N}}$

Predicteur :	SPARTA			SPARTA+		
Motif :	βS	αR	βP	βS	αR	βP
pool random-coil	0.45	0.45	0.40	0.45	0.45	0.40
Simulation 1	0.0645	0.0707	0.3520	0.0874	0.0795	0.2056
Simulation 2	0.0624	0.0783	0.0804	0.037	0.0775	0.1877
Simulation 3	0.1196	0.1139	0.2718	0.1196	0.1139	0 ;2718
Simulation 4	0.0707	0.0546	0.0621	0.0688	0.0888	à.355
Simulation 5	0.05410	0.0582	0.0786	0.0754	0.0729	0.0885
Simulation 6	0.0624	0.0783	0.0804	0.642	0.778	0.1604
Simulation 7	0.0587	0.0386	0.0629	nc	nc	nc

TABLE 5.1 – *Ecart quadratique moyen traduisant la reproduction de l'échantillonnage cible des sélections en fonction des paramètres RMN utilisés. À gauche : le calcul des déplacements chimiques est effectué avec* SPARTA*, à droite avec* SPARTA +*. Seul l'échantillonnage des régions transitoirement structurées est analysé.*

La table 5.1 récapitule les simulations effectuées et l'écart quadratique moyen (EQM)[5] entre l'échantillonnage cible et l'échantillonnage obtenu après sélection pour chaque région comprenant une structure secondaire.

La figure 5.12 montre la reproduction de l'échantillonnage conformationnel de la sélection simulation 4 et 6. Les déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ combinés à un ou quatre couplages dipolaires permettent de caractériser correctement *in-silico* la protéine. Il semblerait donc que les déplacements chimiques ^{15}N et $^{1}H^{N}$, difficiles à prédire, ne soient pas nécessaires.

Pour les résultats obtenus avec le logiciel SPARTA+, la reproduction de la région hélicoïdale et du feuillet est similaire mais nous échantillonnons moins quantitativement la région βP, et ceci, quel que soit le protocole utilisé (5.1). L'EQM vaut 0.15 dans la majorité des cas. La reproduction des données reste excellente pour toutes les simulations effectuées.

5. Il sera parfois noté RMSD pour root-mean-square deviation.

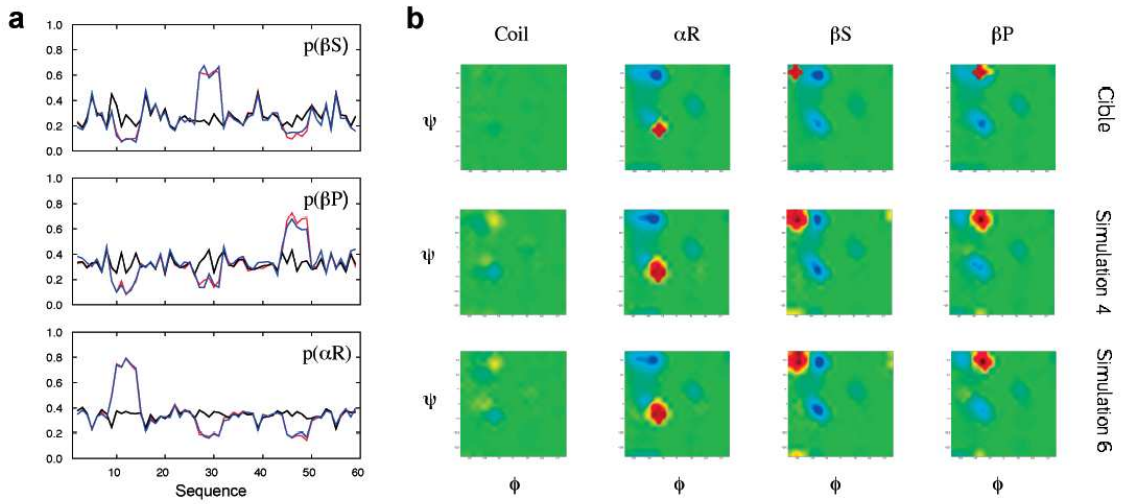


FIGURE 5.12 – ***Reproduction de l'échantillonnage conformationnel cible en combinant CDRs et déplacements chimiques.*** *À gauche, les cadres correspondent au région de l'espace Ramachandran avec de haut en bas : βS, βP, αR. En rouge : échantillonnage cible, en bleu : échantillonnage issu de la sélection avec* ASTEROIDS, *en noir : échantillonnage issu d'un ensemble random-coil. À droite, nous affichons la même information dans l'espace de Ramachandran uniquement sur les motifs structurés, nous avons en haut l'échantillonnage cible, en bas les échantillonnages issus de la sélection. L'abscisse correspond à ϕ, l'ordonnée correspond à ψ. Nous utilisons un gradient de couleur traduisant la densité des angles (ϕ,ψ).*

Predicteur :	SPARTA			SPARTA+		
Motif :	αR	βS	βP	αR	βS	βP
pool random-coil	0.45	0.45	0.40	0.45	0.45	0.40
Simulation 1 Δ	0.115	0.132	0.405	0.20	0.188	0.34
Simulation 2 Δ	0.180	0.172	0.270	0.135	0.183	0.31
Simulation 4 Δ	0.096	0.130	0.147	0.149	0.129	0.21
Simulation 6 Δ	0.126	0.128	0.187	0.148	0.142	0.237

TABLE 5.2 – ***Ecart quadratique moyen de l'échantillonnage conformationnel des sélections réalisées en présence de bruit gaussien.*** *À gauche : le calcul des déplacements chimiques est effectué avec* SPARTA, *à droite avec* SPARTA +. *Seul l'échantillonnage des régions transitoirement structurées est analysé.*

5.2.10 Ajout d'un bruit gaussien

Nous voulons évaluer la capacité de l'algorithme à reproduire l'échantillonnage conformationnel en tenant compte de contraintes ou incertitudes diverses. Nous introduisons alors un bruit gaussien sur les données cibles et répétons les sélections précédentes. Nous pouvons ainsi estimer la robustesse des protocoles avant de les appliquer aux données expérimentales.

La table **??** récapitule l'EQM entre l'échantillonnage cible et l'échantillonnage obtenu après sélection. Le bruit gaussien étant plus élevé pour les valeurs des déplacements chimiques ^{15}N et $^{1}H^{N}$, nous avons un EQM élevé pour la simulation 2 utilisant les 5 déplacements chimiques. L'ajout des couplages dipolaires résiduels améliore ce résultat en compensant les variations de ces déplacements chimiques et permet d'améliorer la reproduction de l'échantillonnage. Pour le motif Polyproline, en ajoutant les 4 CDRs l'EQM passe de 0.4052 à 0.1473. De manière générale, si nous considérons l'ensemble de la séquence (résultat non montré), l'utilisation des couplages dipolaires résiduels combinés aux déplacements chimiques diminue les fluctuations liées au bruit gaussien. Considérant les potentielles erreurs dues à la prédiction des déplacements chimiques, il est donc recommandé d'inclure le couplage dipolaire D_{NH} dans notre protocole pour

obtenir une précision inférieure à 10% en présence de bruit. Par conséquent, considérant l'ensemble des simulations réalisées *in-silico*, le jeu de données minimal pour caractériser l'échantillonnage conformationnel des protéines désordonnées nécessite les déplacements chimiques C^{α}, $^{13}C^{\beta}$, $^{13}C'$ et les CDRs D_{NH}. Nous allons maintenant appliquer cette hypothèse sur des données expérimentales.

5.3 Applications aux données expérimentales

Nous allons maintenant appliquer le protocole à deux jeux de données expérimentaux, l'un étant la construction de Tau : K18, l'autre la région C-terminale N_{tail} de la nucléoprotéine N de la rougeole. L'utilisation de données expérimentales rend la validation du protocole et l'interprétation des résultats plus compliquée. Nous tacherons dans un premier temps de vérifier le bon fonctionnement du protocole puis nous cherchons à identifier le meilleur logiciel de prédiction, c'est-à-dire celui prédisant les données RMN le plus en adéquation avec les données expérimentales. Par ailleurs, de nombreuses validations croisées seront réalisées pour tester la consistance de l'approche. Nous mettrons en évidence les questions clés à résoudre pour obtenir sans ambigüité une description par ensemble des protéines intrinsèquement désordonnées.

5.3.1 K18

La protéine K18 est une construction de la protéine Tau, elle correspond au domaine d'appariement de la protéine avec les microtubules. Elle comprend 130 acides aminés allant du numéro 243 à 372, la numérotation étant celle de l'isoforme le plus long.

Nous effectuons alors les deux simulations suivantes avec les logiciels Sparta ou Sparta+ :

- simulation 1 : la sélection inclut $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{N}$
- simulation 2 : la sélection inclut $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{N}$ + D_{NH}

Nous commençons par inclure toutes les données pour évaluer leur reproduction. Au vu des résultats affichés en figure 5.13, la reproduction est très bonne. Nous pouvons ainsi analyser l'échantillonnage conformationnel obtenu avec la présence de fragments échantillonnant la région PPII et de boucles définissant les répétitions R1, R2, R3, R4.

Echantillonnage conformationnel de K18

L'échantillonnage conformationnel obtenu est atypique, nous observons une diminution globale de la région βS au profit de la région βP ou αR. Les régions ayant été identifiées préalablement par les études précédentes comme étant des feuillets β échantillonne majoritairement la région βP. Cette méthode semble la première pouvant différencier distinctement l'échantillonnage conformationnel des protéines désordonnées au sein de de la région étendue.

Nous effectuons alors la simulation 1, validation croisée des couplages dipolaires résiduels D_{NH}. Après détermination d'un facteur multiplicatif adapté, la reproduction des CDRs D_{NH} en haut de la figure 5.14 est très bonne avec notamment la visualisation

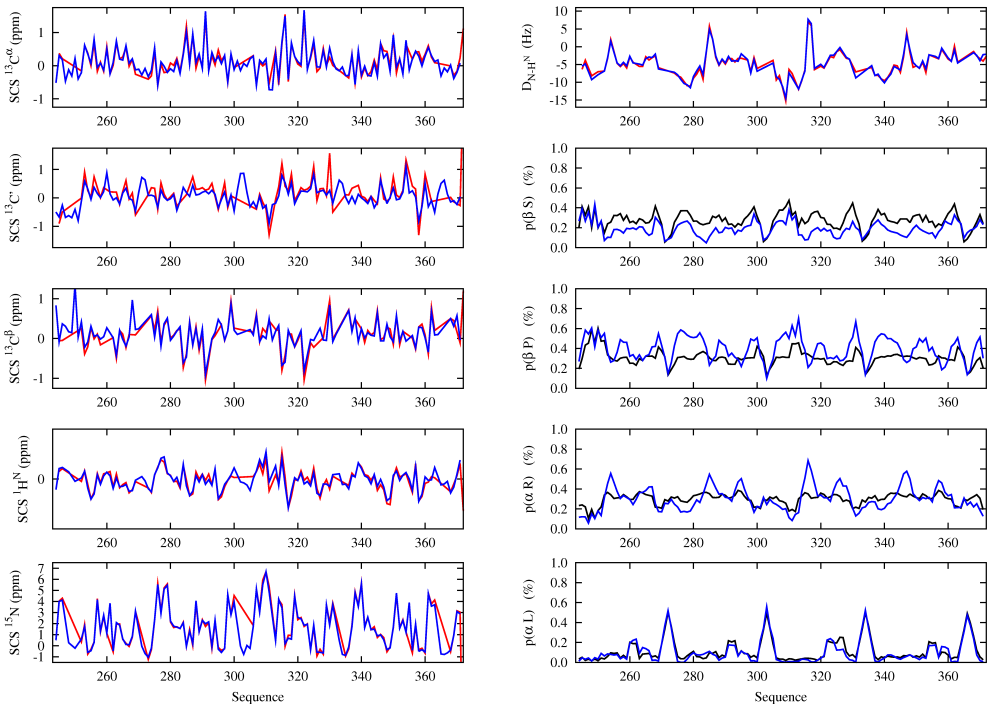


FIGURE 5.13 – ***Déplacements déplacements chimiques et échantillonnage conformationnel issu de la simulation 2.*** *Nous avons de haut en bas $^{13}C^{\alpha}$, $^{13}C'$, $^{13}C^{\beta}$, $^{1}H^{N}$, ^{15}N et à droite les CDRs D_{NH}. Les données expérimentales sont en rouge, les données issues de la sélection en bleu. Les populations correspondent aux 4 régions de l'espace Ramachandran préalablement définies. Ces populations sont moyennées sur 3 résidus afin de gommer les spécificités de chaque acide aminé et ainsi mieux observer les différences par rapport aux populations du random-coil tracées en noir.*

des couplages positifs au niveau des boucles et des couplages plus négatifs au niveau des régions étendues Polyproline.

D'autre part, l'échantillonnage issu des logiciels SPARTA et SPARTA+ est similaire quelle que soit la simulation effectuée. Nous présentons l'échantillonnage obtenu pour le protocole numéro 2 en figure 5.14. Nous avons aussi effectué ce calcul pour la simulation 1 et obtenu une réponse similaire. La reproduction des données à chaque itération est cependant meilleure avec le logiciel SPARTA. Le cas de la simulation 2 est affiché en figure 5.14 à droite.

5.3.2 N_{tail}

La partie C-terminale N_{tail} de la nucléoprotéine N du virus de la Rougeole est intrinsèquement désordonnée, elle a été préalablement étudiée par Jensen et al. [132] en combinant RMN, SAXS et microscopie électronique. Les auteurs caractérisent le domaine N_{tail} isolé et sa transition conformationnelle lors de la liaison avec son partenaire physiologique.

La comparaison des données sans sélection en figure 4.2 montre que la protéine est essentiellement *random-coil* à l'exception faite de la région [485-500] échantillonnant des hélices transitoires. Nous effectuons plusieurs simulations comprenant les données suivantes, dans chaque nous comparons les résultats obtenus avec SPARTA et SPARTA+ :

- simulation 1 : la sélection inclut $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$

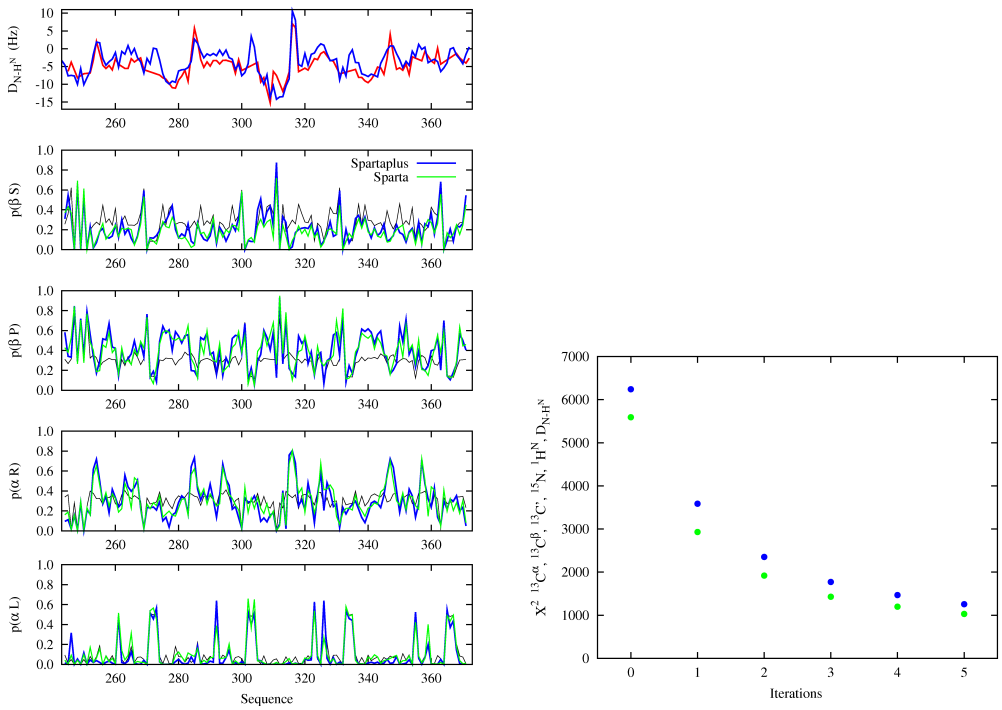


FIGURE 5.14 – ***Validation croisée des CDRs D_{NH}, et différences entre* Sparta *et* Sparta+.** *À gauche en haut : validation croisée des CDRs D_{NH} en bleu, données expérimentales en rouge. Population correspondant à la simulation* 2 *effectuées parallèlement soit avec* SPARTA *(en vert), soit avec* SPARTA+ *(en bleu) . À droite : χ^2_{CDR} correspondant à la simulation* 2 *pour chaque itération du protocole.*

- simulation 2 : la sélection inclut $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{N}$
- simulation 3 : la sélection inclut $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{N}$ + D_{NH}
- simulation 4 : la sélection inclut $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N + D_{NH}
- simulation 5 : la sélection inclut $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{1}H^{N}$ + D_{NH}

La reproduction des données présentées en figure 5.15 à gauche est très bonne. Nous notons une diminution globale de la région βS au profit de la région βP. Le cadre p(αR) permet de visualiser la présence de plusieurs hélices de longueurs différentes, nous retrouvons au moins deux des hélices préalablement identifiées présentées en figure 4.7.

Nous souhaitons de nouveau déterminer la robustesse des simulations, nous effectuons quatre validations croisées indépendantes présentées en figure 5.16. Les données des simulations 1, 2, 4, 5 non incluses dans la sélection sont bien reproduites. Un point délicat est la prise en compte des phénomènes de coopérativité, les déplacements chimiques ne sont pas adaptés à ce problème pourtant les profils obtenus des couplages D_{NH} se calquent sur les données expérimentales (cadres du haut à droite de la 5.16).

La validation croisée des déplacements chimiques azote et proton est bonne, nous notons un léger décalage pour l'azote vers les bas, indiquant qu'il est difficile de retrouver l'information contenu dans ces derniers dans d'autres paramètres mais nous sommes bien en mesure de reproduire les données passives quelque soit la simulation effectuée. D'autre part, l'échantillonnage obtenu avec SPARTA+ est similaire à celui de SPARTA, cette comparaison est effectuée pour toutes les simulations , nous affichons le cas de la simulation 3 en figure 5.16 à gauche.

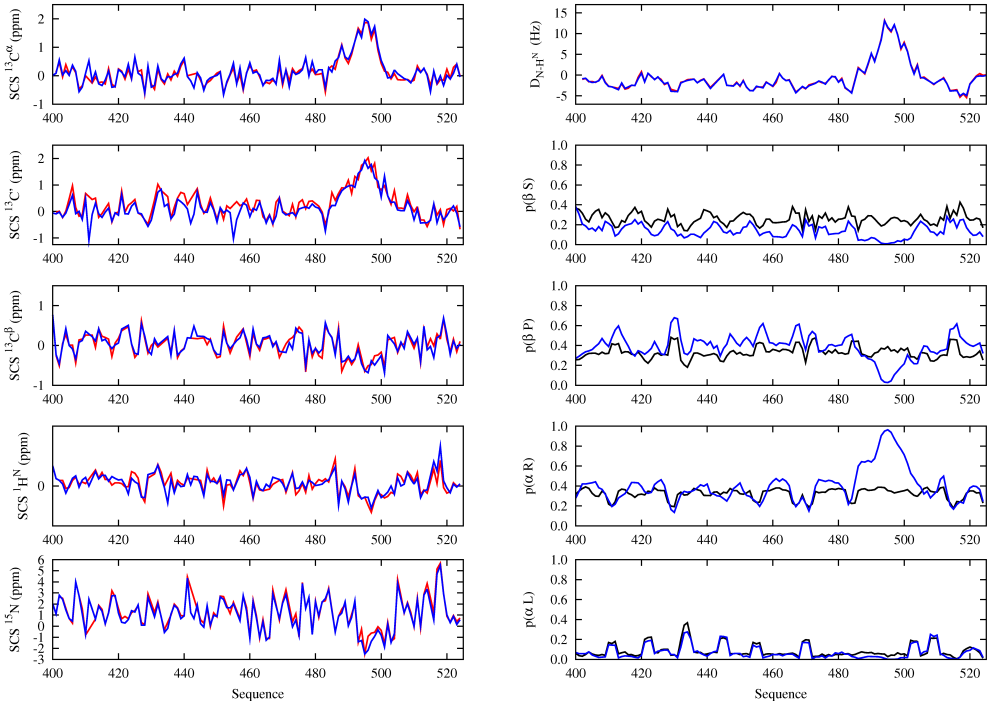


FIGURE 5.15 – ***Déplacements déplacements chimiques et échantillonnage conformationnel issu de la simulation*** 3. *Nous avons de haut en bas $^{13}C^{\alpha}$, $^{13}C'$, $^{13}C^{\beta}$, $^{1}H^{N}$, ^{15}N et à droite les CDRs D_{NH}. Les données expérimentales sont en rouge, les données issues de la sélection en bleu. Les populations correspondent aux* 4 *régions de l'espace Ramachandran préalablement définies. Ces populations sont moyennées sur 3 résidus.*

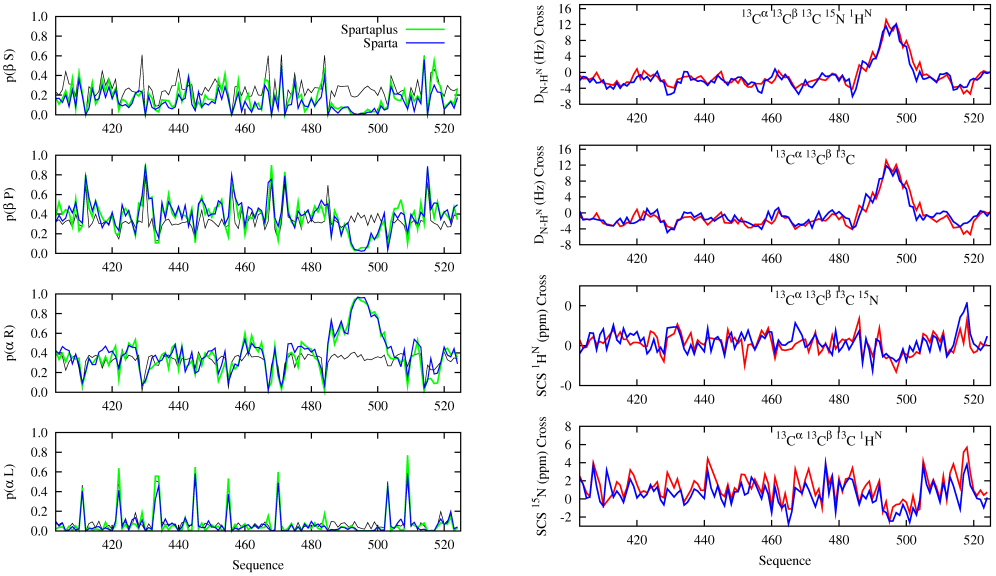


FIGURE 5.16 – ***Validation croisée des CDRs D_{NH}, et différences entre* Sparta *et* Sparta+.** *À gauche : Population respectivement du logiciel* SPARTA *(en vert) et* SPARTA+ *(en bleu) à l'issu des simulations* 3 *effectuées parallèlement. À droite : validation croisée des CDRs D_{NH}, et des déplacements ^{15}N et $^{1}H^{N}$ en bleu, les données expérimentales sont en rouge. Les validations croisées sont indépendantes.*

Conclusion du chapitre

Nous avons dans ce chapitre mis en place un protocole robuste permettant de caractériser quantitativement l'échantillonnage conformationnel des protéines intrinsèquement désordonnées. Nous avons tout d'abord étudié la signature des paramètres RMN en fonction des conformations adoptées. Les spécificités bien connues concernant les hélices α ou les feuillets β se sont révélées peu transposables dans le cas de la région Polyproline. Les valeurs des déplacements chimiques ou des couplages dipolaires ne permettent pas seule de différencier toutes les régions de l'espace Ramachandran, il est essentiel de les combiner pour caractériser quantitativement le paysage énergétique de ces protéines. De plus, n'utilisant que les déplacements chimiques carbones, les valeurs des motifs Polyproline se confondent aux valeurs des motifs *random-coil*. Après vérification de ces résultats lors de sélections avec Asteroids, nous avons identifié un protocole minimal utilisant les déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$ et $^{13}C'$ et les couplages dipolaires résiduels D_{NH} permettant de reproduire les trois motifs structurés. La robustesse de l'approche a alors été testé par l'ajout de bruit gaussien lors des simulations *in-silico*.

L'utilisation de ce protocole sur des données expérimentales se révéla fort intéressant à la fois d'un point méthodologique et biologique. La réponse obtenue fut différente suivant la protéine concernée, ne pouvant statuer pour l'ensemble des données, le protocole applicable aux données expérimentales le plus sûr inclut aussi les déplacements chimiques $^{1}H^{N}$ et^{15}N afin de faciliter l'identification des régimes étendus. Nous avons pu effectuer de nombreuses validations croisées concluantes, et d'autre part identifier plusieurs spécificités conformationnelles des protéines intrinsèquement désordonnées. La poids de la région βS issu de l'analyse des données expérimentales est moins important que celui présent dans la base de donnée d'angles (ϕ,ψ). Cette diminution se fait au profit de la région Polyproline, l'étape suivante est une généralisation de cette étude à d'un plus grand nombre de protéines désordonnées, les objectifs sont la caractérisation des structures transitoires mais aussi de l'état *random-coil*. Il serait ainsi possible de présenter par acide aminé la propension à échantillonner telle région de l'espace Ramachandran.

Caractérisation des interactions à longue portée des protéines intrinsèquement désordonnées

6

L'ordre présent au sein des protéines intrinsèquement désordonnées ne se limite pas à des motifs structurés tels que les hélices ou les feuillets, la présence d'interaction à longue portée est aussi une de leur caractéristique essentielle. Ces interactions seraient potentiellement impliquées dans l'agrégation de plusieurs protéines [149, 150]. Une méthode de choix pour étudier ce phénomène est la relaxation paramagnétique [90, 91], elle met un jeu un électron non apparié qui modifie la relaxation du système de spins. Cette interaction, dépendante de la distance électron-spin, fournit de l'information longue distance (jusqu'à 25Å) pouvant être implémentée dans une description par ensemble afin d'affiner la caractérisation des PIDs.

Nous commencerons par évaluer la capacité de l'algorithme Asteroids à reproduire la présence d'ordre à longue portée en réalisant une série de tests *in-silico*. Dans un second temps, nous analyserons l'influence réciproque de l'ordre local et l'ordre à longue portée et nous insisterons sur l'importance de combiner les paramètres RMN pour obtenir une description correcte de l'état déplié. Dans un troisième temps, nous prédirons *in-silico* l'influence d'interaction à longue portée sur le profit des CDRs et nous affinerons alors la paramétrisation de la ligne de base effectuée précédemment afin de prendre en compte ces prédictions. Nous conclurons en appliquant l'ensemble de ces résultats sur un système expérimental, la protéine α-Synucléine qui est sous forme fibrillaire un des marqueurs de la maladie de Parkinson.

6.1 Matériel et méthodes

6.1.1 Données expérimentales

Les données de relaxation de la protéine α-Synucléine ont été mesurées par Bertoncini et al. [151] sur quatre mutants pour lesquels les alanines suivantes ont été substituées par des cystéines : A18C, A76C, A90C, A140C.

6.1.2 Modélisation de la dynamique de la chaine latérale dans Flexible-Meccano

Utilisant la description, moléculaire par ensemble de Flexible-Meccano, la relaxation paramagnétique est calculée d'après le formalisme de la section précédente. Sauf mention contraire, le temps de corrélation global de la protéine est fixé à 5 ns, comme proposé dans la littérature [116, 151]. Le temps de réorientation de la chaine latérale MTSL est fixé à 500 ps. Pour information, le changement de cette valeur par un facteur multiplicatif de 2 n'implique pas de différence notable. Nous calculons alors explicitement et aléatoirement la position de la chaine latérale MSTL à partir de la base de données de rotamères. 600 conformations sont échantillonnées pour traduire son degré de flexibilité. Les conformations résultant sur un encombrement stérique avec la chaine principale du squelette ne sont pas prises en compte. Nous différencierons par la suite plusieurs degrés de flexibilité de la chaine latérale MTSL :

- Le modèle dynamique utilise le formalisme de la section précédente appliqué à l'ensemble des conformations de la chaine latérale MTSL, à l'exception des conformations incluant un encombrement stérique.

- Le modèle statique suppose l'existence d'un unique électron apparié. La valeur du paramètre d'ordre ${S_{H-e}}^2$ valant 1, la fonction de densité spectrale s'exprime donc :

$$J(\omega) = r_{H-e}^{-6} \left[\frac{\tau_R}{1 + \omega^2 {\tau_R}^2} \right] \tag{6.1}$$

  où r_{H-e} est la distance du vecteur électron-spin et τ_R le temps de corrélation global de la protéine. La position de l'électron apparié est calculée en moyennant l'ensemble des positions des 600 conformations calculées, à l'exception des conformations incluant un encombrement stérique.

- Le modèle dynamique variable traduit une flexibilité moins importante de la chaine latérale MTSL. Pour chaque structure et chaque tag paramagnétique, nous choisissons aléatoirement une conformation MTSL M sur laquelle nous définissons une sphère de rayon R. Nous définissons R_{max} comme la distance entre l'électron non apparié de la conformation M et l'électron non apparié des 599 autres conformations le plus distant. Utilisant le formalisme de la section précédente, nous utilisons uniquement les conformations situées dans une sphère de rayon $R_{max}/10$, $2R_{max}/10$ et ainsi graduellement pour représenter la flexibilité de la chaine latérale.

6.1.3 Sélection d'ensembles avec Asteroids

La fonction guidant la sélection est issue de l'équation 4.1 appliquée à la relaxation paramagnétique et s'écrit :

$$\chi^2 = \sum_{i,m} \left(\left[\frac{I_{para}}{I_{dia}} \right]_{i,m}^{calc} - \left[\frac{I_{para}}{I_{dia}} \right]_{i,m}^{exp} \right)^2 \tag{6.2}$$

où I_{para} est l'intensité liée à la résonance dans le cas où la chaine latérale est paramagnétique (le nitroxide est oxydé) et I_{dia} dans le cas diamagnétique (le nitroxide est réduit). L'estimation de ce rapport est présentée en section **??**. Nous considérons un poids identique pour tous les atomes et acides aminés en jeux.

Nous créons dans chaque cas un ensemble de 10000 structures avec FLEXIBLE-MECCANO sans contact spécifique, le nombre de structures utilisées dans la sélection dépend de la simulation considérée. 4000 itérations sont effectuées avec 100 individus.

6.1.4 Définition d'un contact

Un contact est une contrainte de distance maximale entre deux régions de la protéine imposée lors de la simulation d'un ensemble avec FLEXIBLE-MECCANO. A titre d'exemple, considérant une protéine de 100 acides aminés, l'ensemble associé possède un contact entre la région C-terminale et la région N-terminale lorsque chacune de ces structures a un atome $^{13}C^{\alpha}$ des résidus 1 à 10 distants de moins de 15Å d'un des atomes $^{13}C^{\alpha}$ des résidus 90 à 100. Sauf mention contraire, la distance maximale considérée sera 15Å et un espacement de 20 acides aminés entre les deux régions définissant le contact est requis afin d'éviter toute sur-interprétation liée à la proximité des résidus dans la séquence. Nous serons amenés à considérer des contacts transitoires de propension n% lorsque au moins n% de structures de l'ensemble respectent la contrainte imposée. Lors de sélection de sous-ensembles par ASTEROIDS nous pouvons aussi observer des contacts en comparant la distribution des distances moyennes entre résidus obtenue avec celle d'un ensemble *random-coil*.

6.1.5 La carte de contact

La distance moyenne entre les résidus i et j d'un ensemble est représentée par la métrique suivante :

$$\Delta_{ij} = \log \frac{\langle d_{ij} \rangle}{\langle d^0_{ij} \rangle} \tag{6.3}$$

où $\langle d_{ij} \rangle$ est la distance moyenne de l'ensemble considéré et $\langle d^0_{ij} \rangle$ est la distance moyenne de l'ensemble de référence. Par ensemble de référence, nous sous-entendons un ensemble sans contact spécifique comprenant au minimum 10000 structures. Les distances d_{ij} sont calculées entre l'atome $^{13}C^{\alpha}$ du résidu i et le atome $^{13}C^{\alpha}$ du résidu j.

Nous utilisons un graphique en trois dimensions avec la séquence en abscisse et ordonnée et les valeurs de la métrique selon l'axe z. Un gradient de couleur correspondant à l'axe z représente la valeur de la métrique Δ_{ij} entre deux résidus de la séquence. Sauf mention contraire, les couleurs tendant vers le rouge traduisent une compression de l'ensemble, les couleurs tendant vers le blanc traduisent une extension de l'ensemble tandis que la couleur bleue traduit l'absence de modification des distances moyennes de l'ensemble.

La métrique offre une vision rapide et localisée des interactions à longue portée au sein des protéines désordonnées mais le gradient de couleur ne serait être interprété spatialement de manière absolue. Notamment, la métrique favorise l'importance des contacts à proximité de la diagonale par rapport aux contacts situés au loin.

6.1.6 Données simulées associées à la section PRE

Considérant une protéine de 100 acides aminés dont la séquence est arbitraire, nous générons 10000 structures issues de Flexible-Meccano en calculant les valeurs de relaxation paramagnétique selon le formalisme de la section 6.1 en utilisant le modèle dynamique. Les données simulées sont obtenues en présence de 4 tags paramagnétiques espacés régulièrement sur la séquence en position : 20, 40, 60, 80. Nous définissons alors deux ensembles, l'un possédant un contact avec une propension de 100% entre les régions $11-20$ et $61-70$ et l'autre possédant un contact avec une propension de 100% entre les régions $41-50$ et $81-90$.

Un troisième ensemble est simulé à partir d'une autre séquence arbitraire de 200 acides aminés où est imposé un contact entre les régions $11-20$ et $61-70$ ou entre les régions $141-150$ et $181-190$. Les cystéines sont cette fois-ci situées en position 22, 44, 68, 88, 110, 132, 154 et 176.

Dans chaque cas, 600 conformations de chaines latérales sont échantillonnées, le temps de corrélation global vaut 5 ns et le temps de corrélation interne vaut 500 ps. Ce point est valable pour toutes les simulations de ce chapitre.

6.1.7 Données simulées associées à la section Information local et ...

Trois ensembles de 100000 structures sont générés avec le logiciel Flexible-Meccano à partir d'une séquence polyvaline de 100 acides aminés. Le premier est *random-coil*, le second possède un contact entre la région N-terminale [1 : 20] et le centre de la protéine [40 : 60] présent 100% du temps, le troisième n'inclut pas de contact spécifique mais 4 hélices α situées en position : $10-20$, $30-40$, $50-60$, $70-80$ avec une propension de 75%. Nous analyserons pour chacun la carte de contact, le rayon de giration et la distribution des angles dièdres.

Les CDRs sont calculés avec un tenseur d'alignement global. Le calcul des PRE est issu du modèle dynamique dont le formalisme est décrit en section 6.1. 600 conformations de la chaine latérale MTSL sont échantillonnées par chacune des 9 cystéines réparties uniformément sur la séquence en position : 10, 20, 30, 40, 50, 60, 70, 80, 90.

6.1.8 Données simulées associées à la section CDRs et PRE

Les CDRs sont calculées soit avec le tenseur d'alignement global, soit avec une fenêtre glissante de 15 acides aminés et une ligne de base appropriée. La ligne de base dépend principalement de la longueur de la chaine peptidique et de la présence de contacts spécifique entre deux régions éloignées de la chaine peptidique.

Afin de déterminer les effets liés à la présence d'ordre à longue portée sur les CDRs de protéines désordonnées, neuf ensembles sont générés avec Flexible-Meccano à partir d'une protéine de 100 acides aminés dont la séquence est arbitraire. Chaque ensemble contient 100000 structures, les couplages dipolaires D_{NH} et $D_{C^\alpha H^\alpha}$ sont calculés pour chaque structure et moyennés sur l'ensemble. Nous calculons les CDRs avec le tenseur global d'alignement. Un seul des ensembles ne possède pas de contact spécifique, pour les autres ensembles, les contacts sont introduits à différentes positions de la séquence et sur des régions plus ou moins grandes. Cette opération est ensuite répétée pour une séquence polyvaline afin de paramétrer la ligne de base.

Deux ensembles cibles de la protéine α-Synucléine de 100000 structures sont générées pour tester le protocole combinant CDRs et PRE, l'un contient un contact entre les régions $11-20$ et $61-70$ et l'autre entre les régions $41-50$ et $81-90$. Les CDRs sont calculés avec le tenseur d'alignement global et les PREs correspondant aux cystéines A18C, A76C, A90C, A140C sont calculés avec une chaine latérale dynamique.

6.1.9 Paramétrisation de la ligne de base

La paramétrisation de la ligne de base reprend celle précédemment effectuée pour une protéine dépliée. En présence de contact, nous modulant le profit de la ligne de base en ajoutant une gaussienne centrée sur la position médiane entre les deux régions définissant le contact. Deux autres petites gaussiennes sont rajoutées pour reproduire au mieux la ligne de base. Les paramètres clés sont la longueur de la séquence et les positions médianes n_1 et n_2 des deux régions définissant le contact, la paramétrisation s'exprime alors :

$$\begin{aligned} B(i,L,n_1,n_2) = &\left[2\,b\,\cosh\left(a(m-m_0)\right) - c\right] \\ &\times \left[1 - G\,e^{-\frac{(m-n_0)^2}{2\sigma^2}} + H\right. \\ &\left.\times \left[(D+S)\,e^{-\frac{(m-n_1+S/2))^2}{2\delta^2}} + (D-S)\,e^{-\frac{(m-n_2-S/2))^2}{2\delta^2}}\right]\right] \end{aligned} \tag{6.4}$$

où m_0, a, b et c sont fonction L de la longueur de la chaine :

$$\begin{aligned} m_0 &= \frac{L+1}{2} \\ a &= 0.33 - 0.22\left[1 - e^{-0.015\,L}\right] \\ b &= 1.16\,10^5\,L^{-4} \\ c &= 9.80 - 6.14\left[1 - e^{-0.021\,L}\right] \end{aligned} \tag{6.5}$$

où n_0, D, σ, S, G et H sont fonctions de deux médianes n_1 et n_2 des deux régions définissant le contact :

$$\begin{aligned} n_0 &= \frac{n_1+n_2}{2} \\ D &= |n_1 - n_2| \\ \sigma &= 0.109\,D + 4.6\,10^{-3}\,D^2 \\ S &= n_0 - m_0 \\ H &= 3.87\,10^{-5}\,D \\ G &= 1 - 6.66\;10^{-3}\,D \end{aligned} \tag{6.6}$$

où $\delta = 9.0$.

Cette expression dépend de la longueur de la protéine L et de la position des deux contacts n_1 et n_2. Le couplage du résidu i entre les spins I and S s'exprime alors :

$$D_{\mathrm{IS}}(i) = |B(i,L,n_1,n_2)|D_{\mathrm{IS}}^{\mathbf{LAW}} \tag{6.7}$$

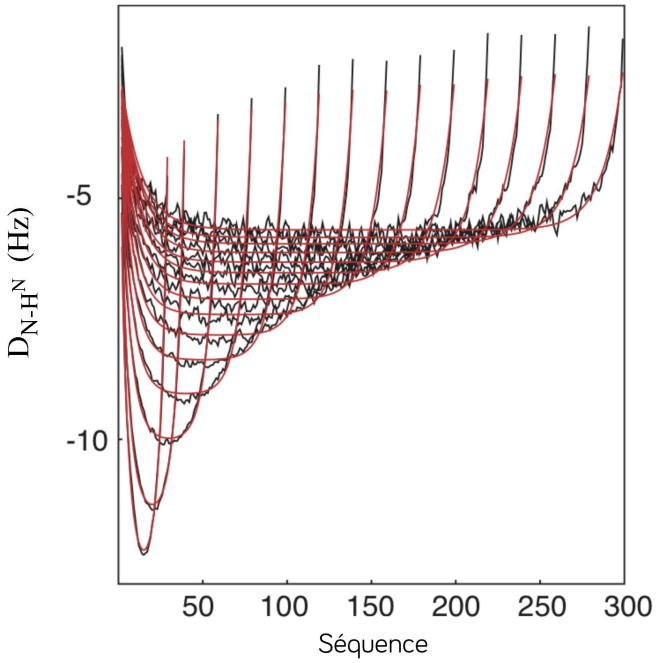


FIGURE 6.1 – ***Reproduction de la ligne de base sans contact pour différentes longueurs de chaine.*** *Les CDRs D_{NH} (en noir) sont calculés avec un tenseur d'alignement global sur 500000 structures, la paramétrisation issue des équations 6.4 et 6.6 est affichée en rouge. Les séquences utilisées sont des polyvalines de 20 à 300 résidus.*

6.2 Résultats

6.2.1 Validation de l'approche Flexible-Meccano Asteroids avec des données de relaxation simulées

En considérant un contact

Pour tester la capacité de l'algorithme génétique Asteroids à détecter des interactions à longue portée, nous réalisons deux tests *in-silico* sur les cibles précédemment présentées incluant des contacts spécifiques. Nous analyserons la reproduction des données mais aussi les caractéristiques biophysiques de l'ensemble telle que la distribution des distances moyennes à l'aide des cartes de contacts.

En premier lieu, quelle que soit la cystéine concernée, nous observons à son voisinage, une modulation du profil de relaxation en raison de la proximité entre l'électron non apparié et les protons amides environnants. Comparant les données *random-coil* et les données cibles, l'ajout d'un contact, en raison d'une proximité accrue entre les deux régions le définissant, modifie les profils de relaxation de l'ensemble qui se caractérisent par une nette diminution de l'intensité dans ces régions. Cette diminution est d'autant plus marquée si la cystéine est localisée sur une des régions définissant le contact. Ainsi, les tags paramagnétiques se comportent comme une sonde à longue distance permettant de détecter les contacts au sein des protéines intrinsèquement désordonnées. Présentée en figure 6.2, la reproduction des données après sélection par Asteroids d'un sous-ensemble de 80 structures est excellente, et ceci, quelle que soit la cystéine considérée.

Nous comparons ensuite les propriétés biophysiques de l'ensemble : la figure 6.3 montre la reproduction des distances moyennes de l'ensemble et la figure 6.4 montre la reproduction de la distribution des rayons de giration. Les cartes de contacts sont très

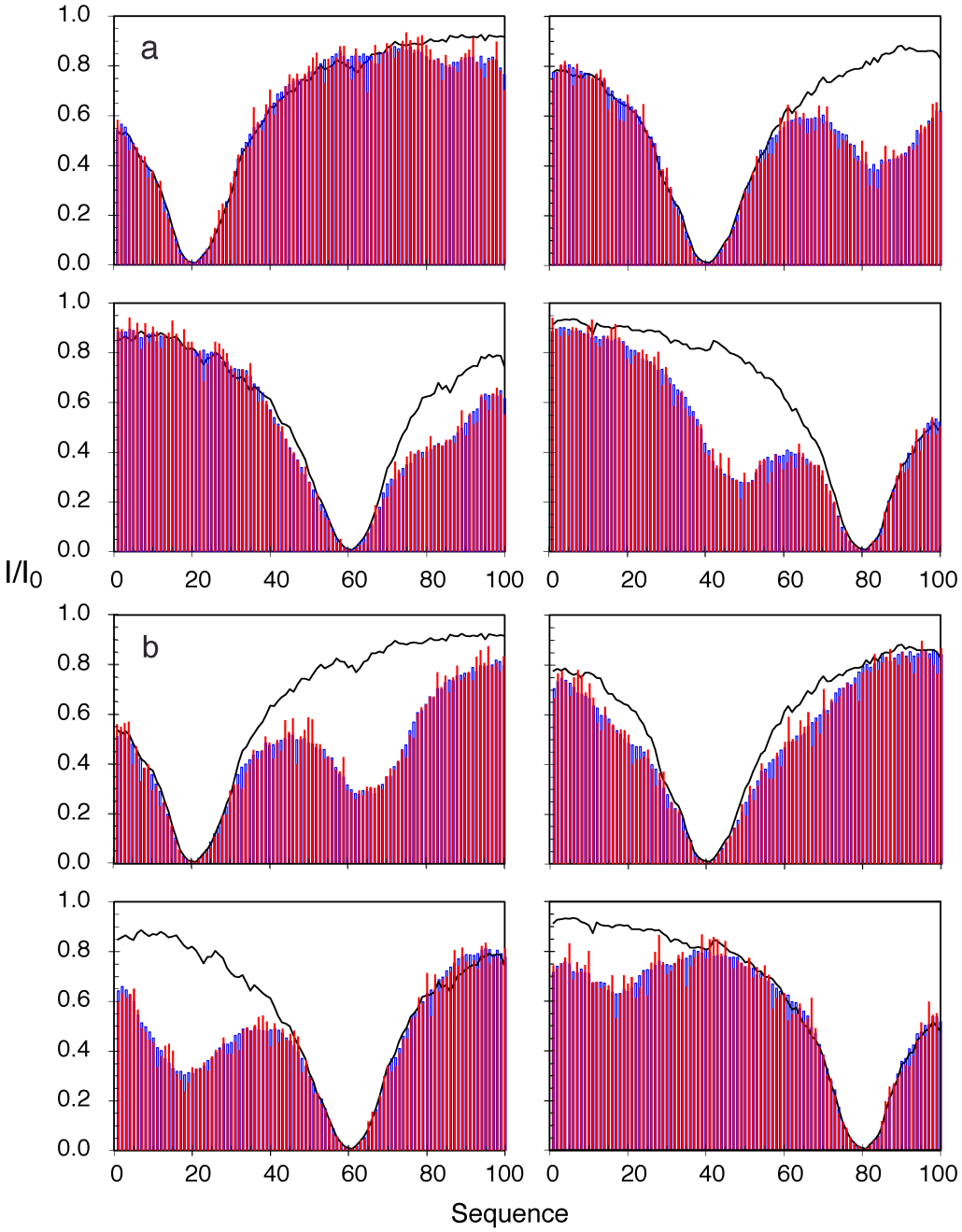


FIGURE 6.2 – ***Profils I/I_0 issus de tests in-silico sur des cibles incluant des contacts spécifiques.*** *Le premier ensemble cible possède un contact entre les acides aminés* 41-50 *et* 81-90 *(a), le second ensemble entre les acides aminés* 11-20 *et* 61-70 *(b). En bleu : données de l'ensemble cible, en rouge : données issues des sous-ensembles sélectionnés avec* ASTEROIDS*, en noir : ensemble random-coil c'est-à-dire sans contact spécifique. Chaque cadre correspond à la position de la chaine latérale MTSL : acide aminé* 20 *(haut gauche), acide aminé* 40 *(haut droite), acide aminé* 60 *(bas gauche), acide aminé* 80 *(bas droit).*

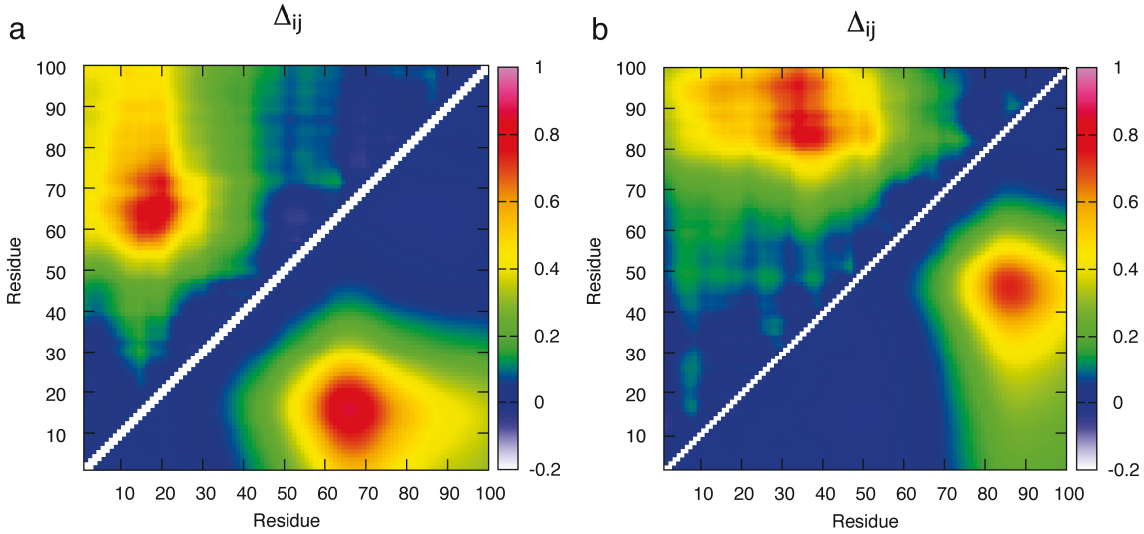


FIGURE 6.3 – ***Comparaison des cartes de contacts cibles et issues des sélections avec* Asteroids.** *En a) contact* 11-20 *et* 61-70, *en b) contact* 41-50 *et* 81-90. *Au-dessus de la diagonale, ensembles sélectionnés avec* ASTEROIDS, *au-dessous de la diagonale, ensembles cibles. L'échelle selon z correspondant aux données au dessus de la diagonale a été multipliée par* 0.5 *pour faciliter l'identification du contact.*

similaires, le contact des ensembles sélectionnés est néanmoins moins bien défini.

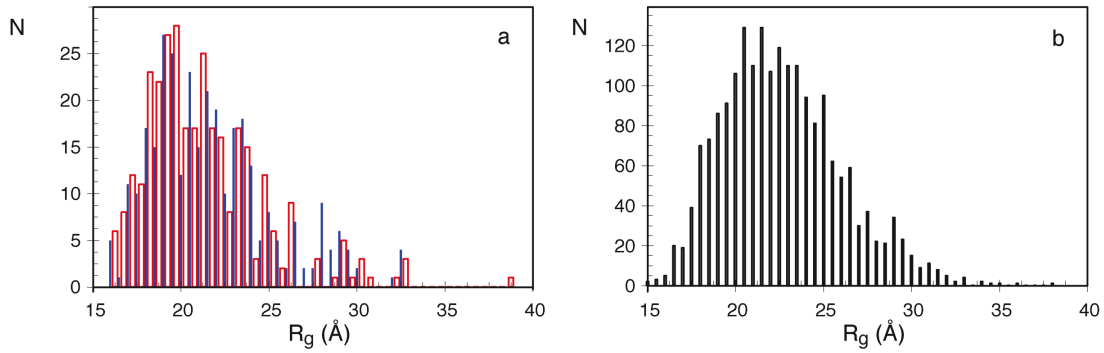


FIGURE 6.4 – ***Distribution du rayon de giration (R_g) pour l'ensemble cible et simulé associée au contact* 11-20 *et* 61-70.** *En a) distribution de R_g après sélection par* ASTEROIDS *d'un ensemble de* 80 *structures (bleu) ou* 160 *structures (en rouge). En b) distribution de R_g pour un ensemble cible de* 2000 *structures incluant le contact* 11-20 *et* 61-70, *ie toutes les structures respectent la contrainte de distance imposée.*

Nous obtenons un rayon de giration moyenne de 21.3 Å avec un ensemble de 80 structures, pour 22.6 Å pour l'ensemble cible. La sélection d'un ensemble de 160 structures augmente la valeur du rayon de giration moyen à 21.7 Å. Le profil de la distribution du rayon de giration de l'ensemble est relativement bien reproduit. La distribution des distances de nos ensembles est légèrement sous-estimée mais la qualité globale de l'approche FLEXIBLE-MECCANO ASTEROIDS reste néanmoins très correcte.

En considérant deux contacts

Nous appliquons le même protocole aux jeux de données de la protéine de 200 acides aminés pour détecter plusieurs contacts au sein d'une même protéine. Le reproduction des données reste excellente et la figure 6.5 présente la reproduction de la carte de contact de l'ensemble sélectionné par rapport à celle de l'ensemble cible. Bien que

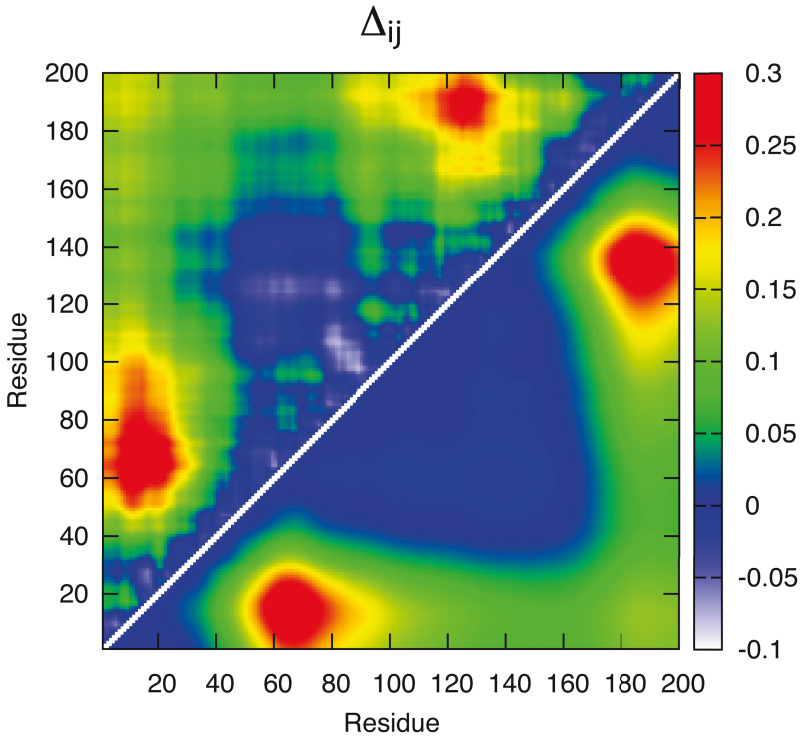


FIGURE 6.5 – ***Cartes de contact issues de données simulées incluant deux contacts entre les acides aminés* 11-20 *et* 61-70 *ou* 41-50 *et* 81-90.** *Au dessus de la diagonale, ensemble sélectionné avec* ASTEROIDS, *au-dessous de la diagonale, ensemble cible. L'échelle selon z correspondant aux données au dessus de la diagonale a été multipliée par* 0.5 *pour faciliter l'identification du contact.*

les contacts soient légèrement plus diffus, la qualité de l'approche reste excellente.

6.2.2 Applications aux données expérimentales de α-Synucléine

Validation du modèle de chaine latérale par validation croisée

Nous appliquons la procédure aux données expérimentales de la protéine α-Synucléine. Afin de déterminer le nombre de structures nécessaires pour reproduire correctement les données et les caractéristiques de l'ensemble, nous retirons de la sélection les données de la chaine latérale MTSL situées en position 76. Seules les données associées aux 3 cystéines paramagnétiques situées en positions 18, 90 et 140 sont donc incluses dans le calcul du χ^2 défini en équation 6.2. Nous effectuons des sélections avec des ensembles allant de 25 à 500 structures et en envisageant deux modèles physiques de chaine latérale MTSL : le modèle statique ou le modèle dynamique.

La figure 6.6 montre la reproduction des données passives du mutant en position 76, c'est-à-dire non prise en compte lors de la sélection, pour les deux modèles de flexibilité de la chaine latérale envisagés après sélection d'un ensemble de 200 structures par ASTEROIDS. Nous constatons une différence notable avec une bien meilleure reproduction des données si la flexibilité de la chaine latérale est prise en compte l'EQM statique vaut 0.24 ± 0.02 et l'EQM dynamique vaut 0.17 ± 0.01.

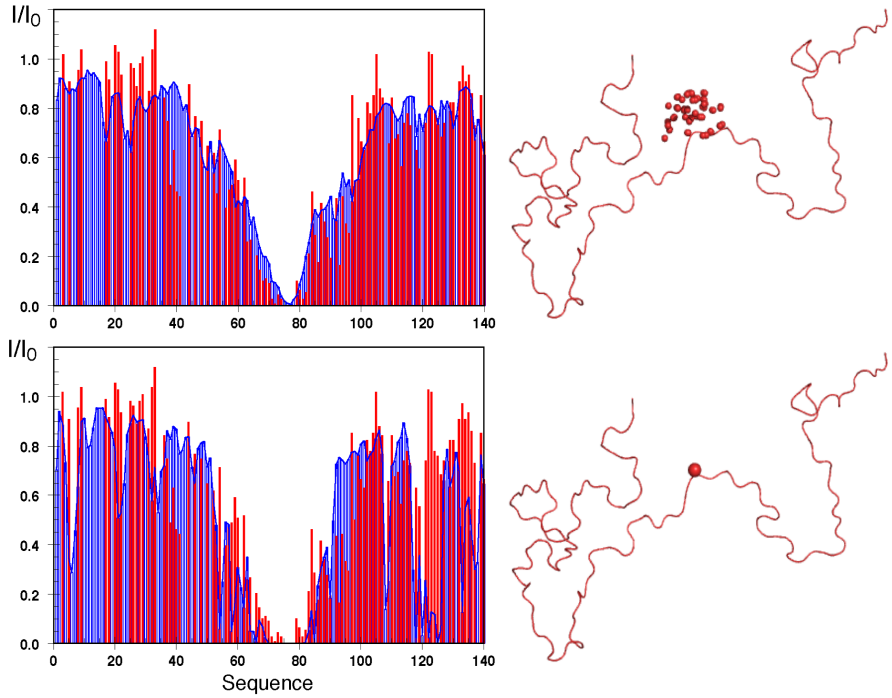


FIGURE 6.6 – ***Reproduction des données passives I/I_0 en utilisant une chaine latérale MTSL statique ou dynamique.*** *Données expérimentales (en rouge). À gauche : données issues de la sélection d'un ensemble de* 200 *structures avec* ASTEROIDS *(en bleu). À droite : deux schémas représentatifs de la flexibilité de la chaine latérale accompagnent les graphiques.*

La dynamique de la chaine latérale

L'utilisation d'une chaine latérale MTSL dynamique se traduit par un meilleur accord avec les données expérimentales. Les échelles de temps définissant l'interconversion des structures et la dynamique de la chaine latérale sont supposées statistiquement indépendantes.

Détermination du nombre de structure par validation croisée

L'évolution du rayon de giration, du χ^2 actif et passif permet de déterminer la taille de l'ensemble (figure 6.7). Ces paramètres évoluent rapidement et se stabilisent sur une plage aux alentours de 100 à 200 structures. Nous choisissons un ensemble de 200 structures, valeur à la fois appropriée avec la reproduction des données PRE et en accord avec la description réalisée au chapitre 4.

Carte de contacts résultante

Le modèle de la chaine latérale et le nombre de structures ayant été déterminés, nous appliquons la sélection aux 4 cystéines pour caractériser au mieux la protéine α-Synucléine. La figure 6.8 montre la reproduction des données après sélection et la carte de contacts associée. En accord avec les études précédentes [151, 123, 150], nous constatons un contact entre la région N-terminale et la région C-terminale. Ce contact pourrait protéger la partie NAC hydrophobique d'une exposition au solvant et inhiberait ainsi l'agrégation et la formation de fibrille.

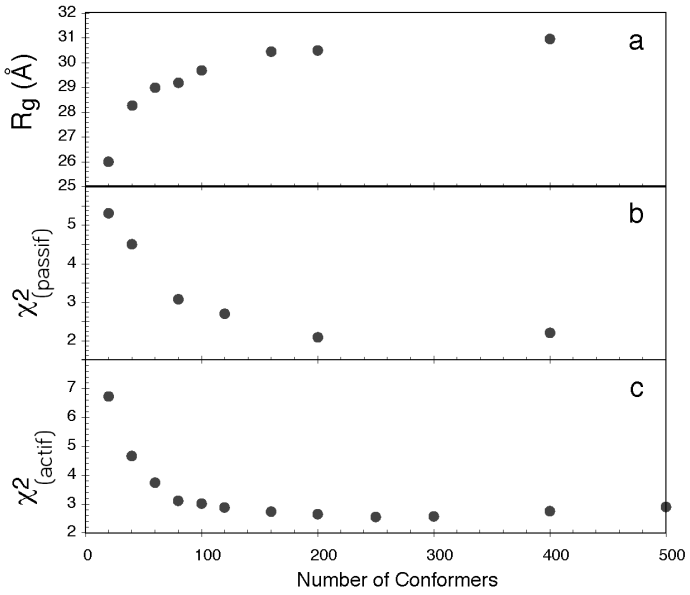


FIGURE 6.7 – ***Rayon de giration moyen et χ^2 actif et passif en fonction du nombre de structures.*** *Evolution du rayon de giration moyen R_g (a), de la reproduction passive (b) et active (c) des données PRE après sélection par* ASTEROIDS *en fonction du nombre de structures dans l'ensemble.*

6.2.3 Information locale et information à longue distance

Philosophie générale

La compréhension des paramètres RMN est fondamentale pour caractériser correctement les protéines désordonnées par description par ensemble. Jusqu'ici, pour déterminer des ensembles représentatifs de l'état déplié, nous avons utilisé séparément les PREs et les CDRs dans une approche utilisant conjointement FLEXIBLE-MECCANO et ASTEROIDS. Une approche combinant les couplages dipolaires résiduels et la relaxation paramagnétique semble un point essentiel à mettre en place pour acquérir une meilleure compréhension de l'état déplié. Avant cette étape, nous cherchons à identifier les implications de changements biophysiques sur les paramètres RMN et les corrélations potentielles entre l'information locale et à longue portée. En effet, nous distinguons deux types d'informations :

Une distinction essentielle

- L'information locale, c'est-à-dire l'échantillonnage conformationnel des résidus.
- L'information à longue portée, la présence de contact à moyenne ou longue portée au sein de la protéine.

La relaxation paramagnétique est fonction de la distance électron-spin et est donc particulièrement sensible à l'information à moyenne et longue portée. Elle permet de déterminer un ensemble de contraintes de distance entre la position effective de l'électron non apparié et l'ensemble des protons amides de la protéine. Les couplages dipolaires résiduels sont extrêmement sensibles à l'échantillonnage local des résidus. Par ailleurs, nous avons présenté en section **??** le formalisme de la ligne de base traduisant les effets à longue portée au sein de la protéine. La compréhension du rôle de ces paramètres en fonction des caractéristiques biophysiques est primordiale pour réaliser toute description par ensemble. Nous allons introduire des biais dans la description random-coil afin

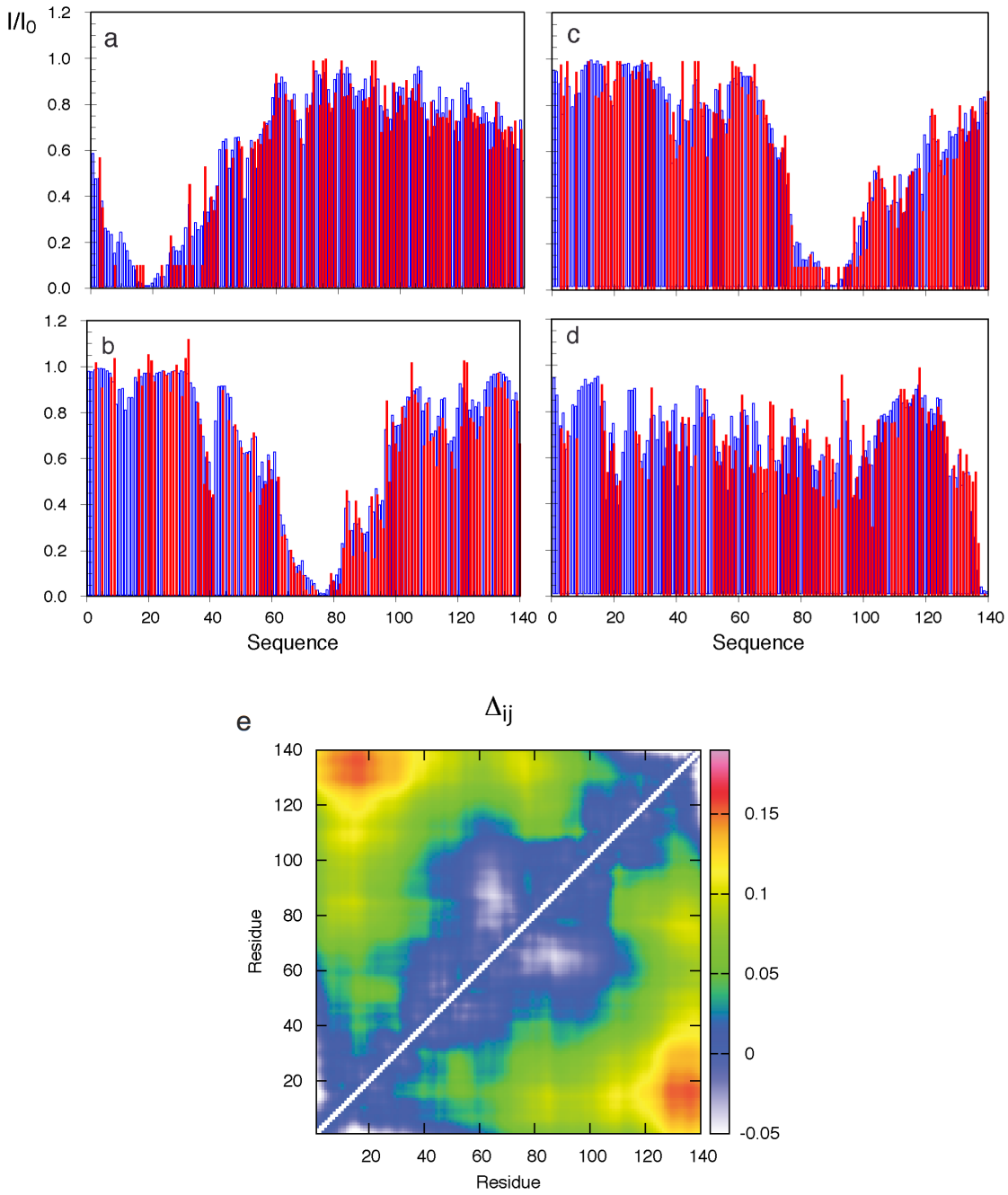


FIGURE 6.8 – ***Reproduction des données expérimentales I/I_0 avec* Asteroids *et carte de contacts associée.*** *(a-d) données expérimentales en rouge et données issues de la sélection avec* ASTEROIDS *en bleu pour les positions de cystéines suivantes : (a) A18C, (b) A76C, (c) A90C and (d) A140C. La carte de contact (e) indique un contact entre la région C-terminale et N-terminale.*

d'étudier l'influence, soit d'un contact, soit des structures transitoires hélicoïdales, sur les paramètres RMN de l'ensemble.

Présence d'un contact

Nous commençons par étudier l'influence d'un contact complètement présent au sein d'un ensemble. Les graphiques présentés sont volontairement denses en information, l'objet de ce paragraphe étant d'insister sur les corrélations existantes ou non entre les paramètres RMN ou caractéristiques étudiées.

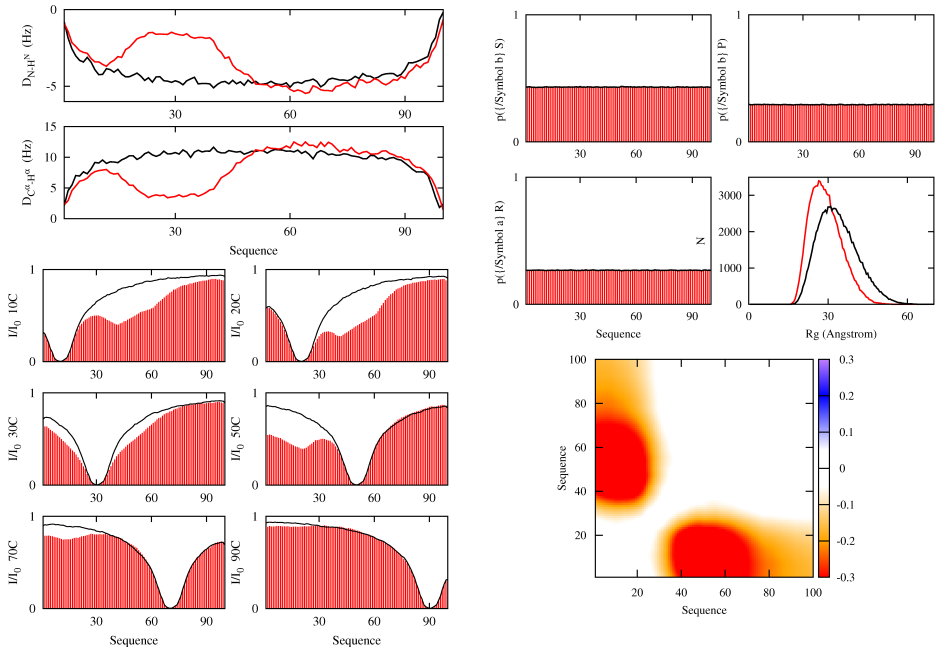


FIGURE 6.9 – ***Influence de la présence d'un contact entre la partie C-terminale et la région centrale d'un ensemble polyvaline.*** *Nous comparons un ensemble random-coil (en noir) avec l'ensemble incluant le contact (en rouge). À gauche : sont présentés les paramètres RMN avec les CDRs D_{NH} et $D_{C^\alpha H^\alpha}$ (en haut) et 6 profils de relaxation paramagnétique des cystéines* 10, 20, 30, 50, 70, 90 *(en bas). À droite : les paramètres biophysiques correspondants sont les populations d'angles dièdres (en haut), la distribution du rayon de giration (en haut à droite) et la carte de contact de l'ensemble normalisé avec le random-coil (en bas).*

La figure 6.9 montre clairement la modification des profils des paramètres RMN, nous observons d'une part pour les CDRs D_{NH} et $D_{C^\alpha H^\alpha}$ un amoindrissement de la ligne de base entre les deux contacts. Concernant les caractéristiques biophysiques : nous avons évidemment une diminution des distances moyennes de l'ensemble et de la distribution de rayon de giration mais surtout la figure à droite ne montre aucun changement notable concernant la distribution des angles (ϕ,ψ) dans l'espace de Ramachandran. Ainsi, l'ajout d'un contact n'implique pas de biais dans l'espace conformationnel en favorisant une orientation privilégiée.

Non corrélation de l'information à longue portée sur l'information locale

La présence de contacts à longue portée n'influence pas l'échantillonnage conformationnel des protéines. En théorie, il existe évidemment une modification de l'échantillonnage pour chaque structure mais l'espace conformationnel accessible étant si vaste qu'il n'est pas possible de détecter cette modification en moyenne sur l'ensemble.

Les profils de la relaxation paramagnétique de l'ensemble *random-coil* sont caractéristiques d'une protéine dépliée. Nous observons une diminution de l'intensité uniquement autour de la cystéine, cette diminution d'intensité traduit la proximité entre l'électron de la chaine latérale MTSL et le spin nucléaire du proton des différents résidus. Dans le cas d'un contact, en raison d'une proximité accrue entre les deux régions le définissant, les profils de relaxation de l'ensemble (en rouge) se caractérisent par une diminution de l'intensité dans ces régions (figure 6.9). Ainsi, les profils de relaxation situés dans la partie N terminale de la protéine sont nettement plus faibles, et ceci, même loin de la région environnante des cystéines. La carte de contact permet de visualiser clairement le contact, c'est-à-dire la contraction de l'ensemble entre la région N-terminale et le domaine central de la protéine.

L'absence de corrélation entre la présence de contact et l'échantillonnage conformationnel augmente la dégénérescence des solutions lors de sélection de sous-ensembles. Considérant une protéine possédant un contact, cherchant à reproduire uniquement la valeur du couplage dipolaire résiduel D_{NH}, Asteroids risque d'introduire un biais dans l'échantillonnage local pour reproduire au mieux les données bien qu'en réalité la modulation soit due à la présence d'interaction à longue portée, l'ensemble résultant de la sélection n'aura donc pas les mêmes caractéristiques. Pour une telle approche, il semble donc essentiel de combiner les D_{NH} avec des paramètres RMN supplémentaires tels que les PREs pour lever la dégénérescence. La combinaison de plusieurs jeux de CDRs peut aussi venir à bout de ce problème mais l'interprétation du résultat est plus délicate dans ce cas, la combinaison de plusieurs paramètres RMN permet de bien différencier l'information.

Présence d'une structure secondaire

Nous souhaitons maintenant analyser la présence d'ordre résiduel symbolisé par 4 hélices coopératives très présentes : trois de propension de 75% et une de 33% de la séquence échantillonnent spécifiquement la région αR.

Présentées en figure 6.10, nous observons une forte modification du profil des CDRs en raison la présence d'ordre résiduel au niveau des hélices, les D_{NH} sont positifs et les $D_{C^\alpha H^\alpha}$ oscillent. La distribution des angles dièdres en forme de créneaux montre clairement une augmentation locale de l'échantillonnage de la région αR.

Utilisant comme référence un ensemble *random-coil*, les profils de la relaxation paramagnétique sont légèrement modifiés, principalement aux alentours des cystéines. Nous observons des variations traduisant une contraction des distances de l'ensemble au niveau des hélices et une légère extension dans les autres régions. Nous ne notons pas de modifications des distances moyennes à longue portée, cette information est corroborée par la carte de contact et le distribution du rayon de giration. La présence de structures secondaires ne peut impliquer de contacts à longue portée mais peut

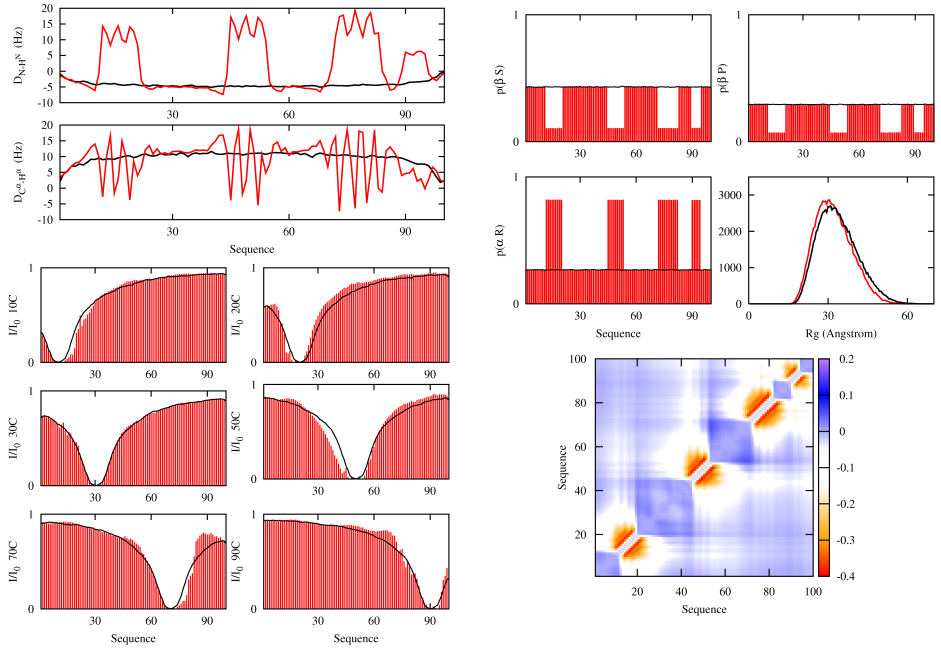


FIGURE 6.10 – ***Influence de la présence de 4 hélices transitoires de propension 75% sur un ensemble polyvaline.*** *Nous comparons un ensemble random-coil (en noir) avec l'ensemble incluant les hélices (en rouge) situées en position* $12-20$, $45-53$, $72-82$, $90-94$. *À droite, sont présentées les paramètres RMN avec les CDRs* D_{NH} *et* $D_{C^\alpha H^\alpha}$ *(en haut) et 6 profils* I/I_0 *des cystéines* 10, 20, 30, 50, 70, 90 *(plus bas). À gauche, les paramètres biophysiques correspondants sont les populations d'angles dièdres (en haut), la distribution du rayon de giration (en haut à droite) et la carte de contact de l'ensemble normalisé avec le random-coil (en bas).*

modifier localement les distances moyennes de l'ensemble, une hélice α impliquera une légère contraction de l'ensemble tandis qu'une hélice PPII une légère extension.

Influence de l'échantillonnage locale : cas expérimental

Comme exposée, la présence d'un échantillonnage spécifique peut influencer le profil des PRE. Les différences dépendront du nombre de motifs, de leur longueur et de leur propension. Pour pouvoir visualiser ces changements, ils doivent avoir lieu à proximité de la cystéine, autrement la moyenne en $\frac{1}{r^6}$ ne permet pas de modification notable des distances de l'ensemble. A titre d'exemple, la cystéine 50 de l'exemple précédent affiche un profil plus bas en raison de l'hélice située en 45-53, ce qui n'est pas le cas pour la cystéine 30.

Pour illustrer cet effet sur des données expérimentales, nous utilisons des résultats du chapitre 5 réalisés sur la protéine K18. Pour information, K18 ne possède pas de contact spécifique et peut être considérée comme une protéine complètement dépliée. Nous déterminons de l'échantillonnage conformationnel en combinant les déplacements $^{13}C^\alpha$, $^{13}C^\beta$, $^{13}C'$, ^{15}N, $^1H^N$ et les couplages dipolaires résiduels D_{NH} dans une sélection avec ASTEROIDS et constatons l'obtention de conformations plus étendues (notamment dans la région 250-265). Nous calculons alors le profil de relaxation correspondant en utilisant les temps de corrélation précédents et une chaine latérale dynamique, le profil de relaxation d'un ensemble *random-coil* et comparons ces jeux de données aux données expérimentales afin de corréler ces informations. L'ensemble comprenant l'échantillonnage conformationnel issu de la sélection reproduit mieux les données PRE. Notamment, le profil des PRE de la cystéine 280C est nettement plus en accord autour

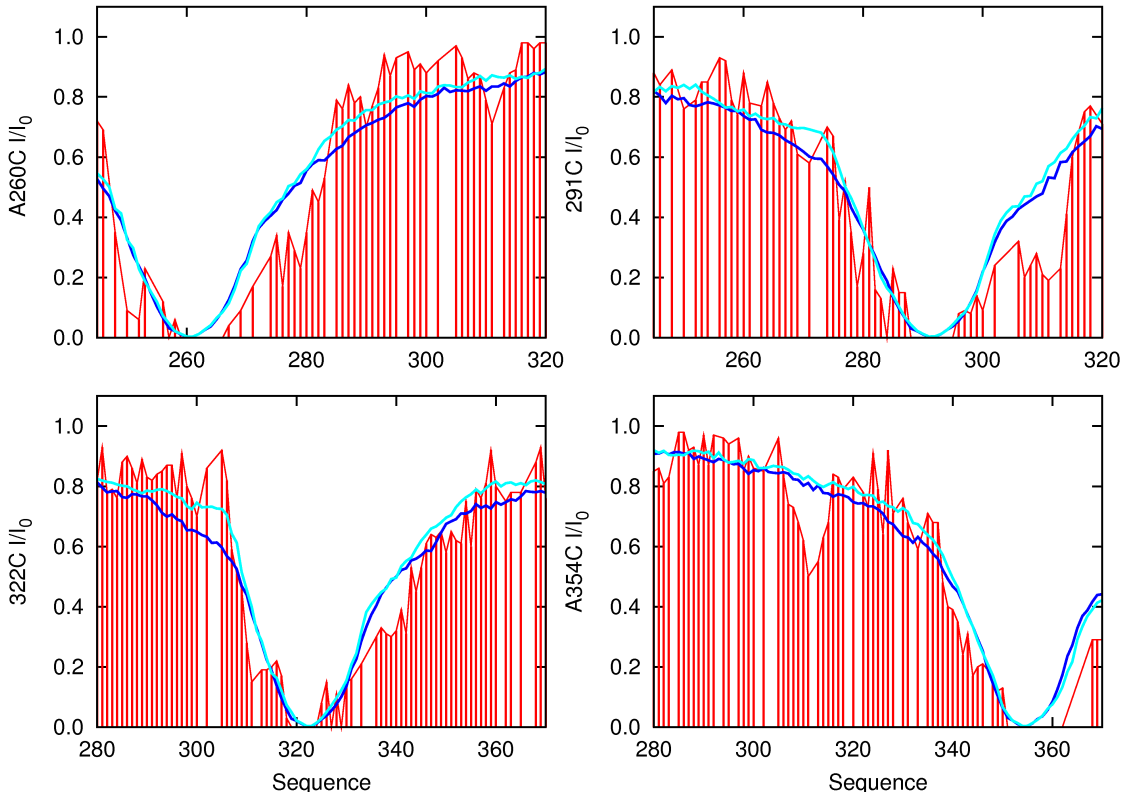


FIGURE 6.11 – ***Influence de l'échantillonnage local sur les profils I/I_0 de la protéine K18.*** *Les données expérimentales correspondantes aux 4 cystéines en position A260C, 291C, 322C et A354C sont en rouge. La simulation random-coil est en bleu, la simulation en cyan utilise une base de données d'angles (ϕ,ψ) issue d'une sélection avec* ASTEROIDS.

des résidus environnant la cystéine.

La détermination de l'échantillonnage conformationnel de la protéine permet d'améliorer la reproduction des données de relaxation paramagnétique, l'assomption contraire n'étant cependant pas valide. Pour la cystéine numéro 322C, nous observons une diminution du profit d'intensité autour du résidu 270 non pris en compte par l'échantillonnage local. Il existe possiblement un repliement très localisé entre la région $265-275$ et la région $310-320$. Pour ce cas, seule l'utilisation des PRE lors de sélection permettra de reproduire ces données d'où la nécessité de combiner l'ensemble des données RMN pour obtenir une description précise et quantitative des protéines désordonnées.

6.2.4 L'introduction d'ordre à longue portée modifie les CDRs des protéines désordonnées

FLEXIBLE-MECCANO permet non seulement d'analyser et d'interpréter les données expérimentales mais aussi de prédire l'influence des caractéristiques biophysiques sur les paramètres RMN. Nous allons donc dans cette partie étudier l'influence des contacts sur les profils des couplages dipolaires résiduels et en particulier de la ligne de base. Cette approche nécessite évidemment une confrontation avec des données expérimentales mais offre dans un premier temps un moyen rapide d'étudier les interactions existant au sein des PIDs.

La figure 6.12 montre les effets liés à la présence d'ordre à longue portée sur le profil des CDRs pour une séquence arbitraire de 100 acides aminés. Nous constatons un déplacement de la courbe entre les deux régions définissant le contact. Ce décalage

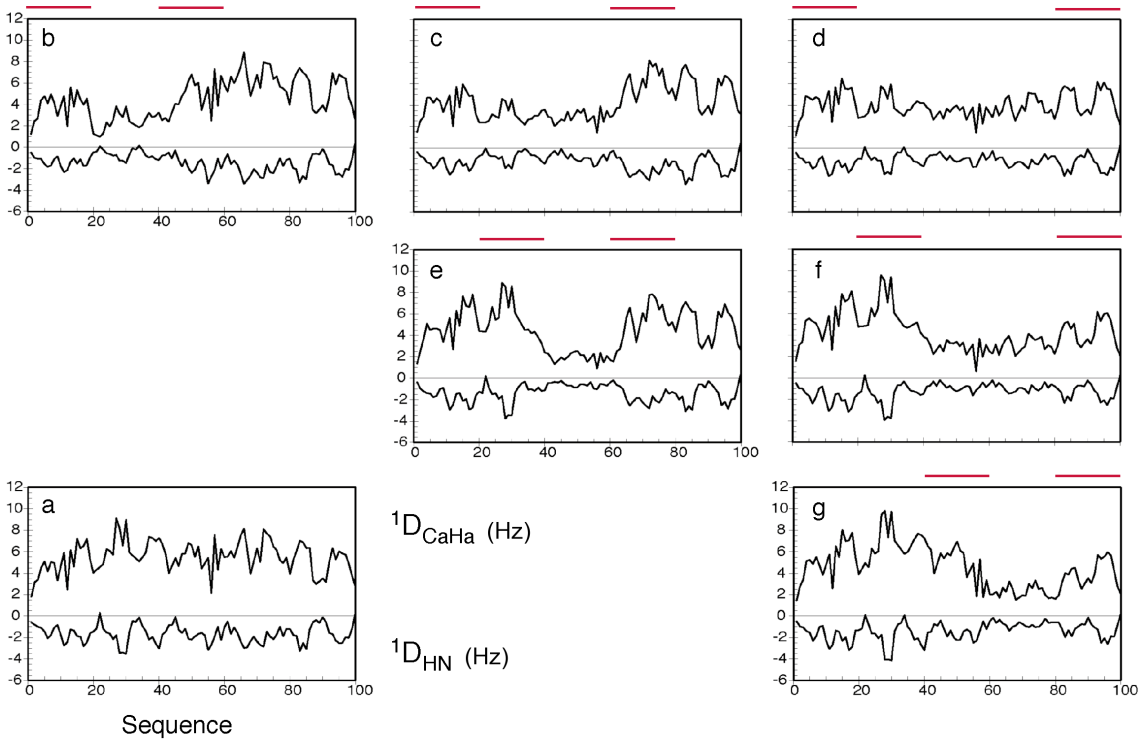


FIGURE 6.12 – ***Profil de CDRs D_{NH} et $D_{C^\alpha H^\alpha}$ simulés en présence de contacts à longue portée pour une séquence arbitraire de 100 résidus.*** *a) Profil des CDRs calculés sur 100000 structures en l'absence de contacts spécifiques. (b-g) Profils des CDRs calculés sur 100000 structures en présence de contacts définis par les régions i et j : (b) $i = 1-20$, $j = 41-60$; (c) $i = 1-20$, $j = 61-80$; (d) $i = 1-20$, $j = 81-100$; (e) $i = 21-40$, $j = 61-80$; (f) $i = 21-40$, $j = 81-100$; (g) $i = 41-60$, $j = 81-100$. Le trait rouge situé sur chaque cadre indique la position des régions définissant le contact.*

dépend de la position du contact : plus les régions définissant le contact sont proches, plus ce décalage est important, et réciproquement. Comme précédemment expliqué, l'échantillonnage conformationnel de la protéine ne semble pas modifié par la présence de contacts.

Nous réitérons le calcul avec une séquence polyvaline afin de supprimer l'échantillonnage local et d'afficher les lignes de bases correspondantes (figure 6.12). Pour l'ensemble *random-coil*, nous retrouvons la ligne de base connue défini par un cosinus hyperbolique, les autres courbes incluant un contact sont modulées, nous observons un amoindrissement de la ligne de base entre les deux régions définissant le contact. L'ajout de fonctions gaussiennes à la fonction cosinus permet en première approximation de reproduire correctement les modulations de la ligne de base (se référer à la section 6.1.9 et à l'équation 6.4).

Pour valider l'utilisation de la fenêtre glissante en présence de contact, nous calculons parallèlement les valeurs des couplages dipolaires D_{NH} et $D_{C^\alpha H^\alpha}$ pour un ensemble possédant un contact entre la région $41-60$ et la région $81-100$, nous utilisons d'une part le tenseur d'alignement global sur un ensemble de 100000 structures de manière à faire converger les CDRs, et d'autre part une fenêtre glissante de 15 résidus sur un ensemble de 200 structures combinée à la ligne de base paramétrée. La figure 6.14 présente les courbes associées aux deux méthodes de calcul et leur parfaite superposition valide la méthode.

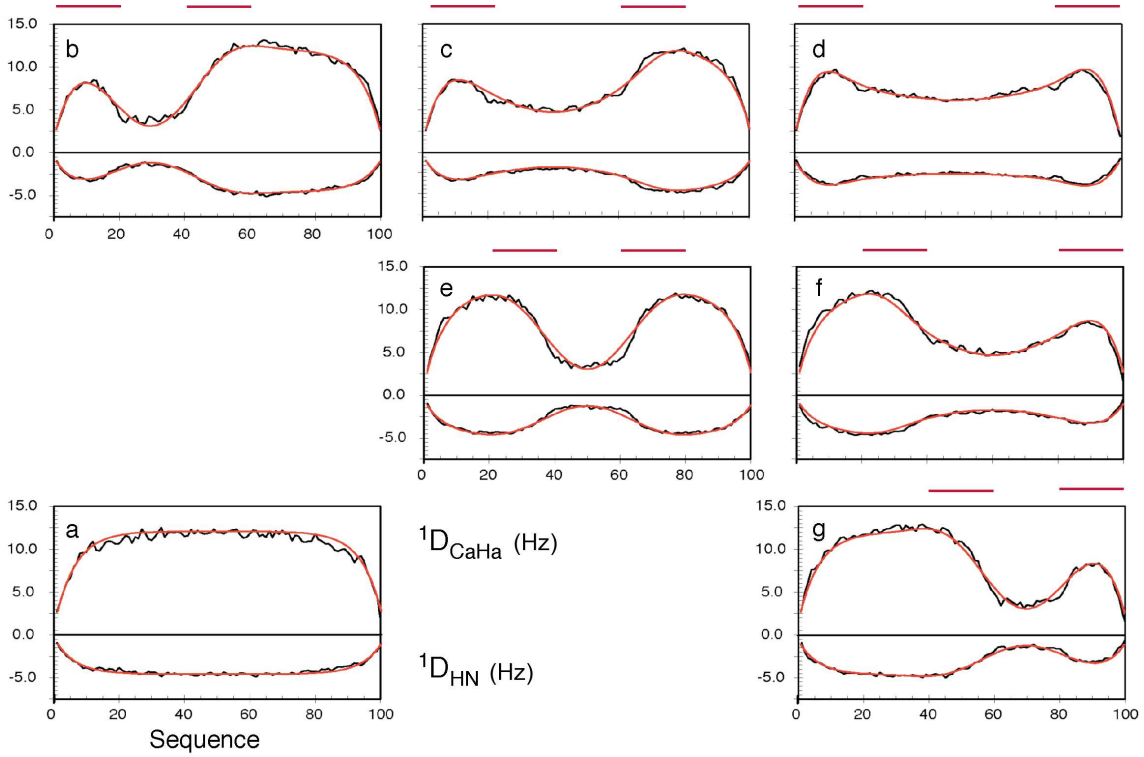


FIGURE 6.13 – ***Profil de CDRs D_{NH} et $D_{C^\alpha H^\alpha}$ simulés en présence de contacts à longue portée pour une séquence polyvaline de 100 résidus.*** *Les données simulées calculées sur 100000 structures sont en noir, la paramétrisation associée en rouge issue de l'équation 6.4. (a) Profil des CDRs en l'absence de contacts spécifiques. (b-g) Profils des CDRs en présence de contacts définis par les régions i et j : (b) $i = 1-20$, $j = 41-60$; (c) $i = 1-20$, $j = 61-80$; (d) $i = 1-20$, $j = 81-100$; (e) $i = 21-40$, $j = 61-80$; (f) $i = 21-40$, $j = 81-100$; (g) $i = 41-60$, $j = 81-100$. Le trait rouge situé sur chaque cadre indique la position des régions définissant le contact.*

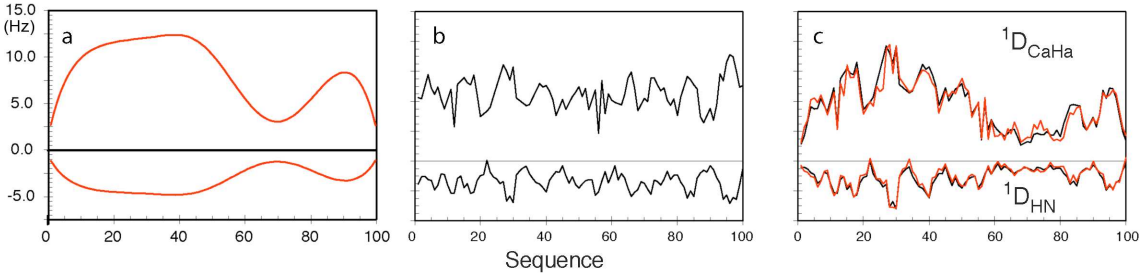


FIGURE 6.14 – ***Utilisation combinée de la ligne de base et de la fenêtre glissante pour calculer les CDRs en présence d'interaction à longue portée.*** *(a) Calcul analytique de la ligne de base possédant un contact entre les positions 50 et 90. (b) CDRs calculés sur 200 structures en utilisant la fenêtre glissante. (c) Multiplication de la ligne de base (a) par les CDRs locaux de (b) (en rouge), le tout comparé aux profils des CDRs calculés sur un ensemble de 100000 structures en utilisant le tenseur d'alignement global et incluant des contacts spécifiques entre les régions $41-60$ et $81-100$ (en noir).*

La fenêtre glissante fonctionne en présence de contact

La présence d'ordre résiduel à longue portée ne modifie pas l'échantillonnage conformationnel. Combinant le calcul des CDRs avec la fenêtre glissante et la paramétrisation de la ligne de base, nous pouvons reproduire avec 200 structures le profil des CDRs calculés en utilisant un tenseur d'alignement global, et ceci, même en présence d'ordre à longue portée.

6.2.5 Utilisation combinée de jeux de données simulées PRE et CDRs

Nous allons dans cette partie combiner les données PRE et CDRs, nous choisissons cependant de ne pas sélectionner conjointement ces paramètres avec Asteroids. La sélection de sous-ensembles étant particulièrement délicate et pouvant mener à des solutions incorrectes, nous préférons introduire plusieurs étapes permettant une analyse de chaque résultat obtenu. Le protocole utilisé sera le suivant :

1 Nous sélectionnons avec Asteroids un sous-ensemble reproduisant les données PRE.

2 Nous localisons la position du contact en analysant la distribution des distances moyennes $^{13}C^{\alpha}$-$^{13}C^{\alpha}$ du sous-ensemble.

3 Nous calculons la ligne de base correspondante avec le formalisme préalablement introduit à l'équation 6.4.

4 Nous combinons cette ligne de base avec les CDRs calculés avec la fenêtre glissante sur un ensemble *random-coil* pour une validation croisée de ces derniers.

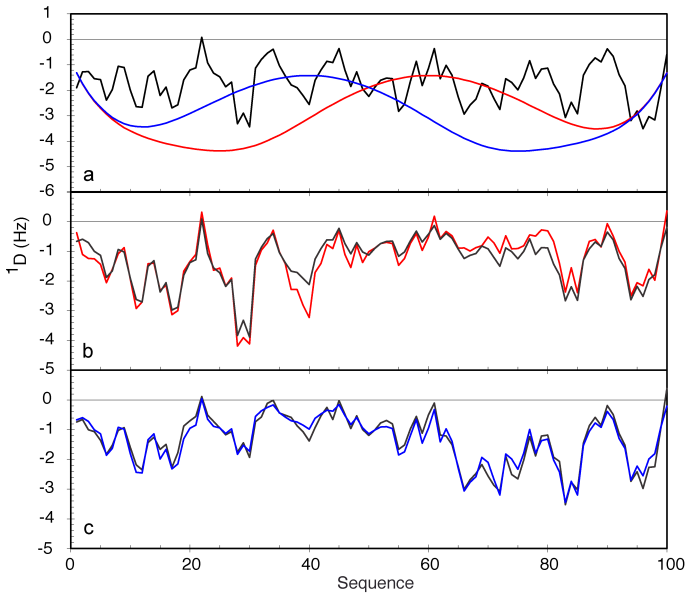


Figure 6.15 – ***Utilisation de données simulées en combinant les CDRs et les PREs.*** *Les PREs sont utilisés pour déterminer les contacts à longue portée (figure 6.3). (a) CDRs calculés avec la fenêtre glissante (en noir), ligne de base paramétrée avec un contact entre les régions* 11-20 *et* 61-70 *(en bleu) ou un contact entre les régions* 41-50 *et* 81-90 *(en rouge). Comparaison des CDRs calculés avec les deux approches : un tenseur d'alignement global (en noir) et la fenêtre glissante multipliée par la ligne de base pour le contact* 45-85 *(en rouge, en b) et pour le contact* 15-65 *(en bleu, en c).*

Nous effectuons tout d'abord deux tests *in-silico* sur deux ensembles de référence, l'un possédant un contact entre les régions $11-20$ et $61-70$ et l'autre entre les régions

41-50 et 81-90 (figure 6.3). La figure montre la validation croisée des CDRs, la reproduction des couplages D_{NH} est très bonne. L'utilisation des PRE pour localiser le contact combiné aux calculs des CDRs avec la ligne de base paramétrée et la fenêtre glissante apparait comme une méthode robuste pour caractériser les PIDs.

6.2.6 Utilisation combinée de jeux de données expérimentaux PRE et CDRs de la protéine α-Synucléine

Le protocole précédent est appliqué aux données expérimentales de la protéine α-Synucléine. La position du contact est issue de la carte de contact déterminée avec Asteroids (figure 6.8). Nous utilisons ensuite 200 structures pour estimer les valeurs des CDRs en utilisant la fenêtre glissante qui sont multipliées par la ligne de base prenant en compte la position du contact.

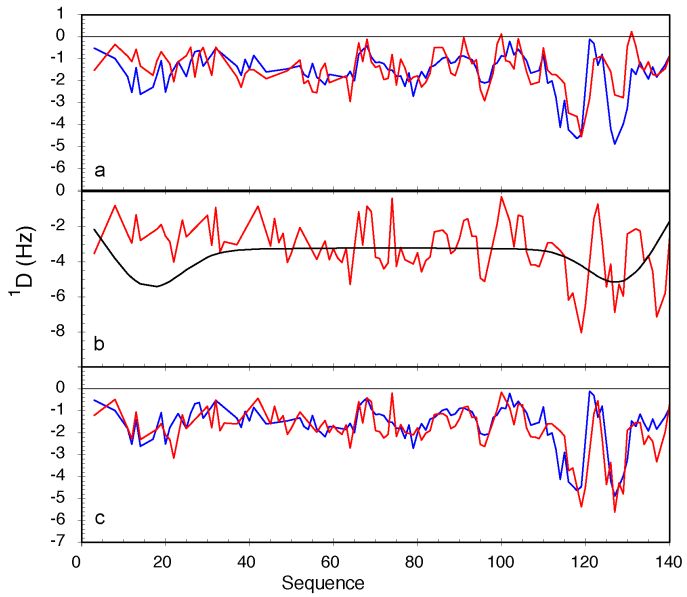


Figure 6.16 – ***Analyse combinée des CDRs et de la PRE en utilisant* Asteroids.** *Comparaison des CDRs D_{NH} expérimentaux avec ceux obtenus en utilisant conjointement la fenêtre glissante et la ligne de base paramétrée en tenant compte du contact déterminé des PREs. (a) CDRs D_{NH} expérimental (en bleu) et issu d'un ensemble random-coil de* Flexible-Meccano *(en rouge). (b) CDRs calculé avec la fenêtre glissante (en rouge) et ligne de base paramétrée en tenant compte du contact déterminé des PREs (en noir). (c) Combinaison des deux courbes du cadre (b) (en rouge) comparé aux CDRs expérimentales D_{NH} (en bleu).*

La combinaison de la fenêtre glissante et de la ligne de base offre une meilleure reproduction des couplages dipolaires résiduels D_{NH} avec un RMSD égal à 0.52 comparé aux cas *random-coil* de Flexible-Meccano pour lequel le RMSD vaut 0.72. Les extrémités de la protéine sont nettement mieux reproduites (figure 6.16).

Ce résultat est fondamental à de nombreux niveaux : d'une part, il valide l'existence de contact à longue portée au sein des PIDs, ces contacts à longue portée peuvent être détectés avec les PREs et les CDRs, d'autre part, les prédictions effectuées avec Flexible-Meccano concernant la modulation de la ligne de base sont correctes. Il est alors possible d'analyser conjointement ces données dans un protocole utilisant Flexible-Meccano et Asteroids afin d'identifier les caractéristiques de la protéine.

Conclusion du chapitre

La relaxation paramagnétique permet de caractériser la présence d'ordre à longue portée au sein des protéines intrinsèquement désordonnées. Nous avons appliqué la description par ensemble sous contraintes en sélectionnant avec l'algorithme génétique Asteroids un ensemble reproduisant les données de relaxation paramagnétique. Nous avions préalablement testé le protocole sur des données simulées en présence de contact et vérifié la reproduction des données et des caractéristiques physiques de l'ensemble : la distribution des distances moyennes et la distribution du rayon de giration. Nous avons ainsi détecté la présence d'un contact entre la région N-terminale et la région C-terminale sur les données PRE expérimentales de α-Synucléine. Parallèlement, nous avons étudié l'influence des interactions à longue portée sur les couplages dipolaires résiduels c'est-à-dire sur la ligne de base. Nous avons intégré la présence de contact dans la formalisme présenté aux chapitres 4 combinant la fenêtre glissante et la ligne de base pour calculer les couplages dipolaires simulés et reproduire les données expérimentales D_{NH}. Cette approche est la première combinant les CDRs et les PREs pour caractériser les propriétés des protéines désordonnées, elles exploitent la sensibilité des paramètres RMN à l'information locale et l'information à longue distance.

Les perspectives de ce chapitre sont nombreuses, il faut d'une part combiner de nouveaux paramètres RMN avec la relaxation paramagnétique, l'incorporation des déplacements chimiques et des couplages dipolaires résiduels étant la prochaine étape. La sélection de structures en combinant directement les CDRs et les PREs nécessite le calcul d'une ligne de base pour chaque structure qui sera alors moyennée sur l'ensemble et multiplié par la valeur des couplages dipolaires résiduels calculés avec la fenêtre glissante (et aussi moyennés sur l'ensemble) pour reproduire les données expérimentales.

Une description structurale de la protéine Tau

7

La protéine Tau est une des plus longues protéines intrinsèquement désordonnées étudiées par RMN à ce jour. L'étude de cette protéine est tout aussi importante d'un point de vue méthodologique que d'un point de vue biologique. Cette protéine présente dans les neurones est notamment connue dans sa forme pathogène comme un des marqueurs clés de la maladie d'Alzheimer. Ce chapitre est divisé en 3 parties dans lesquelles nous caractériserons plusieurs formes de la protéine Tau.

La première partie commencera par une présentation des mécanismes aboutissant à la mort neuronale dans le cadre de la maladie d'Alzheimer afin de mettre en perspective le rôle de cette protéine. Nous présenterons ensuite les travaux passés pour obtenir une description moléculaire de la protéine dans son état physiologique.

Dans une seconde partie, utilisant la relaxation paramagnétique, nous identifierons un contact à longue portée présent à la fois dans la forme native et dans une forme pseudophosphorylée. Nous insisterons sur la méthode et les validations croisées effectuées pour s'assurer de la robustesse de la description.

Dans un dernier temps, nous appliquerons le protocole du chapitre 5 pour caractériser l'échantillonnage conformationnel de la protéine Tau native, nous effectuerons alors une série de validations croisées pour estimer la précision de notre approche. L'approche présentée montrera en particulier comment il est possible d'utiliser la sensibilité et l'interconnectivité des paramètres RMN sur un système complexe comme la protéine Tau.

7.1 Contexte

7.1.1 Enjeux et motivations

Introduction extrait du rapport "World Alzheimer Report 2011" :

- La démence est un syndrome [...] qui affecte la mémoire, le raisonnement, le comportement et la capacité d'effectuer les activités de la vie quotidienne. La maladie d'Alzheimer représente la forme de démence la plus répandue. Les autres formes comprennent la démence vasculaire, la démence à corps de Lewy et la démence frontotemporale.

- La démence affecte surtout les personnes âgées, bien qu'elle puisse apparaitre avant 65 ans, après cet age, la probabilité de développer une démence double tous les cinq ans environ.

- Dans le Rapport mondial Alzheimer de 2009, *Alzheimer's Disease International* estimait qu'en 2010, 35.6 millions de personnes vivaient avec une démence dans le monde, et ce chiffre passera à 65.7 millions en 2030 et 115.4 millions en 2050.

- La démence a des conséquences au niveau émotionnel, financier et social pour les personnes atteintes, leurs familles et leurs amis. Une bonne compréhension des coûts sociaux de la démence, de leur impact sur les familles, le système de santé, la société et les gouvernements pourrait contribuer à résoudre ce problème.

7.1.2 La maladie d'Alzheimer

"La maladie d'Alzheimer est une maladie neurodégénérative incurable du tissu cérébral qui entraîne la perte progressive et irréversible des fonctions mentales et notamment de la mémoire." Elle fut initialement décrite par le médecin allemand Alois Alzheimer (1864-1915). Les causes exactes de la maladie d'Alzheimer restent encore mal connues.

Les deux types de lésions

Le cerveau du patient est victime d'un double processus de dégénérescence et d'inflammation, caractérisé par deux types de lésions locales :

- Les plaques amyloïdes

- Les enchevêtrements neurofibrillaires

Ces lésions locales se diffusent dans le cerveau et entrainent une perte progressive des facultés cognitives. A l'heure actuelle, le diagnostic de la maladie repose sur une évaluation détaillée des facultés cognitives du patient. Cependant, seule l'autopsie avec la mise en évidence des différentes lésions au sein du cerveau permet de diagnostiquer avec certitude la maladie d'Alzheimer. Ce diagnostic repose sur un examen anatomopathologique du cerveau en étudiant la topologie et la progression des lésions considérées. En effet, des études post-mortem ont montré une corrélation entre la propagation des lésions au sein du cerveau et la perte progressive des facultés cognitives. Ces études ont permis à la communauté médicale de proposer un modèle résumant la maladie en

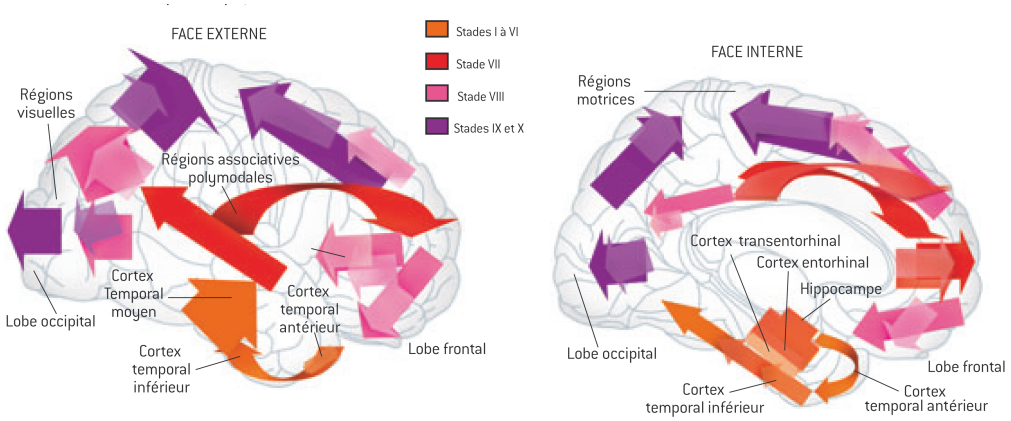


FIGURE 7.1 – ***Propagation des lésions dans le cerveau.*** *Les lésions se propagent des régions inférieures vers les régions dites primaires, qui reçoivent les informations sensorielles ou motrices. La figure extraite de [152]*

10 stades présentés en figure 7.1. Historiquement, la communauté scientifique s'est principalement focalisée sur l'étude des plaques amyloïdes comme marqueurs de la maladie d'Alzheimer. Courant des années 1990, au vu des travaux effectués dans ce domaine, la communauté médicale a révisé ces critères en incluant aussi la dégénérescence neurofibrillaire dans le diagnostic de la maladie. Le modèle actuel de la maladie suggère une forte corrélation entre les deux processus menant à la formation des deux agrégats : plaques amyloïdes et enchevêtrements neurofibrillaires.

7.1.3 Les plaques amyloïdes

Les plaques amyloïdes sont dues à l'agrégation de la protéine β-amyloïdes à l'extérieur des cellules neuronales. Ces plaques ont été identifiées comme des marqueurs de la maladie d'Alzheimer. Les études concernant l'agrégation en fibrilles de la protéine β-amyoïdes sont d'actualité en RMN, nous indiquons à titre d'indicatif quelques publications récentes [59, 60] mais nous ne développerons pas ce sujet plus amplement, nous nous focaliserons sur la protéine Tau et les enchevêtrements neurofibrillaires.

7.1.4 Les enchevêtrements neurofibrillaires

La dégénérescence neurofibrillaire est marquée par la présence d'enchevêtrements fibrillaires. Le composant majeur de ces derniers est la protéine Tau. Cette protéine est abondamment présente dans les neurones et très soluble, elle participe notamment à la stabilisation des microtubules[1]. On n'ignore encore les mécanismes déclenchant la pathogénicité de la protéine mais il a été mis en évidence les caractéristiques suivantes : la protéine Tau devient hyperphosphorylée et agrégée sous forme d'enchevêtrements fibrillaires insolubles. La contribution de cette protéine sous sa forme pathogène à la mort des cellules neuronales est suffisamment significative pour souligner la nécessité de comprendre en détail le rôle de cette protéine dans sa forme physiologique et par la suite déterminer les mécanismes pouvant mener à l'agrégation [16].

7.1.5 Les mécanismes aboutissant à la mort neuronale

La protéine Tau dans son état physiologique participe à la stabilisation des microtubules au sein des neurones en agissant sur la polymérisation ou la dépolymérisation de ces derniers. La protéine interagit avec la tubuline et favorise la polymérisation de

1. Les microtubules sont des fibres constitutives du cytosquelette de la cellule

la tubuline en microtubule. Ce mécanisme est régulé par phosphorylation. Notons que la phosphorylation de Tau agit sur la régulation de l'assemblage des microtubules mais aussi sur la croissance des neurites et le transport au sein des axones.

Dans son état physiologique, dans un neurone, il existe un juste équilibre entre les mécanismes de phosphorylation et de déphosphorylation de la protéine Tau. Bien qu'étant encore inconnus, un facteur, ou de multiples facteurs, peuvent rompre cet équilibre et favoriser l'hyperphosphorylation de la protéine. Cela se traduit par la perte de ses fonctions biologiques. De manière générale, plus la protéine Tau est phosphorylée, moins elle agit avec le microtubule. Un excès de phosphorylation peut même déstabiliser le microtubule. Les protéines Tau hyperphosphylées s'agrègent progressivement sous forme de filaments pathologènes qui deviennent eux-mêmes ce que l'on appelle des enchevêtrements fibrillaires, ces derniers contribuent à la mort des neurones [23, 24, 29].

La chronologie exacte des ces évènements reste hypothétique, cependant des chercheurs ont pu simultanément observer l'accumulation progressive de protéines Tau sous forme de filaments pathogènes et d'autre part l'hyperphophorylation de ces protéines. L'étude des mécanismes de phosphorylation est particulièrement délicate car la protéine Tau contient de nombreux sites de phosphorylation. On compte ainsi 85 résidus potentiellement phosphorylisables (45 serines, 35 thréonines et 5 tyrosines). Tout l'enjeu consiste alors à identifier le rôle de chaque site et les kinases associées afin de comprendre le mécanisme global.

7.1.6 Vers une approche thérapeutique

Les perspectives thérapeutiques se sont récemment focalisées sur la régulation de la phosphorylation de la protéine Tau. L'hypothèse triviale étant qu'en régulant, *i.e.* diminuant, la phosphorylation de la protéine il serait possible de ralentir la propagation de maladie. Cette piste est cependant discutable, une inhibition complète de la phosphorylation pourrait aussi avoir des conséquences néfastes sur le fonctionnement du neurone.

Le premier point consiste à identifier les sites de phosphorylation actifs et les kinases associées. Ces dernières peuvent être divisées en deux groupes :

- Les kinases qui phosphorylent les motifs Ser-Pro et Thr-Pro. Cela comprend la *Glycogen Synthase Kinase*-3-*Beta* (GSK3B), la *Cyclin-dependent Kinase* 5 (CDK5) et des kinases activées par un stress tel que la *c-Jun N-terminal kinase* (JNK) et les kinases de la famille p38 *mitogen-activated kinase*.

- Les sites Sérines et Thréonines non suivis d'une Proline. Cela inclus les protéines kinases A (PKA) et C (PKC), et les kinases *calcium calmodulin-dependent kinase II* (CaM kinases II) [153].

De nombreux sites ont pu être identifiés (figure 7.2) mais il faut supposer l'existence de sites de phosphorylation supplémentaires dans les conditions physiologiques du cerveau humain par rapport à ceux présents lors d'étude post-mortem car la protéine Tau extraite d'une biopsie de tissu devient rapidement déphosphorylée après excision. Les sites de phosphorylation de Tau dans le cerveau semblent avoir un cycle rapide de processus de phosphorylation et déphosphorylation.

Différentes stratégies thérapeutiques sont envisagées : les plus nombreuses

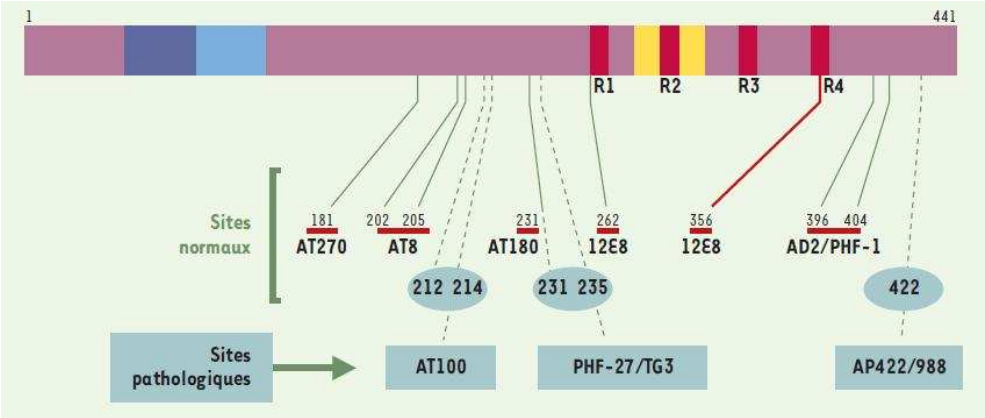


FIGURE 7.2 – ***Les sites potentiels de phosphorylation de la protéine Tau.*** *Il existe* 80 *sites potentiels de phosphorylation (Ser ou Thr) sur la protéine Tau. Certains sont des sites dits "normaux" de phosphorylation reconnus par des anticorps anti-tau dépendant de la phosphorylation. La phosphorylation régule les interactions avec les microtubules. Notamment, les sites reconnus par l'anticorps* 12*E*8 *sont cruciaux dans la liaison de Tau aux microtubules. Les sites dits "pathologiques" sont ceux de la phosphorylation anormale des protéines Tau qui caractérisent les tauopathies. Trois ont été identifiés et reconnus par les anticorps AT*100*, PHF-*27*/TG*3 *et AP*422*/*988*. La numérotation des acides aminés reconnus est celle de l'isoforme la plus longue (*441 *acides aminés). Figure extraite de [154].*

consistent à inhiber indépendamment des kinases spécifiques afin de bloquer le processus de cascade qui implique la phosphorylation. L'hypothèse sous-jacente étant l'existence d'une hiérarchie des différentes kinases, certaines kinases pourraient avoir un rôle primordial dans ce processus. Il a ainsi été mis en évidence le rôle prépondérant de CDK1 qui agit sur CDK5 qui agit à son tour sur la kinase GSK3B. Ces modèles n'ont cependant pas la possibilité d'être validés in vivo. Une stratégie alternative à l'inhibition spécifique d'une kinase serait de cibler une multitude de kinases afin de faire diminuer le niveau de phosphorylation.

7.1.7 Séquence de la protéine Tau

Du point de vue de la séquence, Tau se divise en 4 domaines :

- La domaine N-terminal (résidus 1 à 150) possède 0, 1, ou 2 exons (0N, 1N, 2N).

- Le domaine riche en Proline (résidus 150 à 240) et positivement chargé contrôlant indirectement l'association de Tau aux microtubules via des processus de phosphorylation.

- Le domaine d'appariement de Tau aux microtubules (résidus 240 à 370) formés de 3 ou 4 régions (3R, 4R) de 13 ou 14 acides aminés possédant des séquences très similaires. Ces régions sont par ailleurs séparées par un motif caractéristique Pro-Gly-Gly-Gly.

- Le domaine C-terminal (résidus 370 à 440) contenant des régions acides et basiques qui contrôle indirectement l'appariement de Tau aux microtubules via des processus de phosphorylation

Dans le système nerveux humain, Tau est présente sous six isoformes qui diffèrent par l'inclusion d'exons près de l'extrémité N-terminale et la présence de trois ou quatre régions identiques correspondant aux zones d'interaction avec le microtubule dans la moitié de l'extrémité C-terminale de Tau. Ces isoformes sont communément appelés 3R0N, 3R1N, 3R2N, 4R0N, 4R1N et 4R2N la nomenclature correspondant aux nombres

de régions présentes. Il convient de noter que Tau contient cinq acides aminés (Glycine, Lysine, Proline, Serine et Thréonine) représentant la moitié de sa séquence.

7.1.8 Un résumé des mécanismes connus

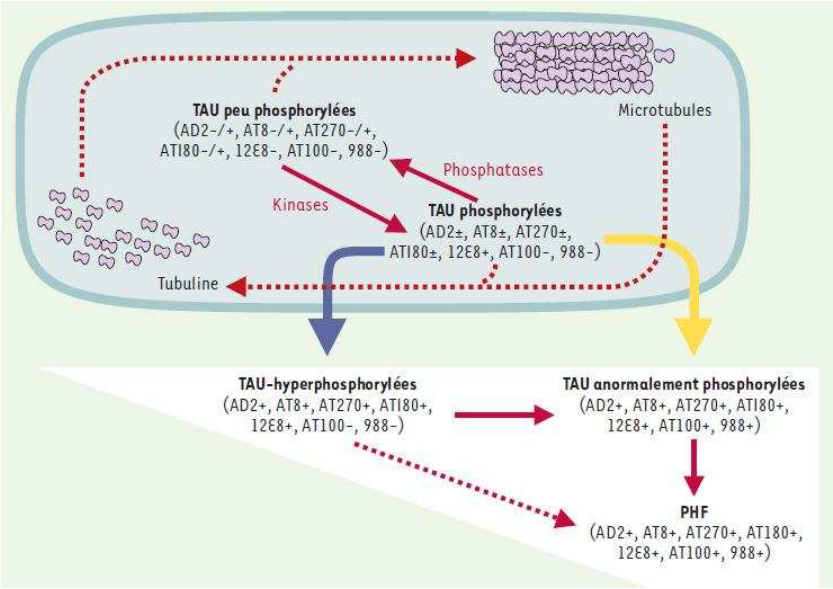


FIGURE 7.3 – ***Schéma récapitulatif des mécanismes associés à Tau.*** *La dynamique des microtubules (équilibre tubuline-microtubules) est assurée par l'échange entre les formes de Tau peu ou non phosphorylées et les formes phosphorylées. Figure extraite de [154].*

Dans un neurone, la dynamique des microtubules (équilibre tubuline-microtubules) est assurée par l'échange entre les formes de Tau peu ou non phosphorylées et les formes phosphorylées. Les protéines Tau présentent donc une immunoréactivité différente pour les anticorps décrits sur la figure précédente selon son degré de phosphorylation. Dans un neurone en dégénérescence neurofibrillaire, il y a hyperphosphorylation et ou phosphorylation anormale des six isoformes de protéine Tau et augmentation de l'immunoréactivité des protéines tau pour certains anticorps et/ou apparition de nouveaux épitopes (reconnus par exemple par les anticorps AT100) (figure 7.3).

7.2 Description moléculaire de la protéine Tau

7.2.1 Attribution de la protéine

La protéine Tau est une des plus longues protéines désordonnées étudiées à ce jour. Une étude dite structurale ne peut donc être réalisée par diffraction aux rayons X, il est nécessaire de recourir à la RMN ou au SAXS afin d'obtenir une description à l'échelle atomique de cette protéine.

Les difficultés rencontrées sont nombreuses, le premier enjeu est l'attribution d'une protéine de 441 acides aminés dont cinq acides aminés représentent environ la moitié de la séquence complète. Cette attribution a été progressivement réalisée en étudiant des troncatures de la protéine Tau : K18 (allant du résidu 243 au résidu 370) et K32 (allant du résidu 201 au résidu 395). Ces troncatures, comprises au niveau du domaine d'interaction de la protéine Tau avec les microtubules, ont été attribuées sans ambiguïté à l'exception près des motifs $^{270}PGGG^{273}$, $^{301}PGGG^{304}$, $^{332}PGGG^{335}$, $^{364}PGGG^{367}$ délimitant respectivement les répétions R1, R2, R3, R4. D'autres stratégies, comme l'utilisation de séquence hétéronucléaire à cinq ou sept dimensions [155], ont été appliquées sur deux isoformes de Tau : hTau23 (352 acides aminés), hTau24 (383 acides aminés) et ont permis d'attribuer l'isoforme le plus long hTau40 avec un ratio de 93% de résidus attribués dans des conditions expérimentales de 25°C et de pH 6.0.

7.2.2 Information provenant des déplacements chimiques secondaires

L'analyse des déplacements chimiques secondaires confirme la nature désordonnée de la protéine Tau. On distingue cependant quelques régions transitoirement structurées :

- Nous identifions deux régions échantillonnant la région hélicoïdale : le motif $^{114}LEDEAAGHVT^{123}$ situé entre l'exone 2 et la région riche en Proline P2 et le motif $^{428}LADEVSASLA^{437}$ situé dans l'extrémité C-terminale. L'analyse quantitative des déplacements chimiques secondaires $^{13}C^{\alpha}$, $^{13}C'$ suggère respectivement une propension de 18% et 25% de population hélicoïdale (Les déplacements chimiques secondaires de Tau complet seront affichés en figure 7.14.).

- Nous notons la présence de motifs PPII dans la région des répétitions. Ces motifs sont dans la littérature identifiés comme des feuillets β cependant l'étude de l'échantillonnage local de la construction K18 au chapitre 5 invalide ce résultat : nous avons montré qu'ils échantillonnaient la région PPII. Nous pouvons identifier trois régions situées respectivement dans les répétitions R2, R3, R4 : le motif $^{275}VQIINK^{280}$, le motif $^{306}VQIVYK^{311}$ et le motif $^{336}QVEVKSEKLD^{345}$. L'analyse quantitative des déplacements chimiques $^{13}C^{\alpha}$ indique des populations respectives de 22%, 25% et 19% pour les résidus $^{275}VQIINK^{280}$, $^{306}VQIVYK^{311}$, $^{336}QVEVKSEKLD^{345}$ [145]. Nous pouvons noter que ces régions R1, R2, R3, R4 sont séparées par des motifs PGGG présentés plus haut qui délimitent l'extension des motifs PPII.

Concernant ces derniers motifs, l'analyse quantitative des déplacements chimiques $^{13}C'$ ou l'analyse conjointe des déplacements chimique $^{13}C^{\alpha}$ et $^{13}C'$ confirme les valeurs précédemment énoncées, et ceci, pour deux températures de mesures différentes (à 5 ou 25°C). Comme le montre la figure 7.4, ces déplacements secondaires sont faibles *i.e.* proches des valeurs *random-coil* mais reproduits quelle que soit la construction de Tau étudiée.

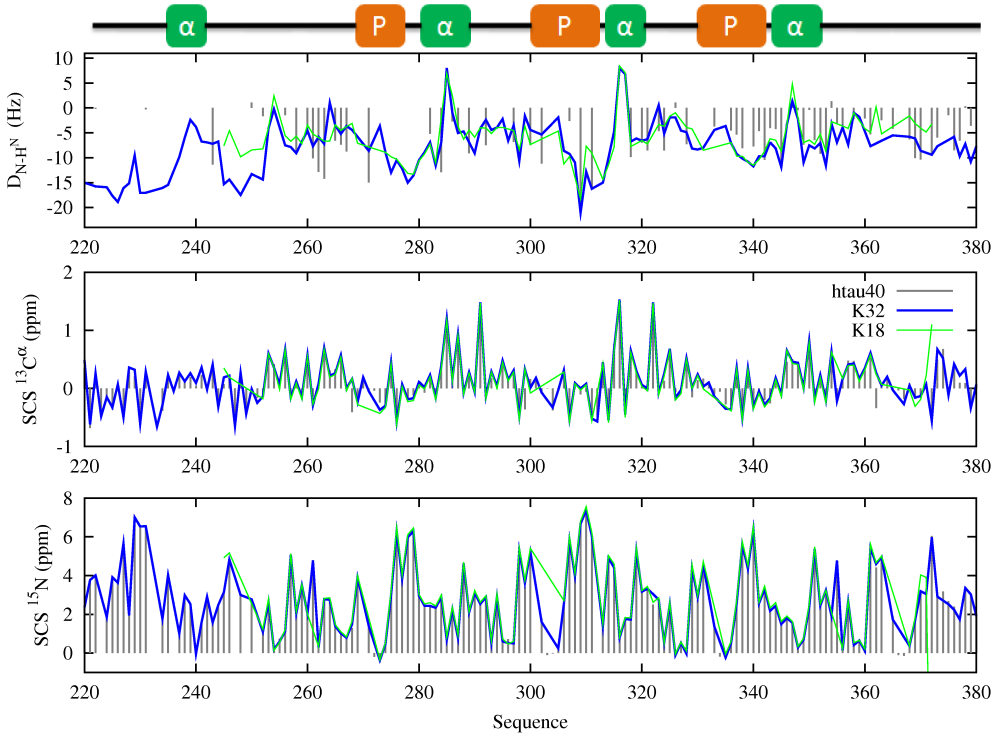


FIGURE 7.4 – ***CDRs D_{NH} et déplacements chimiques $^{13}C^{\alpha}$ et ^{15}N de K18, K32 et htau40.*** *Nous traçons htau40 en gris, K32 en bleu, K18 en vert avec en haut les CDRs D_{NH} suivi des déplacements chimiques secondaires $^{13}C^{\alpha}$ et ^{15}N en bas. Un facteur multiplicatif est utilisé pour ajuster les CDRs.*

7.2.3 Information provenant des couplages dipolaires résiduels

Dans un deuxième temps, nous nous intéressons aux CDRs D_{NH} mesurés sur les troncatures suivantes : K18, K32 et sur l'isoforme de Tau le plus long hTau40. Les valeurs majoritairement négatives traduisent la flexibilité du squelette d'une protéine désordonnée. La valeur des couplages est plus négative dans les régions ayant une plus forte propension pour les motifs PPII (figure 7.4).

Nous pouvons noter par ailleurs la présence de 4 motifs homologues $^{253}LK^{254}$, $^{285}LS^{285}$, $^{315}LS^{316}$, $^{346}LK^{347}$ possédant des couplages positifs. Cette inversion de signe des CDRs dans les protéines désordonnées est caractéristique de conformations échantillonnant des boucles ou la région hélicoïdale. La figure 7.4 compare les CDRs D_{NH} des constructions K18, K32 et htau40. Comme précédemment mentionnés pour les déplacements chimiques, les mesures sont similaires exception faite des effets liés aux lignes de base.

Afin d'étudier plus en détail l'échantillonnage conformationnel du domaine d'appariement de Tau aux microtubules en s'appuyant sur la signature des CDRs, deux approches ont été retenues : la première approche repose sur une description par ensemble de la protéine. Après création d'un ensemble de structure avec le logiciel FLEXIBLE-MECCANO, nous calculons pour chaque structure les valeurs des CDRs qui sont ensuite moyennées sur l'ensemble. Les valeurs simulées sont alors ajustées aux valeurs expérimentales à l'aide d'un facteur multiplicatif. Les différences entre les deux jeux de données traduisent les différences structurales non prises en compte par le modèle *random-coil*. La figure 4.3 montre que ces différences sont localisées dans les régions identifiées auparavant : au niveau des boucles et des motifs PPII.

Réalisée en 2007, une étude approfondie du lien entre l'échantillonnage conforma-

tionnel et la valeur des couplages résiduels dipolaires au niveau des boucles met en évidence l'impossibilité d'extraire l'échantillonnage conformationnel local de la protéine en se basant uniquement sur le couplage D_{NH}. La dynamique moléculaire accélérée a donc été utilisée pour déterminer l'échantillonnage conformationnel reproduisant les données expérimentales [74]. Une comparaison entre l'échantillonnage obtenu et celui issu de la description par ensemble montre de fortes similitudes dans leur distribution à quelques exceptions près. L'utilisation de cet échantillonnage comme base de données dans Flexible-Meccano reproduit correctement les données expérimentales. Cela traduit non seulement le bon caractère conformationnel mais aussi la bonne propension des conformations transitoirement structurées créées dans l'ensemble. Cette étude préliminaire combinant description par ensemble et dynamique moléculaire confirme les hypothèses structurales faites sur la protéine Tau.

7.2.4 Information provenant de la relaxation paramagnétique

Une étude est menée en 2009 pour caractériser le repliement et l'agrégation potentielle de la protéine Tau à partir de la relaxation paramagnétique. La protéine Tau possède deux cystéines (C291 et C322) qui permettent l'attachement d'un radical MTSL. Pour attacher des chaines latérales MTSL supplémentaires et obtenir de plus amples informations, cinq mutations d'Alanine en Cystéine furent réalisées. Les positions retenues étaient : A15C, A72C au niveau de l'extrémité N-terminale, A239C dans la région riche en Proline, et A384C et A416C au niveau de l'extrémité C-terminale. Les auteurs notèrent une diminution des profils d'intensités indiquant la présence d'un contact à longue-portée entre la région N-terminale et la région centrale de la protéine Tau dans son état natif.

Une première description par ensemble comprenant des données PRE fut ensuite réalisée en utilisant le programme X-PLOR. 2288 contraintes de distances furent utilisées dans une description comprenant 10 et 30 structures par ensemble [156]. Les caractéristiques biophysiques des ensembles comme le rayon de giration (65 Å) étaient dans la gamme proposée par des données issues de diffraction par rayon X aux petits angles **??**. Cette étude préliminaire offre une source d'information substantielle sur les interactions à longue portée de la protéine Tau mais doit cependant être approfondie par une validation des protocoles utilisés lors de la description par ensemble.

Durant cette partie nous nous focaliserons sur la méthode et les validations effectuées permettant de garantir la fiabilité de la description par ensemble. Le modèle doit répondre aux points suivants : le nombre de structures à utiliser, la validité du modèle physique, la stabilité de l'algorithme, la prise en compte de l'erreur expérimentale et l'influence de l'échantillonnage conformationnel sur la détermination d'un ensemble issu de contraintes à moyenne et longue distance.

7.3 Matériel et méthodes

7.3.1 Modélisation de la dynamique de la chaine latérale dans Flexible-Meccano

La dynamique de la chaine latérale MTSL a été calculée en échantillonnant 600 conformations pour chaque cystéine. La flexibilité de la chaine latérale est calculée en utilisant une sphère de rayon variable comme présentée en section 6.1. La largeur de raie du proton utilisé pour le calcul de l'intensité vaut 4 Hz (équation 2.14).

7.3.2 Sélection d'ensembles avec Asteroids

Le processus de sélection est identique à celui utilisé pour la protéine α-Synucléine. 4000 itérations sont effectuées avec 50 individus. 30000 structures sont créées pour générer le *pool*. Concernant le calcul du χ^2_{PRE}, le poids est identique pour chacune des cystéines.

7.3.3 Données expérimentales

Les données expérimentales ont été mesurées par Stefan Bibow. Les données comprennent onze jeux de données provenant de la mutation d'alanine en cystéine : A15C, A72C, A125C, A178C, A239C, V256C, 322C, S352C, A384C, et S416C ainsi que le double jeu comprenant les deux cystéines natives de Tau : 291C et 322C.

Deux formes de la protéine Tau sont étudiées : la forme native nommée Wt , la forme pseudo-phosphorylée nommée Emut, réalisée en mutant les acides aminés suivants en acide glutamique : S199E, S202E, T205E, T212E, S214E, S396E, S404E. Les épitopes mutées correspondent aux épitopes spécifiques de la forme hyperphosphorylée de la protéine déterminée par immunoréactivité.

Nous disposons aussi pour la forme native des déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N et $^{1}H^{N}$, des CDRs D_{NH}. Les CDRs sont mesurés à 5°C. Les déplacements chimiques sont mesurés à 5°C et 25°C, en concordance avec les bases de données de Sparta, nous référençons nos spectres sur ceux mesurés à 25°C.

7.3.4 Données simulées

Différents jeux de données simulées seront utilisés dans ce chapitre. Le premier concerne l'ajout d'un bruit gaussien aux données issues de la sélection. Il est réalisé à partir de la marge d'erreur des données expérimentales qui est alors la largeur à mi-hauteur du bruit gaussien.

La dernière partie combinant l'ensemble des paramètres RMN nécessite la réalisation de plusieurs simulations incluant soit la base de données *random-coil* soit la base de données d'angles dièdres issue de sélections selon les déplacements chimiques avec le logiciel Asteroids. Pour chacune, les couplages dipolaires résiduels sont calculés en utilisant le tenseur d'alignement global sur des ensembles de 150000 structures afin d'assurer la convergence. Lors des validations croisées de ces derniers, un facteur d'ajustement est optimisé de manière à reproduire au mieux les données expérimentales. Par ailleurs, les PRE sont calculés d'après le formalisme du Chapitre 6 avec un temps de corrélation global de 5 ns et un temps de corrélation interne de 500 ps.

7.4 Résultats

Nous allons appliquer le protocole de sélection aux données PRE de la forme native et pseudo-phosphorylée de la protéine Tau. Avant cela nous devons répondre à plusieurs points : le premier concerne l'influence de la dynamique de la chaine latérale MTSL sur les profils d'intensité, le second le nombre de structures nécessaires pour caractériser correctement la protéine Tau en utilisant des données de relaxation paramagnétique. Nous testerons la reproduction des données par validation croisée puis la reproduction de l'échantillonnage conformationnel et afin l'influence du bruit expérimentale, qui sera simulé par un bruit gaussien, sur la présence de contacts à

longue portée. A l'issu de ces tests, nous appliquerons notre approche aux deux formes de la protéine Tau.

La deuxième partie de cette section concerne la caractérisation de l'échantillonnage conformationnel de la protéine Tau selon les déplacements chimiques. Nous serons alors en mesure d'analyser les structurations transitoires de la protéine et regarderons l'amélioration de la reproduction des autres paramètres RMN mesurés : les couplages dipolaires D_{NH} et la relaxation paramagnétique.

7.4.1 Influence de la dynamique de la chaine latérale

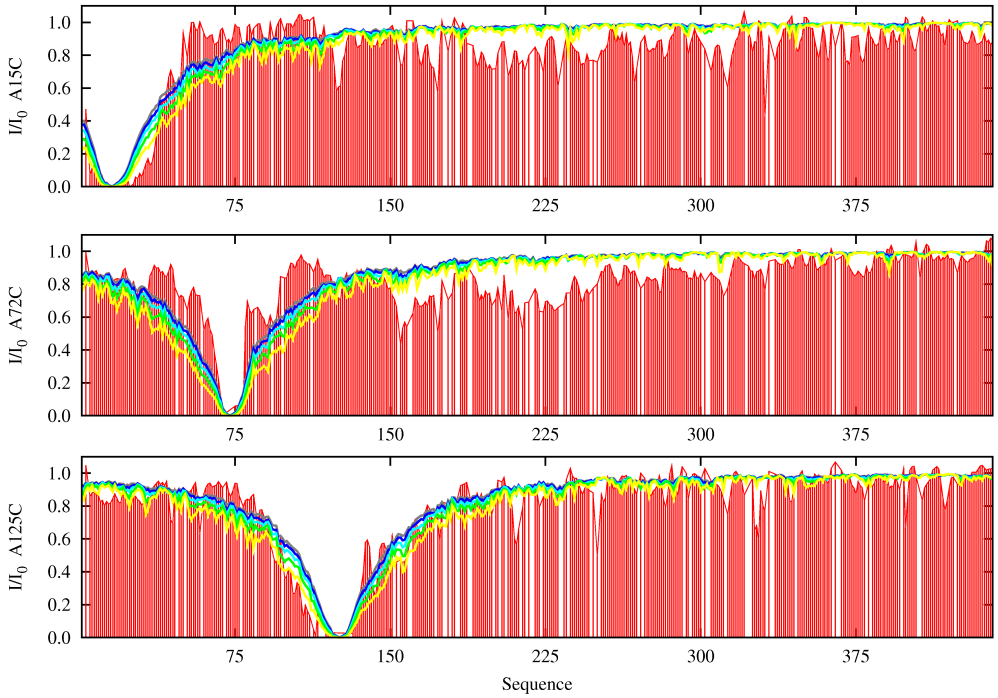


FIGURE 7.5 – ***Influence de la flexibilité de la chaine latérale MTSL sur les profils I/I_0.*** *Nous avons cinq degrés de flexibilité allant d'une chaine latérale faiblement dynamique $R_{max}/10$ (en jaune) à une chaine latérale dynamique $10R_{max}/10$ (en gris). Les données expérimentales sont tracées en rouges. Les cadres correspondent aux cystéines : A15C, A72C, A125C. R_{max} est la distance maximale électron-spin présenté en section 6.1.2*

Afin d'étudier la dynamique de la chaine latérale et son impact sur la reproduction des données, nous restreignons progressivement le rayon de la sphère dans laquelle la chaine latérale peut échantillonner ses conformations et calculons les rapports d'intensité correspondants. Les profils d'intensité expérimentaux et simulés sont affichés en figure 7.5. Concernant les données expérimentales, nous constatons quelques disparités entre les différentes cystéines : le profil d'intensité de la cystéine A72C est très étroit, celui de A125C est très étendu et asymétrique. Cela est probablement du à un échantillonnage local spécifique. Au vu des données simulées, la dynamique de la chaine latérale, n'influence les profils d'intensité qu'à proximité de la cystéine, au-delà 30 résidus nous n'observons pas de changement. Par conséquent, la caractérisation d'un contact à longue porté comme celui présent dans la région 150-225 n'est pas influencée par la dynamique de la chaine latérale MTSL.

Bien que l'influence de l'échantillonnage local ne soit pas négligeable, ce dernier ne peut modifier la reproduction des profils de PRE qu'aux alentours de la cystéine. Par

conséquent le choix le plus adapté aux données et en accord avec le chapitre 6 est le modèle avec le plus de dynamique[2].

7.4.2 Détermination du nombre de structures par validation croisée

Afin de déterminer le nombre de structures nécessaires, nous réalisons une série de sélections d'ensembles comprenant de 25 à 500 structures. Un des jeux de données correspondant à un tag paramagnétique est exclu de la sélection, c'est-à-dire, nous incluons dix jeux de données sur onze lors de la sélection, le jeu de données non incluse, ou passive, est utilisé comme validation croisée.

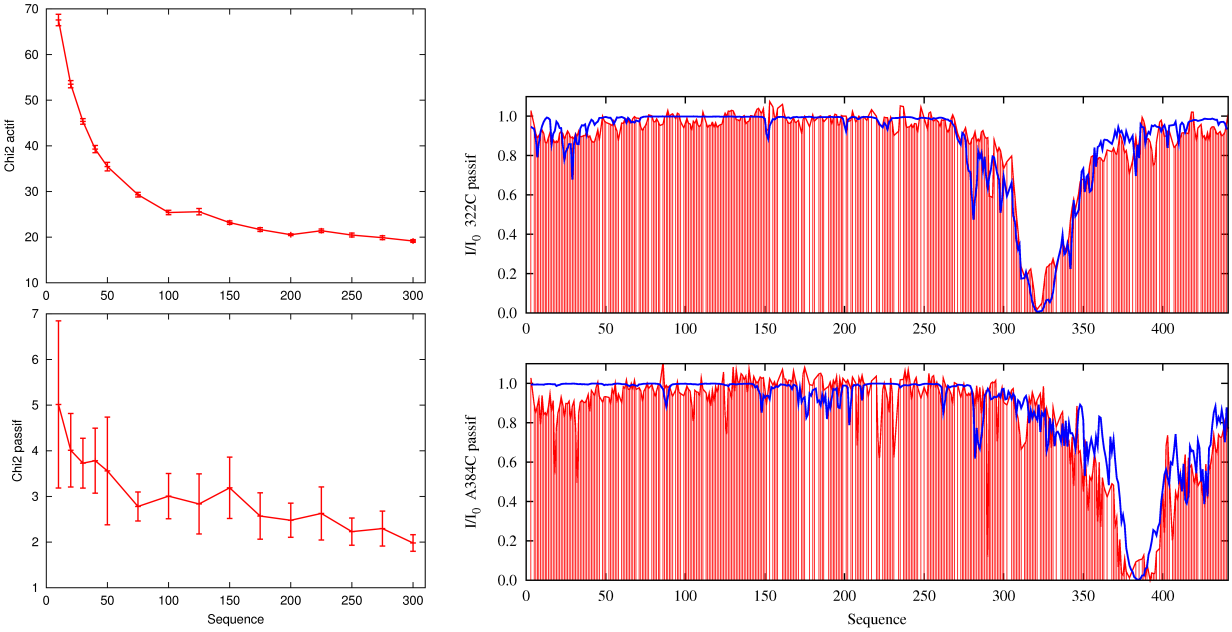


FIGURE 7.6 – ***Reproduction des données actives et passives : χ_2 et validations croisées des cystéines 322C et A384C.*** *9 sélections indépendantes sont réalisées pour chaque taille. À gauche sont affichés, l'évolution du χ^2_{actif} moyen (en haut), du χ^2_{passif} moyen(en bas) et des écarts type respectifs en fonction du nombre de structures inclues dans la sélection. À droite sont affichées les validations croisées correspondantes pour un ensemble de* 200 *structures sélectionnées avec* ASTEROIDS *(en bleu). Les données expérimentales sont en rouge.*

La figure 7.6 montre l'évolution du χ^2 actif et passif en fonction du nombre de structures. Après une variation rapide pour les ensembles de petite taille, les valeurs se stabilisent à partir de 150 structures à la fois pour le χ^2 actif et le χ^2 passif. Nous choisissons d'utiliser un ensemble de 200 structures en accord avec les résultats des chapitres précédents et pour garder un nombre de paramètres restreints pour caractériser l'ensemble.

Les validations croisées d'un sous-ensemble de 200 structures de la cystéine 322C et A384C sont aussi affichées en figure 7.6. La reproduction des données est très satisfaisante : les profils correspondent à ceux mesurés autour de la chaine latérale, pour la région N-terminale nous détectons bien la présence d'un contact pour la cystéine 322C. Pour la cystéine A384C la modulation du profil dans la région N-terminale ne semble

2. Dans la troisième partie du chapitre, nous justifierons ce choix : après prise en compte de l'échantillonnage local de la protéine Tau, la meilleure reproduction des profils I/I_0 nécessitera aussi une chaine latérale hautement flexible.

pas détectable. Ce désaccord s'explique par la nature plus éloignée du contact observé avec la cystéine A384C qui ne peut être complètement reproduit à partir des contraintes des autres cystéines. Nous devons souligner que nous caractérisons une protéine de 441 acides aminés avec 11 cystéines espacées régulièrement sur la séquence. Ce jeu de données est conséquent et a nécessité un lourd travail d'expression de protéines et de spectroscopie RMN, cependant des calculs *in-silico* ont montré la nécessité d'inclure un nombre de cystéines encore plus important pour pouvoir caractériser précisément l'ensemble des contacts à très longue portée (supérieurs à 300 résidus) transitoirement présents au sein d'une protéine de cette taille.

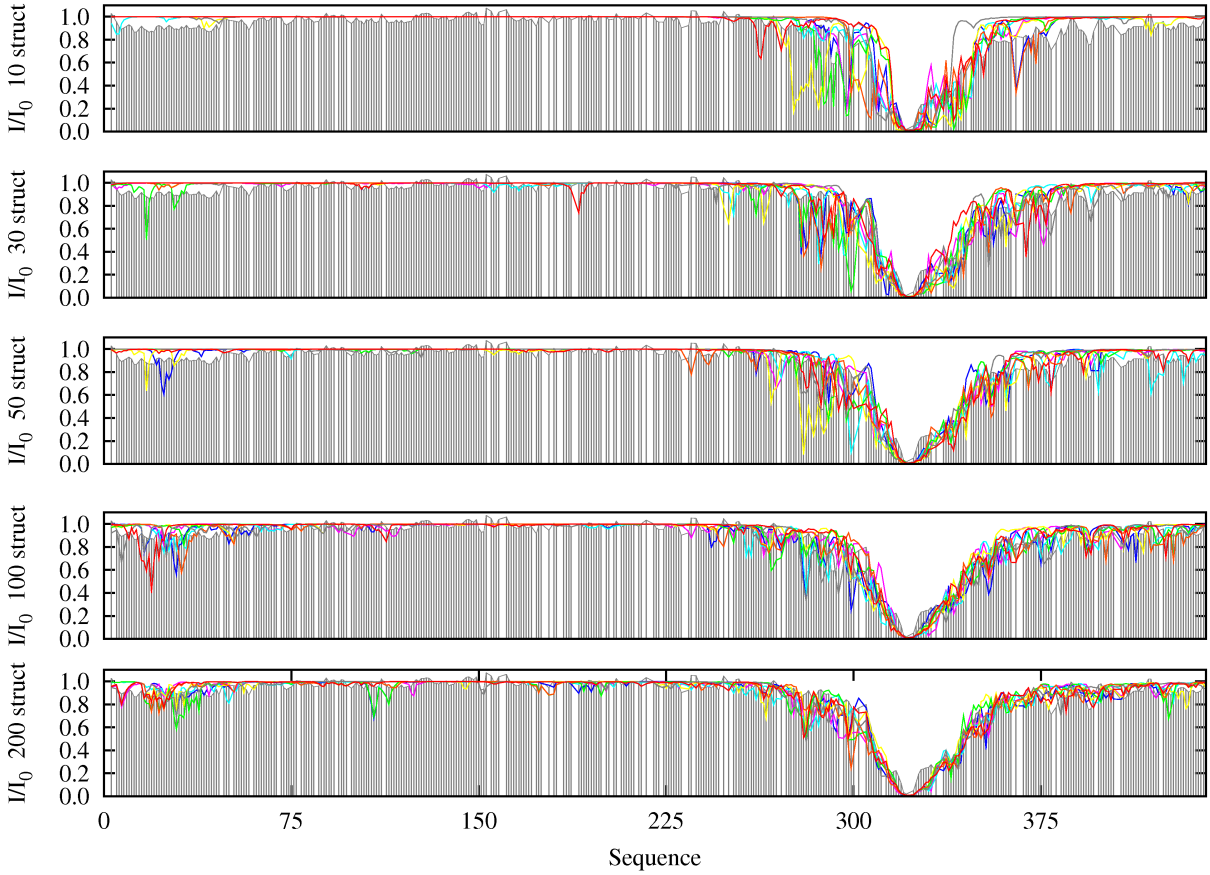


FIGURE 7.7 – ***Validation croisée de la cystéine 322C en fonction du nombre de structures incluses dans la sélection.*** *Les sélections sont réalisées avec* ASTEROIDS, *huit répétitions sont effectuées et tracées en couleur. Les données expérimentales sont en gris. La reproduction de contacts à longue portée nécessite au moins* 100 *structures.*

La figure 7.7 montre l'évolution de la reproduction des données de la validation croisée du mutant 322C en fonction du nombre de structures. dans chaque cas huit répétitions sont effectuées de manière à caractériser la convergence de l'approche. Les ensembles contenant trop peu de structures, de 10 à 100, ne reproduisent ni la présence de contact à longue portée : nous n'observons pas de diminution de l'intensité dans la région C-terminale, ni le profil d'intensité caractéristique environnant la cystéine : la reproduction de l'intensité du voisinage de la cystéine est très variable d'une sélection à l'autre. Nous pouvons en conclure que l'utilisation d'un nombre trop faible de structures ne permet pas de moyenner correctement les données et aboutit à une mauvaise description de la protéine. Un ensemble contenant 200 structures permet de reproduire les données actives et passives.

La validation croisée

La principale difficulté lors d'une telle description consiste à éviter ce qu'on appelle le sur-ajustement (*overfitting*) ou le sous-ajustement (*underfitting*). Pour cela l'utilisation de validation croisée permet de visualiser rapidement l'adéquation avec le modèle choisi.

Dans notre cas, la similarité entre la reproduction de données incluses dans la sélection et de données issue de validation croisée montre que notre description ne contredit pas les données mesurées. Ces simulations permettent en particulier de déterminer le nombre de structures nécessaires dans l'ensemble.

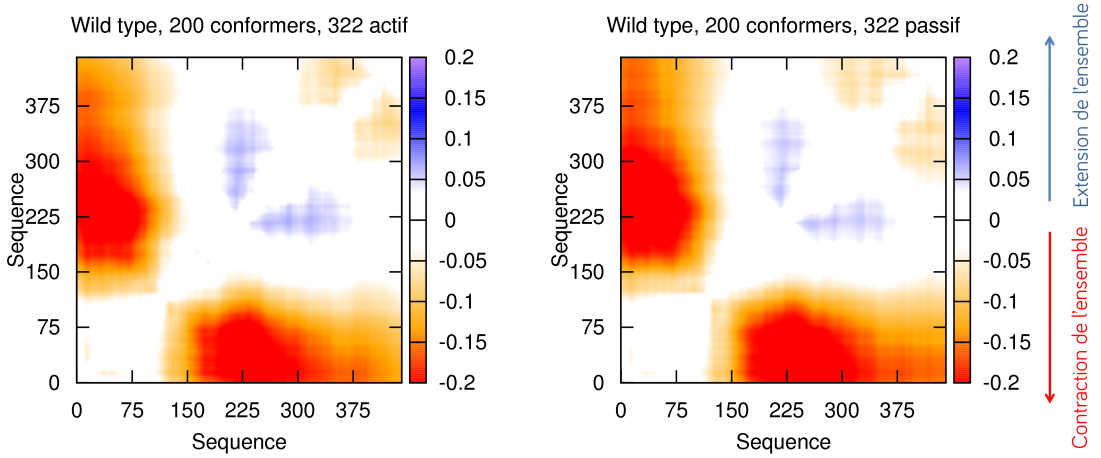


FIGURE 7.8 – ***Cartes de contacts issus de sélections avec* Asteroids *incluant ou non la cystéine* 322C.** *Chaque carte de contacts comprend huit répétitions qui sont normalisées avec un ensemble random-coil. La carte de contact à gauche provient de sélection avec toutes les cystéines, la carte de contacts à droite ne contient pas la cystéine* 322C.

La reproduction des données n'est cependant pas suffisante pour valider une description par ensemble, la forte dégénérescence des données nécessite la comparaison des paramètres biophysiques. La figure 7.8 présente les cartes de contact des ensembles sélectionnés, soit en incluant l'ensemble des données, soit en excluant le jeu de données correspondant à la cystéine 322C. Le fait d'enlever environ 10% des données n'influence pas la reproduction des distances moyennes de l'ensemble. La similitude entre les deux cartes est flagrante et la présence d'ordre à longue portée entre la région N-terminale et le centre de la protéine ne peut pas être remise en cause.

Pour évaluer le niveau de bruit lors de la validation croisée, nous comparons en figure 7.9 la différence entre l'utilisation du jeu de données 322C en tant que donnée active ou en tant que donnée passive, et ceci, pour huit sélections différentes de manière à vérifier de nouveau si le nombre de structures est suffisant et si la réponse fournie par ASTEROIDS est stable. Nous notons une très bonne reproduction des données dans les deux cas. Nous allons maintenant pouvoir appliquer notre protocole à l'ensemble des données expérimentales pour caractériser notre système.

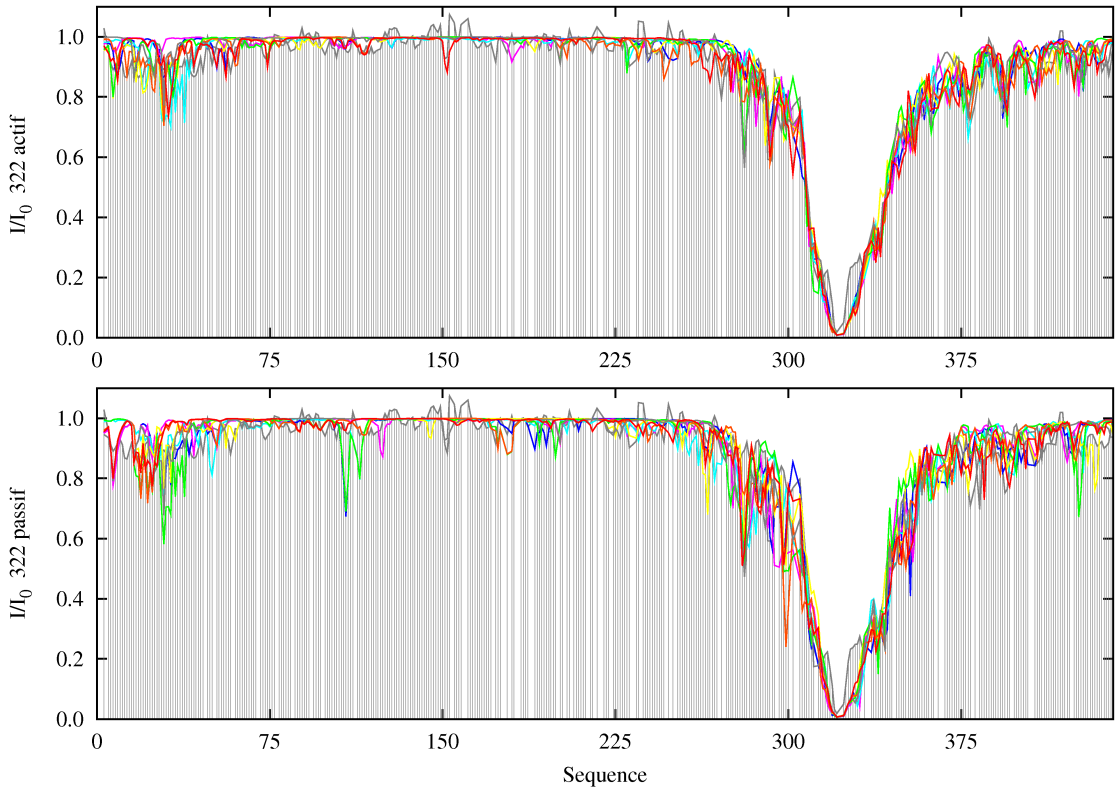


FIGURE 7.9 – ***Reproduction des données associées à la cystéine 322C dans le cas actif ou passif.*** *Si inclus dans la sélection, nous obtenons les courbes en haut, si les données sont passives, c'est-à-dire en validation croisée, les données sont (en bas). Les couleurs vives correspondent à huit répétitions indépendantes issues des sélections avec* ASTEROIDS, *les données expérimentales sont en gris.*

7.4.3 Application aux données complètes de la forme native et pseudo-phosphorylée

Le protocole est appliqué aux jeux de données complets de la forme native et pseudo-phosphorylée en sélectionnant 200 structures pour chaque jeu de données. Les figures 7.10 et 7.11 montrent respectivement la reproduction des données des onze tags paramagnétiques de la forme native et de la forme pseudo-phosphorylée. La reproduction des profils I/I_0 est excellente dans les deux cas.

FIGURE 7.10 – ***Profils I/I_0 des*** 11 ***cystéines de la forme native à l'issu de la sélection avec* Asteroids.** *Les données expérimentales (en rouge), et les données issues de la sélection (en bleu) pour les cystéines suivantes : A*15C, *A*72C *A*125C, *A*178C, *A*239C, *A*256C, 291C-322C, 322C, *A*384C, *A*416C.

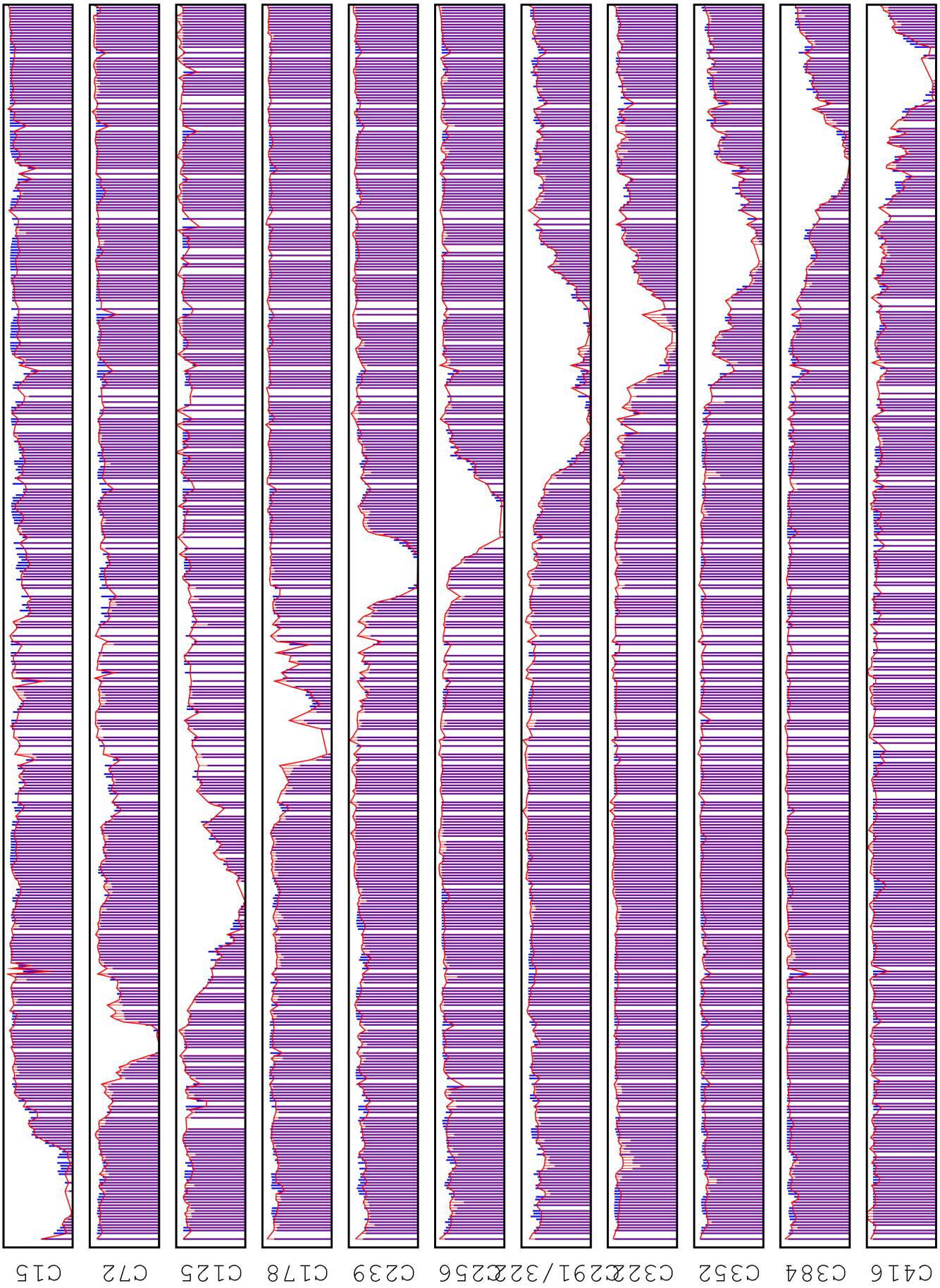


FIGURE 7.11 – ***Profils I/I_0 des* 11 *cystéines de la forme pseudo-phosphorylée à l'issu de la sélection avec* Asteroids.** *Les données expérimentales (en rouge), et les données issues de la sélection (en bleu) pour les cystéines suivantes : A*15*C, A*72*C A*125*C, A*178*C, A*239*C, A*256*C, 291C-322*C*, 322*C*, *A*384*C*, *A*416*C*.

La figure 7.12 montre les cartes de contacts de la forme native et de la forme pseudo-phosphorylée résultante de la sélection des ensembles avec Asteroids [157].

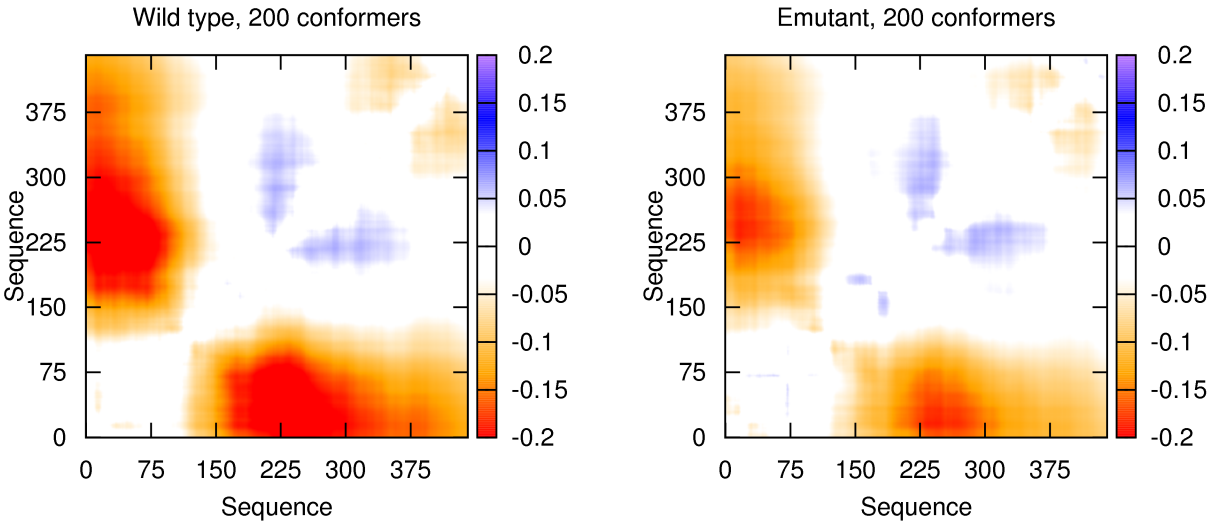


Figure 7.12 – ***Carte de contacts de la forme native et pseudo-phosphorylée de Tau.*** *Les cartes de contacts sont chacune calculées à partir de huit sélections indépendantes réalisées avec* Asteroids *puis normalisées par rapport à l'ensemble random-coil correspondant.*

Native vs phosphorylée

La forme native possède un contact à longue portée entre la région N-terminale [1 : 80] et la région centrale de la protéine [160 : 240]. Nous notons une légère extension de l'ensemble dans la région $225 - 350$ ainsi qu'une légère contraction de l'ensemble dans la partie N-terminale. La forme pseudo-phosphorylée possède ces mêmes caractéristiques mais le contact à longue portée est nettement moins marqué.

La pseudo-phosphorylation modifie la distribution de charge de la protéine Tau et diminue l'ordre résiduel à longue portée existant dans la structure native. La région centrale de Tau est globalement positivement chargée, à l'opposition de la partie C-terminale et la partie contenant les 120 premiers résidus du N-terminale qui présentent un excès de charge négative. L'introduction de 5 acides glutamiques dans la région $199 - 214$ change complètement la distribution de charge de la région riche en proline nommée P2. L'altération des propriétés électrostatiques est due à une diminution de l'interaction coulombienne entre le région P2 et l'extrémité N-terminale positivement chargée.

7.4.4 Sensibilité aux imprécisions expérimentales

La présence de bruit gaussien n'influence pas les caractéristiques majeures obtenues : de légères modulations des distances moyennes entre résidus sont observées mais la figure 7.13 montre la même caractérisation des formes natives Wt et pseudo-phosphorylée Em en présence ou non de données cibles bruitées.

7.4.5 Avant propos

Nous allons par la suite appliquer l'ensemble des méthodes présentées afin de caractériser le plus précisément possible la protéine Tau. Nous souhaitons combiner

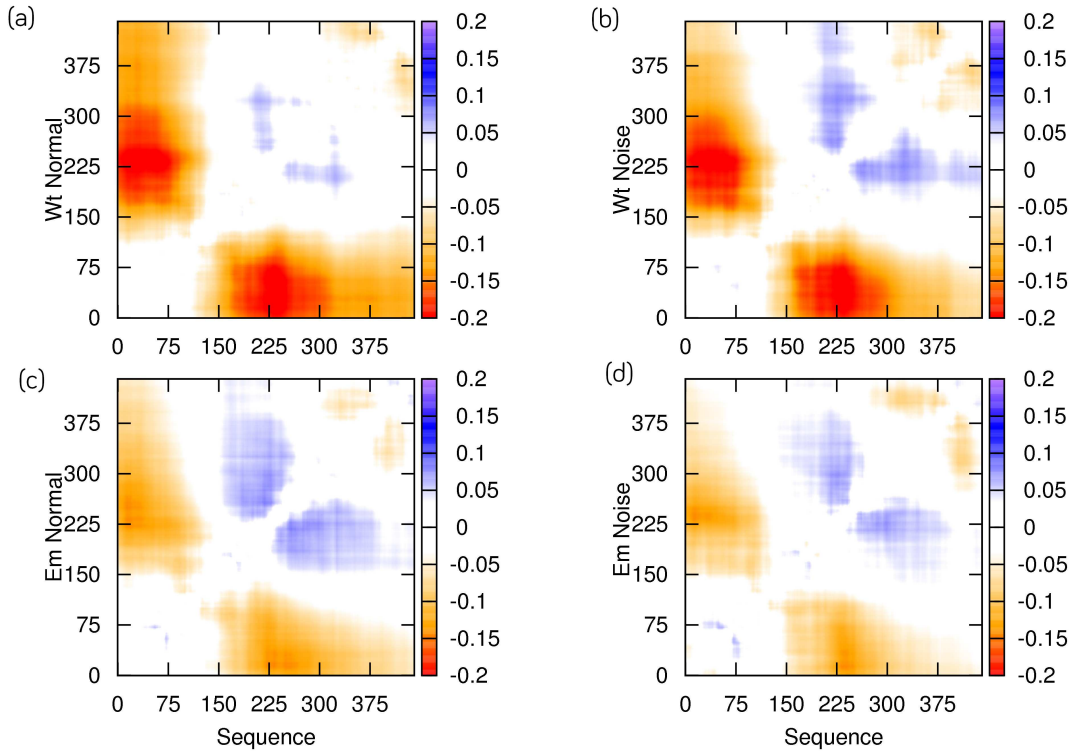


FIGURE 7.13 – ***Carte de contact après perturbation des données cibles avec un bruit gaussien.*** *Nous avons en haut la forme native (a)(b) et en bas la forme pseudo-phosphorylée (c)(d). Les cartes de contact sont issues de sélections réalisées avec* ASTEROIDS *soit avec les données cibles standard (à gauche), soit des données cibles perturbées avec un bruit gaussien de largeur* 0.016 *(à droite).*

dans notre approche le maximum de paramètres RMN disponibles. La solution la plus simple consiste vraisemblablement à reproduire conjointement l'ensemble de ces données lors de la sélection d'un ensemble de structures. Dans un premier temps, nous n'envisagerons pas cette solution, nous souhaitons proposer une méthode alternative à ce qui est généralement suggéré : nous allons étudier l'interconnectivité des données. Nous chercherons à chaque étape à souligner la rigueur de l'approche en insistant notamment sur l'importance de la validation croisée.

7.4.6 Philosophie de l'approche

Nous avons pu à plusieurs reprises souligner le point suivant : la reproduction des données ne garantit pas forcément la validité de l'approche. En particulier lors des sélections, la forte dégénérescence du système ne peut pas toujours être levée en raison du nombre de données expérimentales disponibles. Ainsi, au lieu de chercher à déterminer le meilleur ensemble reproduisant les données après sélection, nous allons déterminer une par une les caractéristiques de l'ensemble puis regarder leur influence sur la reproduction des autres données expérimentales. Nous utiliserons à chaque fois comme référence un ensemble *random-coil*.

L'approche suggérée

L'approche que nous suggérons consiste plutôt à réaliser un minimum d'hypothèses sur l'ensemble qui permettent de reproduire l'ensemble des données. Une fois ce point acquis, il est alors possible et convénient de sélectionner un ensemble reproduisant tous les paramètres RMN disponibles afin d'approcher au mieux les caractéristiques de la protéine étudiée.

Nous commencerons par déterminer l'échantillonnage conformationnel de la protéine Tau native, pour cela nous sélections un sous-ensemble à partir des données $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N et $^{1}H^{N}$ selon le protocole présenté au chapitre 5. Pour déterminer la véracité de cette solution, deux validations sont réalisées : la validation croisée des CDRs D_{NH} et la validation croisée des PREs. Les CDRs étant sensibles à l'échantillonnage local de la protéine, nous devons être en mesure de reproduire les couplages dipolaires D_{NH}. Le second cas est plus subtil, l'échantillonnage local modifiant ponctuellement le profil de la relaxation paramagnétique, nous attendons à proximité de la cystéine une amélioration de la reproduction des données.

7.4.7 Détermination de l'échantillonnage conformationnel de Tau

La détermination de l'échantillonnage conformationnel de la protéine Tau reprend les protocoles présentés précédemment pour la protéine N_{tail}, K18. Cinq itérations sont effectuées, pour chacune 5*200 structures sont sélectionnées puis utilisées pour recréer une nouvelle base de données d'angle dièdres et générer 18000 structures pour une nouvelle sélection.

La reproduction des déplacements chimiques en figure 7.14 est très bonne, nous montrons l'échantillonnage conformationnel correspondant en figure 7.15. Pour plus de lisibilité, la population est moyennée sur résidus i, $i-1$ et $i+1$, et ceci, pour chaque cadre. Nous lissons donc légèrement les spécificités de la séquence pour visualiser plus clairement les modifications conformationnelles. Nous sommes en mesure d'identifier les caractéristiques de Tau mentionnées dans les précédentes publications [145, 74, 156] :

en gris : La région d'appariement de la protéine aux microtubules se distingue par la présence d'une suite de résidus échantillonnant en moyenne jusqu'à 50% la région βP. Nous citons notamment les résidus $^{255}NVKSKI^{260}$, $^{275}VQIINK^{280}$, $^{306}VQIVYK^{311}$ et $^{336}QVEVKSEKLD^{345}$.

en vert : Les boucles entre ces motifs sont présentes en position $^{253}LK^{254}$, $^{285}LS^{285}$, $^{315}LS^{316}$ et $^{346}LK^{347}$ et échantillonnent jusqu'à 50% la région αR.

en bleu : Nous notons la présence de deux hélices transitoires de propension de 60% et et 70% situés respectivement dans la partie N-terminale en position $^{114}LEDEAAGHVT^{123}$ et dans l'extrémité C-terminale en position $^{429}ADEVSASLAXXX^{440}$.

en rose : Le domaine riche en Proline échantillonne fortement la région βP entre les résidus $^{175}TPPAPKTPPSSGEPPKS^{191}$ et $^{213}PSLPTPPTREPKKVAVVRTPPKSP^{236}$. Cette augmentation de la population dans la région βP est très significative pour les Prolines et les résidus précédents les Prolines.

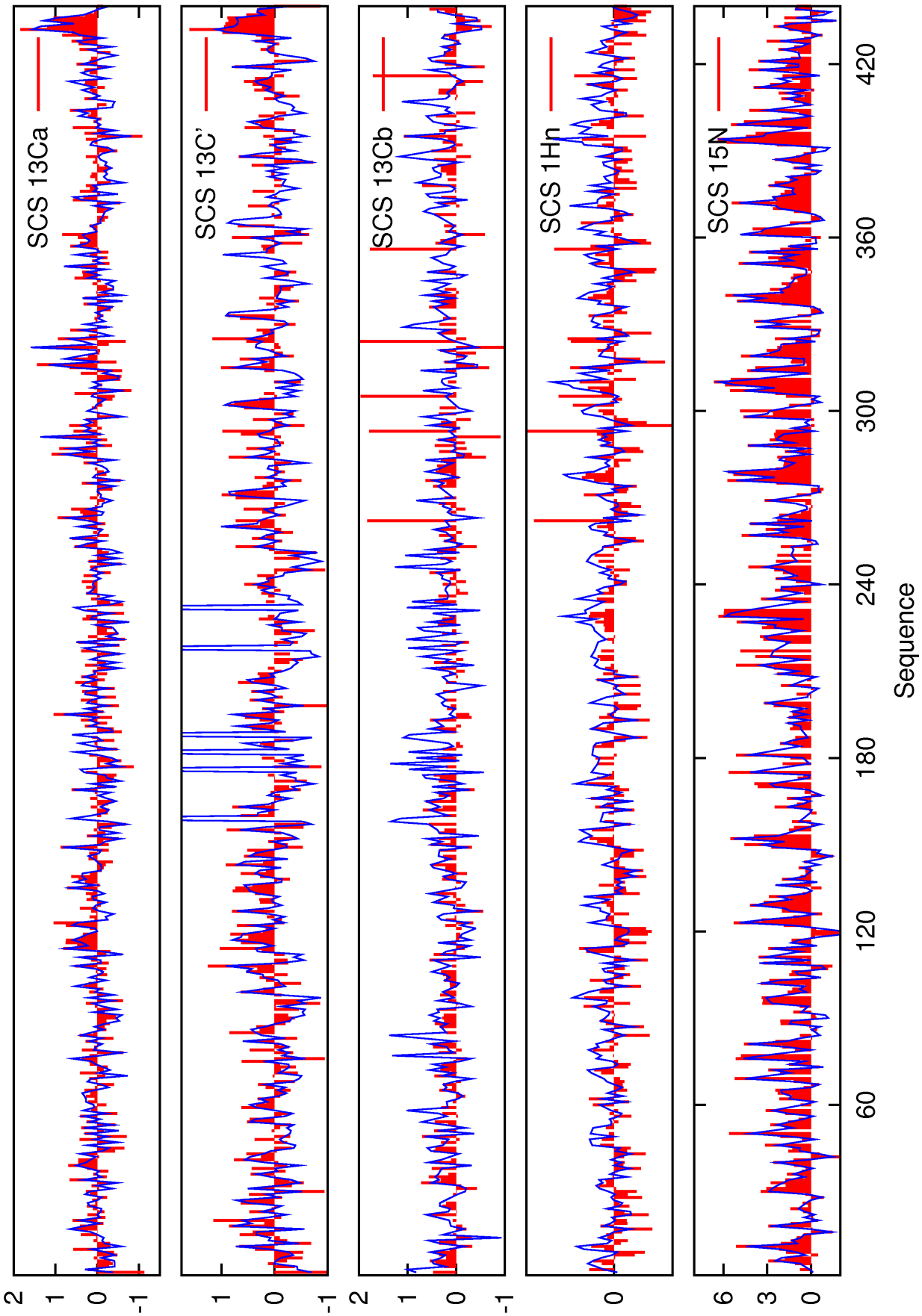


FIGURE 7.14 – ***Application de l'approche* Asteroids *aux données expérimentales de Tau.*** *Les données expérimentales sont tracées en rouge. Les données issues de la sélection sont tracées en bleu.*

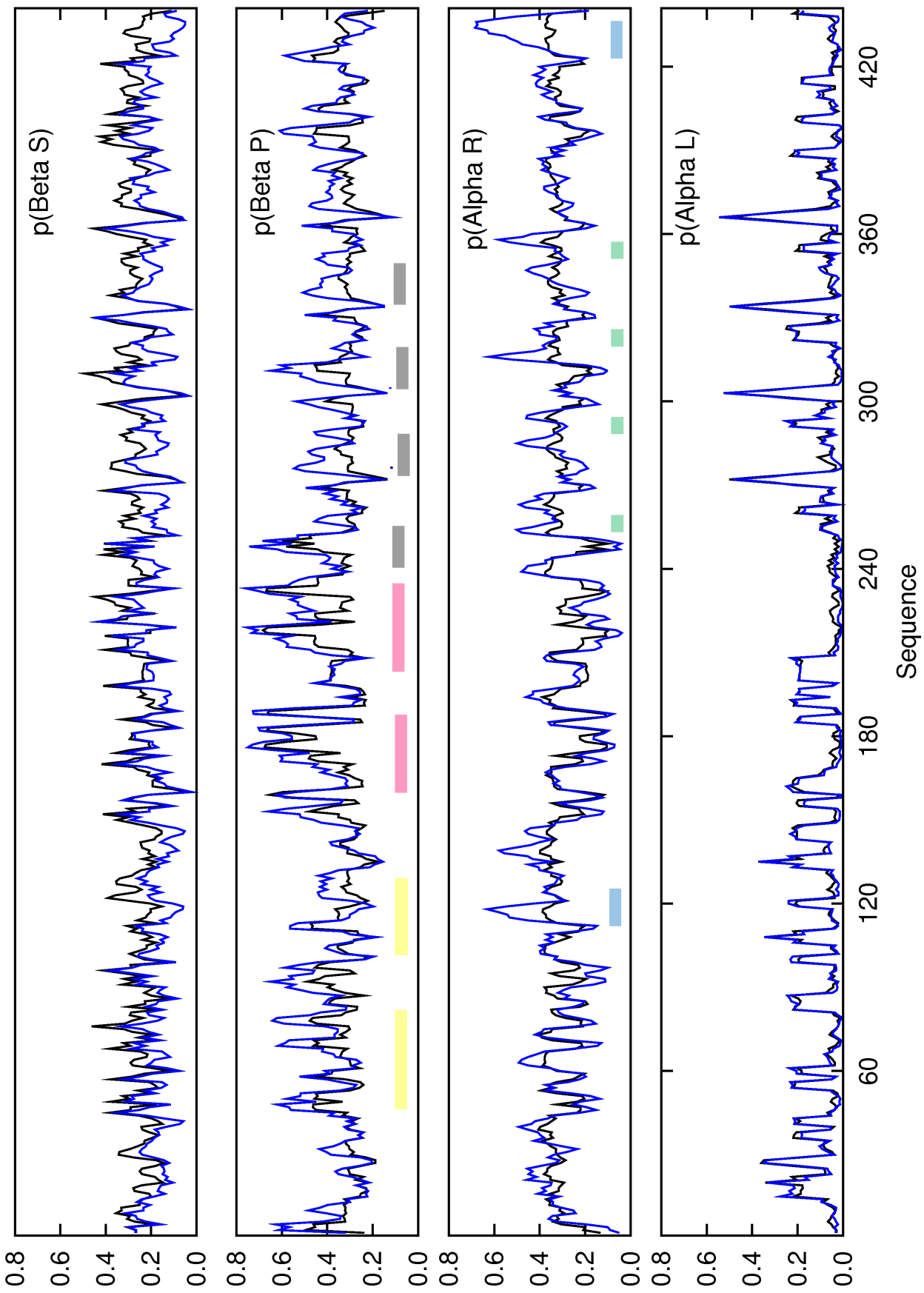


FIGURE 7.15 – ***Distribution de l'échantillonnage conformationnel de la protéine Tau à l'issu de la sélection avec* Asteroids.** *Les données random-coil en noir et les données issues de la sélection en bleu. Les populations sont moyennées sur une fenêtre de 3 résidus.*

en jaune : Nous observons ainsi un rôle plus prépondérant de la région βP pour certains motifs de la partie N-terminale de la protéine Tau

Les logiciels SPARTA et SPARTA+ ont été utilisés en parallèle et apportent des conclusions similaires de manière analogue aux protéines N_{tail} et K18 (se référer à la section 5.3.2 et 5.3.1).

7.4.8 Validation croisée des données PRE

Nous cherchons dans ce paragraphe à analyser l'influence de l'échantillonnage conformationnel sur les profils de relaxation paramagnétique. N'incorporant aucune information à longue distance lors de la validation croisée, nous ne pouvons pas reproduire les profils I/I_0 éloignés des cystéines. Nous pouvons par ailleurs étudier l'incorporation de la distribution d'angles (ϕ,ψ) extraite de la sélection avec les déplacements chimiques sur les profils de relaxation à proximité de la chaine latérale MTSL.

Exposés en figure 7.16, les profils de relaxation sont mieux reproduits en prenant en compte l'échantillonnage spécifique de la protéine Tau. Cette amélioration est néanmoins relative à la région considérée :

- Le profil des cystéines A15C et A416C présentes aux extrémités de la chaine principale est faiblement modifié étant donnée la flexibilité inhérente à ces régions.

- Nous notons une nette amélioration de la reproduction des cystéines suivantes : A72C, A178C, A239C, A256C, A291C-A322C et A322C. La région riche en Proline et la région des répétitions étant majoritairement échantillonnées dans la région βP, nous sommes un présence d'un ensemble localement plus étendu, cette nouvelle distribution de distances est favorable à la reproduction des profils des PRE. Nous soulignons en particulier le cas des cystéines A239C, A256C, A322C qui apparaissent clairement en accord avec ces considérations.

- Le dernier point concerne la cystéine A125C qui n'est pas influencée par la présence de l'hélice en position 113-120 et les 3 cystéines A352C, A384C et A416C légèrement moins en accord par rapport à l'ensemble *random-coil*. La présence de l'hélice en position 428-439 ne modifie pas nettement les profils qui sont plus sensibles aux résidus précédents échantillonnant la région βP : (368-380) (390-398) (402-406). Pour ce cas, l'ensemble semble trop étendu, il semblerait qu'il existe bel et bien un contact entre l'extrémité C-terminale et l'hélice transitoire.

Concernant le dernier cas évoqué, une autre piste à étudier concerne la flexibilité de la chaine latérale, une hypothèse plausible consiste à diminuer la dynamique de la chaine latérale MTSL de manière à reproduire au mieux les données expérimentales. La réponse pourrait ainsi inclure les deux précédentes hypothèses : la présence d'un contact entre l'hélice et l'extrémité C-terminale de la protéine aurait tendance à diminuer la flexibilité spatiale de la chaine latérale MTSL et aboutirait à une meilleure reproduction des PRE aux alentours de la cystéine.

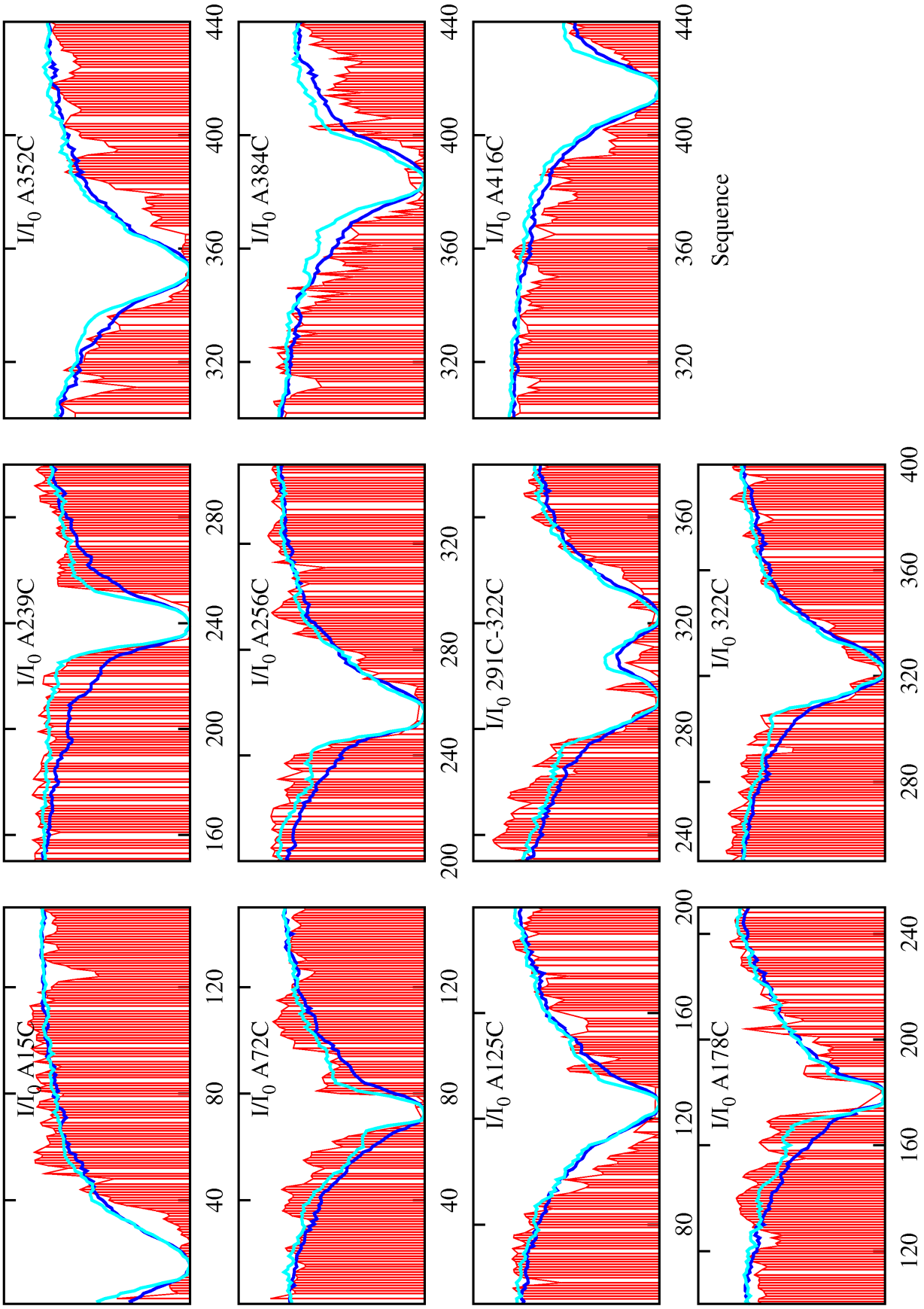


FIGURE 7.16 – ***Validation croisée des profils I/I_0 de la protéine Tau.*** *Les données expérimentales sont tracées en rouge. Nous comparons les profils issus d'un ensemble random-coil (en bleu foncé), avec les profils issus d'un ensemble dont l'échantillonnage conformationnel provient de la sélection des déplacements chimiques (en cyan).*

Validation croisée des PREs

La connaissance de la distribution des angles dièdres de notre ensemble permet d'améliorer le profil des PRE simulés à proximité des cystéines situées dans le domaine central de la protéine (A178C, A239C, A256C, 291C-322C et 322C). Cette validation croisée confirme une nouvelle fois la qualité du protocole déterminant l'échantillonnage conformationnel.

7.4.9 Etude de la flexibilité de la chaine latérale

La validation croisée des paramètres RMN est très importante, elle permet d'une part de s'assurer du bon déroulement de la sélection mais aussi d'identifier d'hypothèses à étudier. Ainsi, en échantillonnant plus souvent la région βP, la protéine Tau est localement plus étendue. Ces mesures semblent en très bon accord avec le profil des PREs. Nous souhaitons maintenant incorporer ce résultat pour étudier la flexibilité de la chaine latérale.

Le modèle de dynamique de la chaine latérale a été présenté au chapitre 6 et appliqué à la protéine α-Synucléine et à la protéine Tau au début du chapitre. Cette simulation utilisait un ensemble *random-coil* avec une chaine latérale hautement dynamique. Nous cherchons maintenant à évaluer si partant d'un échantillonnage spécifique nous pouvons améliorer la reproduction des données en utilisant une chaine latérale plus statique.

La figure 7.17 montre la reproduction des données PRE en validation croisée, l'utilisation d'un échantillonnage local plus étendu améliore toujours la reproduction des données. L'utilisation d'une chaine latérale légèrement plus statique pourrait être envisagée pour les cystéines suivantes A352C, A384C, A416C, où le profil de l'intensité est légèrement supérieur aux données prédites. Pour les autres cystéines, le cas hautement dynamique reste préférable pour reproduite au mieux les données PRE. Notons que le repliement partiel de la protéine peut influencer l'espace disponible de la chaine latérale MTSL entre la partie C-terminale de la protéine et la région riche en Proline. Il faudra ultérieurement prolonger cette étude en effectuant des validations croisées afin de confirmer les hypothèses suggérées, nous pourrions ainsi effectuer ce calcul pour plusieurs tailles d'ensemble et pour différentes cystéines passives dans la sélection et tester si l'incorporation de l'échantillonnage conformationnel permet une meilleur reproduction des données non incluses dans la sélection.

7.4.10 Comparaison des CDRs simulées et expérimentaux

Les CDRs sont à la fois sensibles à l'information locale et à longue distance, nous effectuons donc une validation croisée des couplages D_{NH} mesurés sur la protéine Tau entière ainsi que sur la troncature K32.

Comparant les données simulées d'un ensemble *random-coil* avec les données expérimentales de la protéine Tau, un des premiers points à souligner est la différence d'amplitude entre les couplages expérimentaux de la région C-terminale et N-terminale. La moyenne des couplages dipolaires résiduels vaut -1.3 Hz entre entre le résidu 1 et 150 et vaut -6.2 Hz entre les résidus 151 et 390. Les couplages dipolaires simulés adoptent cependant une ligne de base uniforme, l'échantillonnage conformationnel *random-coil* ne fait pas apparaitre de telle différence, dès lors ayant choisi un facteur d'ajustement

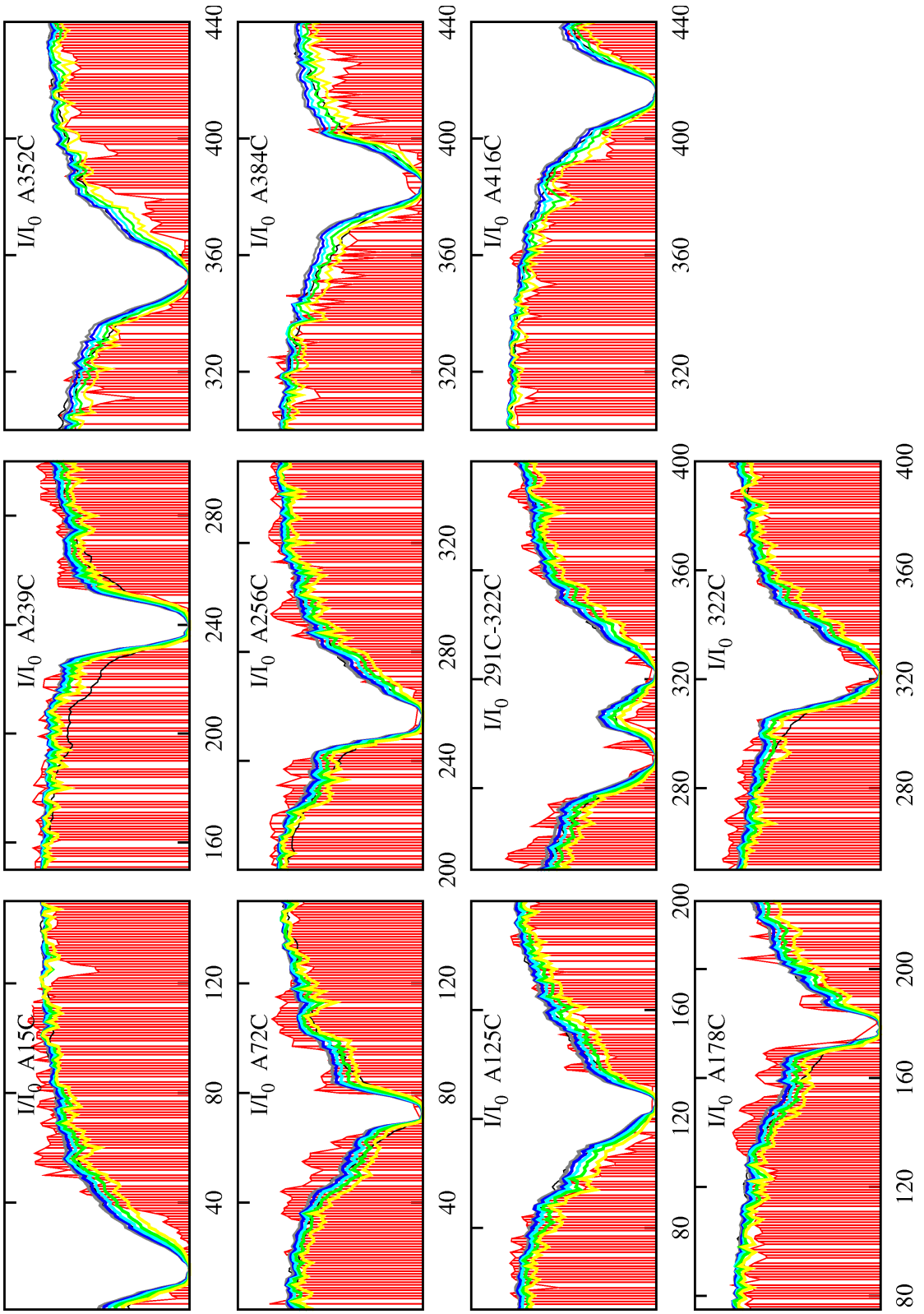


FIGURE 7.17 – ***Validation croisée des profils I/I_0 de la protéine Tau.*** *Les données expérimentales sont tracées en rouge. Nous comparons les profils issus d'un ensemble random-coil (en noir), avec les profils issus d'un ensemble dont l'échantillonnage conformationnel provient de la sélection des déplacements chimiques pour une chaine latérale plus (en gris) ou moins (en jaune) flexible.*

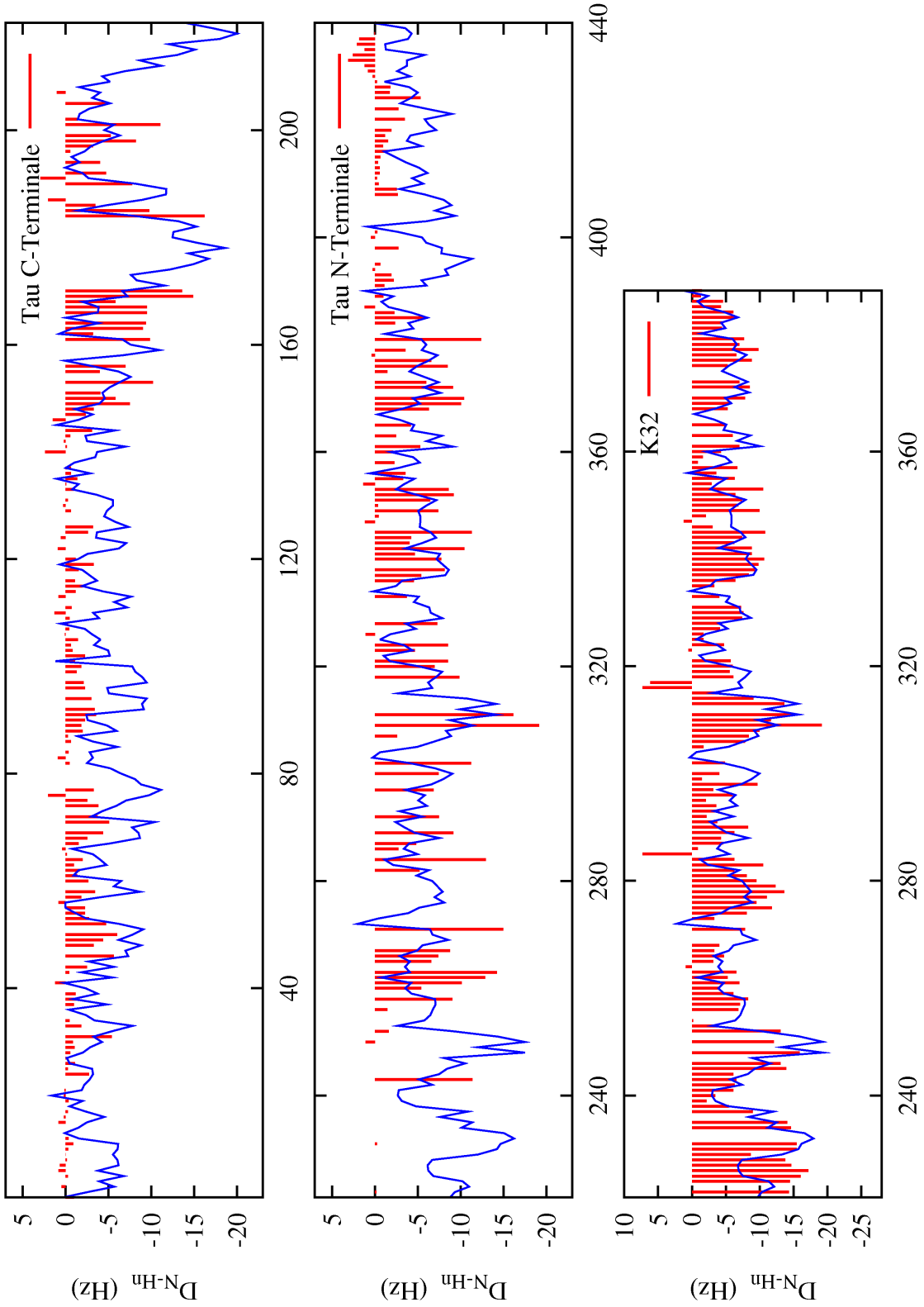


FIGURE 7.18 – ***Comparaison des CDRs simulés et expérimentaux de la protéine Tau et de la protéine K32 en utilisant un ensemble random-coil.*** *Les données expérimentales sont affichées en rouge, les données simulées en bleu.*

en accord avec la partie N-terminale, la partie C-terminale est moins bien reproduite. La mesure des couplages D_{NH} est plus complète sur la troncature K32, nous les utilisons comme repère en figure 7.18 en bas. A l'exception des structures transitoires connues, la reproduction des CDRs est enthousiasmante mais nécessite quelques améliorations notamment au niveau des boucles ou des feuillets.

Les différences existantes dans la partie C-Terminale sont probablement dues à la présence du contact pouvant induire une modulation de la ligne de base qui n'a pas jusqu'ici était pris en compte. Nous choisissons alors de calculer avec Flexible-Meccano de nouveaux ensembles en incorporant la position du contact déterminé en section 7.4.3. Il est difficile de déterminer au résidu près les limites des régions définissant le contact, nous testons plusieurs interactions à longue portée en faisant varier la position des domaines et la distance définissant le contact. Nous présenterons deux contacts définis à 15Å par les régions suivantes [1 :75][150 :225] pour le premier et [1 :75][125 :200] pour le second.

Exposée en figure 7.19, la validation croisée des couplages dipolaires résiduels D_{NH} comprenant les contacts préalablement définis reproduisent mieux les données expérimentales, nous passons pour les données simulées d'un valeur moyenne de couplage de -6.24 Hz à une valeur de -2.99 Hz. La modulation induite par la présence d'une interaction à longue portée entre la région riche en Proline et l'extrémité N-terminale permet une reproduction du profil expérimental. La dernière étape consiste maintenant à inclure l'ensemble des informations dans la description.

Conclusion du chapitre

Pour résumer, la protéine Tau à l'état natif se définit par deux caractéristiques :

- la présence d'un contact à longue portée entre la région riche en Proline et l'extrémité N-terminale

- des structurations locales transitoires avec un poids statistique plus important dans la région βP pour l'ensemble de la séquence, la présence de boucles dans le domaine d'appariement et la présence de deux hélices transitoires.

Incorporant l'échantillonnage issu de la sélection des déplacements chimiques et un contact spécifique entre la région N-terminale et la région centrale de la protéine nous effectuons une nouvelle simulation avec Flexible-Meccano en générant un ensemble de 150000 structures et calculons les paramètres RMN associés.

Nous sommes en mesure de reproduire de façon précise l'ensemble des données expérimentales : la figure 7.20 montre la reproduction des couplages D_{NH} de la partie N-terminale et de cinq profils I/I_0. Les profils des données couplages dipolaires et de la relaxation paramagnétique s'améliorent quantitativement dans les régions mentionnées précédemment. Une parfaite reproduction des données PRE nécessiterait la présence de plusieurs contacts transitoires au sein de l'ensemble ou la sélection d'un ensemble de structures en accord avec les données. Nous nous limitons à ce stade à une description comportant un seul contact entre deux domaines qui, en première approximation, reflète correctement les caractéristiques de la protéine Tau. Nous affichons deux caractéristiques biophysiques de l'ensemble : la carte de contact reflétant la distribution des distances moyennes et la distribution des rayons de gyration (figure 7.20).

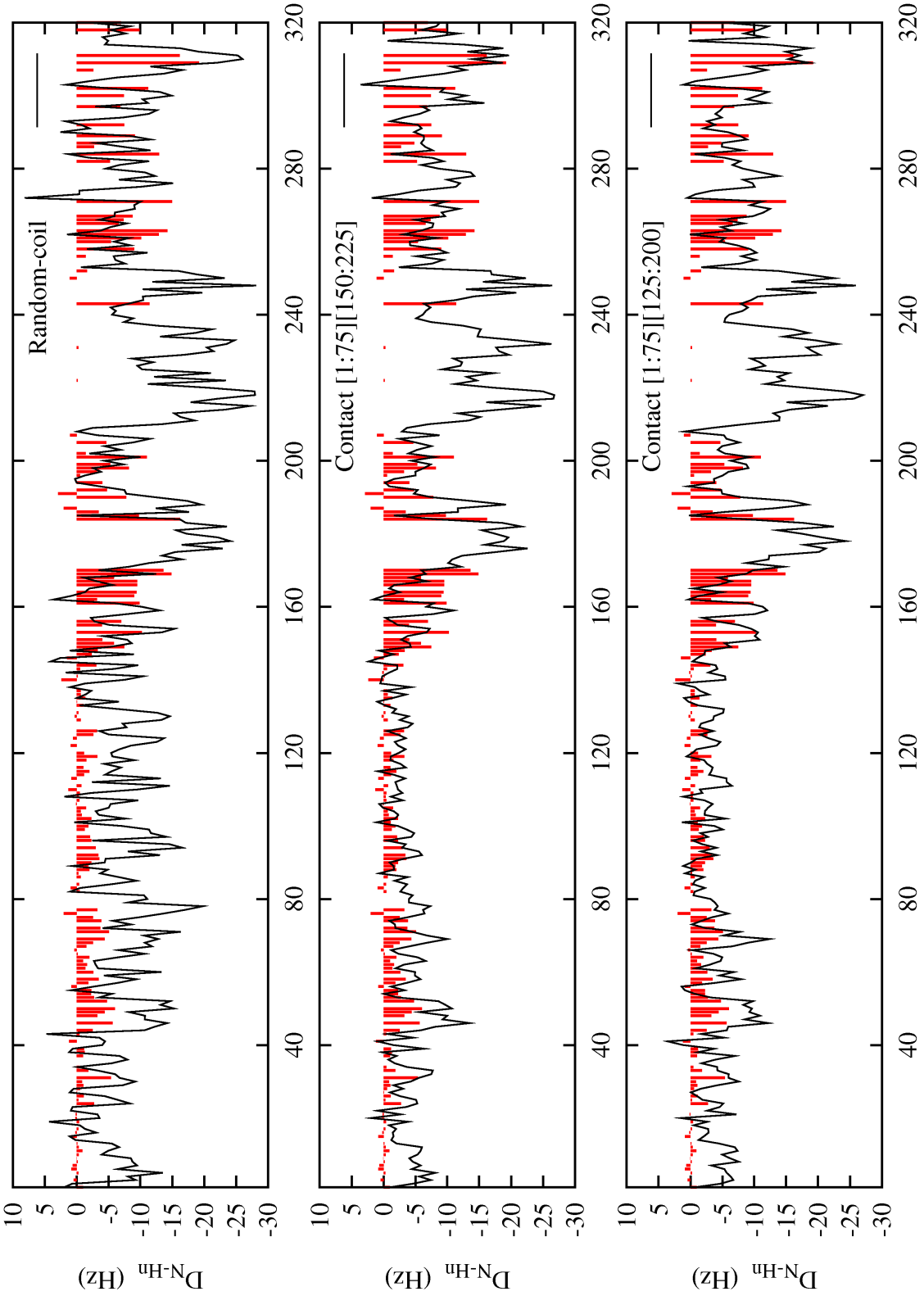


FIGURE 7.19 – ***Comparaison des CDRs simulés et expérimentaux en incorporant ou non un contact à longue portée entre deux domaines de la protéine Tau.*** *Les données expérimentales sont affichées en rouge, les données simulées en noir. Trois régimes sont présentés (de haut en bas) : le cas random-coil, un contact entre les régions [1 :75][150 :225] et un contact entre les régions [1 :75][125 :200].*

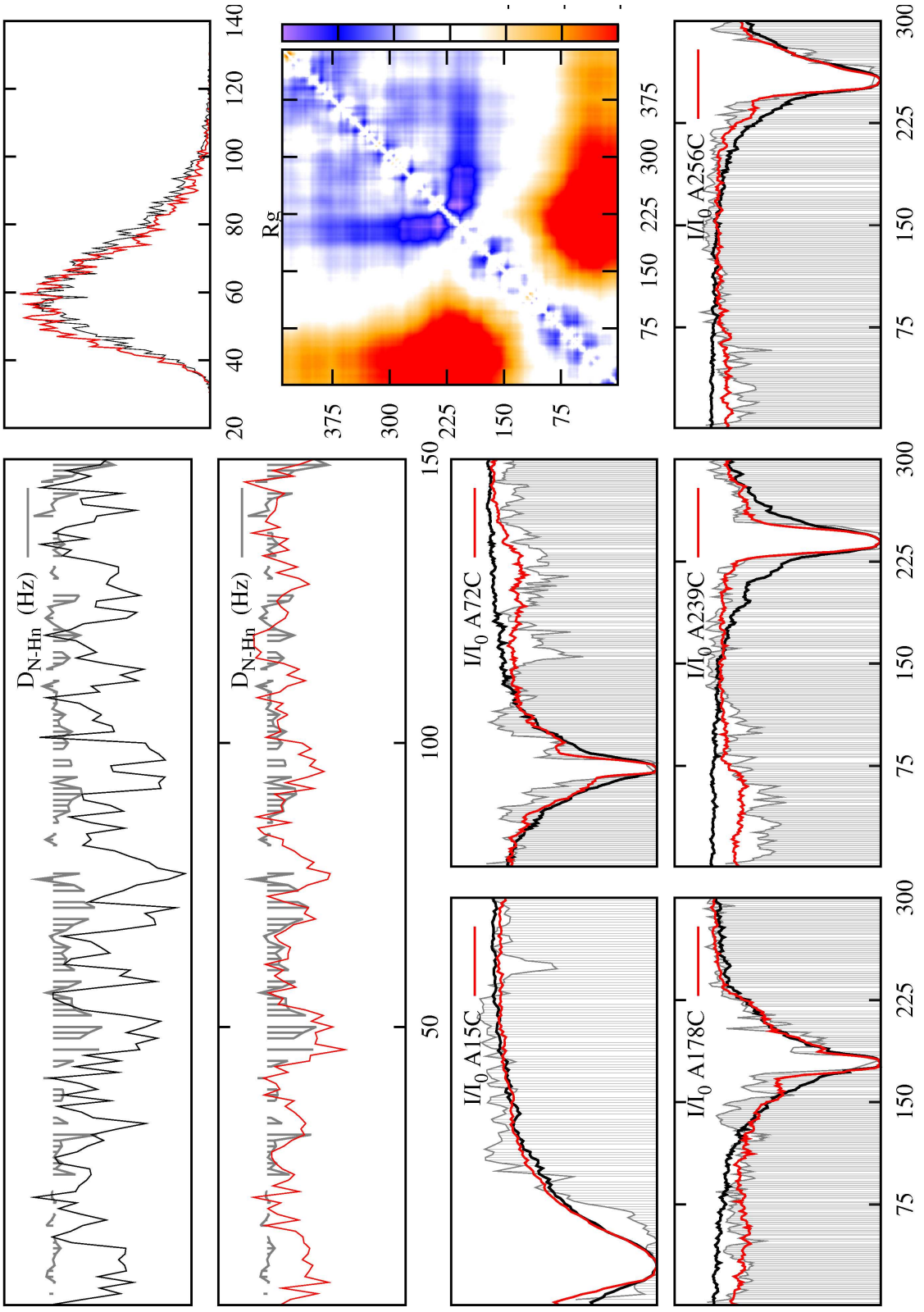


FIGURE 7.20 – ***Reproduction des données après incorporation de l'échantillonnage local et du contact.*** *Les couplages D_{NH} de la partie C-terminale ainsi que les profils* $\mathbf{I/I_0}$ *des cystéines A*15C,*A*72C, *A*178C, *A*239C, *A*256C *sont affichés, les données expérimentales sont en gris, les données de l'ensemble random-coil sont en noir, les données issues de l'ensemble avec un contact et une base de données d'angles dièdres spécifiques sont en rouge. La distribution du rayon de giration et la carte de contact associé aux deux ensembles sont aussi tracées.*

Ainsi, malgré l'existence de dégénérescence, nous sommes en mesure de proposer un modèle simple, et ceci, sans combiner directement les paramètres RMN au sein d'une sélection, Nous pouvons dès lors envisager ce cas, en ajoutant par exemple un pool préalablement construit avec les hypothèses. La sélection d'un sous-ensemble nous permettra alors d'affiner l'étude et de localiser de nouvelles caractéristiques.

Dans cette dernière étude, nous avons pu combiner l'ensemble de nos connaissances concernant la description par ensemble pour caractériser une protéine désordonnée de 441 résidus. Nous sommes maintenant en mesure d'analyser précisément à la fois l'information locale et l'information à longue portée dans une description par ensemble. Les protéines désordonnées présentent de nombreux degrés de liberté, la description par ensemble sous contrainte est une méthode appropriée pour caractériser ces protéines. Le seul pré-requis est la combinaison de plusieurs paramètres RMN : les déplacements chimiques sont très sensibles à l'échantillonnage local mais ne reflètent pas les interactions à longue portée à l'opposé de la relaxation paramagnétique. Ainsi, la combinaison de ces paramètres permet d'identifier l'ensemble des caractéristiques physiques de ces protéines. Un troisième paramètre RMN est fondamental pour analyser et valider les solutions recueillies : les CDRs. En effet, l'interaction dipôle-dipôle directe est sensible à la fois à l'information locale et à l'information à longue portée et fait ainsi d'elle un outil puissant pour évaluer la consistance de l'approche, elle permet de vérifier simplement si le modèle proposé est en accord avec les données expérimentales. Nous disposons maintenant de méthode robuste et sensible pour évaluer les propriétés physiques de ces protéines hautement flexibles.

Développement de l'application Java Flexible-Meccano

8

L'algorithme Flexible-Meccano est un modèle statistique utilisant une description par ensemble de structures explicites pour décrire l'état deplié. Il a été développé par le groupe Flexibilité et Dynamique des Protéines à l'Institut de Biologie Structurale, la section 4.1 présente l'algorithme et ses applications. Dans l'objectif d'une distribution du logiciel à la communauté RMN, nous avons incorporé le calcul des paramètres RMN puis intégré le programme dans une application Java basée sur Netbeans Platform.

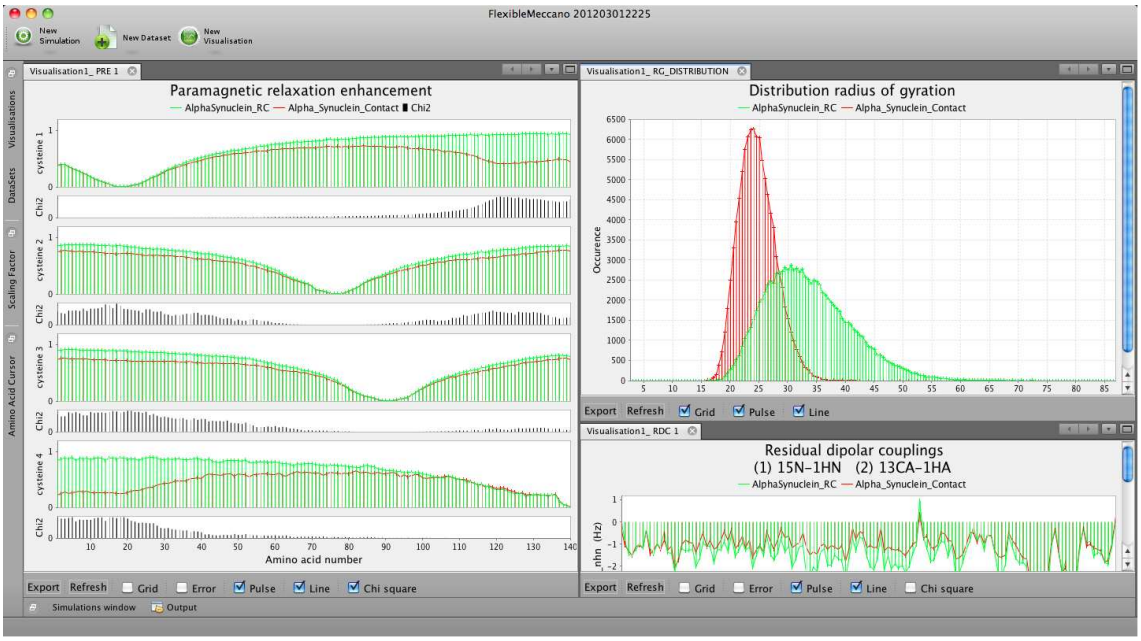


Figure 8.1 – ***Capture d'écran de l'application*** **Flexible-Meccano.** *Simulations de la protéine α-Synucléine avec (en rouge) ou sans (en vert) contact entre la région N-terminale et C-terminale.*

Ce projet de développement logiciel s'est déroulé de décembre 2010 à septembre 2012 en collaboration avec le Groupe Informatique Pour les Scientifiques du sud-est (GIPSE) du Commissariat à l'Energie Atomique (CEA) de Grenoble représenté par Celine CHARAVAY et Stéphane SEGUARD. L'objectif est de fournir une application Java évolutive, distribuable sur Internet intégrant l'application C Flexible-Meccano et proposant une interface graphique conviviale pour analyser les données des simulations en comparaison avec les données expérimentales de l'utilisateur.

8.1 Le projet

8.1.1 Le besoin

L'algorithme Flexible-Meccano a été développé pour générer un ensemble de structures en échange traduisant la flexibilité des protéines désordonnées. Dans l'objectif d'une distribution du logiciel sur internet, nous avons ajouté le calcul des paramètres RMN suivants : les couplages dipolaires résiduels D_{NH}, $D_{C^{\alpha}H^{\alpha}}$, $D_{C'C^{\alpha}}$, $D_{C'H^{N}}$, $D_{C'N}$, la relaxation paramagnétique du proton, le couplage scalaire ^{3}J. Des scripts SHELL sont fournis pour calculer les déplacements chimiques avec le logiciel Sparta et les profils d'intensité SAXS avec le logiciel Crysol. Les paramètres RMN sont calculés pour chaque structure générée et moyennés sur l'ensemble. Le programme fournit aussi les caractéristiques de l'ensemble comme le rayon de giration moyen, la carte des distances moyennes entre résidus et l'information concernant l'échantillonnage conformationnel par acide aminé. Ce programme C est intégré dans une application Java proposant une interface de travail divisée en trois axes :

- Le lancement de simulations en parallèle.

- La mise en mémoire des données simulées et des jeux de données expérimentaux.

- La visualisation de résultats et en particulier la mise en place d'une interface graphique permettant de naviguer rapidement entre les différents jeux de données
  .

8.1.2 Les étapes du projet

Les étapes de ce projet comprennent :

- La faisabilité : spécification des besoins.

- L'élaboration : mise en place de l'architecture technique et fonctionnelle.

- La fabrication : le développement, deux étudiants en Master ont grandement contribué au travail de développement du produit : Fréderic BAUER de mars 2011 à aout 2011 et Madalina GHITA de mai 2012 à aout 2012.

- La transition : mise en service. Une première version de Flexible-Meccano a été distribuée en mai 2012, une seconde version comportant le mode de visualisation complet sera distribuée dès octobre 2012.

8.1.3 L'architecture logicielle

L'architecture logicielle, elle est organisée en 5 modules fonctionnels indépendants, nous distinguons :

- Le module **core** comporte toutes les fonctionnalités communes aux autres modules.

- Le module **simulation** regroupe toutes les fonctionnalités correspondant au lancement d'une simulation.

- Le module **dataSet** permet la lecture de tous les de sorties de l'application C Flexible-Meccano, ces fichiers sont les paramètres RMN ou les caractéristiques des ensembles calculés.

- Le module **visualisation** correspond aux graphique et outils permettant de visualiser et comparer les paramètres RMN.

- Le module **gipse** est le module fonctionnel et technique commun à toutes les applications réalisées par l'équipe GIPSE.

8.2 Le fonctionnement de l'application Flexible-Meccano

Nous allons par la suite brièvement présenter le fonctionnement de l'application. Nous insisterons sur le structure de l'algorithme en C de Flexible-Meccano, la communication entre ce dernier et l'application Java. L'interface a été développée de manière à intégrer facilement par la suite tout nouveau code développé au sein du groupe. L'ajout de nouvelle fonctionnalité ne nécessitant pas de modification de l'application Java.

La communication entre l'interface Java et le code se fait par l'intermédiaire de quatre fichiers texte :

- info.txt : les instructions concernant les chemins des fichiers d'entrées et de sorties et les calculs à effectuer (figure 8.2).

- sequence.txt : la séquence de la protéine et la présence de structures secondaires locales (figure 8.3).

- contacts.txt : les contacts présents au sein de la protéine.

- phipsi.txt : la base de données d'angles (ϕ,ψ) de Flexible-Meccano.

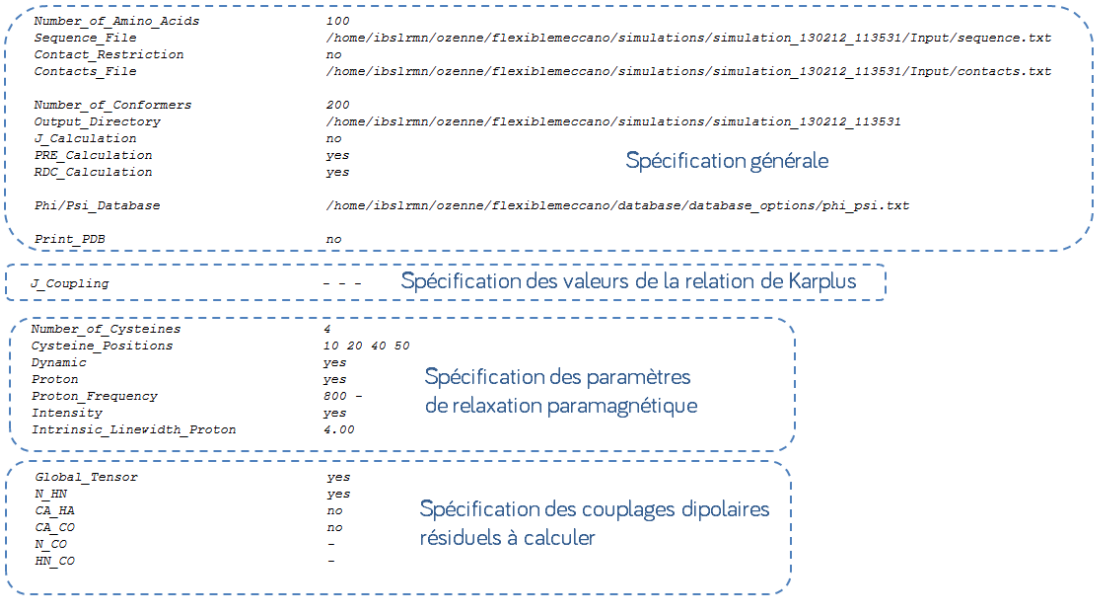


Figure 8.2 – ***Fichier d'entrée des spécifications de la simulation.*** *Ce fichier indique les chemins de l'ensemble des fichiers d'entrées et de sorties de la simulation ainsi que les calculs effectués lors de la simulation.*

(a) : Format CSV

Numéro d'acide aminé	Type d'acide aminé	Type de structures secondaires	Phi	Psi	Propension	Déviation
1	T	0	0.00	0.00	0.00	0.00
2	E	0	0.00	0.00	0.00	0.00
3	D	0	0.00	0.00	0.00	0.00
4	K	0	0.00	0.00	0.00	0.00
5	I	0	0.00	0.00	0.00	0.00
6	S	0	0.00	0.00	0.00	0.00
7	R	0	0.00	0.00	0.00	0.00
8	A	0	0.00	0.00	0.00	0.00
9	V	0	0.00	0.00	0.00	0.00
10	G	0	0.00	0.00	0.00	0.00
11	P	0	0.00	0.00	0.00	0.00
12	R	0	0.00	0.00	0.00	0.00
13	Q	0	0.00	0.00	0.00	0.00
14	A	0	0.00	0.00	0.00	0.00
15	Q	0	0.00	0.00	0.00	0.00
16	V	0	0.00	0.00	0.00	0.00

(b) : Format FASTA

```
>DisProt|DP00303|uniprot|P02185|sp|MYG_PHYCA|gi|118595805 #1-153 #1-
153 #144-149 #82-101 #100-118 #86-94 #43-49 #36-42
VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED
LKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
GDFGADAQGAMNKALELFRKDIAAKYKELGYQG
```

Figure 8.3 – ***Fichier d'entrée de la séquence au format CSV et au format FASTA.*** *(a) : La format CSV contient toute l'information pour modifier l'échantillonnage local de la protéine. (b) Format classique d'un fichier FASTA, celui présenté est l'apo-myoglobine.*

Pour chaque simulation, l'utilisateur charge la séquence de la protéine en format FASTA ou CSV. Il peut ajouter un contact, *i.e.* une contrainte de distance, entre deux régions de la protéine, il peut par ailleurs modifier l'échantillonnage conformationnel de chaque acide aminé. A titre exemple, il est possible d'échantillonner transitoirement en région de l'espace Ramachandran spécifié par la position des angles dièdres pour générer une hélice α, un feuillet β, une hélice α 3_{10}, une hélice PPII ou tout autre région de l'espace Ramachandran inclut dans la base de donnée standard de Flexible-Meccano (une hélice α et un contact transitoire sont imposés en figure 8.4). Ensuite, il spécifie le nom de la simulation, le chemin du répertoire des fichiers de sortie, l'impression ou non des structures en format PDB, les paramètres RMN à calculer et les spécificités de ces derniers. Des paramètres par défaut adaptables sont fournis. L'utilisateur en répétant cette opération qui prend 10 à 30 secondes peut alors lancer plusieurs simulations en parallèle (figure 8.5).

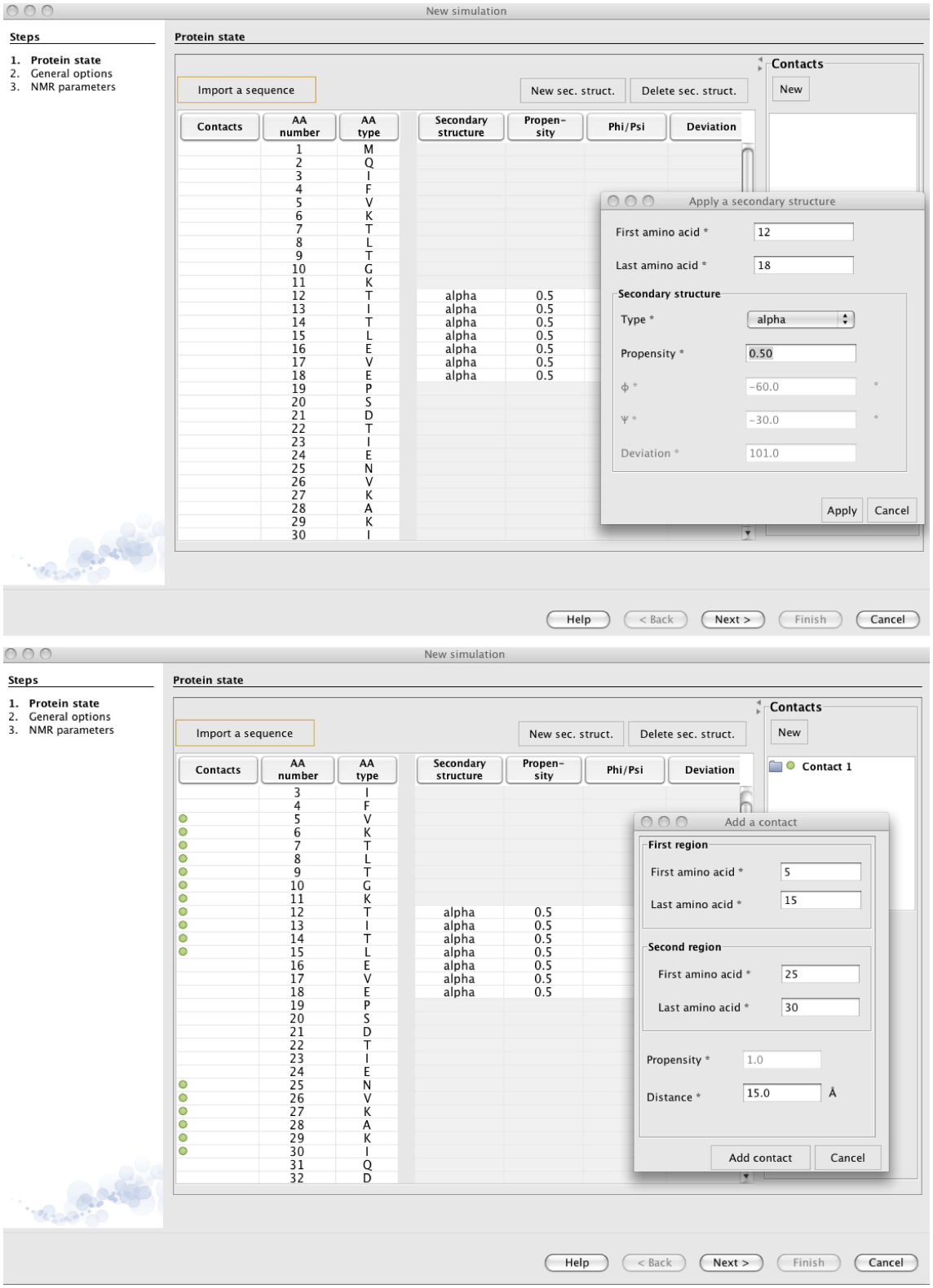


FIGURE 8.4 – ***Spécifications des caractéristiques de la protéine.*** *Chargement de la séquence, ajout des structures secondaires et des contacts transitoires. Une hélice coopérative sera échantillonnée* 50% *du temps entre les résidus* 12 *et* 18*. D'autre part, un contact de 15Å sera imposé entre la région* 5-15 *et la région* 25-30*, un curseur vert apparait pour chaque résidu concerné.*

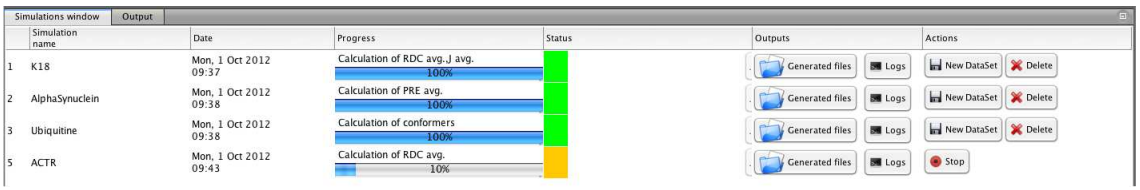


Figure 8.5 – ***Lancement de 4 simulations avec* Flexible-Meccano.** *Une barre d'écoulement indique l'avancement de la simulation. Une fois terminée, la simulation apparait en vert, les données peuvent être sauvegardées ou supprimées. Il est possible d'interrompre une simulation qui s'affichera alors en rouge.*

8.2.1 Le module *dataSet*

Le nombre de données générées lors des simulations pouvant être conséquent nous proposons des outils pour sauvegarder et comparer les jeux de données entre eux et ceci en limitant le nombre d'opérations à faire pour l'utilisateur. La mise en mémoire des données se fait par l'intermédiaire de le module *dataSet*, l'utilisateur peut sauvegarder des données expérimentales ou des données provenant de simulations (figure 8.6).

Ce module gère la lecture de tous les fichiers texte contenant les valeurs des paramètres RMN de chaque protéine. Ces valeurs sont soit issues d'une simulation soit issues d'une expérience. Les valeurs simulées sont stockées dans plusieurs fichiers (un pour chaque paramètre RMN) avec pour chaque acide aminé la valeur moyenne du paramètre calculé sur l'ensemble des conformations. Pour les données expérimentales, ces données doivent être également stockées dans des fichiers par l'utilisateur. Les formats étant similaires à ceci prêt que les données expérimentales possèdent une colonne réservée à l'estimation de l'erreur expérimentale. Deux classes Java (entités) ont été utilisées pour implémenter les concepts :

- *dataSet* : le jeu de données regroupe toutes les valeurs d'un ou de plusieurs paramètres RMN. Ces valeurs sont issues soit d'une même simulation, soit issues d'une même expérience.

- *set of points* : l'ensemble de points représente les valeurs d'un paramètre RMN. Un *dataSet* est composé d'un ou de plusieurs ensembles de points. Pour un simulation donnée, nous pouvons calculer plusieurs paramètres RMN.

8.2.2 Le module de visualisation

L'interface propose ensuite la création de *visualisation* d'un ensemble des graphiques spécifiques aux observables mesurées ou simulées. Ainsi, l'utilisateur sélectionne les jeux de données à comparer puis les paramètres RMN ou biophysiques, si existants, qu'il souhaite comparer (figure 8.7).

Il est possible de comparer de données simulées-simulées, simulées-expérimentales, expérimentales-expérimentales, l'utilisateur choisit le nombre de jeux de données qu'il souhaite comparer, si ce nombre est égal à deux, le χ^2 sera calculé pour chaque paramètre et récapitulé dans un tableau.

Il est possible de modifier le facteur d'ajustement des couplages dipolaires résiduels, d'afficher le χ^2 par résidu, d'afficher un curseur commun à toutes les fenêtres, de déplacer et redimensionner les fenêtres, de changer les couleurs des jeux de données, de modifier l'affichage des points en impulsions, de tracer ou non l'erreur expérimentale (figure 8.8). L'interface propose l'impression des graphiques obtenus en format PDF ou

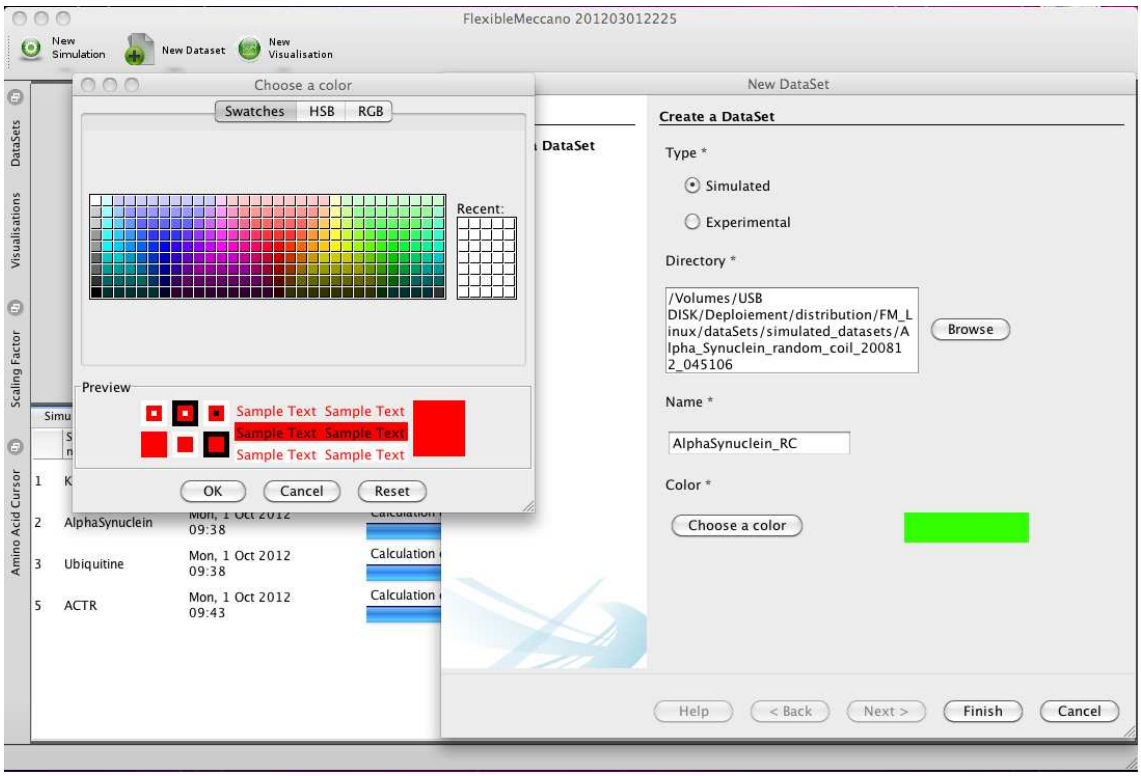


FIGURE 8.6 – ***Mise en mémoire des données.*** *Nous spécifions le type de données : expérimentales ou simulées, le chemin du répertoire où elles se trouvent, le nom du jeu de données et la couleur pour l'affichage.*

PNG (figure 8.9).

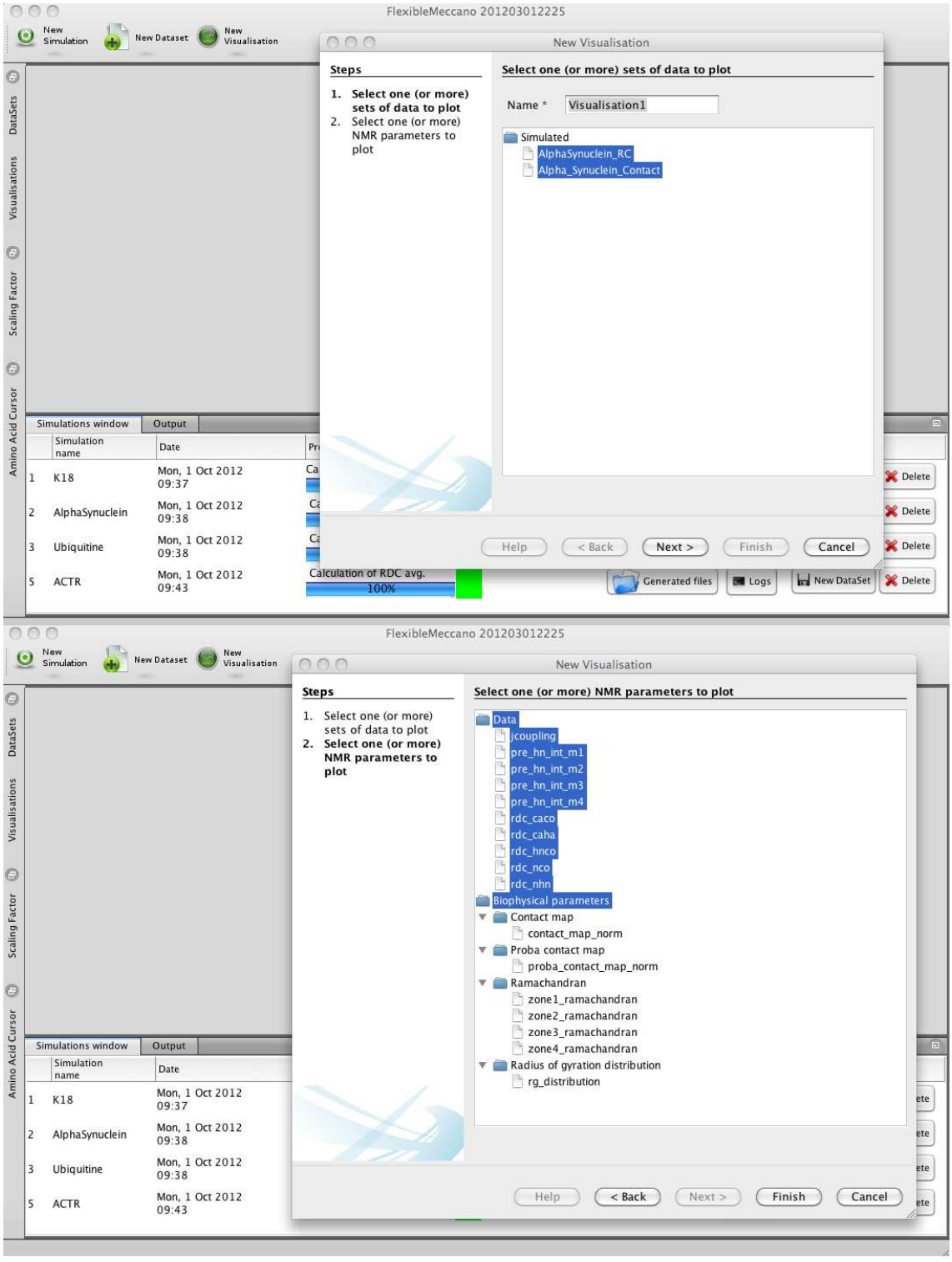


FIGURE 8.7 – ***Création d'une visualisation.*** *Nous sélectionnons les jeux de données à comparer, ces jeux doivent être préalablement mis en mémoire, puis nous choisissons les paramètres RMN ou biophysiques que nous souhaitons affichés.*

FIGURE 8.8 – ***Options supplémentaires.*** *Il est possible de modifier le facteur d'ajustement des couplages dipolaires résiduels, d'afficher le χ^2 par résidu (en noir), d'afficher un curseur commun à toutes les fenêtres, de déplacer et redimensionner les fenêtres et de changer les couleurs des jeux de données.*

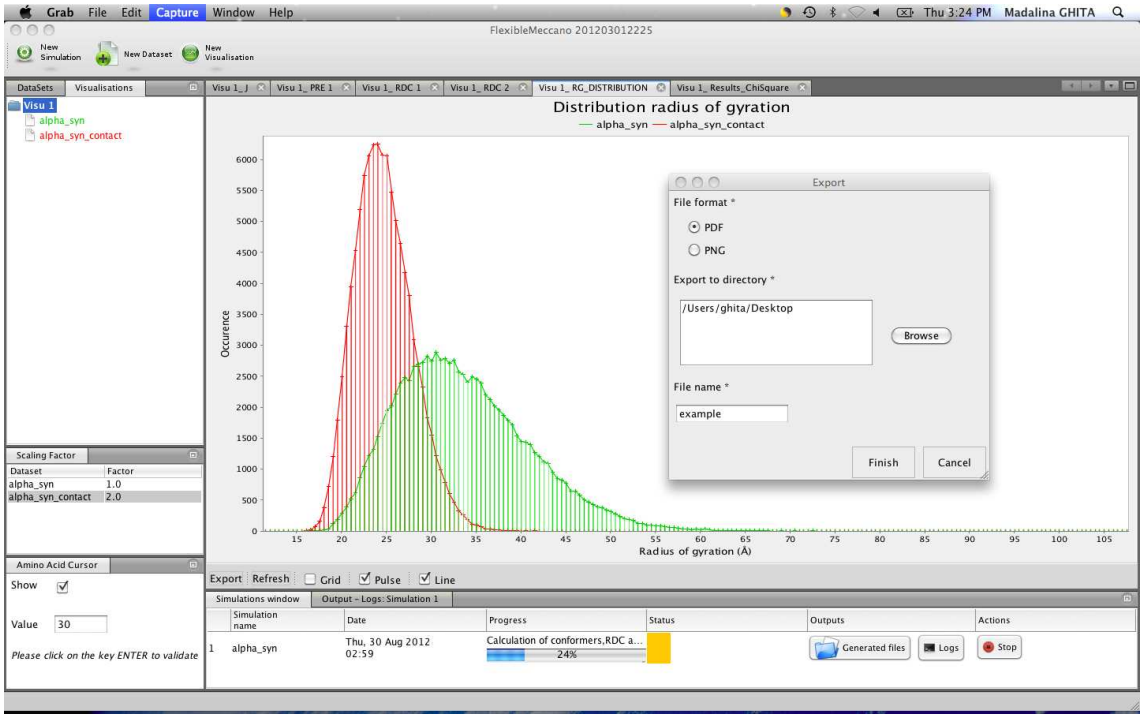


FIGURE 8.9 – ***Impression des graphiques obtenus en format pdf ou png.*** *Nous regardons la distribution du rayon de giration de deux ensembles simulés de la protéine α-Synucléine.*

INSTALLATION GUIDE FLEXIBLE MECCANO

A. SYSTEM REQUIREMENTS

- A recent version (version 6 or later) of Java Runtime Environment (JRE) available from the JAVA site :
- http://www.oracle.com/technetwork/java/javase/downloads/jre-6u27-download-440425.html
- A linux or Mac OS X operating system (OS)

B. FLEXIBLE MECCANO INSTALLATION

- Give execution permission to the script flexiblemeccano linux.sh for linux or FlexibleMeccano Installer.app for Mac OS X
- Execute the installation scripts
- In the installation screen, you will have to choose the installation directory. Keep in mind that this directory contains the script uninstall.sh needed for a clean uninstall of Flexible Meccano
- Follow the instructions in the installation wizard to finish the Flexible Meccano installation

C. FLEXIBLE MECCANO EXECUTION

- Click on the Flexible Meccano shortcut generated by the installation step.
- The first time you execute Flexible Meccano you will have to choose your working directory. This directory will contain a default value file, default database file and the results of your simulations.
- You can now use Flexible Meccano. Some protein sequence examples (in CSV or FASTA format) are provided in the Flexible Meccano archive you downloaded (in the "Sequences" directory).

D. HOW TO DO A CLEAN UNINSTALL OF FLEXBLE MECCANO

IMPORTANT: Do not remove manually the installation directory because it will not cleanly uninstall Flexible Meccano. Follow these steps to do a proper uninstall:

- Go to the installation directory defined in part B of this installation guide and execute the uninstallation script uninstall.sh
- Check the checkbox to delete the /dev Flexible Meccano directory

E. INTERFACE FLEXIBLE--MECCANO WITH EXTERNAL PROGRAMS

Bash script is provided with the Flexible Meccano download package (Scripts directory) in order to interface Flexible Meccano with external programs :

- SCCOMP : adds side chains to pdb files
- http://ignmtest.ccbb.pitt.edu/cgi-bin/sccomp/sccomp3.cgi

- SPARTA :calculates chemical shifts from pdb files (with side chains)
- http://spin.niddk.nih.gov/bax/software/SPARTA/index.html

- CRYSOL: calculates small angle scattering curves from pdb files (with side chains)

- http://www.embl-hamburg.de/biosaxs/crysol.html

Conclusion du chapitre

Ce projet de développement logiciel est maintenant terminé, il a donné lieu à la publication suivante [109], la première version de Flexible-Meccano a déjà été téléchargée plus d'une centaine de fois en quelques mois, la version présentée et contenant une architecture graphique plus complète sera distribuée très prochainement. Le logiciel est disponible pour Unix et Mac sur le site de l'Institut de Biologie Structurale à l'adresse suivante **http ://www.ibs.fr/science-213/scientificoutput/software/flexible-meccano/** avec instructions et exemples. Les applications futures concernent l'ajout de nouveaux paramètres RMN comme la relaxation paramagnétique de l'azote ^{15}N ou du carbone $^{13}C'$ ainsi que l'incorporation de domaines repliés lors de la génération de structures sont en cours et donneront lieu à de nouvelles options.

Conclusion générale

Les protéines intrinsèquement désordonnées ou protéines dépourvues de structure tridimensionnelle à l'état natif, sont devenues en quelques années l'objet de nombreuses études. La découverte de leur existence puis de la place importante qu'elles prenaient au sein du génome a profondément modifié la compréhension des mécanismes du vivant. En effet, la remise en question du dogme structure-fonction a nécessité le renouvellement de nos connaissances et de notre appréhension des fonctions biologiques. Il est nécessaire de repenser la manière d'étudier les protéines et d'interpréter les résultats pour prendre en considération l'existence de flexibilité au sein des protéines et des interactions protéines-protéines. Cette thèse intervient dans ce cadre, dans la mise en place d'une description moléculaire de l'état déplié. Il s'agit de mettre en place des méthodes de portée générale permettant de caractériser les protéines dépliées.

Au cours des premiers chapitres, nous avons présenté le contexte de l'étude et le protocole expérimental utilisé pour caractériser les protéines désordonnées. La méthode retenue est la résonance magnétique nucléaire qui dispose d'une sensibilité unique, d'une résolution à l'échelle atomique et qui permet par de diverses expériences d'accéder à l'ensemble des échelles de temps définissant les mouvements des protéines. Un des intérêts majeurs de cette technique dans notre situation est la possibilité de combiner ces mesures expérimentales à un modèle statistique représentant l'ensemble du paysage énergétique des protéines désordonnées : la description par ensemble explicite de structures. Ce modèle est une représentation discrète des différents états échantillonnés par ces protéines, il permet sous certaines hypothèses le calcul des données RMN et la comparaison aux données expérimentales.

L'algorithme Flexible-Meccano a été implémenté dans ce sens, pour offrir une description structurale de protéines flexibles n'ayant pas de structure propre. Devant ce paradoxe, de nouveaux outils s'imposaient pour décrire les caractéristiques de ces protéines. L'interprétation des données nécessitant le recours à la probabilité de distribution des paramètres définissant ces protéines. Dans un deuxième temps, nous avons cherché à produire une description plus quantitative des caractéristiques connues des protéines désordonnées. Certaines de ces protéines sont impliquées dans ce que l'on appelle des transitions ordre-désordre jouant notamment un rôle lors de la reconnaissance molécule avec un partenaire en pré-configurant le site d'interaction. Nous avons donc cherché à mettre en place un protocole précis quantifiant l'échantillonnage conformationnel de chaque résidu. Pour cela un second algorithme nommé Asteroids a été créé. Il a d'abord été appliqué à une protéine dans un régime déplié : l'Ubiquitine dénaturée dans l'urée (au chapitre 4) puis à deux protéines possédant des régions transitoirement structurées : la partie C-terminale N_{tail} de la nucléoprotéine N du virus de la Rougeole et la construction K18 de la protéine Tau (au chapitre 5). Cette description a nécessité le développement d'outil spécifique pour utiliser et interpréter l'information issue des mesures RMN : les déplacements chimiques et les couplages dipolaires résiduels sont hautement sensibles à l'échantillonnage local des protéines mais nécessitent d'une part une compréhension aiguë de la convergence de ces paramètres et d'autre part l'identification des règles régissant la relation entre l'échantillonnage conformationnel des

résidus et les valeurs associées à cette distribution d'angles dièdres.

La deuxième étape fut la caractérisation de l'information à moyenne et longue distance, une seconde caractéristique des protéines intrinsèquement désordonnées. En dépit de leur caractère structural fluctuant, il existe des contacts transitoires faiblement peuplés entre les différents domaines de la plupart des protéines désordonnées. La relaxation paramagnétique est alors une méthode de choix pour étudier ce phénomène, combiné avec notre méthode de description par ensemble, nous avons pu caractériser l'existence d'interaction à longue portée de deux protéines : la protéine α-synucléine impliquée dans la maladie de Parkinson, la protéine Tau impliquée dans la maladie d'Alzheimer. Un des points essentiels mis en jeu lors de ces études est la complémentarité des paramètres RMN utilisés. La combinaison des paramètres RMN permet d'identifier sans ambigüité les caractéristiques biophysiques de ces protéines et notamment grâce à l'utilisation des couplages dipolaires résiduels. Ces couplages dipôle-dipôle sont uniques, ils offrent une sensibilité accrue à la fois à l'information locale et à l'information à longue distance. Nous avons pu appliquer cette approche à la protéine Tau en combinant l'ensemble des paramètres RMN disponible pour étudier la forme native de la protéine. De plus amples travaux sont en cours pour caractériser aussi des formes phosphorylées.

Cette thèse est un travail de fond sur l'interprétation des paramètres RMN appliqué à un modèle statistique permettant de décrire l'état déplié. Les concepts et méthodes sont applicables et transposables à toute protéine dépliée. C'est sur point que réside un des intérêts majeurs de l'approche, il est désormais possible de caractériser quantitativement les propriétés physiques des protéines désordonnées, ces dernières pouvant combiner ou non la présence de structures secondaires ou contacts transitoires au sein de la même séquence. L'application de cette méthode a une protéine de 441 résidus en étant la démonstration.

La caractérisation des protéines désordonnées ne saurait se restreindre à cette étude, de nombreuses méthodes biophysiques offrent bien des perspectives : nous mentionnerons le SAXS, la microscopie à force atomique (AFM), le transfert d'énergie par résonance de type Förster (FRET). L'incorporation de cette information pourra nous aider à définir au mieux les règles régissant le comportement des protéines désordonnées.

Concernant les perspectives, de nombreux points sont à présenter : le point initial concerne la définition de l'état *random-coil*, la caractérisation croissante de l'échantillonnage conformationnel des protéines désordonnées va nous aider à améliorer la compréhension de ce régime, nous avons ainsi identifié au chapitre 5 la place importante de la région βP sur deux protéines, l'extension de cette méthode à des nouvelles protéines nous permettra de définir le régime *random-coil* pour chaque acide aminé, puis en fonction des conditions environnementales. Nous pourrons d'autre part appliquer ultérieurement ces protocoles à des données RMN mesurées *in-vivo*.

Une autre étape essentielle à la compréhension des mécanismes biologiques est la caractérisation de complexes de protéines désordonnées ou de complexes impliquant des régions désordonnées. La mise en place de méthode pour étudier les repliements transitoires, la reconnaissance moléculaire lors de l'appariement avec un ou plusieurs partenaires, lors des cascades de kinases par exemple, est une des prochaines étapes. Les modifications post-translationnelles sont fortement associées aux protéines désordonnées, l'influence de la phosphorylation sur l'échantillonnage conformationnel ou la présence de contact est un point clé pour comprendre les mécanismes de régulation des protéines intrinsèquement désordonnées. L'implication des protéines désordonnées

dans les maladies neurodégénératives est évidemment un élément moteur. Enfin, la compréhension des mécanismes impliquant des transitions conformationnelles menant à la formation de fibrilles est une question fondamentale pour la mise en place de solutions thérapeutiques.

Publications

Nodet G, Salmon L, Ozenne V, Meier S, Jensen MR, Blackledge M, Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings, *J. Am. Chem. Soc.* 2009, 131 : 17908-17918

Salmon L, Nodet G, Ozenne V, Yin G, Jensen MR, Zweckstetter M, Blackledge M, NMR Characterization of Long-Range Order in Intrinsically Disordered Proteins, *J. Am. Chem. Soc.* 2010, 132 : 8407-8418

Bibow S, Ozenne V, Biernat J, Blackledge M, Mandelkow E, Zweckstetter M. Structural impact of proline-directed pseudophosphorylation at at8, at100, and phf1 epitopes on 441-residue tau. *J. Am. Chem. Soc.* 2011, 133 :15842-15845

Schneider R, Huang JR, Yao M, Communie C, Ozenne V, Mollica L, Salmon L, Jensen MR, Blackledge M. Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. *Mol. Biosyst.* 2012, 8 :58-68

Ozenne V, Bauer F, Salmon L, Huang JR , Jensen MR, Segard S, Bernadó P, Charavay C, Blackledge M. Flexible-meccano : a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* 2012, 28 :1463-1470

Ozenne V, Schneider R, Yao M, Huang JR, Salmon L, Zweckstetter M, Jensen MR, Blackledge M. Mapping the potential energy landscape of intrinsically disordered proteins at amino acid specific resolution. *J. Am. Chem. Soc.* 2012, 134 : 15138-15148

Huang JR, Ozenne V, Jensen MR, Blackledge M. Direct prediction of NMR residual dipolar couplings from the primary sequence of unfolded proteins. *Angew. Chem. Int. Ed. Engl.* 2013, 52 : 687-690

J|A|C|S ARTICLES

Published on Web 11/12/2009

Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings

Gabrielle Nodet,[†] Loïc Salmon,[†] Valéry Ozenne,[†] Sebastian Meier,[‡] Malene Ringkjøbing Jensen,[†] and Martin Blackledge*,[†]

Protein Dynamics and Flexibility, Institut de Biologie Structurale Jean-Pierre Ebel, CEA, CNRS, UJF UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, France, and Carlsberg Laboratory, Gamle Carlsberg Vej 10, 2500 Valby, Denmark

Received August 22, 2009; E-mail: martin.blackledge@ibs.fr

Abstract: An atomic resolution characterization of the structural properties of unfolded proteins that explicitly invokes the highly dynamic nature of the unfolded state will be extremely important for the development of a quantitative understanding of the thermodynamic basis of protein folding and stability. Here we develop a novel approach using residual dipolar couplings (RDCs) from unfolded proteins to determine conformational behavior on an amino acid specific basis. Conformational sampling is described in terms of ensembles of structures selected from a large pool of conformers. We test this approach, using extensive simulation, to determine how well the fitting of RDCs to reduced conformational ensembles containing few copies of the molecule can correctly reproduce the backbone conformational behavior of the protein. Having established approaches that allow accurate mapping of backbone dihedral angle conformational space from RDCs, we apply these methods to obtain an amino acid specific description of ubiquitin denatured in 8 M urea at pH 2.5. Cross-validation of data not employed in the fit verifies that an ensemble size of 200 structures is appropriate to characterize the highly fluctuating backbone. This approach allows us to identify local conformational sampling properties of urea-unfolded ubiquitin, which shows that the backbone sampling of certain types of charged or polar amino acids, in particular threonine, glutamic acid, and arginine, is affected more strongly by urea binding than amino acids with hydrophobic side chains. In general, the approach presented here establishes robust procedures for the study of all denatured and intrinsically disordered states.

Introduction

Despite decades of experimental and theoretical advances in the characterization of structure, kinetics, dynamics, and thermodynamics of many thousands of soluble, folded proteins, the mechanism of protein folding, the conformational transition from a flexibl unfolded polypeptide chain to a stable folded protein structure, remains largely unexplained.[1] One reason for this is that one side of the protein folding equation is essentially impossible to characterize in atomic detail using classical approaches to structural biology, requiring instead the development of approaches that explicitly invoke the highly dynamic nature of the unfolded state.[2−5] An atomic-resolution characterization of the structural properties of unfolded proteins is therefore an essential prerequisite for a quantitative understanding of the thermodynamic basis of protein folding and stability.

The importance of developing techniques that are capable of describing the conformational sampling of unfolded polypeptide chains in solution has gained further importance with the gradual realization, over the past decade, that a large fraction of eukaryotic genomes codes for proteins that are intrinsically disordered in their native state.[6−9] Of particular relevance is the relationship between intrinsic structural characteristics of the unfolded chain and the mechanisms of protein folding upon binding, underlining the need for a basic understanding of the conformational space that is populated by a protein in the unfolded state.[10,11] The role that intrinsically disordered proteins (IDPs) play in neurodegenerative disease and cancer further emphasizes the importance of understanding conformational transitions from physiological to pathological forms of the same protein.[12]

Nuclear magnetic resonance (NMR) spectroscopy is probably the most powerful biophysical tool for studying IDPs due to

[†] Institut de Biologie Structurale Jean-Pierre Ebel.
[‡] Carlsberg Laboratory.

(1) Dill, K. A.; Shortle, D. *Annu. Rev. Biochem.* **1991**, *60*, 795–825.
(2) Daggett, V.; Fersht, A. R. *Natl. Rev. Mol. Cell Biol.* **2003**, *4*, 497–502.
(3) Vendruscolo, M.; Paci, E.; Karplus, M.; Dobson, C. M. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 14817–14821.
(4) Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 3–14.
(5) Eliezer, D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 23–30.
(6) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739–756.
(7) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. *Biochemistry* **2002**, *41*, 6573–6582.
(8) Tompa, P. *TIBS.* **2002**, *27*, 527–533.
(9) Fink, A. L. *Curr. Opin. Struct. Biol.* **2005**, *15*, 35–41.
(10) Dyson, H. J.; Wright, P. E. *Curr. Opin. Struct. Biol.* **2002**, *12*, 54–60.
(11) Fuxreiter, M.; Simon, I.; Friedrich, P.; Tompa, P. *J. Mol. Biol.* **2004**, *338*, 1015–1026.
(12) Dobson, C. M. *Trends Biol. Sci.* **1999**, *24*, 329–332.

10.1021/ja9069024 CCC: $40.75

the remarkable sensitivity of different NMR phenomena to dynamics occurring on time scales varying from picoseconds to hours and the ability to report on both local and long-range structure.[13] In particular, residual dipolar couplings (RDCs), which become measurable when a protein is dissolved in an anisotropic alignment medium or matrix,[14,15] have been shown to be very sensitive reporters of local and long-range structure,[16] even in highly disordered systems.[17] Since the initial demonstration that RDCs can be measured in proteins even under highly denaturing conditions,[18−25] it has been recognized that RDCs provide unique site-specifi probes of orientational order in disordered states.[17,26]

A recently developed explicit ensemble description of IDPs, flexible-meccano [27] constructs multiple copies of the protein in different states, designed to represent all possible conformational states that exchange on time scales relevant to the NMR time scale. Using a statistical coil description that samples amino acid-specifi backbone dihedral angle $\{\phi/\psi\}$ propensities, a conformational ensemble is created, and RDCs are calculated for each conformer and then averaged over the ensemble. This approach implicitly assumes that all conformers are in rapid exchange on time scales faster than a millisecond, an assumption based on the presence of a single set of NMR signals detected in ^{1}H and ^{15}N spectra of denatured and intrinsically disordered proteins. The absence of conformational exchange broadening excludes the presence of exchange between significantl populated conformational states occurring on slower time scales. RDCs simulated using these approaches present reasonable agreement with experimental couplings measured in both intrinsically disordered and chemically denatured proteins.[28−32] These studies have been used to provide evidence that site-specifi differences in RDCs measured along the primary chain can result from native differences in the rigidity of different amino acid types in an otherwise fully disordered chain,[27] from the presence of transiently populated local secondary structural elements[31] or from the presence of transient interactions between sites distant in the chain.[28]

While ^{15}N−^{1}H^{N} RDCs are by far the most commonly measured dipolar couplings, for reasons of experimental facility and precision, the advantages of measuring more RDCs from different spin-pairs in the peptide unit were recently demonstrated by Meier et al., who determined up to seven RDCs per amino acid from urea-unfolded ubiquitin at pH 2.5, including ^{15}N−^{1}H^{N}, ^{13}C$^{\alpha}$−^{1}H$^{\alpha}$, and ^{13}C$^{\alpha}$−^{13}C′ RDCs, inter- and intraresidue ^{1}H^{N}−^{1}H$^{\alpha}$ RDCs, and ^{1}H^{N}−^{1}H^{N} RDCs measured using quantitative *J*-type experiments[33] on perdeuterated ubiquitin. In combination, these data indicated that the standard description of the statistical coil behavior was inappropriate for urea unfolded proteins and that a modificatio of the random coil description was necessary to account simultaneously for all data.[34] On the basis of extensive simulation, the authors proposed that, in the presence of urea, the backbone dihedral angles definin the conformational behavior of the unfolded chain have a significantl higher propensity to sample more extended regions of Ramachandran space ($\psi > 50°$, $\phi < 0°$). This indication is supported by a comparison of extensive experimental small angle scattering (SAS) and pulse fiel gradient (PFG) dependences measured from urea-denatured proteins, with predicted data from conformational ensembles constructed using statistical coil models sampling increasing levels of this extended region (P. Bernado, personal communication). These independent biophysical techniques concur to substantiate an overall description of conformational bias respected by disordered polypeptide chains in the presence of high concentrations of denaturant.[35−38] RDCs measured between different spins within the peptide unit have also been shown to exhibit complementary dependences on the presence of local structure, an observation that has been shown to be crucial for the quantitative determination of the nature and extent of helical sampling present in molecular recognition elements of intrinsically disordered viral proteins[31] and the disordered N-terminal domain of p53.[39]

These studies have mainly used a rational, hypothesis-based approach, calculating explicit ensembles containing tens of thousands of conformers from different conformational sampling regimes and comparing the ensemble-averaged couplings to experimental data. In this study, we are interested in taking the analysis of RDCs one crucial step further, by investigating the possibility of definin the conformational sampling of the peptide chain directly from the experimental NMR data at amino

(13) Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, *104*, 3607–3622.
(14) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111–1114.
(15) Prestegard, J. H.; al-Hashimi, H. M.; Tolman, J. R. *Q. Rev. Biophys.* **2000**, *33*, 371–424.
(16) Blackledge, M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *46*, 23–61.
(17) Meier, S.; Blackledge, M.; Grzesiek, S. *J. Chem. Phys.* **2008**, *128*, 052204.
(18) Shortle, D.; Ackerman, M. S. *Science* **2001**, *293*, 487–489.
(19) Alexandrescu, A. T.; Kammerer, R. A. *Protein Sci.* **2003**, *12*, 2132–2140.
(20) Mohana-Borges, R.; Goto, N. K.; Kroon, G. J. A.; Dyson, H. J.; Wright, P. E. *J. Mol. Biol.* **2004**, *340*, 1131–1142.
(21) Fieber, W.; Kristjansdottir, S.; Poulsen, F. M. *J. Mol. Biol.* **2004**, *339*, 1191–1199.
(22) Meier, S.; Güthe, S.; Kiefhaber, T.; Grzesiek, S. *J. Mol. Biol.* **2004**, *344*, 1051–1069.
(23) Ohnishi, S.; Lee, A. L.; Edgell, M. H.; Shortle, D. *Biochemistry* **2004**, *43*, 4064–4070.
(24) Sallum, C. O.; Martel, D. M.; Fournier, R. S.; Matousek, W. M.; Alexandrescu, A. T. *Biochemistry* **2005**, *44*, 6392–6403.
(25) Ding, K.; Louis, J. M.; Gronenborn, A. M. *J. Mol. Biol.* **2004**, *335*, 1299–1307.
(26) Jensen, M. R.; Markwick, P.; Griesinger, C.; Zweckstetter, M.; Meier, S.; Grzesiek, S.; Bernado, P.; Blackledge, M. *Structure* **2009**, *17*, 1169–1185.
(27) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.
(28) Bernado, P.; Bertoncini, C.; Griesinger, C.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2005**, *127*, 17968–17969.
(29) Mukrasch, M. D.; Markwick, P. R. L.; Biernat, J.; von Bergen, M.; Bernado, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 5235–5243.
(30) Dames, S. A.; Aregger, R.; Vajpai, N.; Bernado, P.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2006**, *128*, 13508–13514.
(31) Jensen, M. R.; Houben, K.; Lescop, E.; Blanchard, L.; Ruigrok, R. W. H.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 8055–8061.
(32) Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 11266–11267.
(33) Meier, S.; Häussinger, D.; Jensen, P.; Rogowski, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2003**, *125*, 44–45.
(34) Meier, S.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 9799–9807.
(35) Kohn, J. E.; Millett, I. S.; Jacob, J.; Zagrovic, B.; Dillon, T. M.; Cingel, N.; Dothager, R. S.; Seifert, S.; Thiyagarajan, P.; Sosnick, T. R.; Hasan, M. Z.; Pande, V. S.; Ruczinski, I.; Doniach, S.; Plaxco, K. W. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12491–12496.
(36) Merchant, K. A.; Best, R. B.; Louis, J. M.; Gopich, I. V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1528–1533.
(37) Möglich, A.; Joder, K.; Kiefhaber, T. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 12394–12399.
(38) Gabel, F.; Jensen, M. R.; Zaccai, G.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 8769–8771.
(39) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. *Proc. Natl. Acad. Sci. (U.S.A.)* **2008**, *105*, 5762–5767.

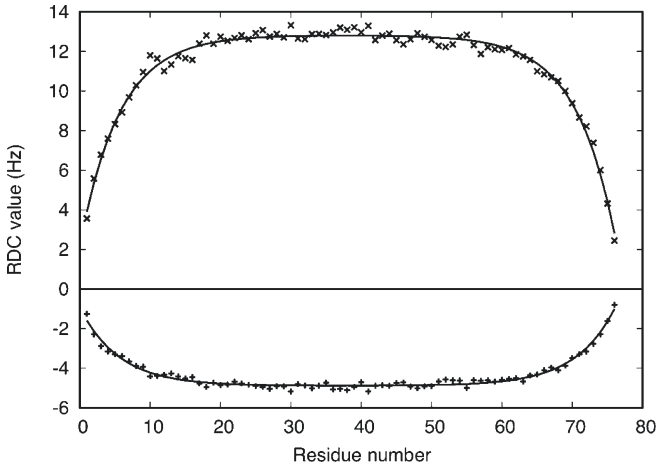


Figure 1. Residual dipolar coupling baselines in unfolded chains. Baseline effects underlying simulated ensemble-averaged RDCs from 100K copies of a polyvaline chain of 76 amino acids in length (crosses) and predicted RDCs following a hyperbolic cosine curve of the form given in eq 1 (line). $^{15}N-^{1}H^{N}$ couplings are shown below zero and $^{13}C^{\alpha}-^{1}H^{\alpha}$ RDCs are shown above zero.

acid-specifi or even atomic resolution, as have recently been developed in the Bonvin and Forman-Kay laboratories.[40,41] In order to do this, we develop a novel algorithm to select from a large pool of possible conformers, created using the algorithm flexible-meccano to best describe the system.

We test this approach, using extensive simulation, to determine how well the fittin of RDCs to reduced conformational ensembles containing few copies of the molecule can correctly reproduce the backbone conformational behavior of the protein. We also use cross-validation of data not employed in the fi to determine the most appropriate ensemble size to characterize the highly fluctuatin molecule. Having established approaches that allow accurate mapping of conformational space from RDCs, we apply these methods to the amino acid-specifi description of backbone conformational sampling in ubiquitin denatured in 8 M urea at pH 2.5.

Results and Discussion

RDCs from Disordered Proteins Modeled by Multiplication of Local Sampling Profiles and Underlying Baseline. RDCs can be simulated from explicit molecular ensembles of disordered proteins using shape-based considerations of the alignment properties of each copy of the molecule, and the average couplings can be predicted by taking the mean over the entire ensemble.[27,42] Comparison of such predictions with experimental data has revealed the unique sensitivity of RDCs to local and global sampling properties of highly disordered proteins. A key disadvantage of this approach is the number of structures that need to be treated, before the average RDC value converges to a nonfluctuatin value. This number can reach many tens of thousands in proteins of 100 amino acids. It has recently been proposed that convergence of RDCs toward experimental data can be achieved with a smaller number of conformers if the protein is divided into short, uncoupled segments (Local Alignment Windows, LAWs) and the RDCs are calculated using the alignment tensor of these segments.[43,44] The ability to describe the conformational properties with ensembles containing fewer structures will of course make any ensemble selection procedure more tractable and is therefore an attractive prospect. In general, however, RDCs are affected both by the local conformational sampling and the chain-like nature of the unfolded protein, which induce an effective baseline reflectin the increasing degrees of freedom available toward the ends of the chain.[45,46] Long-range information is therefore necessarily absent from an approach that only employs LAWs to predict the RDCs. If this approach is employed, the simulated data need to be corrected for the effects of the unfolded chain.

We have simulated ensemble-averaged RDCs for polyvaline chains of differing lengths. The predicted RDCs can be relatively well fitte to a hyperbolic cosine curve of the form (Figure 1)

$$B(i) = 2b\cosh(a(i - d)) - c \qquad (1)$$

where i is the residue number and d is half the number of residues. a, b, and c are optimized for each different coupling type, where $(2b - c)$ is the RDC value at position d. This baseline dependence can be used to correct RDCs calculated using LAWs as described below.

RDCs are simulated for the central residue of LAWs of equal length, sliding the LAW one amino acid at a time along the chain (note that the termini are treated in the same way by adding dummy residues beyond the ends of the chain; see Experimental Section). These RDCs are then averaged over all structures. RDCs simulated for LAWs of m amino acids in length will exhibit a fla baseline, because each calculated RDC is at the center of a fragment of m amino acids and is therefore at the middle of the same local effective baseline. The RDC distribution resulting from the LAWs therefore depends on amino acid type but does not contain the baseline effects. It can be shown (Figure 2) that this amino acid-specifi distribution can be multiplied with the baseline predicted in eq 1, to closely reproduce RDCs predicted from the explicit full-length description of the protein, which contains both amino acid-specifi effects and the chain nature of the full length protein.

In order to determine the convergent characteristics when RDCs are simulated using LAWs of different lengths, we have compared the average values taken over an increasing number of conformers. Examples are shown in Figure 3a of the same $^{1}D_{NH}$ RDC when the RDC is calculated for the central amino acid of LAWs of different lengths (3, 9, 15, 25, and full length protein of 76 amino acids). Further simulations of $^{1}D_{C\alpha H\alpha}$, $^{1}D_{C\alpha C'}$, $D_{NHH\alpha}$, and D_{NHNH} RDCs show similar convergent characteristics (data not shown). It is clear that for the full-length protein the average is only converged when more than 10 000 structures are taken into account, while for LAWs of 15 amino acids this number falls to a few hundred. Figure 3b shows the strong dependence of the range of sampled RDCs on the length of the LAW. As the LAW gets longer, the individual structures can have larger RDC values, rendering the average less and less stable (vide infra).

(40) Marsh, J. A.; Neale, C.; Jack, F. E.; Choy, W.-Y.; Lee, A. Y.; Crowhurst, K. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2007**, *367*, 1494–1510.
(41) Krzeminski, M.; Fuentes, G.; Boelens, R.; Bonvin, A. M. J. J. *Proteins: Struc. Funct. Bioinform.* **2009**, *74*, 894–905.
(42) Jha, A. K.; Colubri, A.; Freed, K.; Sosnick, T. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099–13105.
(43) Marsh, J. A.; Baker, J. M. R.; Tollinger, M.; Forman-Kay, J. D. *J. Am. Chem. Soc.* **2008**, *130*, 7804–7805.
(44) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359–374.
(45) Louhivuori, M.; Pääkkönen, K.; Fredriksson, K.; Permi, P.; Lounila, J.; Annila, A. *J. Am. Chem. Soc.* **2003**, *125*, 15647–15650.
(46) Obolensky, O. I.; Schlepckow, K.; Schwalbe, H.; Solov'yov, A. V. *J. Biomol. NMR* **2007**, *39*, 1–16.

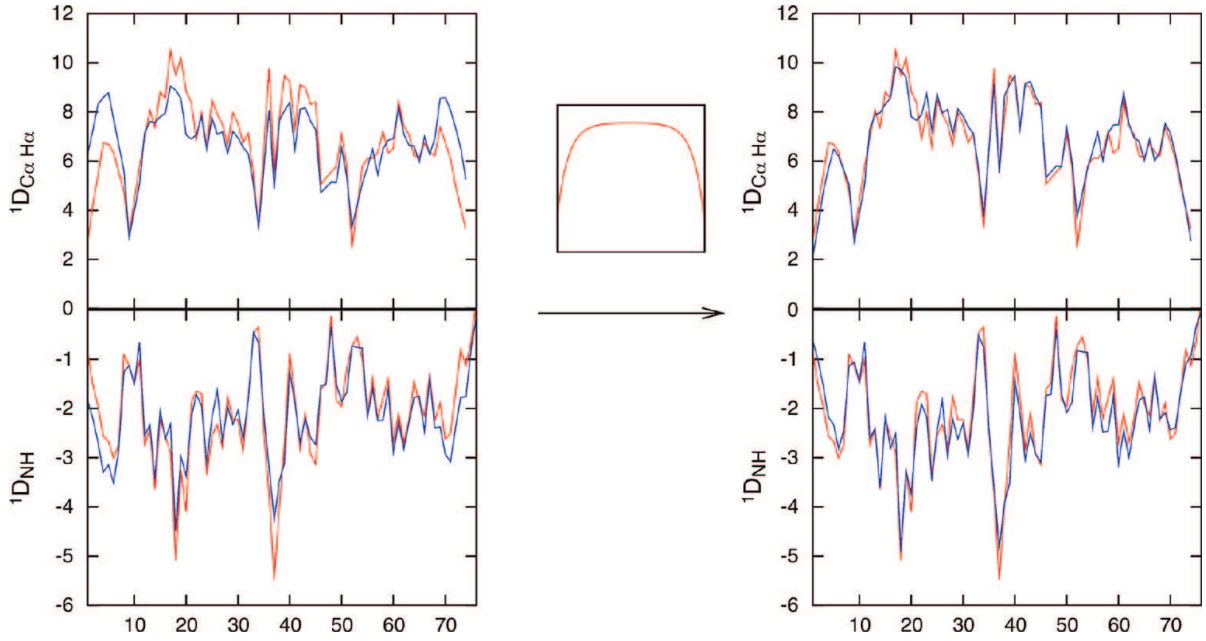


Figure 2. Multiplication of RDCs calculated using LAWs with RDC baselines in unfolded chains. $^{15}N-^1H^N$ and $^{13}C^\alpha-^1H^\alpha$ RDCs calculated from the central amino acid of a 15 amino acid LAW (blue, left) contain no baseline information and therefore diverge from the RDCs calculated from an explicit ensemble using a global alignment tensor (red). When multiplied with the hyperbolic cosine curve (eq 1), RDCs from the LAW (blue, right) more closely resemble the RDCs calculated from the global alignment tensor (red).

Alignment Strand Length Required To Define Accurately Conformational Sampling. In order to further determine the accuracy of describing RDCs using LAWs, we have compared the ability of LAWs of different lengths (after multiplication with the baseline described by eq 1) to reproduce RDCs simulated using a global alignment tensor (Figure 4). Not surprisingly, the shortest LAWs (three amino acids in length) never correctly reproduce average RDCs, due to the effects of neighboring amino acids (beyond nearest neighbors), on the local conformational sampling. The influenc of neighboring residues on local conformational sampling is commonly estimated in terms of a so-called "persistence length", beyond which the remainder of the chain can be considered to exert a negligible effect. The persistence length depends on the relative rigidity of the local primary sequence. The relevance of taking full account of the persistence length on the local conformational sampling is further demonstrated by simulations that have been performed using a more rigid statistical coil model for which RDCs simulated using LAWs of nine amino acids fail to reproduce the averaged RDCs calculated using the global alignment tensor (data not shown). These simulations therefore indicate that while convergence characteristics of the predicted RDCs improve with shorter LAWs, the shortest strands can never fully reproduce the correct average, even if a very large number of structures were used in the average. On the basis of these simulations, we consider that a LAW length of 15 amino acids should be an acceptable compromise between efficienc and accuracy for the subsequent analyses.

How Many Structures Are Required for RDCs To Define Accurately Conformational Sampling? The next question concerns the number of structures required to describe correctly the conformational sampling. The averaging of RDCs is particularly demanding in terms of numbers of structures for two main reasons: first because of the large number of backbone dihedrals whose relevant conformational space must be efficientl sampled before the overall shape and dimensions of the protein, and therefore the associated alignment tensor, average to convergent values. A second consideration is less obvious, but potentially more important: each dipolar coupling calculated from a single conformer of the entire molecule will sample a value within a range that can be orders of magnitude higher than the range spanned by the average values (Figure 3b). This dynamic-range problem can induce significan instability in the fittin procedure when using an ensemble containing too few structural models.

In order to numerically estimate the minimum number of structures that can accurately reproduce the true structural propensities of a conformational equilibrium, we have undertaken the following simulation: Two distinct statistical coil sampling regimes were defined and entire sets of RDCs were calculated from flexible-meccan using these regimes with the global alignment tensor. The first regime S, define the standard statistical coil model employed in flexible-meccano where amino acid-specifi conformational distributions are extracted from populations of coil regions found in the protein structural database. The second sampling regime (E) samples a more extended region of Ramachandran space, populating the region $\{50° < \psi < 180°\}$ with a higher propensity than the S regime (see Experimental Section), while retaining the amino acid specifi sampling from the S database. These data sets were then used as targets for the ensemble selection algorithm ASTEROIDS (A Selection Tool for Ensemble Representations Of Intrinsically Disordered States) described in the Experimental Section.

The ability of the algorithm to reproduce the correct conformational sampling and the correct RDCs for two different LAWs and the global alignment tensor is summarized in Figure 5 as a function of the number of structures constituting the ensemble. Using the target function χ^2_{Ram}, which measures the population of four different regions of Ramachandran space define in Figure 6, we measure the ability of the protocol to reproduce amino acid-specifi conformational sampling throughout the molecule (see Experimental Section). In each of the three considered window lengths, (9, 15, and full length protein), the

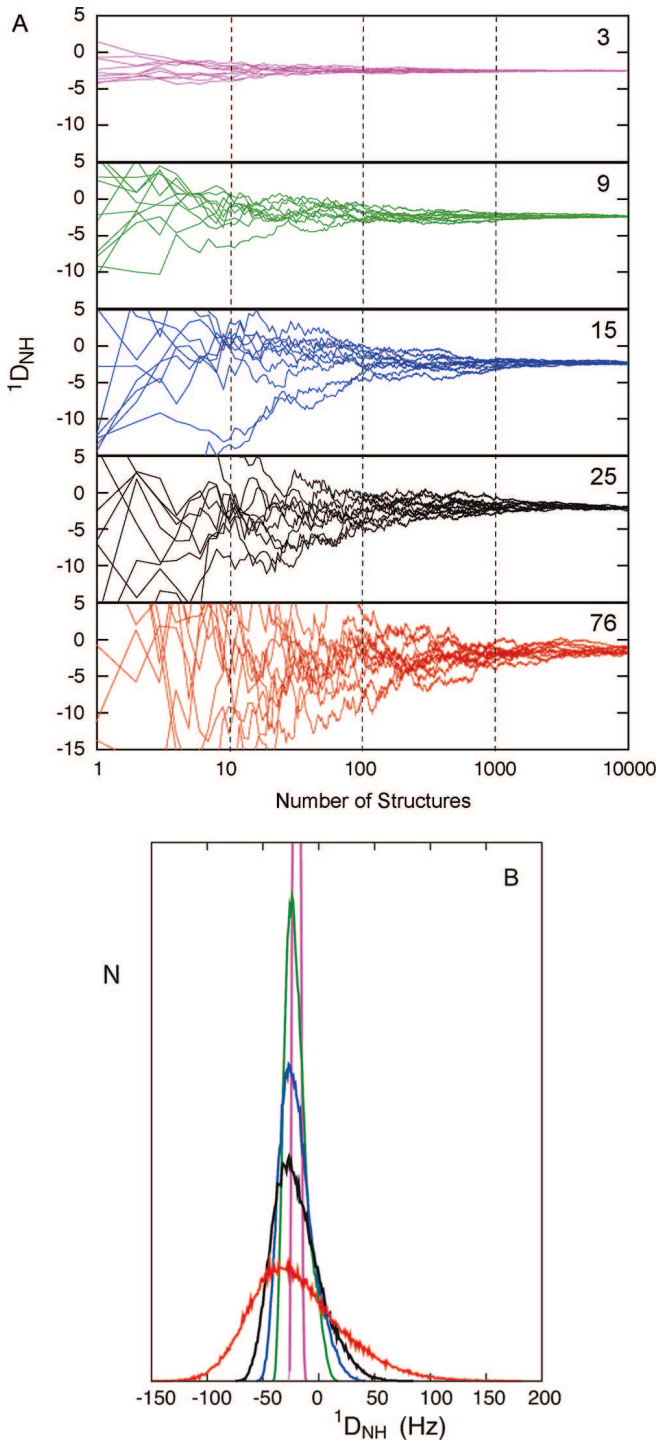


Figure 3. Convergence of $^{15}N-^1H^N$ RDCs calculated using LAWs of different lengths. (a) Comparison of 10 simulations of the central amino acid of an m amino acid LAW. The same $^1D_{NH}$ RDC (amino acid 41 of ubiquitin) is calculated using LAWs of $m = 3$, 9, 15, 25 or from the full length (76 amino acid) protein using a global alignment tensor. The x-axis represents the number of structures used to calculate the average. (b) Range and distribution of RDCs from the simulations shown in part a. Color code is the same in both cases (purple, three amino acid window; green, nine amino acids; blue, 15 amino acids; black, 25 amino acids; red, 76 amino acids).

reproduction of the RDCs improves rapidly with the number of structures included in the ensemble average. Simultaneously, the reproduction of the correct conformational sampling (the sampling used to simulate the RDC data) improves in all cases. These simulations, and those applied to the more extended

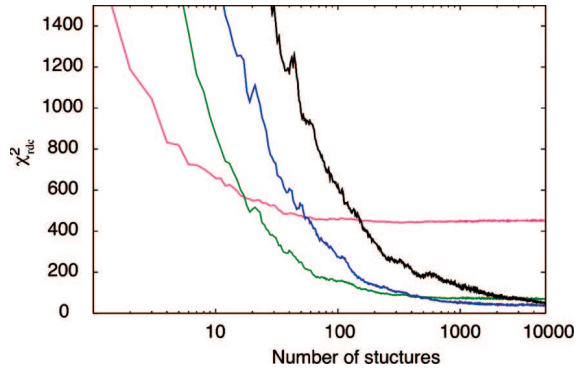


Figure 4. Accuracy of RDCs calculated using LAWs compared to a full length description. Equation 4 was used to directly compare the ability of RDCs calculated using the convolution of baseline and LAWs to reproduce RDCs calculated using an explicit description of the full length protein. The x-axis define the number of averaged RDCs. χ^2_{RDC} was calculated over the entire protein. Color code: purple, three amino acid window; green, nine amino acids; blue, 15 amino acids; black, 25 amino acids.

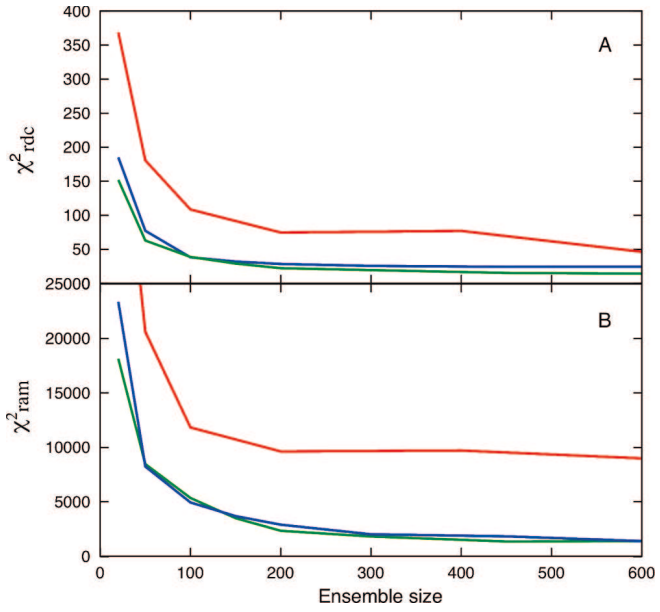


Figure 5. Accuracy of ensembles of structures calculated using LAWs of different lengths. The ability of ASTEROIDS to reproduce the correct conformational sampling and the correct RDCs for LAWs of different lengths is summarized as a function of the number of structures constituting the ensemble. (a) χ^2_{RDC} measures the reproduction of the target RDCs calculated using the full length 50 000-strong explicit description of the global alignment tensor. (b) χ^2_{Ram} measures the ability of the protocol to reproduce conformational sampling throughout the molecule. Color code: green, nine amino acid LAWs; blue, 15 amino acid LAWs; red, 76 amino acids (global alignment tensor). The x-axis define the number of structures used.

sampling regime (data not shown), indicate that the optimal combination for an accurate description of conformational behavior of the protein backbone requires a window length of at least 15 amino acids and 200 structures.

The site-specifi reproduction of the different RDCs comprising the χ^2_{RDC} using an ensemble of 200 and 20 structures is shown in Figure 7, for a LAW of 15 amino acids. Although the fi is significantl poorer in the case of 20 structures, the overall features are actually quite well reproduced, and the quality of the fi would probably be considered acceptable in the presence of commonly encountered levels of experimental noise. The conformational sampling is, however, very poorly reproduced, throughout the protein, when only 20 structures are

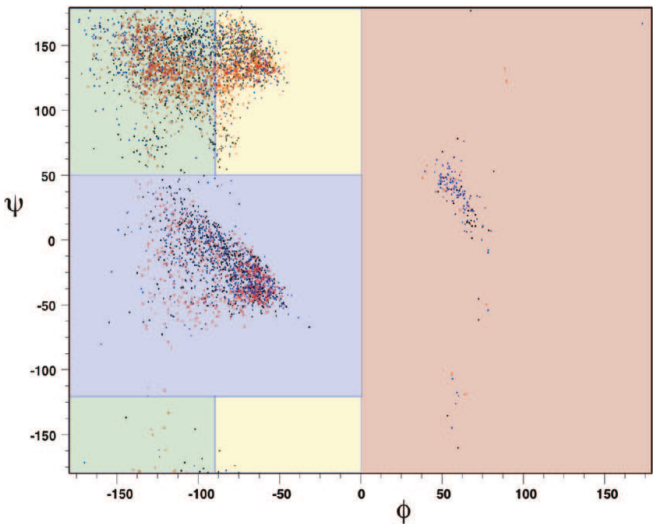


Figure 6. In order to quantify the similarity between conformational sampling between different ensembles, Ramachandran space is divided into four quadrants and define as follows: α_L, $\{\phi > 0°\}$; α_R, $\{\phi < 0, -120° < \psi < 50°\}$; β_P, $\{-90° < \phi < 0°, \psi > 50°$ or $\psi < -120°\}$; β_S, $\{-180° < \phi < -90°, \psi > 50°$ or $\psi < -120°\}$. The population of these quadrants is indicated as p_{α_L}, p_{α_R}, p_{β_P}, and p_{β_S}. Dots represent standard statistical coil distributions of valine (red), lysine (blue), and leucine (black).

included. This is graphically underlined in Figure 8, where the populations of the four quadrants of conformational space present in the 200- and 20-fold ensembles are compared with those present in the ensemble used to create the simulated data. Discrepancies in the population of the different quadrants of up to 30% compared to the value present in the original ensemble are found throughout the primary sequence for the 20-fold ensemble. These differences do not appear to be correlated to amino acid type. The 200-fold ensembles, on the other hand, closely reproduce the original sampling (figur 8b) for every region of primary sequence. It is therefore evident that, in cases where too few structures are included in the average, achieving acceptable reproduction of experimental data does not guarantee that the resulting ensemble accurately represents the correct conformational distribution.

Application of ASTEROIDS to Experimental RDCs from Urea-Unfolded Ubiquitin. Using the optimal parameters determined on the basis of the simulations described above, we have applied the ASTEROIDS approach to the determination of a representative ensemble to describe the conformational behavior of the protein ubiquitin under denaturing conditions (pH 2.5 and 8 M urea). In the initial analysis, ensembles of 200 structures were selected from a set of 12 000 conformers for which LAWs of 15 amino acids in length were used to calculate the dipolar couplings. The results, shown in Figure 9a, indicate a reasonable reproduction of experimental data but reveal notable systematic effects, in particular that the $D_{NHH\alpha(i-1)}$, $D_{NHNH(i+1)}$ RDCs are overestimated when the other couplings, effectively the $^1D_{NH}$ and $^1D_{C\alpha H\alpha}$ RDCs agree optimally with simulation. These observations agree qualitatively with identificatio of differential scaling of $^1H-^1H$ couplings compared to covalently bound spins in the analysis of these RDCs. In order to allow for this possibility in the current analysis, we allowed for two independent scaling factors, K_1 for the $^1D_{NH}$, $^1D_{C\alpha H\alpha}$, and $^1D_{C\alpha C'}$ and K_2 for the $D_{NHH\alpha}$, $D_{NHH\alpha(i-1)}$, $D_{NHNH(i+1)}$, and $D_{NHNH(i+2)}$. These factors are optimized uniformly for the covalently bound and through-space dipolar interactions, resulting in the data reproduction shown in Figure 9b. The two scaling factors $K_1 = 0.58$ and $K_2 = 0.96$ differ by approximately 0.6, a difference that may result from additional local conformational dynamics that are not taken into account by the statistical coil model and that scale the $D_{NHH\alpha(i-1)}$, $D_{NHNH(i+1)}$ RDCs differentially to the RDCs between spins whose distances are effectively fixed This possibility is currently under more detailed investigation.

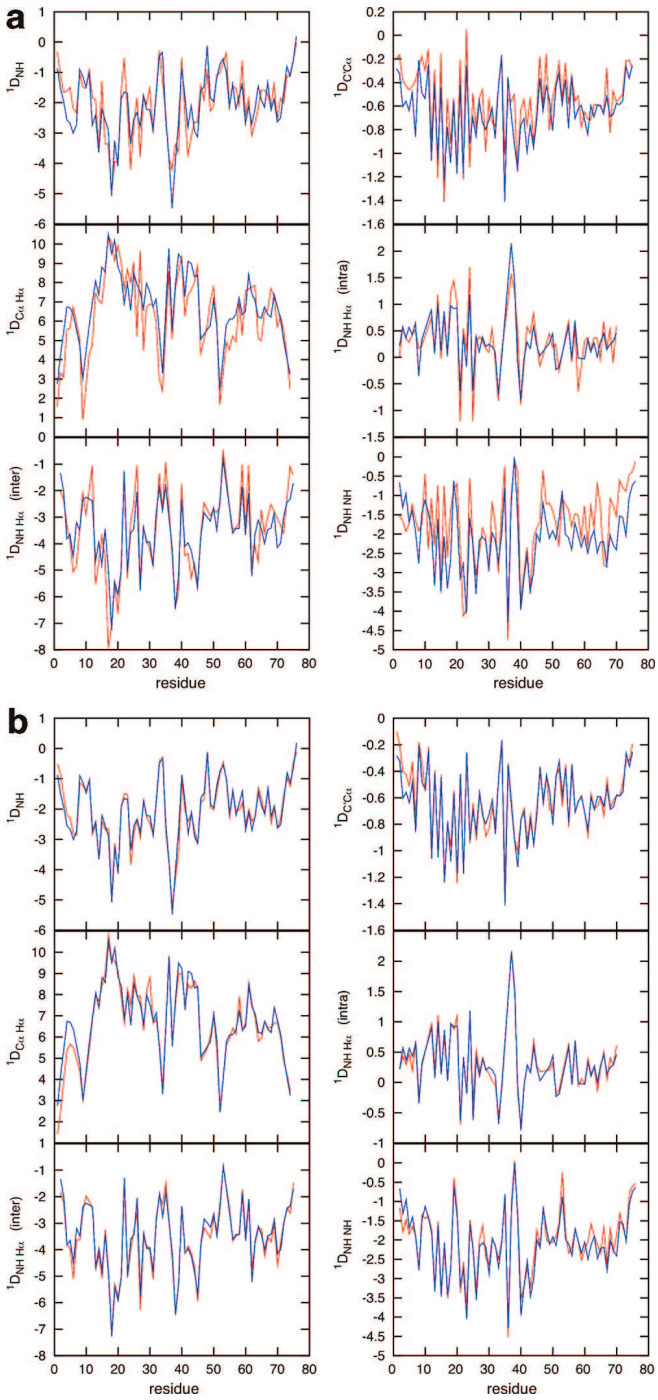


Figure 7. Site-specifi reproduction of the RDCs simulated using an explicit ensemble of 50 000 structures. (a) Reproduction of the target data (blue) using an ensemble of 20 structures (red) for a window length of 15 amino acids. (b) Reproduction of the target data (blue) using an ensemble of 200 structures (red) for a window length of 15 amino acids. In both cases, the genetic algorithm ASTEROIDS was used to select the optimal ensemble.

In order to test the validity of the approaches shown here for the analysis of experimental data, we have repeated the ASTEROIDS ensemble selection procedure, taking 10% of the

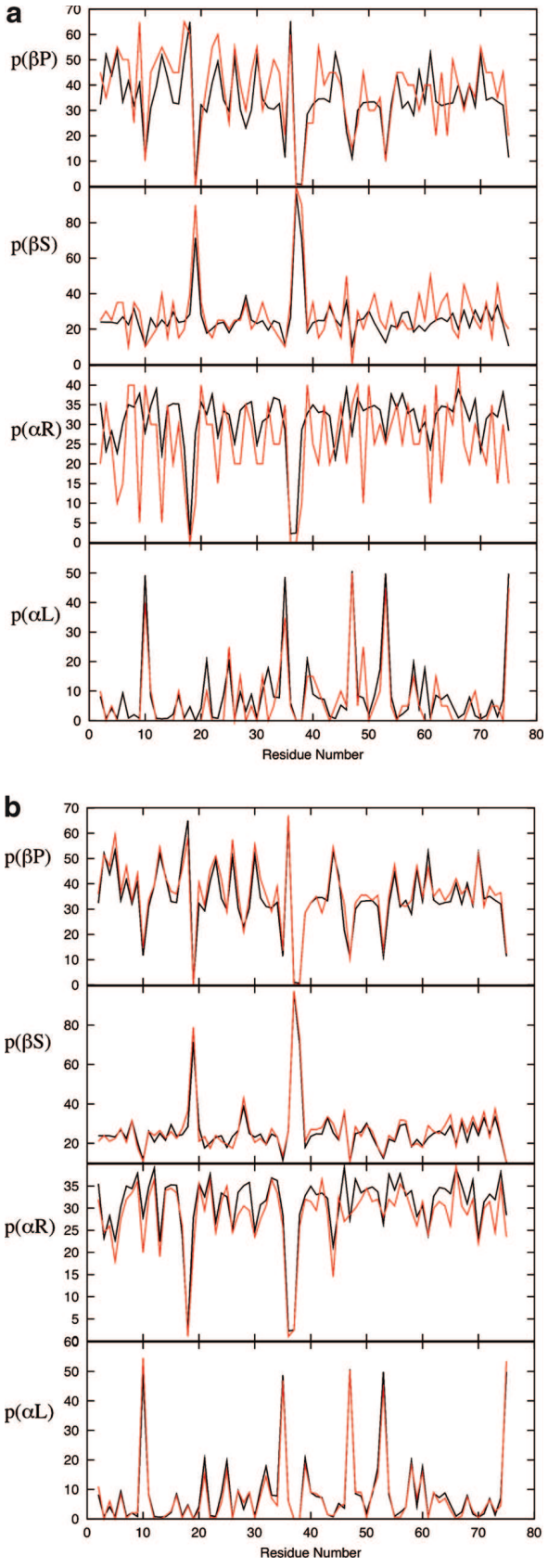


Figure 8. Accuracy of the reproduction of conformational sampling using the ASTEROIDS approach with ensembles of 20 and 200 structures. Populations of the four quadrants of conformational space define in Figure 6 using the (a) 20-fold and (b) 200-fold ensembles (red) compared with those present in the ensemble used to create the simulated data (black). Discrepancies in the population of the different quadrants of up to 30% compared to the value present in the original ensemble are found for ensembles of size 20.

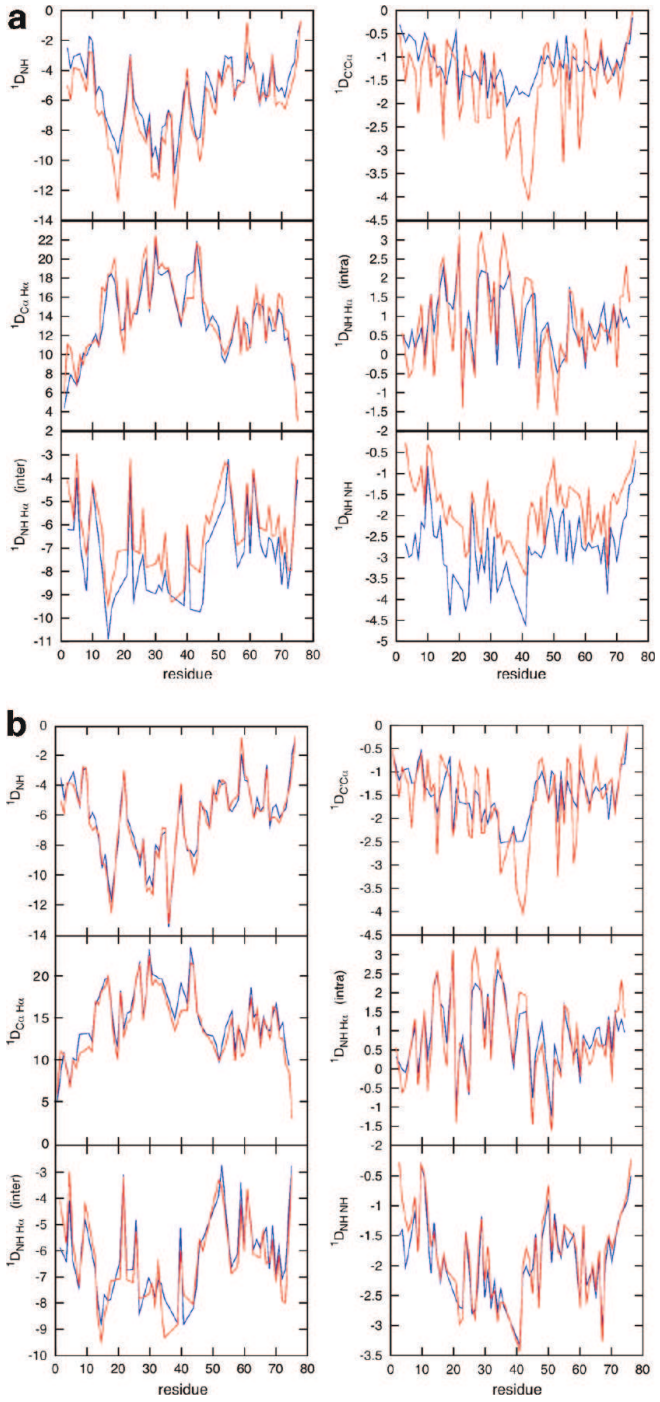


Figure 9. Application of ASTEROIDS to experimental RDCs from urea-unfolded ubiquitin. (a) Reproduction of experimental data (red) using an ensemble of 200 structures (blue). (b) Reproduction of experimental data (red) using an ensemble of 200 structures (blue) with differential scaling of the covalently bound and interproton RDCs.

RDCs out of the analysis and comparing the predicted values using the resulting ensemble with the experimental RDCs. The results are shown in Figure 10, where the back-calculated RDCs are found to be in reasonable agreement with the experimentally determined values. The calculation was repeated 10 times at seven different ensemble sizes. The average cross-validated χ^2 is plotted as a function of ensemble size (Figure 10b). The size of 200-fold ensembles used in the current approach is within the range where the cross validation target function is essentially flat

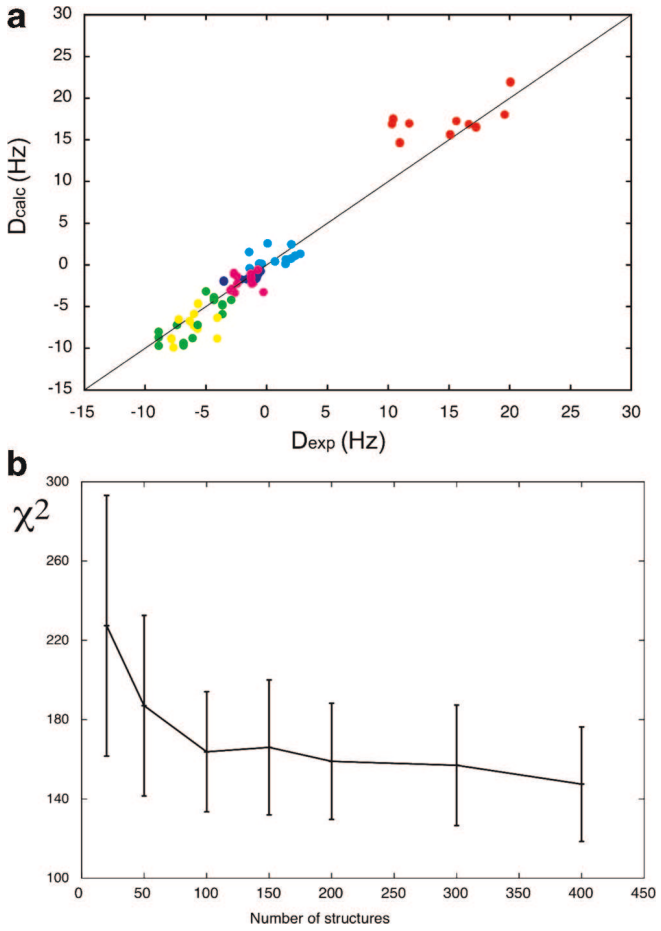


Figure 10. Reproduction of data not used in the fittin procedure. (a) The ASTEROIDS ensemble selection procedure was repeated, taking 10% of the RDCs out of the analysis and comparing the predicted values using the resulting ensemble with the experimental RDCs. Color code: green, $^1D_{NH}$; red, $^1D_{C\alpha H\alpha}$; dark blue, $^1D_{C\alpha C'}$; cyan, $D_{NHH\alpha}$; yellow, $D_{NHH\alpha(i-1)}$; magenta, $D_{NHNH(i+1)}$.(b) Average χ^2 over 10 cross-validation calculations at each of seven different ensemble sizes. The 200-fold ensemble size used in the current approach is within the range where the cross-validation target function is essentially flat

The precision with which the RDCs can defin the conformational behavior of the backbone has been assessed using noise-based Monte Carlo simulations (see Experimental Section) based on estimates of experimental uncertainty. The results are summarized in Figure S1 of the Supporting Information, and show that the average uncertainty in the populations of the different quadrants is approximately ±3%. We have also repeated the entire analysis in the absence of one experimental data set to assess the relative importance of each data set for the conformational description. The results are shown in Figure S2 and summarized in Tables S1 and S2 of the Supporting Information, where the backbone sampling is compared to the populations determined using all data. The root-mean-square deviation of the four populations define in Figure 6 and the average differences demonstrate that although we fin that the most important RDCs are the $D_{NHHai,i+1}$ and D_{NHNH}, the effects are actually not very large when these RDCs are removed (maximum rmsd of 5%, and average difference in populations of 3%). These results suggest that both covalently bound and interproton RDCs are important for an accurate description of conformational sampling but that none of the RDC types are critical for the validity of the description or the conclusions drawn from it.

The amino acid Ramachandran sampling has been used to calculate expected $^3J_{NHH\alpha}$ scalar couplings, reporting on the sampling of the ϕ backbone dihedral angle. These values have been compared to experimentally determined couplings[47] (Figure S3, Supporting Information), in comparison to the reproduction of the data using the standard coil database. The *J*-coupling data reproduction is quite good in both cases, but only slightly better in the case of the selected ensemble ($\chi^2 = 11.5$ compared to 12.6), probably reflectin the fact that the differences in the two descriptions are often found in the distribution of the ψ backbone dihedral angle. However, this analysis does demonstrate that the local analysis of RDCs in terms of Ramachandran distributions does not contradict independent experimental data in a significan way.

Urea Preferentially Affects the Conformational Sampling of Amino Acids with Side Chain Hydrogen-Bonding Moieties. Figure 11 shows the backbone dihedral angle distributions resulting from the analysis of experimental data of urea-unfolded ubiquitin and the normalized difference compared to the distribution of angles derived using an ensemble of structures produced using the standard statistical coil model of the unfolded state. Figure S4 of the Supporting Information shows the amino acid specifi populations of all amino acids for the standard statistical coil model. The sampling of the different regions of the Ramachandran space define in Figure 6 is summarized in Figure 12.

In general, the results indicate that the sampling of backbone dihedral angles in Ramachandran space is more extended, sampling the β_P and β_S regions with higher propensity and the α_R region with lower propensity than the statistical coil database. This result is in agreement with a previous study of the more general characteristics of conformational sampling, using the same experimental data.[34] In this study, a hypothesis-driven approach was used to suggest a general extension of conformational sampling of the peptide chain. With the new techniques developed here, we are able to extract amino acid-specifi conformational sampling directly from the RDC data. This approach relies on the supposition that the database from which structures are selected contains enough conformational diversity to allow for a representative description to be constructed from its population. Under these conditions, the method is relatively hypothesis-free in comparison to previous approaches. This reveals that the effects of urea on backbone conformational sampling are far from uniform. The extended nature of the chain is more apparent in localized contiguous segments of primary sequence: the regions 30−36 and 70−73 sample the β_P region more extensively than both the statistical coil and the remainder of the protein, while extended β regions are preferentially sampled in the region 14−18. This latter tendency may be correlated with the previously observed presence of a small (around 20%) residual population of β hairpin in this region of the molecule.[48] Amino acids preceding prolines (18 and 36) are found to better reproduce experimental RDCs with a more uniform sampling of propensities in the β_P and β_S regions, compared to the statistical coil database that preferentially samples the α_R region.

The comparison with the statistical coil model clarifie detail that may be masked by amino acid-specifi sampling of backbone dihedral angle and allows the identificatio of sites

(47) Peti, W.; Henning, M.; Smith, L. J.; Schwalbe, H. *J. Am. Chem. Soc.* **2000**, *122*, 12017–12018.
(48) Meier, S.; Strohmeier, M.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2007**, *129*, 754–755.

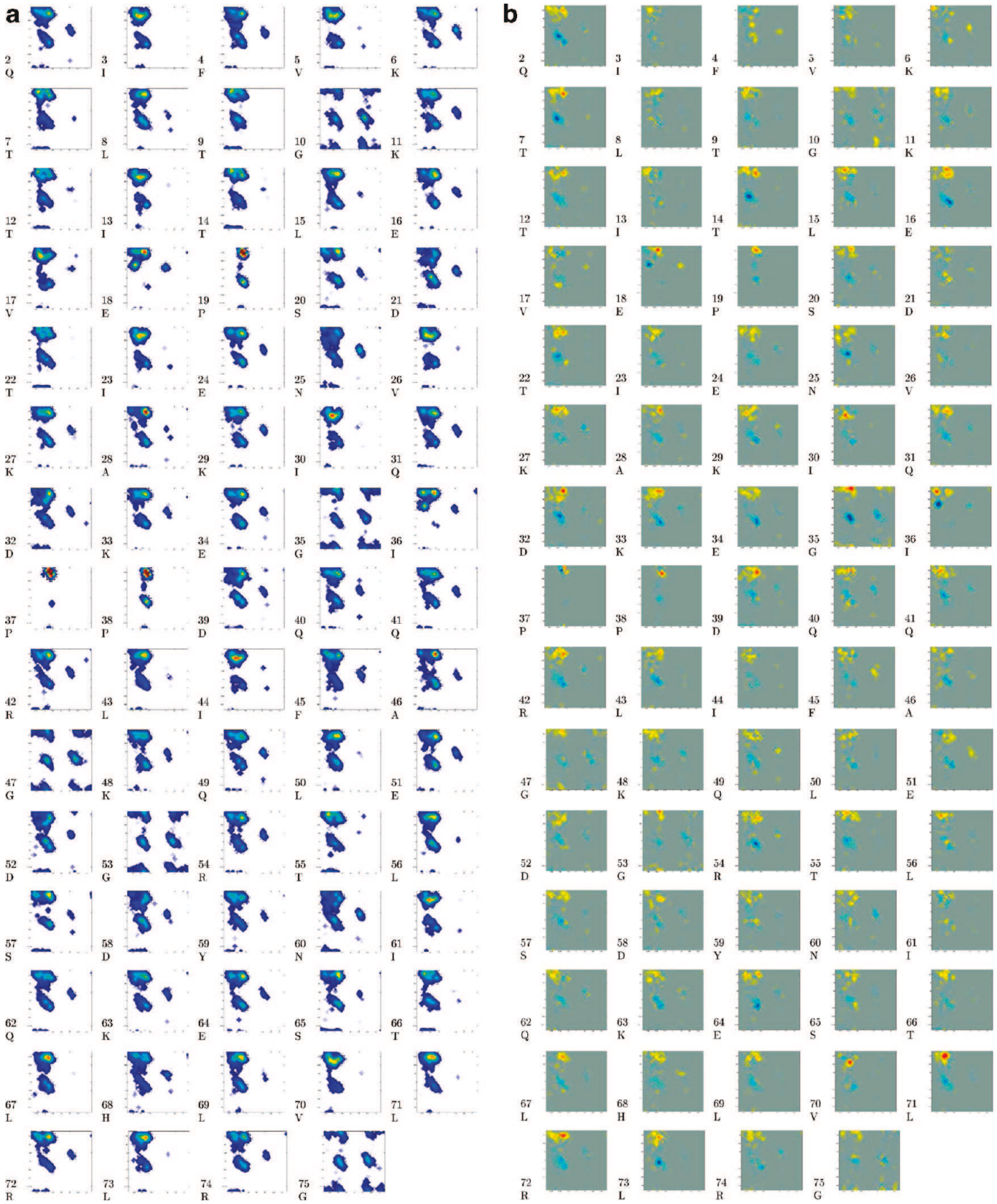


Figure 11. Amino acid-specifi Ramachandran distributions for unfolded ubiquitin in 8 M urea at pH 2.5 in comparison with a standard statistical coil distribution. The populations increase from dark blue, via cyan, green, and yellow, to red. (a) Conformational sampling determined from the ASTEROIDS analysis of experimental RDC data (10 calculations were combined to produce 2000 conformers for the sake of figur resolution). (b) Difference between the conformational sampling distributions shown in panel a and the conformational sampling for the flexible-meccan statistical coil distribution. In this case, blue to green corresponds to negative values (population is lower in the urea unfolded sampling than in the statistical coil) and green (via yellow) to red corresponds to positive values (population is higher in the urea-unfolded sampling than in the statistical coil). Gray corresponds to equal populations.

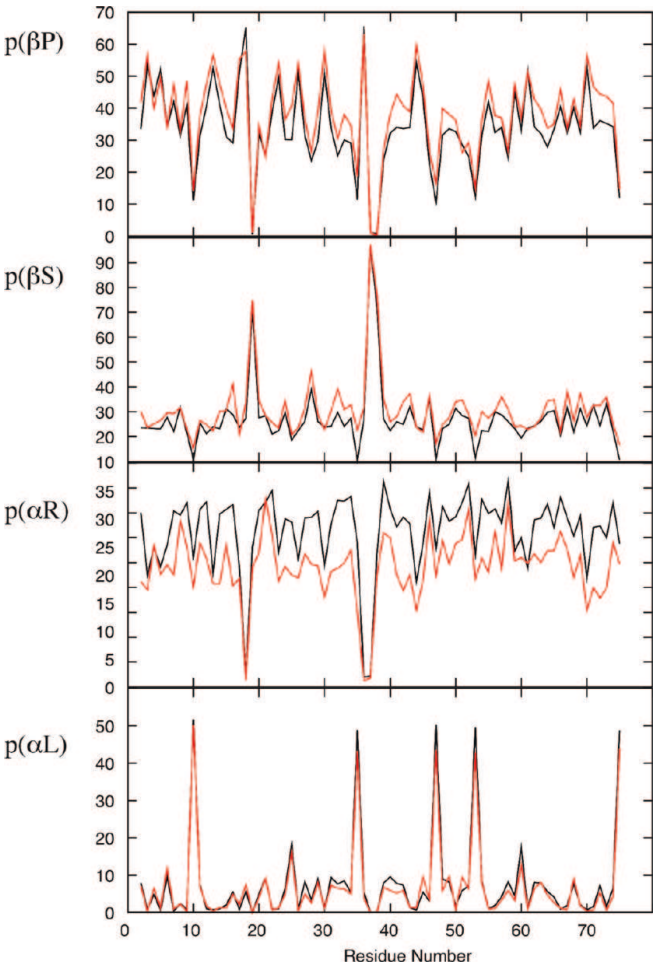


Figure 12. Populations of the four quadrants of conformational space define in Figure 6 using the amino acid-specifi Ramachandran distributions for unfolded ubiquitin in 8 M urea at pH 2.5 shown in Figure 11 (red) in comparison to a standard statistical coil distribution (black).

whose behavior deviates from random coil in the presence of urea. In this context, it is interesting to note that the amino acids whose backbone conformational sampling are most systematically affected by the presence of urea are threonine (four out of seven have a notably more extended backbone sampling than in the statistical coil model), glutamic acid (three out of fiv are more extended than the statistical coil model), and arginine (three out of four are more extended than the statistical coil model). These amino acids all contain potential hydrogen-bond-donor moieties on their side chains. A recent study using vibrational spectroscopy demonstrated that at low pH urea orients with the carboxyl group pointing toward the protein surface, an observation that supports the suggestion that hydrogen-bond-donor groups may interact preferentially with urea.[49] By contrast, only three of a total of 24 hydrophobic amino acids (valine, leucine, isoleucine, alanine, tyrosine, and phenylalanine) exhibit significantl different conformational sampling between the urea-denatured and the statistical coil states. The specifi amino acid composition may therefore be responsible for the apparent localization of differential backbone sampling properties in the different regions of the protein. A recent study used small angle scattering to estimate the number of additional urea molecules that are preferentially recruited during the unfolding transition of ubiquitin from neutral to acidic pH to be approximately 20, a number that correlates qualitatively with the observation here that the backbone behavior of approximately a third of the amino acids are preferentially affected by the presence of urea.[38]

Conclusions

In this study, we have used extensive simulation to optimize an approach that exploits experimental RDCs measured from unfolded proteins to determine conformational sampling on an amino acid-specifi basis. Previous applications have used a full-length description of the protein, averaging RDCs over an unrestrained ensemble that is large enough to allow for convergence of the coupling values. Although providing important insight into the behavior of a number of disordered proteins for which conformational information is otherwise difficul to measure, these studies are hypothesis-based, testing different conformational sampling regimes and comparing them to experimental data, an approach that severely limits both the scope and application as well as the potential for discovery. Here we develop a general approach that allows one to select an ensemble directly from the experimental data. Our combination of analytical baseline descriptor and numerical averaging of smaller alignment windows is tested against simulation, and on the basis of these simulations, parameters such as window length and number of structures are calibrated. We fin that a combination of LAWs of 15 amino acids in length, with ensemble sizes of 200, accurately describes conformational space, while ensembles of 20 structures reproduce the experimental data but, critically, do not reproduce the correct conformational sampling. Using this approach we can describe conformational sampling at an amino acid resolution.

These approaches have been applied to the amino acid-specifi description of backbone conformational sampling in ubiquitin denatured in 8 M urea at pH 2.5. Having established the precision that the approach is expected to offer, we are able to analyze in fin detail the local conformational differences between the standard statistical coil description and the sampling define by the experimental data measured in the presence of urea, and we interpret this in the context of urea binding or interacting with specifi types of amino acids in the peptide chain.

Experimental Section

Experimental methods for measuring the RDCs included in the analysis have been presented elsewhere. All data were taken from the earlier study by Meier et al.[34]

Flexible-Meccano Calculations. Simulated RDCs were calculated using the program flexible-meccan interfaced to the program PALES[50] as described. The program was run in two modes: For calculations using a global alignment tensor for the entire molecule, the standard procedure was used. For calculations using the local alignment windows (LAWs) the RDC for the central amino acid of the local *m* amino acid segment (3, 9, 15, or 25) was calculated for each individual structure. For the terminal amino acids, alanine amino acids were added to the N or C terminus during the building of the protein, such that the *m* amino acid segment was always present. The resulting RDC profil along the primary sequence is calculated by averaging each value and multiplying with the effective baseline given in eq 1. If RDCs were calculated using the full length protein, they were averaged over all conformers as previously described.

(49) Chen, X.; Sagle, L. B.; Cremer, P. S. *J. Am. Chem. Soc.* **2007**, *129*, 15104–15105.

(50) Zweckstetter, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 3791–3792.

A pool of 12 000 structures is generated with flexible-meccano Half of the structures were calculated using the standard statistical coil model S, and the other half using a more extended regime E. The sampling regime (E) samples a more extended region of Ramachandran space, populating the region $\{50° < \psi < 180°\}$ with a higher propensity than the S regime (78% compared to 59%).

ASTEROIDS Ensemble Selection. ASTEROIDS uses a genetic algorithm[51−53] to build a representative ensemble of structures of fixe size N from a large database. The algorithm selects an ensemble of N structures using the following fitnes function compared to the experimental data.

$$\chi^2_{\text{asteroids}} = \sum_i w_i^2 (D_{i,\text{calc}} - D_{i,\text{exp}})^2 \quad (2)$$

where w_i is the weight of coupling D_i. The weights were set according to coupling type and determined by the range of each type of coupling in hertz. Values of w were set to 1.0 for $^1D_{\text{NH}}$ and $D_{\text{NHH}\alpha(i-1)}$, 0.5 for $^1D_{\text{C}\alpha\text{H}\alpha}$, 2.0 for $^1D_{\text{C}\alpha\text{C}'}$, $D_{\text{NHH}\alpha}$, and $D_{\text{NHNH}(i+1)}$, and 3.0 for $D_{\text{NHNH}(i+2)}$. The fina ensemble is obtained from generations of ensembles that undergo evolution and selection using this fitnes function. Each generation comprises 100 different ensembles of size N.

Evolution can proceed in three different ways: random, mutation, and crossing. At each evolution step, the protocol ensures that a structure does not appear more than once in a given ensemble and that a given ensemble is not repeated in a generation. Random evolution proceeds by randomly selecting structures in the complete database. Mutation occurs by taking an ensemble and replacing 1% of the structures (or at least one structure) by structures randomly selected from the complete database (external mutation) or from a new database containing all the structures selected at least once in the previous generation (internal mutation). Crossing is achieved by randomly pairing ensembles from the previous generation. New ensembles are generated by selecting N structures in a pool made of the structures present in the previously define pairs.

The firs generation is always obtained using random evolution. Evolution of this generation is achieved by the following procedure. New ensembles are generated (100 by random evolution, 100 by external mutation, 100 by internal mutation and 100 by crossing). Among these new ensembles and the previous generation, 100 different ensembles representing minima with respect to the fitnes function are selected using tournaments to provide the next generation. Ensembles are randomly split into groups and then ordered using the fitnes function to determine the winners of the tournament. The best ensembles of each tournament are retained to form the next generation. The number of tournaments and the number of winners of each tournament are adjusted such that 100 ensembles are selected. Selection pressure increases as the number of tournaments decreases. To avoid premature convergence in local minima, the selection pressure is gradually increased during evolution. The number of tournaments therefore successively goes from 100 to 50, 25, 20, 10, 2, and to 1. To ensure robustness of the fittin procedure, the evolution and selection processes are repeated over 2000 successive generations.

Ramachandran Segment Division. In order to describe the sampling of conformational space in the different ensembles and their agreement with known distributions, Ramachandran space is divided into four quadrants indicated in Figure 6 and define as follows: α_L, $\{\phi > 0°\}$; α_R, $\{\phi < 0, -120° < \psi < 50°\}$; β_P, $\{-90° < \phi < 0°, \psi > 50°$ or $\psi < -120°\}$; β_S, $\{-180° < \phi < -90°, \psi > 50°$ or $\psi < -120°\}$.

The population of these quadrants is indicated as p_{α_L}, p_{α_R}, p_{β_P}, and p_{β_S}. The Ramachandran similarity factor χ^2_{Ram} of the entire molecule is measured by the following function:

$$\chi^2_{\text{Ram}} = \sum_i \sum_q (p_{i,q,\text{ref}} - p_{i,q,\text{fit}})^2 \quad (3)$$

where p_q are the four different populations of the quadrants q, i are the different amino acids, and $_{\text{ref}}$ and $_{\text{fi}}$ signify the target and fitte Ramachandran distributions.

Comparison of RDCs. In order to compare RDCs calculated using different window lengths with those calculated using 50 000 conformers from the full length description of the protein, the following function χ^2_{RDC} is used:

$$\chi^2_{\text{RDC}} = \sum_i (D_{i,\text{LAW}} - D_{i,\text{fl}})^2 \quad (4)$$

where $D_{i,\text{LAW}}$ represents the RDC calculated using LAWs, after multiplication with the baseline function given in eq 1, and $D_{i,\text{f}}$ is the RDC calculated using the full length description.

Monte Carlo Simulations and Error Analysis. In order to estimate the precision with which the conformational sampling can be define on the basis of experimental RDCs, we have run noise-based Monte Carlo simulations, using random sampling of Gaussian distributions whose width is based on experimentally estimated uncertainties for each RDC. Fifty Monte Carlo simulations were run, and the effective uncertainty of the Ramachandran quadrant population was calculated on the basis of this.

In order to estimate the importance of the different RDC types, we have repeated the analysis of experimental data with one entire data set removed from the ASTEROIDS approach.

***J*-Coupling Analysis.** $^3J_{\text{NHH}\alpha}$ scalar couplings were calculated by averaging over the amino acid-specifi ϕ backbone dihedral angle distributions and compared to experimentally measured values, using recently derived Karplus relationships.[54]

Acknowledgment. L.S. received a grant from the French Ministry of Education. This work was supported by the French Research Ministry through ANR-PCV07_194985. M.R.J. benefite from an EMBO fellowship and Lundbeckfonden support.

Supporting Information Available: A figur showing the standard statistical coil distribution on a residue-specifi basis. Residue-specifi populations of Ramachandran space resulting from Monte Carlo simulations. A figur and tables showing conformational sampling of the different quadrants of Ramachandran space when specifi RDC types are removed. A figur showing calculated and experimental ^3J scalar couplings. This material is available free of charge via the Internet at http://pubs.acs.org.

JA9069024

(51) Fraser, A. S. *Austr. J. Biol. Sci.* **1957**, *10*, 484–491.
(52) Holland, J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, 1975.
(53) Jones, G. *Genetic and Evolutionary Algorithms. Encyclopedia of Computational Chemistry*; Wiley: Chichester, U.K., 1998.
(54) Markwick, P. R. L.; Showalter, S. A.; Bouvignies, G.; Brüschweiler, R.; Blackledge, M. *J Biomol NMR* **2009**, *45*, 17–21.

J|A|C|S ARTICLES

NMR Characterization of Long-Range Order in Intrinsically Disordered Proteins

Loïc Salmon,[†,§] Gabrielle Nodet,[†,§] Valéry Ozenne,[†] Guowei Yin,[‡] Malene Ringkjøbing Jensen,[†] Markus Zweckstetter,[‡] and Martin Blackledge*[,†]

Protein Dynamics and Flexibility, Institut de Biologie Structurale Jean-Pierre Ebel, CEA; CNRS; UJF UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, France, and NMR-Based Structural Biology, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany

Received February 25, 2010; E-mail: martin.blackledge@ibs.fr

Abstract: Intrinsically disordered proteins (IDPs) are predicted to represent a significant fraction of the human genome, and the development of meaningful molecular descriptions of these proteins remains a key challenge for contemporary structural biology. In order to describe the conformational behavior of IDPs, a molecular representation of the disordered state based on diverse sources of structural data that often exhibit complex and very different averaging behavior is required. In this study, we propose a combination of paramagnetic relaxation enhancements (PREs) and residual dipolar couplings (RDCs) to define both long-range and local structural features of IDPs in solution. We demonstrate that ASTEROIDS, an ensemble selection algorithm, faithfully reproduces intramolecular contacts, even in the presence of highly diffuse, ill-defined target interactions. We also show that explicit modeling of spin-label mobility significantly improves the reproduction of experimental PRE data, even in the case of highly disordered proteins. Prediction of the effects of transient long-range contacts on RDC profiles reveals that weak intramolecular interactions can induce a severe distortion of the profiles that compromises the description of local conformational sampling if it is not correctly taken into account. We have developed a solution to this problem that involves efficiently combining RDC and PRE data to simultaneously determine long-range and local structure in highly flexible proteins. This combined analysis is shown to be essential for the accurate interpretation of experimental data from α-synuclein, an important IDP involved in human neurodegenerative disease, confirming the presence of long-range order between distant regions in the protein.

Introduction

The realization that a large fraction of functional proteins encoded by the human genome are intrinsically disordered or contain long disordered regions has revealed a fundamental limitation of classical structural biology.[1−4] Intrinsically disordered proteins (IDPs) are functional despite their lack of well-define structure, imposing a new perspective on the relationship between primary protein sequence and function and necessitating the development of an entirely new set of experimental and analytical techniques.[5,6] The importance of developing new methodologies to study these proteins is underlined by the fact that IDPs are associated with many human diseases, including cancer, cardiovascular disease, amyloidosis, neurodegenerative disease, and diabetes.

NMR spectroscopy is exquisitely suited to the study of IDPs,[7] primarily because heteronuclear chemical shift assignment remains possible even for very large disordered proteins.[8] NMR analysis can then be used to precisely study the specifi local conformational preferences that encode biological function.[9−11] In spite of their highly dynamic nature, IDPs also exhibit transient or persistent long-range tertiary structure that may be related to biological activity (e.g., via so-called fly-castin interactions[12]) or simply confer protection from proteolysis or amyloidosis. It is precisely the transient nature of such contacts that precludes straightforward NMR detection using standard techniques such as $^{1}H-^{1}H$ cross-relaxation. However, long-range information can be measured via the effects of dipolar relaxation between the observed spin and an unpaired electron, which can be artificiall introduced into the protein by attaching a nitroxide group to a strategically placed cysteine mutant.[13,14]

[†] Institut de Biologie Structurale Jean-Pierre Ebel.
[‡] Max Planck Institute for Biophysical Chemistry.
[§] These authors contributed equally.

(1) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739–756.
(2) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. *Biochemistry* **2002**, *41*, 6573–6582.
(3) Tompa, P. *Trends. Biochem. Sci.* **2002**, *27*, 527–533.
(4) Dyson, H. J.; Wright, P. E. *Curr. Opin. Struct. Biol.* **2002**, *12*, 54–60.
(5) Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 3–14.
(6) Eliezer, D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 23–30.
(7) Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, *104*, 3607–3622.
(8) Mukrasch, M. D.; Bibow, S.; Korukottu, J.; Jeganathan, S.; Biernat, J.; Griesinger, C.; Mandelkow, E. M.; Zweckstetter, M. *PLoS Biol.* **2009**, *7*, 399–414.
(9) Meier, S.; Blackledge, M.; Grzesiek, S. *J. Chem. Phys.* **2008**, *128*, 052204.
(10) Wright, P. E.; Dyson, H. J. *Curr. Opin. Struct. Biol.* **2009**, *19*, 31–38.
(11) Jensen, M. R.; Markwick, P.; Griesinger, C.; Zweckstetter, M.; Meier, S.; Grzesiek, S.; Bernado, P.; Blackledge, M. *Structure* **2009**, *17*, 1169–1185.
(12) Shoemaker, B. A.; Portman, J. J.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 8868–8873.
(13) Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 158–169.
(14) Clore, G. M.; Tang, C.; Iwahara, J. *Curr. Opin. Struct. Biol.* **2007**, *17*, 603–616.

The gyromagnetic ratio of the electron spin is sufficientl high that the observed line broadening due to the paramagnetic relaxation enhancement (PRE) affords sensitive long-range probes of intra- and intermolecular distances and distance distribution functions. The interpretation of experimental PREs can be relatively straightforward in the case of folded proteins, where an assumption of a static probe localized at a single point in space can be applied to extract approximate distance constraints.[15] It has also been shown that simple modeling of spin-label side-chain mobility in terms of an average over three positions can significantl improve the accuracy of the distance information.[16] Detailed information about transient encounter complexes and their role in protein−protein interactions can also be extracted by combining paramagnetic effects and ensemble-averaged restrained molecular dynamics (MD).[17−19]

In the case of partially folded and unfolded proteins, paramagnetic effects are particularly powerful, as the interactions are sufficientl strong to allow the identificatio of fluctuating weakly populated tertiary structural contacts. In this case, the treatment of the intrinsic dynamics of the system is of considerable importance. PREs have thus been interpreted in terms of average distance restraints between the unpaired electron and the observed spin, and these distances have been incorporated directly as constraints into restrained MD or ensemble-averaged restrained MD approaches.[20−27] Explicit relaxation rates can also be incorporated as constraints,[28] and more recently, PREs have been interpreted in terms of probability distributions.[26,29,30] PREs can also be used to select representative ensembles from a large pool of possible conformers.[31−33]

In this study, we have applied to the interpretation of PRE data from disordered proteins a recently introduced approach for modeling highly dynamic and disordered systems that derives explicit molecular ensembles determined on the basis of experimental data. Ensemble selection is based on the creation of a large number of conformers using an amino acid-specifi random coil database known as *flexible-meccano*.[34] *Flexible-meccano* allows for very efficien restraint-free sampling of the available conformational space and was initially demonstrated and refine to provide structural ensembles in agreement with experimentally measured NMR and small-angle X-ray scattering (SAXS) data.[35−42] In parallel, the ensemble selection algorithm ASTEROIDS has been developed to directly determine appropriate regions of conformational space populated by the IDP through selection of conformers from the *flexible-meccano* ensemble using inferential analysis of experimental NMR data.[43] To date, the approach has been applied to experimental measurements that depend essentially on local structural behavior, such as residual dipolar couplings (RDCs) and chemical shifts.[44] Here we have adapted the approach to incorporate the interpretation of PREs. In order to allow for flexibilit of the spin label with respect to the backbone conformation, explicit rotameric libraries that have been parametrized against experimental electron spin resonance (ESR) measurements and MD simulations[45] are used to map the allowed position of the electron spin. We then account for the dynamics of the electron spin within this envelope by evoking a model for the autocorrelation function of the relaxation-active interaction that was originally proposed for the interpretation of $^1H-^1H$ cross-relaxation effects.[46] This allows the motion of the relaxation-active dipole−dipole interaction between the electron spin and the observed nucleus to be modeled for each conformer in the ensemble.

The observation that RDCs can be measured in disordered proteins has been followed by the rapid development of techniques for interpreting experimental data in terms of local structure.[38,40,41,47−60] Comparison of experimental data with

(15) Battiste, J. L.; Wagner, G. *Biochemistry* **2000**, *39*, 5355–5365.
(16) Iwahara, J.; Schwieters, C. D.; Clore, G. M. *J. Am. Chem. Soc.* **2004**, *126*, 5879–5896.
(17) Tang, C.; Schwieters, C. D.; Clore, G. M. *Nature* **2007**, *449*, 1078–1082.
(18) Volkov, A. N.; Worrall, J. A.; Holtzmann, E.; Ubbink, M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18945–18950.
(19) Bashir, Q.; Volkov, A. N.; Ullmann, G. M.; Ubbink, M. *J. Am. Chem. Soc.* **2010**, *132*, 241–247.
(20) Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 170–184.
(21) Lindorff-Larsen, K.; Kristjansdottir, S.; Teilum, K.; Fieber, W.; Dobson, C. M.; Poulsen, F. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 3291–3299.
(22) Dedmon, M. M.; Lindorff-Larsen, K.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. *J. Am. Chem. Soc.* **2005**, *127*, 476–477.
(23) Bertoncini, C. W.; Jung, Y. S.; Fernandez, C. O.; Hoyer, W.; Griesinger, C.; Jovin, T. M.; Zweckstetter, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1430–1435.
(24) Kristjansdottir, S.; Lindorff-Larsen, K.; Fieber, W.; Dobson, C. M.; Vendruscolo, M.; Poulsen, F. M. *J. Mol. Biol.* **2005**, *347*, 1053–1062.
(25) Song, J.; Guo, L. W.; Muradov, H.; Artemyev, N. O.; Ruoho, A. E.; Markley, J. L. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1505–1510.
(26) Allison, J. R.; Varnai, P.; Dobson, C. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 18314–18326.
(27) Ganguly, D.; Chen, J. *J. Mol. Biol.* **2009**, *390*, 467–477.
(28) Huang, J.-R.; Grzesiek, S. *J. Am. Chem. Soc.* **2010**, *132*, 694–705.
(29) Felitsky, D. J.; Lietzow, M. A.; Dyson, H. J.; Wright, P. E. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 6278–6283.
(30) Xue, Y.; Podkorytov, I. S.; Rao, D. K.; Benjamin, N.; Sun, H.; Skrynnikov, N. R. *Protein Sci.* **2009**, *18*, 1401–1424.
(31) Marsh, J. A.; Neale, C.; Jack, F. E.; Choy, W.-Y.; Lee, A. Y.; Crowhurst, K. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2007**, *367*, 1494–1510.
(32) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359–374.
(33) Cho, M. K.; Nodet, G.; Kim, H. Y.; Jensen, M. R.; Bernado, P.; Fernandez, C. O.; Becker, S.; Blackledge, M.; Zweckstetter, M. *Protein Sci.* **2009**, *18*, 1840–1846.
(34) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.
(35) Bernado, P.; Bertoncini, C.; Griesinger, C.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2005**, *127*, 17968–17969.
(36) Dames, S. A.; Aregger, R.; Vajpai, N.; Bernado, P.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2006**, *128*, 13508–13514.
(37) Skora, L.; Cho, M. K.; Kim, H.-Y.; Fernandez, C.; Blackledge, M.; Zweckstetter, M. *Angew. Chem., Int. Ed.* **2006**, *45*, 7012–7015.
(38) Mukrasch, M. D.; Markwick, P. R. L.; Biernat, J.; von Bergen, M.; Bernado, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 5235–5243.
(39) Meier, S.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 9799–9807.
(40) Jensen, M. R.; Houben, K.; Lescop, E.; Blanchard, L.; Ruigrok, R. W. H.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 8055–8061.
(41) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 5762–5767.
(42) Bernado, P.; Blackledge, M. *Biophys. J.* **2009**, *97*, 2839–2845.
(43) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 16968–16975.
(44) Jensen, M. R.; Salmon, L.; Nodet, G.; Blackledge, M. *J. Am. Chem. Soc.* **2010**, *132*, 1270–1272.
(45) Sezer, D.; Freed, J. H.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 5755–5767.
(46) Brüschweiler, R.; Roux, B.; Blackledge, M.; Griesinger, C.; Karplus, M.; Ernst, R. R. *J. Am. Chem. Soc.* **1992**, *114*, 2289–2302.
(47) Shortle, D.; Ackerman, M. S. *Science* **2001**, *293*, 487–489.
(48) Alexandrescu, A. T.; Kammerer, R. A. *Protein Sci.* **2003**, *12*, 2132–2140.
(49) Mohana-Borges, R.; Goto, N. K.; Kroon, G. J. A.; Dyson, H. J.; Wright, P. E. *J. Mol. Biol.* **2004**, *340*, 1131–1142.
(50) Fieber, W.; Kristjansdottir, S.; Poulsen, F. M. *J. Mol. Biol.* **2004**, *339*, 1191–1199.
(51) Meier, S.; Güthe, S.; Kiefhaber, T.; Grzesiek, S. *J. Mol. Biol.* **2004**, *344*, 1051–1069.

predictions from calculated ensembles of random-coil conformers has indicated that RDCs are sensitive to amino acid-specifi backbone dihedral angle distributions. The ability to defin random-coil RDC values has led to firs the identificatio and then the quantificatio of the level of secondary structure propensity in IDPs, initially by comparison with ensemble averages reporting on different sampling regimes[35−42] and more recently by using RDCs to determine conformational sampling on an amino acid-specifi basis using ASTEROIDS.[43] In the latter case, a highly efficien local alignment window (LAW) approach to the simulation of RDCs was used to account for local-sampling and near-neighbor effects.[43,59] This demonstrated that in order to correctly defin the conformational behavior for a LAW with a length of 15 amino acids, at least 200 structures are needed to average the RDCs.[43] In addition, it was noted that in contrast to chemical shifts and scalar couplings, RDCs are also sensitive to the degree and nature of transient long-range order, and even in the absence of specifi contacts, it was found to be necessary to combine the local prediction from the LAWs with a generic baseline profil along the primary sequence that accounts for the chainlike nature of the protein.

In this study, ASTEROIDS and *flexible-meccano* were adapted to allow for transient long-range order and combined with experimental PREs to determine an ensemble description of α-synuclein, a paradigm of the IDP family, whose conformational properties in free solution have been characterized extensively using NMR spectroscopy and associated biophysical techniques.[22,23,26,61−66] We demonstrate that even in the presence of highly diffuse, ill-define target interactions, explicit modeling of spin-label mobility significantl improves the prediction of experimental data not used in the analysis. We also show that even weak intramolecular interactions can induce a severe distortion of the expected RDC values that compromises the description of local conformational sampling if not correctly taken into account. The expected modulation of the RDCs is parametrized in such a way that it can be analytically introduced into the predicted RDC profile and we demonstrate that incorporation of long-range contacts from the PRE-derived ensemble significantl improves the prediction of experimental RDCs from α-synuclein.[23] This novel approach allows for the direct and efficien introduction of long-range contacts into ensemble-averaged RDCs and provides for the simple and powerful combination of RDCs and PREs into a single ensemble description.

Theoretical Aspects

Dynamic Averaging of PREs. IDPs are highly flexibl on diverse time scales, and this flexibilit must be taken into account in the analysis of the measured PREs. The transverse relaxation rate due to the presence of the unpaired electron, Γ_2, can be expressed as follows:[67]

$$\Gamma_2 = \frac{2}{5}\left(\frac{\mu_0}{4\pi}\right)^2 \gamma_H{}^2 g_e{}^2 \mu_B{}^2 s_e(s_e + 1)[4J(0) + 3J(\omega_H)] \quad (1)$$

where g_e is the electron g-factor, γ_H is the gyromagnetic ratio of the observed nucleus (proton), s_e is the electron spin, ω_H is the proton frequency, μ_B is the Bohr magneton, and μ_0 is the permittivity of free space. It has been shown[14,46] that the spectral density function $J(\omega)$ can be described using a model-free expression of the order parameter comprising the orientational and distance-dependent components of the internal motion, both of which strongly depend on the motion of the spin label with respect to the observed nuclear spin:

$$J(\omega) = \langle r_{H-e}{}^{-6}\rangle \left[\frac{S^2_{H-e}\tau_c}{1 + \omega^2\tau_c{}^2} + \frac{(1 - S^2_{H-e})\tau_e}{1 + \omega^2\tau_e{}^2}\right] \quad (2)$$

where the order parameter S^2_{H-e} describes the motion of the dipolar interaction vector, $\tau_c = \tau_r\tau_s/(\tau_r + \tau_s)$ is define in terms of the electron spin and rotational correlation times τ_s and τ_r, respectively, τ_e is given by the expression $\tau_e = 1/(\tau_i{}^{-1} + \tau_r{}^{-1} + \tau_s{}^{-1})$, in which τ_i represents the effective correlation time of the spin label, and r_{H-e} is the instantaneous distance between the proton and electron spins. The order parameter can be usefully decomposed into radial and angular components as

$$S^2_{H-e} = S^2_{ang}S^2_{rad} \quad (3a)$$

where

$$S^2_{rad} = \langle r_{H-e}{}^{-6}\rangle^{-1}\langle r_{H-e}{}^{-3}\rangle^2 \quad (3b)$$

and

$$S^2_{ang} = \frac{4\pi}{5}\sum_{m=-2}^{2} |\langle Y_2^m(\Omega^{mol})\rangle|^2 \quad (3c)$$

in which Ω^{mol} refers to the orientation of the interaction vector in the frame of the *flexible-meccano* conformer. These expressions are used to calculate the effective transverse relaxation rate for each backbone conformation produced with the *flexible-meccano* algorithm.

The electron spin label is attached to the molecule via a thiol-reactive methanethiosulfonate (MTSL) attached to a cysteine side chain. MTSL conformations are built explicitly for each *flexible-meccano* backbone conformer by randomly sampling known rotameric descriptions.[45] Only conformations that do not result in steric overlap with the remainder of the chain are retained in the *N*-conformer ensemble that is used to represent the position of the side chain. Thus, for each backbone

(52) Ohnishi, S.; Lee, A. L.; Edgell, M. H.; Shortle, D. *Biochemistry* **2004**, *43*, 4064–4070.
(53) Sallum, C. O.; Martel, D. M.; Fournier, R. S.; Matousek, W. M.; Alexandrescu, A. T. *Biochemistry* **2005**, *44*, 6392–6403.
(54) Ding, K.; Louis, J. M.; Gronenborn, A. M. *J. Mol. Biol.* **2004**, *335*, 1299–1307.
(55) Louhivuori, M.; Pääkkönen, K.; Fredriksson, K.; Permi, P.; Lounila, J.; Annila, A. *J. Am. Chem. Soc.* **2003**, *125*, 15647–15650.
(56) Jha, A. K.; Colubri, A.; Freed, K.; Sosnick, T. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099–13105.
(57) Obolensky, O. I.; Schlepckow, K.; Schwalbe, H.; Solov'yov, A. V. *J. Biomol. NMR* **2007**, *39*, 1–16.
(58) Betancourt, M. R. *J. Phys. Chem. B* **2008**, *112*, 5058–5069.
(59) Marsh, J. A.; Baker, J. M. R.; Tollinger, M.; Forman-Kay, J. D. *J. Am. Chem. Soc.* **2008**, *130*, 7804–7805.
(60) Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 11266–11267.
(61) Eliezer, D.; Kutluay, E.; Bussell, R., Jr.; Browne, G. *J. Mol. Biol.* **2001**, *307*, 1061–1073.
(62) Fernandez, C. O.; Hoyer, W.; Zweckstetter, M.; Jares-Erijman, E. A.; Subramaniam, V.; Griesinger, C.; Jovin, T. M. *EMBO J.* **2004**, *23*, 2039–2046.
(63) Sung, Y.-h.; Eliezer, D. *J. Mol. Biol.* **2007**, *372*, 689–707.
(64) Wu, K. P.; Kim, S.; Fela, D. A.; Baum, J. *J. Mol. Biol.* **2008**, *378*, 1104–1115.
(65) Li, C.; Lutz, E. A.; Slade, K. M.; Ruf, R. A.; Wang, G. F.; Pielak, G. J. *Biochemistry* **2009**, *48*, 8578–8584.
(66) Lendel, C.; Damberg, P. *J. Biomol. NMR* **2009**, *44*, 35–42.
(67) Solomon, I. *Phys. Rev.* **1955**, *99*, 559–565.

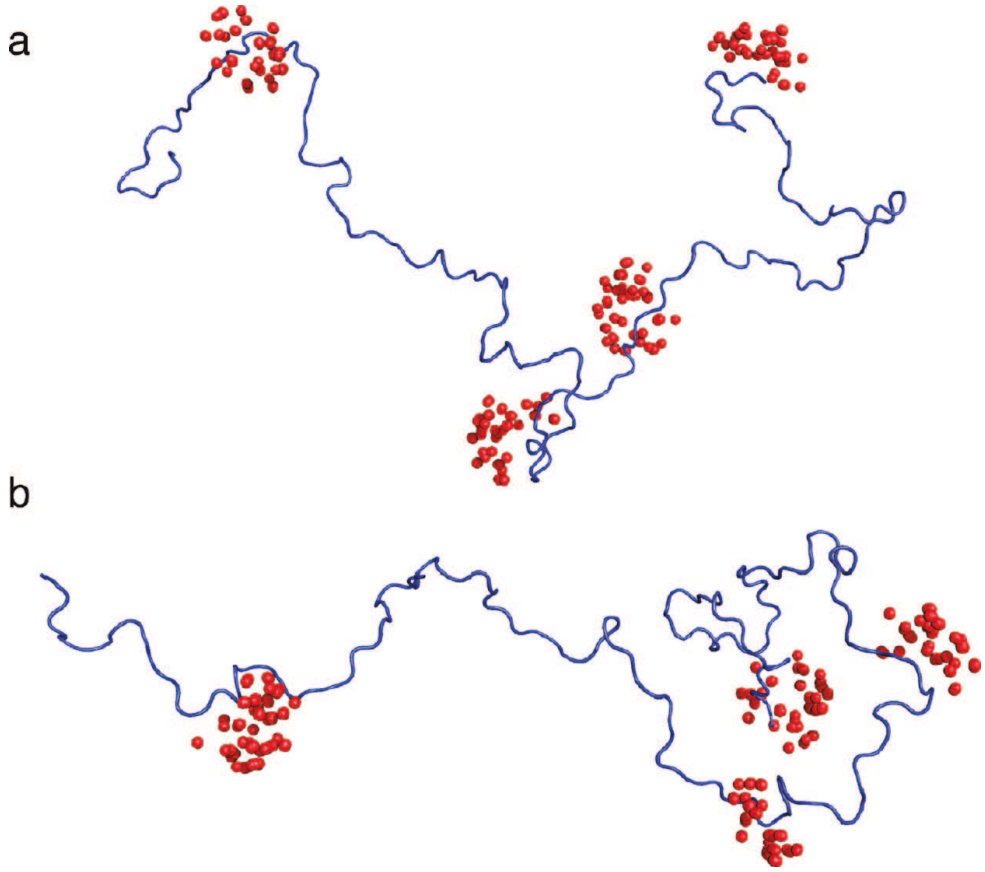


Figure 1. Representation of the possible nitroxide spin label positions relative to the backbone of individual structures calculated using the conformational sampling algorithm *flexible-meccano*. Two representative conformers are shown. The positions of the heavy atoms are represented by the blue ribbon, while allowed MTSL side-chain positions are shown in red for each of four paramagnetic probes used in the α-synuclein study (amino acids 18, 76, 90, and 140). Previously proposed MTSL rotameric libraries[45] were randomly sampled for a total of 600 conformers for each site. Each position was retained and included in the averaging procedure if no steric clashes were found with respect to the given backbone conformation.

conformation, the MTSL side chain is represented by a population-weighted sampling of the available rotameric states. The effective relaxation rate for each amide proton is taken as the average of the rates $\Gamma_{2,c}^{fm}$ for the N retained *flexible-meccano* conformers:

$$\Gamma_2^{\text{total}} = \frac{1}{N}\sum_{c=1}^{N}\Gamma_{2,c}^{fm} \qquad (4)$$

Effective intensities are then calculated as described in Methods.

The assumption made here are that the interconversion between different side-chain conformations is independent of (and faster than) the interconversion between different discrete conformers. In common with previous applications,[12,23,28] we estimated τ_c to be 5 ns, and the internal motion describing the sampling of the different side-chain conformations was assumed to have a correlation time of 500 ps. This is in broad agreement with values derived from earlier MD/ESR-based studies,[68] and we note that changing the internal correlation time by a factor of 2 in either direction had no noticeable influenc on the resulting analysis.

Figure 1 shows the possible positions of the spin label for each of four paramagnetic probes attached to cysteine mutants of the protein α-synuclein in two *flexible-meccano* conformers (amino acids 18, 76, 90, and 140, which are the positions used in the experimental study).[23] The spin label can clearly occupy a large volume space that could potentially affect the effective relaxation behavior of the observed spins.

Results and Discussion

Our aim in this study was to analyze the effects of long-range transient contacts on experimentally observable NMR parameters from unfolded proteins and to develop a formalism that allowing their use for the meaningful characterization of both local and long-range structure in these highly flexibl systems. In order to do this, we initially used molecular simulations to investigate the expected effects in systems with either one or two dominant long-range contacts. Although these simulated systems were intentionally oversimplifie for the sake of clarity, the application of the observed results to more complex networks of long-range transient interactions is expected to be straightforward.

Paramagnetic Relaxation Enhancement in Highly Disordered Systems: Simulation. We initially determined whether it is possible to detect weakly specifi long-range interactions via the combined ASTEROIDS and *flexible-meccano* analysis applied to simulated PREs. Figure 2 shows PREs calculated for a simulated model protein of 100 amino acids with paramagnetic spin labels attached at positions 20, 40, 60, and 80 (red bars). In Figure 2a, each conformer contains a contact between 41−50 and 81−90. The definitio of a contact is given in Methods. The solid line shows the expected broadening in the absence of *specific* contacts (the reference ensemble where all conformers are allowed). We note that the effective broaden-

(68) Sezer, D.; Freed, J. H.; Roux, B. *J. Chem. Phys.* **2008**, *128*, 165106.

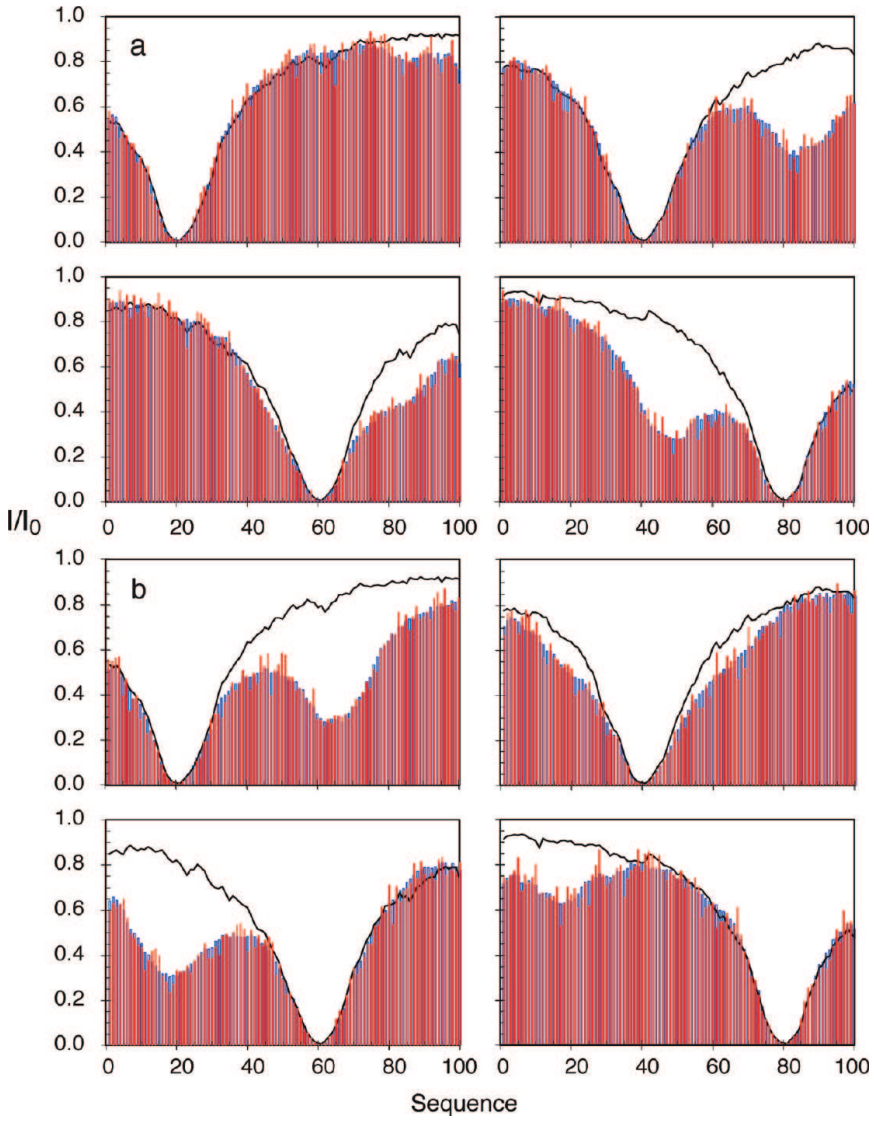


Figure 2. Reproduction of simulated sample PRE data for ensembles containing specifi contacts using the ASTEROIDS ensemble selection algorithm.[43] (a) Blue: data averaged over the target ensemble in which each conformer has a contact between 41−50 and 81−90. Red: data averaged over an ensemble of 80 structures selected using ASTEROIDS. The four boxes show the PRE data for simulated spin labels at residues 20 (top left), 40 (top right), 60 (bottom left), and 80 (bottom right). Lines show the PREs calculated from a control ensemble with no specifi contacts. (b) Blue: as in (a) for a target ensemble in which each conformer has a contact between 11−20 and 61−70. Red: data averaged over an ensemble of 80 structures selected using ASTEROIDS.

ing, even in the absence of specifi contacts, is quite significan as a result of the large volume space sampled by the spin label. Figure 2b shows a similar representation of an ensemble with contacts between positions 11−20 and 61−70. The ASTEROIDS algorithm targeting these simulated PREs was then used to select 80-member conformational ensembles from a pool of 10 000 structures without specifi contacts calculated using the *flexible-meccano* Monte Carlo sampling approach (see Methods). The resulting ensembles reproduced the simulated PREs well, as shown by the blue bars in Figure 2. It should be noted that these simulations used examples that were quite demanding, with 20% of the chain involved in weakly specifi contacts. It is nevertheless a reasonable reproduction of the situation that one may encounter when studying intrinsically disordered or partially folded proteins, with long-range interactions occurring between strands carrying complementary electrostatic charge or containing hydrophobic side-chains. It was therefore of interest to determine whether the broad averaging effects predicted from such a simulation would allow the extraction of meaningful information concerning the long-range contacts.

ASTEROIDS Reproduces the Overall Biophysical Features of the Target Ensemble. Figure 3 shows the effective contacts present in the ASTEROIDS ensembles that matched the simulated data. This representation compares interatomic (C^{α}) distances present in the reference ensemble with those in the selected ensemble (see Methods). The contacts that were used to simulate the data are well identifie in both cases. The exact values of the distances were not reproduced (the distances were underestimated), but this is not considered a serious drawback in view of the ill-define nature of the contact. We also compared the overall distributions of the selected ensembles relative to the reference ensemble. Figure 4 shows that the ASTEROIDS ensemble of structures selected using the simulated PREs from the reference ensemble containing contacts between regions 11−20 and 61−70 (Figure 2a) reproduced the distribution of the radiii of gyration (R_g) for members of the reference ensemble quite closely. The average R_g of the ASTEROIDS ensembles increased slightly with increasing number of structures, from 21.3 Å for the 80-member ensemble to 21.7 Å for the 160-member ensemble, compared with 22.6

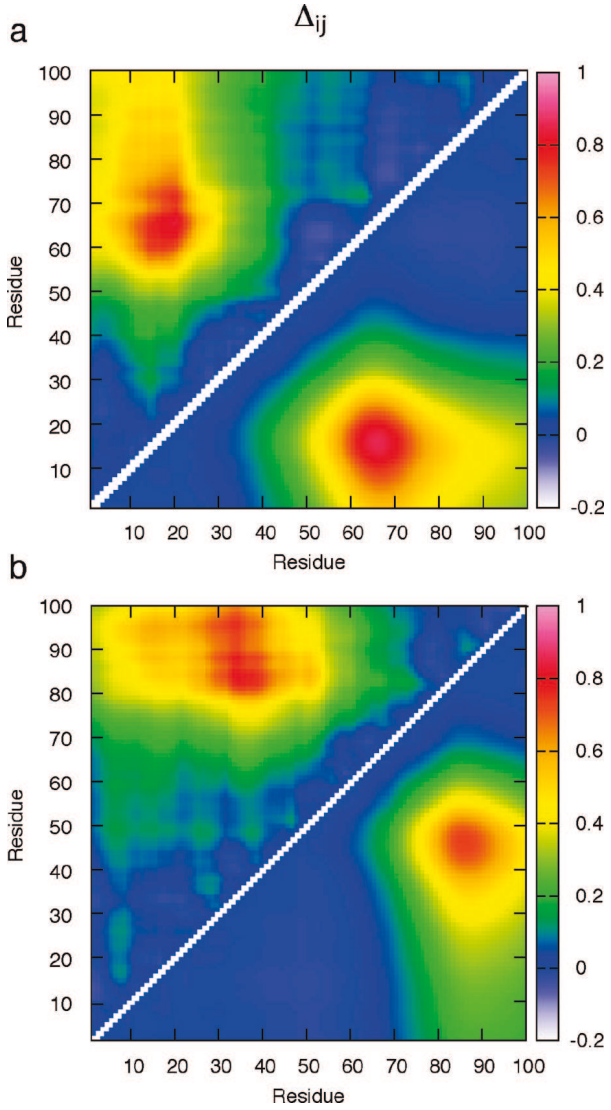


Figure 3. Contact maps showing chain proximity in the ensembles selected using ASTEROIDS on the basis of the data shown in Figure 2 (above the diagonal) in comparison with target ensembles (below the diagonal). In (a), the contact was between 11−20 and 61−70, while for (b), the contact was between 41−50 and 81−90. The scale for the data above the diagonal in each panel has been multiplied by a factor of 0.50 for ease of identificatio of the contact.

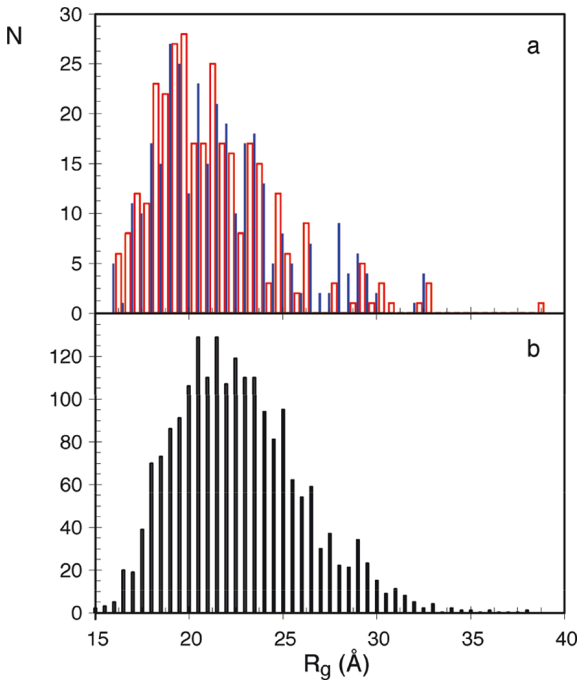


Figure 4. Ability of the ASTEROIDS approach to accurately reproduce the distribution of radii of gyration (R_g) in the selected ensembles. (a) Histogram showing the overall dimensions of the structures in ASTEROIDS ensembles selected on the basis of PREs shown in Figure 2b (contacts between 11−20 and 61−70). Blue: distribution of R_g in ensembles of size 80 (average R_g = 21.3 Å). Red: distribution of radii of gyration in ensembles of size 160 (average R_g = 21.7 Å). (b) Distribution of R_g for a set of 2000 structures from the target ensembles in which all of the structures contain a contact between 11−20 and 61−70 (average R_g = 22.6 Å).

Å for the reference ensemble. The previously noted tendency of PRE-based analysis to produce unrealistically compact ensembles of unfolded states, although present, was apparently less pronounced using the combined ASTEROIDS and *flexible-meccano* approach than in the case of restrained MD-based studies.[26−28] The exact origin of this observation is not clear and will require further comparative studies, but the improvement may be related to the explicit modeling of side-chain flexibilit or to the fact that this approach uses the data to select representative ensembles rather than fittin the conformational sampling directly to the data.

We tested the ability of the combined ASTEROIDS and *flexible-meccano* approach to reproduce more than one contact. Clearly, the accuracy of this reproduction depends strongly on the number of paramagnetic probes and their specifi distribution in the protein as well as the nature of the contacts (diffuse or well-defined) We performed an additional simulation, in this case for a protein containing 200 amino acids, where the target ensemble consisted of conformers with a contact between 11−20 and 61−70 or between 141−150 and 181−190. Simulated data from eight paramagnetic probes allowed ASTEROIDS to accurately and unambiguously fin both contacts (Figure 5). The simulated target and fitte data from the eight sites are shown in Figure S1 in the Supporting Information.

Paramagnetic Relaxation Enhancement in Highly Disordered Systems: Experimental Data. In order to test the ensemble selection procedure further, we applied this approach to an experimental data set measured by Bertoncini et al.[23] for the intrinsically disordered protein α-synuclein. We employed these experimental data to determine how the use of an explicit flexibl side-chain description of the spin label compares to using a fixe single position for each *flexible-meccano* conformer. In order to do this, we used ASTEROIDS to select ensembles based on the PRE data from cysteine mutants 18, 90, and 140 and then used these ensembles to predict the PREs measured for the spin label at position 76. It should be noted that this involved removing 25% of the available experimental data. The ensembles determined using a flexibl side-chain description and a static side-chain description both fi the experimental data from the three "active" labels to within the experimental uncertainty, with the flexibl side-chain model affording a slightly better fi (data not shown). More importantly, the reproduction of the "passive" data (i.e., the data not used in the ensemble selection) was systematically and significantl better when the flexibl side-chain model was employed: the root-mean-square deviation (rmsd) for the flexibl side-chain model was 0.17 ± 0.01, compared with an rmsd of 0.24 ± 0.02 for the static description. An example is shown in Figure 6, where the data reproductions of the PREs induced by the spin label at position 76 are compared for the two descriptions. This

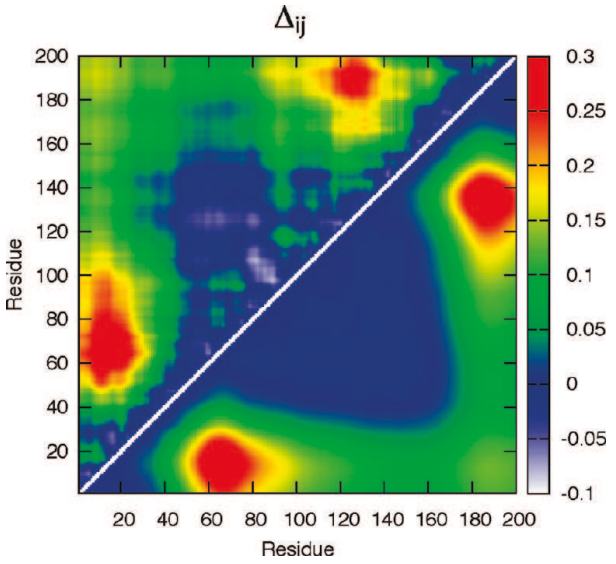


Figure 5. Contact maps showing chain proximity in the presence of two contacts. Above the diagonal: contact map for an ASTEROIDS ensemble selected to reproduce simulated PRE data averaged over an ensemble in which each 200 amino acid conformer has a contact between 11−20 and 61−70 or between 141−150 and 181−190. In this case, eight PRE sites were simulated (sequence numbers 22, 44, 66, 88, 110, 132, 154, and 176). Below the diagonal: contact map for the target ensemble used to simulate the PRE data. The scale for the data above the diagonal has been multiplied by a factor of 0.66 for ease of identificatio of the contact.

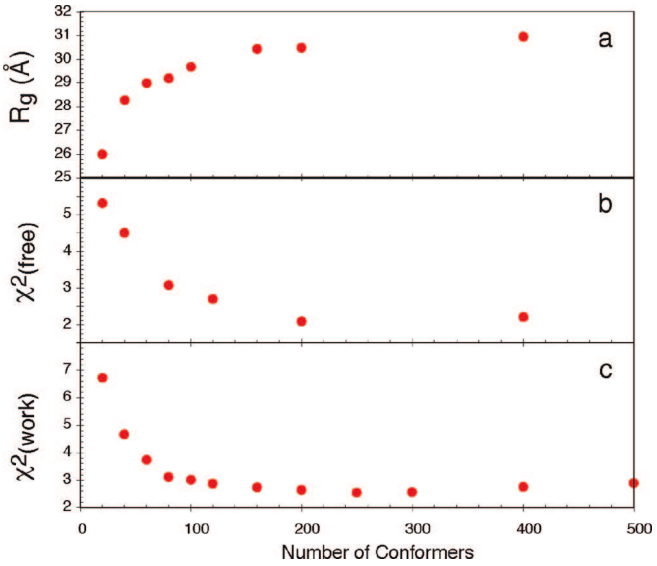


Figure 7. Ensemble characteristics as a function of selected ensemble size, targeting experimental PRE data measured in α-synuclein. (a) Average radius of gyration as a function of the number of structures in the selected ensemble. (b) χ^2 for the free data as a function of the number of structures in the selected ensemble. The free data in this case consists of the entire A76C data set. Only data from A18C, A90C, and A140C were used in the ensemble selection for the cross-validated reproduction of the "free" data set. (c) χ^2 for the active data as a function of the number of structures in the selected ensemble.

example was chosen at random and is representative of the observed improvement. This result demonstrates the importance of incorporating local MTSL side-chain dynamics into the ensemble interpretation of the PREs, even for highly dynamic systems. These motions are predicted to occur on a relaxation-active time scale[45] and therefore require the use of the model-free or equivalent description that can explicitly account for the effect of local motions on the spectral density function. If fast motions of the spin label relative to the backbone are not included in the analysis, time-scale-dependent modulation of the observed relaxation interaction may be aliased into the effective intramolecular distance distribution.

The quality of the cross-validated data reproduction using the dynamic description allowed us to use this approach to probe the optimal number of structures required to describe the ensemble. This number depends on the complexity of the system (including the number of long-range contacts) as well as the number and position of the spin labels, but in this case, both the active and passive χ^2 values indicated that ensembles of ~200 structures were appropriate (Figure 7). This was supported by analysis of the effective radius of gyration, which rises until it reaches a plateau at approximately the same number of structures. Figure 8 shows the data reproduction when data from all four sites were included in the analysis; also shown is the resulting contact map comparing average interatomic distances in the ensemble with those from a control ensemble in which no selection on the basis of experimental data was made. In line with previous studies, a long-range contact between the C- and N-terminal domains was observed [as well as a weaker contact between the so-called NAC region (residues 65−95) and the C-terminal domain].[22,23,35]

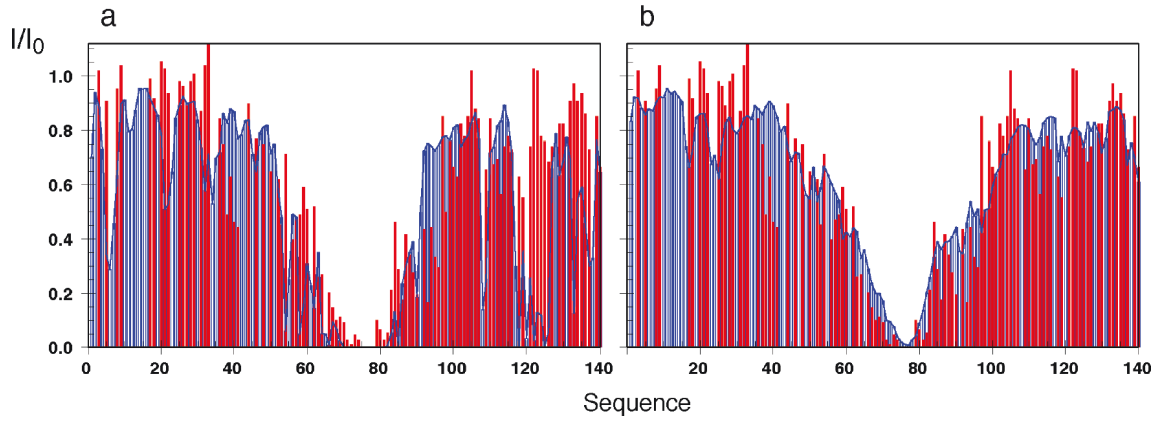


Figure 6. Cross-validation of "free" α-synuclein PRE data. Only data from A18C, A90C, and A140C were used in the ensemble selection. (a) Example of the reproduction of the PRE data from the A76C site using the static position of the C^{β} atom as a representation of the average position of the spin label. (b) Example of the reproduction of the PRE data from the A76C site using the explicit MTSL side-chain dynamic averaging model described in the text. In both cases, the experimental PREs are shown in red and the calculated ratios in blue.

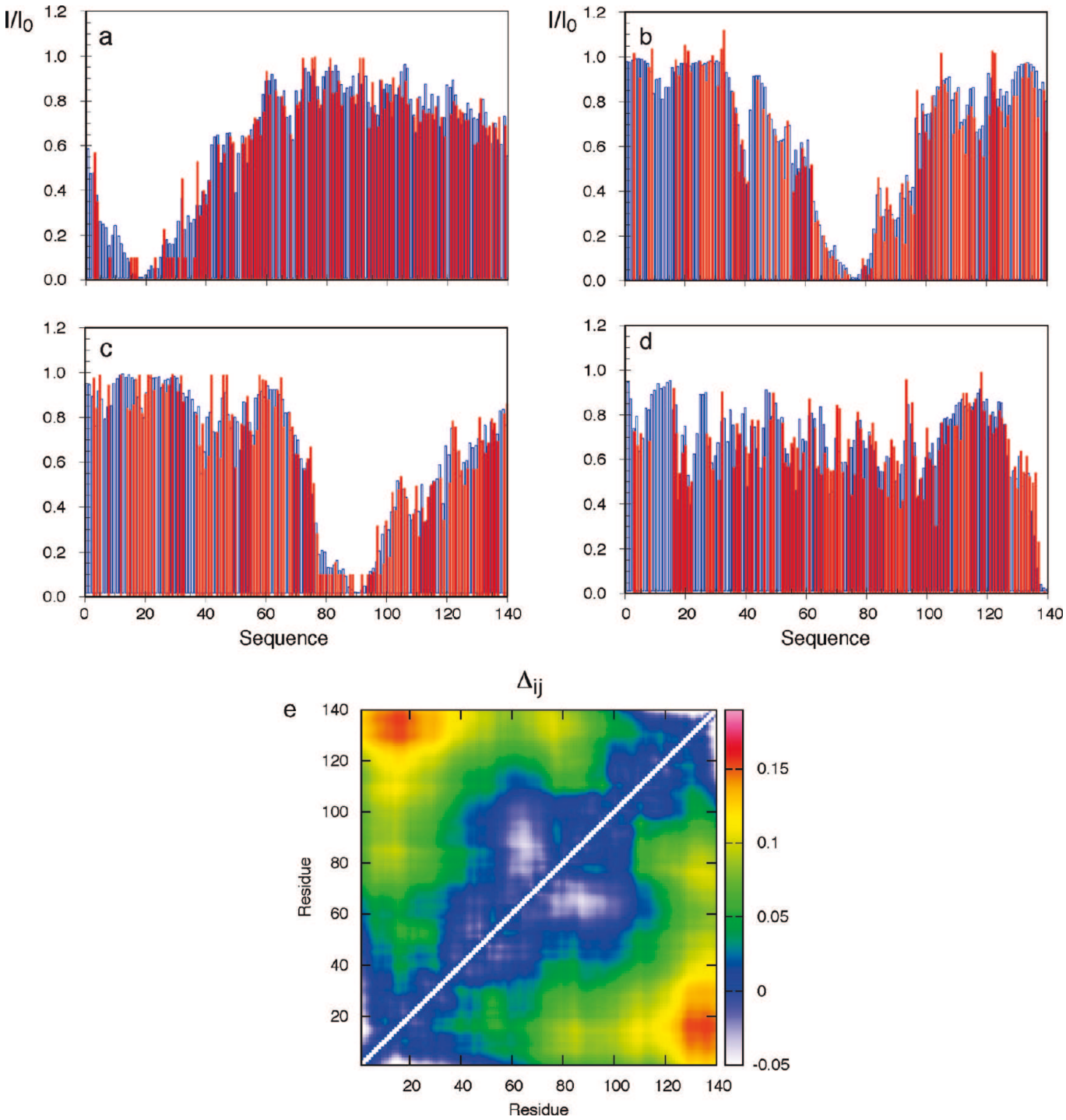


Figure 8. Reproduction of PRE data measured for α-synuclein. (a−d) Comparison of experimental and ensemble-averaged data for an example calculation. (e) Resulting contact map showing the relative proximity of different parts of the chain.

Effects of Weak Long-Range Contacts on RDCs Measured in Highly Disordered Systems. In order to obtain a unifie representation of the behavior of disordered proteins in solution, it is necessary to incorporate data from different sources that exhibit different structural and dynamic dependences. Here we investigate the effects of weak long-range contacts on the expected values of RDCs that are generally assumed to report mainly on local conformational propensities in disordered chains, and we propose appropriate guidelines for combining PREs and RDCs when using ensemble descriptions of flexibl proteins.

The *flexible-meccano* approach was used to predict RDCs from 100 000-member ensembles of the 100 amino acid model sequence in the presence of weakly define long-range contacts (Figure 9). The expected profile when no specifi contacts were present are also shown (Figure 9a). Figure 9b−g shows profile of the expected $^{15}N-^{1}H^{N}$ ($^{1}D_{NH}$) and $^{13}C^{\alpha}-^{1}H^{\alpha}$ ($^{1}D_{C\alpha H\alpha}$) RDCs when a contact between two 20 amino acid strands (e.g., regions 1−20 and 81−100) was present. The effect of even such diffuse long-range contacts is surprisingly strong, resulting in significan quenching of the RDC values in regions between the two contact regions and some reinforcement of RDCs in the region of the contacting parts of the chain. Amino acids in all regions had essentially identical conformational sampling in all cases, but the RDCs were very different, indicating very clearly that

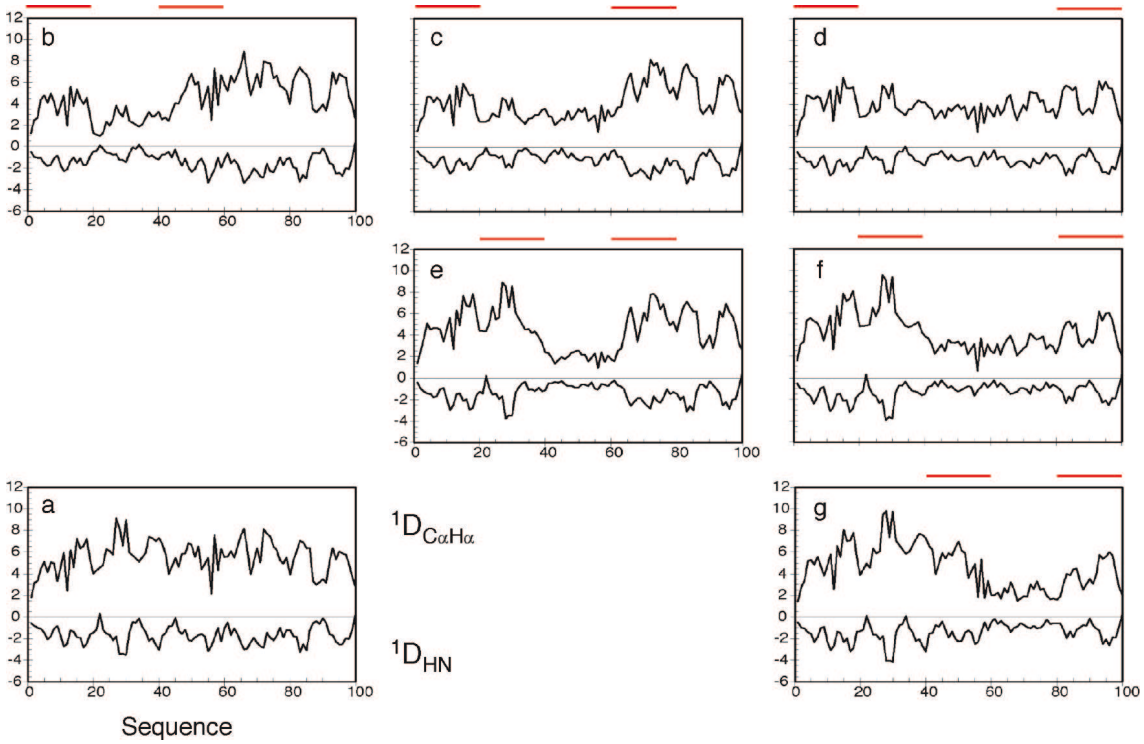


Figure 9. Simulation of $^1D_{NH}$ and $^1D_{C\alpha H\alpha}$ RDC profile for a disordered protein with an arbitrary sequence in the presence of contacts between different sections of the chain. (a) Profil of couplings in the absence of specifi contacts. The program PALES was used to calculate RDCs for each conformer; 100 000 conformers were used in this average and the ones shown in panels (b−g). (b−g) Profile of couplings in the presence of contacts between regions i and j: (b) $i = 1-20$, $j = 41-60$; (c) $i = 1-20$, $j = 61-80$; (d) $i = 1-20$, $j = 81-100$; (e) $i = 21-40$, $j = 61-80$; (f) $i = 21-40$, $j = 81-100$; (g) $i = 41-60$, $j = 81-100$. The two continuous red bars above each plot indicate the positions of the contacting regions.

caution needs to be exercised when interpreting RDCs uniquely in terms of local structure if long-range contacts are also present. This would potentially lead to significan error in the cases shown in Figure 9.

In order to further clarify the origin of these effects, the same analysis was carried out for a homopolymer (polyvaline), resulting in the expected bell-shaped curve for the ensembles without contact-specifi selection (Figure 10a) and clear modification occurring for the ensembles with specifi contacts (Figure 10b−g). The effect of diffuse long-range contacts is apparently to superpose a more complex baseline upon the local structure of the expected RDCs. This baseline has peaks in the interacting regions and a trough in the intervening region. We believe that the effect has a similar origin as that found in the presence of helical elements in disordered chains, where $^1D_{NH}$ values become positive as a result of the effective average alignment of the $^{15}N-^1H^N$ bond vectors with the average chain direction and thereby the magnetic field [60] The same effect may occur here, although in this case, the helix has a very long period in terms of amino acids and therefore would create a very broad inverted curve relative to the bell-shaped curve, whose shallowness depends on the distance between the interacting segments, as observed from the numerical simulation.

Parametrization of the Effect of Long-Range Contacts on RDCs in Disordered Systems. It has previously been shown that RDCs from unfolded chains with no specifi interacting regions can be expressed in terms of the product of a generic baseline and RDCs derived from sampling of conformational space that can be define using short local alignment windows (LAWs):[59,43]

$$D_{ml} = |b_{ml}| D_{ml}^{\mathrm{LAW}} \qquad (5)$$

where m and l represent the pair of nuclei (e.g., N and H^N). In Figure 10, the red curves were obtained using the parametrization of a generic baseline expression that reproduces the numerically predicted baselines shown for the polyvaline chain (see Methods for the full expression). This can be described as a combination of the baseline expression for no specifi contacts (a hyperbolic cosine function introduced previously[43]) with Gaussian curves between the contact points. Importantly, the curves depend only on the position of the contacts and the length of the chain.

This expression can then be combined with RDCs predicted using LAWs accounting for short-range conformational behavior. This is illustrated in Figure 11, where the LAW-derived profil (Figure 11b), which was calculated using 200 structures, is combined with the baseline predicted for long-range contacts between segments 41−60 and 81−100 (Figure 11a). The prediction agrees essentially identically with the explicit simulations calculated using 100 000 conformers containing the required contact (Figure 11c). In the case of more than one contact (as shown in Figure 5, for example), the baseline effects are combined as shown in eq 11 and can again be shown to accurately reproduce the effects simulated from explicit averages over 100 000 conformers containing these contacts (see Figure S2 in the Supporting Information).

Simultaneous Analysis of PRE-Derived Long-Range Contacts and RDC-Derived Local Information. The above results show that it is possible in principle to combine PRE-derived long-range information with RDC-derived local information while accounting for possibly significan long-range effects on RDCs and preserving a relatively small number of structures. This latter point is of particular importance when using ensemble selection approaches. In order to test this possibility further, we analyzed the ensembles presented in Figure 3, where the target contacts were between positions 11−20 and 61−70 and between positions 41−50 and 81−90. The contact matrices were analyzed

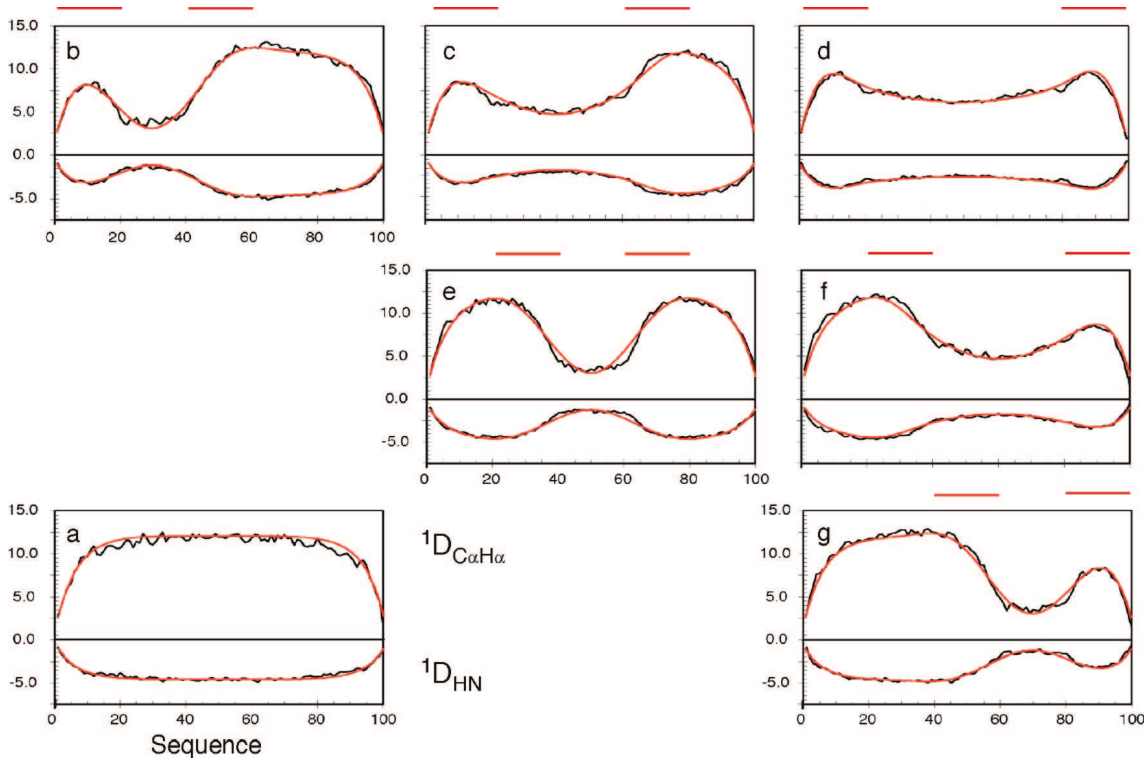


Figure 10. Simulation of RDC profile for a homopolymer (polyvaline) in the presence of contacts between different sections of the chain. (a) Profil of calculated couplings in the absence of specifi contacts. The program PALES was used to calculate RDCs from each conformer; 100 000 conformers were used in this average and the ones shown in panels (b−g). (b−g) Profile of couplings in the presence of contacts between regions i and j: (b) $i = 1-20$, $j = 41-60$; (c) $i = 1-20$, $j = 61-80$; (d) $i = 1-20$, $j = 81-100$; (e) $i = 21-40$, $j = 61-80$; (f) $i = 21-40$, $j = 81-100$; (g) $i = 41-60$, $j = 81-100$. The two continuous bars above each plot indicate the positions of the contacting regions. The red curves were computed using eq 11 with the contact positioned in the center of each region.

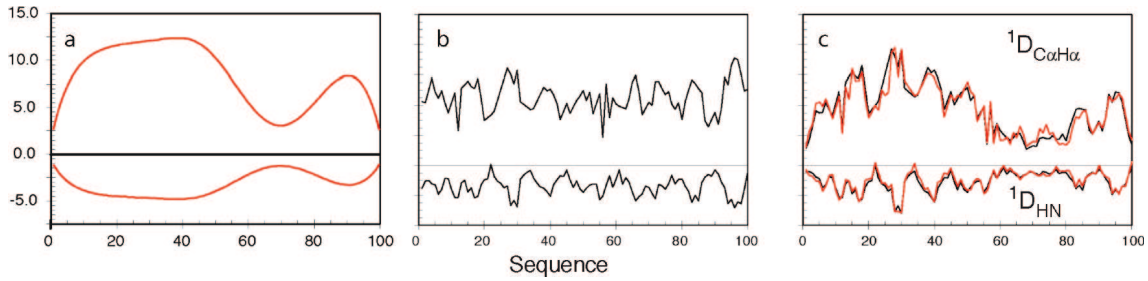


Figure 11. Example of the combination of analytically calculated baselines and RDCs averaged using the local alignment window (LAW) approach. (a) Baseline contribution calculated analytically using eq 11 for contacts between the regions centered on residues 50 and 90. (b) RDCs calculated using the previously proposed LAW approach with windows 15 amino acids in length; each RDC was averaged over 200 structures. (c) Combination of the baseline from (a) and the local RDCs from (b) (red curves) compared to the RDCs averaged over 100 000 full-length conformers in which each structure has a contact between 41−60 and 81−100 (black curves).

to fin the maximum of the difference between the PRE-derived ensemble and the reference ensemble containing no specifi contacts (see Methods). The results are shown in Figure 12. In Figure 12a, the red and blue curves indicate the RDC baselines derived using this approach (calculated using eq 11), and the black curve shows the $^1D_{NH}$ RDCs calculated using the LAW approach. In Figure 12b,c, the combination of the baseline and the locally calculated RDCs is compared to RDCs calculated explicitly from 100 000 conformers, all of which fulfil the contact criterion. The good agreement demonstrates that one can combine PREs and RDCs in a meaningful way for the ensemble description of disordered proteins using experimental data.

Combining Experimental PREs and RDCs in α-Synuclein Validates RDC Baseline Analysis. Finally, we applied this analysis to the contact matrix determined on the basis of experimental PRE data from α-synuclein (shown in Figure 8e). Experimentally measured RDCs are shown in Figure 13a and compared to RDCs calculated from an explicit representation of full-length α-synuclein. The RDC baseline derived from analysis of the contact matrix is shown in Figure 13b, superimposed on the RDCs calculated using the LAW approach. The two curves were combined using eq 5, and the result is compared to the experimental data (after appropriate scaling) in Figure 13c. The RDC profil reproduces the experimental data significantl better than the ensemble derived in the absence of specifi contacts (rmsd of 0.51 Hz compared with 0.75 Hz). This study therefore not only validates the predicted effects on RDC profile due to long-range transient contacts in unfolded systems but also demonstrates that PREs and RDCs can be usefully combined in an experimental context. This provides further support for previously published observations that RDCs

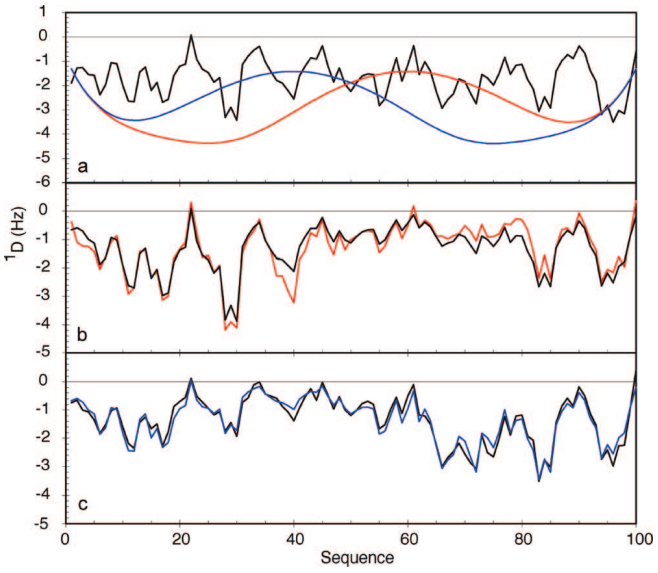


Figure 12. Example of a combined analysis of PREs and RDCs in the context of simulated data. PREs were used to determine long-range contacts. RDC profile were calculated using baselines determined on the basis of PRE analysis and LAWs. Contacts were identifie from distance matrices as described in the text. The reproduction of the PREs and the resulting distance matrix from this simulation are shown in Figures 2 and 3. (a) Black curve: LAW-averaged RDCs. Blue curve: RDC baseline extracted from the contact matrix shown in Figure 3a (contact between 11−20 and 61−70). Red curve: RDC baseline extracted from the contact matrix shown in Figure 3b (contact between 41−50 and 81−90). (b) Black curve: RDCs calculated from an explicit ensemble calculation using 100 000 conformers. Red curve: the combination of the LAW curve and red baseline curve shown in (a) (contact between regions 41−50 and 81−90). (c) Black curve: RDCs calculated from an explicit ensemble calculation using 100 000 conformers. Red curve: combination of the LAW curve and blue baseline curve shown in (a) (contact between regions 11−20 and 61−70).

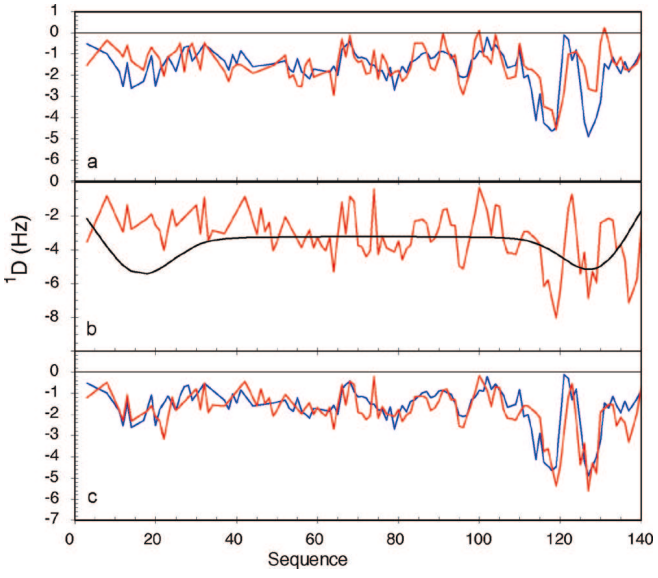


Figure 13. Example of a combined analysis of PREs and RDCs in the context of experimental data: comparison of experimental $^1D_{NH}$ RDCs measured from α-synuclein aligned in PEG-hexanol with values obtained using the combination of LAW and baseline prediction from PRE analysis. (a) Comparison of experimental $^1D_{NH}$ RDCs (blue) with couplings calculated using a standard *flexible-meccano* prediction (red). The rmsd between the two distributions was 0.78 Hz. (b) LAW-predicted RDCs (red) and effective baseline derived from the contacts shown in Figure 8e using eq 11 (black). (c) Combination of the curves shown in (b) (red) compared to the experimental $^1D_{NH}$ RDCs (blue). The rmsd in this case was 0.52 Hz.

have been correctly reproduced only in the presence of long-range contacts.[35]

Conclusions

In order to understand the conformational behavior of IDPs, a molecular representation of the partially folded state is required. Because of the very large number of degrees of conformational freedom available to such a disordered system, this representation should be based on extensive sets of experimental data using novel analytical tools designed to exploit the specifi conformational sensitivity of the different experimental parameters. Each experimental parameter is sensitive to different aspects of the structural and dynamic behavior of the disordered state and requires specifi consideration of the relevant averaging properties of the physical interaction. In this study, we have taken another step toward the development of a unifie molecular representation of the disordered state by combining complementary data sets with novel analytical tools designed to exploit the specifi conformational sensitivity of the different experimental parameters.

Having recently demonstrated that multiple RDCs can be combined with an efficien ensemble selection algorithm (ASTEROIDS) to defin local conformational sampling directly from the experimental data, we have extended the approach to incorporate the possible presence of long-range contacts. We have demonstrated the use of ASTEROIDS to analyze PREs and faithfully reproduce intramolecular proximity even in the presence of highly diffuse, ill-define contacts that give rise to broad PRE profiles We have also demonstrated that the combination of numerical and analytical modeling of spin-label mobility significantl improves the reproduction of the experimental data. The effects of long-range contacts on RDCs have been shown to produce severe distortion of RDC profile predicted on the basis of local sampling alone. We have demonstrated that this distortion can be generally parametrized and combined with RDC prediction based on local sampling alone to provide an efficien and reliable tool for interpreting RDCs in flexibl chains containing preferred long-range contacts.

We thus have shown that it is possible to combine NMR data that exhibit very different averaging properties and structural dependences in a meaningful way, providing the perspective of characterizing the essential local and long-range conformational characteristics of unfolded proteins using PREs and RDCs. In the example we provided, the reproduction of experimental RDCs from the protein α-synuclein was significantl improved when baseline effects derived from the PRE analysis were introduced into the analysis, demonstrating the feasibility of combining these experimental parameters into an informative ensemble description.

Methods

Experimental Data. Details of experimental measurements of RDCs and PREs have been published elsewhere.[23,33]

PRE Calculations with *Flexible-Meccano*. Sterically allowed MTSL side-chain conformations were sampled using previously published rotameric distributions[68] and built explicitly for each spin-label site of each *flexible-meccano* backbone; 600 side-chain conformers were calculated, and the sterically allowed conformers were retained. Relaxation effects were averaged over these conformers as described in Theoretical Aspects.

Definition of Contacts. We considered a contact to be present between two different parts of the polypeptide chain if the C^β of an amino acid in one contiguous strand (e.g., residues 11−20) was

located less than 15 Å from any C^{β} in another contiguous strand (e.g., residues 51−60).

Contact Matrices. Contact matrices were analyzed to determine $l_{ij}^{\max}$, the maximum of the difference between the PRE-derived ensemble and the reference ensemble containing no specifi contacts:

$$l_{ij}^{\max} = \max_{i,j\in[1,n]} \{\log(d_{ij}^{\mathrm{PRE}}/d_{ij}^{\mathrm{ref}})\} \tag{6}$$

The matrix was then divided into segments of 5 × 5 amino acids and searched for the highest-populated segment fulfillin the following criterion:

$$0.9l_{ij}^{\max} \le \{\log(d_{ij}^{\mathrm{PRE}}/d_{ij}^{\mathrm{ref}})\} \le l_{ij}^{\max} \tag{7}$$

This approach identifie the highest-populated contacting region. The center of this region was then used to calculate the baseline effects on the RDC profil using eq 11.

Average distances between sites were represented in terms of the metric Δ_{ij}, define as

$$\Delta_{ij} = \log(\langle d_{ij}\rangle/\langle d_{ij}^{0}\rangle) \tag{8}$$

where d_{ij} is the distance between sites i and j in any given structure of the ASTEROIDS ensemble and d_{ij}^{0} is the distance between sites i and j in any given structure of the reference ensemble (with no specifi selection). This metric was used to highlight a higher propensity to form contacts than in a molecule that has no specifi contacts. It should be noted that this representation of average interatomic distances naturally (and artificially enhances contacts that are further apart in the chain, so the observed contacts are "smeared" away from the diagonal.

RDC Calculations with *Flexible-Meccano* Using a Global Alignment Tensor. Simulated RDCs were calculated using the program *flexible-meccano* interfaced to PALES.[69] Profile of RDCs in the presence of long-range order were simulated by retaining only conformers for which the desired contact was present.

RDC Calculations with *Flexible-Meccano* Using a Local Alignment Window. For calculations using a LAW, the RDC for the central amino acid of the local 15 amino acid segment was calculated for each individual structure.[43] For the terminal amino acids, seven alanines were added to the N- or C-terminus during the building of the protein to ensure that a 15 amino acid segment was always present. The resulting RDC profil along the primary sequence was calculated by averaging each value over the whole ensemble and multiplying by the corresponding scaled absolute value of the effective baseline given in eq 11. RDCs calculated using full-length descriptions of the protein were averaged over all conformers as previously described.[11]

ASTEROIDS Ensemble Selection. ASTEROIDS uses a previously described genetic algorithm to build a representative ensemble of structures of fixe size N from a large database. The algorithm selects an ensemble of N structures by compared with experimental data using the following fitnes function:

$$\chi^2_{\mathrm{ASTEROIDS}} = \sum_k (\Delta I^k_{\mathrm{calcd}} - \Delta I^k_{\mathrm{exptl}})^2 \tag{9}$$

where

$$\Delta I^k_{\mathrm{calcd}} = \frac{I^k_{\mathrm{ox}}}{I^k_{\mathrm{red}}} = \frac{\Gamma^k_{2,\mathrm{red}} \exp(-\Gamma^k_{2,\mathrm{para}} t_{\mathrm{m}})}{\Gamma^k_{2,\mathrm{red}} + \Gamma^k_{2,\mathrm{para}}} \tag{10}$$

in which $\Gamma_{2,\mathrm{red}}$ is the intrinsic transverse relaxation rate of the observed proton spin and t_{m} is the mixing time, for which a value of 10 ms was used. The fina ensemble is obtained from generations of ensembles that undergo evolution and selection using this fitnes function. Each generation comprises 100 different ensembles of size N. Remaining parameters are treated as previously described.

Parametrization of a Generic RDC Baseline Expression for Transiently Contacting Chains. A generic RDC baseline expression for transiently contacting chains can be obtained by convoluting the baseline expression for no specifi contacts (a hyperbolic cosine function introduced previously[43]) with a Gaussian curve between the contact points and then correcting this with Gaussian curves in the vicinity of the contacting points. Importantly, the Gaussian curves depend only on the position of the contacts and the length of the chain. This results in the following analytical expression for the baseline RDC, D_{ij}^{BL}:

$$\begin{aligned} D_{ij}^{\mathrm{BL}}\{\alpha,\beta\} = \{&2b(L)\cosh[-a(L)(m - m_0)] - \\ &c(L)\}\Big(1 - \sum_i \{G_i e^{-(m-n_{0,i})^2/2\sigma_i^2} + \\ &H_i\big[(D_i + S_i)e^{-(m-n_{1,i}+S_i/2)^2/2\delta^2} + \\ &(D_i - S_i)e^{-(m-n_{2,i}+S_i/2)^2/2\delta^2}\big]\}\Big) \end{aligned} \tag{11}$$

where L is the length of the chain, the contact occurs between positions n_1 and n_2, and the sum includes all of the independent contacts i. Other parameters are define as follows: $m_0 = (L + 1)/2$, $n_0 = (n_1 + n_2)/2$, $D = |n_1 - n_2|$, and $S = n_0 - m_0$. The parametrizations of a, b, c, G, H, σ, and δ are given in the Supporting Information.

Acknowledgment. L.S. received a grant from the French Ministry of Education. This work was supported by the French Research Ministry through ANR Protein Motion PCV07_194985 PCVI 0013 and the Deutsche Forschungsgemeinschaft Heisenberg Scholarship Z.W. 71/2-1 and 3-1 (to M.Z). M.R.J. benefite from a long-term EMBO fellowship and Lundbeckfonden support.

Supporting Information Available: Figure S1 showing a reproduction of simulated sample PRE data for ensembles containing two specifi contacts (produced using the ensemble selection algorithm ASTEROIDS) and the associated baseline effects; Figure S2 showing a comparison between RDCs calculated by ensemble averaging and the baseline contribution calculated using eq 11; Figure S3 showing RDCs measured in A76C cysteine mutant and wild-type α-synuclein; Figure S4 showing calculated and experimental ^3J scalar couplings from α-synuclein; and parametrization of a generic RDC baseline expression. This material is available free of charge via the Internet at http://pubs.acs.org.

JA101645G

(69) Zweckstetter, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 3791.

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

COMMUNICATION

pubs.acs.org/JACS

Structural Impact of Proline-Directed Pseudophosphorylation at AT8, AT100, and PHF1 Epitopes on 441-Residue Tau

Stefan Bibow,[†] Valéry Ozenne,[⊥] Jacek Biernat,[§,∥] Martin Blackledge,[⊥] Eckhard Mandelkow,[§,∥] and Markus Zweckstetter*[,†,‡]

[†]Department of NMR-based Structural Biology, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

[‡]DFG Center for the Molecular Physiology of the Brain, 37073 Göttingen, Germany

[§]Max Planck Unit for Structural Molecular Biology, c/o DESY, Notkestrasse 85, 22607 Hamburg, Germany

[∥]DZNE, German Center for Neurodegenerative Diseases, c/o CAESAR, Ludwig-Erhard-Allee 2, 53175 Bonn, Germany

[⊥]Institut de Biologie Structurale Jean-Pierre Ebel, CEA-CNRS-UJF UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, France

Supporting Information

ABSTRACT: The intrinsically disordered protein tau becomes excessively phosphorylated and aggregates into neurofibrillary tangles in Alzheimer's disease. To obtain insight into the structural consequences of phosphorylation, we characterized a mutant protein of tau in which epitopes recognized by Alzheimer diagnostic antibodies were mimicked by mutation to glutamic acid [AT8 (S199E, S202E, T205E), AT100 (T212E and S214E), and PHF1 (S396E and S404E)]. A large number of distance restraints obtained from NMR paramagnetic relaxation enhancement in combination with ensemble conformer calculations demonstrate that pseudophosphorylation causes an opening of the transient folding of tau. Together with previous studies on the Parkinson-related protein α-synuclein, our data indicate that networks of transient long-range interactions are common properties of intrinsically disordered proteins and that their modulation is important for aggregation.

Aggregation of the microtubule-associated protein tau into neurofibrillary tangles is the pathological hallmark of a variety of dementias.[1,2] For reasons not yet known, tau becomes excessively phosphorylated in Alzheimer's brains and as a result no longer binds properly to microtubules. The unbound tau is free to undergo abnormal aggregation. In vivo, hyperphosphorylation of tau precedes tangle formation.[3] At least 30 phosphorylation sites in tau filaments have been identified.[4] Phosphorylation at serine/proline and threonine/proline motifs in the flanking regions of the repeat domain of tau has only a moderate influence on tau−microtubule interactions but is upregulated in Alzheimer's disease.[5]

Tau is a prototypical intrinsically disordered protein that does not assume a rigid tertiary or secondary structure but populates an ensemble of interconverting structures in solution.[6,7] Because of the inherent flexibility of tau, NMR spectroscopy is the only method that allows a description of its conformations and dynamics with high resolution.[8] We have recently shown that it is possible to obtain the complete backbone resonance assignment of the longest isoform of human tau and demonstrated that 441-residue tau has a distinct domain character with an intricate network of long-range interactions.[7]

Here we investigate changes in the local and global structure of 441-residue tau related to phosphorylation in the epitopes recognized by the Alzheimer diagnostic antibodies AT8 (S199E, S202E, T205E), AT100 (T212E and S214E), and PHF1 (S396E and S404E). To avoid the ambiguities of heterogeneous phosphorylation, we cloned "pseudophosphorylation" mutants of tau in which serine and threonine residues were converted into glutamic acid. The same mutant protein was previously shown to aggregate slightly faster in comparison with wild-type (wt) tau.[9] In addition, a different six-site pseudophosphorylation mutant (S199E, S202E, T205E, T231E, S396E, and S404E) had a decreased rate of elongation and a pronounced lag time of aggregation.[10] We determined cross-validated ensembles of wild-type (wt) and pseudophosphorylated tau using a large number of long-range distance restraints. Our data reveal that pseudophosphorylation weakens the transient folding of tau.

To obtain insight into transient long-range interactions in wt and pseudophosphorylated (so-called E-mutant) tau, we used the technique of paramagnetic relaxation enhancement (PRE).[11] Covalent attachment of a spin label to a cysteine induces PRE of the NMR signals of the protein within a distance of ∼25 Å. The relaxation enhancement can then be quantified by comparison of NMR signal intensities in two-dimensional correlation experiments recorded in the paramagnetic (I_{para}) and diamagnetic (I_{dia}) states. An I_{para}/I_{dia} intensity ratio of >0.95 indicates a distance exceeding 25 Å. We attached the paramagnetic nitroxide label MTSL to 10 different positions (C15, C72, C125, C178, C239, C256, C322, C352, C384, and C416) uniformly distributed along the primary sequence of wt and E-mutant 441-residue tau (Figures 1 and 2). In addition, PRE measurements were performed in which a spin label was attached simultaneously to the two native cysteines C291 and C322 of tau. Our previous characterization of wt tau was based on six MTSL positions.[7]

In Figure 1, a comparison between the PRE broadening profiles of the amide protons of wt and E-mutant tau is shown for MTSL attached at positions 239 and 384, which are close to the sites of pseudophosphorylation. In agreement with our previous study,[7] the spin label at position 239 strongly attenuated

Received: June 23, 2011

   dx.doi.org/10.1021/ja205836j | J. Am. Chem. Soc. XXXX, XXX, 000–000

the signals of 60 residues at the N-terminus with weaker broadening extending up to residue 140. In addition, weak paramagnetic broadening was observed for about 60 residues upstream and downstream of the site of the spin label (Figure 1a). Very similar PRE profiles were obtained when the diamagnetic state was obtained by addition of ascorbic acid instead of dithiothreitol (DTT) (Figure S1 in the Supporting Information). For residues 150−441 of the E-mutant, a highly similar PRE profile was observed. However, at the N-terminus of the E-mutant, the PRE intensity ratios were higher than in wt tau by up to 34% for residues M31−D34. The reduced paramagnetic effect demonstrates that pseudophosphorylation in the regions flanking the repeat domain of tau attenuate their transient interaction with the N-terminal domain. MTSL attached at position 384 induced a broad PRE profile around the attachment site and weak paramagnetic enhancement at the N-terminus for both proteins. In the E-mutant, the paramagnetic effect was reduced for residues 410−441 (Figure 1b).

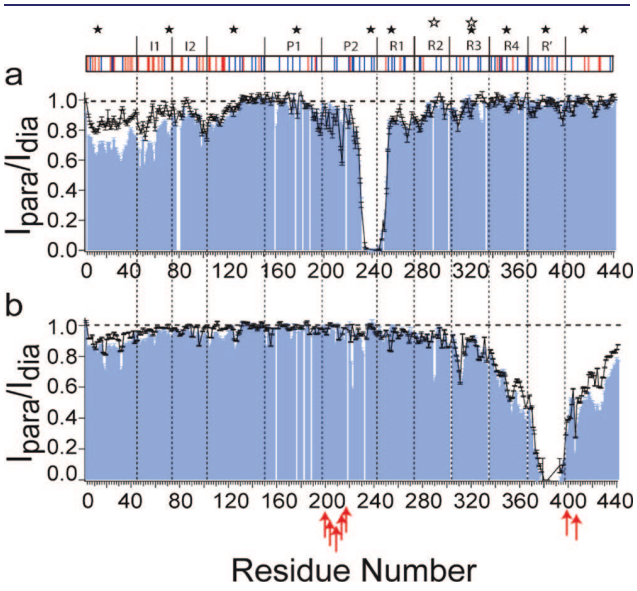


Figure 1. PRE broadening profiles of amide protons of wt (blue) and E-mutant 441-residue tau (black line) with MTSL attached to positions (a) 239 and (b) 384. Arrows mark sites of mutation to glutamic acid. Intensity ratios were averaged over a three-residue window. Decreases in peak intensity ratios that occur far from the site of spin-labeling (>10 residues) are indicative of long-range contacts. The domain organization shown at the top (inserts I1, I2; proline-rich regions P1, P2; domain of repeats R1−R4; pseudorepeat R′) highlights the location of negative (red) and positive (blue) charges. The sites of spin labeling are indicated by solid (single spin label) and open (double spin label) stars.

The PRE profiles in Figure 1 provide information only about changes in transient long-range structure involving residues 239 and 384. To probe the effect of pseudophosphorylation on the ensemble of conformations in other regions of tau, nine additional PRE profiles were compared (Figure S2). The NMR data demonstrate that residue stretches harboring each of the nine sites are involved in long-range interactions in both wt and E-mutant tau. For all attachment sites, the paramagnetic effect was at least slightly reduced in distinct regions upon pseudophosphorylation. According to the PRE profile for MTSL at position 15, pseudophosphorylation in the regions flanking the repeat domain causes a weakening of the long-range contact between the N- and C-termini (Figure S2). A weakened contact between the two termini is in agreement with a lower fluorescence resonance energy transfer (FRET) efficiency between residues 432 and 17 in the E-mutant relative to wt tau.[9] However, the FRET efficiencies between residues 310 and 17 and between residues 432 and 322 were larger in the E-mutant than in the wt protein,[9] in apparent contrast to the NMR PRE profiles. We attribute the differences to the use of two hydrophobic labels (tryptophan and dansyl group) in the FRET studies.

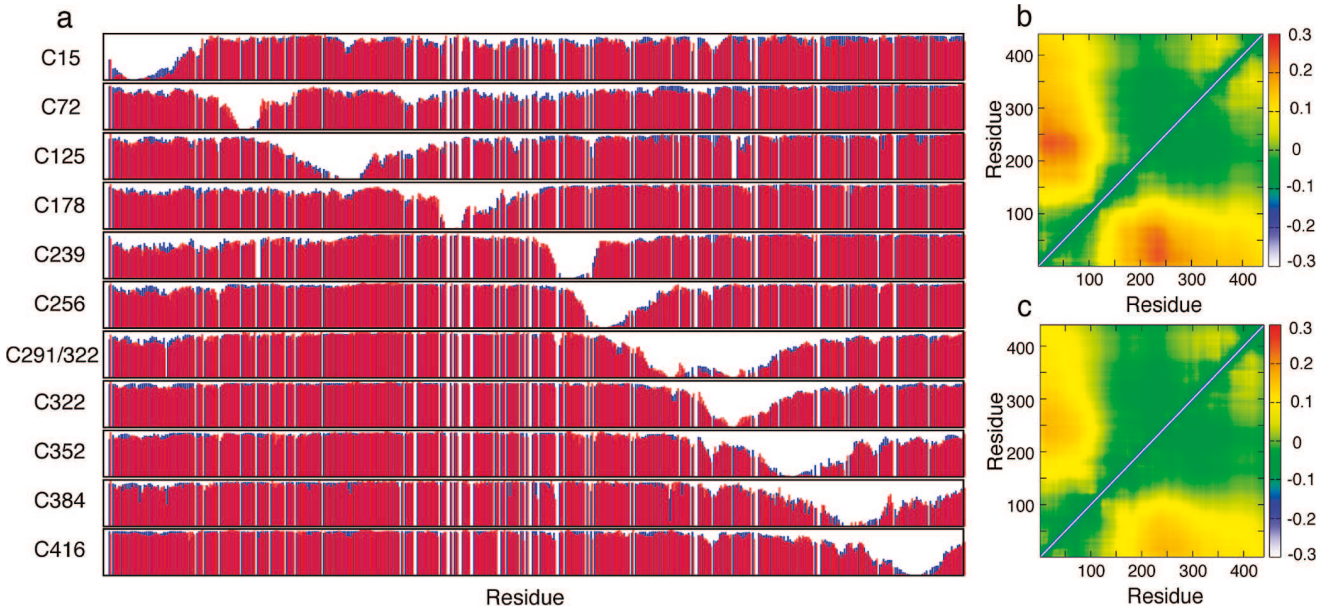


Figure 2. PREs and long-range contacts in the representative ensembles of (a, b) wt and (c) E-mutant 441-residue tau. (a) Comparison of experimental PREs (red) and values back-calculated from the representative ensemble of wt tau (blue). (b, c) Contacts are plotted as $\log(d_{ij}/d_{ij,\mathrm{ref}})$, where $d_{ij,\mathrm{ref}}$ refers to the distance in the reference ensemble of 27 000 structures and d_{ij} refers to the distance in the selected ensemble [see Figure S14 for a difference plot of (b) and (c)].

dx.doi.org/10.1021/ja205836j |J. Am. Chem. Soc. XXXX, XXX, 000–000

A hint concerning the mechanism of the pseudophosphorylation-induced weakening of transient long-range interactions was provided by an analysis of the charge distribution of tau. While the central region of tau is predominantly positively charged, an excess of negative charge was found for the 120 N-terminal residues as well as at the C-terminus. Introduction of five glutamic acid residues in the region 199−214 changes the net charge in the proline-rich region P2 from +2 to −3 (Figure 1a). The altered electrostatic properties result in a reduced Coulombic attraction between P2 and the positively charged N-terminal domain.

The overall dimensions of the ensemble of conformations populated by tau can be estimated using diffusion NMR spectroscopy.[12] We estimated ensemble-averaged hydrodynamic radii of 56.8 ± 0.2 Å for wt tau and 60.2 ± 0.3 Å for the E mutant, both in 90% H_2O/10% D_2O (Figure S3), supporting a less compact ensemble of conformers for the E-mutant. The value for wt tau is slightly larger than the hydrodynamic radius estimated previously in 100% D_2O,[7] potentially as a result of different viscosities and stronger intramolecular hydrophobic interactions in D_2O.[13]

The raw PRE data do not constrain a single conformer but should be converted into ensembles of structures for which the PRE-derived distance restraints must be fulfilled by the complete ensemble.[14] Previously, we calculated an ensemble of 10 conformers of wt tau that was in agreement with 2288 distance restraints derived from six MTSL positions.[7] Here we calculated one ensemble for the wt protein and one for the E-mutant using 4646 and 4545 distance restraints, respectively. In addition, while in our previous study simulated annealing was used to drive the compaction of the ensemble under the influence of the PRE restraints,[7] here subensembles of 200 structures were selected from a statistical coil ensemble of 30 000 conformers using the program ASTEROIDS.[14−16] The flexibility of the MTSL side chain was taken into account by rotamer modeling as described previously (Figure S4). In order to determine the optimal number of structures for the ensemble, complete PRE data sets for single MTSL positions were removed from the analysis, and these values were back-calculated from subensembles that were determined using the remaining 10 PRE data sets (Figures S5−S8). In addition, random samples of Gaussian noise were added to the input data to test the sensitivity of the ensembles to experimental uncertainties (Figures S9 and 10). To test the effect of local structural preferences on the properties of the ensemble, we performed two sets of test calculations: (i) local sampling for residues 395−425 with alternating 20% more α-helical and 20% more β-sheet interspersed with standard sampling and (ii) 25% helix content for residues 428−437 in agreement with NMR chemical shifts[7] (Figures S11 and S12).

The calculations showed that the mobility of the MTSL label, experimental uncertainties in the PREs, and local conformational preferences of the backbone do not affect the overall properties of the ensemble, although details in the ensembles might change. In addition, selection against the data resulted in slightly better reproduction than cross-validation. The contrary would be expected only if the system were overdetermined and the data that were left out were completely redundant. Thus, very large data sets are required for a full description of ensembles of disordered proteins. On the other hand, the cross-validation showed that the model, which used all of the data except that for one mutant, was still consistent with the data from this mutant. It also showed that subensembles of 200 structures describe the experimental data well and reproduce the back-calculated PREs with high quality (Figure 2 and Figure S13). Although each run of the program produces a different ensemble in terms of individual conformers, biophysical parameters such as long-range contacts and radii of gyration are constant.

The calculated ensembles support the interpretation of the raw data (Figure 2 and Figure S14): In the E-mutant, the N-terminal domain has a decreased probability to be in proximity to the proline-rich region, and the compaction of the C-terminus is reduced. The overall dimensions of the ensemble of structures were slightly different, with hydrodynamic radius values of 58.9 and 60.0 Å for wt and E mutant tau, respectively (as predicted by HydroPRO[17]), which is in qualitative agreement with the experimental values. The structure calculations demonstrated that the experimental PRE profiles obtained for different MTSL positions are consistent. Importantly, we looked at differences between two data sets and therefore two ensembles with exactly the same extent and distribution of data.

Next, we asked whether changes in the global structure are connected to changes in the local structure of tau. To answer this, we assigned and measured ^{15}N and $^{13}C^{\alpha}$ chemical shifts[18] as well as $^{3}J_{HNHA}$ scalar couplings in the E-mutant and compared them to those of wt tau. Both proteins showed only small deviations from random coil values of $^{13}C^{\alpha}$ chemical shifts, in agreement with their disordered nature (Figure S15). Direct comparison of the experimental $^{13}C^{\alpha}$ chemical shifts for wt and E-mutant tau demonstrated that local conformational changes induced by pseudophosphorylation are small and restricted to the vicinity of the mutation. Rigid secondary structure is not induced, in agreement with $^{3}J_{HNHA}$ scalar couplings (Figure S15) and previous work on short tau peptides.[19] Despite the fact that the PHF1 epitope near the C-terminus contains only two phosphorylation sites (396 and 404), $^{13}C^{\alpha}$ chemical shift changes were more pronounced and affected a larger set of residues than did glutamic acid mutations in the proline-rich regions, suggesting that the PHF1 epitope is more prone to conformational changes.

In summary, we have determined an ensemble of conformers of 441-residue tau on the basis of a large number of distance restraints derived from paramagnetic relaxation enhancement and compared it to an equally well defined ensemble for a mutant protein mimicking phosphorylation. Our data demonstrate that pseudophosphorylation at the AT8, AT100, and PHF1 epitopes reduces the electrostatic attraction between the N-terminal domain and the proline-rich region of tau and causes a weakening of the network of transient long-range interactions of 441-residue tau. Aggregation studies have previously suggested that the identical mutant protein aggregates faster than wt tau.[9] Together with studies on the Parkinson-related protein α-synuclein, in which long-range interactions between the C-terminal tail and the hydrophobic central part delay aggregation,[20,21] our data indicate that networks of transient long-range interactions are common properties of intrinsically disordered proteins and important for aggregation.

■ ASSOCIATED CONTENT

Supporting Information. Eleven PRE broadening profiles, chemical shifts, and ^{3}J couplings for wt and E-mutant tau. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
mzwecks@gwdg.de

■ ACKNOWLEDGMENT

We thank Ilka Lindner for excellent technical support, Eva-Maria Mandelkow and Christian Griesinger for discussions, and the Max Planck Society and the DFG (ZW 71/2-2 and 7-1 to M.Z.) and TAUSTRUCT - ANR MALZ 2010 (to M.B.) for financial support.

■ REFERENCES

(1) Ballatore, C.; Lee, V. M.; Trojanowski, J. Q. *Nat. Rev. Neurosci.* **2007**, *8*, 663–672.
(2) Garcia, M. L.; Cleveland, D. W. *Curr. Opin. Cell Biol.* **2001**, *13*, 41–48.
(3) Bancher, C.; Brunner, C.; Lassmann, H.; Budka, H.; Jellinger, K.; Wiche, G.; Seitelberger, F.; Grundke-Iqbal, I.; Iqbal, K.; Wisniewski, H. M. *Brain Res.* **1989**, *477*, 90–99.
(4) Mi, K.; Johnson, G. V. *Curr. Alzheimer Res.* **2006**, *3*, 449–463.
(5) Gong, C. X.; Liu, F.; Grundke-Iqbal, I.; Iqbal, K. *J. Neural Transm.* **2005**, *112*, 813–838.
(6) Cleveland, D. W.; Hwo, S. Y.; Kirschner, M. W. *J. Mol. Biol.* **1977**, *116*, 227–247.
(7) Mukrasch, M. D.; Bibow, S.; Korukottu, J.; Jeganathan, S.; Biernat, J.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. *PLoS Biol.* **2009**, *7*, No. e34.
(8) Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, *104*, 3607–3622.
(9) Jeganathan, S.; Hascher, A.; Chinnathambi, S.; Biernat, J.; Mandelkow, E. M.; Mandelkow, E. *J. Biol. Chem.* **2008**, *283*, 32066–32076.
(10) Sun, Q.; Gamblin, T. C. *Biochemistry* **2009**, *48*, 6002–6011.
(11) Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 170–184.
(12) Wilkins, D. K.; Grimshaw, S. B.; Receveur, V.; Dobson, C. M.; Jones, J. A.; Smith, L. J. *Biochemistry* **1999**, *38*, 16424–16431.
(13) Cioni, P.; Strambini, G. B. *Biophys. J.* **2002**, *82*, 3246–3253.
(14) Salmon, L.; Nodet, G.; Ozenne, V.; Yin, G.; Jensen, M. R.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2010**, *132*, 8407–8418.
(15) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 17908–17918.
(16) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.
(17) Garcia De La Torre, J.; Huertas, M. L.; Carrasco, B. *Biophys. J.* **2000**, *78*, 719–730.
(18) Wishart, D. S.; Sykes, B. D. *Methods Enzymol.* **1994**, *239*, 363–392.
(19) Bielska, A. A.; Zondlo, N. J. *Biochemistry* **2006**, *45*, 5527–5537.
(20) Bertoncini, C. W.; Jung, Y. S.; Fernandez, C. O.; Hoyer, W.; Griesinger, C.; Jovin, T. M.; Zweckstetter, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1430–1435.
(21) Hoyer, W.; Cherny, D.; Subramaniam, V.; Jovin, T. M. *Biochemistry* **2004**, *43*, 16233–16242.

Molecular BioSystems

Cite this: DOI: 10.1039/c1mb05291h

www.rsc.org/molecularbiosystems

REVIEW

Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy†

Robert Schneider, Jie-rong Huang, Mingxi Yao, Guillaume Communie, Valéry Ozenne, Luca Mollica, Loïc Salmon, Malene Ringkjøbing Jensen and Martin Blackledge*

Received 14th July 2011, Accepted 8th August 2011
DOI: 10.1039/c1mb05291h

In order to understand the conformational behaviour of Intrinsically Disordered Proteins (IDPs), it is essential to develop a molecular representation of the partially folded state. Due to the very large number of degrees of conformational freedom available to such a disordered system, this problem is highly underdetermined. Characterisation therefore requires extensive experimental data, and novel analytical tools are required to exploit the specific conformational sensitivity of different experimental parameters. In this review we concentrate on the use of nuclear magnetic resonance (NMR) spectroscopy for the study of conformational behaviour of IDPs at atomic resolution. Each experimental NMR parameter is sensitive to different aspects of the structural and dynamic behaviour of the disordered state and requires specific consideration of the relevant averaging properties of the physical interaction. In this review we present recent advances in the description of disordered proteins and the selection of representative ensembles on the basis of experimental data using statistical coil sampling from *flexible-meccano* and ensemble selection using ASTEROIDS. Using these tools we aim to develop a unified molecular representation of the disordered state, combining complementary data sets to extract a meaningful description of the conformational behaviour of the protein.

Introduction

One of the most remarkable discoveries of protein science over the last decade concerns the revelation that a large fraction of functional proteins encoded by the human genome is either fully disordered or contains long disordered regions.[1–4] Intrinsically disordered proteins (IDPs) remained beyond the scope of classical structural biology, and therefore escaped the attention of the multiplication of structural genomics projects that have emerged in the hope of classifying all protein folds. IDPs are biologically functional despite a lack of stable, well-defined three-dimensional structural fold, and as such they impose a different perspective on the relationship between primary protein sequence and function. IDPs are also strongly involved in numerous human pathologies, and the development of pharmacological solutions to these problems awaits a molecular description of the role of flexibility in the development of disease.[5–7] Proteins present a vast spectrum of flexibility in their physiological states, from stable enzymes to highly flexible chains. In analogy to folded proteins, the primary sequence predetermines the functional behaviour of the protein, but in this case, rather than focussing on a unique fold that stabilizes the protein, and considering the role of local structure and dynamics relative to this scaffold, we are forced to consider the more central role that conformational flexibility plays in the function of the intrinsically disordered state. The determination of a single structure has no real physical relevance, at least in the free form of such proteins, and there is therefore a pressing need for the development of an entirely new set of experimental and descriptive approaches to describe the conformational behaviour of IDPs.[8–11]

One obvious aim of a structural description of IDPs is to determine rules that define the behaviour of the flexible protein in terms of probability to populate a defined region of conformational space. This is often achieved by evoking an explicit ensemble description of interconverting structures, whose populations are interpreted in terms of a population-weighted distribution that represents the true conformational equilibrium. However the definition of this distribution is no easy task. IDPs populate a vast conformational space, and the mapping of this potential energy landscape represents a classical ill-posed problem, in which the number and complexity of the available degrees of conformational freedom far outweigh the accessible experimental data that can be

Protein Dynamics and Flexibility, Institut de Biologie Structurale Jean-Pierre Ebel, CEA, CNRS, UJF UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, France. E-mail: martin.blackledge@ibs.fr; Tel: +33 4 38789554

† Published as part of a Molecular BioSystems themed issue on Intrinsically Disordered Proteins: Guest Editor M. Madan Babu.

measured for a particular system. Some caution therefore needs to be exercised when treating such under-determined systems, where the development of an ensemble description that is in agreement with the experimental data may not ensure that the associated conformational sampling is correct. The development of robust procedures that address this issue is of paramount importance.

NMR of intrinsically disordered proteins

Characterisation of the diverse conformational properties of the unfolded protein cannot be based solely on a single experimental technique, but necessarily relies on the exploitation of complementary approaches reporting on both short range and long-range structural parameters. It is also essential to consider the time scales that characterise local and global motions and the inter-conversion rates of different members of a conformational ensemble. Nuclear magnetic resonance (NMR) spectroscopy is particularly rich in both short range and long-range structural information that can be exploited to accurately define the behaviour of IDPs.[12] Despite a comparatively restricted amide proton chemical shift dispersion, NMR signals retain the spectroscopic characteristics of small molecules, because of the flexibility of the chain, so that heteronuclear chemical shift assignment remains possible, even for very large intrinsically disordered proteins.[13] Molecular weight restrictions that apply to folded proteins therefore do not extend in the same way to intrinsically disordered proteins of the same number of amino acids.

Most importantly NMR provides access to ensemble and time averaged conformationally dependent parameters at atomic resolution. The measurement of structurally dependent parameters inherently provides a basic tool to study local conformational propensities that may be important for folding upon binding,[14] and transient or persistent long-range contacts or tertiary structure that may also play a role in molecular interactions.[14–16] In this article we describe advances of some NMR-based techniques that have taken place in recent years for the description of the conformational behaviour of IDPs.[17–19]

The chemical shift of a specific nucleus reports on the local physico-chemical environment of the nucleus, and in the presence of conformational flexibility, depends on a population-weighted average over local conformations sampled by all molecules in the ensemble that are exchanging on timescales faster than the millisecond. This timescale therefore dictates our interpretation of all NMR parameters that are measured from this chemical shift averaging process. The chemical shift can also provide information about the local structural propensity[20] that can be detected in intrinsically disordered proteins by analyzing the deviation of measured parameters from the expected value that would be measured in the absence of any local structure (the so-called 'random coil' value).[21,22] The absolute definition of a random coil remains open to argument, in most cases amino-acid specific values are measured experimentally from small peptides with no apparent local structure.[23–25] The chemical shift provides a sensitive probe of local structural sampling, in particular ^{13}C shifts, whose values depend, in order of importance, on the covalent structure ($^{13}C^{\alpha}$, $^{13}C^{\beta}$ or $^{13}C'$), the type of amino acid, and finally on the local structural propensity which is the parameter of interest. The difference between the measured shift and the amino-acid specific random coil shift, known as the 'secondary' chemical shift, is commonly used to identify the presence of transient structure in flexible chains.[26–28] Scalar couplings between nuclei on the backbone of the protein also depend on backbone dihedral angles and average in a similar way to chemical shifts.[29–31] Again random coil values have been measured in small peptides and these values can be compared to experimental values to determine the level of transient local structure.

Residual dipolar couplings (RDCs), measured between pairs of nuclei, are also extremely promising tools for studying the conformational behaviour of disordered proteins.[32–36] RDCs become measurable when the protein of interest is dissolved in a dilute liquid crystalline medium, such that the average dipolar coupling, normally averaged to zero in free solution, has a residual, non-zero value.[37–39] Under these conditions RDCs depend on the average over the ensemble of orientations of the vector connecting the two spins in the following way:

$$D_{ij} = -\frac{\gamma_i \gamma_j \hbar \mu_0}{8\pi^2 r^3} \left\langle \frac{3\cos^2 \Omega - 1}{2} \right\rangle \qquad (1)$$

where Ω is the orientation of the internuclear vector with respect to the static magnetic field and r is the vibrationally averaged distance. The angular parentheses again describe an average over conformations that exchange with rates faster than the millisecond timescale. RDCs are highly sensitive probes of time and ensemble-averaged conformational equilibria on timescales up to the millisecond in folded proteins,[40–44] but can also be used to characterize the conformational behaviour of unfolded proteins. The sensitivity of RDCs to the local structure in an otherwise unfolded chain can be best illustrated by considering the orientation of an amide bond vector. The expected average orientation of the amide vectors present in an unfolded chain aligned in a direction parallel to the magnetic field is approximately orthogonal to the field, resulting in coupling with a negative sign. If a helical element is present, this will induce a change in sign of the measured coupling, because the bond vector would be aligned rather in an average parallel direction with respect to the average chain direction. The angular averaging term in eqn (1) changes sign and so does the dipolar coupling (Fig. 1). Over the last decade significant progress has been made in developing an understanding of the nature of RDCs in the unfolded state, and the potential for exploiting this information has generated considerable interest in the development of new approaches to exploit this experimental parameter.[45–47]

Disordered proteins often exhibit evidence of fluctuating long-range tertiary structure, that may be important for physiological interactions, for example *via* so-called fly-casting interactions,[16] in the control of early folding events, or to provide protection from aggregation or proteolysis. While it is difficult to detect such transient contacts *via* standard approaches to the measurement of internuclear distances, using ^{1}H–^{1}H cross relaxation, the detection of such long-range information is possible by exploiting the strength of

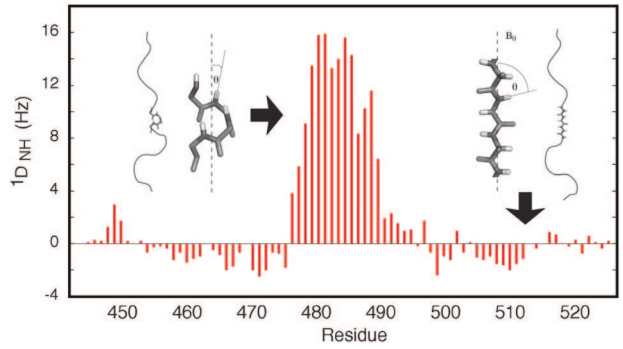


Fig. 1 Illustration of the sensitivity of RDCs to the presence of local structure. The orientational dependence shown in eqn (1) results in positive $^1D_{NH}$ RDCs for the central helical element, where the NH bond vectors tend to be aligned with the field, while in the disordered regions the RDCs are negative, because the average orientation is perpendicular to the direction of the chain.

the dipolar relaxation between the nuclear spin and an unpaired electron that can be introduced into the protein by attaching a nitroxide group to a cysteine mutant.[48–50] Because the gyromagnetic ratio of the electron spin is over 600 times higher than the proton spin, the observed line broadening due to the paramagnetic relaxation enhancement provides long-range probes of intra- and intermolecular distances and distance distribution functions that can be detected even if only weakly populated.[51–57]

A number of additional NMR parameters can be used to characterize the unfolded state: the most common are pulse-field gradient spin echo experiments,[58] that report on the population weighted average translational diffusion properties of the chain and heteronuclear spin relaxation, that report on local order on picosecond to nanosecond timescales.[59–61] The complementary information available from small angle X-ray scattering that reports on the average mass distribution in three dimensional space, and therefore the dimensions of the ensemble of structures, is also often exploited in combination with NMR data to provide a more complete picture of the disordered state.[62–68]

Ensemble descriptions of IDPs from NMR data

Despite remarkable progress in recent years, the transformation of these highly diverse experimental parameters into a meaningful conformational description remains a key challenge for contemporary structural biologists. The most common approach that has been applied over recent years borrows tools developed for the determination of the structure of proteins in solution, where additional terms are incorporated into a physical potential energy function to bias the conformational sampling. A restrained molecular dynamics (MD) simulation, run in parallel over different members of the ensemble, is then used to drive the ensemble into a region of conformational space that is in agreement with experimental data.[69–74] Despite the popularity of such techniques, a number of key questions remain open with respect to their generalisation. It is not clear how the introduction of non-physical parameters into the force field will affect the ability of the molecular dynamics engine to efficiently search conformational phase space, or its ability to sample a Boltzmann-weighted distribution of conformers. It is also unclear how to optimize the number of structures present in the ensemble average, a feature that will depend strongly on the density and information content of the experimental parameters. A more general problem, that is shared by all approaches to the interpretation of experimental data from disordered states, concerns the characteristic averaging timescales of each experimental parameter that must be properly accounted for within the conformational ensemble.

An entirely different approach does not use the experimental data to drive the individual members of the ensemble into a conformation in agreement with the experimental data, but instead samples conformational space as broadly as possible, and then exploits the experimental data to define the region of conformational space that is appropriate for the system under investigation. Enhanced molecular dynamics approaches such as accelerated molecular dynamics have been used in this way to study intrinsic dynamics in folded proteins,[44,75,76] although the potential extent of conformational space available to IDPs complicates the successful application of such approaches to these highly flexible systems. An alternative strategy is to attempt to flood conformational space by creating a statistical coil model of the protein based on the intrinsic conformational behaviour of each amino acid, derived for example from backbone dihedral angle distributions found in loop regions of protein structures.[77–79]

An explicit ensemble description of IDPs, called *flexible-meccano*, builds multiple copies of the protein that are ensemble designed to represent all possible states that are relevant for the NMR observable.[35] *Flexible-meccano* randomly samples amino-acid-specific backbone dihedral angle $\{\phi/\psi\}$ propensities derived from non-secondary structural elements of high-resolution X-ray crystallographic structures,[80] and thereby assembles a conformational ensemble from which experimental values can be calculated. Amino-acid specific hard-sphere steric clashes are used to provide a physically reasonable model of repulsive interatomic forces, and no attractive forces are explicitly used. The simplicity of the model allows for highly efficient structure ensemble assembly (100 000 structures of a 100 amino acid protein can be created in 30 minutes on a single processor). The ensembles are randomly sampled from population-weighted distributions that are taken to represent the potential energy surface of each amino acid. Although this does not guarantee a Boltzmann distribution, the absence of additional constraints in this sampling phase avoids distortions due to additional potential energy terms such as those used in restrained MD calculations.

The presence of a single set of signals detected in NMR spectra of denatured and intrinsically disordered proteins imposes the implicit assumption that all conformers used to predict an experimental value are in rapid exchange on timescales faster than the millisecond. The ensemble of structures can then be used to predict experimental values that would be measured if the statistical coil model were relevant. For the prediction of chemical shifts and scalar couplings, local structural information is sufficient to predict the expected value, while for RDCs the calculation of the expected alignment of each conformer is necessary before averaging over the ensemble.

In the most common case of steric alignment this calculation is performed on the basis of the three dimensional shape of the protein.[81]

RDCs simulated using this very simple approach predict values in reasonable agreement with experimental couplings measured in both intrinsically disordered and chemically denatured proteins. Initial studies already indicated that the orientational space sampled by inter-nuclear bond-vectors from RDCs is sensitive enough to pick up differences in amino-acid specific backbone dihedral angle distributions, even in the absence of secondary structural propensity.[10,35] *Flexible-meccano* has also been used in combination with molecular dynamics based simulations, to quantify the level of β-turn propensity in the K18 domain of the protein Tau[82] and α-helical propensity in the transactivation domain of the protein p53.[83]

While $N–H^N$ RDCs alone have been shown to provide evidence for local structural propensity, the measurement of multiple RDCs from each peptide unit provides the necessary information to make quantitative estimates of the detail and population of the structural elements. Thus, the combination of RDCs from different bond-vectors ($N–H^N$, $H^\alpha–C^\alpha$, $C'–H^N$, $C^\alpha–C'$) was also shown to be crucial to the description of the length and population of different helical structures that form the rapidly exchanging conformational equilibrium of the molecular recognition element of the disordered C-terminal domain of the nucleoproteins from *Sendai* and *measles* viruses.[84,85] In this case entire ensembles of all possible helical elements were calculated, and the minimum combination that could reproduce the experimental data was determined, along with their associated populations. Remarkably, in both cases, the helical elements present in the molecular recognition elements that were significantly populated in solution were found to follow amino acids with known propensity to stabilize helices in free solution.[85] An extensive set of RDCs, including a large number of long-range $^1H–^1H$ couplings, were measured in the protein Ubiquitin in its denatured state,[87] and used in combination with *flexible-meccano* to identify modifications of the statistical coil model that are appropriate to account for conformational sampling of the unfolded chain in the presence of the denaturant.[88,89]

The statistical coil description of the disordered state thus provides a relatively straightforward approach for calculating RDC profiles that would be expected if the protein behaved as a random coil. The establishment of such approaches is essential in order to develop a clear understanding of the origin of experimentally observed fluctuations in the absence and in the presence of specific or persistent local or long-range structure. However the next step, requiring the quantitative interpretation of departures from expected random coil values in terms of specific local or long-range conformational behaviour, is of equal importance and fundamentally more challenging.[10,90,91]

Determination of meaningful ensembles in agreement with experimental data

A number of studies have applied a rational, hypothesis-based approach, calculating explicit ensembles containing tens of thousands of conformers from different conformational sampling regimes and comparing the ensemble-averaged couplings to experimental data. In some case this is achieved with the aid of molecular dynamics simulation to create alternative conformational sampling that provides agreement with experimental data.[82,83] While these studies are informative and important to advance our understanding of the field, in order to generalize the methodology it is necessary to take the analysis one step further, and develop approaches that can accurately define the conformational sampling of the peptide chain directly from the experimental NMR data.

In order to address this issue, the ensemble selection algorithm, ASTEROIDS (A Selection Tool for Ensemble Representation Of Intrinsically Disordered States) has been developed to determine appropriate regions of conformational space populated by the IDP by selection of conformers from the *flexible-meccano* ensemble using experimental NMR data.[92–94] The ASTEROIDS algorithm is based on an efficient genetic algorithm that is used to propose conformational ensemble descriptions selected from a large pool of possible conformers that are in agreement with the experimental data. In order to identify conditions under which an approach that evokes a sub-ensemble of structures can be accurately applied to describe a pseudo-continuum of conformers, we systematically adopt the following simple procedure that clearly quantifies the conformational accuracy of such approaches: (1) Data are simulated under specific conditions of conformational sampling and appropriately averaged over an ensemble of a very large number of conformers (between 50 and 100 thousands). (2) Sub-ensembles of tractable size are generated using ASTEROIDS to be in agreement with these data, and the conformational sampling represented in these ensembles is compared to the target sampling used in step (1) to generate the data.

One of the most important problems encountered in the treatment of RDCs derives from the large number of structures required before a simple arithmetic average reaches convergence. The reason for this is that, in addition to the obvious dependence on local conformational sampling, the RDCs for each individual conformer depend on conformational degrees of freedom throughout the molecule, that each define the shape of the protein, and therefore the size and distribution of the RDCs. Indeed, convergence of RDCs from a 76 amino acid chain is not yet achieved in 10 000 structures. More rapid convergence of RDCs can be achieved using a smaller number of conformers if the protein were divided into short, uncoupled segments (Local Alignment Windows—LAWs) and the RDCs are calculated using the alignment tensor of these segments.[95] This is an important result: the ability to describe the conformational properties using fewer structures renders ensemble selection more tractable.

However there are important aspects that need to be addressed before such approaches can be used to explain experimental data. Adopting the procedures described above, RDCs were calculated using specific conformational sampling regimes averaged over a large ensemble.[92] The average RDCs were then used, in combination with a 15 amino acid window, to select different sized ensembles of conformers from a large pool in agreement with the data. The results demonstrated that

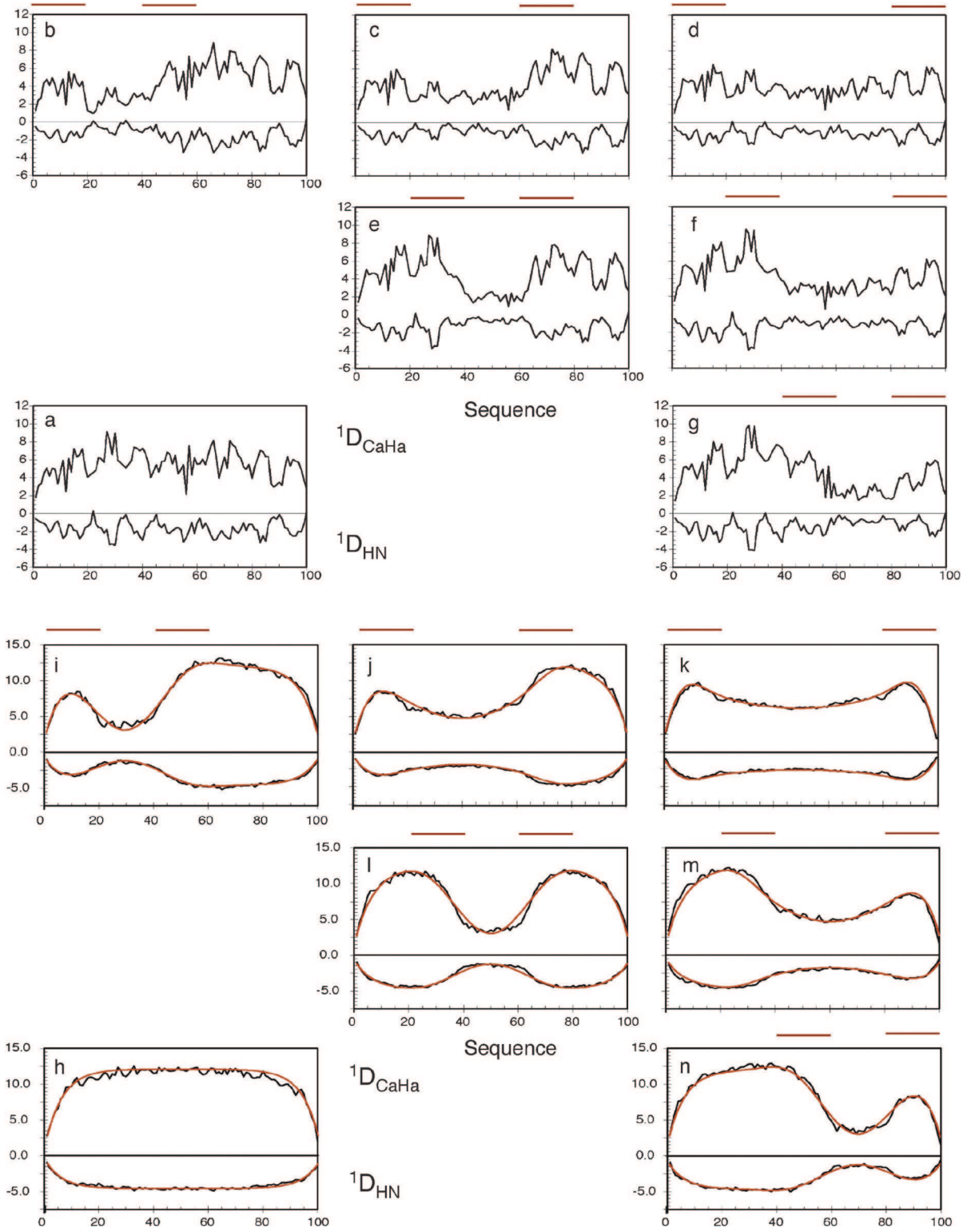


Fig. 2 The effects of long-range contacts on expected RDC profiles. Top: (a) $^1D_{NH}$ and $^1D_{CaHa}$ RDCs calculated for a 100 amino acid sequence in the absence of specific contacts. The program PALES was used to calculate RDCs from each conformer. 100 000 conformers were used in this and each average shown in figures (b–n). (b–g) The same calculation is performed, but conformers are only retained in the ensemble if at least one inter-C^β distance exists between the primary sequence ranges shown below the red lines: (b) i = 1–20, j = 41–60, (c) i = 1–20, j = 61–80, (d) i = 1–20, j = 81–100, (e) i = 21–40, j = 61–80, (f) i = 21–40, j = 81–100, (g) i = 41–60, j = 81–100. Bottom: (h) $^1D_{NH}$ and $^1D_{CaHa}$ RDCs calculated for a 100 amino acid poly-valine sequence in the absence of specific contacts. (i–n) The same calculation is performed, but conformers are only retained in the ensemble if at least one inter-C^β distance exists between the primary sequence ranges shown below the red lines: (i) i = 1–20, j = 41–60, (j) i = 1–20, j = 61–80, (k) i = 1–20, j = 81–100, (l) i = 21–40, j = 61–80, (m) i = 21–40, j = 81–100, (n) i = 41–60, j = 81–100. The dark red curves show the analytical reproduction of the long-range effects on the RDCs with the contact positioned in the centre of each region. Reprinted with permission from the *Journal of the American Chemical Society*.[93]

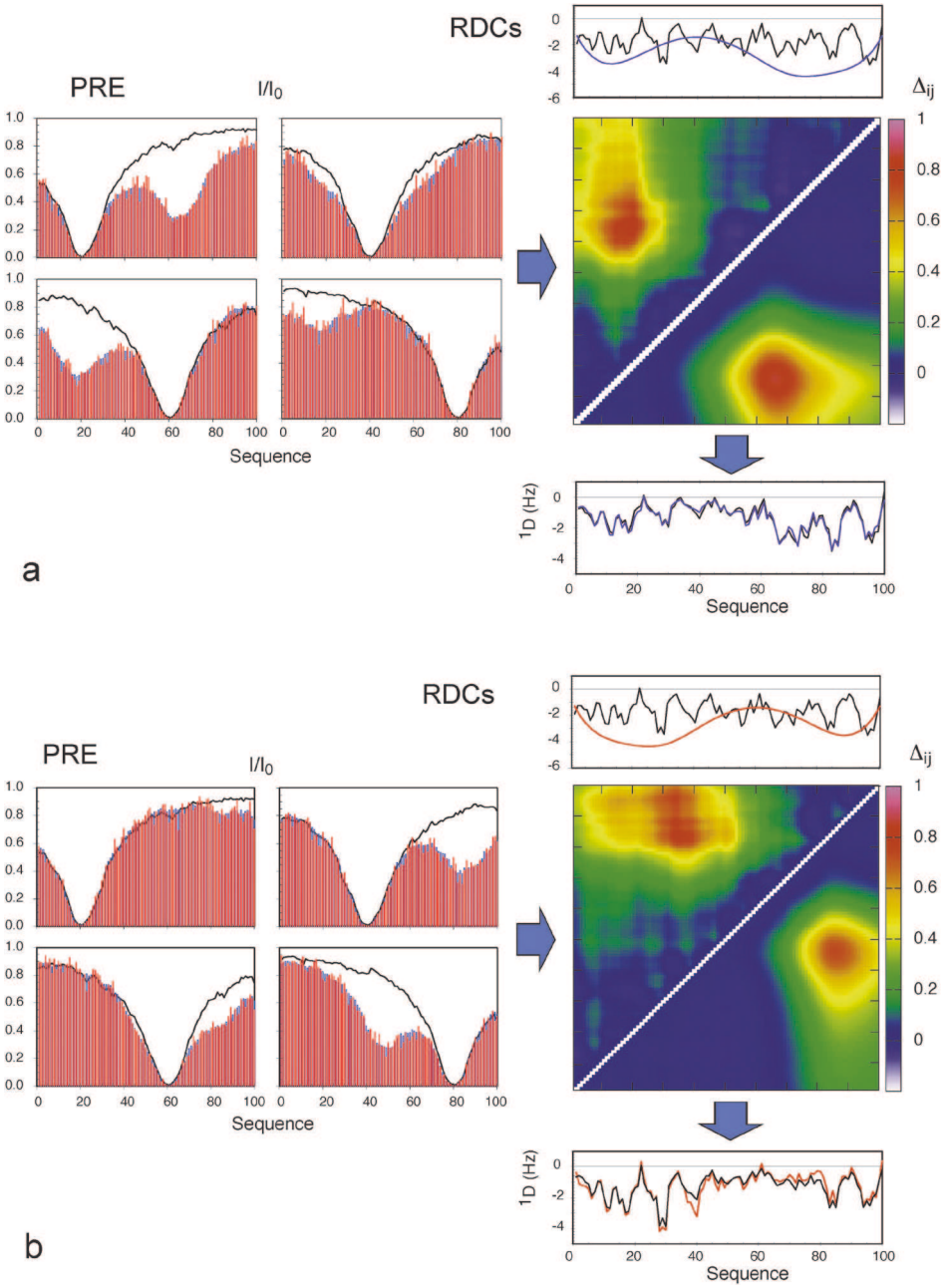


Fig. 3 Combination of effects of long-range order derived from PREs with local conformational sampling using local alignment windows for the interpretation of RDCs. (a) Blue: Data averaged over the target ensemble where each conformer has a contact between 11–20 and 61–70. Red: Average PREs over an ensemble of 80 structures selected using ASTEROIDS. The four boxes show the PRE data for simulated spin labels at residues 20 (top left), 40 (top right), 60 (bottom left) and 80 (bottom right). Lines show the PREs calculated from a control ensemble with no specific contacts. The distance matrix shows the chain proximity in the ensembles selected using ASTEROIDS (above the diagonal), compared to target ensembles (below the diagonal). Average distances between sites are shown in terms of: $\Delta_{ij} = \log(\langle d^0_{ij}\rangle / \langle d_{ij}\rangle)$ where d_{ij} is the distance in any given structure of the ASTEROIDS ensemble between sites i and j, and d^0_{ij} is the distance in any given structure of the reference ensemble between sites i and j. Values above the diagonal have been multiplied by 2 for ease of identification of the contact. Top: Black: RDCs calculated using the local alignment window (LAW). Blue: Predicted effect of the long-range contact detected using the ASTEROIDS interpretation of the PREs. Bottom: Combination (purple) of the two curves shown in the top panel and RDCs averaged over 100 000 full length conformers where each structure has a contact between 41–50 and 81–90 (black). (b) Blue: Data averaged over the target ensemble where each conformer has a contact between 41–50 and 81–90. Red: Average PREs over an ensemble of 80 structures selected using ASTEROIDS. The four boxes show the PRE data for simulated spin labels at residues 20 (top left), 40 (top right), 60 (bottom left) and 80 (bottom right). Lines show the PREs calculated from a control ensemble with no specific contacts. The distance matrix shows the chain proximity in the ensembles selected using ASTEROIDS (above the diagonal), compared to target ensembles (below the diagonal). Values above the diagonal have been multiplied by 2 for ease of identification of the contact. Top: Black: RDCs calculated using the local alignment window (LAW). Blue: Predicted effect of the long-range contact detected using the ASTEROIDS interpretation of the PREs. Bottom: Combination (purple) of the two curves shown in the top panel and RDCs averaged over 100 000 full length conformers where each structure has a contact between 11–20 and 61–70 (black). Reprinted with permission from the *Journal of the American Chemical Society*.[93]

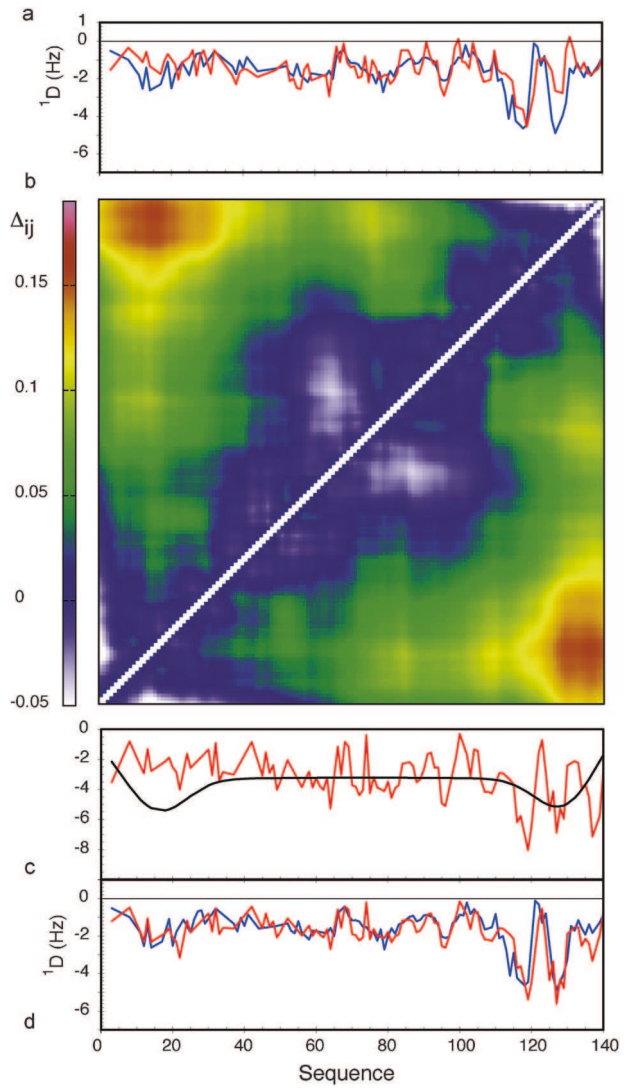


Fig. 4 Combined analysis of PREs and RDCs in the context of experimental data from α-synuclein. (a) Comparison of experimental $^1D_{NH}$ RDCs with couplings calculated using a standard *flexible-meccano* prediction (red). The rmsd between the two distributions is 0.78 Hz. (b) Contact map showing the relative proximity of different parts of the chain in α-synuclein, derived from experimental PRE data. Average distances between sites are shown in terms of: $\Delta_{ij} = \log(\langle d_{ij}\rangle\langle d_{ij}^0\rangle)$ where d_{ij} is the distance in any given structure of the ASTEROIDS ensemble between sites *i* and *j*, and d_{ij}^0 is the distance in any given structure of the reference ensemble between sites *i* and *j*. (c) LAW-predicted RDCs (red) and effective baseline derived from the distance matrix shown in (b). (d) Combination of the curves shown in (c) (red) in comparison to the experimental $^1D_{NH}$ RDCs (rmsd = 0.52 Hz). Reprinted with permission from the *Journal of the American Chemical Society*.[93]

ensembles that evoked only 20 structures reproduced the experimental data, but critically did not reproduce the backbone dihedral angle distributions that were at the origin of the average. Only when at least 200 structures were used in the average was the conformational behaviour sufficiently well reproduced. The reason for this is the instability of adding additional RDCs to an ensemble where the average is not yet converged.

The revelation that experimental data can be reproduced by an ensemble of structures that does not represent the correct

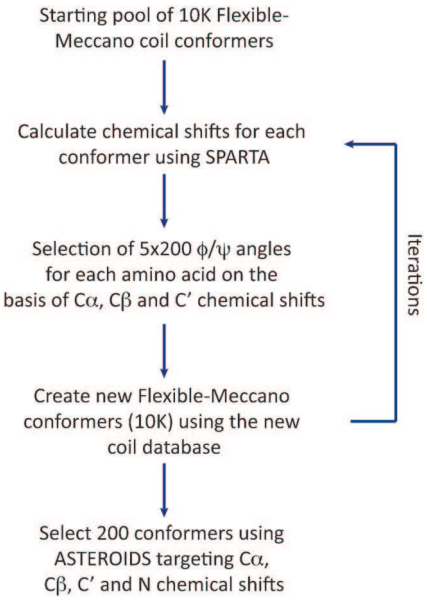


Fig. 5 Flowchart showing the iterative construction of a conformational ensemble using ASTEROIDS on the basis of heteronuclear chemical shifts.

conformational sampling was initially surprising to us, although this appears to be a predictable manifestation of the potential pit-falls of deriving ensembles under such underdetermined conditions. The result has particular importance, and highlights the risks of reducing the number of members of a conformational average until the data are reproduced. Such a procedure can clearly produce ensembles whose local conformational sampling is quantitatively incorrect, while reproducing experimental data.

Secondly, and possibly more critically, approaches that only use a LAW to analyze RDCs patently ignore the fact that RDCs are affected both by the local conformational sampling and long-range order. This is important even in the absence of specific long-range contacts, because the chain-like nature of the unfolded protein induces an effective baseline reflecting the increasing degrees of freedom available towards the ends of the chain (Fig. 2a and h). Long-range information is necessarily absent from an approach that only employs LAWs to predict the RDCs. If this approach is employed the simulated data need to be corrected for the effects of the unfolded chain. This can be achieved when LAW-predicted RDCs are multiplied by the expected baseline of an unfolded chain, whose bell-shaped dependence can be parameterised by fitting to numerical simulation.

The effects of ignoring long-range contacts when analyzing RDCs from disordered chains can however be much more severe when preferential long-range contacts exist in the protein, as demonstrated by the following simulations: RDCs were predicted from 100 000 strong ensembles using the *flexible-meccano* simulations of a 100 amino acid model sequence in the presence of weakly defined long-range contacts, defined as a contact between any of two 20 amino acid strands (Fig. 2). In comparison to the expected values for a chain with no specific long-range contacts, the effect is significant, even for such diffuse long-range contacts. Simulation predicts significant quenching

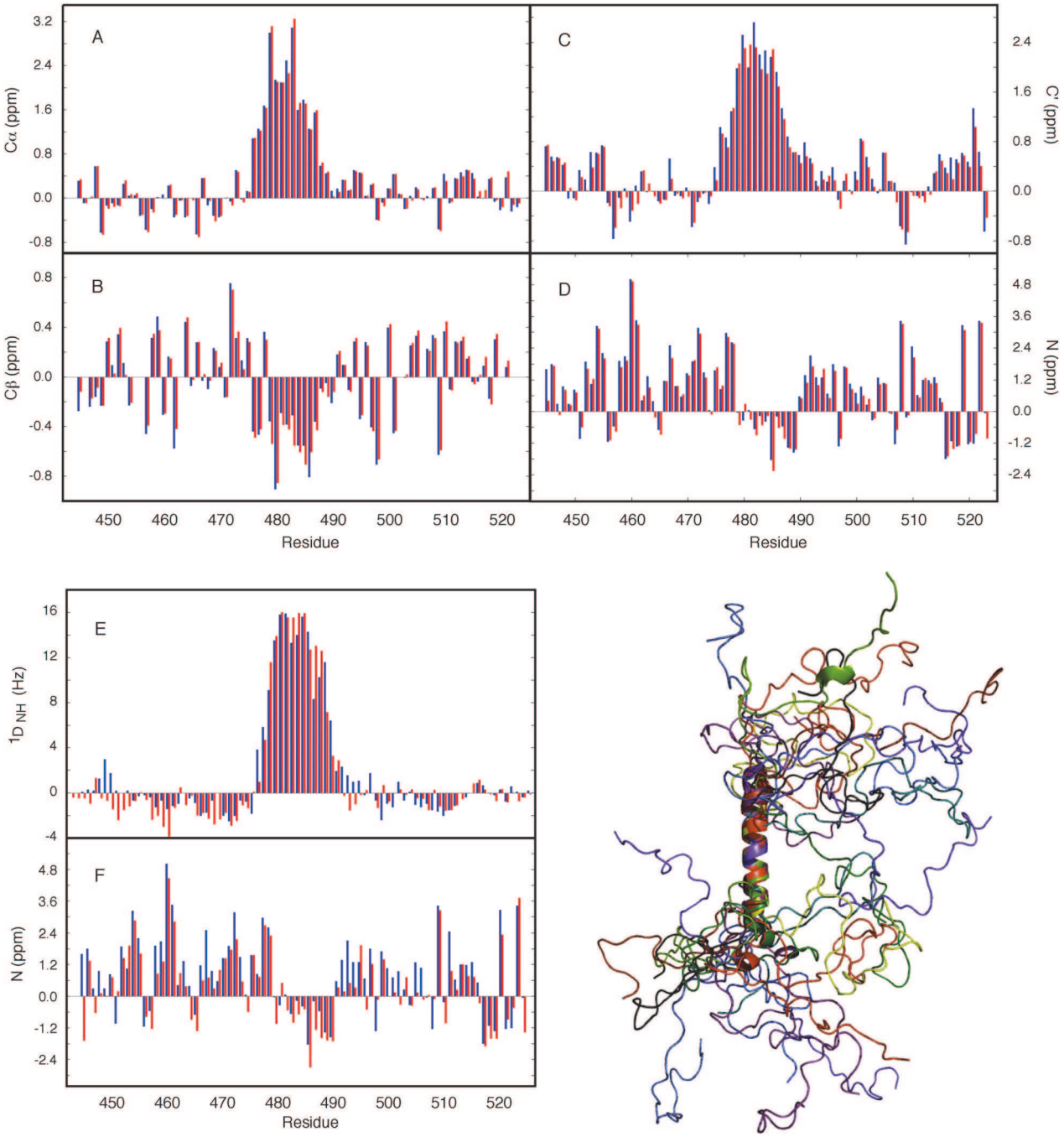


Fig. 6 Application of ASTEROIDS to ensemble representation on the basis of chemical shifts. Secondary chemical shifts from an ensemble of 200 structures determined using the ASTEROIDS algorithm compared to experimental secondary chemical shifts (blue). Red: secondary chemical shifts averaged over the final ensemble. (A) α carbon, (B) β carbon, (C) carbonyl, (D) amide nitrogen. (E, F) Reproduction of independent parameters by the ensemble based on chemical shift selection. (E) ^{15}N–^{1}H residual dipolar couplings (RDCs) measured in sterically aligned N_{TAIL} compared to averages over 50 000 conformers calculated using the amino acid specific description of N_{TAIL} determined from the chemical shifts. Simulated data (red) were scaled uniformly to best match experiment (blue). (F) Reproduction of ^{15}N secondary chemical shifts (blue: experiment, red: simulation), calculated using an ensemble determined from only ^{13}C shifts. Reprinted with permission from the *Journal of the American Chemical Society*.[97]

of RDC values in regions between the two contact regions. Importantly, although the local conformational sampling is not measurably affected by the contacts, the resulting RDCs are very different because of the transient long-range order that is also present. This again demonstrates that extreme caution needs to be exercised when interpreting RDCs uniquely in terms of the local structure. Comparison with identical simulations for a poly-valine indicates that the actual effect of diffuse long-range contacts is to convolute a more complex 'baseline' on the local structure of the expected RDCs. Fortunately a generic mathematical expression that accurately models the form of this baseline can be derived that reproduces the numerically predicted baselines shown in Fig. 2, which depends only on the position of the contacts and the length of the chain.

The consequences of this are that long-range information, for example derived from paramagnetic relaxation enhancement (*vide infra*), can be combined with the efficient sliding window approach, to simultaneously account for both aspects within the same ensemble average (Fig. 3).[93]

Combining RDCs and PREs in a single conformational ensemble

Similar analyses were applied to the interpretation of paramagnetic relaxation enhancements in disordered systems. We again use *flexible-meccano*, in combination with ASTEROIDS, to model intermolecular contacts giving rise to experimental PRE in disordered proteins. One important result demonstrates that even in the presence of highly diffuse, ill-defined target interactions, explicit modelling of spin label mobility significantly improves reproduction of conformational sampling, both for experimental and simulated data. We find that intermolecular contacts can be identified using 4 strategically placed spin labels in a 100 amino acid protein (Fig. 3) (and that two contacts can be identified using 8 spin labels in a 200 amino acid protein). Of course the ability to detect the transient contacts, and more importantly to estimate their population, strongly depends on the number of cysteine mutants that are available for the study.[71] Using cross validation of an entire data set that is not used in the analysis, we are also able to determine the appropriate number of structures necessary to define the system.

The ability to combine long-range information from PREs and RDCs in this way represents a major step forward in our ability to describe highly disordered systems. As an example, we applied these methods to experimental PRE and RDC data from α-Synuclein.[57,96] Experimentally measured RDCs agree significantly better when a long-range contact between the N and C terminal domains, derived from PREs, is included in the RDC analysis (rmsd of 0.51 compared to 0.75). This not only validates the predicted effects on RDC profiles due to long-range transient contacts in disordered systems, but also demonstrates that PREs and RDCs can be meaningfully combined to understand experimental data (Fig. 4).

Defining conformational ensembles of IDPs from chemical shifts

Finally we have applied the *flexible-meccano*/ASTEROIDS combination to explore the possibility of using chemical shifts alone to map local backbone conformational sampling of intrinsically disordered and partially folded proteins (Fig. 5).[97] $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ and ^{15}N chemical shifts have different backbone ϕ/ψ dihedral angle dependences that are complementary in terms of the mapping of different regions of the Ramachandran space. $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ secondary shifts report essentially on the Ramachandran space sampled by the observed amino acid, while both $^{13}C'$ and ^{15}N are also sensitive to the sampling properties of the neighbouring amino acids. ASTEROIDS is used to select a 200-strong sub-ensemble out of a larger pool (typically 10 000 structures) constructed by *flexible-meccano* that is in agreement with the experimental $^{13}C\alpha$, $^{13}C\beta$, $^{13}C'$ and ^{15}N chemical shifts (Fig. 6). The program SPARTA[98] is used to calculate chemical shifts for each member of the ensemble. No assumptions are made about the secondary structural propensity, with the first ensemble containing only unfolded structures derived from the statistical coil database. The local conformational bias is identified automatically on the basis of chemical shifts, and the resulting propensities are then used to assemble the new database for the next iteration. The algorithm thus automatically resolves the backbone dihedral angle distributions for the construction of entire secondary structural elements, as well as identifying local conformational sampling in unfolded domains. The analysis was applied to the study of N_{TAIL}, the C-terminal domain of the Sendai virus nucleoprotein, which contains a conformationally fluctuating helical element at its centre. Excellent agreement with experimental shifts is observed throughout the protein. Here again we are able to cross-validate the conformational description against independent data sets ($^{1}D_{NH}$ dipolar couplings or ^{15}N chemical shifts) to demonstrate both the accuracy of the description and the predictive power of the approach (Fig. 6). Although the conformational information is not as rich as that provided by RDCs, this approach raises the exciting prospect of probing the conformational behaviour of disordered proteins under more demanding conditions where additional parameters cannot be easily measured, for example when studying IDPs *in situ*.[86]

Conclusions

In order to understand the conformational behaviour of IDPs, a molecular representation of the partially folded state is required. We have developed ensemble approaches that characterize the disordered state, initially comparing free statistical coil simulations with measured data in order to understand expected random coil values of the different experimental parameters. Deviations from expected values allowed us to identify the presence of secondary structural propensity in a number of IDPs. We have then developed an ensemble description approach, initially for the study of helical elements in viral proteins from Sendai and Measles, then applied more generally for any disordered system. This problem is highly underdetermined, and each experimental NMR parameter requires specific consideration of the relevant averaging properties of the physical interaction responsible for the experimental observable, before valid parameter ranges and procedures can be established. The resulting approach, ASTEROIDS, can now be used to combine different sources of experimental NMR data, for example RDCs, PREs and chemical shifts, to define the conformational behaviour of the protein, and hopefully to follow the changes in conformational equilibrium that accompany physiologically relevant interactions.

References

1 V. N. Uversky, *Protein Sci.*, 2002, **11**, 739–756.
2 A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva and Z. Obradović, *Biochemistry*, 2002, **41**, 6573–6582.
3 P. Tompa, *Trends Biochem. Sci.*, 2002, **27**, 527–533.
4 H. J. Dyson and P. E. Wright, *Curr. Opin. Struct. Biol.*, 2002, **12**, 54–60.
5 A. K. Dunker and V. N. Uversky, *Curr. Opin. Pharmacol.*, 2010, **10**, 782–788.
6 M. M. Babu, R. van der Lee, N. S. de Groot and J. Gsponer, *Curr. Opin. Struct. Biol.*, 2011, **21**, 432–440.

7 M. Vendruscolo and C. M. Dobson, *Nature*, 2007, **449**, 555.
8 T. Mittag and J. D. Forman-Kay, *Curr. Opin. Struct. Biol.*, 2007, **17**, 3–14.
9 D. Eliezer, *Curr. Opin. Struct. Biol.*, 2009, **19**, 23–30.
10 M. R. Jensen, P. R. L. Markwick, S. Meier, C. Griesinger, M. Zweckstetter, S. Grzesiek, P. Bernadó and M. Blackledge, *Structure*, 2009, **17**, 1169–1185.
11 C. K. Fisher and C. M. Stultz, *Curr. Opin. Struct. Biol*, 2011, **21**, 426–431.
12 H. J. Dyson and P. E. Wright, *Nat. Rev. Mol. Cell Biol.*, 2005, **6**, 197–208.
13 M. D. Mukrasch, S. Bibow, J. Korukottu, S. Jeganathan, J. Biernat, C. Griesinger, E. Mandelkow and M. Zweckstetter, *PLoS Biol.*, 2009, **7**, e34.
14 K. Sugase, H. J. Dyson and P. E. Wright, *Nature*, 2007, **447**, 1021–1025.
15 V. N. Uversky, *Chem. Soc. Rev.*, 2011, **40**, 1623–1634.
16 B. A. Shoemaker, J. J. Portman and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 8868–8873.
17 S. Meier, M. Blackledge and S. Grzesiek, *J. Chem. Phys.*, 2008, **128**, 052204.
18 P. E. Wright and H. J. Dyson, *Curr. Opin. Struct. Biol.*, 2009, **19**, 31–38.
19 P. Tompa, *Curr. Opin. Struct. Biol.*, 2011, **21**, 419–425.
20 D. S. Wishart and B. D. Sykes, *J. Biomol. NMR*, 1994, **4**, 171–180.
21 H. Zhang, S. Neal and D. S. Wishart, *J. Biomol. NMR*, 2003, **25**, 173–195.
22 J. A. Marsh, V. K. Singh, Z. Jia and J. D. Forman-Kay, *Protein Sci.*, 2006, **15**, 2795–2804.
23 S. Schwarzinger, G. J. Kroon, T. R. Foss, J. Chung, P. E. Wright and H. J. Dyson, *J. Am. Chem. Soc.*, 2001, **123**, 2970–2978.
24 Y. Wang and O. Jardetzky, *J. Am. Chem. Soc.*, 2002, **124**, 14075–14084.
25 W. Peti, L. J. Smith, C. Redfield and H. Schwalbe, *J. Biomol. NMR*, 2001, **19**, 153–165.
26 J. Yao, J. Chung, D. Eliezer, P. E. Wright and H. J. Dyson, *Biochemistry*, 2001, **40**, 3561–3571.
27 M. Kjaergaard, K. Teilum and F. M. Poulsen, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 12535–12540.
28 K. Modig, V. W. Jürgensen, K. Lindorff-Larsen, W. Fieber, H. G. Bohr and F. M. Poulsen, *FEBS Lett.*, 2007, **581**, 4965–4971.
29 L. J. Smith, K. A. Bolin, H. Schwalbe, M. W. MacArthur, J. M. Thornton and C. M. Dobson, *J. Mol. Biol.*, 1996, **255**, 494–506.
30 H. Schwalbe, K. M. Fiebig, M. Buck, J. A. Jones, S. B. Grimshaw, A. Spencer, S. J. Glaser, L. J. Smith and C. M. Dobson, *Biochemistry*, 1997, **36**, 8977–8991.
31 L. Serrano, *J. Mol. Biol.*, 1995, **254**, 322–333.
32 D. Shortle and M. S. Ackerman, *Science*, 2001, **293**, 487–489.
33 R. Mohana-Borges, N. K. Goto, G. J. A. Kroon, H. J. Dyson and P. E. Wright, *J. Mol. Biol.*, 2004, **340**, 1131–1142.
34 W. Fieber, S. Kristjansdottir and F. M. Poulsen, *J. Mol. Biol.*, 2004, **339**, 1191–1199.
35 P. Bernadó, L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok and M. Blackledge, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 17002–17007.
36 A. K. Jha, A. Colubri, K. F. Freed and T. R. Sosnick, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 13099–13104.
37 N. Tjandra and A. Bax, *Science*, 1997, **278**, 1111–1114.
38 M. Blackledge, *Prog. Nucl. Magn. Reson. Spectrosc.*, 2005, **46**, 23–61.
39 J. Tolman and K. Ruan, *Chem. Rev.*, 2006, **106**, 1720–1736.
40 J. Meiler, J. J. Prompers, W. Peti, C. Griesinger and R. Brüschweiler, *J. Am. Chem. Soc.*, 2001, **123**, 6098–6107.
41 S. A. Showalter and R. Brüschweiler, *J. Am. Chem. Soc.*, 2007, **129**, 4158–4159.
42 O. F. Lange, N.-A. Lakomek, C. Farès, G. F. Schröder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger and B. L. de Groot, *Science*, 2008, **320**, 1471–1475.
43 L. Salmon, G. Bouvignies, P. Markwick, N. Lakomek, S. Showalter, D.-W. Li, K. Walter, C. Griesinger, R. Brüschweiler and M. Blackledge, *Angew. Chem., Int. Ed.*, 2009, **48**, 4154–4157.
44 P. R. L. Markwick, G. Bouvignies, L. Salmon, J. A. McCammon, M. Nilges and M. Blackledge, *J. Am. Chem. Soc.*, 2009, **131**, 16968–16975.
45 M. Louhivuori, K. Pääkkönen, K. Fredriksson, P. Permi, J. Lounila and A. Annila, *J. Am. Chem. Soc.*, 2003, **125**, 15647–15650.
46 K. Fredriksson, M. Louhivuori, P. Permi and A. Annila, *J. Am. Chem. Soc.*, 2004, **126**, 12646–12650.
47 O. I. Obolensky, K. Schlepckow, H. Schwalbe and A. V. Solov'yov, *J. Biomol. NMR*, 2007, **39**, 1–16.
48 J. L. Battiste and G. Wagner, *Biochemistry*, 2000, **39**, 5355–5365.
49 J. R. Gillespie and D. Shortle, *J. Mol. Biol.*, 1997, **268**, 170–184.
50 G. M. Clore, C. Tang and J. Iwahara, *Curr. Opin. Struct. Biol.*, 2007, **17**, 603–616.
51 A. N. Volkov, J. A. R. Worrall, E. Holtzmann and M. Ubbink, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 18945–18950.
52 C. Tang, J. Iwahara and G. M. Clore, *Nature*, 2006, **444**, 383–386.
53 G. M. Clore and J. Iwahara, *Chem. Rev.*, 2009, **109**, 4108–4139.
54 J. Iwahara and G. M. Clore, *Nature*, 2006, **440**, 1227–1230.
55 S. Kristjansdottir, K. Lindorff-Larsen, W. Fieber, C. M. Dobson, M. Vendruscolo and F. M. Poulsen, *J. Mol. Biol.*, 2005, **347**, 1053–1062.
56 K. Lindorff-Larsen, S. Kristjansdottir, K. Teilum, W. Fieber, C. Dobson, F. Poulsen and M. Vendruscolo, *J. Am. Chem. Soc.*, 2004, **126**, 3291–3299.
57 C. W. Bertoncini, Y.-S. Jung, C. O. Fernandez, W. Hoyer, C. Griesinger, T. M. Jovin and M. Zweckstetter, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 1430–1435.
58 D. K. Wilkins, S. B. Grimshaw, V. Receveur, C. M. Dobson, J. A. Jones and L. J. Smith, *Biochemistry*, 1999, **38**, 16424–16431.
59 B. Brutscher, R. Brüschweiler and R. R. Ernst, *Biochemistry*, 1997, **36**, 13043–13053.
60 J. Klein-Seetharaman, M. Oikawa, S. B. Grimshaw, J. Wirmer, E. Duchardt, T. Ueda, T. Imoto, L. J. Smith, C. M. Dobson and H. Schwalbe, *Science*, 2002, **295**, 1719–1722.
61 S. Schwarzinger, P. E. Wright and H. J. Dyson, *Biochemistry*, 2002, **41**, 12681–12686.
62 W.-Y. Choy, F. A. A. Mulder, K. A. Crowhurst, D. R. Muhandiram, I. S. Millett, S. Doniach, J. D. Forman-Kay and L. E. Kay, *J. Mol. Biol.*, 2002, **316**, 101–112.
63 J. E. Kohn, I. S. Millett, J. Jacob, B. Zagrovic, T. M. Dillon, N. Cingel, R. S. Dothager, S. Seifert, P. Thiyagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pande, I. Ruczinski, S. Doniach and K. W. Plaxco, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 12491–12496.
64 I. S. Millett, S. Doniach and K. W. Plaxco, *Adv. Protein Chem.*, 2002, **62**, 241–262.
65 P. Bernadó and M. Blackledge, *Biophys. J.*, 2009, **97**, 2839–2845.
66 J. Lipfert and S. Doniach, *Annu. Rev. Biophys. Biomol. Struct.*, 2007, **36**, 307–327.
67 E. Mylonas, A. Hascher, P. Bernado, M. Blackledge, E. Mandelkow and D. Svergun, *Biochemistry*, 2008, **47**, 10345–10353.
68 P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge and D. I. Svergun, *J. Am. Chem. Soc.*, 2007, **129**, 5656–5664.
69 A. M. Bonvin, J. A. Rullmann, R. M. Lamerichs, R. Boelens and R. Kaptein, *Proteins*, 1993, **15**, 385–400.
70 J. Gsponer, H. Hopearuoho, S. B.-M. Whittaker, G. R. Spence, G. R. Moore, E. Paci, S. E. Radford and M. Vendruscolo, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 99–104.
71 D. Ganguly and J. Chen, *J. Mol. Biol.*, 2009, **390**, 467–477.
72 J. R. Allison, P. Varnai, C. M. Dobson and M. Vendruscolo, *J. Am. Chem. Soc.*, 2009, **131**, 18314–18326.
73 K.-P. Wu, D. S. Weinstock, C. Narayanan, R. M. Levy and J. Baum, *J. Mol. Biol.*, 2009, **391**, 784–796.
74 S. Esteban-Martín, R. B. Fenwick and X. Salvatella, *J. Am. Chem. Soc.*, 2010, **132**, 4626–4632.
75 D. Hamelberg, J. Mongan and J. A. McCammon, *J. Chem. Phys.*, 2004, **120**, 11919–11929.
76 P. R. L. Markwick, G. Bouvignies and M. Blackledge, *J. Am. Chem. Soc.*, 2007, **129**, 4724–4730.
77 N. C. Fitzkee, P. J. Fleming and G. D. Rose, *Proteins*, 2005, **58**, 852–854.
78 J. F. Leszczynski and G. D. Rose, *Science*, 1986, **234**, 849–855.
79 A. K. Jha, A. Colubri, M. H. Zaman, S. Koide, T. R. Sosnick and K. F. Freed, *Biochemistry*, 2005, **44**, 9691–9702.
80 S. C. Lovell, I. W. Davis, W. B. Arendall 3rd, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson and D. C. Richardson, *Proteins*, 2003, **50**, 437–450.

81 M. Zweckstetter and A. Bax, *J. Am. Chem. Soc.*, 2000, **122**, 3791–3792.
82 M. D. Mukrasch, P. Markwick, J. Biernat, M. von Bergen, P. Bernadó, C. Griesinger, E. Mandelkow, M. Zweckstetter and M. Blackledge, *J. Am. Chem. Soc.*, 2007, **129**, 5235–5243.
83 M. Wells, H. Tidow, T. Rutherford, P. Markwick, M. Jensen, E. Mylonas, D. Svergun, M. Blackledge and A. Fersht, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 5762–5767.
84 M. R. Jensen and M. Blackledge, *J. Am. Chem. Soc.*, 2008, **130**, 11266–11267.
85 M. R. Jensen, K. Houben, E. Lescop, L. Blanchard, R. W. H. Ruigrok and M. Blackledge, *J. Am. Chem. Soc.*, 2008, **130**, 8055–8061.
86 M. R. Jensen, G. Communie, E. A. Ribeiro Jr, N. Martinez, A. Desfosses, L. Salmon, L. Mollica, F. Gabel, M. Jamin, S. Longhi, R. W. H. Ruigrok and M. Blackledge, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 9839–9844.
87 S. Meier, M. Strohmeier, M. Blackledge and S. Grzesiek, *J. Am. Chem. Soc.*, 2007, **129**, 754–755.
88 S. Meier, S. Grzesiek and M. Blackledge, *J. Am. Chem. Soc.*, 2007, **129**, 9799–9807.
89 F. Gabel, M. R. Jensen, G. Zaccaï and M. Blackledge, *J. Am. Chem. Soc.*, 2009, **131**, 8769–8771.
90 J. A. Marsh and J. D. Forman-Kay, *J. Mol. Biol.*, 2009, **391**, 359–374.
91 C. K. Fisher, A. Huang and C. M. Stultz, *J. Am. Chem. Soc.*, 2010, **132**, 14919–14927.
92 G. Nodet, L. Salmon, V. Ozenne, S. Meier, M. R. Jensen and M. Blackledge, *J. Am. Chem. Soc.*, 2009, **131**, 17908–17918.
93 L. Salmon, G. Nodet, V. Ozenne, G. Yin, M. R. Jensen, M. Zweckstetter and M. Blackledge, *J. Am. Chem. Soc.*, 2010, **132**, 8407–8418.
94 M. R. Jensen, P. Bernadó, K. Houben, L. Blanchard, D. Marion, R. W. H. Ruigrok and M. Blackledge, *Protein Pept. Lett.*, 2010, **17**, 952–960.
95 J. A. Marsh, J. M. R. Baker, M. Tollinger and J. D. Forman-Kay, *J. Am. Chem. Soc.*, 2008, **130**, 7804–7805.
96 P. Bernadó, C. W. Bertoncini, C. Griesinger, M. Zweckstetter and M. Blackledge, *J. Am. Chem. Soc.*, 2005, **127**, 17968–17969.
97 M. R. Jensen, L. Salmon, G. Nodet and M. Blackledge, *J. Am. Chem. Soc.*, 2010, **132**, 1270–1272.
98 Y. Shen and A. Bax, *J. Biomol. NMR*, 2007, **38**, 289–302.

BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 11 2012, pages 1463–1470
doi:10.1093/bioinformatics/bts172

Structural bioinformatics

Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables

Valéry Ozenne[1], Frédéric Bauer[2], Loïc Salmon[1], Jie-rong Huang[1], Malene Ringkjøbing Jensen[1], Stéphane Segard[2], Pau Bernadó[3], Céline Charavay[2] and Martin Blackledge*[,1]

[1]Protein Dynamics and Flexibility, Institut de Biologie Structurale Jean-Pierre Ebel, CEA; CNRS; UJF UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, [2]Groupe Informatique pour les Scientifiques du Sud Est (GIPSE), IRTSV / Laboratoire Biologie à Grande Echelle, CEA - INSERM U1038 - UJF, 17 avenue des Martyrs, 38054 Grenoble Cedex 9 and [3]Centre de Biochimie Structurale, CNRS UMR 5048 - UM 1 - INSERM UMR 1054, 34090, Montpelier, France

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Intrinsically disordered proteins (IDPs) represent a significant fraction of the human proteome. The classical structure function paradigm that has successfully underpinned our understanding of molecular biology breaks down when considering proteins that have no stable tertiary structure in their functional form. One convenient approach is to describe the protein in terms of an equilibrium of rapidly inter-converting conformers. Currently, tools to generate such ensemble descriptions are extremely rare, and poorly adapted to the prediction of experimental data.

Results: We present *flexible-meccano*—a highly efficient algorithm that generates ensembles of molecules, on the basis of amino acid-specific conformational potentials and volume exclusion. Conformational sampling depends uniquely on the primary sequence, with the possibility of introducing additional local or long-range conformational propensities at an amino acid-specific resolution. The algorithm can also be used to calculate expected values of experimental parameters measured at atomic or molecular resolution, such as nuclear magnetic resonance (NMR) and small angle scattering, respectively. We envisage that *flexible-meccano* will be useful for researchers who wish to compare experimental data with those expected from a fully disordered protein, researchers who see experimental evidence of deviation from 'random coil' behaviour in their protein, or researchers who are interested in working with a broad ensemble of conformers representing the flexibility of the IDP of interest.

Availability: A fully documented multi-platform executable is provided, with examples, at http://www.ibs.fr/science-213/scientific-output/software/flexible-meccano/

Contact: martin.blackledge@ibs.fr

Received on February 29, 2012; revised on March 23, 2012; accepted on April 2, 2012

1 INTRODUCTION

The realization that a significan percentage of the functional proteins encoded in eukaryotic genomes are fully or partially disordered in their functional state has revolutionized our understanding of structural and molecular biology (Babu, 2012; Dunker *et al*., 2002; Dyson and Wright, 2005; Tompa, 2002; Uversky, 2002). Intrinsically disordered proteins (IDPs), or proteins containing long intrinsically disordered regions, do not adopt a stable 3D fold, and therefore fall beyond the scope of classical structural biology. IDPs are biologically functional in the disordered state imposing a very different perspective on the relationship between primary protein sequence and function compared with the standard structure/function relationship that underpins our understanding of molecular biology. IDPs are implicated in a large number of human pathologies, and the development of pharmacological solutions to these problems awaits a molecular description of the role of flexibilit in a number of diseases (Babu *et al*., 2011; Dunker and Uversky, 2010; Vendruscolo and Dobson, 2007).

In order to understand the conformational behaviour of IDPs it is essential to develop a molecular description of the disordered state. The structural biology paradigm shifts when we consider disordered proteins, so that the determination of a single structure has no real physical relevance, or at best can only describe isolated sub-states on a vast potential energy landscape. A structural description of IDPs rather aims to determine rules that defin the behaviour of the flexibl protein in terms of probabilities of populating different regions of conformational space, and to correlate these probabilities with the function of the protein. The description of conformational propensities can be conveniently achieved by evoking an explicit ensemble description of inter-converting structures in equilibrium.

Due to the very large number of degrees of freedom available to such a disordered system, the problem of definin conformational space is highly underdetermined, requiring extensive experimental data to delimit the structural propensities of a given protein. Novel analytical tools are required to exploit the specifi conformational sensitivity of different experimental parameters. Each experimental NMR parameter for example, is sensitive to different aspects of the structural and dynamic behaviour of the disordered state and requires specifi consideration of the relevant averaging properties of the physical interaction (Schneider *et al*., 2012).

*To whom correspondence should be addressed.

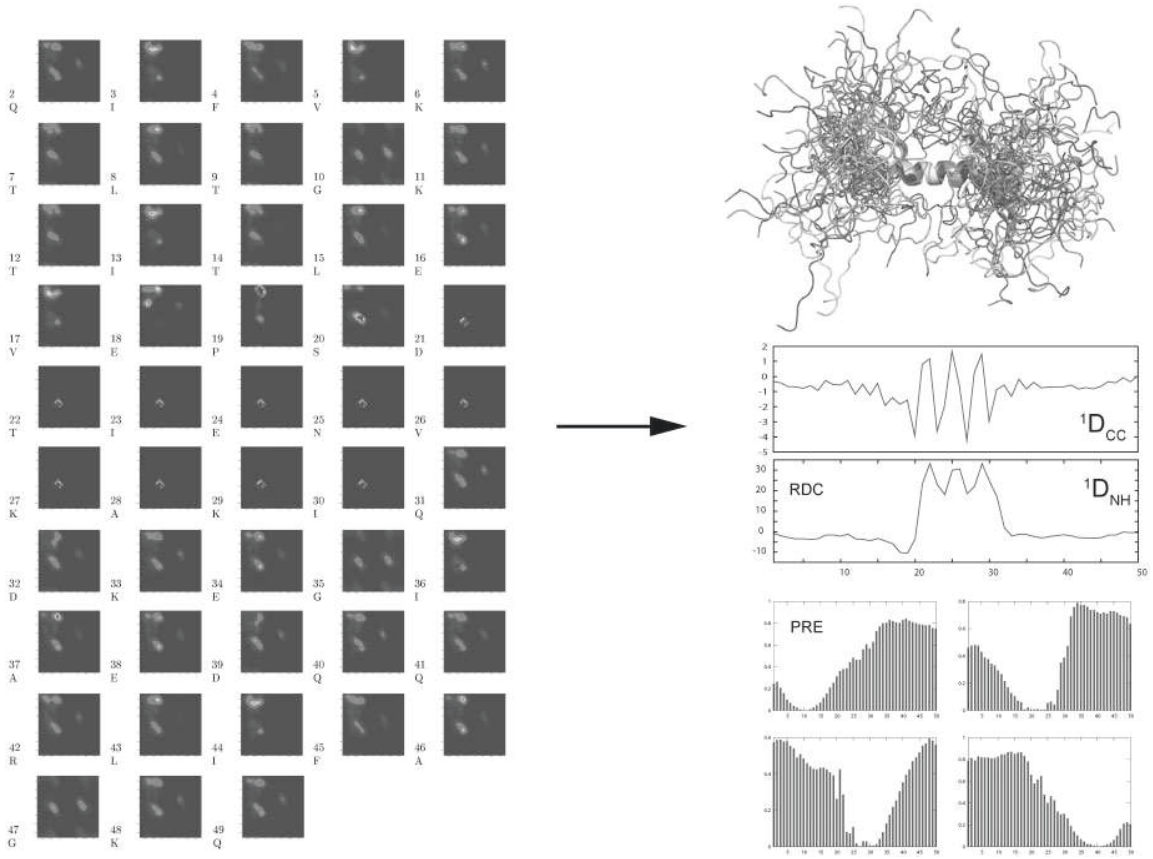


Fig. 1. Schematic representation of the *flexible-meccano* algorithm. Conformational sampling details are define by the user (shown figurativel on the left in terms of Ramachandran sampling for each amino acid)—the default is the *flexible-meccano* statistical coil description. This can be modifie in terms of additional propensities to form secondary structure, by modifying the amino acid-specifi potentials or by introducing long-range contacts. Biophysical parameters (radius of gyration and ϕ/ψ sampling), and expected experimental parameters (shown figurativel on the right, for example NMR RDCs, PREs etc.) are calculated for this conformational regime, as well as explicit conformational ensemble of backbone conformers.

2 SYSTEM AND METHODS

The protocol that we present here describes the generation of explicit ensemble descriptions of proteins, using the program *flexible-meccano* that has been explicitly developed to describe the behaviour of IDPs and denatured proteins. Expected experimental parameters, such as NMR and small angle scattering data, are calculated from the conformational ensembles (Bernadó *et al*., 2005a, b; Jensen *et al*., 2008; Jensen *et al*., 2011; Mukrasch *et al*., 2007; Nodet *et al*., 2009; Salmon *et al*., 2010; Wells *et al*., 2008). These parameters can then be compared with experimental measurements. Amino acid-specifi statistical coil sampling is used to describe the unfolded state on the basis of the primary sequence with the possibility of introducing additional local conformational propensities, or long-range tertiary contacts. The *flexible-meccano* approach was exploited for the firs time to study the C-terminal domain of Sendai virus phosphoprotein, and has been gradually refined extended and tested on a number of different experimental systems, including the disordered proteins α-synuclein, Tau and p53, involved in Parkinson's and Alzheimer's diseases and human cancer, respectively. Currently, no alternative tools exist that can provide ensembles and ensemble-averaged parameters for user-define conformational sampling regimes. The program is interfaced to a fully interactive and robust graphical interface as described below.

3 ALGORITHM

3.1 Flexible-meccano ensemble generation

Flexible-meccano uses a highly efficien minimization algorithm to build multiple, different copies of the same polypeptide chain by randomly sampling amino acid-specifi backbone dihedral angle $\{\phi/\psi\}$ potential wells. The population-weighted amino acid-specifi potentials are derived from a compilation of non-secondary structural elements of high-resolution X-ray crystallographic protein structures. The peptide chain is constructed by using the selected $\{\phi/\psi\}$ pairs to sequentially connect peptide planes.

The algorithm is based on tools developed for Meccano and Dynamic Meccano approaches (Bouvignies *et al*., 2006a, b; Hus *et al*., 2001; Hus *et al*., 2008; Salmon *et al*., 2009). These algorithms were used to determine the average orientation of peptide planes, and their associated dynamics, in folded proteins on the basis of optimization of parameters definin the orientation of each place on

the basis of NMR residual dipolar couplings (RDCs). The difference in this case is that, rather than determining the orientation of the peptide plane on the basis of experimental data, the unique constraint used to orient each peptide unit is the backbone dihedral angle pair, that is randomly selected from the database potential for each amino acid. Each tetrahedral junction is constructed with optimal geometry.

The coordinates of the generic peptide plane were derived from high-resolution X-ray crystallographic structures (Salmon *et al*., 2009). Amino acid-specifi hard-spheres are used to avoid steric clashes, to provide an efficient but physically reasonable model of repulsive interatomic forces. No attractive forces are explicitly used. A total of 23 potential energy wells are sampled: one for each of the 20 different amino acid types, and specifi potentials accounting for the particular backbone conformational propensities of residues that precede prolines, prolines that precede prolines, and glycines that precede prolines. The simplicity of the model makes the structure ensemble generation highly efficien (100 000 structures of a 100-amino acid protein can be created in 30 min on a single processor, although the time increases with the number of experimental parameters that are simultaneously predicted). The complete absence of experimental constraints in this sampling phase avoids distortions due to additional potential energy terms such as those used in restrained MD calculations. Although this statistical coil model of the unfolded state has been tested with respect to its predictive power of diverse experimental parameters (for example RDCs and chemical shifts, and small angle scattering curves), it is simple for the user to replace the statistical coil potentials by an alternative description of the unfolded state.

Additional conformational propensities can be added to influenc the sampling of the protein in the following ways (Figure 1):

(1) Local conformational propensities can be modifie on an amino acid-specifi basis to include an additional potential, centred on a specifi Gaussian shaped region of backbone dihedral angle space $\{\phi_{\text{target}}/\psi_{\text{target}}\}$, of width $\{\Delta\phi_{\text{target}}/\Delta\psi_{\text{target}}\}$, populated with a propensity p_{res}. The widths of the additional Gaussian shaped potentials and their propensities can be set by the user, and are simply added to, or replace, the existing potential.

(2) Conformational propensities of regions of the primary sequence can be modifie to include the presence of continuous secondary structure—either α-helical, β-sheet or polyproline II, with propensity p_{sec}. Propensities can be introduced in a cooperative, or independent manner, for example a complete helix can be constructed for 20% of conformers, (cooperative) or 20% additional helical sampling can be introduced randomly in the same sequence (non-cooperative).

(3) Long-range contacts can be included in the conformational description by specifying that a certain percentage of structures (p_{dist}), must contain at least one C^{α} atom from an amino acid between residues i and j, that is closer than $d_{\max}$ away from a C^{α} atom in any amino acid between residues k and l.

The calculation can be used to generate explicit coordinates of each conformer in the ensemble, or simply to calculate appropriately averaged observables that would be expected in the presence of such a conformational regime (see below). At present the introduction of folded domains into the ensemble is restricted to the presence of individual secondary structural elements, or regions that can be uniquely encoded by their $\{\phi/\psi\}$ values.

3.2 Application to the prediction of experimental parameters

After construction of the conformational ensemble, expected experimental values can be calculated from this ensemble.

3.2.1 Residual dipolar couplings RDCs report on local conformational sampling of each amino acid in the sequence, and are exquisitely sensitive to the presence of even weakly populated secondary structural elements (Jensen and Blackledge, 2008; Jensen *et al*., 2009). The dipolar coupling of a given magnetic moment with any other magnetic moment in its surroundings is given by (Blackledge, 2005):

$$D_{IS} = -\frac{\gamma_I \gamma_S \mu_0 h}{16\pi^3 r_{IS}^3} \langle P_2 \cos(\theta_{IS}) \rangle \quad (1)$$

where γ is the gyromagnetic ratio for the two spins I and S, r_{IS} is the distance between the spins, μ_0 is the permeability of free space, and h is Planck's constant. In the fast exchange regime, the measured RDC reports on the arithmetic average over all conformations sampled up to the millisecond timescale. It has been shown that averaging of expected RDCs from each conformer in the ensemble using the expression in Equation (2) gives reasonable reproduction of the distribution of experimental RDCs measured in IDPs and denatured proteins. The RDCs from each internuclear vector IS are calculated from the orientation (θ, φ) with respect to the alignment tensor of each individual conformer (j) with axial and rhombic components (A_a, A_r) :

$$D_{IS}^{j} = -\frac{\gamma_I \gamma_S \hbar \mu_0}{8\pi^2 r_{IS}^3} \left[A_a (3\cos^2\theta - 1) + \frac{3}{2} A_r \sin^2\theta \cos(2\varphi) \right] \quad (2)$$

The alignment tensors are calculated using an in-house routine, based on previous published algorithms (Berlin *et al*., 2009, Zweckstetter and Bax, 2000). The average shown in Equation (1) is then approximated by the calculation of the mean of the RDCs over all structures in the ensemble:

$$D_{IS} = \left\langle D_{IS}^{j} \right\rangle \quad (3)$$

As discussed in previous publications, this average has highly unfavourable convergence characteristics (Nodet *et al*., 2009). As a rule of thumb, $n \times 1000$ structures are required in the ensemble, where n is the number of amino acids in the sequence, before convergence of the RDC of a specifi amino acid has been achieved.

It is possible to alleviate the convergence problem by calculating the alignment characteristics of uncorrelated 'local alignment windows' (LAWs) (Marsh *et al*., 2008), but this approach requires knowledge of the explicit modulation of the underlying baseline of the RDC profile The baseline profil normally exhibits a bell-shaped distribution of RDCs, however, this profil may be significantl modulated in the presence of persistent long-range contacts between regions of the chain that are distant in the primary sequence (Nodet *et al*., 2009; Salmon *et al*., 2010). When selecting ensembles of structures in agreement with experimental data, it is useful to average over a smaller number of structures, and in this case the combination

of LAWs and explicit modulation of the underlying baseline has been shown to provide a means of combining RDCs and paramagnetic relaxation enhancements (PREs) for ensemble selection (Salmon *et al.*, 2010). Here, we concentrate on the prediction of values of experimental parameters that would be expected under given conformational sampling regimes, so we have chosen to retain the explicit global molecular description for all parameters. The algorithm allows for the definitio of long-range contacts between different regions of the primary sequence at a given propensity (*vide supra*), alone, or in combination with given populations of secondary structural motifs. This allows the prediction of expected RDC profile even in the presence of complex levels of local and long-range structure.

3.2.2 Scalar $^3J_{NHH\alpha}$ couplings Scalar couplings between amide and alpha protons ($^3J_{NHH\alpha}$) report on the conformationally averaged ϕ dihedral angle (Pardi *et al.*, 1984; Ludvigsen *et al.*, 1991; Vuister and Bax, 1993). The following Karplus relationship is used to calculate the values for each conformer:

$$^3J_{\mathrm{NHH}\alpha}(\phi) = A\cos^2(\phi - 60^\circ) + B\cos(\phi - 60^\circ) + C \qquad (4)$$

which are then averaged over the ensemble. A, B and C have been optimized using coupling constants measured in several proteins of known structure and therefore provide a constraint on the distribution of ϕ angles in conformational ensembles of IDPs (Mukrasch *et al.*, 2007; Smith *et al.*, 1996).

3.2.3 Paramagnetic relaxation enhancements A coherent picture of the conformational behaviour of IDPs and partially folded proteins requires not only a mapping of local structure but also long-range order. Long-range interactions in IDPs are often transient in nature and their detection, therefore, requires a strong probe that is active over a few nanometers such as that provided by an unpaired electron. One of the most efficien ways of introducing an unpaired electron is by attaching a thiol reactive methanethiosulfonate (MTSL) spin label to the protein through a cysteine residue. The dipolar interaction between the unpaired electron and the protein nuclei induces PREs that strongly depend on the electron-nucleus distances. By introducing spin labels at several different positions in the protein, a mapping of long-range interactions in the disordered state becomes possible.

The transverse relaxation rate due to the presence of the unpaired electron can be expressed as follows (Abragam, 1994; Gillespie and Shortle, 1997):

$$\Gamma_{2,H} = \frac{1}{15}\left(\frac{\mu_0}{4\pi}\right)^2 \gamma_H^2 g_e^2 \mu_B^2 s_e (s_e+1)\{4J(0) + 3J(\omega_H)\} \qquad (5)$$

where g_e is the electron g-factor, γ_H is the gyromagnetic ratio of the observed nucleus (proton), s_e is the electron spin quantum number, ω_H is the proton frequency, μ_B is the Bohr magneton and μ_0 is the permittivity of free space. PREs can be calculated over structural ensembles by considering a fixe position of the MTSL side-chain (for example on the Cβ atom of the cysteine) and by invoking the spectral density function:

$$J(\omega) = \left\langle r_{H-e}^{-6} \right\rangle \left\{ \frac{\tau_c}{1+\omega^2\tau_c^2} \right\} \qquad (6)$$

where τ_c is the correlation time of the relaxation active interaction. This simple description does not account for the potentially high level of flexibilit of the spin label itself (the electron spin label is attached to the molecule via MTSL attached to a cysteine side-chain). In order to address this, in the *flexible-meccano* approach, MTSL conformations are built explicitly for each backbone conformer by randomly sampling available rotamers and retaining only conformations that are sterically allowed. $J(\omega)$ can be described using a model-free expression of the order parameter, comprising the orientational and distance-dependent components of the internal motion that both strongly depend on the motion of the spin label with respect to the observed nuclear spin (Bruschweiler *et al.*, 1992; Clore and Iwahara, 2009):

$$J(\omega) = \left\langle r_{H-e}^{-6} \right\rangle \left\{ \frac{S_{H-e}^2 \tau_c}{1+\omega^2\tau_c^2} + \frac{\left(1-S_{H-e}^2\right)\tau_e}{1+\omega^2\tau_e^2} \right\} \qquad (7)$$

where the order parameter S_{H-e}^2 describes the motion of the dipolar interaction vector, $\tau_c = \tau_r\tau_s/(\tau_r+\tau_s)$ is a function of τ_s and τ_r the electron spin and rotational correlation times, respectively, and $\tau_e = 1/(\tau_i^{-1}+\tau_r^{-1}+\tau_s^{-1})$ where τ_i is the local correlation time of the spin label. r_{H-e} is the instantaneous distance between the proton and the electron spins. Order parameters can be expressed in terms of radial and angular components:

$$S_{H-e}^2 = S_{\mathrm{ang}}^2 S_{\mathrm{rad}}^2 \qquad (8)$$

where:

$$S_{\mathrm{rad}}^2 = \left\langle r_{H-e}^{-6} \right\rangle^{-1} \left\langle r_{H-e}^{-3} \right\rangle^2 \qquad (9)$$

$$S_{\mathrm{ang}}^2 = \frac{4\pi}{5} \sum_{m=-2}^{2} \left| \left\langle Y_2^m\left(\Omega^{\mathrm{mol}}\right) \right\rangle \right|^2 \qquad (10)$$

Ω^{mol} describes the orientation of the interaction vector in the frame of each conformer. The above expressions are used to calculate the transverse relaxation rate for each backbone conformation produced with the *flexible-meccano* algorithm, and the effective relaxation rate for each amide proton is then averaged over all retained conformers:

$$\Gamma_2^{\mathrm{total}} = \frac{1}{n}\sum_{i=1}^{n} \Gamma_{2,i}^{\mathrm{fm}} \qquad (11)$$

PREs are often described in terms of the ratio between the intensity measured in a standard HSQC experiment in the presence of the reduced and oxidized forms of the MTSL label:

$$\frac{I}{I^0} = \frac{\Gamma_2^{\mathrm{red}} \exp\left(-\Gamma_2^{\mathrm{calc}}\tau_{\mathrm{mix}}\right)}{\Gamma_2^{\mathrm{red}} + \Gamma_2^{\mathrm{calc}}} \qquad (12)$$

Here Γ_2^{red} is the intrinsic relaxation rate of the amide proton and τ_{mix} is the mixing time during which relaxation occurs in the HSQC pulse sequence (typically 10 ms).

3.2.4 Chemical shift prediction Chemical shifts are the most accessible NMR parameter, and provide powerful probes of conformation, in particular of secondary structural propensity (Eliezer *et al.*, 2001; Kjaergaard *et al.*, 2011; Marsh *et al.*, 2006; Modig *et al.*, 2007; Schwarzinger *et al.*, 2001). Remarkable progress has been made in recent years in the prediction of chemical shifts from protein conformation (Wishart and Sykes, 1994), and

Fig. 2. Screenshot of the *flexible-meccano* program, showing the interface allowing the user to defin conformational propensities [in this case an α-helical propensity between residues 12 and 18 (top), and a long-range contact between positions 5–15 and 25–30 (bottom)].

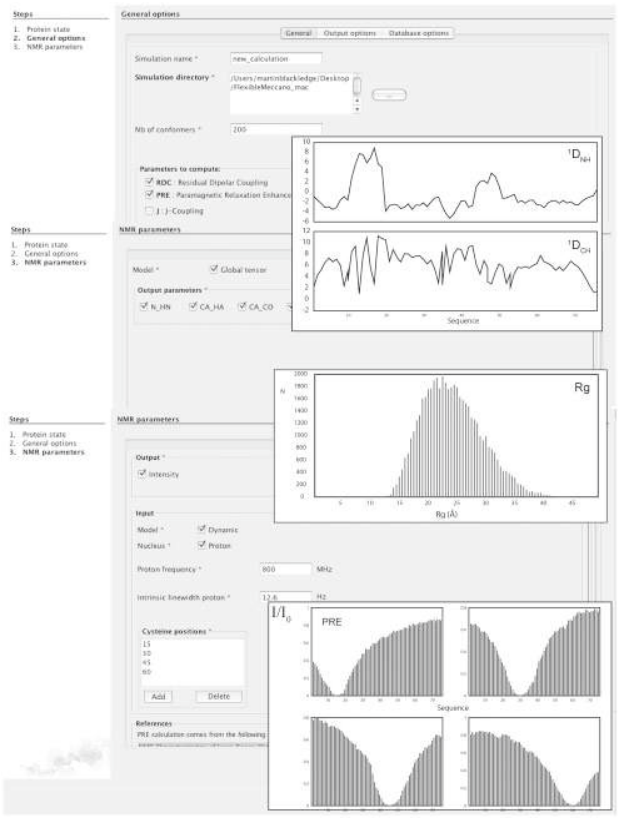


Fig. 3. Screenshot montage of the *flexible-meccano* program, showing the interface allowing the user to select which experimental and biophysical parameters to be calculated from the ensemble. As examples, $^1D_{NH}$ and $^1D_{C\alpha H\alpha}$ RDCs, a distribution of radius of gyration, and PRE profiles are shown in black, light grey and dark grey, respectively. This calculation corresponds to the top screen in Figure 2, with 50% of conformers containing a helical element form residue 12–18 (giving rise to positive $^1D_{NH}$ RDCs in this region).

vice-versa (Berjanskii *et al*., 2009; Cavalli *et al*., 2007; Shen *et al*., 2008). We have previously investigated the possibility of combining *flexible-meccano* with chemical shift prediction to analyze experimental chemical shifts in IDPs (Jensen *et al*., 2010). While $^{13}C\alpha$, $^{13}C\beta$, $^{13}C'$ and Hα chemical shifts depend strongly on ϕ and ψ, ^{15}N and $^1H^N$ chemical shifts show a more or less uniform dependence on the two dihedral angles. In addition, the $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts display an almost inverse dependence on the ϕ/ψ distribution and therefore report on the populations of α-helix and β-sheet in the disordered state. Scripts are provided with the *flexible-meccano* program that can be directly interfaced to well-known chemical shift prediction algorithms such as Sparta (Shen and Bax, 2007) and ShiftX (Han *et al*., 2011; Neal *et al*., 2003).

3.2.5 SAXS prediction Small angle X-ray scattering provides complementary information about the extent of conformational sampling of the unfolded protein. Scripts are again provided that allow the user to interface the *flexible-meccano* program with the SAXS program Crysol (Svergun *et al*., 1995) and calculate expected SAXS curves that would be associated with the given ensemble (Bernadó and Blackledge, 2010; Bernadó and Svergun, 2012; Bernadó *et al*., 2007).

3.2.6 Ramachandran segment division In order to describe the sampling of conformational space in the different ensembles, the *flexible-meccano* program provides a statistical analysis of the Ramachandran space sampled by the ensemble. In order to do this, ϕ,ψ space is divided into four quadrants as follows; α_L: $\{\phi>0°\}$, α_R : $\{\phi<0, -120°<\psi<50°\}$, β_P : $\{-100°<\phi<0°\}$, $\psi>50°$ or $\psi<-120°\}$, β_S : $\{-180°<\phi<-100°\}$, $\psi>50°$ or $\psi<-120°\}$. The population of these quadrants is indicated as $p_{\alpha L}$,$p_{\alpha R}$,$p_{\beta P}$ and $p_{\beta S}$.

4 IMPLEMENTATION

Examples of the implementation of *flexible-meccano* are shown in Figures 2 and 3. The following steps are shown:

(1) Reading of the primary sequence of the protein. This provides a scrollable table describing the conformational potentials to be used in the statistical sampling for each ϕ/ψ pair. The default potentials are those provided by the *flexible-meccano* statistical coil library.

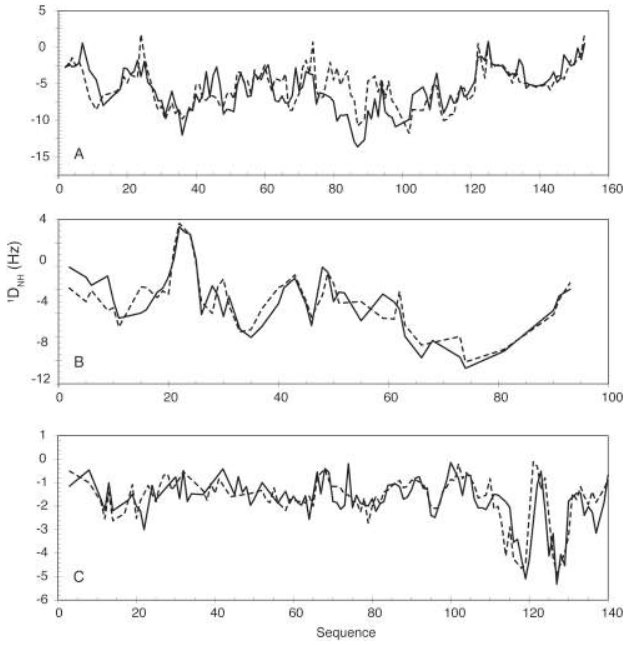


Fig. 4. RDCs calculated using *flexible-meccano* (solid lines), compared with experimental values (dashed lines) for (**A**) apomyoglobin in 8 M Urea (simulation used statistical coil for all residues; Mohana-Borges *et al*., 2004), (**B**) N-terminal disordered transactivation domain of p53 (1–93; Wells *et al*., 2008), simulation used coil sampling except for residues 22–24 which populate 30% more helix, and 58–91, which populate 20% more PPII and (**C**) alpha-synuclein. Simulation used coil sampling except for the presence of a long range contact between residues 1–10 and 130–140.

(2) It is possible to modify the potentials by adding additional propensities. In the example shown in Figure 2 the strand from residue 12–18 populates an α-helix for 50% of the conformers. The remaining conformers follow the amino acid-specifi potentials. It is also possible to introduce specifi sampling for each amino acid, focusing on specifi regions of Ramachandran space. All modification can be introduced 'by-hand' in the input fil containing the sequence information, or via the graphical interface.

(3) Long-range contacts between different regions of the primary sequence are also specifie using the graphical interface, where the range of amino acids involved in the contact, and the maximum distance between any residues in the two ranges, are introduced.

(4) The user can choose between the following output options: the number of structures in the ensemble to be calculated, whether or not explicit structural coordinates should be written for each conformer in the ensemble, and which data types (RDCs, PREs, J-couplings etc.) should be predicted from the ensemble. In the case of RDC prediction, the user can select the RDCs to be calculated, whereas for PRE prediction the position of the cysteine mutants must be specified the magnetic fiel strength and the intrinsic (diamagnetic) linewidth in the proton dimension.

(5) The ensemble calculation and data prediction algorithm runs entirely as a background calculation, so that the user can either run numerous calculations from the same interface, or analyze the results of one calculation while running another. All results are robustly classifie in terms of date and time of initialization of each calculation.

(6) Data output is provided in text format, and in graphical form (postscript). In addition to the specifie data types, the program also provides a distribution of the radius of gyration over the ensemble.

(7) Robust scripts are provided that interface the resulting coordinate file with protocols for calculating ensemble averaged chemical shifts and small angle scattering curves.

Examples of RDCs calculated using *flexible-meccano* in comparison to experimental values from apomyoglobin (Mohana-Borges *et al*., 2004), alpha-synuclein (Bernadó *et al*., 2005a) and p53 (Wells *et al*., 2008) are shown in Figure 4 along with the conformational sampling regime used to reproduce the data.

5 DISCUSSION

The classical structure function paradigm that has successfully underpinned our understanding of molecular biology breaks down when considering proteins that have no stable tertiary structure in their functional form. The determination of a 3D structure can provide only a single snapshot of such a highly flexibl system, and alternative methods are essential to study the behaviour of these disordered proteins (Fisher and Stultz, 2011; Marsh *et al*., 2010; Mittag *et al*., 2010; Tompa, 2011). One promising approach is to describe the protein in terms of an equilibrium of rapidly inter-converting conformers. Currently, tools to generate such ensemble descriptions are rare, and poorly adapted to the prediction of experimental data. In this article, we present an algorithm that generates ensembles of molecules, on the basis of amino acid-specifi conformational potentials and volume exclusion. Conformational sampling depends uniquely on the primary sequence, with the possibility of introducing additional local or long-range conformational propensities. We show how the algorithm can be used to calculate expected values of experimental NMR parameters measured at atomic or molecular resolution, for a broad range of user-define conformational sampling regimes.

We envisage three generic levels of interest: (i) Researchers who have measured experimental data from a particular IDP of interest and are motivated to compare their data with those expected from a fully disordered protein with this specifi primary sequence. No such tool exists at this time. (ii) Researchers who see evidence of deviation from 'random coil' behaviour in their protein, and who would like to determine whether these data are in agreement with a particular conformational sampling regime (e.g. a particular propensity of helical sampling in a given region of the chain, or the presence of weak long-range contacts between parts of the chain that are distant in primary sequence; Nodet *et al*., 2009; Salmon *et al*., 2010; Schneider *et al*., 2012). (iii) Researchers who are interested in working with a broad ensemble of conformers representing the flexibilit of the IDP of interest, either to use 'sample and select' approaches to develop a sub-ensemble in agreement with experimental data, or to seed molecular dynamics or restrained

ensemble molecular dynamics simulations of their protein. We add a fina note of caution: any analysis that involves detailed inspection of the structure of specifi conformers must of course be performed in the knowledge that no ensemble of highly disordered proteins can ever be considered to be unique. Agreement with experiment only confirm that a given ensemble does not violate a specifi data type, which itself is only sensitive to particular aspects of the conformational sampling.

Funding: Agence National de Recherche for financia support from TAUSTRUCT—ANR MALZ 2010 (to M.B.), ProteinDisorder—ANR JCJC 2010 (to M.R.J.), Spin-HD—ANR CHEX 2011 (to P.B.) and the GIPSE computational support group of the Commissariat à l'Energie Atomique et aux energies alternatives (CEA).

Conflict of Interest: none declared.

REFERENCES

Abragam,A. (1994) *The Principles of Nuclear Magnetism Reprint*. Clarendon Press, Oxford, UK.

Babu,M.M. (2012) Intrinsically disordered proteins. *Mol. Biosyst.*, **8**, 21.

Babu,M.M. *et al*. (2011) Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.*, **21**, 432–440.

Berjanskii,M. *et al*. (2009) GeNMR: a web server for rapid NMR-based protein structure determination. *Nucleic Acids Res.*, **37**, W670–W677.

Berlin,K. *et al*. (2009) Improvement and analysis of computational methods for prediction of residual dipolar couplings. *J. Magn. Reson.*, **201**, 25–33.

Bernadó,P. and Blackledge,M. (2010) Structural biology: proteins in dynamic equilibrium. *Nature*, **468**, 1046–1048.

Bernadó,P. and Svergun,D.I. (2012) Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol. Biosyst.*, **8**, 151–167.

Bernadó,P. *et al*. (2005a) Definin long-range order and local disorder in native alpha-synuclein using residual dipolar couplings. *J. Am. Chem. Soc.*, **127**, 17968–17969.

Bernadó,P. *et al*. (2005b) A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc. Natl Acad. Sci. USA*, **102**, 17002–17007.

Bernadó,P. *et al*. (2007) Structural characterization of flexibl proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.*, **129**, 5656–5664.

Blackledge,M. (2005) Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. *Prog. Nucl. Magn. Reson. Spectrosc.*, **46**, 23–61.

Bouvignies,G. *et al*. (2006a) Simultaneous determination of protein backbone structure and dynamics from residual dipolar couplings. *J. Am. Chem. Soc.*, **128**, 15100–15101.

Bouvignies,G. *et al*. (2006b) Ultrahigh-resolution backbone structure of perdeuterated protein GB1 using residual dipolar couplings from two alignment media. *Angew. Chem. Int. Ed. Engl.*, **45**, 8166–8169.

Bruschweiler,R. *et al*. (1992) Influenc of rapid intramolecular motion on NMR cross-relaxation rates - a molecular-dynamics study of Antamanide in solution. *J. Am. Chem. Soc.*, **114**, 2289–2302.

Cavalli,A. *et al*. (2007) Protein structure determination from NMR chemical shifts. *Proc. Natl Acad. Sci. USA*, **104**, 9615–9620.

Clore,G.M. and Iwahara,J. (2009) Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. *Chem. Rev.*, **109**, 4108–4139.

Dunker,A.K. and Uversky,V.N. (2010) Drugs for "protein clouds": targeting intrinsically disordered transcription factors. *Curr. Opin. Pharmacol.*, **10**, 782–788.

Dunker,A.K. *et al*. (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.

Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.

Eliezer,D. *et al*. (2001) Conformational properties of alpha-synuclein in its free and lipid-associated states. *J. Mol. Biol.*, **307**, 1061–1073.

Fisher,C.K. and Stultz,C.M. (2011) Constructing ensembles for intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, **21**, 426–431.

Gillespie,J.R. and Shortle,D. (1997) Characterization of long-range structure in the denatured state of staphylococcal nuclease. II. Distance restraints from paramagnetic relaxation and calculation of an ensemble of structures. *J. Mol. Biol.*, **268**, 170–184.

Han,B. *et al*. (2011) SHIFTX2: significantl improved protein chemical shift prediction. *J. Biomol. NMR*, **50**, 43–57.

Hus,J.-C. *et al*. (2001) Determination of protein backbone structure using only residual dipolar couplings. *J. Am. Chem. Soc*, **123**, 1541–1542.

Hus,J.-C. *et al*. (2008) 16-fold degeneracy of peptide plane orientations from residual dipolar couplings: analytical treatment and implications for protein structure determination. *J. Am. Chem. Soc.*, **130**, 15927–15937.

Jensen,M.R. and Blackledge,M. *et al*. (2008) On the origin of NMR dipolar waves in transient helical elements of partially folded proteins. *J. Am. Chem. Soc.*, **130**, 11266–11267.

Jensen,M.R. *et al*. (2008) Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein. *J. Am. Chem. Soc.*, **130**, 8055–8061.

Jensen,M.R. *et al*. (2009) Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Structure*, **17**, 1169–1185.

Jensen,M.R. *et al*. (2010) Definin conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *J. Am. Chem. Soc.*, **132**, 1270–1272.

Jensen,M.R. *et al*. (2011) Intrinsic disorder in measles virus nucleocapsids. *Proc. Natl Acad. Sci. USA*, **108**, 9839–9844.

Kjaergaard,M. *et al*. (2011) Random coil chemical shift for intrinsically disordered proteins: effects of temperature and pH. *J. Biomol. NMR*, **49**, 139–149.

Ludvigsen,S. *et al*. (1991) Accurate measurements of coupling constants from two-dimensional nuclear magnetic resonance spectra of proteins and determination of phi-angles. *J. Mol. Biol.*, **217**, 731–736.

Marsh,J.A. *et al*. (2006) Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation *Protein Sci.*, **15**, 2795–2804.

Marsh,J.A. *et al*. (2008) Calculation of residual dipolar couplings from disordered state ensembles using local alignment. *J. Am. Chem. Soc.*, **130**, 7804–7805.

Marsh,J.A. *et al*. (2010) Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators. *Structure*, **18**, 1094–1103.

Mittag,T.*et al*. (2010) Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure*, **18**, 494–506.

Modig,K.*et al*. (2007) Detection of initiation sites in protein folding of the four helix bundle ACBP by chemical shift analysis. *FEBS Lett.*, **581**, 4965–4971.

Mohana-Borges,R. *et al*. (2004) Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. *J. Mol. Biol.*, **340**, 1131–1142.

Mukrasch,M.D. *et al*. (2007) Highly populated turn conformations in natively unfolded tau protein identifie from residual dipolar couplings and molecular simulation. *J. Am. Chem. Soc.*, **129**, 5235–5243.

Neal,S. *et al*. (2003) Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. *J. Biomol. NMR*, **26**, 215–240.

Nodet,G. *et al*. (2009) Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *J. Am. Chem. Soc.*, **131**, 17908–17918.

Pardi,A. *et al*. (1984) Calibration of the angular dependence of the amide proton-C alpha proton coupling constants, 3JHN alpha, in a globular protein. Use of 3JHN alpha for identificatio of helical secondary structure. *J. Mol. Biol.*, **180**, 741–751.

Salmon,L. *et al*. (2009) Protein conformational flexibilit from structure-free analysis of NMR dipolar couplings: quantitative and absolute determination of backbone motion in ubiquitin. *Angew. Chem. Int. Ed. Engl.*, **48**, 4154–4157.

Salmon,L. *et al*. (2010) NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.*, **132**, 8407–8418.

Schneider,R. *et al*. (2012) Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. *Mol. Biosyst.*, **8**, 58–68.

Schwarzinger,S. *et al*. (2001) Sequence-dependent correction of random coil NMR chemical shifts. *J. Am. Chem. Soc.*, **123**, 2970–2978.

Shen,Y. and Bax,A. (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR*, **38**, 289–302.

Shen,Y. *et al*. (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl Acad. Sci. USA*, **105**, 4685–4690.

Smith,L.J. *et al*. (1996) Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations. *J. Mol. Biol.*, **255**, 494–506.

Svergun,D. *et al*. (1995) CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.*, **28**, 768–773.

Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.

Tompa,P. (2011) Unstructural biology coming of age. *Curr. Opin. Struct. Biol.*, **21**, 419–425.

Uversky,V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, **11**, 739–756.

Vendruscolo,M. and Dobson,C.M. (2007) Chemical biology: more charges against aggregation. *Nature*, **449**, 555.

Vuister,G. and Bax,A. (1993) Quantitative J correlation - a new approach for measuring homonuclear 3-bond J(H(N)H(alpha) coupling-constants in N-15-enriched proteins. *J. Am. Chem. Soc.*, **115**, 7772–7777.

Wells,M. *et al*. (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc. Natl Acad. Sci. USA*, **105**, 5762–5767.

Wishart,D.S. and Sykes,B.D. (1994) The ^{13}C chemical-shift index: a simple method for the identificatio of protein secondary structure using ^{13}C chemical-shift data. *J. Biomol. NMR*, **4**, 171–180.

Zweckstetter,M. and Bax,A. (2000) Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. *J. Am. Chem. Soc.*, **122**, 3791–3792.

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

pubs.acs.org/JACS

Mapping the Potential Energy Landscape of Intrinsically Disordered Proteins at Amino Acid Resolution

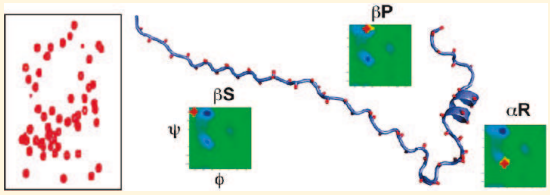
Valéry Ozenne,[†] Robert Schneider,[†] Mingxi Yao,[†] Jie-rong Huang,[†] Loïc Salmon,[†] Markus Zweckstetter,[‡] Malene Ringkjøbing Jensen,[†] and Martin Blackledge*[,†]

[†]CEA, CNRS, and UJF-Grenoble 1, Protein Dynamics and Flexibility, Institut de Biologie Structurale Jean-Pierre Ebel, 41 Rue Jules Horowitz, Grenoble 38027, France

[‡]Department of NMR-based Structural Biology, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, German Center for Neurodegenerative Diseases (DZNE), 37077 Göttingen, Germany

Supporting Information

ABSTRACT: Intrinsically disordered regions are predicted to exist in a significant fraction of proteins encoded in eukaryotic genomes. The high levels of conformational plasticity of this class of proteins endows them with unique capacities to act in functional modes not achievable by folded proteins, but also places their molecular characterization beyond the reach of classical structural biology. New techniques are therefore required to understand the relationship between primary sequence and biological function in this class of proteins. Although dependences of some NMR parameters such as chemical shifts (CSs) or residual dipolar couplings (RDCs) on structural propensity are known, so that sampling regimes are often inferred from experimental observation, there is currently no framework that allows for a statistical mapping of the available Ramachandran space of each amino acid in terms of conformational propensity. In this study we develop such an approach, combining highly efficient conformational sampling with ensemble selection to map the backbone conformational sampling of IDPs on a residue specific level. By systematically analyzing the ability of NMR data to map the conformational landscape of disordered proteins, we identify combinations of RDCs and CSs that can be used to raise conformational degeneracies inherent to different data types, and apply these approaches to characterize the conformational behavior of two intrinsically disordered proteins, the K18 domain from Tau protein and Ntail from Measles virus nucleoprotein. In both cases, we identify the enhanced populations of turn and helical regions in key regions of the proteins, as well as contiguous strands that show clear and enhanced polyproline II sampling.



INTRODUCTION

The realization that a large fraction of proteins encoded in eukaryotic genomes contain a significant level of functional disorder[1−4] has engendered considerable interest in the development of experimental and analytical techniques to describe this disorder.[5−8] The conformational plasticity of intrinsically disordered proteins (IDPs) endows them with unique capabilities to act in functional modes not achievable by folded, globular proteins. A number of different scenarios have been identified for the binding of IDPs to their partner proteins, including folding-upon-binding[9] or the formation of dynamic, so-called fuzzy complexes[10] where the IDP samples various states on the surface of the partner. However, a number of open questions remain, for example, it is unclear how the intrinsic structural propensity is defined by the primary sequence of an IDP, and how this propensity is related to the thermodynamics and kinetics of the interaction and the conformation adopted in the complex. A full understanding of how IDPs carry out their function in the absence of a stable tertiary fold requires a description of the potential energy landscape sampled by each amino acid in the protein. In order to achieve this end, ensemble representations of a continuum of rapidly interconverting structures have emerged as a convenient tool for representing the structural and dynamic properties of IDPs and their complexes.[11−19] In this context, the determination of representative descriptions of the behavior of IDPs remains one of the major challenges for the study of the molecular basis of biological function in these highly disordered systems.

Nuclear magnetic resonance (NMR) spectroscopy represents a tool of choice to address this challenge, providing experimental measurement of site-specific ensemble averages over all conformers sampled up to the millisecond time scale. Of these, the chemical shift (CS) is the most accessible, reporting on the local chemical and electronic environment, as well as medium and long-range interactions.[20−23] Unfortunately, this conformational dependence is poorly defined at a theoretical level. A popular empirical alternative is to compile experimental chemical shifts measured in folded proteins for which three-dimensional coordinates are available and to establish conformational dependences on this basis.[24,25] This approach has led to the observation that secondary structural elements such as α-helices and β-sheets can be readily identified

Received: July 15, 2012

   dx.doi.org/10.1021/ja306905s | J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

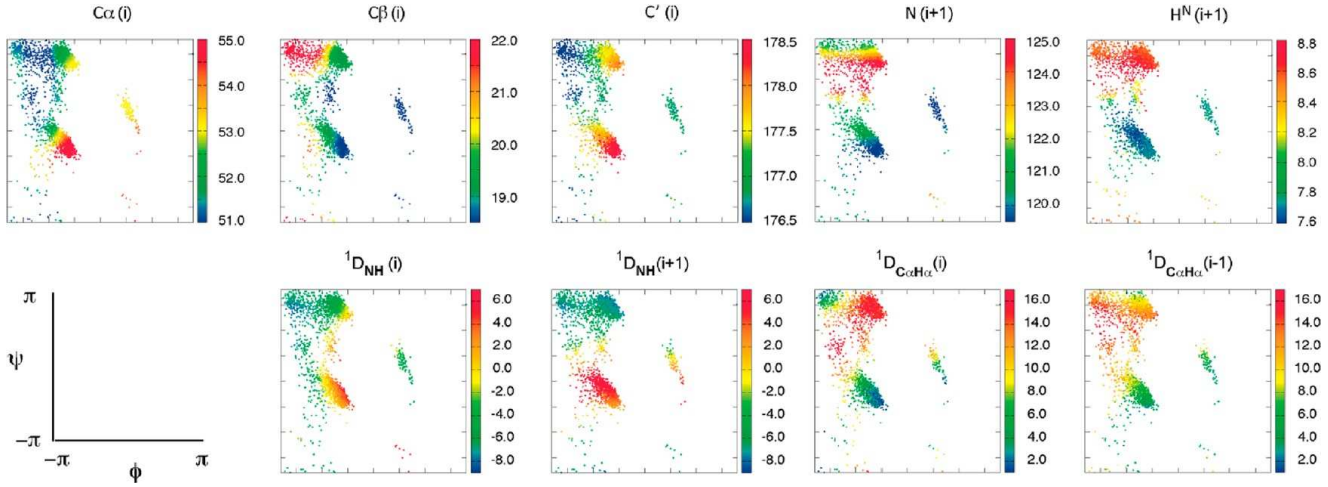


Figure 1. Dependence of primary experimental data on backbone dihedral angle sampling. (A) Distribution of predicted chemical shifts (in ppm) for the central residue i = 8 of a poly alanine 15-mer chain as function of the conformational sampling $\{\phi, \psi\}$ of residue i. (B) Ensemble averaged backbone RDCs for the poly alanine 15-mer chain plotted against average $\{\phi, \psi\}$ values of residue i. Values are shown in hertz (Hz) in all cases, assuming an arbitrary level of overall alignment.

on the basis of the ^{13}C backbone CS.[20,26−28] Structural restraints based on CS have also been introduced into structure determination algorithms, and the power of CS prediction using database-dependent approaches was further exemplified via their combination with molecular modeling to achieve full structure determination.[29−31]

The application of CS to the study of disordered systems, where deviation of the shift from its coil value—the secondary shift—is expected to be smaller than in a folded protein, requires a more subtle approach.[21,27,32,33] Nevertheless, the strong and complementary dependence of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shifts on the presence of α-helix and β-sheet conformations has led to the development of simple and accurate algorithms for the determination of the propensity of regions of the protein to form secondary structure in solution.[34] Recently CSs have been combined with ensemble selection algorithms[14,15,35,36] or expressed as the population weighted average of generic CSs from three regions of Ramachandran space (α-helix, β-sheet and polyproline II) and a random coil shift,[37] to solve for the populations of these regions. Residual dipolar couplings (RDCs), measured under conditions of weak molecular alignment, are sensitive to the reorientational sampling properties of internuclear bond-vectors, and are therefore also sensitive reporters of the local conformational behavior of IDPs.[16,38−41] Most applications of RDCs to the studies of disordered systems have exploited the particular ability of RDCs to identify the presence of α-helical and turn elements in otherwise disordered systems,[42−45] while the combination of different RDCs measured throughout the peptide plane can also detect enhanced sampling of more extended backbone conformations (either β-sheet or polyproline II).[15,39,46]

Despite intense contemporary interest in this question however, it remains unclear how accurately NMR CSs and RDCs can be used to uniquely define backbone conformational sampling in intrinsically disordered proteins, principally because no analytical or numerical framework for the determination of the potential energy landscape of unfolded proteins at amino acid specific resolution is yet available. This question is of additional importance because of the proposed relevance, derived from vibrational spectroscopy and circular dichroism as well as homonuclear NMR, of the polyproline II (PPII) region of Ramachandran space for the behavior of disordered proteins.[47−49] The development of a method that unequivocally maps the population of the entire backbone conformational space sampled by amino acid is therefore of considerable importance.

In this study, we develop an approach to address the ability of primary experimental NMR data, specifically CSs and RDCs, to map the conformational behavior of IDPs on an amino acid specific basis. To achieve this aim, we combine the ensemble selection algorithm ASTEROIDS,[15] with *flexible-meccano*[50,51] and SPARTA[25] to systematically map the sensitivity of different CSs and RDCs to determine the population distribution of each backbone dihedral angle in the protein. This approach provides clear insight into conformational propensities that can be distinguished on the basis of experimental data, and simultaneously identifies regions of Ramachandran space whose populations cannot be resolved. Finally, we propose combinations of RDCs and CSs that can be used to raise these degeneracies and determine populations of all regions of Ramachandran space. The approach is applied to the two experimental cases, Ntail, the intrinsically disordered C-terminal domain of the nucleoprotein from measles virus, and the K18 domain of the protein Tau, an IDP that is implicated in the development of Alzheimer's disease. In both systems, we identify turn and helical regions as well as the presence of contiguous regions exhibiting enhanced PPII sampling.

RESULTS AND DISCUSSION

Variation of Backbone Chemical Shifts over $\{\phi, \psi\}$ Space. One of the advantages of using CSs as structural probes is that resonances from different nuclei exhibit complementary dependences on backbone dihedral angles $\{\phi,\psi\}$. In principle, this complementarity may allow for a site-specific mapping of the conformational sampling in disordered proteins. The predicted dihedral angle dependence of five experimentally measurable CSs is shown in Figure 1 for an alanine tripeptide. The conformers were generated using *flexible-meccano* on the basis of the statistical coil model, and the chemical shifts were predicted for each conformer using the program SPARTA.[25] To simplify the subsequent discussion, we divide the

dx.doi.org/10.1021/ja306905s | J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

Ramachandran plot into four regions: β-sheet (βS), PPII (βP), α-helical (αR) and left handed helix (αL) (Figure 2). We note

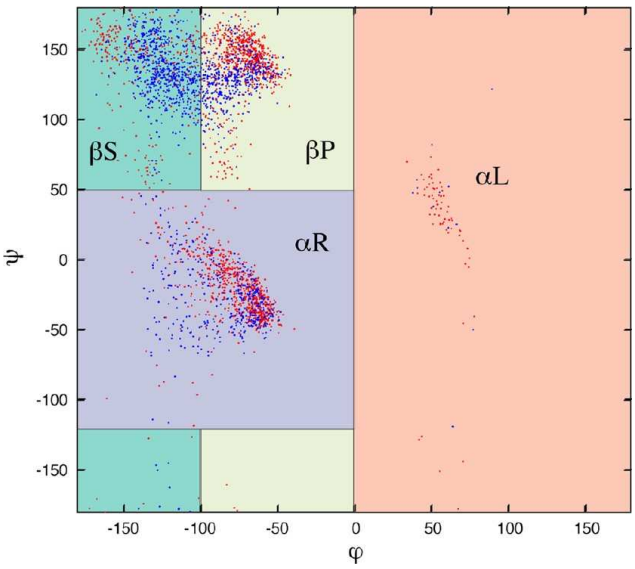


Figure 2. Definition of the regions of Ramachandran space used throughout the study. Points shown are from valine (blue) and alanine (red) residues in statistical coil conformations.

that this definition of conformational space avoids the appearance of bias when mapping specific conformations due to the arbitrary definition of an additional sampling regime termed 'random coil' that represents the remaining sampling. In this study, the entire Ramachandran space is mapped in terms of population distributions, or described in terms of these 4 regions, obviating the need to define an additional 'random coil' region.

Well-known dependences are immediately identifiable from Figure 1, with higher values of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shifts uniquely populating αR and βS conformations, respectively. The determination of the populations in other regions of Ramachandran space appears less straightforward. Thus, similar shifts are predicted in the βP and the upper left αR region for $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$, making it difficult, on the basis of the ^{13}C CSs alone, to map the populations in these regions. This degeneracy is partially raised by considering the influence of the $\{\phi,\psi\}$ sampling on the CSs of the neighboring amino acids. In particular, ^{15}N and $^{1}H^{N}$ shifts of the following residue provide additional differentiation of the βP and upper αR regions.

The prediction for the alanine tripeptide shown in Figure 1 is relevant for this specific sequence. While overall features will be retained for different sequences, considerable variation is observed as a function of the identity of the three amino acids. To develop a better understanding of the ability of ensemble descriptions to define conformational propensities on the basis of CSs, we have therefore performed explicit simulations using synthetic data derived from specific conformational sampling regimes.

Ensemble Mapping of Conformational Propensities from Chemical Shifts. Conformationally biased ensembles obeying specific sampling properties were generated using the *flexible-meccano* algorithm, and averaged CSs were predicted from these ensembles using the program SPARTA. These synthetic data were then used as the target for the ASTEROIDS approach to select subensembles in agreement with these values (see Methods). Subensembles are selected from a pool of 20 000 structures calculated using the amino acid specific potential energy surfaces derived from the statistical coil model. An iterative procedure is then used to modify the potentials to enhance the sampling as a function of each selection until convergence is achieved. It is important to note here that the *flexible-meccano*/ASTEROIDS approach is used as a means to describe the potential energy landscape sampled by the protein backbone. Repetition of the selection procedure determines ensembles containing different structures, which are therefore not unique in this sense; however, the backbone sampling characteristics do not vary from one ensemble to another, which are therefore converged and unique in terms of conformational substates and their populations. This also demonstrates that pool sampling is sufficiently complete.

The modulation of the predicted CSs when sampling a specific conformational propensity is compared to statistical coil values in Figure 3a. Three regimes that are significantly different from the statistical coil model were tested, comprising a higher tendency to sample the βS, βP or αR regions (see Methods). Simple inspection reveals that while well-known deviations are seen for ^{13}C shifts in the presence of βS and αR propensity, these CSs are hardly modified by the presence of raised βP population. This is evidently because the mean values of the statistical coil shifts are essentially indistinguishable from βP values (Figure 1). The uncertainties for each CS as determined from predictions for folded proteins are also shown on this Figure 3a.[25] It is notable that the expected changes for ^{15}N and $^{1}H^{N}$ shifts in the presence of enhanced βP sampling are relatively small compared to this uncertainty.

We initially consider two scenarios for selection on the basis of CSs, simulating data sets comprising either $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ or 'full' CS sets including $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N and $^{1}H^{N}$. Figure 3b presents the ability of ASTEROIDS to reproduce conformational tendencies present throughout the protein when using these different combinations of CSs in the target function. In all cases, the simulated data are well reproduced by the selected ensemble (Supporting Information Figure S1). When using CSs from $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ the ASTEROIDS algorithm accurately reproduces the propensity of enhanced conformational sampling in the βS and αR regions (see also Table 1). The population of the βP region is however poorly reproduced, with additional sampling of the upper αR region that appears to compensate for insufficient sampling of βP. Figure 3c shows the comparison of the average Ramachandran space of the 5 amino acids from each strand (βS, αR and βP) and from the coil regions in between these strands, for the target and selected ensembles. This further highlights the degeneracy of the upper αR and βP regions when only $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ CSs are used in the selection. As expected from consideration of Figure 1, the addition of ^{15}N and $^{1}H^{N}$ improves this situation considerably; however, the dependence of these shifts on additional factors such as temperature, ionic strength and pH, renders them potentially volatile in terms of conformational mapping. To determine the levels of confidence that can be derived from different CSs, we have therefore applied the same approach to simulated data with Gaussian-based noise levels reflecting the relative accuracy of predictions for the different nuclei (see Methods). The results are summarized in Table 1, and demonstrate that the accuracy of the determination of the populations of βS and αR regions is significantly more robust to the presence of noise than βP,

dx.doi.org/10.1021/ja306905s | J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

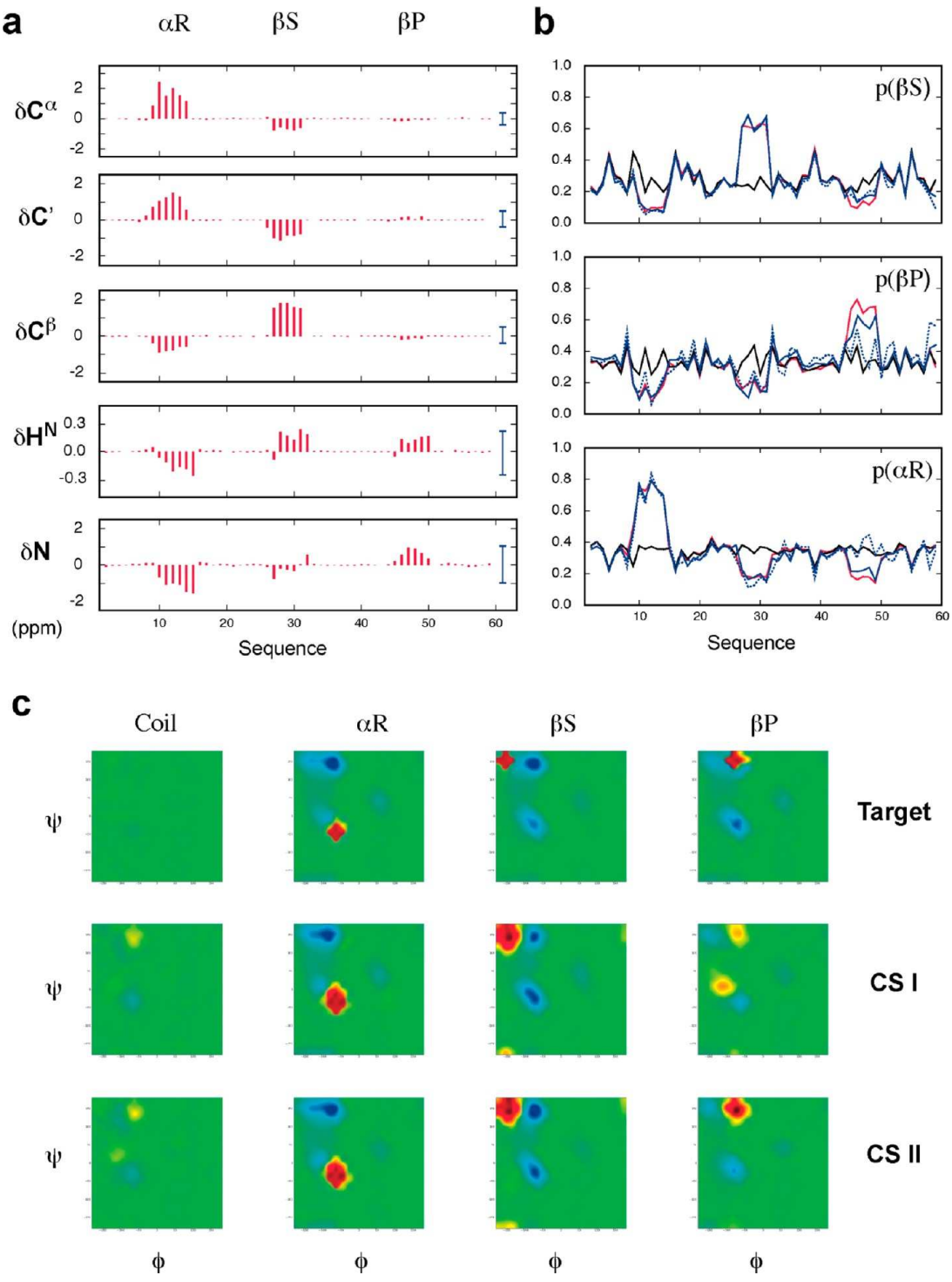


Figure 3. Mapping of conformational space in disordered systems using CS. (a) Modification of predicted chemical shifts for enhanced conformational propensities in different regions of Ramachandran space compared to statistical coil values. Three regimes that are significantly different from the statistical coil model were tested, comprising a higher tendency to sample the βS, αR and βP regions. Blue error bars indicate the average accuracy to which each chemical shift is predicted for folded proteins. (b) Reproduction of conformational sampling by an ASTEROIDS-selected ensemble comprising 200 conformers obtained by targeting the synthetic chemical shift data set shown in panel a. The pool from which the structures were selected was created using the standard coil library of *flexible-meccano*. Selection carried out using $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts or $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ and ^{15}N, $^{1}H^{N}$. Red: populations of conformational space in the target ensemble. Blue: populations in the selected ensemble (dashed line ^{13}C shifts only, solid line all shifts). Black: populations in the starting (statistical coil) ensemble. (c) Ramachandran plots showing the difference between the population in the regions sampling coil, αR, βP and βS regions. Top line, target ensemble; middle line, selection using only ^{13}C CS, bottom line selection using all CS. Red, increased sampling; blue, reduced sampling compared to statistical coil.

dx.doi.org/10.1021/ja306905s | *J. Am. Chem. Soc.* XXXX, XXX, XXX−XXX

Table 1. Ability of CSs To Reproduce Conformational Sampling in the Presence and Absence of Noise

Δ^a	βS^b	αR^b	βP^b
Coil[c]	0.45	0.45	0.41
CS I[d]	0.065	0.07	0.35
CS II[e]	0.06	0.08	0.08
CS I σ[f]	0.11	0.13	0.41
CS II σ[g]	0.18	0.17	0.27
RDC[h]	0.12	0.11	0.27
RDC CS[i]	0.07	0.05	0.06
RDC CS σ[j]	0.13	0.13	0.19
RDC CS σ[k]	0.10	0.13	0.15

[a]All values in the table show average absolute differences between target and selection, averaged over the 5 amino acid regions experiencing selective enhanced sampling. [b]Populations averaged over the five amino acids oversampling these regions. [c]Difference between target population and statistical coil average. [d]Difference between target population and selection using $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ CSs. [e]Difference between target population and selection using $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{N}$ CSs. [f,g]As in *d*, *e* in the presence of Gaussian weighted noise using errors estimated from 25% of the rmsd's of SPARTA predictions of CSs from folded proteins. [h]Difference between target population and selection using $^{1}D_{N-H}$, $^{2}D_{C'-HN}$, $^{1}D_{C\alpha-H\alpha}$ and $^{1}D_{C\alpha-C'}$ RDCs. [i]Difference between target population and selection using RDCs listed in *h* and $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ CSs. [j]Difference between target population and selection using $^{1}D_{N-H}$ and $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ CSs in the presence of noise. [k]As in *i* in the presence of noise.

mainly due to the higher predictive imprecision of ^{15}N and $^{1}H^{N}$ shifts.

These calculations highlight two important points concerning the use of CSs to map local conformational sampling in disordered systems. The first concerns the inherent degeneracy of CSs for the upper αR and βP regions, which is partially raised by the ^{15}N and $^{1}H^{N}$ shifts. Second, and more importantly, the expected ^{13}C CSs in the presence of enhanced βP sampling are strongly degenerate with the statistical coil values that are expected from intrinsic sampling in the absence of specific conformational propensity.

Variation of Residual Dipolar Couplings over {ϕ, ψ} Space. RDCs measured in disordered systems have also been shown to depend strongly on the nature of the backbone conformational sampling. This is illustrated in Figure 1 where different ensemble averaged backbone RDCs are plotted against average {ϕ, ψ} values (see Methods). The sensitivity of RDCs both to the conformational sampling of the amino acid of interest and its immediate neighbors complicates interpretation of this representation, and underlines the importance of using the ASTEROIDS approach to select ensembles of entire structures. Nevertheless, the most commonly measured RDCs, $^{1}D_{N-H}$ and $^{1}D_{C\alpha-H\alpha}$, clearly exhibit the expected sensitivity to αR, but also show degeneracy between βS and βP, either for the amino acid of interest or an immediate neighbor. Expected values for RDCs simulated from the sequence containing additional populations of βS, βP and αR presented above are shown in Figure 4a. In this case, all three additional propensities modulate the expected values of RDCs, averaging to different values than the statistical coil, although this modulation is similar for βS and βP.

An ASTEROIDS analysis was performed on the same system, using $^{1}D_{N-H}$, $^{2}D_{C'-HN}$, $^{1}D_{C\alpha-H\alpha}$ and $^{1}D_{C\alpha-C'}$ RDCs in the selection procedure. Figure 4b, 4c and table 1 present the ability of a combination of these four RDC types to define the conformational potentials. The ASTEROIDS-selected ensemble accurately reproduces the propensity of enhanced conformational sampling in the αR region, and in the extended region (βS and βP together). However the data do not distinguish between these extended regions, in particular the enhanced βP population is not correctly determined. Similarly, upper and lower αR regions are found to be degenerate when using only RDCs.

From the above it is evident that combination of CSs and RDCs should raise the upper αR/βP/coil and βS/βP degeneracies observed for ^{13}C CSs and RDCs respectively, and thereby allow for a more accurate mapping of Ramachandran space. In the following we test this hypothesis and identify generally accessible and conformationally informative combinations of CS and RDCs that can be usefully applied to the study of a large number of disordered proteins.

Ensemble Mapping of Conformational Propensities by Combining CSs and RDCs. An ASTEROIDS analysis of the same system as illustrated earlier was performed combining $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ CSs with $^{1}D_{N-H}$, $^{2}D_{C'-HN}$, $^{1}D_{C\alpha-H\alpha}$ and $^{1}D_{C\alpha-C'}$ RDCs. In this case (Figure 5), a more precise mapping of Ramachandran space is achieved, raising all degeneracies identified for CSs and RDCs alone. Removal of some RDCs, so that only $^{1}D_{N-H}$ RDCs are included, still provides good reproduction of all regions of conformational space. As shown in Table 1, the populations are still correctly reproduced in the presence of significant levels of noise (equivalent to 1 Hz error for the $^{1}D_{N-H}$ RDCs).

The combination of ^{13}C, ^{15}N and $^{1}H^{N}$ CSs and $^{1}D_{N-H}$ RDCs represents a tractable solution for many experimental studies that is evidently information rich, while remaining robust with respect to uncertainty of experimental conditions, spectral calibration, noise and prediction error. We have therefore applied this approach to two experimental systems.

Application to the Disordered Domain of the Nucleoprotein from Measles Virus. $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N and $^{1}H^{N}$ CSs and $^{1}D_{N-H}$ RDCs were used to define the conformational sampling of the 125 amino acid intrinsically disordered C-terminal domain of the nucleoprotein of measles virus (Figure 6a). In addition to characterizing the molecular recognition element that comprises a high population of helix as described recently,[52,53] the 105 unfolded amino acids appear to indicate the presence of a lower population of βS in localized regions of this domain, compared to the statistical coil description (Figure 6b). This reduction is mainly due to higher βP population, in particular for the three continuous regions (435−445), (448−453) and (518−524), where close to 50% of conformers populate this region of Ramachandran space. Figure 8 shows the reproduction of the $^{1}D_{N-H}$ RDCs when only $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N and $^{1}H^{N}$ CSs are used, testifying that the analysis is both predictive, and not noticeably prone to overfitting.

Application to the K18 Domain of Tau Protein. The same method was applied to the 130 amino acid K18 domain of Tau protein using $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N and $^{1}H^{N}$ CSs and $^{1}D_{N-H}$ RDCs (Figure 7a). This domain contains four highly homologous repeat sequences, so that the sampling profile necessarily exhibits a repetitive nature. In this case the βS population is again depleted compared to the statistical coil (Figure 7b). The four previously described type I β-turns and the four triglycine sequences account for the eight regions of significantly increased αR population. The turns are found to

dx.doi.org/10.1021/ja306905s | J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

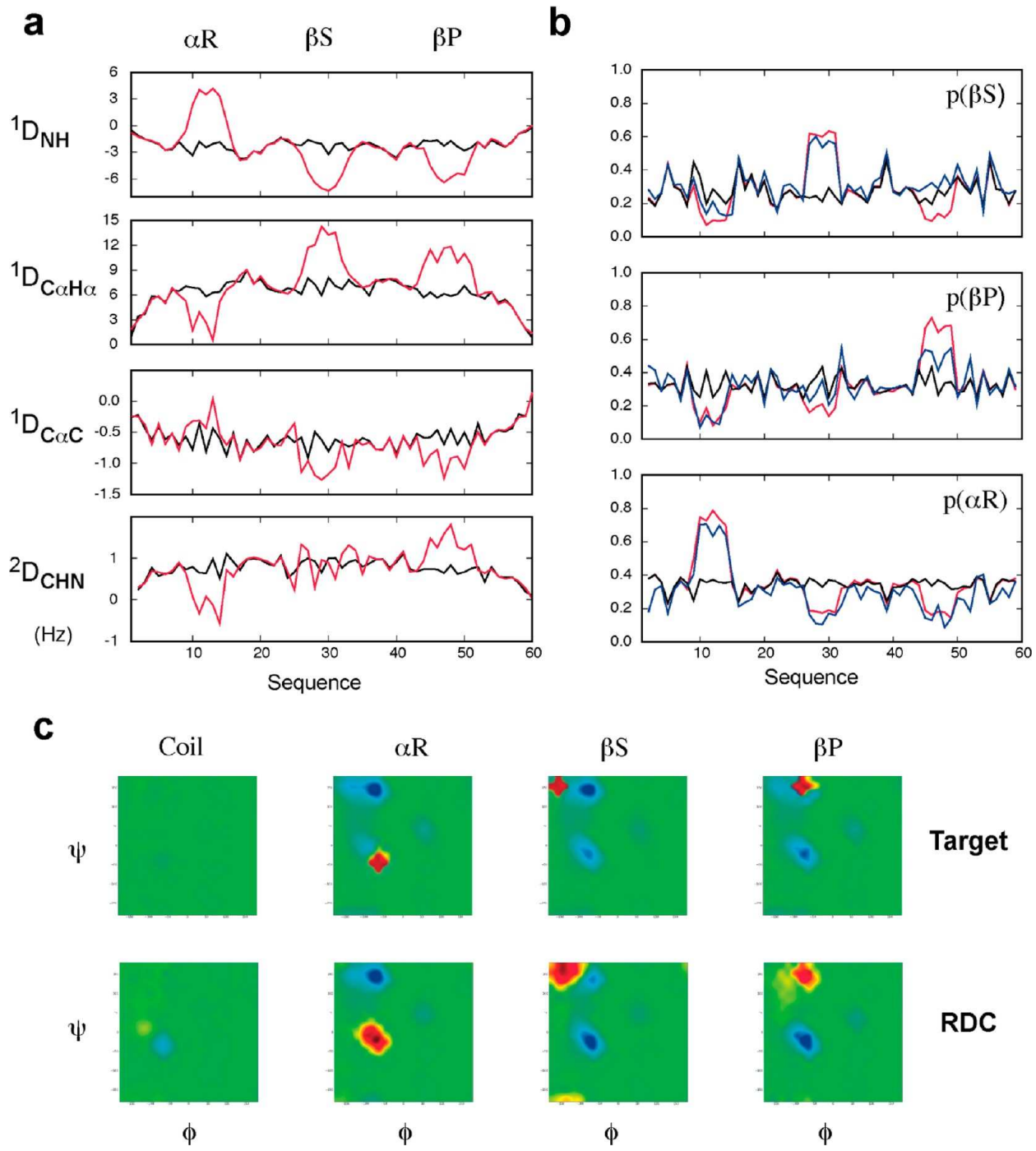


Figure 4. Mapping of conformational space in disordered systems using RDCs. (a) Modification of predicted RDCs for enhanced conformational propensities in different regions of Ramachandran space compared to expected values for statistical coil sampling (see Figure 2). An arbitrary level of alignment was assumed for the absolute scaling of the RDCs. (b) Amino acid specific difference in population between the ASTEROIDS selection and target using simulated RDC data shown in panel a. Red: populations in the target ensemble. Blue: populations in the selected ensemble. Black: populations in the starting (statistical coil) ensemble. (c) Ramachandran plots showing the difference between the population in the regions sampling coil, αR, βP and βS regions. Top line, target ensemble; bottom line, selection using simulated RDC data shown in panel a . Color coding as in Figure 2.

be populated between 15 and 25%, spanning very similar ranges to those determined using a combination of accelerated molecular dynamics and RDCs.[42,55] Outside these localized regions, a higher population of βP is observed, in particular in the aggregation nucleation sites, between residues (256−261), (275−282), (307−313) and (338−346). These strands, the central two of which mediate binding to microtubules and have been identified as aggregation nucleation sites important for the formation of Tau oligomers, have previously been proposed to sample extended populations.[42,55] The results shown here clearly indicate that this extended sampling is due to strongly enhanced sampling of the βP region of conformational space

dx.doi.org/10.1021/ja306905s | J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

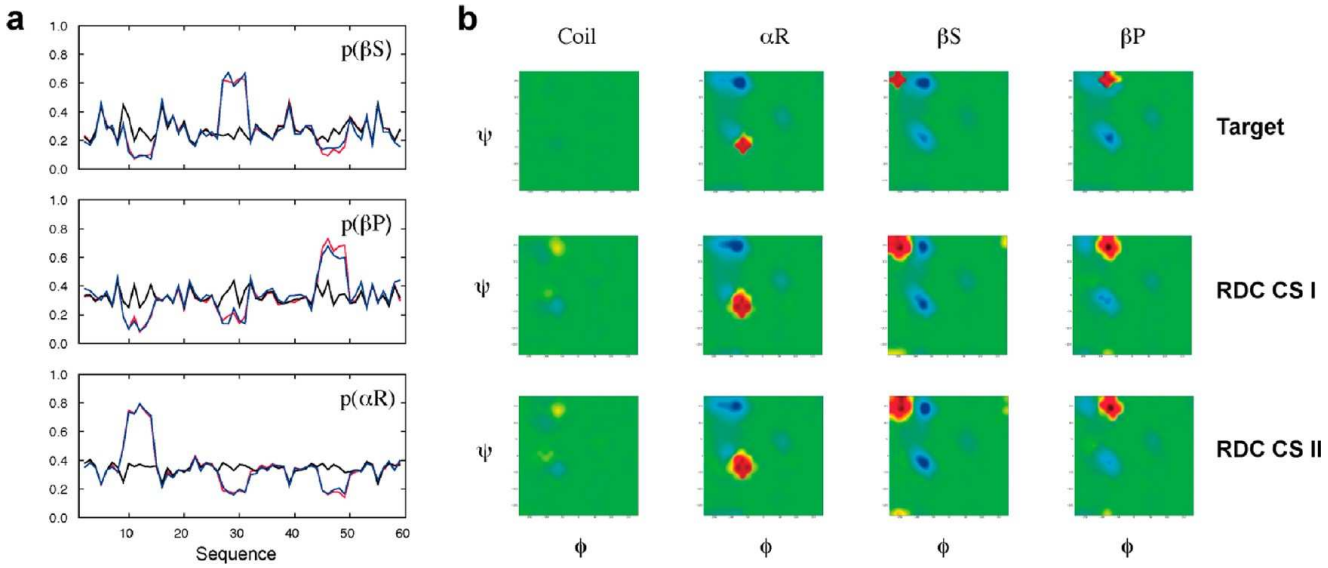


Figure 5. Mapping of conformational space in disordered systems using a combination of RDCs and CSs. (a) Amino acid specific difference in population between the ASTEROIDS selection and target applied to simulated CS and RDC data shown in Figures 2a and 3a. Red: populations in the target ensemble. Blue: populations in the selected ensemble. Black: populations of different regions of conformational space in the starting (statistical coil) ensemble. (b) Ramachandran plots showing the average difference between the population in the regions sampling coil, αR, βP and βS regions. Top line, target ensemble; middle line, selection using ^{13}C CS and $^{1}D_{NH}$ RDCs; bottom, selection using ^{13}C CS and all RDCs shown in Figure 3. Color coding as in Figure 2.

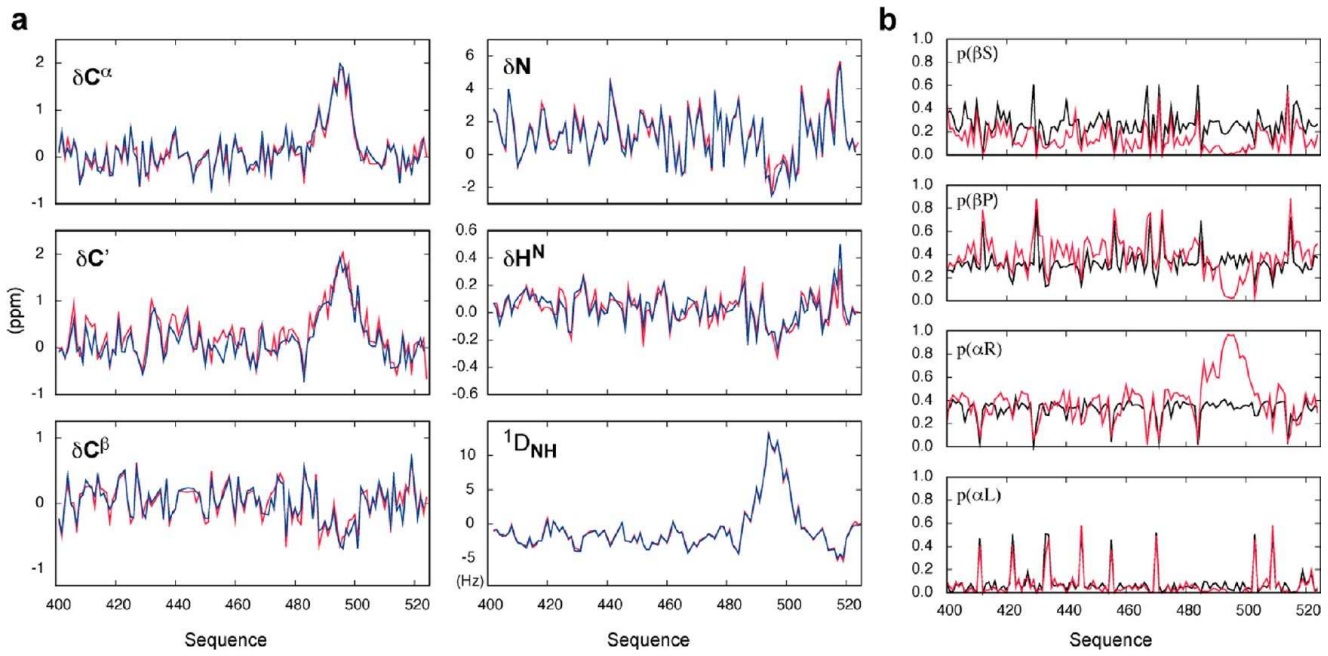


Figure 6. Characterization of intrinsically disordered proteins using RDCs and CSs. ASTEROIDS CS-RDC approach applied to experimental data from the disordered C-terminal domain, N_{TAIL}, of the nucleoprotein from measles virus. (a) Reproduction of experimental data (red experimental, blue ensemble average). (b) Population of different regions of conformational space for each amino acid in the N_{TAIL} sequence (red selected ensemble, black statistical coil).

f8 over a continuous range of 6−9 amino acids. Figure 8 shows the reproduction of the $^{1}D_{N-H}$ RDCs when only $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N and $^{1}H^{N}$ CSs are used; the 'free' data are again closely reproduced.

The amino acid conformational potentials for the region f9 273−287 of K18 are shown in Figure 9, in comparison to the statistical coil sampling. The raised βP sampling in the region 275−282 is evident, as is the partially populated β-turn that immediately follows this. We note that this conformational sampling, determined in this case uniquely from the experimental data, is very similar to that predicted by accelerated molecular dynamics simulation in a previous study,[42] populating enhanced αR in Leu284 and Ser 285 to very similar levels.

Finally, we note that this entire study was repeated using the program SPARTA+,[54] and the results concerning both

dx.doi.org/10.1021/ja306905s | J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

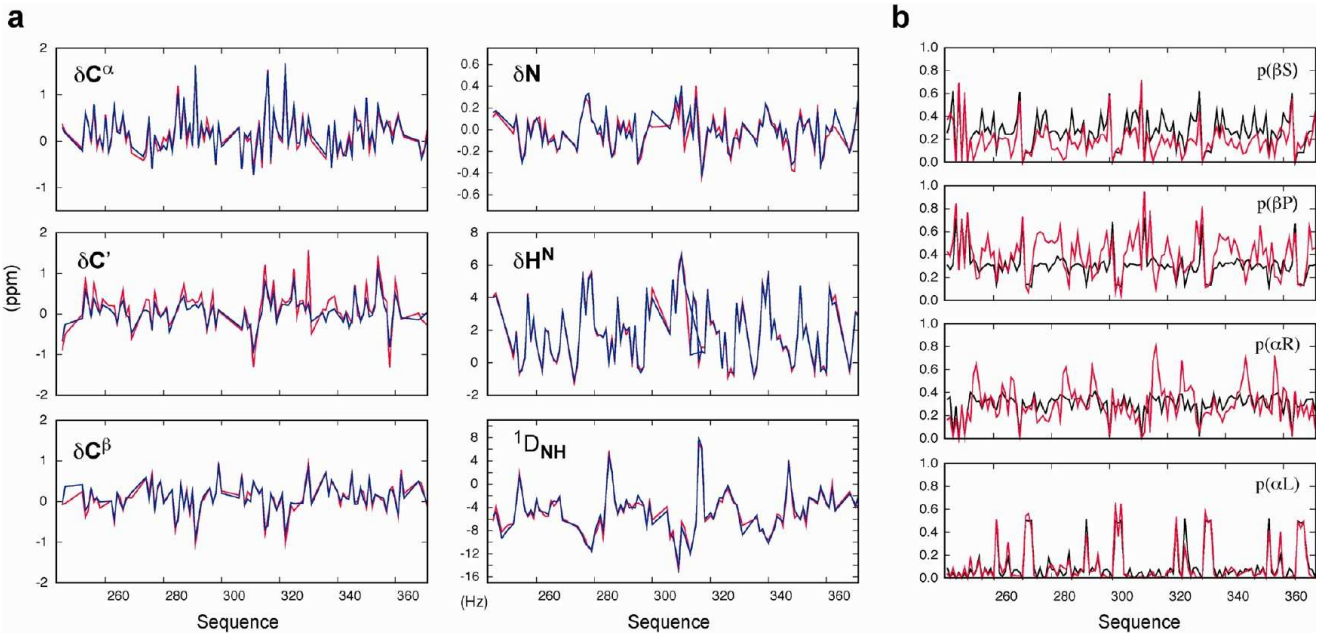


Figure 7. ASTEROIDS CS-RDC approach applied to experimental data from the K18 fragment of Tau protein. (a) Reproduction of experimental data (red experimental, blue ensemble average). In the case of chemical shifts, the reproduction of experimental data is shown resulting from the selection procedure described. (b) Population of different regions of conformational space for each amino acid in the K18 sequence (red selected ensemble, black statistical coil).

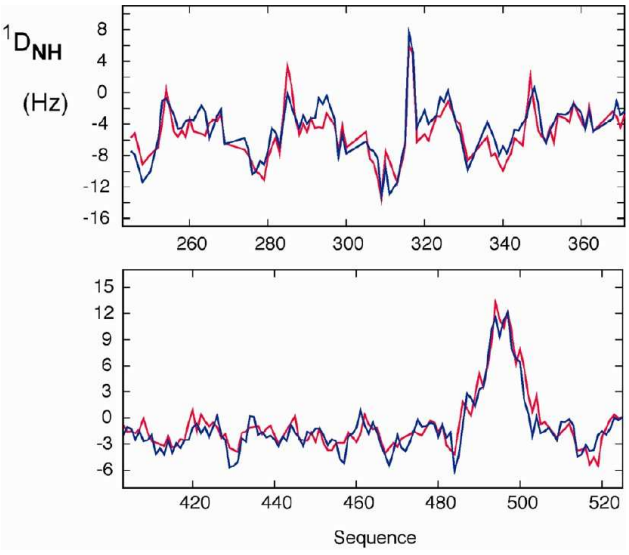


Figure 8. Cross validation of data not used in the ensemble selection procedure. Top: K18 fragment of Tau protein. Bottom: Disordered C-terminal domain, N_{TAIL}, of the nucleoprotein from measles virus. In both cases, $^{1}D_{N-H}$, back-calculated values from the ensemble selected against only $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N and $^{1}H^{N}$ CSs are shown.

experimental systems are essentially indistinguishable in terms of conformational sampling (data not shown), indicating that the analysis is robust at least with respect to the differences between these two prediction programs.

CONCLUSION

It is becoming increasingly clear that intrinsic disorder plays a central role in the function of a significant fraction of both eukaryotic and prokaryotic proteins. The development of an atomic resolution description of the conformational behavior of disordered proteins is a fundamental requirement if we are to understand their biological activity on a molecular level, and NMR represents potentially the most powerful source of this information. However, the actual resolution to which the amino acid specific potential energy surface can be mapped from experimental data remains obscure. Although dependences of some NMR parameters on structural propensities in disordered systems are known, so that sampling regimes are often inferred from experimental observations, there is currently no framework that allows for a statistical mapping of the available Ramachandran space of each amino acid in terms of conformational propensity. In this study, we address this question by combining highly efficient conformational sampling with ensemble selection to systematically investigate the ability of different sources of NMR data to map the backbone conformational sampling of IDPs on a residue specific level.

The results provide clear insight into conformational propensities that can be distinguished on the basis of experimentally available data. While backbone ^{13}C chemical shifts can be used to accurately determine the populations of βS and αR regions of Ramachandran space, clear degeneracies exist, in particular concerning the βP region, which is degenerate with average values predicted for random statistical coil sampling. This degeneracy can be raised by ^{15}N and $^{1}H^{N}$ shifts, although the prediction accuracy of these shifts is lower. Extending our analysis to commonly measured RDCs confirms the ability of this kind of measurement to distinguish between extended and helical bias, but also identifies a distinct degeneracy, this time between the βS and βP regions.

We demonstrate that a simple combination of RDCs and CSs raises inherent degeneracies to accurately resolve backbone conformational propensities. On the basis of these results, we propose a robust and generally applicable approach for the mapping of conformational potentials uniquely from experimental data, that is applied to two different biological systems.

dx.doi.org/10.1021/ja306905s | J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

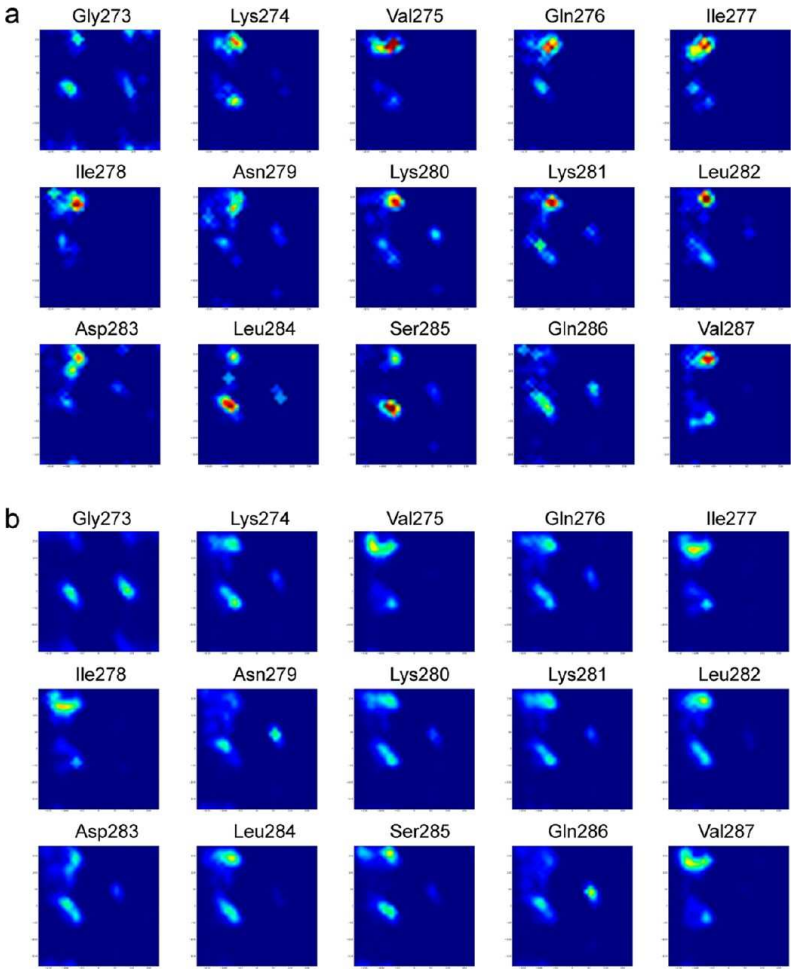


Figure 9. Ramachandran plots showing the amino acid specific conformational potentials in the 273−287 section of K18. (a) Selection from $^{1}D_{N-H}$, $^{13}C\alpha$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N and $^{1}H^{N}$ CSs using the ASTEROIDS approach for which the results are shown in Figure 7. (b) Conformational sampling from the statistical coil model. Dark blue represents lowest population, and red represents maximal sampling.

In both cases, we detect an increase of conformational sampling in the βP region compared to the standard statistical coil description, supporting previous experimental indications from vibrational spectroscopy and circular dichroism for the importance of this region in IDPs. Although the approach is amino acid specific, in many cases these regions are continuous, strongly suggesting that the observation is physically meaningful, but also suggesting that this is not simply a general feature, rather dependent on an underlying dependence on primary sequence. Using these approaches, a more extensive study of a broad range of experimentally available IDPs is currently underway in our laboratory, to determine whether general trends can be identified relating primary sequence composition to backbone conformational behavior.

More generally we are confident that the results from this study will pave the way to a more accurate understanding of the conformational propensities of disordered proteins in solution, and thereby provide hitherto inaccessible insight into the relationship between primary sequence and protein function in this fascinating family of proteins.

METHODS

Calculation of Average Chemical Shifts and RDCs in Ramachandran Space. The information content of the different chemical shifts was investigated by generating a 50 000-strong ensemble of poly alanine pentadecapeptide chains using the ensemble generation algorithm *flexible-meccano*.[50,51] For each conformer, the CSs were calculated using the prediction algorithm SPARTA,[25] and conformers were clustered into bins with a radius of 1° according to the $\{\phi,\psi\}$ values of the central amino acid (residue 8). The CSs within each cluster were then averaged and plotted against the $\{\phi,\psi\}$ value of the central amino acid.

Similarly, the information content of different types of RDCs was investigated. An ensemble consisting of 1 000 000 conformers of the poly alanine pentadecapeptide was created using *flexible-meccano*. RDCs were predicted using PALES[55] for each conformer and averaged in a similar way as described above for the CSs. The averaged RDCs of the central or neighboring amino acids were plotted against the $\{\phi,\psi\}$ sampling of the central amino acid.

Generation of Synthetic CS and RDC Data Sets in the Presence of Specific Conformational Sampling Regimes. To test the ability of different experimental CSs and RDCs to map conformational space, ensemble selections were carried out using ASTEROIDS targeting synthetic data sets. A model protein of 60 amino acids of arbitrary sequence was chosen sampling the statistical coil model except for three regions of five amino acids, where enhanced propensity was introduced in the αR (aa 10−14), βS (aa 27−31) or βP (aa 45−49) regions. Each propensity was introduced such that 50% of the selected conformers in the strand are specified to populate the region of interest, and the remaining 50% populate the statistical coil. An ensemble comprising 10 000 conformers of this model protein was generated using *flexible-meccano*, and CSs were predicted for each conformer using SPARTA. The CSs were

dx.doi.org/10.1021/ja306905s | J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

subsequently averaged over the ensemble and used as the target for the ASTEROIDS protocol.

To generate the synthetic RDC data set, an ensemble comprising 100 000 conformers of the same sequence was generated. A global alignment tensor was calculated for each conformer using an in-house written routine based on steric exclusion volume and the RDCs were calculated using this tensor. The RDCs were subsequently averaged over the ensemble and used as the target for the ASTEROIDS protocol.

To test the robustness of the ASTEROIDS protocol for mapping conformational space using CSs and RDCs, Gaussian-based noise was added to the synthetic CS and RDC data sets. The noise levels were based on the relative accuracy of SPARTA predictions for the different nuclei[25] and the predicted range of each dipolar coupling type. The following noise levels were applied: Cα (0.22 ppm), Cβ (0.24 ppm), C′ (0.25 ppm), N (0.6 ppm), H^N (0.12 ppm), $^1D_{N-H}$ (0.5 Hz), $^2D_{C'-HN}$ (0.25 Hz), $^1D_{C\alpha-H\alpha}$ (1 Hz) and $^1D_{C\alpha-C'}$ (0.25 Hz).

Ensemble Selections Using ASTEROIDS. Initially, a large pool of statistical coil conformers (20 000) was generated using *flexible-meccano*[50,51] and the genetic algorithm ASTEROIDS was used to select a subset of conformers in agreement with the experimental (or synthetic) data as described previously.[15] This procedure was repeated in an iterative manner in order to enhance the presence of conformational propensities of interest within the pool. Thus, in each step, a new pool was generated using the residue-specific $\{\phi, \psi\}$ angles derived from the selected ASTEROIDS ensembles in the previous iteration. Five independent ensemble selections comprising 200 conformers were carried out at each iteration step and iterations were continued until convergence. RDCs were calculated from a given member of an ensemble using the local alignment window (LAW) of 15 amino acids in length combined with a generic baseline as described previously.[15,36] The alignment tensor was calculated for each LAW using an in-house written routine based on steric alignment. A uniform scaling is applied to the entire predicted set to best-reproduce the experimental data. CSs were calculated for each structure using the program SPARTA and random coil values for calculation of secondary shifts were taken from RefDB.[27]

Experimental Data: C-Terminal Domain of Measles Virus Nucleoprotein. Experimental CSs of the intrinsically disordered C-terminal domain of Measles virus nucleoprotein were obtained previously at 25 °C in a buffer consisting of 50 mM sodium phophate at pH 6.5, 50 mM NaCl, 1 mM EDTA and 0.02% NaN.[53] $^1D_{N-H}$ RDCs were measured previously under the same conditions in a liquid crystal composed of poly ethylene glycol and 1-hexanol.[52]

Experimental Data: K18 Construct of Tau Protein. Experimental CSs of the K18 construct of Tau were obtained as described previously.[56] CS prediction using SPARTA relies on a database of 200 high-resolution structures for which nearly complete sets of chemical shift assignments are available. These CS assignments were obtained at temperatures above 20 °C with the vast majority lying between 20 and 30 °C. To avoid any bias, we calculated the CSs of K18 corresponding to 25 °C by comparing the 5 °C assignment of K18 to the 25 °C assignment of full-length Tau[57] and subsequently applying a uniform shift to each nucleus type independently. These new experimental data were used as the target for the ASTEROIDS protocol. $^1D_{N-H}$ RDCs of the K18 construct were measured previously in stretched poly acrylamide gels.[42]

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figures showing the reproduction of synthetic data from the fits shown in Figures 3−5. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

martin.blackledge@ibs.fr

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the Commissariat à l'energie atomique, the CNRS and the Université Joseph Fourier (Grenoble). This work was supported financially by the ANR under the following projects: ProteinDisorder (JCJC 2010), TAUSTRUCT (MALZ 2010) and by FINOVI.

■ REFERENCES

(1) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739−756.
(2) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradović, Z. *Biochemistry* **2002**, *41*, 6573−6582.
(3) Tompa, P. *Curr. Opin. Struct. Biol* **2011**, *21*, 419−425.
(4) Dyson, H. J.; Wright, P. E. *Chem. Rev* **2004**, *104*, 3607−3622.
(5) Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell Biol* **2005**, *6*, 197−208.
(6) Meier, S.; Blackledge, M.; Grzesiek, S. *J. Chem. Phys.* **2008**, *128*, 052204.
(7) Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol* **2007**, *17*, 3−14.
(8) Schneider, R.; Huang, J.; Yao, M.; Communie, G.; Ozenne, V.; Mollica, L.; Salmon, L.; Jensen, M. R.; Blackledge, M. *Mol. BioSyst.* **2012**, *8*, 58−68.
(9) Wright, P. E.; Dyson, H. J. *Curr. Opin. Struct. Biol* **2009**, *19*, 31−38.
(10) Tompa, P.; Fuxreiter, M. *Trends Biochem. Sci.* **2008**, *33*, 2−8.
(11) Smith, L. J.; Bolin, K. A.; Schwalbe, H.; MacArthur, M. W.; Thornton, J. M.; Dobson, C. M. *J. Mol. Biol.* **1996**, *255*, 494−506.
(12) Lindorff-Larsen, K.; Kristjansdottir, S.; Teilum, K.; Fieber, W.; Dobson, C.; Poulsen, F.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 3291−3299.
(13) Kristjansdottir, S.; Lindorff-Larsen, K.; Fieber, W.; Dobson, C. M.; Vendruscolo, M.; Poulsen, F. M. *J. Mol. Biol.* **2005**, *347*, 1053−1062.
(14) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359−374.
(15) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 17908−17918.
(16) Jensen, M. R.; Markwick, P. R. L.; Meier, S.; Griesinger, C.; Zweckstetter, M.; Grzesiek, S.; Bernadó, P.; Blackledge, M. *Structure* **2009**, *17*, 1169−1185.
(17) Bernadó, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. *J. Am. Chem. Soc.* **2007**, *129*, 5656−5664.
(18) Esteban-Martín, S.; Fenwick, R. B.; Salvatella, X. *J. Am. Chem. Soc.* **2010**, *132*, 4626−4632.
(19) Huang, J.; Grzesiek, S. *J. Am. Chem. Soc.* **2010**, *132*, 694−705.
(20) Wishart, D. S.; Sykes, B. D. *J. Biomol. NMR* **1994**, *4*, 171−180.
(21) Schwarzinger, S.; Kroon, G. J.; Foss, T. R.; Chung, J.; Wright, P. E.; Dyson, H. J. *J. Am. Chem. Soc.* **2001**, *123*, 2970−2978.
(22) Wang, Y.; Jardetzky, O. *J. Am. Chem. Soc.* **2002**, *124*, 14075−14084.
(23) Osapay, K.; Case, D. A. *J. Biomol. NMR* **1994**, *4*, 215−230.
(24) Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S. *J. Biomol. NMR* **2003**, *26*, 215−240.
(25) Shen, Y.; Bax, A. *J. Biomol. NMR* **2007**, *38*, 289−302.
(26) Yao, J.; Chung, J.; Eliezer, D.; Wright, P. E.; Dyson, H. J. *Biochemistry* **2001**, *40*, 3561−3571.
(27) Zhang, H.; Neal, S.; Wishart, D. S. *J. Biomol. NMR* **2003**, *25*, 173−195.
(28) Cornilescu, G.; Delaglio, F.; Bax, A. *J. Biomol. NMR* **1999**, *13*, 289−302.
(29) Cavalli, A.; Salvatella, X.; Dobson, C.; Vendruscolo, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 9615−9620.
(30) Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J.; Liu, G.; Eletsky, A.; Wu, Y.; Singarapu, K.; Lemak, A.; Ignatchenko, A.;

602 Arrowsmith, C.; Szyperski, T.; Montelione, G.; Baker, D.; Bax, A. *Proc.*
603 *Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4685−4690.
604 (31) Berjanskii, M.; Tang, P.; Liang, J.; Cruz, J. A.; Zhou, J.; Zhou, Y.;
605 Bassett, E.; MacDonell, C.; Lu, P.; Lin, G.; Wishart, D. S. *Nucleic Acids*
606 *Res.* **2009**, *37*, W670−677.
607 (32) De Simone, A.; Cavalli, A.; Hsu, S.-T. D.; Vranken, W.;
608 Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 16332−16333.
609 (33) Tamiola, K.; Acar, B.; Mulder, F. A. A. *J. Am. Chem. Soc.* **2010**,
610 *132*, 18000−18003.
611 (34) Marsh, J. A.; Singh, V. K.; Jia, Z.; Forman-Kay, J. D. *Protein Sci.*
612 **2006**, *15*, 2795−2804.
613 (35) Jensen, M. R.; Salmon, L.; Nodet, G.; Blackledge, M. *J. Am.*
614 *Chem. Soc.* **2010**, *132*, 1270−1272.
615 (36) Salmon, L.; Nodet, G.; Ozenne, V.; Yin, G.; Jensen, M.;
616 Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2010**, *132*, 8407−
617 8418.
618 (37) Camilloni, C.; De Simone, A.; Vranken, W. F.; Vendruscolo, M.
619 *Biochemistry* **2012**, *51*, 2224−2231.
620 (38) Mohana-Borges, R.; Goto, N. K.; Kroon, G. J. A.; Dyson, H. J.;
621 Wright, P. E. *J. Mol. Biol.* **2004**, *340*, 1131−1142.
622 (39) Meier, S.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2007**,
623 *129*, 9799−9807.
624 (40) Obolensky, O. I.; Schlepckow, K.; Schwalbe, H.; Solov'yov, A. V.
625 *J. Biomol. NMR* **2007**, *39*, 1−16.
626 (41) Louhivuori, M.; Pääkkönen, K.; Fredriksson, K.; Permi, P.;
627 Lounila, J.; Annila, A. *J. Am. Chem. Soc.* **2003**, *125*, 15647−15650.
628 (42) Mukrasch, M. D.; Markwick, P.; Biernat, J.; Bergen, M.; von;
629 Bernadó, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M.;
630 Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 5235−5243.
631 (43) Jensen, M. R.; Houben, K.; Lescop, E.; Blanchard, L.; Ruigrok,
632 R. W. H.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 8055−8061.
633 (44) Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*,
634 11266−11267.
635 (45) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen,
636 M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. *Proc.*
637 *Natl. Acad. Sci. U.S.A.* **2008**, *105*, 5762−5767.
638 (46) Huang, J.; Gabel, F.; Jensen, M. R.; Grzesiek, S.; Blackledge, M.
639 *J. Am. Chem. Soc.* **2012**, *134*, 4429−4436.
640 (47) Shi, Z.; Chen, K.; Liu, Z.; Kallenbach, N. R. *Chem. Rev.* **2006**,
641 *106*, 1877−1897.
642 (48) Maiti, N. C.; Apetri, M. M.; Zagorski, M. G.; Carey, P. R.;
643 Anderson, V. E. *J. Am. Chem. Soc.* **2004**, *126*, 2399−2408.
644 (49) Woody, R. W. *J. Am. Chem. Soc.* **2009**, *131*, 8234−8245.
645 (50) Bernadó, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok,
646 R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*,
647 17002−17007.
648 (51) Ozenne, V.; Bauer, F.; Salmon, L.; Huang, J.-R.; Jensen, M. R.;
649 Segard, S.; Bernadó, P.; Charavay, C.; Blackledge, M. *Bioinformatics*
650 **2012**, *28*, 1463−1470.
651 (52) Jensen, M. R.; Communie, G.; Ribeiro, E. A., Jr; Martinez, N.;
652 Desfosses, A.; Salmon, L.; Mollica, L.; Gabel, F.; Jamin, M.; Longhi, S.;
653 Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2011**,
654 *108*, 9839−9844.
655 (53) Gely, S.; Lowry, D. F.; Bernard, C.; Jensen, M. R.; Blackledge,
656 M.; Costanzo, S.; Bourhis, J.-M.; Darbon, H.; Daughdrill, G.; Longhi,
657 S. *J. Mol. Recognit.* **2010**, *23*, 435−447.
658 (54) Shen, Y.; Bax, A. *J. Biomol. NMR* **2010**, *48*, 13−22.
659 (55) Zweckstetter, M. *Nat. Protoc.* **2008**, *3*, 679−690.
660 (56) Mukrasch, M. D.; Biernat, J.; von Bergen, M.; Griesinger, C.;
661 Mandelkow, E.; Zweckstetter, M. *J. Biol. Chem.* **2005**, *280*, 24978−
662 24986.
663 (57) Mukrasch, M. D.; Bibow, S.; Korukottu, J.; Jeganathan, S.;
664 Biernat, J.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. *PLoS Biol.*
665 **2009**, *7*, e34.

Bibliographie

[1] I. D. Kuntz. Structure-based strategies for drug design and discovery. *Science*, 257 (5073) :1078–1082, Aug 1992. (Cité page 2.)

[2] C. Boesch, A. Bundi, M. Oppliger, and K. Wüthrich. 1h nuclear-magnetic-resonance studies of the molecular conformation of monomeric glucagon in aqueous solution. *Eur J Biochem*, 91(1) :209–214, Nov 1978. (Cité page 2.)

[3] M. P. Williamson, T. F. Havel, and K. Wüthrich. Solution conformation of proteinase inhibitor iia from bull seminal plasma by 1h nuclear magnetic resonance and distance geometry. *J Mol Biol*, 182(2) :295–315, Mar 1985. (Cité page 2.)

[4] Bin Xue, Robert W. Williams, Christopher J. Oldfield, A Keith Dunker, and Vladimir N. Uversky. Archaic chaos : intrinsically disordered proteins in archaea. *BMC Syst Biol*, 4 Suppl 1 :S1, 2010. (Cité pages 2 et 13.)

[5] Megan Sickmeier, Justin A. Hamilton, Tanguy LeGall, Vladimir Vacic, Marc S. Cortese, Agnes Tantos, Beata Szabo, Peter Tompa, Jake Chen, Vladimir N. Uversky, Zoran Obradovic, and A Keith Dunker. Disprot : the database of disordered proteins. *Nucleic Acids Res*, 35(Database issue) :D786–D793, Jan 2007. (Cité page 2.)

[6] Helen M. Berman, Tammy Battistuz, T. N. Bhat, Wolfgang F. Bluhm, Philip E. Bourne, Kyle Burkhardt, Zukang Feng, Gary L. Gilliland, Lisa Iype, Shri Jain, Phoebe Fagan, Jessica Marvin, David Padilla, Veerasamy Ravichandran, Bohdan Schneider, Narmada Thanki, Helge Weissig, John D. Westbrook, and Christine Zardecki. The protein data bank. *Acta Crystallogr D Biol Crystallogr*, 58(Pt 6 No 1) : 899–907, Jun 2002. (Cité page 2.)

[7] H Jane Dyson and Peter E. Wright. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, 6(3) :197–208, Mar 2005. (Cité pages 5 et 11.)

[8] Vladimir N. Uversky and A Keith Dunker. Understanding protein non-folding. *Biochim Biophys Acta*, 1804(6) :1231–1264, Jun 2010. (Cité page 5.)

[9] Zoran Obradovic, Kang Peng, Slobodan Vucetic, Predrag Radivojac, Celeste J. Brown, and A Keith Dunker. Predicting intrinsic disorder from amino acid sequence. *Proteins*, 53 Suppl 6 :566–572, 2003. (Cité page 5.)

[10] M Madan Babu, Robin van der Lee, Natalia Sanchez de Groot, and Jörg Gsponer. Intrinsically disordered proteins : regulation and disease. *Curr Opin Struct Biol*, 21 (3) :432–440, Jun 2011. (Cité page 7.)

[11] Vladimir Vacic, Christopher J. Oldfield, Amrita Mohan, Predrag Radivojac, Marc S. Cortese, Vladimir N. Uversky, and A Keith Dunker. Characterization of molecular recognition features, morfs, and their binding partners. *J Proteome Res*, 6(6) :2351–2366, Jun 2007. (Cité pages 7 et 12.)

[12] Peter E. Wright and H Jane Dyson. Linking folding and binding. *Curr Opin Struct Biol*, 19(1) :31–38, Feb 2009. (Cité page 7.)

[13] Yuefeng Wang, John C. Fisher, Rose Mathew, Li Ou, Steve Otieno, Jack Sublet, Limin Xiao, Jianhan Chen, Martine F. Roussel, and Richard W. Kriwacki. Intrinsic disorder mediates the diverse regulatory functions of the cdk inhibitor p21. *Nat Chem Biol*, 7(4) :214–221, Apr 2011. (Cité page 7.)

[14] Lilia M. Iakoucheva, Celeste J. Brown, J David Lawson, Zoran Obradovi ?, and A Keith Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol*, 323(3) :573–584, Oct 2002. (Cité page 7.)

[15] Mark Wells, Henning Tidow, Trevor J. Rutherford, Phineus Markwick, Malene Ringkjobing Jensen, Efstratios Mylonas, Dmitri I. Svergun, Martin Blackledge, and Alan R. Fersht. Structure of tumor suppressor p53 and its intrinsically disordered n-terminal transactivation domain. *Proc Natl Acad Sci U S A*, 105(15) : 5762–5767, Apr 2008. (Cité pages 7 et 92.)

[16] E. M. Mandelkow and E. Mandelkow. Tau in alzheimer's disease. *Trends Cell Biol*, 8(11) :425–427, Nov 1998. (Cité pages 7 et 130.)

[17] Eva-Maria Mandelkow and Eckhard Mandelkow. Biochemistry and cell biology of tau protein in neurofibrillary degeneration. *Cold Spring Harb Perspect Med*, 2(7) : a006247, Jul 2012. (Cité page 7.)

[18] Robert Bussell, Jr and David Eliezer. Effects of parkinson's disease-linked mutations on the structure of lipid-associated alpha-synuclein. *Biochemistry*, 43(16) : 4810–4818, Apr 2004. (Cité page 7.)

[19] Vladimir N. Uversky, Christopher J. Oldfield, and A Keith Dunker. Intrinsically disordered proteins in human diseases : introducing the d2 concept. *Annu Rev Biophys*, 37 :215–246, 2008. (Cité page 7.)

[20] Yugong Cheng, Tanguy LeGall, Christopher J. Oldfield, A Keith Dunker, and Vladimir N. Uversky. Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry*, 45(35) :10448–10460, Sep 2006. (Cité page 7.)

[21] Mark R. Cookson. alpha-synuclein and neuronal cell death. *Mol Neurodegener*, 4 : 9, 2009. (Cité page 7.)

[22] E. M. Mandelkow and E. Mandelkow. Tau as a marker for alzheimer's disease. *Trends Biochem Sci*, 18(12) :480–483, Dec 1993. (Cité page 7.)

[23] Diane P. Hanger, Brian H. Anderton, and Wendy Noble. Tau phosphorylation : the therapeutic challenge for neurodegenerative disease. *Trends Mol Med*, 15(3) : 112–119, Mar 2009. (Cité pages 7 et 131.)

[24] Dorthe Matenia and Eva-Maria Mandelkow. The tau of mark : a polarized view of the cytoskeleton. *Trends Biochem Sci*, 34(7) :332–342, Jul 2009. (Cité pages 7 et 131.)

[25] Peter Tompa and Monika Fuxreiter. Fuzzy complexes : polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci*, 33(1) :2–8, Jan 2008. (Cité page 7.)

[26] Peter Tsvetkov, Nina Reuven, and Yosef Shaul. The nanny model for idps. *Nat Chem Biol*, 5(11) :778–781, Nov 2009. (Cité page 7.)

[27] Philip M. Kim, Andrea Sboner, Yu Xia, and Mark Gerstein. The role of disorder in interaction networks : a structural analysis. *Mol Syst Biol*, 4 :179, 2008. (Cité page 7.)

[28] Kana Shimizu and Hiroyuki Toh. Interaction between intrinsically disordered proteins frequently occurs in a human protein-protein interaction network. *J Mol Biol*, 392(5) :1253–1265, Oct 2009. (Cité page 7.)

[29] Hyoung-Gon Lee, George Perry, Paula I. Moreira, Matthew R. Garrett, Quan Liu, Xiongwei Zhu, Atsushi Takeda, Akihiko Nunomura, and Mark A. Smith. Tau phosphorylation in alzheimer's disease : pathogen or protector ? *Trends Mol Med*, 11(4) :164–169, Apr 2005. (Cité pages 7 et 131.)

[30] H Jane Dyson and Peter E. Wright. Elucidation of the protein folding landscape by nmr. *Methods Enzymol*, 394 :299–321, 2005. (Cité page 8.)

[31] Guillaume Bouvignies, Pramodh Vallurupalli, D Flemming Hansen, Bruno E. Correia, Oliver Lange, Alaji Bah, Robert M. Vernon, Frederick W. Dahlquist, David Baker, and Lewis E. Kay. Solution structure of a minor and transiently formed state of a t4 lysozyme mutant. *Nature*, 477(7362) :111–114, Sep 2011. (Cité page 8.)

[32] L. E. Kay, D. A. Torchia, and A. Bax. Backbone dynamics of proteins as studied by 15n inverse detected heteronuclear nmr spectroscopy : application to staphylococcal nuclease. *Biochemistry*, 28(23) :8972–8979, Nov 1989. (Cité page 8.)

[33] Loïc Salmon, Guillaume Bouvignies, Phineus Markwick, Nils Lakomek, Scott Showalter, Da-Wei Li, Korvin Walter, Christian Griesinger, Rafael Brüschweiler, and Martin Blackledge. Protein conformational flexibility from structure-free analysis of nmr dipolar couplings : quantitative and absolute determination of backbone motion in ubiquitin. *Angew Chem Int Ed Engl*, 48(23) :4154–4157, 2009. (Cité page 8.)

[34] Christopher J. Oldfield, Jingwei Meng, Jack Y. Yang, Mary Qu Yang, Vladimir N. Uversky, and A Keith Dunker. Flexible nets : disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics*, 9 Suppl 1 :S1, 2008. (Cité page 9.)

[35] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins : re-assessing the protein structure-function paradigm. *J Mol Biol*, 293(2) :321–331, Oct 1999. (Cité page 9.)

[36] Vladimir N. Uversky and A Keith Dunker. Multiparametric analysis of intrinsically disordered proteins : looking at intrinsic disorder through compound eyes. *Anal Chem*, 84(5) :2096–2104, Mar 2012. (Cité page 9.)

[37] Malene Ringkjøbing Jensen, Pau Bernadó, Klaartje Houben, Laurence Blanchard, Dominque Marion, Rob W H. Ruigrok, and Martin Blackledge. Structural disorder within sendai virus nucleoprotein and phosphoprotein : insight into the structural basis of molecular recognition. *Protein Pept Lett*, 17(8) :952–960, Aug 2010. (Cité page 9.)

[38] Thomas Kiefhaber, Annett Bachmann, and Kristine Steen Jensen. Dynamics and mechanisms of coupled protein folding and binding reactions. *Curr Opin Struct Biol*, 22(1) :21–29, Feb 2012. (Cité page 10.)

[39] Nasrollah Rezaei-Ghaleh, Martin Blackledge, and Markus Zweckstetter. Intrinsically disordered proteins : from sequence and conformational properties toward drug discovery. *Chembiochem*, 13(7) :930–950, May 2012. (Cité page 11.)

[40] Predrag Radivojac, Lilia M. Iakoucheva, Christopher J. Oldfield, Zoran Obradovic, Vladimir N. Uversky, and A Keith Dunker. Intrinsic disorder and functional proteomics. *Biophys J*, 92(5) :1439–1456, Mar 2007. (Cité page 11.)

[41] A Keith Dunker, Christopher J. Oldfield, Jingwei Meng, Pedro Romero, Jack Y. Yang, Jessica Walton Chen, Vladimir Vacic, Zoran Obradovic, and Vladimir N. Uversky. The unfoldomics decade : an update on intrinsically disordered proteins. *BMC Genomics*, 9 Suppl 2 :S1, 2008. (Cité page 12.)

[42] Bálint Mészáros, István Simon, and Zsuzsanna Dosztányi. Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol*, 5(5) :e1000376, May 2009. (Cité page 12.)

[43] Lilia M. Iakoucheva, Predrag Radivojac, Celeste J. Brown, Timothy R. O'Connor, Jason G. Sikes, Zoran Obradovic, and A Keith Dunker. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*, 32(3) :1037–1049, 2004. (Cité page 12.)

[44] Hongbo Xie, Slobodan Vucetic, Lilia M. Iakoucheva, Christopher J. Oldfield, A Keith Dunker, Zoran Obradovic, and Vladimir N. Uversky. Functional anthology of intrinsic disorder. 3. ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J Proteome Res*, 6(5) :1917–1932, May 2007. (Cité page 12.)

[45] Rune Linding, Robert B. Russell, Victor Neduva, and Toby J. Gibson. Globplot : Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*, 31 (13) :3701–3708, Jul 2003. (Cité page 12.)

[46] Jaime Prilusky, Clifford E. Felder, Tzviya Zeev-Ben-Mordehai, Edwin H. Rydberg, Orna Man, Jacques S. Beckmann, Israel Silman, and Joel L. Sussman. Foldindex : a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, 21(16) :3435–3438, Aug 2005. (Cité page 12.)

[47] Zoran Obradovic, Kang Peng, Slobodan Vucetic, Predrag Radivojac, and A Keith Dunker. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, 61 Suppl 7 :176–182, 2005. (Cité page 12.)

[48] Zsuzsanna Dosztányi, Veronika Csizmók, Péter Tompa, and István Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol*, 347(4) :827–839, Apr 2005. (Cité page 12.)

[49] Rune Linding, Lars Juhl Jensen, Francesca Diella, Peer Bork, Toby J. Gibson, and Robert B. Russell. Protein disorder prediction : implications for structural proteomics. *Structure*, 11(11) :1453–1459, Nov 2003. (Cité page 12.)

[50] Takashi Ishida and Kengo Kinoshita. Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics*, 24(11) :1344–1348, Jun 2008. (Cité page 13.)

[51] Adam L. Rucker, Cara T. Pager, Margaret N. Campbell, Joseph E. Qualls, and Trevor P. Creamer. Host-guest scale of left-handed polyproline ii helix formation. *Proteins*, 53(1) :68–75, Oct 2003. (Cité pages 13 et 14.)

[52] E. W. Blanch, L. A. Morozova-Roche, D. A. Cochran, A. J. Doig, L. Hecht, and L. D. Barron. Is polyproline ii helix the killer conformation ? a raman optical activity study of the amyloidogenic prefibrillar intermediate of human lysozyme. *J Mol Biol*, 301(2) :553–563, Aug 2000. (Cité page 14.)

[53] Fatma Eker, Kai Griebenow, and Reinhard Schweitzer-Stenner. Abeta(1-28) fragment of the amyloid peptide predominantly adopts a polyproline ii conformation in an acidic solution. *Biochemistry*, 43(22) :6893–6898, Jun 2004. (Cité page 14.)

[54] Zhengshuang Shi, C Anders Olson, George D. Rose, Robert L. Baldwin, and Neville R. Kallenbach. Polyproline ii structure in a sequence of seven alanine residues. *Proc Natl Acad Sci U S A*, 99(14) :9190–9195, Jul 2002. (Cité page 14.)

[55] S. Williams, T. P. Causgrove, R. Gilmanshin, K. S. Fang, R. H. Callender, W. H. Woodruff, and R. B. Dyer. Fast events in protein folding : helix melting and formation in a small peptide. *Biochemistry*, 35(3) :691–697, Jan 1996. (Cité page 15.)

[56] R. Gilmanshin, S. Williams, R. H. Callender, W. H. Woodruff, and R. B. Dyer. Fast events in protein folding : relaxation dynamics of secondary and tertiary structure in native apomyoglobin. *Proc Natl Acad Sci U S A*, 94(8) :3709–3713, Apr 1997. (Cité page 15.)

[57] Stefan Bibow, Marco D. Mukrasch, Subashchandrabose Chinnathambi, Jacek Biernat, Christian Griesinger, Eckhard Mandelkow, and Markus Zweckstetter. The dynamic structure of filamentous tau. *Angew Chem Int Ed Engl*, 50(48) :11520–11524, Nov 2011. (Cité page 17.)

[58] Venita Daebel, Subashchandrabose Chinnathambi, Jacek Biernat, Martin Schwalbe, Birgit Habenstein, Antoine Loquet, Elias Akoury, Katharina Tepper, Henrik Müller, Marc Baldus, Christian Griesinger, Markus Zweckstetter, Eckhard Mandelkow, Vinesh Vijayan, and Adam Lange. ?-sheet core of tau paired helical filaments revealed by solid-state nmr. *J Am Chem Soc*, Aug 2012. (Cité page 17.)

[59] Ivano Bertini, Leonardo Gonnelli, Claudio Luchinat, Jiafei Mao, and Antonella Nesi. A new structural model of a ?40 fibrils. *J Am Chem Soc*, 133(40) :16013–16022, Oct 2011. (Cité pages 17 et 130.)

[60] Robert Tycko. Solid-state nmr studies of amyloid fibril structure. *Annu Rev Phys Chem*, 62 :279–299, 2011. (Cité pages 17 et 130.)

[61] Guohua Lv, Ashutosh Kumar, Karin Giller, Maria L. Orcellet, Dietmar Riedel, Claudio O. Fernández, Stefan Becker, and Adam Lange. Structural comparison of mouse and human ?-synuclein amyloid fibrils by solid-state nmr. *J Mol Biol*, 420 (1-2) :99–111, Jun 2012. (Cité page 17.)

[62] Anthony K. Mittermaier and Lewis E. Kay. Observing biological dynamics at atomic resolution using nmr. *Trends Biochem Sci*, 34(12) :601–611, Dec 2009. (Cité page 21.)

[63] Lewis E. Kay. Nmr studies of protein structure and dynamics - a look backwards and forwards. *J Magn Reson*, 213(2) :492–494, Dec 2011. (Cité page 21.)

[64] J. Balbach, V. Forge, N. A. van Nuland, S. L. Winder, P. J. Hore, and C. M. Dobson. Following protein folding in real time using nmr spectroscopy. *Nat Struct Biol*, 2 (10) :865–870, Oct 1995. (Cité page 21.)

[65] Loïc Salmon, José-Luis Ortega Roldan, Ewen Lescop, Antoine Licinio, Nico van Nuland, Malene Ringkjøbing Jensen, and Martin Blackledge. Structure, dynamics, and kinetics of weak protein-protein complexes from nmr spin relaxation measurements of titrated solutions. *Angew Chem Int Ed Engl*, 50(16) :3755–3759, Apr 2011. (Cité page 21.)

[66] G Marius Clore and Junji Iwahara. Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. *Chem Rev*, 109(9) :4108–4139, Sep 2009. (Cité pages 21 et 34.)

[67] Pramodh Vallurupalli, D Flemming Hansen, and Lewis E. Kay. Structures of invisible, excited protein states by relaxation dispersion nmr spectroscopy. *Proc Natl Acad Sci U S A*, 105(33) :11766–11771, Aug 2008. (Cité page 22.)

[68] Magnus Kjaergaard, Søren Brander, and Flemming M. Poulsen. Random coil chemical shift for intrinsically disordered proteins : effects of temperature and ph. *J Biomol NMR*, 49(2) :139–149, Feb 2011. (Cité page 23.)

[69] Magnus Kjaergaard and Flemming M. Poulsen. Disordered proteins studied by chemical shifts. *Prog Nucl Magn Reson Spectrosc*, 60 :42–51, Jan 2012. (Cité page 23.)

[70] Haiyan Zhang, Stephen Neal, and David S. Wishart. Refdb : a database of uniformly referenced protein chemical shifts. *J Biomol NMR*, 25(3) :173–195, Mar 2003. (Cité pages 23 et 53.)

[71] A. Pardi, M. Billeter, and K. Wüthrich. Calibration of the angular dependence of the amide proton-c alpha proton coupling constants, 3jhn alpha, in a globular protein. use of 3jhn alpha for identification of helical secondary structure. *J Mol Biol*, 180(3) :741–751, Dec 1984. (Cité page 25.)

[72] G. W. Vuister and A. Bax. Measurement of four-bond hn-h alpha j-couplings in staphylococcal nuclease. *J Biomol NMR*, 4(2) :193–200, Mar 1994. (Cité page 25.)

[73] L. J. Smith, K. M. Fiebig, H. Schwalbe, and C. M. Dobson. The concept of a random coil. residual structure in peptides and denatured proteins. *Fold Des*, 1(5) :R95–106, 1996. (Cité pages 25 et 41.)

[74] Marco D. Mukrasch, Phineus Markwick, Jacek Biernat, Martin von Bergen, Pau Bernadó, Christian Griesinger, Eckhard Mandelkow, Markus Zweckstetter, and Martin Blackledge. Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. *J Am Chem Soc*, 129(16) :5235–5243, Apr 2007. (Cité pages 25, 60, 84, 136 et 147.)

[75] J. H. Prestegard, H. M. al Hashimi, and J. R. Tolman. Nmr structures of biomolecules using field oriented media and residual dipolar couplings. *Q Rev Biophys*, 33 (4) :371–424, Nov 2000. (Cité page 27.)

[76] N. Tjandra, J. G. Omichinski, A. M. Gronenborn, G. M. Clore, and A. Bax. Use of dipolar 1h-15n and 1h-13c couplings in the structure determination of magnetically oriented macromolecules in solution. *Nat Struct Biol*, 4(9) :732–738, Sep 1997. (Cité page 27.)

[77] Loïc Salmon, Malene Ringkjøbing Jensen, Pau Bernadó, and Martin Blackledge. Measurement and analysis of nmr residual dipolar couplings for the study of intrinsically disordered proteins. *Methods Mol Biol*, 895 :115–125, 2012. (Cité page 27.)

[78] Ronaldo Mohana-Borges, Natalie K. Goto, Gerard J A. Kroon, H Jane Dyson, and Peter E. Wright. Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. *J Mol Biol*, 340(5) :1131–1142, Jul 2004. (Cité pages 28, 30 et 92.)

[79] Y. Ishii, M. A. Markus, and R. Tycko. Controlling residual dipolar couplings in high-resolution nmr of proteins by strain induced alignment in a gel. *J Biomol NMR*, 21(2) :141–151, Oct 2001. (Cité page 28.)

[80] GM Clore, MR Starich, and AM Gronenborn. Measurement of residual dipolar couplings of macromolecules aligned in the nematic phase of a colloidal suspension of rod-shaped viruses. *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, 120(40) :10571–10572, OCT 14 1998. ISSN 0002-7863. (Cité page 28.)

[81] GM Clore, MR Starich, CA Bewley, ML Cai, and J Kuszewski. Impact of residual dipolar couplings on the accuracy of NMR structures determined from a minimal number of NOE restraints. *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, 121(27) :6513–6514, JUL 14 1999. ISSN 0002-7863. (Cité page 28.)

[82] Markus Zweckstetter. Nmr : prediction of molecular alignment from structure using the pales software. *Nat Protoc*, 3(4) :679–690, 2008. (Cité pages 29, 51, 66 et 82.)

[83] Guillaume Bouvignies, Sebastian Meier, Stephan Grzesiek, and Martin Blackledge. Ultrahigh-resolution backbone structure of perdeuterated protein gb1 using residual dipolar couplings from two alignment media. *Angew Chem Int Ed Engl*, 45 (48) :8166–8169, Dec 2006. (Cité page 29.)

[84] Loïc Salmon, Levi Pierce, Alexander Grimm, Jose-Luis Ortega Roldan, Luca Mollica, Malene Ringkjøbing Jensen, Nico van Nuland, Phineus R L. Markwick, J Andrew McCammon, and Martin Blackledge. Inside back cover : Multi-timescale conformational dynamics of the sh3 domain of cd2-associated protein using nmr spectroscopy and accelerated molecular dynamics (angew. chem. int. ed. 25/2012). *Angew Chem Int Ed Engl*, 51(25) :6279, Jun 2012. (Cité page 29.)

[85] Martti Louhivuori, Kimmo Pääkkönen, Kai Fredriksson, Perttu Permi, Juhani Lounila, and Arto Annila. On the origin of residual dipolar couplings from denatured proteins. *J Am Chem Soc*, 125(50) :15647–15650, Dec 2003. (Cité pages 30 et 69.)

[86] O. I. Obolensky, Kai Schlepckow, Harald Schwalbe, and A. V. Solov'yov. Theoretical framework for nmr residual dipolar couplings in unfolded proteins. *J Biomol NMR*, 39(1) :1–16, Sep 2007. (Cité pages 30 et 69.)

[87] Wolfgang Fieber, Sigridur Kristjansdottir, and Flemming M. Poulsen. Short-range, long-range and transition state interactions in the denatured state of acbp from residual dipolar couplings. *J Mol Biol*, 339(5) :1191–1199, Jun 2004. (Cité pages 30 et 92.)

[88] A Abragam. The principles of nuclear magnetism. *Oxford University Press*, page 599, Jan 1989. (Cité page 31.)

[89] G Lipari and A Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. theory and range of validity. *J Am Chem Soc*, 104(17) :4546–4559, 1982. (Cité page 32.)

[90] J. R. Gillespie and D. Shortle. Characterization of long-range structure in the denatured state of staphylococcal nuclease. ii. distance restraints from paramagnetic relaxation and calculation of an ensemble of structures. *J Mol Biol*, 268(1) :170–184, Apr 1997. (Cité pages 33 et 105.)

[91] J. R. Gillespie and D. Shortle. Characterization of long-range structure in the denatured state of staphylococcal nuclease. i. paramagnetic relaxation enhancement by nitroxide spin labels. *J Mol Biol*, 268(1) :158–169, Apr 1997. (Cité pages 33 et 105.)

[92] Yi Xue, Ivan S. Podkorytov, D Krishna Rao, Nathan Benjamin, Honglei Sun, and Nikolai R. Skrynnikov. Paramagnetic relaxation enhancements in unfolded proteins : theory and application to drkn sh3 domain. *Protein Sci*, 18(7) :1401–1424, Jul 2009. (Cité page 33.)

[93] Yi Xue and Nikolai R. Skrynnikov. Motion of a disordered polypeptide chain as studied by paramagnetic relaxation enhancements, 15n relaxation, and molecular dynamics simulations : how fast is segmental diffusion in denatured ubiquitin ? *J Am Chem Soc*, 133(37) :14614–14628, Sep 2011. (Cité pages 33 et 34.)

[94] Junji Iwahara, Charles D. Schwieters, and G Marius Clore. Ensemble approach for nmr structure refinement against (1)h paramagnetic relaxation enhancement data arising from a flexible paramagnetic group attached to a macromolecule. *J Am Chem Soc*, 126(18) :5879–5896, May 2004. (Cité page 33.)

[95] Bertil Halle. The physical basis of model-free analysis of nmr relaxation data from proteins and complex fluids. *J Chem Phys*, 131(22) :224507, Dec 2009. (Cité page 34.)

[96] Friederike Sziegat, Robert Silvers, Martin Hähnke, Malene Ringkjøbing Jensen, Martin Blackledge, Julia Wirmer-Bartoschek, and Harald Schwalbe. Disentangling the coil : Modulation of conformational and dynamic properties by site-directed mutation in the non-native state of hen egg white lysozyme. *Biochemistry*, Apr 2012. (Cité page 34.)

[97] Deniz Sezer, Jack H. Freed, and Benoît Roux. Parametrization, molecular dynamics simulation, and calculation of electron spin resonance spectra of a nitroxide spin label on a polyalanine alpha-helix. *J Phys Chem B*, 112(18) :5755–5767, May 2008. (Cité page 34.)

[98] G Marius Clore, Chun Tang, and Junji Iwahara. Elucidating transient macromolecular interactions using paramagnetic relaxation enhancement. *Curr Opin Struct Biol*, 17(5) :603–616, Oct 2007. (Cité page 34.)

[99] Charles K. Fisher and Collin M. Stultz. Constructing ensembles for intrinsically disordered proteins. *Curr Opin Struct Biol*, 21(3) :426–431, Jun 2011. (Cité pages 39 et 46.)

[100] Jonathan E. Kohn, Ian S. Millett, Jaby Jacob, Bojan Zagrovic, Thomas M. Dillon, Nikolina Cingel, Robin S. Dothager, Soenke Seifert, P. Thiyagarajan, Tobin R. Sosnick, M Zahid Hasan, Vijay S. Pande, Ingo Ruczinski, Sebastian Doniach, and Kevin W. Plaxco. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc Natl Acad Sci U S A*, 101(34) :12491–12496, Aug 2004. (Cité page 40.)

[101] Pau Bernadó and Martin Blackledge. A self-consistent description of the conformational behavior of chemically denatured proteins from nmr and small angle scattering. *Biophys J*, 97(10) :2839–2845, Nov 2009. (Cité page 40.)

[102] Abhishek K. Jha, Andres Colubri, Muhammad H. Zaman, Shohei Koide, Tobin R. Sosnick, and Karl F. Freed. Helix, sheet, and polyproline ii frequencies and strong nearest neighbor effects in a restricted coil library. *Biochemistry*, 44(28) :9691–9702, Jul 2005. (Cité pages 41 et 42.)

[103] Abhishek K. Jha, Andrés Colubri, Karl F. Freed, and Tobin R. Sosnick. Statistical coil model of the unfolded state : resolving the reconciliation problem. *Proc Natl Acad Sci U S A*, 102(37) :13099–13104, Sep 2005. (Cité page 41.)

[104] V. L. Arcus, S. Vuilleumier, S. M. Freund, M. Bycroft, and A. R. Fersht. A comparison of the ph, urea, and temperature-denatured states of barnase by heteronuclear nmr : implications for the initiation of protein folding. *J Mol Biol*, 254(2) :305–321, Nov 1995. (Cité page 43.)

[105] L. J. Smith, K. A. Bolin, H. Schwalbe, M. W. MacArthur, J. M. Thornton, and C. M. Dobson. Analysis of main chain torsion angles in proteins : prediction of nmr coupling constants for native and random coil conformations. *J Mol Biol*, 255(3) : 494–506, Jan 1996. (Cité page 43.)

[106] C. J. Penkett, C. Redfield, I. Dodd, J. Hubbard, D. L. McBay, D. E. Mossakowska, R. A. Smith, C. M. Dobson, and L. J. Smith. Nmr analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein. *J Mol Biol*, 274 (2) :152–159, Nov 1997. (Cité page 41.)

[107] Joseph A. Marsh and Julie D. Forman-Kay. Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. *J Mol Biol*, 391(2) :359–374, Aug 2009. (Cité page 43.)

[108] Pau Bernadó, Laurence Blanchard, Peter Timmins, Dominique Marion, Rob W H. Ruigrok, and Martin Blackledge. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc Natl Acad Sci U S A*, 102(47) :17002–17007, Nov 2005. (Cité pages 43, 49 et 53.)

[109] Valéry Ozenne, Frédéric Bauer, Loïc Salmon, Jie-Rong Huang, Malene Ringkjøbing Jensen, Stéphane Segard, Pau Bernadó, Céline Charavay, and Martin Blackledge. Flexible-meccano : a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*, 28(11) :1463–1470, Jun 2012. (Cité pages 43, 45 et 171.)

[110] Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, and David E. Shaw. How fast-folding proteins fold. *Science*, 334(6055) :517–520, Oct 2011. (Cité page 44.)

[111] Robert Konrat. The protein meta-structure : a novel concept for chemical and molecular biology. *Cell Mol Life Sci*, 66(22) :3625–3639, Nov 2009. (Cité pages 44 et 45.)

[112] Pau Bernadó, Efstratios Mylonas, Maxim V. Petoukhov, Martin Blackledge, and Dmitri I. Svergun. Structural characterization of flexible proteins using small-angle x-ray scattering. *J Am Chem Soc*, 129(17) :5656–5664, May 2007. (Cité pages 45, 49 et 53.)

[113] Loïc Salmon, Gabrielle Nodet, Valéry Ozenne, Guowei Yin, Malene Ringkjøbing Jensen, Markus Zweckstetter, and Martin Blackledge. Nmr characterization of long-range order in intrinsically disordered proteins. *J Am Chem Soc*, 132(24) : 8407–8418, Jun 2010. (Cité pages 45 et 60.)

[114] Sigridur Kristjansdottir, Kresten Lindorff-Larsen, Wolfgang Fieber, Christopher M. Dobson, Michele Vendruscolo, and Flemming M. Poulsen. Formation of native and non-native interactions in ensembles of denatured acbp molecules from paramagnetic relaxation enhancement studies. *J Mol Biol*, 347(5) :1053–1062, Apr 2005. (Cité page 45.)

[115] Gabrielle Nodet, Loïc Salmon, Valéry Ozenne, Sebastian Meier, Malene Ringkjøbing Jensen, and Martin Blackledge. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from nmr residual dipolar couplings. *J Am Chem Soc*, 131(49) :17908–17918, Dec 2009. (Cité page 45.)

[116] Jie-rong Huang and Stephan Grzesiek. Ensemble calculations of unstructured proteins constrained by rdc and pre data : a case study of urea-denatured ubiquitin. *J Am Chem Soc*, 132(2) :694–705, Jan 2010. (Cité pages 46 et 107.)

[117] Charles K. Fisher, Austin Huang, and Collin M. Stultz. Modeling intrinsically disordered proteins with bayesian statistics. *J Am Chem Soc*, 132(42) :14919–14927, Oct 2010. (Cité page 46.)

[118] Joseph A. Marsh, Chris Neale, Fernando E. Jack, Wing-Yiu Choy, Anna Y. Lee, Karin A. Crowhurst, and Julie D. Forman-Kay. Improved structural characterizations of the drkn sh3 domain unfolded state suggest a compact ensemble with native-like and non-native structure. *J Mol Biol*, 367(5) :1494–1510, Apr 2007. (Cité page 46.)

[119] Joseph A. Marsh and Julie D. Forman-Kay. Ensemble modeling of protein disordered states : Experimental restraint contributions and validation. *Proteins*, Oct 2011. (Cité page 46.)

[120] Robert Schneider, Jie-rong Huang, Mingxi Yao, Guillaume Communie, Valéry Ozenne, Luca Mollica, Loïc Salmon, Malene Ringkjøbing Jensen, and Martin Blackledge. Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. *Mol Biosyst*, 8(1) :58–68, Jan 2012. (Cité page 46.)

[121] Peter Tompa. Intrinsically disordered proteins : a 10-year recap. *Trends Biochem Sci*, Sep 2012. (Cité page 46.)

[122] Vladimir N. Uversky. Natively unfolded proteins : a point where biology waits for physics. *Protein Sci*, 11(4) :739–756, Apr 2002. (Cité page 49.)

[123] Pau Bernadó, Carlos W. Bertoncini, Christian Griesinger, Markus Zweckstetter, and Martin Blackledge. Defining long-range order and local disorder in native alpha-synuclein using residual dipolar couplings. *J Am Chem Soc*, 127(51) :17968–17969, Dec 2005. (Cité pages 49, 51 et 115.)

[124] Pau Bernadó, Martin Blackledge, and Javier Sancho. Sequence-specific solvent accessibilities of protein residues in unfolded protein ensembles. *Biophys J*, 91 (12) :4536–4543, Dec 2006. (Cité page 49.)

[125] Konstantin Berlin, Dianne P. O'Leary, and David Fushman. Improvement and analysis of computational methods for prediction of residual dipolar couplings. *J Magn Reson*, 201(1) :25–33, Nov 2009. (Cité page 51.)

[126] Yang Shen and Ad Bax. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR*, 38(4) :289–302, Aug 2007. (Cité pages 53 et 81.)

[127] Yang Shen and Ad Bax. Sparta+ : a modest improvement in empirical nmr chemical shift prediction by means of an artificial neural network. *J Biomol NMR*, 48(1) : 13–22, Sep 2010. (Cité pages 53 et 81.)

[128] Eran Eyal, Rafael Najmanovich, Brendan J. McConkey, Marvin Edelman, and Vladimir Sobolev. Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *J Comput Chem*, 25(5) :712–724, Apr 2004. (Cité page 53.)

[129] D Svergun, C Barberato, and MHJ Koch. CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *JOURNAL OF APPLIED CRYSTALLOGRAPHY*, 28(Part 6) :768–773, DEC 1 1995. ISSN 0021-8898. (Cité page 53.)

[130] Malene Ringkjøbing Jensen, Klaartje Houben, Ewen Lescop, Laurence Blanchard, Rob W H. Ruigrok, and Martin Blackledge. Quantitative conformational analysis of partially folded proteins from residual dipolar couplings : application to the molecular recognition element of sendai virus nucleoprotein. *J Am Chem Soc*, 130 (25) :8055–8061, Jun 2008. (Cité page 60.)

[131] Min-Kyu Cho, Hai-Young Kim, Pau Bernado, Claudio O. Fernandez, Martin Blackledge, and Markus Zweckstetter. Amino acid bulkiness defines the local conformations and dynamics of natively unfolded alpha-synuclein and tau. *J Am Chem Soc*, 129(11) :3032–3033, Mar 2007. (Cité page 60.)

[132] Malene Ringkjøbing Jensen, Guillaume Communie, Euripedes Almeida Ribeiro, Jr, Nicolas Martinez, Ambroise Desfosses, Loïc Salmon, Luca Mollica, Frank Gabel, Marc Jamin, Sonia Longhi, Rob W H. Ruigrok, and Martin Blackledge. Intrinsic disorder in measles virus nucleocapsids. *Proc Natl Acad Sci U S A*, 108(24) :9839–9844, Jun 2011. (Cité pages 60, 61, 84 et 100.)

[133] Sebastian Meier, Stephan Grzesiek, and Martin Blackledge. Mapping the conformational landscape of urea-denatured ubiquitin using residual dipolar couplings. *J Am Chem Soc*, 129(31) :9799–9807, Aug 2007. (Cité page 65.)

[134] Joseph A. Marsh, Jennifer M R. Baker, Martin Tollinger, and Julie D. Forman-Kay. Calculation of residual dipolar couplings from disordered state ensembles using local alignment. *J Am Chem Soc*, 130(25) :7804–7805, Jun 2008. (Cité page 66.)

[135] Jie-rong Huang, Frank Gabel, Malene Ringkjøbing Jensen, Stephan Grzesiek, and Martin Blackledge. Sequence-specific mapping of the interaction between urea and unfolded ubiquitin from ensemble analysis of nmr and small angle scattering data. *J Am Chem Soc*, 134(9) :4429–4436, Mar 2012. (Cité page 75.)

[136] Malene Ringkjøbing Jensen, Loïc Salmon, Gabrielle Nodet, and Martin Blackledge. Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *J Am Chem Soc*, 132(4) :1270–1272, Feb 2010. (Cité pages 79, 83 et 84.)

[137] Joseph A. Marsh, Vinay K. Singh, Zongchao Jia, and Julie D. Forman-Kay. Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein : implications for fibrillation. *Protein Sci*, 15(12) :2795–2804, Dec 2006. (Cité page 79.)

[138] Carlo Camilloni, Alfonso De Simone, Wim F. Vranken, and Michele Vendruscolo. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry*, 51(11) :2224–2231, Mar 2012. (Cité pages 79 et 84.)

[139] Yang Shen, Oliver Lange, Frank Delaglio, Paolo Rossi, James M. Aramini, Gaohua Liu, Alexander Eletsky, Yibing Wu, Kiran K. Singarapu, Alexander Lemak, Alexandr Ignatchenko, Cheryl H. Arrowsmith, Thomas Szyperski, Gaetano T. Montelione, David Baker, and Ad Bax. Consistent blind protein structure generation from nmr chemical shift data. *Proc Natl Acad Sci U S A*, 105(12) :4685–4690, Mar 2008. (Cité page 81.)

[140] Yang Shen, Philip N. Bryan, Yanan He, John Orban, David Baker, and Ad Bax. De novo structure generation using chemical shifts for proteins with high-sequence identity but different folds. *Protein Sci*, 19(2) :349–356, Feb 2010. (Cité page 81.)

[141] Andrea Cavalli, Xavier Salvatella, Christopher M. Dobson, and Michele Vendruscolo. Protein structure determination from nmr chemical shifts. *Proc Natl Acad Sci U S A*, 104(23) :9615–9620, Jun 2007. (Cité page 81.)

[142] Beomsoo Han, Yifeng Liu, Simon W. Ginzinger, and David S. Wishart. Shiftx2 : significantly improved protein chemical shift prediction. *J Biomol NMR*, 50(1) : 43–57, May 2011. (Cité page 81.)

[143] Stephen Neal, Alex M. Nip, Haiyan Zhang, and David S. Wishart. Rapid and accurate calculation of protein 1h, 13c and 15n chemical shifts. *J Biomol NMR*, 26 (3) :215–240, Jul 2003. (Cité page 81.)

[144] Kai J. Kohlhoff, Paul Robustelli, Andrea Cavalli, Xavier Salvatella, and Michele Vendruscolo. Fast and accurate predictions of protein nmr chemical shifts from interatomic distances. *J Am Chem Soc*, 131(39) :13894–13895, Oct 2009. (Cité page 81.)

[145] Marco D. Mukrasch, Jacek Biernat, Martin von Bergen, Christian Griesinger, Eckhard Mandelkow, and Markus Zweckstetter. Sites of tau important for aggregation populate beta-structure and bind to microtubules and polyanions. *J Biol Chem*, 280 (26) :24978–24986, Jul 2005. (Cité pages 84, 134 et 147.)

[146] Magnus Kjaergaard, Ann-Beth Nørholm, Ruth Hendus-Altenburger, Stine F. Pedersen, Flemming M. Poulsen, and Birthe B. Kragelund. Temperature-dependent structural changes in intrinsically disordered proteins : formation of alpha-helices or loss of polyproline ii ? *Protein Sci*, 19(8) :1555–1564, Aug 2010. (Cité page 84.)

[147] Scott A. Showalter and Rafael Brüschweiler. Quantitative molecular ensemble interpretation of nmr dipolar couplings without restraints. *J Am Chem Soc*, 129 (14) :4158–4159, Apr 2007. (Cité page 92.)

[148] Sonja Alexandra Dames, Regula Aregger, Navratna Vajpai, Pau Bernado, Martin Blackledge, and Stephan Grzesiek. Residual dipolar couplings in short peptides reveal systematic conformational preferences of individual amino acids. *J Am Chem Soc*, 128(41) :13508–13514, Oct 2006. (Cité page 92.)

[149] Judith Klein-Seetharaman, Maki Oikawa, Shaun B. Grimshaw, Julia Wirmer, Elke Duchardt, Tadashi Ueda, Taiji Imoto, Lorna J. Smith, Christopher M. Dobson, and Harald Schwalbe. Long-range interactions within a nonnative protein. *Science*, 295 (5560) :1719–1722, Mar 2002. (Cité page 105.)

[150] Matthew M. Dedmon, Kresten Lindorff-Larsen, John Christodoulou, Michele Vendruscolo, and Christopher M. Dobson. Mapping long-range interactions in alpha-synuclein using spin-label nmr and ensemble molecular dynamics simulations. *J Am Chem Soc*, 127(2) :476–477, Jan 2005. (Cité pages 105 et 115.)

[151] Carlos W. Bertoncini, Young-Sang Jung, Claudio O. Fernandez, Wolfgang Hoyer, Christian Griesinger, Thomas M. Jovin, and Markus Zweckstetter. Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein. *Proc Natl Acad Sci U S A*, 102(5) :1430–1435, Feb 2005. (Cité pages 107 et 115.)

[152] André Delacourte. Le retour de la protéine tau. *La recherche*, 10(3 Suppl) :44–48, 2003. (Cité page 130.)

[153] Maria Jose Metcalfe and Maria E. Figueiredo-Pereira. Relationship between tau pathology and neuroinflammation in alzheimer's disease. *Mt Sinai J Med*, 77(1) : 50–58, 2010. (Cité page 131.)

[154] André Delacourte. La maladie d'alzheimer, une tauopthie parmi d'autres ? *Médecine/Science*, 18(3 Suppl) :727–736, 2002. (Cité pages 132 et 133.)

[155] Rhagavendran L. Narayanan, Ulrich H N. Dürr, Stefan Bibow, Jacek Biernat, Eckhard Mandelkow, and Markus Zweckstetter. Automatic assignment of the intrinsically disordered protein tau with 441-residues. *J Am Chem Soc*, 132(34) :11906–11907, Sep 2010. (Cité page 134.)

[156] Marco D. Mukrasch, Stefan Bibow, Jegannath Korukottu, Sadasivam Jeganathan, Jacek Biernat, Christian Griesinger, Eckhard Mandelkow, and Markus Zweckstetter. Structural polymorphism of 441-residue tau at single residue resolution. *PLoS Biol*, 7(2) :e34, Feb 2009. (Cité pages 136 et 147.)

[157] Stefan Bibow, Valéry Ozenne, Jacek Biernat, Martin Blackledge, Eckhard Mandelkow, and Markus Zweckstetter. Structural impact of proline-directed pseudophosphorylation at at8, at100, and phf1 epitopes on 441-residue tau. *J Am Chem Soc*, 133(40) :15842–15845, Oct 2011. (Cité page 145.)

Résumé Près de 40% des protéines présentes dans les cellules sont prédites partiellement ou complètement désordonnées. Ces protéines dépourvues de structure tridimensionnelle à l'état natif sont impliquées dans de nombreux mécanismes biologiques, la flexibilité jouant un rôle moteur dans les mécanismes de reconnaissance moléculaire. La prise en considération de l'existence de flexibilité au sein des protéines et des interactions protéines-protéines a nécessité le renouvellement de nos connaissances, de notre appréhension des fonctions biologiques ainsi que des approches pour étudier et interpréter ces phénomènes. La méthode retenue pour étudier ces transitions conformationnelles est la spectroscopie par résonance magnétique nucléaire. Elle dispose d'une sensibilité unique, d'une résolution à l'échelle atomique et permet par diverses expériences d'accéder à l'ensemble des échelles de temps définissant les mouvements de ces protéines. Nous combinons ces mesures expérimentales à un modèle statistique représentant l'ensemble du paysage énergétique des protéines désordonnées : la description par ensemble explicite de structures. Ce modèle est une représentation discrète des différents états échantillonnés par ces protéines. Il permet, combinant les déplacements chimiques, les couplages dipolaires et la relaxation paramagnétique, de développer une description moléculaire de l'état déplié en caractérisant à la fois l'information locale et l'information à longue portée présente dans les protéines intrinsèquement désordonnées.

MOTS-CLES Résonance Magnétique Nucléaire, Couplages Dipolaires Résiduels, Déplacement Chimique, Relaxation Paramagnétique, Désordre Conformationnel, Protéines Intrinsèquement Désordonnées, Description par Ensemble, protéine Tau.

Abstract Around 40% of the human génome does not fold into stable three-dimensional structures but are either unfolded, or contain unfolded regions of significant length. The inherent flexibility of this class of proteins is essential for their function in a vast range of biomolecular process such as molecular recognition. In order to take into account the specificity of these interactions, it has been necessary to invent new approaches to study and interpret their behaviour. Nuclear magnetic resonance spectroscopy is a unique atomic resolution probe which is sensitive to a very large range of time scales. We combine experimental NMR data with a statistical model describing the energy landscape of unfolded state : the explicit ensemble description. This model is a discrete representation of the different states of theses proteins. Combining chemical shifts, residual dipolar couplings and paramagnetic relaxation enhancement, it is then possible to develop a molecular description of the unfolded state caracterising both the local and long-range information of intrinsically disordered proteins.

KEY WORDS Nuclear Magnetic Resonance, Residual Dipolar Coupling, Chemical Shift, Paramagnetic Relaxation Enhancement, Conformational Disorder, Intrinsically Disordered Proteins, Ensemble Description, Tau protein

---

LABORATOIRE DE THESE Institut de Biologie Structurale Jean-Pierre Ebel, UMR 5075, CEA-CNRS-UJF. Equipe Flexibilité et Dynamique des Protéines. 41, rue Jules Horowitz, 38027 Grenoble Cedex.