Algorithms and comptutational tools for the study of Intrinsically Disordered Proteins

Alejandro Estaña Garcia

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ FÉDÉRALE TOULOUSE MIDI-PYRÉNÉES

Délivré par :

l'Institut National des Sciences Appliquées de Toulouse (INSA de Toulouse)

Présentée et soutenue le *27/02/2020* **par :**

ALEJANDRO ESTAÑA GARCÍA

ALGORITHMS AND COMPUTATIONAL TOOLS FOR THE STUDY OF INTRINSICALLY DISORDERED PROTEINS

JURY

THOMAS SCHIEX	Directeur de Recherche	Président du Jury
ORA SCHUELER-FURMAN	Professeure Adjointe	Rapporteur
MICHAEL NILGES	Professeur	Rapporteur
LOÏC SALMON	Directeur de Recherche	Examinateur
PAU BERNADÓ	Directeur de Recherche	Directeur de thèse
JUAN CORTÉS	Directeur de Recherche	Directeur de thèse

École doctorale et spécialité :

MITT : Domaine STIC : Réseaux, Télécoms, Systèmes et Architecture

Unité de Recherche :

Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS-CNRS) et Centre de Biochimie Structurale (CBS-CNRS)

Directeur(s) de Thèse :

Juan CORTÉS et *Pau BERNADÓ*

Rapporteurs :

Ora SCHUELER-FURMAN et *Michael NILGES*

Acknowledgments

Including my master's internship, I have spent five years working on this project. I find pertinent the analogy between the thesis and a long-distance run. There are very different moments, the energy and the emotions fluctuate during the race, but you are not alone, and that gives you extra strength all the way through. This experience has been very positive for me, mainly thanks to the people I have had by my side. I hope I have given back to them at least a small part of what they have given me.

I want to start by thanking the two main people responsible of this adventure, my two excellent advisors. Thanks to Juan Cortés for trusting me from the very beginning. We have shared many moments of all kinds and your wise advice has been there when I have needed it. Thanks to Pau Bernadó for always being ready to help me and support me. All in all, thank you both for the moments you have given me. The passion you have for science is admirable. Being able to participate in meetings with you both has been a privilege. Meetings full of interesting ideas to find the best way to address the scientific problem we were facing. I could not be more grateful knowing that thanks to you I have grown on a personal and professional level.

I want to thank my team in Toulouse for all these years together. Especially Marc Vaisset, who has been willing to give me a hand and teach me with his natural sympathy. It's been a pleasure for me to work with Amélie Bazoret, who has been an excellent teammate and always has accurate advice. Also to Laurent Denarie and Kevin Molloy for having advised and helped me in a climate of conviviality. I don't want to forget the rest of the RIS team with whom I have shared many moments both inside and outside of the lab. Although I cannot name all of you, you have been a very important factor in making me feel at home in the LAAS.

I would also like to thank all the members of the Montpellier CBS team. I felt very welcome with each visit I made to you. I want to highlight and thank Nathalie Sibille who has actively participated in my thesis and has helped me a lot in this project.

I don't want to forget the CALMIP supercomputing center team, particularly Emmanuel Courcelle and Nicolas Renon who have given me a hand to improve the efficiency of my parallel code.

Finally, I would like to thank my family and friends who have been the main driving force behind me in this ambitious project. Especially to my mother and brother for being by my side unconditionally.

Summary

Intrinsically Disordered Proteins (IDPs) are involved in many biological processes. Their inherent plasticity facilitates very specialized tasks in cell regulation and signalling, and their malfunction is linked to severe pathologies. Understanding the functional roles of IDPs requires their structural characterization, which is extremely challenging, and needs a tight coupling of experimental and computational methods. In contrast to structured/globular proteins, IDPs cannot be represented by a single conformation, and their models must be based on ensembles of conformations representing a distribution of states that the protein adopts in solution. While purely random coil ensembles can be reliably constructed by available bioinformatics tools, these tools fail to reproduce the conformational equilibrium present in partially-structured regions.

In this thesis, we propose several computational methods that, combined with experimental data, provide a better structural characterization of IDPs. These methods can be grouped in two main categories: methods to construct conformational ensemble models, and methods to simulate conformational transitions.

A key methodological development of this thesis, and the basis of most algorithmic contributions, has been the construction of a three-residue fragments (tripeptide) database from high-resolution protein structures. We demonstrate that these tripeptide building blocks are highly rich in information and represent a solid foundation to accurately describe the structure and fluctuations in disordered chains. One advantage of this approach, which we exploit along the thesis, is the capacity to apply tailored filtering to obtain the most appropriate tripeptide database for each specific purpose.

Contributing to the first type of methods, we propose a new approach to generate realistic conformational ensembles that improves previously existing methods, being able to reproduce the partially-structured regions in IDPs. This method exploits structural information encoded in a database of three-residue fragments (tripeptides) extracted from high-resolution experimentally resolved protein structures. We have shown that conformational ensembles generated by our method accurately reproduce structural data obtained from NMR and SAXS experiments for a benchmark set of nine IDPs. Also exploiting the tripeptide database, we have developed an algorithm to predict the propensity to form secondary structure elements of fragments inside IDPs. This new method provides more accurate results than those of the most commonly-used predictors available on our benchmark set of well-characterized IDPs.

Contributing to the second type of methods, we have developed an original approach to model the folding mechanism of secondary structural elements. The computation of conformational transitions is formulated as a discrete path search problem using the tripeptide database. To evaluate the approach, we have applied the strategy to two small synthetic polypeptides mimicking two common structural motifs in proteins. The folding mechanisms extracted are very similar to those obtained when using traditional, computationally expensive approaches. Finally,

we have developed a more general method to compute transition paths between a (possibly large) set of conformations of an IDP. This method builds on a multi-tree variant of the TRRT algorithm, developed at LAAS-CNRS, and which provided good results for small and middle-sized biomolecules. In order to apply this method to IDPs, we have proposed a hybrid strategy for the parallelization of the algorithm, enabling an efficient execution in computer clusters.

In addition to the aforementioned methodological work, I have been actively involved in multidisciplinary work, together with biophysicists and biologists, where I have applied these methods to the investigation of important biological systems, in particular the huntingtin protein, the causative agent of Huntington's disease.

In conclusion, the work carried out during my PhD thesis has enabled a better understanding of the relationship between sequence and structural properties in IDPs, paving the way to novel applications. For example, this deeper understanding of sequence-structure relationships will enable us to anticipate structural perturbations exerted by sequence mutations, and subsequently, the rational design of IDPs with tailored properties for biotechnological applications.

Resume

Les protéines intrinsèquement désordonnées (IDPs, acronyme en anglais de Intrinsically Disordered Proteins) sont essentielles dans de nombreux processus biologiques. Leur plasticité inhérente les assigne à des tâches complémentaires de celles des protéines globulaires, dans la régulation et dans la signalisation cellulaire ; leurs dysfonctionnements sont associés des pathologies sévères. Comprendre leur rôle fonctionnel exige de caractériser la structure des IDPs et des complexes qu'elles forment. Modéliser les IDPs est extrêmement difficile et exige un couplage étroit des méthodes expérimentales et informatiques. Contrairement aux protéines structurées/globulaires, les IDPs ne peuvent pas être représentées par une seule conformation, et leurs modèles doivent être fondés sur des ensembles de conformations représentatifs des états que la protéine adopte en solution.

Il existe de multiples outils bioinformatiques qui permettent d'identifier à priori les éléments partiellement structurés au sein des IDPs. Cependant, les caractéristiques structurelles détectées par ces programmes dépendent fortement de la méthodologie utilisée, et les différentes méthodes produisent souvent des résultats contradictoires. Alors que des ensembles purement composés par "random coil" peuvent être construits de manière, par des outils bioinformatiques accessibles, l'équilibre conformationnel présent dans les régions partiellement structurées est mal reproduit.

Dans cette thèse, nous proposons plusieurs méthodes de calcul qui, combinées à des données expérimentales, permettent une meilleure caractérisation structurelle des IDPs. Ces méthodes peuvent être regroupées en deux grandes catégories : les méthodes de construction de modèles d'ensembles conformationnels et les méthodes de simulation des transitions conformationnelles.

Un développement méthodologique clé de cette thèse, et la base de la plupart des contributions algorithmiques, a été la construction d'une base de données de fragments de trois résidus (tripeptides) à partir de structures protéiques à haute résolution. Nous démontrons que ces blocs de construction tripeptidiques sont très riches en informations et constituent une base solide pour décrire avec précision la structure et les fluctuations des protéines désordonnées. L'un des avantages de cette approche, que nous exploitons tout au long de la thèse, est la capacité d'appliquer un filtrage sur mesure pour obtenir la base de données de tripeptides la mieux adaptée à chaque objectif spécifique.

Dans le premier ensemble de méthodes, nous présentons une nouvelle approche pour générer des ensembles conformationnels réalistes, qui améliore les approches existantes, et permet de reproduire les régions partiellement structurées des IDPs. Cette méthode exploite les informations structurelles codées dans les bases de données de tripeptides. Nous avons montré que les ensembles conformationnels construits par notre méthode reproduisent fidèlement les descripteurs structurels obtenus à partir d'expériences RMN et SAXS. En tant que composante nécessaire de l'algorithme de construction d'ensemble, nous avons développé un algorithme pour prédire la propension de certains fragments à l'intérieur des IDPs à former des éléments de structure secondaire. Cette nouvelle méthode, qui exploite également la base de données de tripeptides, fournit des résultats plus précis que ceux des prédicteurs les plus couramment utilisés sur plusieurs IDPs bien caractérisées. Bien que le prédicteur structurel ait été principalement développé pour compléter notre méthode de modélisation d'ensembles, il peut également être très utile comme outil indépendant.

Dans un deuxième type de méthodes, nous avons développé une approche originale pour modéliser le mécanisme de repliement des éléments structuraux secondaires. Le calcul des transitions conformationnelles menant à la formation des éléments structuraux est formulé comme un problème de recherche de chemin discret à l'aide de la base de données de tripeptides. Pour évaluer l'approche, nous avons appliqué la stratégie à deux petits polypeptides synthétiques imitant deux motifs structurels communs dans les protéines. Les mécanismes de repliement extraits sont très similaires à ceux obtenus en utilisant des approches traditionnelles et coûteuses en calcul. Enfin, nous avons mis au point une méthode plus générale pour calculer les chemins de transition entre un ensemble (éventuellement important) de conformations d'IDPs. Cette méthode s'appuie sur une variante multi-arbre de l'algorithme Transition-based Rapidly-exploring Random Tree (Multi-TRRT), récemment développé au LAAS-CNRS, et qui a donné de bons résultats pour les biomolécules de petites et moyennes tailles. Afin d'appliquer cette méthode aux IDPs, nous avons proposé une stratégie hybride (mémoire partagée/distribuée) pour la parallélisation de l'algorithme, permettant une exécution efficace dans les clusters de calcul.

Outre le travail méthodologique susmentionné, nous avons également participé activement à des travaux multidisciplinaires, en collaboration avec des biophysiciens et des biologistes, où nous avons appliqué ces méthodes à l'étude de systèmes

biologiques d'importance, en particulier la protéine huntingtin, l'agent responsable de la maladie de Huntington.

En conclusion, les travaux menés dans le cadre de cette thèse de doctorat ont permis de mieux comprendre la relation entre la séquence et la structure des IDPs, ouvrant la voie à de nouvelles applications. Grâce à cette compréhension plus approfondie des relations séquence-structure il sera possible d'anticiper les perturbations structurelles engendrées par les mutations dans la séquence, ainsi que la conception rationnelle des IDPs ayant des propriétés spécifiques pour des applications dans les biotechnologies.

Contents

Chapter 1

Introduction

Intrinsically Disordered Proteins or Regions (IDPs/IDRs) play crucial roles in multiple biological processes and are directly involved in several pathologies, including cancer and neurodegeneration [229, 38, 7]. The inherent plasticity of this family of proteins facilitates a range of functions that are complementary to those of their folded counterparts [245]. In most cases, the activity of IDPs is manifested when interacting with globular partners to trigger signaling or metabolic cascades [227]. These interactions are mediated by Short Linear Motifs (SLiMs) that recognize regions of the partner surface in a highly specific manner [232]. The presence of transiently formed structural motifs in SLiMs facilitates partner recognition and tunes the thermodynamics and kinetics of interactions [150, 167, 192]. To understand these functional mechanisms, it is pivotal to identify and characterize these partially structured elements inserted into IDPs.

The relatively flat conformational energy landscape of IDPs has notably hampered their structural characterization. Modelling IDPs is extremely challenging, and requires a tight coupling of experimental and computational methods [11 , 12]. In contrast to structured/globular proteins, IDPs cannot be represented by a single conformation, their flexibility rends impossible the crystallization of the protein. Even in the case when a fragment of the IDP is crystallized in interaction with its globular partner, the resulting conformation is not enough to characterize its structure in solution. IDP models must be based on ensembles, usually involving thousands of conformations representing a distribution of states that the protein adopts in solution [13, 14]. Experimental data obtained by Nuclear Magnetic Resonance (NMR) and Small-Angle X-ray Scattering (SAXS) provide information on conformational trends at the residue level, the presence of transient long-range contacts, and the overall size of the ensemble of conformations [59]. Then, the quantitative interpretation of these data requires the use of computational approaches that account for their ensemble averaged properties. These computational approaches are based on the construction of large conformational ensembles, which are subsequently refined by integrating the experimental data using restrained Molecular Dynamics (MD) simulations [44, 208], sub-ensemble selection [164, 119, 16], or Bayesian statistics [65]. Chemical Shifts (CSs) and Residual Dipolar Couplings (RDCs) measured in partially aligned media are the most sensitive probes to quantify conformational restrictions at the residue level and to define secondary structural elements [58, 99]. Conversely, ensembles refined with SAXS data describe the overall properties of the protein in solution [18, 182]. Consequently, conformational ensembles that simultaneously describe both sources of complementary information are excellent structural

models of proteins in solution [207, 33].

Multiple computational tools using distinct levels of description have been developed to characterize IDPs when no or limited experimental information is available. Current disorder prediction tools, which are based on the statistical analysis of protein sequences, provide rough estimations of partly structured regions in IDPs [47], although the exact secondary structure classes are poorly defined. In principle, a more accurate characterization can be provided by MD-based methods. However, despite significant advances in the extension of MD methods to IDPs [173, 83], their applicability to exhaustively explore the conformational space of these proteins is still limited. Knowledge-based approaches have emerged as an alternative to overcome some of these limitations. These approaches usually describe the conformational properties of individual residues using the so-called coil libraries, which contain residue-specific $\{\phi, \psi\}$ angles from fragments of experimentally determined protein structures that do not form secondary structural elements [209, 103, 15, 66, 223, 203]. Despite their simplicity, coil models provide an accurate description of NMR parameters such as J-couplings [209, 203] and RDCs [15, 99], and SAXS curves [18] for flexible peptides and disordered proteins. Nevertheless, these approaches fail to identify secondary structural elements in IDPs. This limitation is caused by the chain building strategy, which sequentially appends individual residues accounting for the amino acid type and overlooks the sequence and structural context [103, 15]. Consequently, approaches such as Flexible-Meccano [15, 163] provide excellent models for the random-coil but do not capture structural features involving multiple consecutive residues. The omission of coordinated effects precludes the capacity of current approaches to predict structural classes and their populations, and hamper their application for advanced purposes.

During the last two decades, significant progress has been made to understand the complex behavior of IDPs. Nevertheless many aspects related with the structure-dynamics/function paradigm remain poorly understood. Our work aims at developing algorithms specially adapted to the study of IDPs to bring new knowledge about the relationship between sequence and conformational behavior for this family of proteins. It should also be mentioned that, although the first objective of the methods developed in this thesis concerns IDPs, the algorithmic advances can be useful for structural biology in general.

Contributions of the thesis

The goal of this thesis was to develop computational methods to model IDPs, making special emphasis in partially structured fragments. More precisely, our aim was to better understand the relationship between amino acid sequence and structural properties of IDPs from different perspectives. For this, we have developed several algorithms that can be grouped into two main categories:

- The first group is formed by algorithms for the prediction and sampling of IDP conformation. These algorithms rely on an extensive database of three-residue fragments (also called tripeptides) extracted from experimentally determined high-resolution protein structures, described in Chapter 3. The information encoded in this database captures local sequence-dependent structural properties that can be exploited for IDP modeling. First, we propose in Chapter 4 a method to predict secondary structure propensities in IDPs based in a simple statistical approach using information in the tripeptide database. Then, also using this database, Chapter 5 presents a sampling algorithm that, combined with information about structural propensities, is able to build realistic conformational ensembles of IDPs.

- The second group includes algorithms for sampling conformational transition paths. To better understand the transitions between order and disorder in partially-structured regions, we developed a path search algorithm that finds likely conformational transitions between two states applying a heuristic that exploits structural information in the tripeptide database. Chapter 6 presents the algorithm and a proof of concept using two well-characterized mini-proteins. Finally, Chapter 7 presents a novel approach to globally explore the conformational space of highly-flexible molecules such as IDPs. It is based on an efficient algorithm originating from robotics, called Multi-TRRT. We propose a hybrid paralellization strategy enabling the computation of a roadmap of transition paths between a (possibly very large) set of conformations in very short time. Preliminary results presented in this manuscript show the potential of this technique.

In addition to the methodological developments, in the context of this thesis, we have applied the new methods to several systems in the framework of inter/pluridisciplinary work together with biophysicists and biologists. The most relevant one has been the structural analysis of the huntingtin protein, the causative agent of Huntington's disease. The structural properties of the poly-Q region in huntingtin were studied by coupling the backbone NMR chemical shifts and an optimized combination of ensemble models generated by the sampling algorithm described in Chapter 5. The results show the potential of the algorithm to reproduce structural properties of challenging systems. The article describing this work (submitted) is included as an annex of this thesis.

Chapter 2

Basic concepts and Background

Contents

2.1 Introduction

The high flexibility of IDPs makes their study highly complex. Many experimental and computational methods have been developed to gain knowledge of this type of proteins. This chapter presents basic concepts of structural biology with special

emphasis on IDPs, and the most widely used experimental methods for their characterization: Nuclear Magnetic Resonance (NMR) and Small-Angle X-ray Scattering (SAXS). Then, an overview of computational protein modelling and the most popular simulation algorithms are presented. In addition, a special type of algorithms coming from robotics, which I parallelized during my thesis (see Chapter 7), is also explained. Finally, I highlight the importance of combining different types of experimental and computational methods to obtain a more detailed description of IDPs.

2.2 Structural biology

2.2.1 Protein structure

The amino acids are the constituent units of proteins. An amino acid is a molecule composed of a carbon atom (α-carbon) attached to a carboxyl group (-COOH), an amine group (-NH2), one hydrogen H and a radical R also called side chain, see Figure 2.1. Notice that the Cα is a stereogenic center (bound to four different groups) and, therefore two enantiomers are possible. In proteins L-amino acids are used almost exclusively. A detailed explanation of the main concepts in structural biology can be found in textbooks, such as references [36, 75].

Consecutive amino acid residues are linked together through a peptide bond, a double bond between the carbon C of the carboxyl group of one residue and the N of the following amine group, see red lines in Figure 2.2. When the peptidic bonds are formed, two parts can be distinguished in a proteins: the backbone and the side chains, see Figure 2.2 .

The side chain determines the physico-chemical properties of each amino acid type. Although about 500 naturally occurring amino acids are known, only 20 appear in proteins. The sequence of these 20 amino acid residues in a peptidic chain is called primary structure (Figure 2.3, a). Sequence is the fingerprint of the protein and determines its final structure. The transformation from an elongated amino acid chain to a compact 3D structure is called *folding* and starts with each region of the protein adopting a specific secondary structure depending on the sequence. The

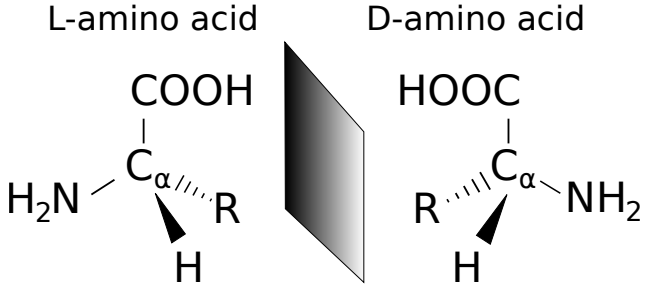


Figure 2.1. L-amino acid and D-amino acid representation with its α-carbon, the amine group, the carboxylic group, the hidrogen and the side chain

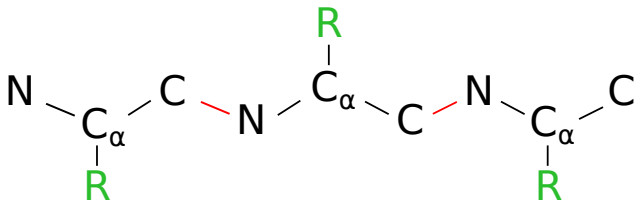


Figure 2.2. Simplified representation of a protein. In red the peptidic bonds, in black the backbone atoms and in green the residue side chains

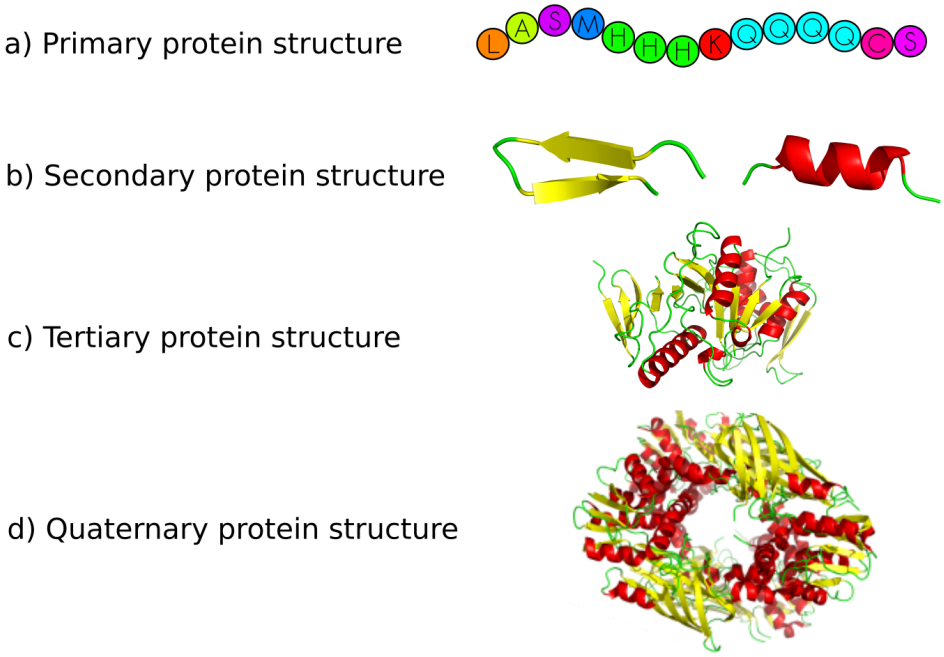


Figure 2.3. Representation of the four levels of protein structure: a) Primary protein structure, b) Secondary protein structure, c) Tertiary protein structure, d) Quaternary protein structure

main secondary structures are: α-helix, β-sheet, turns and coil regions (β-sheet and α-helix are represented in Figure 2.3b). These secondary structures interact between them to finally form the tertiary structure (Figure 2.3c). Some of the proteins interact with other proteins and come together to form complex called quaternary structure (Figure 2.3d). Intrinsically Disordered Proteins (IDPs) do not experience the folding process and remain disordered in physiological conditions. Despite their inherent plasticity IDPs, are fully functional [242].

2.2.2 The Structure-function paradigm

Proteins play many important biological functions in living organisms. Enzymes catalyze all types of chemical reactions, some are nutrient and storage proteins, which are vital in many plants for the growth and survival of the seeds, others provide cells with the ability to contract, some bind and transport substances,

they can be structural proteins giving a defined shape to cells, and some govern regulatory processes of the cell.

All these functions depend on the resulting structure of the protein, also called native state. The native state is not a rigid conformation but a combination of accessible states that the protein can adopt depending on the solvent and the temperature. Therefore, proteins are dynamic systems and their conformational movements are important for function [84]. Dynamics in flexible proteins, such as IDPs, is an inherent feature: they are constantly moving and changing from one conformational state to another. The energies of the ensemble of all possible conformations is called conformational landscape and it is characteristic of each protein. Depending on the sequence, the shape of the conformational landscape changes. While proteins that fold in a well defined 3D structure will present a funneled landscape with a clear stable global minimum state, proteins that do not fold into a specific shape, such as IDPs, have a relatively flat energy landscape with multiple local minima and the protein is constantly changing its shape moving between different states [100].

2.3 Intrinsically Disordered Proteins and their functions

Intrinsically Disordered Proteins or Regions (IDPs/IDRs) have emerged as key actors for a large variety of biological functions such as cell signalling and regulation [242, 56, 243]. The main feature of IDPs and IDRs is their lack of permanent secondary or tertiary structure that provides them with an inherent malleability enabling highly specialized biological functions [56]. The disordered nature of IDPs is encoded in their characteristic amino acid sequence. IDPs are enriched in charged and polar (Arg, Gly, Gln, Ser, Pro, Glu and Lys) and structure-breaking (Gly and Pro) amino acids, whereas they are significantly depleted in bulky hydrophobic (Ile, Leu and Val) and aromatic (Trp, Tyr and Phe) amino acids [4,5]. Eukaryotic genomes are highly enriched in genes coding for disordered proteins, and this observation has been linked to the major complexity of these organisms. In particular, the human genome has 44% of the genes encoding proteins containing disordered fragments with a length greater than 30 residues [8]. The capacity of IDPs to adapt their conformation to specifically recognize one or several partners, and the low to moderate affinity for partners make IDPs ideal for protein-protein interactions [227]. In fact, it has been shown that interactome hubs are enriched in this family of proteins [57, 110]. Partner recognition is normally performed through conserved and partially structured motifs of the protein, and their individual properties can be modulated by post-translational modifications or alternative splicing. IDRs are highly flexible regions connecting well-folded globular proteins forming the so-called multi-domain proteins. Multi-domain protein topology, which is highly prevalent in eukaryotes, enables the presence of multiple biological activities performed by the globular domains in close proximity [81, 128]. In many of these cases, IDRs behave as entropic linkers with an inherent plasticity that can be tuned depending on the

length and the specific amino acid sequence of the region.

IDPs/IDRs perform a large diversity of biological functions mainly exploiting their inherent flexibility. These functions have been classified in six categories (see Figure 2.4) in a recent study [231]:

1) **Entropic chains:** This category corresponds to all IDPs that are not structured and their function is directly linked to their disorder. These IDPs act as springs, bristles or linkers, and their functions cannot be performed by a rigid structures, as its ability to fluctuate from one state to another in a conformational ensemble is fundamental.

2) **Display sites:** These disordered regions are in transient interaction with one or more ligand(s) to induce a chemical modification (phosphorilation, acetylation, ...) promoting post-translational modifications.

3) **Chaperones:** Their function is to assist in the folding, assembly and cellular transport of newly formed proteins. The level of disorder of these proteins is very high, allowing them to interact with different partners. It has been recently discovered that RNA chaperones have a greater percentage of disorder in their sequence than other types of chaperones (40% disorder for RNA chaperones compared to 15% disorder for protein chaperones). To further emphasize the importance of disorder, the main function of these proteins depends directly on the disorder since it is the disordered fragments that recognize, solubilize or loosen the structure of misfolded proteins.

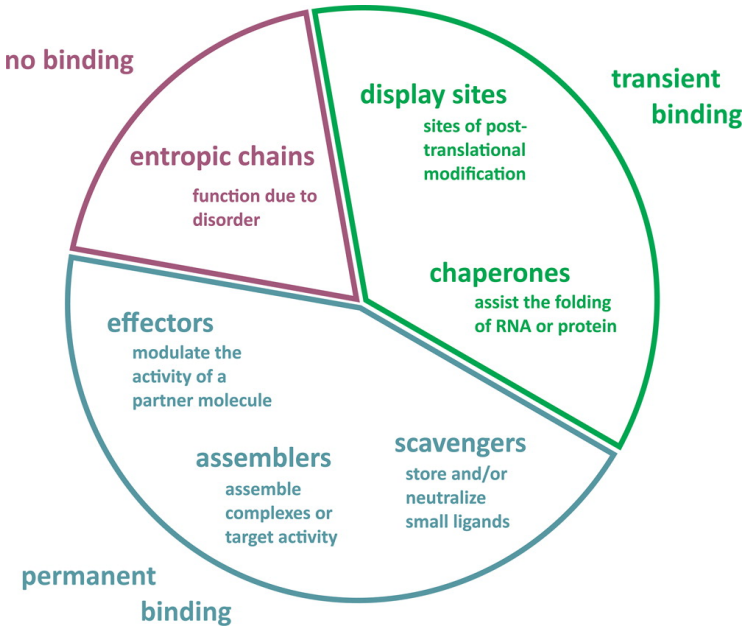


Figure 2.4. Functional classification of IDRs. Figure obtained from reference [231]

4) **Effectors:** Some proteins having IDRs act as effectors causing a modification in the behavior of a protein, either by activation or by repression of its function.

5) **Assemblers:** Some proteins, through their IDRs, have several binding sites with multiple partner proteins, and then act as molecular assemblers by promoting the formation, stabilization and regulation of large protein complexes.

6) **Scavengers:** These IDPs or IDRs store and/or neutralize small ligands. For example, casein traps calcium phosphate and thus prevent salt precipitation.

Under certain circumstances (mutations or environmental conditions) IDPs can not properly perform their function. Indeed, IDP malfunction is linked to a large number of diseases including cancer, neurodegeneration and cardiovascular diseases [229].

The biological relevance of IDPs has fostered their structural characterization [59]. Identification of the conformational preferences of binding motifs, the detection of transient long-range contacts within the chain, the structural perturbations exerted by post-translational modifications (PTMs), the shape of biomolecular complexes with disordered partners, and the spatial distribution of globular domains in multi-domain protein are structural features that must be characterized to understand the molecular bases of biological function. This characterization is far from being trivial as the inherent disorder of IDPs/IDRs precludes their crystallization.

In the following sections, NMR and SAXS experiments, which will be used along the thesis, will be explained in more detail, focusing on their application to the study of IDPs. In addition, I will emphasize how their structural information can be integrated or combined with computational approaches to characterize IDPs from a structural and dynamical perspective.

2.4 Nuclear Magnetic Resonance spectroscopy

Nuclear Magnetic Resonance (NMR) spectroscopy is an established analytical method in many scientific fields such as physics, chemistry, biology and medicine. A detailed explanation of the main concepts in NMR can be found in textbooks such as references [86, 78]. NMR is one of the principal structural biology techniques. It has atomic resolution and also has the unique ability to accurately probe protein-protein interactions and to measure the dynamic properties of proteins [149].

2.4.1 Nuclear Magnetic Resonance spectroscopy of proteins

Solving the 3D structure of a macromolecule with atomic resolution is crucial to understand the details of its the function. NMR together with X-ray crystallography and cryo-Electron Microscopy (cryo-EM) are the best means of analyzing protein structures at atomic resolution. However, NMR experimental conditions allow to

study proteins in near-natural conditions and capture the dynamics and the flexibility of particles in solution. However, the main limitation of the method is the size of the proteins that can be studied. Large globular proteins have short NMR signal relaxation times, thus reducing the sensitivity of the spectra. Moreover, large proteins present more peaks in the spectra, complicating the process of the frequency assignment. Important in the context of this thesis, NMR is the only technique allowing the high-resolution structural characterization of IDPs [58]. The first step to study an IDP by NMR is to assign a resonance frequency to all magnetically active nuclei (^{1}H, ^{15}N, and ^{13}C) of the protein. Due to very low amide proton dispersion, assignment of IDP spectra is challenging. However, the use of high magnetic field spectrometers and several methodological developments allow to routinely assign NMR frequencies of large IDPs [156]. In the last two decades, novel NMR experiments combined with modelling strategies have been developed to interpret experimental parameters measured in IDPs in terms of structure [101, 102, 100, 243].

2.4.2 General NMR theory

The principle of NMR spectroscopy is that a nucleus is placed in a very strong magnetic field and it is then exposed to electromagnetic radiation making the nucleus resonate at a specific frequency of that radiation. The absorption produced by the resonance of the nucleus is detected by radio receivers. A nucleus needs to have a spin to be detected. All nuclei having an odd number of protons or/and an odd number of neutrons have a net spin different of zero, the rest of atoms have their spin compensated and they can not be detected by NMR. In biomolecules, the atoms that can directly be observed are ^{1}H but not ^{12}C, ^{14}N and ^{16}O. To be able to obtain more information about the proteins, isotope labeling with ^{13}C and ^{15}N is necessary. As nuclei are charged particles, their movement generates a magnetic field corresponding to a magnetic moment μ proportional to the spin angular momentum I and a constant specific for each atom γ, known as gyromagnetic ratio.

$$\mu = \gamma I \tag{2.1}$$

For the case of a 1/2 spin nucleus and in the absence of a magnetic field, the two states are energetically degenerate. If the nucleus is in a magnetic field B_0 in the Z direction, the spins will align and the two states will have a different energy as a result of the interaction between the nuclear magnetic dipole moment and the external magnetic field. The spin states positioned against the field have higher energy compared to these aligned with the field. The energy gap between the two states, δE, increases with the applied magnetic field B_0. The NMR spectrum is the result of applying energy to the system in the form of varying radio frequencies. The nuclei can only absorb energy which matches the δE. As $E = Hv$, only a specific frequency is absorbed. The absorbed radio frequency induces a resonance, producing a peak in the spectrum at that specific precession frequency. In other words, the magnetic moment of the nucleus rotates around the Z axis at the resonance

frequency. The motion is called Larmor precession, and the precession frequency, v_0, is given by equation:

$$v_0 = \frac{|\gamma| B_0}{2\pi} \tag{2.2}$$

2.4.3 Population of nuclei in a sample

The sample exposed to NMR not only contains one nucleus or one molecule but a large number of them. The resulting spectrum corresponds to the information of all the molecules in the sample, and the behavior of the population has to be understood to interpret the final result. When the magnetic field is applied, there are more nuclei in the low energy spin state than in the high energy spin state. The spin populations follow the Boltzmann distribution:

$$\frac{n_{high}}{n_{low}} = e^{-\frac{\delta E}{kT}} \tag{2.3}$$

where n_{high} is the number of nuclei with high spin energy, n_{low} is the number of nuclei with low spin energy, δE is the energy gap between the two energetic levels, k is the Boltzmann constant ($k = 1.38066 \cdot 10^{-23} JK^{-1}$) and T is the temperature. The population excess generates an overall magnetization called bulk magnetization (M). Increasing the bulk magnetization improves the sensitivity of the experiment. Therefore, following the equation 2.3, we can increase the population difference by lowering the temperature or increasing δE by applying a stronger magnetic field B_0.

The bulk magnetization is the average of all the individual nuclear magnetic moments and, if no perturbations are applied, the resulting vector points in the direction of the external magnetic field B_0, see Figure 2.5a. When the radio frequency is pulsed into the sample, all the nuclear magnetic moments are displaced in the same direction and the resulting average is not pointing to the Z direction but with a certain angle to this axis. As all the individual nuclei are precessing at the resonance frequency in a coherent way, the resulting bulk magnetization is moved from its original axis, and starts to precess in the X,Y plane see Figure 2.5b. When the radio frequency is turned off, the precession of the individual nuclei becomes disordered again and, the bulk magnetization gradually returns to the Z axis. This process is called spin-spin relaxation, and the time it takes is T_2. During relaxation, the bulk magnetisation vector moves around the Z axis, and this oscillating magnetic moment generates a electrical current that is detected in Y or X axis. The resulting periodic signal over time is then treated using Fourier Transform to be expressed in terms of frequencies.

2.4.4 Electron shielding

Nuclei are surrounded by electrons, which are moving charges that obey to the laws of electronic induction. The applied magnetic field, B_0, induces circulation in the electron cloud, and a magnetic field in opposite direction of B_0 is induced. As a consequence, the local magnetic field that the nucleus experiences is smaller than

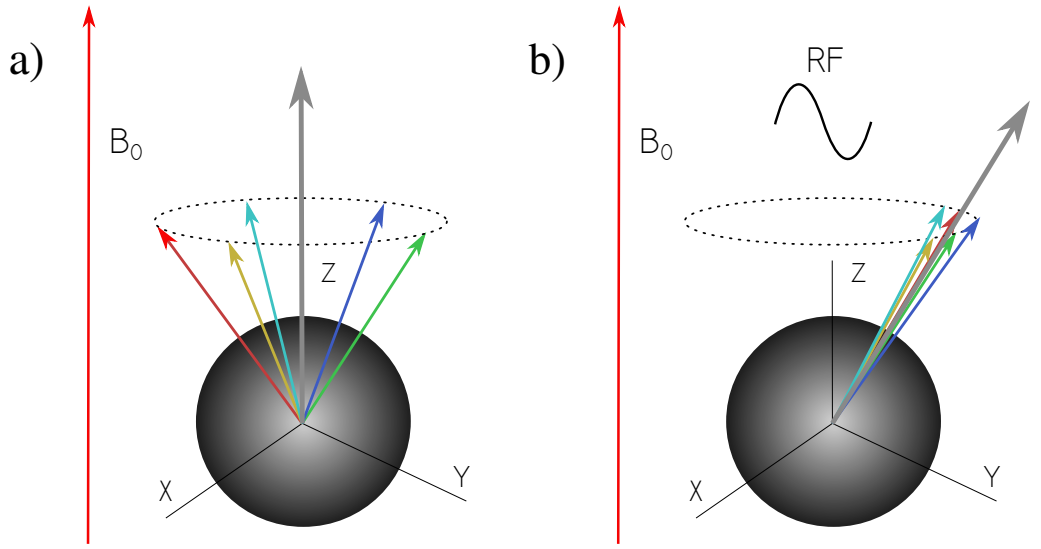


Figure 2.5. Schematic representation of a Nuclear Magnetic Resonance experiment. The applied magnetic field (vertical red vector) orientates the atom spins to precess around the direction of field. a) More nuclei are oriented in the sense of the field represented with vectors of different colors, but because the nuclei are precessing incoherently the resulting bulk magnetization, grey vector, is pointing in the direction of the magnetic field. b) After pulsing the radio frequency, all nuclei precess coherently around the magnetic field direction and a net bulk magnetization appears.

the applied field:

$$B_{local} = B_0(1 - \sigma) \tag{2.4}$$

where σ is known as the shielding or screening constant, which is a dimensionless quantity. The electrons are shielding the nucleus and a higher external field is required to meet the resonance condition. This effect has a crucial importance as the electronic distribution surrounding an atom depends on the electronegativity of near atoms allowing to identification of the chemical compounds, or the frequency assignment.

2.4.5 NMR observables

There are multiple structural observables that can be obtained from NMR spectroscopy and that have been used for the study of proteins, such as J-coupling, Nuclear Overhauser effects (NOEs), relaxation rates, residual dipolar coupling (RDCs) and chemical shifts (CS). In this section, we will focus on explaining those used throughout the thesis: CSs and RDCs.

2.4.5.1 Chemical Shifts

The local magnetic field generated by the local electron density surrounding each nucleus generates small differences in the absorption frequency from the standard

one. The observed resonance frequency value of each nucleus is known as chemical shift (CS) and it is used to identify each nuclei in the molecule. CSs are expressed in function of the reference resonance frequency v_{ref} and in parts per million (ppm):

$$ppm = \frac{(v_o - v_{ref})10^6}{v_{ref}} \tag{2.5}$$

where v_o is the observed resonance frequency. As CSs are small variations from the reference frequency, all the observed CSs for the same nuclei appear clustered in a relative small region of the spectrum. The reference resonance frequency v_{ref} corresponds to the bare proton nucleus, which is a non convenient reference. Therefore the shifts are quoted relative to the atoms of standard compounds, such as Tetramethylsilane $Si(CH_3)_4$ (TMS).

CSs are very sensitive to the chemical environment sensed by each nucleus. Therefore, their analysis allows, in principle, the determination of protein structures. However, these strategies require the combination of the experimental data with advanced computational tools [204].

IDPs have an inherent flexibility that results in a large number of protein conformations present in the solution. For this reason, the CS that is measured corresponds to the average all the conformations present in solution. When compared with the expected CSs from a random coil, these average values reveal the presence of secondary structural elements. A chemical shift index (CSI) has been established to highlight regions that deviate from pure random coils to form secondary structural elements [240, 239]. With the growing relevance of IDPs, the interest in using CSs to detect partially structured elements has been renewed, and several databases have appeared based on small synthetic peptides [195, 111] or IDPs [221] to identify these regions.

2.4.5.2 Residual dipolar couplings (RDCs)

Residual dipolar couplings (RDCs) probe the relative orientation of different pairs of nuclei within a molecule [176, 133]. Dipolar couplings are sensitive to the distance between both nuclei and the angle that their connecting vector forms with the external static magnetic field. In solution, where molecules tumble isotropically, dipolar couplings are cancelled. However, if an alignment medium is introduced in the sample, molecules behave anisotropically, hampering the signal to be completely cancelled and RDCs can be measured [69, 205].

For a couple of nuclei i and j with spin 1/2, the following equation describes the splitting of the signal [224]:

$$D_{ij} = \frac{\mu_0 \gamma_i \gamma_j h}{(2\pi r_{ij})^3} \left\langle \frac{3\cos(\theta - 1)^2}{2} \right\rangle \tag{2.6}$$

where D_{ij} is expressed in Hz, γ_i and γ_j are the gyromagnetic constants of the nuclei i and j, respectively, μ_0 is the vacuum magnetic permeability and h is the Planck's

constant. The brackets around the angular term means the average over all the motions of the particles in the solution. The structural variables of the equation correspond to the internuclei distance r_{ij}, and the angle θ is the angle between the vector connecting atom i and j and the applied magnetic field vector B_0, see Figure 2.6.

In the case of non-flexible or rigid molecules, the angular average in equation 2.7 can be replaced by a geometric sum of terms describing the orientation of the internuclear vector with respect to the protein and the corresponding average describing the order of the entire molecule.

$$\left\langle \frac{3\cos\left(\theta-1\right)^2}{2} \right\rangle = \sum_{kl=xyz} S_{kl}\cos\left(\alpha_k^{ij}\right)\cos\left(\alpha_l^{ij}\right) \tag{2.7}$$

The position of the protein is defined with respect to arbitrary Cartesian coordinate axes being α_n^{ij} the angle between the internuclear vector and the axes. S_{kl} is a Cartesian $3X3$ tensor describing the ordering of the protein and depends on the orientation of the magnetic field relative to fixed coordinate axes in the protein. As a result, with the RDCs it is possible to obtain local and distant information about the orientation that has been widely used for structural determination [89]. A detailed explanation of the equations can be found in [225].

Many liquid-crystalline aligning media have been used to generate the anisotropic sample of the proteins in solution. The common alignment media are charged polyacrylamide gels that include bicelles made of dimyristoylphosphatidylcholine (DMPC) and dihexanoylphosphatidylcholine (DHPC) [11, 162]; filamentous phages $Pf1$ [80] or fd [31]; stretched polyacrylamide gel [190, 30]; compressed polyacrylamide gel [228]; polietilenglicol/hexanol mixture [188]. More recently, alignment media based on DNA [54, 135] and collagen [138] have also been developed. More detailed information about the most commonly used alignment means can be found in reviews [10, 225].

IDP conformations in solution differ in shape and size, and therefore, when they are in an alignment media, they experience different degrees of alignment. In addition, the internuclear vector has a different orientation with respect to the alignment tensor for each conformation. All this variability is condensed in an average RDC that reports on the conformational sampling of an individual vector with respect to the biopolymeric chain [136]. Nevertheless, RDCs measured in partially aligned samples are the most sensitive experimental measurement to probe conformational sampling in IDPs [101]. Slighly negative NH RDC values are observed in random coil regions [136]. Interestingly, more positive and more negative RDCs than expected for a random coil are associated to α-helices and extended conformations, respectively [151]. This is an excellent indication to qualitatively assess the presence of distinct types of secondary structural elements. More quantitative interpretation of RDCs can be derived when applying atomistic models of disordered chains [103, 15, 141]. The measurement of multiple backbone RDCs enriches the description of residue-specific structural preferences [98].

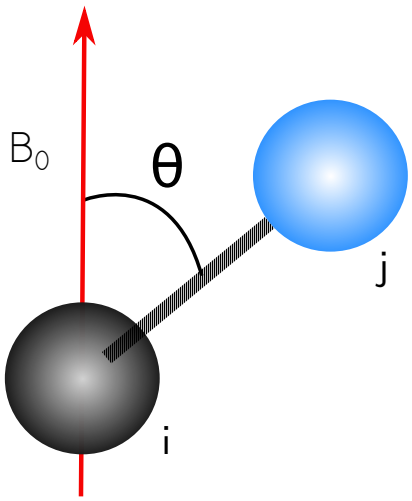


Figure 2.6. Representation of the angle θ formed between the magnetic field B_0 and the vector connecting the two bonded atoms i and j

2.5 Small-Angle X-ray Scattering

Small-angle scattering (SAS) of X-rays (SAXS) or neutrons (SANS) is a biophysical technique used to determine the low resolution structure of particles with a size ranging from 1 nm up to around 300 nm. A detailed description of the main concepts in SAS can be found in textbooks, such as references [159, 218]. SAS is a versatile technique used in many different fields, and many types of samples can be analyzed: solid objects, dust, gels or solution samples. In structural biology, SAS is a useful tool to study the overall shape and structural transitions of macromolecules in solution. Work in this thesis, as well as the explanations in this section, are focused on SAXS, but the theory and analysis would be equivalent for SANS data.

2.5.1 Small angle X-ray scattering for proteins

Small-angle scattering of X-rays (SAXS) is a method capable of giving overall information about the structure and the conformational changes of biological macromolecules in solution [64, 219, 113, 177, 93]. Major advances in instrumentation and computational methods in the last decade have led to a tremendous increase in the applications of SAXS in structural biology [171, 145, 169, 179, 72]. While lower in resolution than NMR, X-ray crystallography and cryo-EM, SAXS has the advantage that it does not require crystallization and it does not have molecular weight limitations. The sample is analyzed under near native conditions, allowing its use not only for static structural modelling but also for analyzing dynamical processes, such as folding/unfolding or assembly/dissociation, and also to understand the response to changes in the experimental conditions (pH, temperature, pressure, ionic strength, binding...). The experimental conditions of SAXS allows the characterization of proteins that are impossible to crystallize, like IDPs.

The biophysical information of the particles in the sample provided by SAXS are: radius of gyration (R_g), the maximum intra-particle distance (D_{max}) and the molecular weight (MW). The scattering data are also able to provide structural information that can be exploited to generate low-resolution 3D structures. Due to its low resolution nature, SAXS becomes more informative in combination with other structural, hydrodynamic, computational or biochemical methods. In the following sections, the method and its application to IDPs will be described.

2.5.2 General SAXS theory

In a typical SAXS experiment, a highly focused monochromatic (with a well-defined wavelength, λ) and collimated X-ray beam is directed orthogonally onto a flow cell or static flat sample holder containing the biological sample. The beam is scattered while passing thought the sample before impacting the 2D detector.

The photons of the monochromatic plane wave, with a wave vector module $|k| = 2\pi/\lambda$, impact with the electrons of the molecule and are deviated. We only consider elastic collisions between the photons and the electrons because inelastic collisions have smaller effect and do not yield any structural information. Elastic scattering means that the scattered photons have the same energy or wavelength than the incident photons $|k_i| = |k_f|$. In a solution of proteins, the scattering is isotropic and the resulting intensity, $I_{total}(s)$, depends only on the modulus of the momentum transfer $s = k_i - k_f$, $|s| = 4\theta \sin\theta/\lambda$, see Figure 2.7.

At any point r inside the protein, the scattering length density $\rho(r)$ is the total scattering length of the protein per unit of solution volume. The scattering amplitude $A(s)$ is the Fourier transform of $\rho(r)$ over the protein volume.

As the particles in solution are randomly situated and oriented, they scatter incoherently and the scattering pattern is isotropic. The observed intensity is the (spherical) average of the intensities due to individual particles in all possible positions and orientations. This resulting intensity is the product of the scattering amplitude and its complex conjugate:

$$A(s) \cdot A^*(s) = I(s) \tag{2.8}$$

Interpretation of the scattering result relies on the formalism of contrast variation [177]. Not only the proteins in the solution contribute to the total scattering intensity, but the bulk solvent also scatters the beam. The intensity due to the solvent needs to be subtracted to isolate the contribution of the protein, $\Delta\rho(r) = \rho(r) - \rho_s$. $\rho(r)$ being the total electron density in r and ρ_s the electron density of the solvent. The scattering amplitude $A(s)$ is a Fourier transform of the excess electron density:

$$A(s) = \mathscr{F}[\Delta\rho(r)] = \int \Delta\rho(r) e^{-isr} \mathrm{d}r \tag{2.9}$$

Relevant structural information of the particle can be obtained by the correlation function, $\gamma(r)$ [43]. $\gamma(r)$ expresses the correlation of electron density function

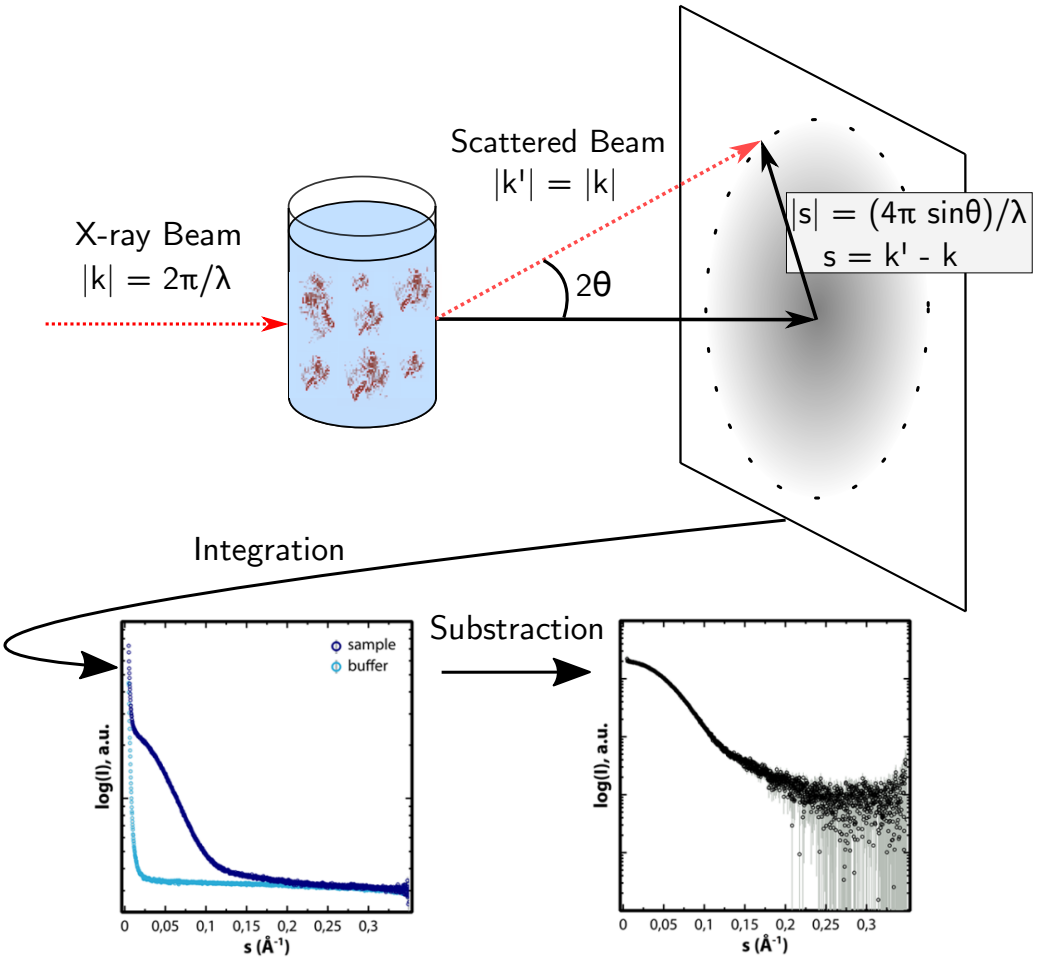


Figure 2.7. Schematic representation of a SAXS experiment. The incident X-ray beam is scattered when passing through. The resulting scattered photons are captured in a 2D detector and then integrated as a function of the scattering angle (2θ). This process is performed for the buffer and for the sample that are then subtracted to obtain the protein scattering profile.

between two points within the system separated by a distance r. $\gamma(r)$ is related with $I(s)$ by the next equation:

$$I(s) = 4\pi \int \gamma(r) r^2 \frac{\sin sr}{sr} \mathrm{d}r \tag{2.10}$$

Where $\frac{\sin sr}{sr}$ corresponds to the radially average $\langle e^{-isr} \rangle$ by applying the Debye formula to the isotropic scattering intensity recorded in the detector [42].

Experimentally, the final scattered curve is obtained by subtracting the SAXS profile of the buffer from that of the sample containing the macromolecule, see Figure 2.7.

2.5.3 Structure and form factors

There are two types of scattering interactions that contribute to the final SAXS intensity, the form factor, $I(s)$, and the structure factor, $S(s)$, $I_{total}(s) = I(s) \cdot S(s)$. The form factor corresponds to the scattering produced by the particle and therefore gives information about its structure, in our case the structure of the protein. The structure factor is the intensity due to the fact that our sample is composed by more than one particle. The structure factor gives a general information about how the particles are distributed within the sample. In structural biology, we are interested on the shape of the particle, so we try to minimize the effect of the structure factor by diluting the sample to reduce the inter-particle interference and making the structural factor as close as possible to 1. However, samples must contain enough protein ($>$ 1 mg/mL) to obtain a correct signal to noise. As a consequence, a dillution series of the protein of interest is performed, and the resulting curves are merged (or extrapolated to infinite dilution) to increase the signal to noise while avoiding the structure factor contribution.

2.5.4 Radius of gyration and forward scattering

The R_g is a parameter directly extracted from the SAXS curve that provides a measure of the overall size of the protein. The R_g is the average root-mean-square of the distances to the center of density in the macromolecule weighted by the scattering length density. R_g is an indicator of the compactness of the protein. With the same chain length, if the protein adopts a compact shape the resulting R_g will be smaller compared to an extended shape. The R_g can be estimated using Guinier's approximation [77]:

$$I(s) \approx I(0) \exp \frac{-s^2 R_g^2}{3} \tag{2.11}$$

This Gaussian function approximation of the intensity for small angles, $s < 1.3$, and plotted as $lnI(s)$ vs. s^2 should be a linear function from where the R_g can be obtained. In the same extrapolation process, the forward scattering (or the scattering at 0 angle), I(0), can be estimated. This parameter, which is proportional to the number of electrons of the particle, can be used to estimate the molecular weight and the oligomerization state of the macromolecule in solution.

2.5.5 Pair-wise distance distribution

For non-interacting particles in dilute solution the scattering intensity can be represented by an integral over the particle:

$$I(s) = 4\pi \int_0^{D_{max}} p(r) \frac{sin(sr)}{sr} dr \tag{2.12}$$

where r is the distance between two points scattered within the sample and D_{max} is the maximal dimension of the particle. $p(r) = \gamma(r) r^2$ represents the histogram

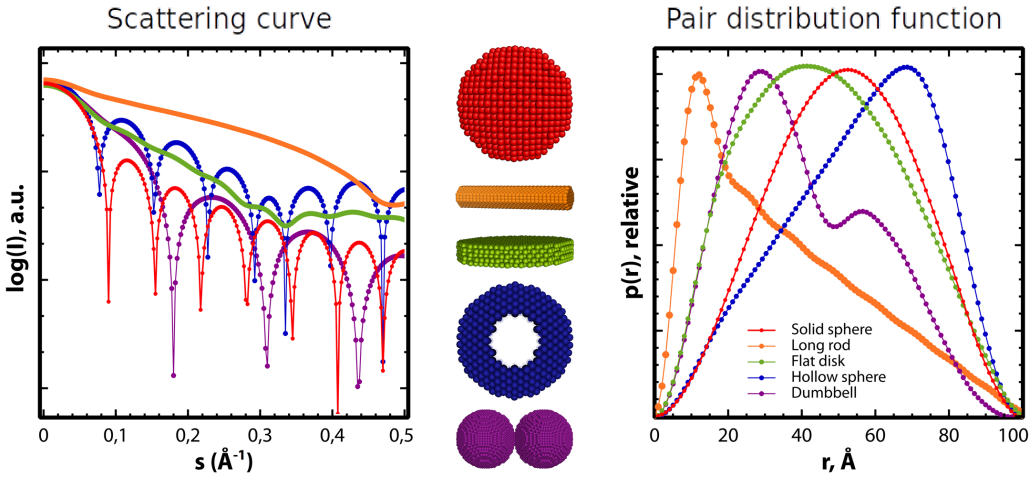


Figure 2.8. Scattering curves and pair-wise distance distribution functions for 5 particles (sphere, long rod, flat disk, hollow sphere, and dumbbell) with the same size and different shapes. From pair distribution functions is easier to intuitively identify the original shapes of the particles than using the scattering curve. Figure inspired from [113]

of distances between pairs of points within the macromolecule and can be obtained by the indirect Fourier transformation of the SAXS curve:

$$p(r) = \frac{1}{2\pi^2} \int_0^\infty srI(s) \sin srds \tag{2.13}$$

As $p(r)$ is described in real space, it is more intuitive to interpret the structural properties using $p(r)$ rather than using $I(s)$. This is illustrated in figure 2.8 where particles of the same size but different shapes yield distinct SAXS profiles and $p(r)$ functions. Notice that from $p(r)$, one can intuitively identify the shapes of the original particles.

2.5.6 SAXS applied to Intrinsically Disordered Proteins

One of the major advances of SAXS in the last decade has been its extension to address biomolecular dynamics [53, 18, 182, 17, 109, 106]. Although used in the past to study protein flexibility [6], the availability of robust protocols to interpret SAS data in terms of ensembles of conformations have generalized these studies and, therefore, have enriched the spectrum of applications of the technique [14].

The fact that IDPs sample an astronomical number of conformations has a strong impact on the scattering profiles measured and their comprehensive analysis in terms of structure. The experimental SAXS profile of an IDP corresponds to the average of all the conformations that the protein adopts in solution, inducing special features to the curves. Figure 2.9A displays the synthetic SAXS curves for seven conformations of p15PAF, a 111 residue-long IDP, selected from a large pool of 5,000 conformations [40]. The individual conformations display several

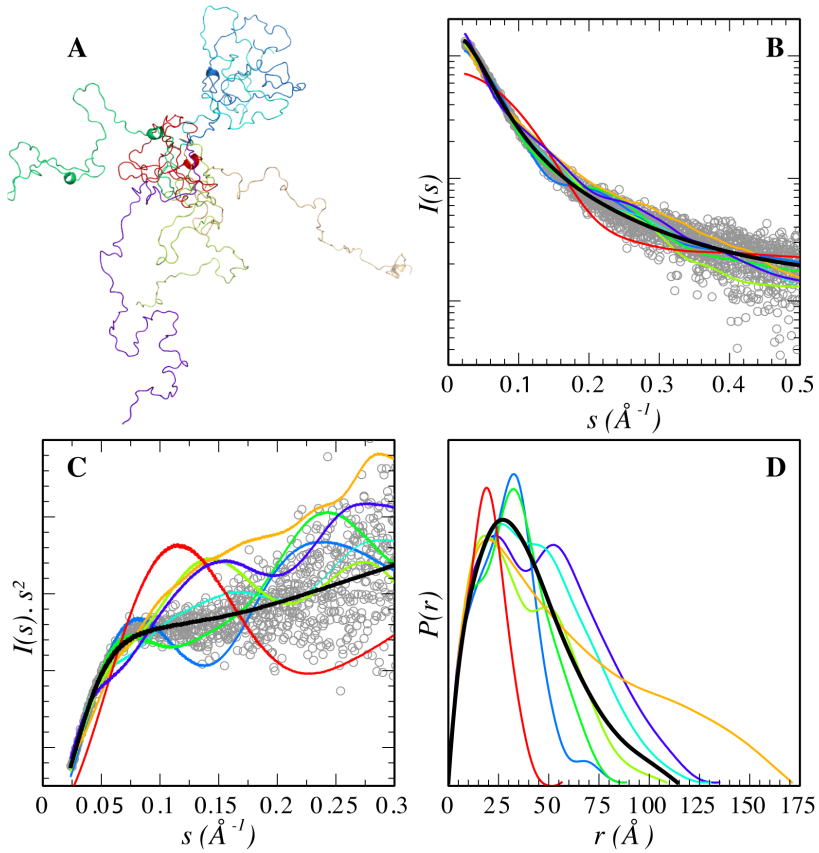


Figure 2.9. (A) Seven representative conformers randomly selected from an ensemble of 5.000 explicit all-atoms models generated for p15PAF [40]. Solid lines correspond to their computed curves (B) and Kratky plots (C) and are colored as in panel A. The average over the ensemble of 5,000 conformations yields a featureless curve that is in very good agreement with the experimental data (gray circles). (D) $p(r)$ functions computed for the 7 conformers and the complete ensembles in the same color code that in panels (A-C). Figure extracted from [32]

features along the complete momentum transfer range simulated. The initial part of the simulated curves, containing the lowest resolution structural information, presents distinct slopes indicating a large variety of possible sizes and shapes that an unstructured chain can adopt. The SAXS profile, obtained after averaging curves for the 5,000 conformations, presents a smoother behavior with essentially no features (Figure 2.9B).

Traditionally, Kratky plots ($I(s)\Delta s^2$ as a function of s) have been used to qualitatively identify disordered states and distinguish them from globular particles. The scattering intensity of a globular protein behaves approximately as $1/s^4$ conferring a bell-shaped Kratky plot with a well-defined maximum. Conversely, an ideal Gaussian chain has a $1/s^2$ dependence of $I(s)$ and therefore presents a plateau at large s values. In the case of a chain with no thickness, the Kratky plot also presents a plateau over a specific range of s, which is followed by a monotonic increase.

This last behavior is normally observed experimentally in unfolded proteins. The Kratky representation has the capacity to enhance particular features of scattering profiles that allows an easier identification of different degrees of compactness [53]. This is shown in Figure 2.9C where different degrees of compactness for the conformations are observed. Multi-domain proteins present in the same molecule a dual (folded/disordered) behavior and, consequently SAXS profiles and Kratky plots present contributions from both structurally distinct regions. Pair-wise distance distributions, $p(r)$, derived from disordered proteins also present specific properties (Figure 2.9D). The most characteristic feature is the smooth decrease towards large intramolecular distances. Maximum intramolecular distance, $Dmax$, are very large in disordered proteins. It is worth noting that due to the low population of highly extended conformations in the ensembles, experimental $Dmax$ values are systematically underestimated [14]. Unstructured proteins, due to the presence of extended conformations, are characterized by large average sizes compared to globular proteins. The radius of gyration, R_g, is the most common descriptor to quantify the overall size of molecules in solution and it is normally obtained using Guinier's approximation, eq. 2.11). Debye's approximation (eq. 2.14) can be more precise than Guinier's one to derive R_g values as its validity extends to larger momentum transfer ranges [24].

$$\frac{I(s)}{I(0)} = \frac{2}{x^2}\left(x - 1 + e^{-x}\right); x = s^2 \cdot R_g^2 \tag{2.14}$$

Alternatively, $p(r)$ function calculated from the complete scattering profile using a Fourier transformation also yields precise R_g values for disordered proteins. The experimental R_g is a single value representation of the size of the molecule, which for disordered states represents a z-average over all accessible conformations in solution [64]. The most common quantitative interpretation of R_g for unfolded proteins, which is based on Flory's studies in polymer science, relates this parameter to the length of the protein chain through a power law [237],

$$R_g = R_0 \cdot N^{\nu} \tag{2.15}$$

where N is the number of residues in the polymer chain, R_0 is a constant that depends on several factors, in particular, on the persistence length, and ν is an exponential scaling factor. For an excluded-volume polymer, Flory estimated ν to be ≈ 0.6, and more accurate theoretical estimates established a value of 0.588 [76]. A recent compilation of R_g values measured for 26 chemically denatured proteins sampling broad range of chain lengths found a ν value of 0.598 ± 0.028, and a R_0 value of 1.927 ± 0.27 [114]. The agreement between the ν value obtained experimentally and the theoretical models demonstrates the random coil nature of the chemically denatured proteins. However, the question whether the conformational sampling in the chemically denatured state is equivalent to that found for IDPs in native conditions must be clarified, see reference [114] and references therein. Using atomistic ensemble models of several disordered proteins, Flory's equation has been

parametrized for IDPs [13]:

$$R_g = (2.54 \pm 0.01) \cdot N^{(0.522 \pm 0.01)} \tag{2.16}$$

The exponential value obtained from the parametrization, $\nu = 0.522 \pm 0.01$, is notably smaller than that derived from the dataset of denatured proteins, $\nu = 0.598 \pm 0.028$, indicating that IDPs are more compact than chemically denatured proteins. This observation is in line with NMR studies that indicated that urea denatured proteins have an enhanced sampling (around 15%) of extended conformations compared with IDPs [144]. As some IDPs are expected to have certain populations of secondary or tertiary structure, this relationship can be used as an interpretative tool. Thus, deviations from expected IDP random coil model indicate enhanced degrees of compactness or extendedness within the protein.

2.6 Modelling Intrinsically Disordered Proteins

SAXS data and the majority of NMR structural parameters are ensemble averages. In order to fully exploit the structural and dynamic information encoded in experimental data, the use of theoretical models and computational methods is necessary. Indeed, as further discussed in the next section, the suitable coupling of experimental data and theoretical/computational methods is essential to build realistic models of IDPs in order to better understand their structural and dynamic properties. However, modelling disordered proteins is extremely challenging [247]. As mentioned above, IDPs present a relatively flat (non-funnelled) energy landscape, with an extremely large number of local minima separated by low-energy barriers. This, combined with their large size, makes the analysis of their energy landscape a challenging problem for computational methods.

In this section, we briefly present computational methods used to model and simulate IDPs. They are grouped in three categories: (1) knowledge-based approaches to build conformational ensemble models, (2) physics-based (traditional) methods to sample states and simulate dynamics, (3) robotics-inspired methods to explore the conformational space.

2.6.1 Knowledge-based approaches to build conformational ensemble models

Computational methods for structural investigations of IDPs are mainly aimed at producing an ensemble representation of disordered proteins. This requires an extensive and statistically correct exploration of the conformational space to obtain a representative set of states. Information extracted from an statistical analysis of known protein structures can be used for this purpose. The most representative knowledge-based method for the generation of atomistic models of disordered proteins is Flexible-Meccano (FM) [15, 163], although other similar methods have been described [103]. In FM, each conformation is built by assembling peptide plane units

in a consecutive manner using a residue-specific coil library derived from crystallographic structures. To avoid the collapse of the chain, a coarse-grained description of side chains is also used. Based on this set of conformations, experimentally measurable NMR parameters and SAXS curves can be estimated, which has permitted the validation of the resulting models. FM provides excellent models for the random-coil but do not capture structural features involving multiple consecutive residues, such as secondary structural elements. To solve this problem, FM allows to add the percentage of secondary structure population that the user considers appropriate to fit the experimental data. NMR observables are obtained using appropriate computational methods and compared with the experimental ones. If the result does not satisfy the process is restarted again adjusting the artificially added structure. Although this method has given good results, it is a time-consuming and laborious strategy [238, 40, 213]. FM has also been used with $\{\phi, \psi\}$ values derived from MD simulations [154]. The long-range contacts can also be simulated with FM, since it allows to force the contact between the residues of the chain and to see the effect in the experimental parameters [12]. In this thesis, we present an approach that exploits the structural information encoded in an extensive coil library of three-residue fragments (Chapter 3) to create IDP ensemble models that capture relevant structural features, therefore overcoming some of the limitations of FM. This method is explained in Chapter 5.

2.6.2 Physics-based methods to sample states and to simulate dynamics

Different methods based on physical models have been proposed to sample the conformational space of IDPs and to simulate their dynamic behaviour. The most frequently-used methods are based on molecular dynamics (MD) simulations. MD simulations analyze the evolution of the system under study by solving Newton's equations of motion [107, 174]. Theoretically, MD is a suitable method to correctly sample the conformational space of IDPs. Nevertheless, in practice, the high-dimensionality and the wideness of the energy landscape hampers its exhaustive exploration. Several approaches have been proposed to enhance conformational exploration with MD methods. A particularly effective one is Replica Exchange MD (REMD) that runs multiple simulations in parallel with different settings (usually different temperatures) and exchanges states between these processes [222, 246, 28]. Going further in this direction, a recent method called Multiscale Enhance Sampling (MSES) couples temperature replica exchange and Hamiltonian replica exchange, using a coarse-grained model to guide atomistic conformational sampling [125]. The performance of MD-based method can also be improved by the integration of experimental data to restrain the exploration of the most relevant regions of the conformational space [131, 44, 244].

Monte Carlo (MC) methods are a classical alternative to MD, being the Markov chain Metropolis scheme [146] the most widely used MC sampling technique [235]. The system is randomly perturbed and the new conformation is accepted with a

probability that depends on the energy change between the new conformation and the previous one. Particular mention deserves a recently proposed variant called Hamiltonian Switch Metropolis Monte Carlo (HS-MMC), which has been specially conceived to study IDRs tethered to globular domains. Proteins including IDRs present energy minima due to the contact of the disordered and ordered regions. To avoid being trapped in such minima, the HS-MMC switches between an all-atom Hamiltonian to an excluded volume Hamiltonian to push the IDR away from the ordered domain.

Both MD- and MC-based approaches may suffer from inaccuracies of current energy models, which are better suited to globular proteins and tend to provide structurally biased ensembles that do not properly reflect the conformational behaviour of unstructured proteins in solution [20, 83]. The development of more suitable force-fields and solvation models for IDPs are key issues for a correct performance of computational methods. Indeed, this is a very active field of research [234, 60].

When applied to large molecules, MD and MC methods are computationally demanding. Thus, parallel computing is almost mandatory. Basic MD and MC methods are sequential processes, but parallel computation is usually applied at a lower level for energy evaluation (and derivatives). More sophisticated variants of these methods, such as REMD, admit parallelization at a higher level. This will be discussed in Section 7.2.1.

2.6.3 Robotics-inspired methods to explore the conformational space

Algorithms originating from robotics, which compute feasible paths between two configurations for multi-body systems in a constrained space, have also been applied to model conformational transitions in biomolecules such as proteins and peptides [2, 71, 199]. One of these methods is the Rapidly-exploring Random Tree (RRT) algorithm [122], a path planning algorithm that can tackle complex problems in high-dimensional spaces. The basic principle of the RRT algorithm is to construct incrementally a tree whose origin is located at the initial configuration q_{init} to explore the space of accessible configurations and find a feasible path connecting q_{init} to the final configuration q_{goal}. A more sophisticated extension of the RRT, the Transition-based Rapidly explorer Random Tree (TRRT) algorithm [35], was designed to find the high-quality (low-cost) paths, when a cost/energy function to evaluate configurations is provided. The pseudo-code of TRRT is sketched in Algorithm 1. Note that the `TransitionTest` inside the TRRT is inspired from the Metropolis test in MC methods. The originality if TRRT is that the temperature parameter, which modulates the difficulty of this transition test, is a self-adaptive parameter that evolves in order to avoid local minima traps. This class of algorithms, and in particular the multi-tree version of the TRRT, will be further explained in Chapter 7.

Algorithm 1: Transition-based Rapidly-exploring Random Tree

input : the configuration space $\mathscr{C}$
the initial configuration q_{init} and the final configuration q_{goal}
output: the tree $\mathscr{T}$

```
1 T ← InitTree(q_init)
2 while not stoppingCriteria (T, q_goal) do
3     q_rand ← sampleRandomConf(C)
4     q_near ← findNearestNeighbor(T, q_rand)
5     q_new ← extend(q_near, q_rand)
6     if q_new ≠ null and
7     TransitionTest(cost(q_near), cost(q_new)) then
8         addNewNode(T, q_new)
9         addNewEdge(T, q_near, q_new)
```

2.7 Combined use of SAXS, NMR and computational methods

To better understand the structure and dynamics in IDPs, the combination of different experimental and computational methods is necessary. The complementarity between NMR and SAS is based on the distinct resolution of the information provided. Whereas SAS probes the overall properties of molecules, NMR information reports on atomic or residue-specific information. Therefore, the simultaneous description of both observables strongly suggests the appropriateness of the derived model. In that context, SAXS can be used to validate structural models of IDPs refined with NMR data. In this approach, the residue-specific conformational preferences of an IDP are refined using RDCs and CSs using Flexible-Meccano [15, 163]. The final model contains percentages of secondary structural elements in localized regions that have been imposed to properly describe the NMR data. The resulting ensemble can be validated by simply comparing the average SAXS curve computed from the ensemble with that experimentally measured one. This strategy has been applied to the partially folded Sendai virus PX [15], the transactivation domain of p53 [238], the K18 construct of Tau protein [155, 154], and the oncogene p15PAF [40]. A similar approach has been performed to study PaaA2 antitoxin [213]. In this last study, the NMR-derived ensemble was used as starting pool for a SAXS EOM refinement, demonstrating that the protein exists in solution as two preformed helices connected by a flexible linker.

The best manner to exploit the complementarity of both techniques is to integrate the experimental data into the same refinement protocol. Some of these integrative approaches have been applied to IDPs. One of them is ENSEMBLE, a program that derives ensembles of disordered proteins by collectively describing SAXS curves in addition to several NMR observables: CS, J-couplings, RDCs, PREs, Nuclear Overhauser effects, hydrodynamic radius, solvent accessibility restraints, hydrogen-exchange protection factors, and 15N R2 relaxation

rates [142, 119]. A large number of random structures are computed with FOLDTRAJ or TRADES [63, 62], and a Monte Carlo algorithm is used to select a subset of these structures that are collectively consistent with the experimental restraints. This subset is used as a basis for the generation of new structures, and the process is repeated until a final ensemble consistent with all of the experimental measurements is obtained. This approach addresses the intrinsic problem of under-restraining and consequent over-fitting by finding the smallest ensemble that is consistent with all experimental restraints imposed. ENSEMBLE has been applied to characterize the protein Sic1 and its hexaphosphorylated version pSic1 by combining SAXS data with several NMR parameters, including CS, PREs, RDCs, and ^{15}N R2 [148]. Moreover, a structural model of the complex between pSic1, which contains several binding regions, and its partner Cdc4 was generated by combining restraints of the free form of pSic1 with sparse NMR data of the complex suggesting a fuzzy interaction. ASTEROIDS is another program that allows the synergistic interpretation of NMR and SAXS data [98]. The power of ASTEROIDS is illustrated in a recent study of Tau and α-synuclein using NMR (CSs, RDCs, PREs) and SAXS data [194]. Using extensive cross-validation, the authors showed that five different types of independent experimental parameters are predicted more accurately by selected ensembles than by statistical coil descriptions. With this method, they could highlight that Tau and α-synuclein sample polyproline-II region in the aggregation-nucleation sites.

Chapter 3

Tripeptide database

Contents

3.1 Introduction

One of the basic components of the algorithms presented in following chapters of this manuscript is a database of three-residue fragments (called *tripeptides* from now on). We built this tripeptide database from a large set of experimentally determined high-resolution protein structures with the aim of exploiting the structural information encoded in these fragments. Note that the use of tripeptides enriches the structural information of the database compared with traditional amino acid-specific databases usually used for IDP conformational sampling [15]. With this additional information, we expected to build more accurate models of disordered proteins including partially formed secondary structural elements. Whereas libraries involving larger fragments have been shown to be powerful tools for the prediction of probable (stable) conformations of globular proteins and peptides [79, 116, 184, 8, 202, 140], our results (see Chapters 4 and 5) highlight that our extensive database of tripeptides is enough to accurately represent the conformational variability and local structural propensities in IDPs. Note that representing the conformational variability of disordered chains requires a broad sampling of structures, which would not be guaranteed using databases of larger (penta- or hepta-peptide) fragments. In this regard, tripeptides emerge as an optimal compromise for our purpose of exploring the conformational sampling in IDPs.

This chapter explains the building process of the tripeptide database. Then, two analyses of the tripeptide database are presented to better understand the importance of the local sequence context and the structure of the flanking residues.

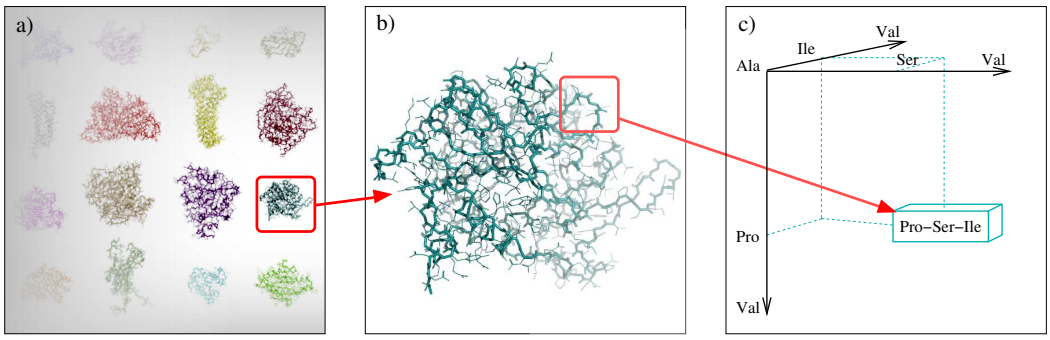


Figure 3.1. Construction of the tripeptide database: (a) A non-redundant set of experimentally determined protein structures is used as input. (b) For each protein, fragments of three consecutive residues (called tripeptides) are analyzed. (c) The structural information is stored in a database containing one record for each tripeptide (8,000 in total).

3.2 Database construction

The tripeptide database was built from a large set of experimentally-determined high-resolution protein structures. We used the SCOPe [67] 2.06 release, with entries having less than 95% sequence identity to each other. A total of 8,907,065 three-residue fragments were extracted from these protein structures and classified on the basis of their sequence (8,000 tripeptide classes). The database construction process is illustrated in Figure 3.1. The number of their instances ranges between 9 for the less frequent tripeptide (Cys-Cys-Trp) to 4,512 for the most frequent one (Ala-Ala-Ala). The average number of instances is about 688. The tripeptides suitable for IDP modeling are mainly those contained in random-coil regions of globular proteins. Conformations adopted by residues were assigned using the program DSSP [105], which allowed us to filter out fragments corresponding to α-helices and β-strands. More precisely, we removed all tripeptides containing at least one residue involved in these types of secondary structures (i.e. DSSP types H, G, I, E and B) from the database. This applied to approximately 60% of the total number of tripeptides extracted from the SCOPe database. The remaining 40% of the tripeptides (3,645,381), which contained residues in coil/loop regions (i.e. DSSP types T, S and blank/C), were included in the coil database.

We adopt a rigid geometry simplification [198], which assumes constant bond lengths and angles. Indeed, the standard deviation for the bond lengths and the bond angles in our database is two orders of magnitude smaller that their average value, and therefore, we can neglect their variation. Therefore, the only variables required to determine the conformation of a protein backbone correspond to the ω, ϕ and ψ dihedral angles of each amino acid residue. The values of ω usually fluctuate around 180°, corresponding to the *trans* conformation of the peptide bond. Values of ω around 0° are much less frequent, corresponding to the *cis* conformation. *cis* conformations are mainly observed in proline residues. A specific section of this chapter (Section 3.5) has been devoted to the study of proline *cis* conformations and

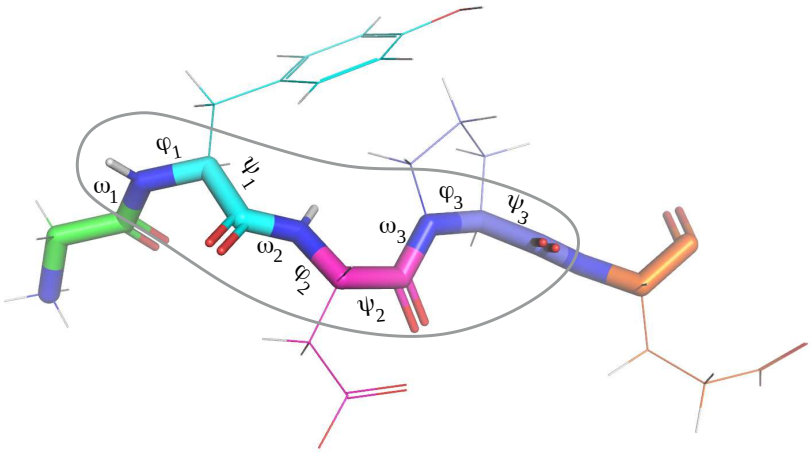


Figure 3.2. Illustration of a protein fragment involving 5 residues. Each residue is represented using a different color for the carbon atoms. The backbone is represented using thicker lines. Considering constant bond lengths, bond angles and peptide bond torsions, the protein backbone conformation can be defined from three angles (ω, ϕ and ψ) for each residue. The gray line indicates the tripeptide fragment.

the effect that the nature of the flanking residues has on its population. Depending on the value of the ϕ and ψ angles, we classify the residues as belonging to the α, β or γ region. The regions of the Ramachandran plot that we use in this thesis are shown in Figure 3.3 and are defined using the same angular intervals as in previous studies [164]:

$$\begin{aligned}
&\alpha : \phi \leq 0^\circ \ , \ -120^\circ < \psi \leq 50^\circ \\
&\beta : \phi \leq 0^\circ \ , \ 50^\circ < \psi \leq 240^\circ \\
&\gamma : \phi > 0^\circ \ .
\end{aligned}$$

The database stores these angular values for each tripeptide extracted from the ensemble of protein structures (*i.e.*, nine angles for each tripeptide). Figure 3.2 represents a protein fragment involving 5 residues, from which 3 tripeptides are extracted. The angles defining the conformation of each residue are represented on the corresponding bonds.

3.3 Sequence-dependent structural preferences

The nature of the neighboring residues has a strong impact on the distribution of the ϕ-ψ angles of the observed residue. The conformation of a given residue depends on its physico-chemical properties as well as these of the flanking residues. This is illustrated for four tripeptides all having alanine as a middle residue (X-ALA-Y) in Figure 3.4. One can clearly observe that a change on the flanking residues directly affects the ϕ-ψ distribution of the central residue. When the alanine is preceded by

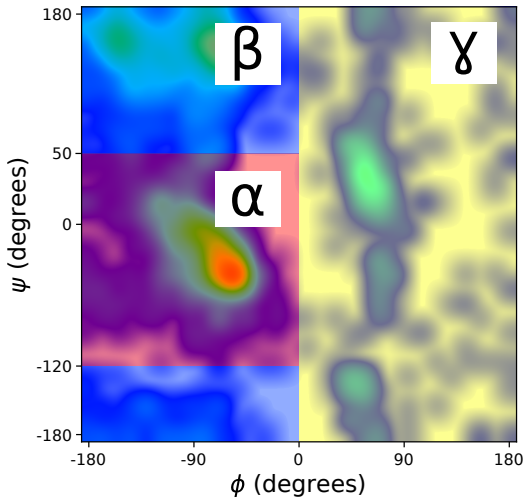


Figure 3.3. The Ramachandran regions were identified using definitions in related work [164]. Concretely, $[\alpha : \phi \leq 0°; -120° \leq \psi < 50°;]$, $[\beta : \phi < 0°; 50° < \psi \leq 240°]$, $[\gamma : \phi > 0]$.

other alanine and followed by histidine almost all the Ramachandran angles of the central alanine are in the extended region β region. However, when the flanking residues are threonine and serine, the central alanine has a strong tendency to be in the α region.

This observation is in agreement with previous studies [88], where the information of the closest neighbors is enough to well reproduce the RDCs of an IDP. They show that the influence of the second nearest residues is weak, unless local structure is present. Note that the impact of relatively close residues in the sequence is much more important in structured proteins because they are spatially closer and may form hydrogen-bonds, like in the case of α-helices.

Finally, we should mention that in addition to the sequence-dependent structural preferences encoded in the database, the sampling method presented in Chapter 5 uses a coarse-grained model for the side chains, that partially captures repulsive interactions with non adjacent residues.

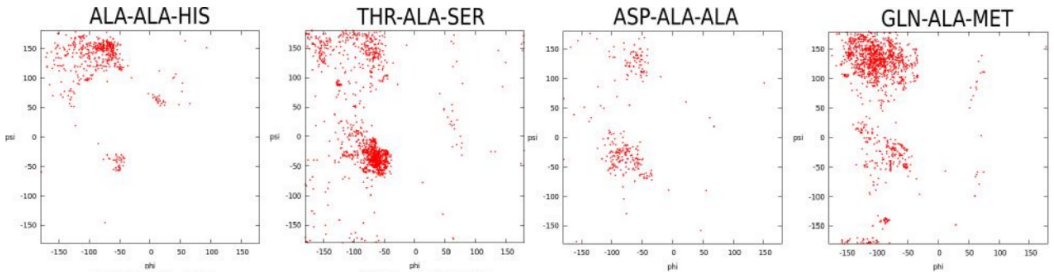


Figure 3.4. Distributions of the ϕ-ψ angles of the central residue for four tripeptides having alanine as central residue. (a) Ala-Ala-His. (b) Thr-Ala-Ser. (c) Asp-Ala-Ala. (d) Gln-Ala-Met.

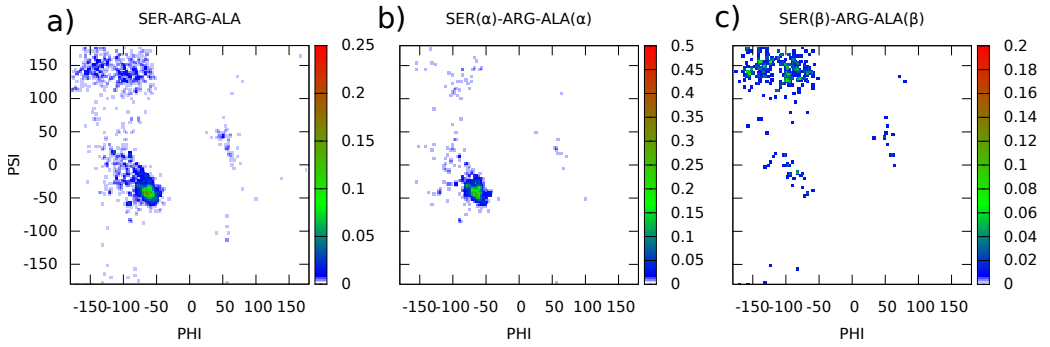


Figure 3.5. Distributions of the ϕ-ψ angles of the central residue in a tripeptide, Ser-Arg-Ala, depending on the structure of the neighboring residues. (a) All the ϕ-ψ angles for the central residue Arg, independently on the structure of Ser and Ala. (b) Angles for Arg when Ser and Ala are in the α region. (c) Values for Arg when Ser and Ala are in the β/polyproline-II region.

3.4 Context-dependent structural preferences

In addition to the sequence-dependent structural preferences, the conformation of the neighboring residues, ϕ-ψ angles, also has a direct influence in the structural propensities of a given residue. This is illustrated for the Ser-Arg-Ala tripeptide in Figure 3.5. One can clearly observe that when the ϕ-ψ angles of the neighboring Ser and Ala residues are constrained to be in the α region, the central Arg residue has a high probability to be also in this region. The same happens for the β/polyproline-II region corresponding to extended conformations. This result shows that the tripeptide database displays some degree of structural cooperativity that is incorporated in our structural models of IDPs thanks to the information encoded in the tripeptide database.

3.5 Effects of the neighboring residues on the *cis*/*trans* proline isomerization

The tripeptide database has multiple applications in structural biology. To illustrate its usefulness, we statistically quantified the neighboring effects in the proline *cis*/*trans* equilibrium. This analysis is part of a broader study about this isomerization in poly-proline tracts present in the protein huntingtin (manuscript in preparation).

Our database, containing 754,308 proline-centered tripeptides, had examples for the 400 existing X-Pro-Y tripeptides (see Table 3.1). The most represented tripeptide was Ala-Pro-Gly with 7,310 examples, while Trp-Pro-Met was the less represented tripeptide with only 26 examples. The average number of tripeptides was 1,886. The population of *cis* conformations (the *cis-trans* isomerization in checked for the peptide bond X-Pro) derived from our database, 6.48%, is slightly larger to these found in [96] and [201] with 4.63% and 4.50%, respectively. The

difference probably arises from the nature and size of the databases used or the analysis. We could derive reliable *cis* populations for the 400 existing X-Pro-Y. According to our calculations, we could identify the Trp-Pro-Arg tripetide as the most prone to present *cis* conformations, 56.10%, followed by Trp-Pro-Pro and Trp-Pro-Ile with 47.30% and 42.60% respectively. On the other extreme, no *cis* conformations were found for tripeptides Cys-Pro-Met, Gln-Pro-Cys, Met-Pro-Asp and Met-Pro-Cys.

In the context of the study of poly-proline homo-repeats, it is important to analyse the effects of neighboring prolines in the isomerization. The populations found in our database for the X-Pro-Pro and Pro-Pro-X tripeptides were 7.05 % and 8.56 %, respectively. These percentages indicate that proline in $i-1$ position slightly favours the probability of the *cis* isomer compared with a proline in position $i+1$, and they are similar to the average *cis* in the whole database. Interestingly, when both flanking positions are occupied by proline, Pro-Pro-Pro, the percentage of the *cis* isomer decreases dramatically to 0.80%. This result suggests a strong bias towards the *trans* isomer in poly-P tracts. This observation is in excellent agreement with the experimental observations done in the group at the CBS Montpellier. Concretely, individual prolines were isotopically labeled in two poly-proline tracts of huntingtin, with 11 and 3 consecutive prolines, and their *cis*/*trans* populations were quantified in a residue-specific manner (manuscript in preparation). Only the first proline of both poly-proline tracts displayed a measurable population of the *cis* conformation, while this isomer could not be detected ($<$ 1.5 %) for all the other prolines, including the last one. The experimental results suggest a cooperative effect within poly-proline tracts that strongly disfavours the *cis* population, in very good agreement with the calculations from the database. This cooperative effect presents a defined directionality (from N- to C-terminus) with the first proline (X-Pro-Pro) as the only one experiencing the *cis*/*trans* isomerization. Notice that according to our statistical calculations, no significant differences should be observed between the first and the last proline of a poly-Proline tract.

3.6 Discussion about the structural filtering in the tripeptide database

We performed a simple experiment to demonstrate the interest of using a database built from fragments in coil regions with respect to a database including fragments placed in secondary structure elements, in the context of IDP modeling. The experiment consisted of generating conformational ensembles, using an approach similar to Flexible-Meccano (FM) [15, 163], for an intrinsically disordered fragment of the N-tail protein from measles virus (ntailMV) [97] using data (distributions of ϕ-ψ angles) with and without secondary structure filtering. The results are presented in Figure 3.6. The experimental N-HN RDCs profile (black solid lines) is compared with the theoretical RDCs computed using the unfiltered database (red solid line) and the coil database (blue solid line). It is clear that the sampling performed using

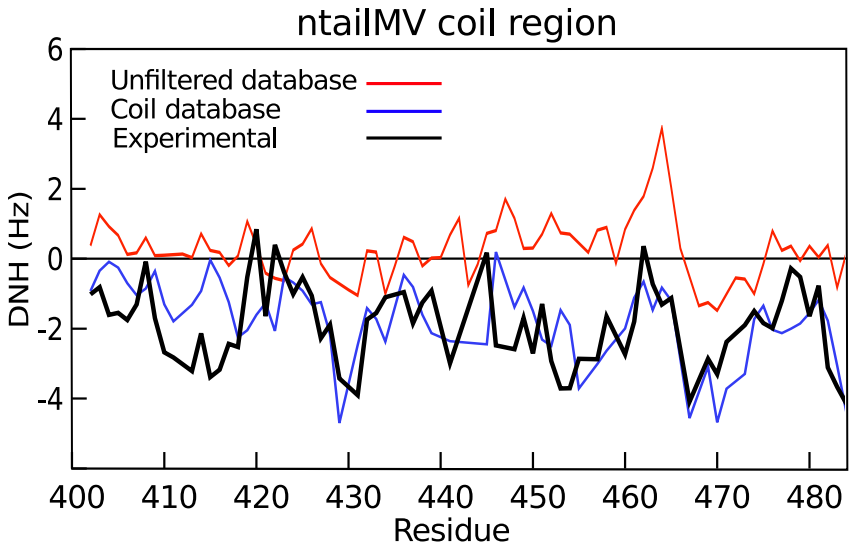


Figure 3.6. For the coil region of ntailMV: Experimental N-HN RDCs (black solid lines) compared with the theoretical RDCs computed from FM-like sampling using the unfiltered database (red solid line) and the coil database (blue solid line).

the unfiltered database overestimates the α-helix population while the coil database nicely reproduces the experimentally determined RDCs for this fragment.

It is important to highlight that although the database only includes fragments extracted from coil regions (i.e. tripeptides contained in secondary structures elements according to DSSP assignments are filtered out), secondary structures can still be sampled from it. Indeed, values of ϕ-ψ angles corresponding to α-helical regions or extended (β or polyproline-II) regions are still contained in the database despite the structural filtering, since DSSP also considers other criteria (related to hydrogen bonds) to assign secondary structures. Therefore, protein conformations sampled from partially-overlapping tripeptides extracted from the coil database can contain partially-formed or even fully-formed (canonical) secondary structure elements. This will be shown through the examples presented in Chapters 5 and 6.

Table 3.1: List of all 400 possible X-Pro-Y tri-peptides (X and Y = any amino acid) extracted from a coil database built from the crystallographic entries in PDB with a resolution $\leq$ 2.0Å. The total occurrence of each tripeptide (*#total*), as well as the number of prolines in *cis* (*#cis*) or *trans* (*#trans*) configuration and the calculated percentage of prolines in *cis* configuration (*%cis*) in the respective tripeptide are listed. Tripeptides with less than 100 instances (low confidence level) are highlighted in grey.

Sequence	*#cis*	*#trans*	*#Total*	*%cis*	Sequence	*#cis*	*#trans*	*#Total*	*%cis*
ALA_PRO_ALA	184	3789	3973	4.63	ASN_PRO_PRO	280	2253	2533	11.10
ALA_PRO_ARG	136	2273	2409	5.65	ASN_PRO_SER	133	2874	3007	4.42
ALA_PRO_ASN	134	2263	2397	5.59	ASN_PRO_THR	298	2669	2967	10.00
ALA_PRO_ASP	136	3442	3578	3.80	ASN_PRO_TRP	292	489	781	37.40
ALA_PRO_CYS	29	427	456	6.36	ASN_PRO_TYR	312	1368	1680	18.60
ALA_PRO_GLU	207	3555	3762	5.50	ASN_PRO_VAL	221	1671	1892	11.70
ALA_PRO_GLN	173	1488	1661	10.40	ASP_PRO_ALA	80	5403	5483	1.46
ALA_PRO_GLY	356	6954	7310	4.87	ASP_PRO_ARG	41	3816	3857	1.06
ALA_PRO_HIS	83	1591	1674	4.96	ASP_PRO_ASN	113	4237	4350	2.60
ALA_PRO_ILE	64	1854	1918	3.34	ASP_PRO_ASP	58	3602	3660	1.58
ALA_PRO_LEU	344	3974	4318	7.97	ASP_PRO_CYS	3	423	426	0.70
ALA_PRO_LYS	94	1726	1820	5.16	ASP_PRO_GLU	119	3547	3666	3.25
ALA_PRO_MET	15	689	704	2.13	ASP_PRO_GLN	57	2071	2128	2.68
ALA_PRO_PHE	224	1533	1757	12.70	ASP_PRO_GLY	139	1938	2077	6.69
ALA_PRO_PRO	69	2699	2768	2.49	ASP_PRO_HIS	18	1238	1256	1.43
ALA_PRO_SER	156	3612	3768	4.14	ASP_PRO_ILE	138	1160	1298	10.60
ALA_PRO_THR	110	2107	2217	4.96	ASP_PRO_LEU	88	2926	3014	2.92
ALA_PRO_TRP	57	1019	1076	5.30	ASP_PRO_LYS	47	2851	2898	1.62
ALA_PRO_TYR	213	1586	1799	11.8	ASP_PRO_MET	28	601	629	4.45
ALA_PRO_VAL	109	2598	2707	4.03	ASP_PRO_PHE	61	1388	1449	4.21
ARG_PRO_ALA	67	2917	2984	2.25	ASP_PRO_PRO	182	1813	1995	9.12
ARG_PRO_ARG	94	1118	1212	7.76	ASP_PRO_SER	67	6095	6162	1.09
ARG_PRO_ASN	110	1177	1287	8.55	ASP_PRO_THR	32	3391	3423	0.94
ARG_PRO_ASP	313	3199	3512	8.91	ASP_PRO_TRP	43	543	586	7.34
ARG_PRO_CYS	8	176	184	4.35	ASP_PRO_TYR	141	1257	1398	10.10
ARG_PRO_GLU	63	2940	3003	2.10	ASP_PRO_VAL	195	2324	2519	7.74
ARG_PRO_GLN	49	1176	1225	4.00	CYS_PRO_ALA	27	1117	1144	2.36
ARG_PRO_GLY	183	4116	4299	4.26	CYS_PRO_ARG	21	663	684	3.07
ARG_PRO_HIS	28	819	847	3.31	CYS_PRO_ASN	19	702	721	2.64
ARG_PRO_ILE	45	827	872	5.16	CYS_PRO_ASP	32	1003	1035	3.09
ARG_PRO_LEU	284	2395	2679	10.60	CYS_PRO_CYS	3	84	87	3.45
ARG_PRO_LYS	124	1026	1150	10.80	CYS_PRO_GLU	6	951	957	0.62
ARG_PRO_MET	26	424	450	5.78	CYS_PRO_GLN	5	242	247	2.02
ARG_PRO_PHE	213	1342	1555	13.70	CYS_PRO_GLY	61	951	1012	6.03
ARG_PRO_PRO	167	1599	1766	9.46	CYS_PRO_HIS	8	362	370	2.16
ARG_PRO_SER	68	2139	2207	3.08	CYS_PRO_ILE	12	499	511	2.35
ARG_PRO_THR	108	1996	2104	5.13	CYS_PRO_LEU	15	533	548	2.74
ARG_PRO_TRP	29	632	661	4.39	CYS_PRO_LYS	2	640	642	0.31
ARG_PRO_TYR	53	804	857	6.18	CYS_PRO_MET	0	278	278	0.00
ARG_PRO_VAL	75	1542	1617	4.64	CYS_PRO_PHE	58	502	560	10.40
ASN_PRO_ALA	120	2717	2837	4.23	CYS_PRO_PRO	148	658	806	18.40
ASN_PRO_ARG	137	1637	1774	7.72	CYS_PRO_SER	24	660	684	3.51
ASN_PRO_ASN	114	3357	3471	3.28	CYS_PRO_THR	7	634	641	1.09
ASN_PRO_ASP	21	4629	4650	0.45	CYS_PRO_TRP	4	272	276	1.45
ASN_PRO_CYS	33	476	509	6.48	CYS_PRO_TYR	22	433	455	4.84
ASN_PRO_GLU	49	3600	3649	1.34	CYS_PRO_VAL	11	659	670	1.64
ASN_PRO_GLN	23	1822	1845	1.25	GLU_PRO_ALA	194	1449	1643	11.80
ASN_PRO_GLY	120	2670	2790	4.30	GLU_PRO_ARG	281	1572	1853	15.20
ASN_PRO_HIS	107	1168	1275	8.39	GLU_PRO_ASN	582	1308	1890	30.80
ASN_PRO_ILE	126	1044	1170	10.80	GLU_PRO_ASP	63	2341	2404	2.62
ASN_PRO_LEU	256	2243	2499	10.20	GLU_PRO_CYS	8	521	529	1.51
ASN_PRO_LYS	25	2344	2369	1.06	GLU_PRO_GLU	253	2597	2850	8.88
ASN_PRO_MET	23	719	742	3.10	GLU_PRO_GLN	108	555	663	16.30
ASN_PRO_PHE	102	1464	1566	6.51	GLU_PRO_GLY	176	3378	3554	4.95

Sequence	#cis	#trans	#Total	%cis
GLU_PRO_HIS	130	867	997	13.00
GLU_PRO_ILE	69	1664	1733	3.98
GLU_PRO_LEU	161	2917	3078	5.23
GLU_PRO_LYS	111	1615	1726	6.43
GLU_PRO_MET	37	747	784	4.72
GLU_PRO_PHE	220	1250	1470	15.00
GLU_PRO_PRO	221	1956	2177	10.20
GLU_PRO_SER	373	2207	2580	14.50
GLU_PRO_THR	215	1502	1717	12.50
GLU_PRO_TRP	63	444	507	12.40
GLU_PRO_TYR	329	1204	1533	21.50
GLU_PRO_VAL	875	2012	2887	30.30
GLN_PRO_ALA	126	1499	1625	7.75
GLN_PRO_ARG	55	763	818	6.72
GLN_PRO_ASN	36	976	1012	3.56
GLN_PRO_ASP	73	3423	3496	2.09
GLN_PRO_CYS	0	326	326	0.00
GLN_PRO_GLU	36	2259	2295	1.57
GLN_PRO_GLN	77	871	948	8.12
GLN_PRO_GLY	87	3572	3659	2.38
GLN_PRO_HIS	14	874	888	1.58
GLN_PRO_ILE	29	1262	1291	2.25
GLN_PRO_LEU	44	2500	2544	1.73
GLN_PRO_LYS	21	1278	1299	1.62
GLN_PRO_MET	51	289	340	15.00
GLN_PRO_PHE	78	961	1039	7.51
GLN_PRO_PRO	100	1187	1287	7.77
GLN_PRO_SER	83	1907	1990	4.17
GLN_PRO_THR	58	1185	1243	4.67
GLN_PRO_TRP	14	548	562	2.49
GLN_PRO_TYR	50	562	612	8.17
GLN_PRO_VAL	94	2126	2220	4.23
GLY_PRO_ALA	111	2588	2699	4.11
GLY_PRO_ARG	179	2548	2727	6.56
GLY_PRO_ASN	178	3485	3663	4.86
GLY_PRO_ASP	203	2975	3178	6.39
GLY_PRO_CYS	65	498	563	11.50
GLY_PRO_GLU	77	2989	3066	2.51
GLY_PRO_GLN	88	1427	1515	5.81
GLY_PRO_GLY	310	4321	4631	6.69
GLY_PRO_HIS	106	1276	1382	7.67
GLY_PRO_ILE	96	1304	1400	6.86
GLY_PRO_LEU	285	3912	4197	6.79
GLY_PRO_LYS	87	2162	2249	3.87
GLY_PRO_MET	76	669	745	10.20
GLY_PRO_PHE	397	1155	1552	25.60
GLY_PRO_PRO	94	1581	1675	5.61
GLY_PRO_SER	187	2526	2713	6.89
GLY_PRO_THR	179	3103	3282	5.45
GLY_PRO_TRP	121	631	752	16.10
GLY_PRO_TYR	447	951	1398	32.00
GLY_PRO_VAL	157	2596	2753	5.70
HIS_PRO_ALA	111	1921	2032	5.46
HIS_PRO_ARG	30	963	993	3.02
HIS_PRO_ASN	16	1947	1963	0.81
HIS_PRO_ASP	72	2392	2464	2.92
HIS_PRO_CYS	8	209	217	3.69
HIS_PRO_GLU	26	2239	2265	1.15
HIS_PRO_GLN	47	665	712	6.60
HIS_PRO_GLY	47	3177	3224	1.46
HIS_PRO_HIS	27	502	529	5.10
HIS_PRO_ILE	37	575	612	6.05
HIS_PRO_LEU	75	1662	1737	4.32

Sequence	#cis	#trans	#Total	%cis
HIS_PRO_LYS	24	1289	1313	1.83
HIS_PRO_MET	46	865	911	5.05
HIS_PRO_PHE	73	1314	1387	5.26
HIS_PRO_PRO	193	989	1182	16.30
HIS_PRO_SER	433	1990	2423	17.90
HIS_PRO_THR	97	1343	1440	6.74
HIS_PRO_TRP	0	489	489	0.00
HIS_PRO_TYR	46	1170	1216	3.78
HIS_PRO_VAL	60	1109	1169	5.13
ILE_PRO_ALA	119	2945	3064	3.88
ILE_PRO_ARG	37	1353	1390	2.66
ILE_PRO_ASN	33	2023	2056	1.61
ILE_PRO_ASP	40	3243	3283	1.22
ILE_PRO_CYS	28	910	938	2.99
ILE_PRO_GLU	48	3138	3186	1.51
ILE_PRO_GLN	196	1617	1813	10.80
ILE_PRO_GLY	203	3379	3582	5.67
ILE_PRO_HIS	43	1200	1243	3.46
ILE_PRO_ILE	8	1340	1348	0.59
ILE_PRO_LEU	27	1923	1950	1.38
ILE_PRO_LYS	24	2222	2246	1.07
ILE_PRO_MET	5	617	622	0.80
ILE_PRO_PHE	14	1561	1575	0.89
ILE_PRO_PRO	56	2626	2682	2.09
ILE_PRO_SER	284	2058	2342	12.10
ILE_PRO_THR	119	1871	1990	5.98
ILE_PRO_TRP	101	645	746	13.50
ILE_PRO_TYR	47	1099	1146	4.10
ILE_PRO_VAL	51	1553	1604	3.18
LEU_PRO_ALA	177	5669	5846	3.03
LEU_PRO_ARG	94	2714	2808	3.35
LEU_PRO_ASN	50	3195	3245	1.54
LEU_PRO_ASP	56	6276	6332	0.88
LEU_PRO_CYS	41	553	594	6.90
LEU_PRO_GLU	141	6134	6275	2.25
LEU_PRO_GLN	142	1910	2052	6.92
LEU_PRO_GLY	291	6735	7026	4.14
LEU_PRO_HIS	28	1336	1364	2.05
LEU_PRO_ILE	109	2362	2471	4.41
LEU_PRO_LEU	373	3592	3965	9.41
LEU_PRO_LYS	97	3547	3644	2.66
LEU_PRO_MET	39	721	760	5.13
LEU_PRO_PHE	192	2228	2420	7.93
LEU_PRO_PRO	106	4539	4645	2.28
LEU_PRO_SER	127	5315	5442	2.33
LEU_PRO_THR	57	3721	3778	1.51
LEU_PRO_TRP	36	738	774	4.65
LEU_PRO_TYR	94	2098	2192	4.29
LEU_PRO_VAL	186	3462	3648	5.10
LYS_PRO_ALA	117	1977	2094	5.59
LYS_PRO_ARG	55	1194	1249	4.40
LYS_PRO_ASN	315	1619	1934	16.30
LYS_PRO_ASP	46	3039	3085	1.49
LYS_PRO_CYS	281	239	520	54.00
LYS_PRO_GLU	76	3742	3818	1.99
LYS_PRO_GLN	31	1294	1325	2.34
LYS_PRO_GLY	89	6721	6810	1.31
LYS_PRO_HIS	55	604	659	8.35
LYS_PRO_ILE	59	1185	1244	4.74
LYS_PRO_LEU	191	3006	3197	5.97
LYS_PRO_LYS	161	1871	2032	7.92
LYS_PRO_MET	46	931	977	4.71
LYS_PRO_PHE	326	1974	2300	14.20

Sequence	$\#cis$	$\#trans$	$\#Total$	$\%cis$
LYS_PRO_PRO	54	1479	1533	3.52
LYS_PRO_SER	104	3120	3224	3.23
LYS_PRO_THR	33	1709	1742	1.89
LYS_PRO_TRP	36	728	764	4.71
LYS_PRO_TYR	94	945	1039	9.05
LYS_PRO_VAL	86	2779	2865	3.00
MET_PRO_ALA	15	1026	1041	1.44
MET_PRO_ARG	9	496	505	1.78
MET_PRO_ASN	19	720	739	2.57
MET_PRO_ASP	0	970	970	0.00
MET_PRO_CYS	0	95	95	0.00
MET_PRO_GLU	85	728	813	10.50
MET_PRO_GLN	8	629	637	1.26
MET_PRO_GLY	118	1400	1518	7.77
MET_PRO_HIS	42	450	492	8.54
MET_PRO_ILE	14	436	450	3.11
MET_PRO_LEU	9	785	794	1.13
MET_PRO_LYS	8	645	653	1.23
MET_PRO_MET	13	266	279	4.66
MET_PRO_PHE	20	435	455	4.40
MET_PRO_PRO	12	574	586	2.05
MET_PRO_SER	33	562	595	5.55
MET_PRO_THR	23	465	488	4.71
MET_PRO_TRP	1	935	936	0.11
MET_PRO_TYR	21	414	435	4.83
MET_PRO_VAL	11	564	575	1.91
PHE_PRO_ALA	154	2578	2732	5.64
PHE_PRO_ARG	167	967	1134	14.70
PHE_PRO_ASN	258	1513	1771	14.60
PHE_PRO_ASP	142	2902	3044	4.66
PHE_PRO_CYS	18	184	202	8.91
PHE_PRO_GLU	1123	2453	3576	31.40
PHE_PRO_GLN	117	1211	1328	8.81
PHE_PRO_GLY	298	3777	4075	7.31
PHE_PRO_HIS	167	894	1061	15.70
PHE_PRO_ILE	62	1629	1691	3.67
PHE_PRO_LEU	221	1715	1936	11.40
PHE_PRO_LYS	55	1850	1905	2.89
PHE_PRO_MET	38	160	198	19.20
PHE_PRO_PHE	130	880	1010	12.90
PHE_PRO_PRO	211	1735	1946	10.80
PHE_PRO_SER	198	1863	2061	9.61
PHE_PRO_THR	84	1489	1573	5.34
PHE_PRO_TRP	34	436	470	7.23
PHE_PRO_TYR	62	809	871	7.12
PHE_PRO_VAL	128	1806	1934	6.62
PRO_PRO_ALA	154	2115	2269	6.79
PRO_PRO_ARG	72	1183	1255	5.74
PRO_PRO_ASN	65	1464	1529	4.25
PRO_PRO_ASP	78	2003	2081	3.75
PRO_PRO_CYS	188	289	477	39.40
PRO_PRO_GLU	164	3497	3661	4.48
PRO_PRO_GLN	107	1498	1605	6.67
PRO_PRO_GLY	206	4761	4967	4.15
PRO_PRO_HIS	73	1312	1385	5.27
PRO_PRO_ILE	50	1213	1263	3.96
PRO_PRO_LEU	1060	2576	3636	29.20
PRO_PRO_LYS	90	1949	2039	4.41
PRO_PRO_MET	23	752	775	2.97
PRO_PRO_PHE	511	1195	1706	30.00
PRO_PRO_PRO	25	3074	3099	0.81
PRO_PRO_SER	110	2720	2830	3.89
PRO_PRO_THR	198	1748	1946	10.20

Sequence	$\#cis$	$\#trans$	$\#Total$	$\%cis$
PRO_PRO_TRP	29	450	479	6.05
PRO_PRO_TYR	167	1477	1644	10.20
PRO_PRO_VAL	78	1571	1649	4.73
SER_PRO_ALA	271	2238	2509	10.80
SER_PRO_ARG	99	1583	1682	5.89
SER_PRO_ASN	137	2421	2558	5.36
SER_PRO_ASP	129	4183	4312	2.99
SER_PRO_CYS	92	390	482	19.10
SER_PRO_GLU	53	2837	2890	1.83
SER_PRO_GLN	214	1324	1538	13.90
SER_PRO_GLY	139	2931	3070	4.53
SER_PRO_HIS	53	896	949	5.58
SER_PRO_ILE	213	1359	1572	13.50
SER_PRO_LEU	143	3239	3382	4.23
SER_PRO_LYS	95	1581	1676	5.67
SER_PRO_MET	43	655	698	6.16
SER_PRO_PHE	166	2835	3001	5.53
SER_PRO_PRO	425	1428	1853	22.90
SER_PRO_SER	112	3270	3382	3.31
SER_PRO_THR	606	1902	2508	24.20
SER_PRO_TRP	28	805	833	3.36
SER_PRO_TYR	467	1403	1870	25.00
SER_PRO_VAL	762	2157	2919	26.10
THR_PRO_ALA	649	2713	3362	19.30
THR_PRO_ARG	151	1671	1822	8.29
THR_PRO_ASN	275	1805	2080	13.20
THR_PRO_ASP	26	3882	3908	0.67
THR_PRO_CYS	19	551	570	3.33
THR_PRO_GLU	65	3300	3365	1.93
THR_PRO_GLN	25	1146	1171	2.13
THR_PRO_GLY	52	4884	4936	1.05
THR_PRO_HIS	32	1104	1136	2.82
THR_PRO_ILE	60	2198	2258	2.66
THR_PRO_LEU	114	3272	3386	3.37
THR_PRO_LYS	68	1764	1832	3.71
THR_PRO_MET	8	916	924	0.87
THR_PRO_PHE	92	1984	2076	4.43
THR_PRO_PRO	186	4206	4392	4.23
THR_PRO_SER	145	2883	3028	4.79
THR_PRO_THR	56	3070	3126	1.79
THR_PRO_TRP	26	1078	1104	2.36
THR_PRO_TYR	22	1547	1569	1.40
THR_PRO_VAL	102	3414	3516	2.90
TRP_PRO_ALA	115	692	807	14.30
TRP_PRO_ARG	266	208	474	56.10
TRP_PRO_ASN	157	394	551	28.50
TRP_PRO_ASP	80	960	1040	7.69
TRP_PRO_CYS	4	73	77	5.19
TRP_PRO_GLU	45	875	920	4.89
TRP_PRO_GLN	58	278	336	17.30
TRP_PRO_GLY	119	902	1021	11.70
TRP_PRO_HIS	33	153	186	17.70
TRP_PRO_ILE	109	147	256	42.60
TRP_PRO_LEU	104	360	464	22.40
TRP_PRO_LYS	44	224	268	16.40
TRP_PRO_MET	4	22	26	15.4
TRP_PRO_PHE	101	833	934	10.80
TRP_PRO_PRO	216	241	457	47.30
TRP_PRO_SER	364	596	960	37.90
TRP_PRO_THR	84	663	747	11.20
TRP_PRO_TRP	21	177	198	10.60
TRP_PRO_TYR	54	143	197	27.40
TRP_PRO_VAL	203	557	760	26.70

Sequence	$\#cis$	$\#trans$	$\#Total$	$\%cis$	Sequence	$\#cis$	$\#trans$	$\#Total$	$\%cis$
TYR_PRO_ALA	818	1240	2058	39.70	VAL_PRO_ALA	124	3295	3419	3.63
TYR_PRO_ARG	461	945	1406	32.80	VAL_PRO_ARG	72	2086	2158	3.34
TYR_PRO_ASN	495	2008	2503	19.80	VAL_PRO_ASN	47	2287	2334	2.01
TYR_PRO_ASP	214	3147	3361	6.37	VAL_PRO_ASP	30	4242	4272	0.70
TYR_PRO_CYS	27	318	345	7.83	VAL_PRO_CYS	16	339	355	4.51
TYR_PRO_GLU	97	1989	2086	4.65	VAL_PRO_GLU	78	3368	3446	2.26
TYR_PRO_GLN	84	919	1003	8.37	VAL_PRO_GLN	25	1561	1586	1.58
TYR_PRO_GLY	323	3282	3605	8.96	VAL_PRO_GLY	304	5442	5746	5.29
TYR_PRO_HIS	85	643	728	11.70	VAL_PRO_HIS	23	982	1005	2.29
TYR_PRO_ILE	62	793	855	7.25	VAL_PRO_ILE	47	1235	1282	3.67
TYR_PRO_LEU	241	1287	1528	15.80	VAL_PRO_LEU	113	2900	3013	3.75
TYR_PRO_LYS	484	1052	1536	31.50	VAL_PRO_LYS	34	1854	1888	1.80
TYR_PRO_MET	34	294	328	10.40	VAL_PRO_MET	98	620	718	13.60
TYR_PRO_PHE	200	1132	1332	15.00	VAL_PRO_PHE	111	1595	1706	6.51
TYR_PRO_PRO	179	899	1078	16.60	VAL_PRO_PRO	60	3782	3842	1.56
TYR_PRO_SER	236	1454	1690	14.00	VAL_PRO_SER	89	2853	2942	3.03
TYR_PRO_THR	193	1439	1632	11.80	VAL_PRO_THR	101	2072	2173	4.65
TYR_PRO_TRP	29	614	643	4.51	VAL_PRO_TRP	16	387	403	3.97
TYR_PRO_TYR	191	891	1082	17.70	VAL_PRO_TYR	119	1734	1853	6.42
TYR_PRO_VAL	68	971	1039	6.54	VAL_PRO_VAL	63	2337	2400	2.62

Chapter 4

Prediction of secondary structure propensities in IDPs

Contents

4.1 Introduction

For over more than 40 years, numerous methods have been developed to predict secondary structure in proteins from their sequence [104, 168]. However, current secondary structure predictors are in general trained and evaluated on folded/globular proteins, and thus, are not necessarily appropriate to identify partially-structured regions in IDPs. Furthermore, numerous methods have also been proposed to predict structural disorder from protein sequence (see [134] and references therein). Available disorder predictors mostly focus on the identifications of disordered regions in predominantly folded proteins. In general, they only provide a binary output (i.e. ordered/disordered) or a disorder probability for each residue, but do not identify structural classes. Traditionally, secondary structure and disorder predictors have been developed independently from each other, since they aim at providing different information. One exception is the s2D method [211], which predicts secondary structure populations and disorder in a unified framework. The s2D

method, as well as the work presented here, relies on a more holistic view of structural biology of IDPs by exploring descriptors that span the continuum between the two extremes: ordered, disordered [212, 45, 39].

In contrast to the most advanced approaches, which are based on intricate machine-learning techniques, here we present an extremely simple strategy to identify secondary structural propensities from protein sequences. As machine-learning-based approaches, our method exploits structural information contained in databases. However, our approach performs simple statistical operations on the conformational preferences of three-residue fragments extracted from coil regions of experimentally-determined high-resolution protein structures (explained in Chapter 3). The main advantage of our strategy with respect to most of the machine-learning-based methods, specially those using neural networks, is that it enables a comprehensible connection between sequence and structural preferences/propensities. We have called our method Local Structural Propensity Predictor (LS2P). The code of the predictor (in Python) will be freely provided.

4.2 Material and Methods

4.2.1 Structural classification of three-residue fragments

The prediction method proposed in this work, LS2P, exploits the statistical information on the structural preferences of three-residue fragments using the tripeptide database presented in Chapter 3.

To simplify the structural classification, the conformational space of each residue r_i was subdivided according to the values of the Ramachandran angles, ϕ and ψ, into three regions $S = \{\alpha, \beta, \gamma\}$, as explained in Chapter 3, and illustrated in Figure 3.3

The combinations of these three structural classes at the single residue level lead to 27 structural classes $\mathscr{S}$ for a tripeptide: $\alpha\alpha\alpha, \alpha\alpha\beta, \alpha\alpha\gamma, \alpha\beta\alpha, \ldots, \gamma\gamma\gamma$. The number of conformations per class was retrieved and stored for each of the 8,000 tripeptide types. These numbers are used by the LS2P predictor as explained below.

4.2.2 Statistical analysis of local structural propensities

LS2P predicts secondary structure propensities for a given protein sequence. For each residue r_i in the sequence, the secondary structure propensity is calculated using statistical information for the tripeptide t_i centered at this residue and for its neighbors: t_{i-2}, t_{i-1}, t_{i+1} and t_{i-2}.

Let n^i denote the total number of structures present in the tripeptide database for t_i. The number of structures for each one of the 27 structural classes is indicated using the corresponding Greek letters in subscript. For instance, $n^i_{\beta\gamma\alpha}$ is the number of structures of t_i with the first residue of the tripeptide in the β region, the second in γ and the third in α. We use lower-case Latin letters, for instance x or y, as variables when the three structural classes have to be considered for one or several residues. This notation is used below within summation equations.

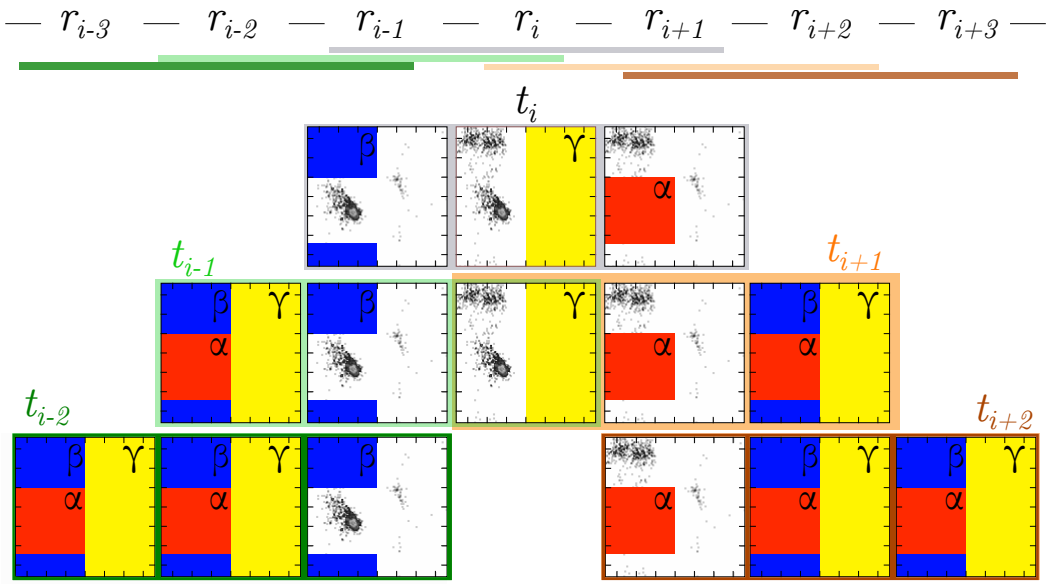


Figure 4.1. Ramachandran plots of residues in t_i, and in the neighboring tripeptides t_{i-2}, t_{i-1}, t_{i+1} and t_{i+2}. Colored regions correspond to the case where t_i is in the structural class $\mathscr{S} = \beta\gamma\alpha$. Notice that overlapping residues in consecutive tripeptides must be in the same structural class.

For a tripeptide t_i, independently of the rest of the sequence, the number of structures present in each of the 27 structural classes with respect to the total number of structures already gives us an idea of its conformational preferences. For example, for the particular case $\mathscr{S} = \beta\gamma\alpha$, and considering t_i independently of the rest of the sequence:

$$p(\beta\gamma\alpha)_i = \frac{n^i_{\beta\gamma\alpha}}{n_i} \ , \ n_i = \sum_{w,x,y \in S} n^i_{wxy} \tag{4.1}$$

However, in order to better take into account the sequence context, the compatibility of the structural preferences of t_i with those of the neighboring tripeptides has to be considered. This is illustrated in Figure 4.1. In this particular case, the probability of t_i to adopt a conformation of type $\beta\gamma\alpha$ must consider the probability of the two last residues of t_{i-1} to adopt conformations $\beta\gamma$, of the two first residues of t_{i+1} to adopt conformations $\gamma\alpha$, of the last residue of t_{i-2} to adopt conformations β, and of the first residues of t_{i+2} to adopt conformations α. The structural preferences conditioned by the neighbors can be easily computed operating with the numbers of structures in the tripeptide database. For the example of $\mathscr{S} = \beta\gamma\alpha$, the equation can be written as:

$$p(\beta\gamma\alpha)_i = \frac{\sum_{t,u,y,z \in S} n^{i-2}_{tu\beta}\ n^{i-1}_{u\beta\gamma}\ n^{i}_{\beta\gamma\alpha}\ n^{i+1}_{\gamma\alpha y}\ n^{i+2}_{\alpha yz}}{\sum_{t,u,v,w,x,y,z \in S} \left(n^{i-2}_{tuv}\ n^{i-1}_{uvw}\ n^{i}_{vwx}\ n^{i+1}_{wxy}\ n^{i+2}_{xyz} \right)} \tag{4.2}$$

To compute the propensity of tripeptide t_i to adopt a particular structural class e.g. $\mathscr{S} = \beta\gamma\alpha$ with respect to the observations in our database, we have to divide $p(\beta\gamma\alpha)_i$ by the overall probability to observe this structural class for all tripeptides:

$$p(\beta\gamma\alpha)_{\text{all}} = \frac{n^{\text{all}}_{\beta\gamma\alpha}}{N} \tag{4.3}$$

where "all" implies the sum for the 8,000 tripeptide sequences, and N is the total number of tripeptide structures in the database. Thus, the structural propensity can be written as:

$$P(\beta\gamma\alpha)_i = \frac{p(\beta\gamma\alpha)_i}{p(\beta\gamma\alpha)_{\text{all}}} \tag{4.4}$$

Note that the values $P(\mathscr{S})_i$ do not correspond to an estimation of the populations of structural classes for the protein in solution. They are an indicator of the structural propensity of different regions within the IDP. Values larger than 1 for a given structural class indicate that this class in favored for the tripeptide in the local sequence context. On the contrary, values below 1 indicate unlikelihood for that class.

4.3 Results

4.3.1 Identification of secondary structure propensities in IDPs: Overall picture

A benchmark set of nine structurally-characterized IDPs was used to evaluate the performance of our approach. Table 4.1 provides the list of these proteins together with the references to structural studies of these systems based on NMR experiments. Note that the same benchmark will be used to evaluate the ensemble modeling method presented in Chapter 5. Concretely, MAPK Kinase 7 (MKK7) [117], the fragment 955-1097 of the Erythrocyte binding antigen 181 (EBA-181) [22], p15 [40], sic1 [148], Measles virus ntail (ntailMV) [97], Sendai virus ntail (ntailSV) [98], the unique domain of the src kinase (USrc) [170], K18 construct of Tau protein

Table 4.1: The nine IDPs in the benchmark set together with the reference to the articles from which the experimental RCDs used in this thesis where obtained.

Protein	RDCs data
MKK7	[117]
EBA-181	[22]
p15	[40]
sic1	[148]
ntailMV	[97]
ntailSV	[98]
USrc	[170]
K18	[154]
Tau	[194]

(K18) [154], and full-length Tau protein [194] were used in our study. Predictions of secondary structure propensities predicted by LS2P were compared to the NMR RDCs, which are extremely sensitive to small conformational bias at residue level [99]. Other structural analyses of these 9 proteins from related literature were also considered for this evaluation.

First, we analysed the number of conformations in our database for all the tripeptides of the benchmark set. Results are summarized in Figure 4.2. The average number of conformations per tripeptide ranges from 637 to 792 for sic1 and Tau, respectively. The minimum number of conformations found is 22, which corresponds to the tripeptide Cys-Met-His in EBA-181. These observations indicate that we have an excellent sampling for the vast majority of the tripeptides and that

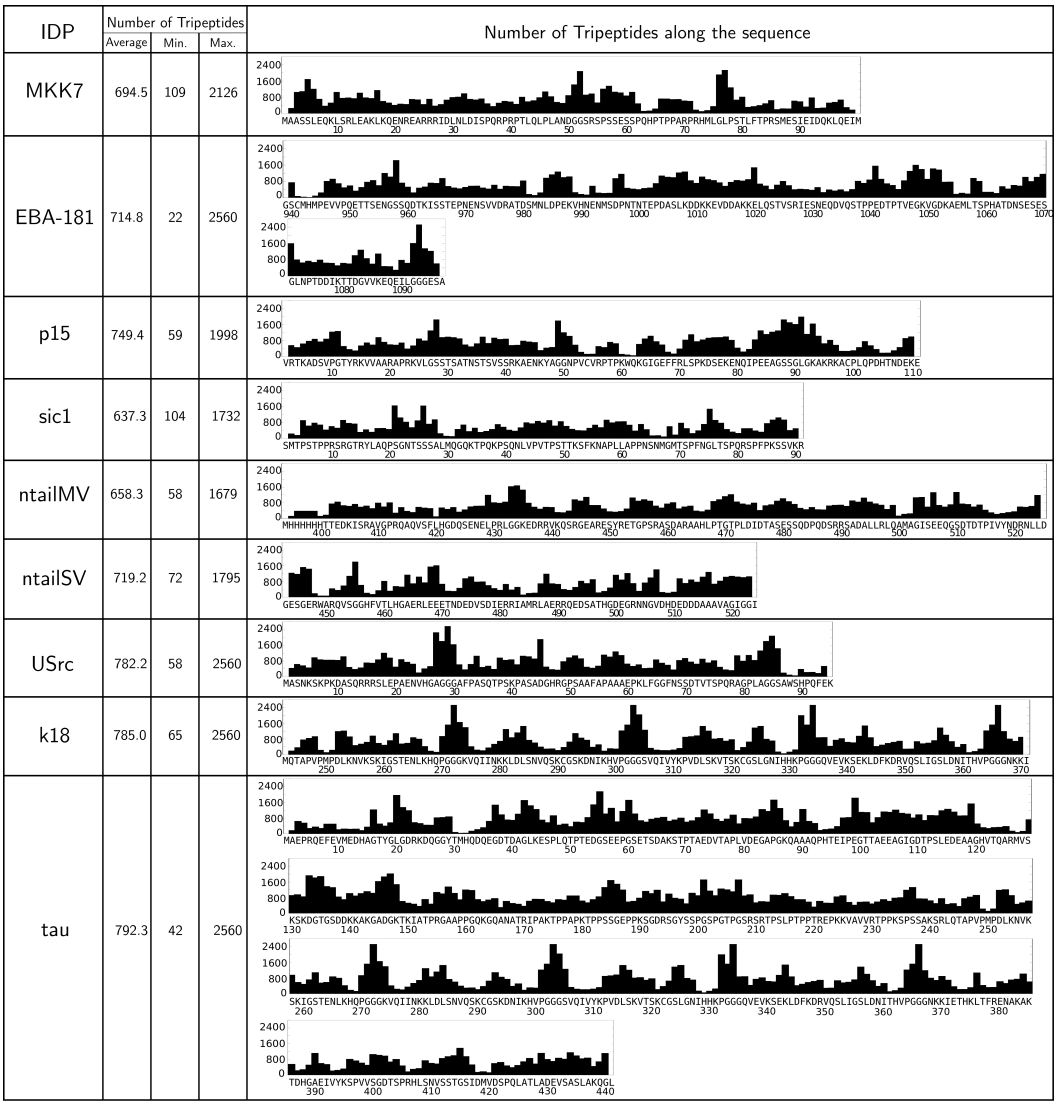


Figure 4.2. Number of structures in the tripeptide database for the tripeptides on the 9 IDPs benchmark. The value for each tripeptide is shown on its central residue.

reliable statistics can be derived from the analysis.

In order to illustrate the application of LS2P, results for the nine benchmark IDPs are presented at the end of this chapter. MKK7 (see Figure 4.4) and EBA-181 (see Figure 4.5) are representative examples to explain the results provided by LS2P, and therefore they are commented in more detail here, while the rest of the IDPs are presented in the following sections. From a structural point of view, MKK7 and EBA-181 present very different features. While MKK7 involves relatively long regions with helical or extended propensity, EBA-181 is almost fully disordered, only presenting short partially-structured fragments.

MKK7 fragment analysed involves three MAPK binding domains that have been structurally characterized by NMR: D1 (residues 25-34), D2 (residues 38-47) and D3 (residues 70-79) [117]. These three domains have been shown to present different conformational propensities. The N-terminus of MKK7 (residues 5-30) presents an α-helical structure that is characterized by the positive values of the RDC profile. This helical propensity is well predicted by the LS2P method. From residue 26, LS2P predicts the following region to be highly extended. Interestingly, this regions display very negative RDC values, which is a signature of extended conformations. Moreover, the extendedness of D2 was captured by the ensemble refinement performed in the original study [117]. The rest of the sequence appears, according to LS2P, as preferentially extended, although some α-helical propensity is observed at the C-terminus. Moreover, some MKK7 stretches around residues 54, 65 and 80 are dominated by less abundant structures involving γ-type conformations. Our secondary structure prediction is in very good agreement with the structural conformation found for the three MAPK binding motifs. D1 lies in the transition between helical and extended conformations at the N-terminus of MKK7, and this dual behavior was captured by the ensemble refinement done in the original study. D2, which is inserted in the long extended region of MKK7 according to LS2P, was experimentally shown to sample β-strand and polyproline-II (PPII) conformations. Conversely, predictions of the D3 indicate that this region has no special enrichment in helical or extended conformations, with the exception of residues 73 and 74, in line with the original experimentally-derived ensemble model.

Structural investigations of EBA-181 have shown that the fragment involving residues 945-1097, which is part of the RIII–V region, behaves essentially as a random coil with the presence of several turn motifs or short single-turn α-helices [22]. These short helical elements, corresponding to positive RDCs around residues 987–988, 998, 1006–1007 and 1016–1019, are correctly identified by LS2P. Note that the identification of turns will be described in more detail below. Other short regions present some propensity to adopt extended conformations, in particular regions around prolines P945, P949, P1003, P1039, P1040 and P1044. These short extended regions are also well predicted by LS2P. When analysing the enrichment of the 27 structural groups, we observe that most of them are present along the sequence and only regions around 958, 1050 and the C-terminus seem to be highly enriched in less common structural classes. Sequences allowing more heterogeneous conformations seem to be an indicator of disorder and absence of

stable secondary structural elements.

These results for MKK7 and EBA-181, which showcase different types of IDPs from a structural point of view, are an indicator of the good performance of LS2P. The following sections will describe more specifically the ability of LS2P to identify different types of secondary structural elements within IDPs.

4.3.2 Identification of helical elements within IDPs

LS2P is able to detect helical elements of different lengths, even though the method operates from structural preferences of three-residue fragments. This highlights the importance of the local sequence context, which implicitly encodes the cooperative formation of structural elements along the polypeptide chain. In addition of the previously described examples, our benchmark contains other examples of IDPs involving relatively long fragments with helical propensity. The two most prominent ones are ntailMV (see Figure 4.8) and ntailSV (see Figure 4.9). These two proteins have similar sequences and perform the same function by interacting with the phosphoprotein in two related viruses through a highly stable α-helix.

LS2P identifies several regions displaying an enrichment in helical conformations in both proteins, including the experimentally characterized functional α-helix. When comparing with s2D predictions (note that the comparison with s2D is further commented in a separate section below), we observe different levels of agreement. In ntailSV, both algorithms identify four α-helices and, interestingly, two of these regions (around residues 450 and 515) display positive RDCs, suggesting the presence of helical populations. Conversely, only the functional helix is identified by s2D for ntailMV. The most surprising result is that our approach predicts that the two functional helices, specially ntailSV, contain a non-negligible proportion of extended conformations in the middle of the α-helix. This observation is in contrast with the experimental data [98, 97] and the predictions done with s2D. These contradictory observation underlines a fundamental difference between both methods. While s2D was trained using experimental data and therefore captures propensities in longer protein stretches, LS2P is sensitive to local conformational bias. In that sense, LS2P could encounter more problems to correctly identify large helices, which have a strong cooperative nature. However, LS2P can probe short structures, such as turns and N-caps, that would remain invisible for s2D. It has been shown that the functional helices of both ntail proteins are highly stabilized by N-capping serine and aspartic acid residues placed upstream of the helix [98, 97]. The inspection of the conformational propensities in these regions identifies several residues with a strong propensity for $\beta\alpha\alpha$ structures. Concretely, tripeptides centered in 473, 474 and 479 in ntailSV, and 485, 488 and 491 in ntailMV display a strong enrichment in $\beta\alpha\alpha$ conformations. We speculate that this conformation stabilizes downstream helices in solution, but our predictor may loose the structural cooperativity due to its local nature.

4.3.3 Identification of extended regions in IDPs

Several regions are identified as extended ($\beta\beta\beta$) in the analysis of the benchmark set. Note that the current implementation of LS2P does not make the difference between β-strand-type and PPII-type conformations, both of them being classified as "extended". Note also that the possible presence of hydrogen bonds to stabilize β-strands is not considered as this is an uncommon situation in IDPs. Instead, extended regions are identified only on the basis of the local amino acid sequence.

Protein Tau (see Figure 4.12), and in particular the K18 (see Figure 4.11) construct, is an excellent example to illustrate the ability of LS2P to predict the propensity of some regions within IDPs to adopt extended conformations. The method identifies extended regions described in related literature [154, 194]: at the N-terminal region of Tau (around residue 50, in particular), within the proline-rich region (residues 212-232), and inside the pseudorepeat domains contained in K18 (residues 275-282, 307-313 and 338-346, approximately). All the regions correspond to negative RDCs in Figures 4.12 (Tau) and 4.11 (K18). Note that for the last region, residues 338-346, LS2D predicts a combination of extended and helical propensities for the second half of this region, followed by an increase of helical propensity, showing a peak around residues 348. This prediction agrees with the RDC profile in this region.

LS2P also identifies extended regions in other proteins such as sic1 [148, 147] (see Figure 4.7) and p15 [40] (see Figure 4.6, showing a good agreement with RDC profiles and structural descriptions from related literature.

There are two main factors, relaying only on the local sequence, that induce extended conformations. One of them is the presence of prolines, which enriches neighbouring residues with extended conformations, as is the case for the the proline-rich region in Tau or the short extended regions in EBA-181. Amino acid bulkiness is another property that has a strong effect on the conformational preferences of

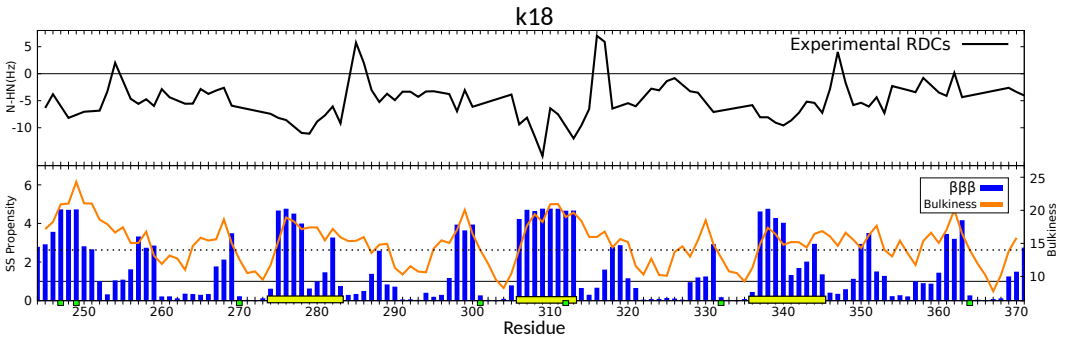


Figure 4.3. RDC profile (top, black line), $\beta\beta\beta$ propensity (bottom, blue bars) and bulkiness profiles (bottom, orange line) for the K18 construct of Tau protein. The prolines are indicated with a green square. The dashed line situated at bulkiness = 14 indicates the threshold adobe which the amino acid are considered as bulky [29]. The three experimentally-characterized extended regions involving residues 275-282, 307-313 and 338-346 are highlighted in yellow.

neighboring residues [248]. Amino acids with large side chains enrich extended conformations in neighboring residues as a conformational mechanism to avoid steric clashes. To illustrate the importance of amino acid bulkiness in the identification of extended conformations, we have computed the bulkiness profile for all proteins in the benchmark [29]. Note that the presence of prolines is accounted for this calculation increasing by 60% its theoretical volume.

Figure 4.3 shows experimental RDCs, the predicted extended propensity and the bulkiness profile for the the K18 construct of Tau. We observe a correlation between the regions having highly negative RDCs, displaying a enhanced population of $\beta\beta\beta$ propensity, and large bulkiness. This correlation suggests that LS2P properly identifies regions with extended conformations, and that our statistical approach captures the steric influence encoded in the sequence.

4.3.4 Identification of turns in IDPs

Various turn types have been defined in folded/globular proteins. Turns have also been identified in some IDPs. Usually, these turns are only partially formed, complicating their identification. Experimentally, turns can be identified based on RDCs, which display anomalous values with respect to the neighbouring residues.

Here, we focus type I and type VIII β-turns, which are the most common form of turns. Disregarding the actual presence of a hydrogen bond that stabilizes such structural elements, they can be defined in a simple and general way as follows: type I and type VIII β-turns are characterized by two consecutive amino acid residues with conformations in the α region of Ramachandran space, preceded and succeeded by residues with more extended conformations [154]. The possibility to distinguish other turn types from the 27 structural classes of tripeptides used by LS2P remains to be further investigated.

Following the definition above, turns could be predicted from the results of LS2P by identifying consecutive (overlapping) tripeptides with high propensities in structural classes $\beta\alpha\alpha$ and $\alpha\alpha\beta$. As shown in figure 4.11, the results provided by LS2P fit the well-characterized turns in the K18 construct of Tau [154], involving residues 252-255, 283-286, 314-317, and 345-348. In addition to the aforementioned signature $\beta\alpha\alpha$-$\alpha\alpha\beta$ for the two middle residues, these structural elements are often characterized by a peak of $\alpha\alpha\alpha$ propensity for the second of these middle residues, which can be higher than the $\alpha\alpha\beta$ propensity. In such cases, the $\alpha\alpha\beta$ propensity increases for the next residues in the sequence.

Coming back to the example of EBA-181 (see figure 4.5, the representation of $\beta\alpha\alpha$ and $\alpha\alpha\beta$ propensities, in addition to $\alpha\alpha\alpha$, allows to highlight differences between turn motifs around residues 987 and 998, and small helices involving residues 1006-1007 and 1016-1019. Note that the $\beta\alpha\alpha$-$\alpha\alpha\beta$ propensity for pairs of consecutive residues is also high in other regions of EBA-181 showing positive peaks in the RDC profile, such as residues 971-972 and 1031-1032.

4.3.5 Comparison with state-of-the-art methods for structural propensity prediction

As mentioned in the introduction, the vast majority of secondary structure predictors are aimed to predict structural elements within proteins that are mostly globular, and produce a one-letter code (corresponding only to α-helices and β-strands in most of the cases) that simplifies the analysis. These methods usually fail to identify partially-structured regions in IDPs, especially when the structural propensity is relatively low. On the other hand, disorder predictors, which are aimed to identify regions lacking secondary structure, do not provide information about structural propensities at the frontier between order and disorder. Tests using our benchmark set of 9 IDPs (results of these tests are not presented here since they lack of interest), as well as results presented in the literature [211, 212], show the limitations of these "traditional" predictors when applied to IDP sequences. A remarkable exception is the s2D method [211], which was especially conceived to simultaneously predict secondary structure and disorder propensities, and which is particularly well suited to the structural study if IDPs. Here, we compare the performance of s2D and LS2P to predict secondary structure propensities for the 9 proteins considered in this work.

LS2P and s2D agree in many cases, particularly when the structural elements are know to have relatively high propensity to be formed in solution. This is the case for instance for the helical region at the N-terminal side of MKK7 and for the helical region in ntailSV and ntailMV. Both methods also agree on the prediction of the extended regions in K18. As mentioned before, the underlying principle of s2D may make this method more suitable that LS2P for identifying relatively large and highly-populated structural elements.

However, s2D generally fails to identify transient secondary structure in several cases for which LS2P clearly provides this information. This is for instance the case for sic1, which has been shown to concatenate regions with significant propensity to adopt extended or helical conformations [147]. s2D also fails to identify small structural motifs such as turns or short helices, whereas LS2P is able to find them, as it has been illustrated for EBA-181 and K18(Tau). Another interesting case is p15 (see Figure 4.6, an IDP that has been shown experimentally to have two extended regions, one at the N-terminal involving residues 15-24 and another at the C-terminal residues involving 94-104 [40]. LS2P identify the two regions while s2D does not predict extended propensity for these residues. The rest of structural preferences found for this protein are not strong when looking at the RDCs profile, and they are not detected neither by LS2D nor by s2D.

Surprisingly, both methods s2D and LS2P fail to identify a few regions for which previous work suggest some structural propensity. This is the case for the regions involving residues 60-75 in USrc, for which helical propensity has been suggested, as also indicate the positive RDC values in this region (see Figure 4.10). Another example is the small helix involving residues 72-75 in sic1. The fact that both methods fail to identify transient structural elements in the same regions, despite

their very different underlying principles, shows that this is a challenging problem probably, mainly due to the very particular sequences that can be found in IDPs. Indeed, the structural prediction of IDPs is still an open problem that required further research efforts, as discussed in the following section.

4.3.6 Exhaustive structural prediction of poly-Q flanking regions

In addition to the study of the benchmark set of nine proteins, the predictor presented in this chapter was also used for the investigation of poly-Q regions in proteins. More precisely, we explored the secondary structure propensity in the N-flanking region of long poly-Q tracts in human proteins. For that, four hundred fragments with ten or more glutamines and containing a maximum of two non-glutamine residues were collected from 309 different human proteins, and the ten preceding (-10 to -1) residues were structurally analysed using our secondary structure predictor. Our analysis shows a general tendency to adopt α-helical conformations in this flanking region. Interestingly, this tendency presents its largest percentage when close to the poly-Q homo-repeat (residues -1 and -2), and systematically decreases when more residues of the N-flanking region are incorporated in the analysis. This analysis, which reinforces with a structural perspective the compositional analysis done on poly-Q flanking regions, is further described in the article included in the annexes entitled: “Flanking regions define the conformation of the poly-glutamine homo-repeat in huntingtin through opposite structural mechanisms".

4.4 Conclusion

In this chapter, we have investigated the ability to predict secondary structure propensities within IDPs using local sequence-dependent information encoded in small protein fragments extracted from coil regions in experimentally-determined high-resolution protein structures. We have developed an extremely simple statistical approach based on a coarse classification of tripeptide structures. In contrast with nowadays popular neural-network-based secondary structure predictors, this approach enables a comprehensive connection between sequence and structural propensities. Moreover, thanks to this simplicity, the proposed predictor LS2P is very computationally inexpensive. This should allow the fast scanning of large databases or complete proteomes.

Results presented in the previous section show that our method is able to predict the main secondary structural elements: α-helices and extended conformations. These are detected regardless of the length of the secondary structural element. This is a clear advantage with respect to state-of-the-art secondary structure predictors, including those suited to IDPs such as s2D, which mainly identify relatively long and highly populated secondary structure elements. In addition to the detection of the most common conformations, the statistical analyses of our tripeptide database provide further information. We detect N-capping structures from the amino acid

sequence, as exemplified by the highly stable helices of N-tail proteins. Another unique feature of our approach in the detection of certain types of turns. Provided by the correct assignment of turn-type using experimental methods, we can connect them with the sequence of structural classes predicted by L2SP, enabling an easy scan through other protein sequences.

Despite the good overall performance of the method, it should be noted that LS2P may predict structural propensity is some regions for which there is no experimental evidence of secondary structure. This is the case for instance of the C-terminal region (residues 95-99) of MMK7 (see Figure 4.4), as mentioned above. The method can also fail to predict helical propensities in a few cases (i.e. it may produce false negatives). An example is the low populated helical structure involving residues 60-75 in USrc (see Figure 4.10), which has been characterized by NMR experiments [170]. Such under-performance in some regions can be due to inaccuracies or lack of information in the tripeptide database.

Indeed, as machine-learning-based methods strongly rely on the data-set used for training, the results provided by LS2D are dependent on the tripeptide database. Our current database was constructed from coil regions in a large set of structured proteins mostly determined by X-ray crystallography. The available information in this database can be inaccurate or limited for sequences that are seldom observed in globular proteins but that may appear in IDPs. With the enlargement of repositories of high-resolution structures and data obtained from NMR experiments, we expect to enrich our database and achieve more robust predictions. A more extensive and accurate structural database would also enable us to further refine the structural classes with respect to the three classes per residue α, β, γ considered in this work. In particular, it would be interesting to distinguish β-strand-type and PPTT-type conformations.

Finally we must mention that, due to the simplicity of the approach, LS2P is unable to accurately predict populations of structural elements (i.e. expected percentage of the different structural classes for the protein in solution). This would be an interesting extension for future work. A possible approach in this direction would be to couple LS2P with the method to construct conformational ensemble models of IDPs presented in Chapter 5, and which requires information to distinguish fully-disordered and partially-structured regions.

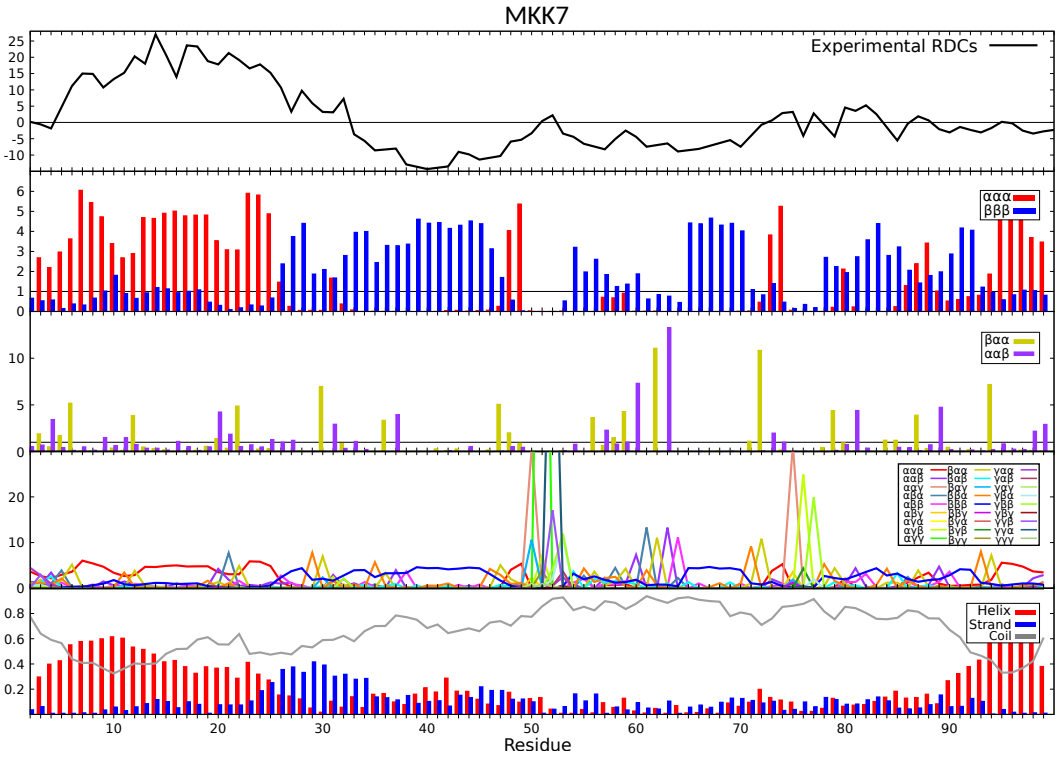


Figure 4.4. 5 panels with results for MKK7. From top to bottom: Experimental RDCs; $\alpha\alpha\alpha$ and $\beta\beta\beta$ propensity useful to identify helical and extended regions; $\beta\alpha\alpha$ and $\alpha\alpha\beta$ propensity useful to identify turns; The 27 structural classes propensities; s2D result.

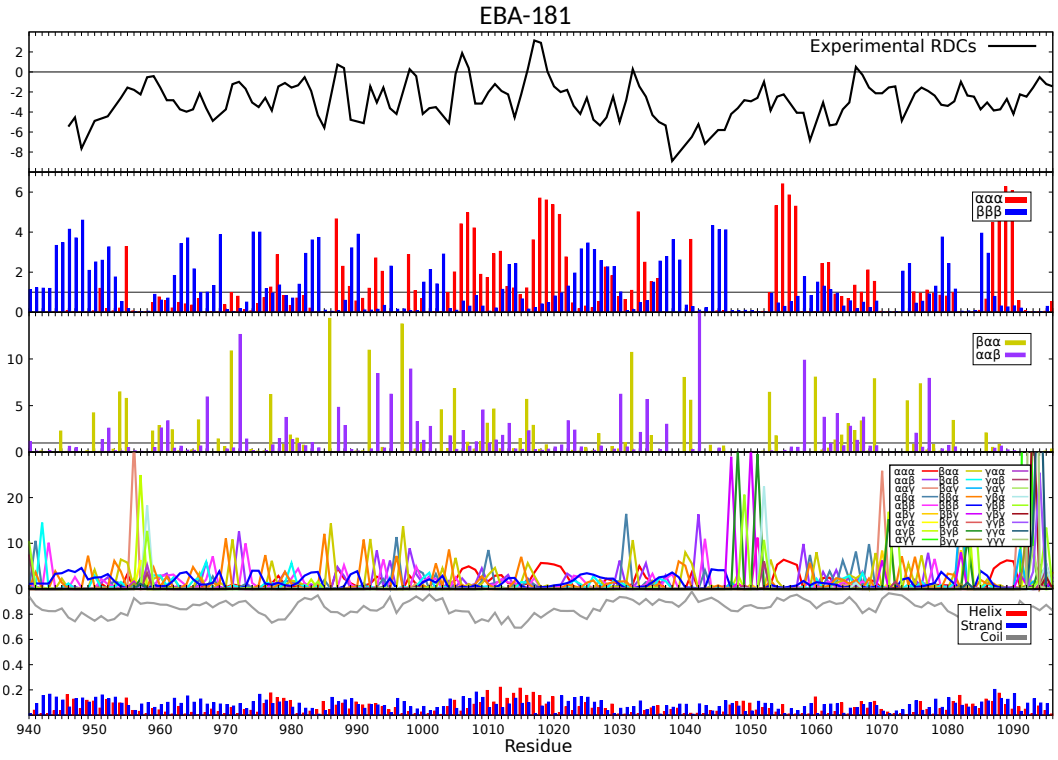


Figure 4.5. 5 panels with results for eba181. From top to bottom: Experimental RDCs; $\alpha\alpha\alpha$ and $\beta\beta\beta$ propensity useful to identify helical and extended regions; $\beta\alpha\alpha$ and $\alpha\alpha\beta$ propensity useful to identify turns; The 27 structural classes propensities; s2D result.

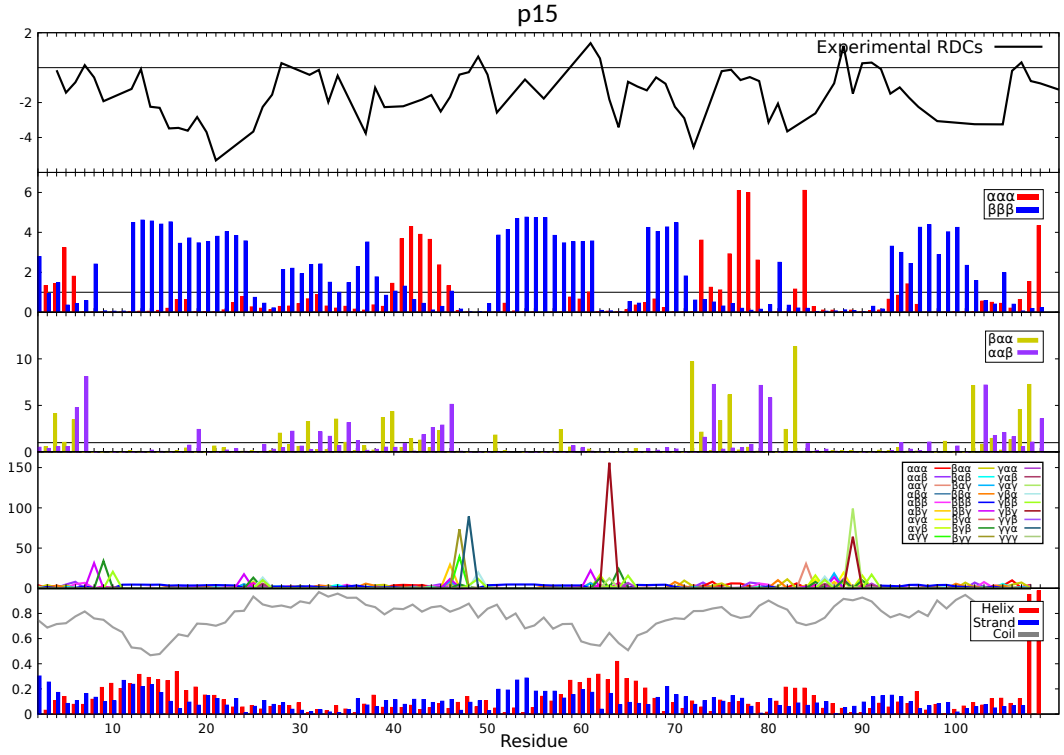


Figure 4.6. 5 panels with results for p15. From top to bottom: Experimental RDCs; $\alpha\alpha\alpha$ and $\beta\beta\beta$ propensity useful to identify helical and extended regions; $\beta\alpha\alpha$ and $\alpha\alpha\beta$ propensity useful to identify turns; The 27 structural classes propensities; s2D result.

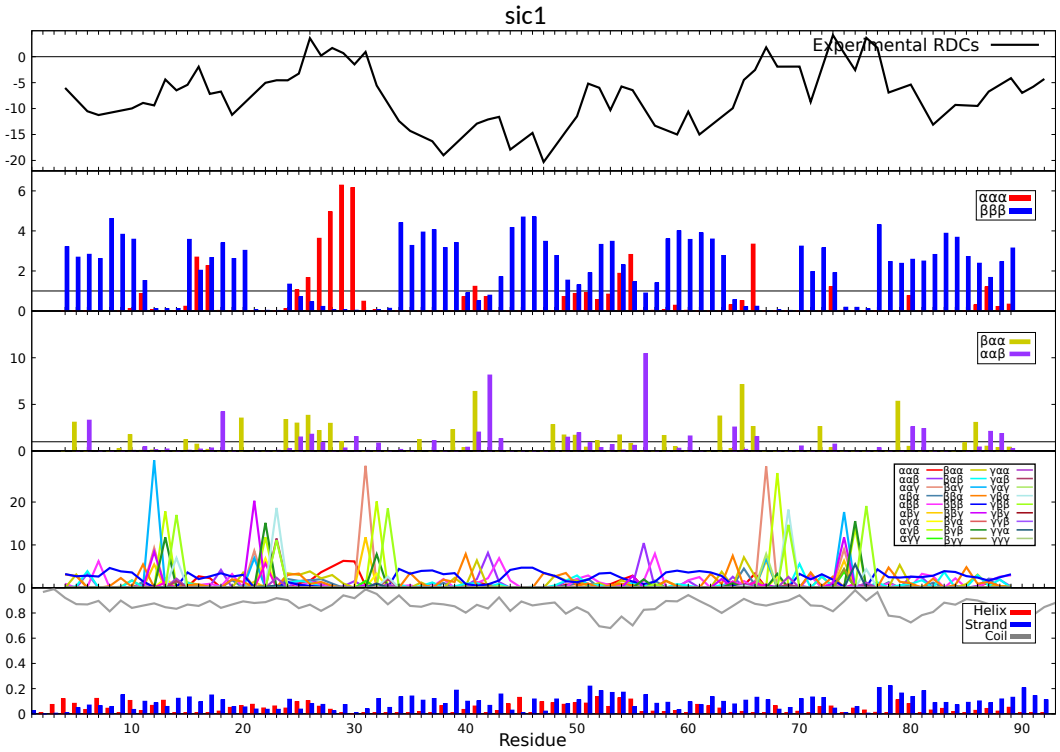


Figure 4.7. 5 panels with results for sic1. From top to bottom: Experimental RDCs; $\alpha\alpha\alpha$ and $\beta\beta\beta$ propensity useful to identify helical and extended regions; $\beta\alpha\alpha$ and $\alpha\alpha\beta$ propensity useful to identify turns; The 27 structural classes propensities; s2D result.

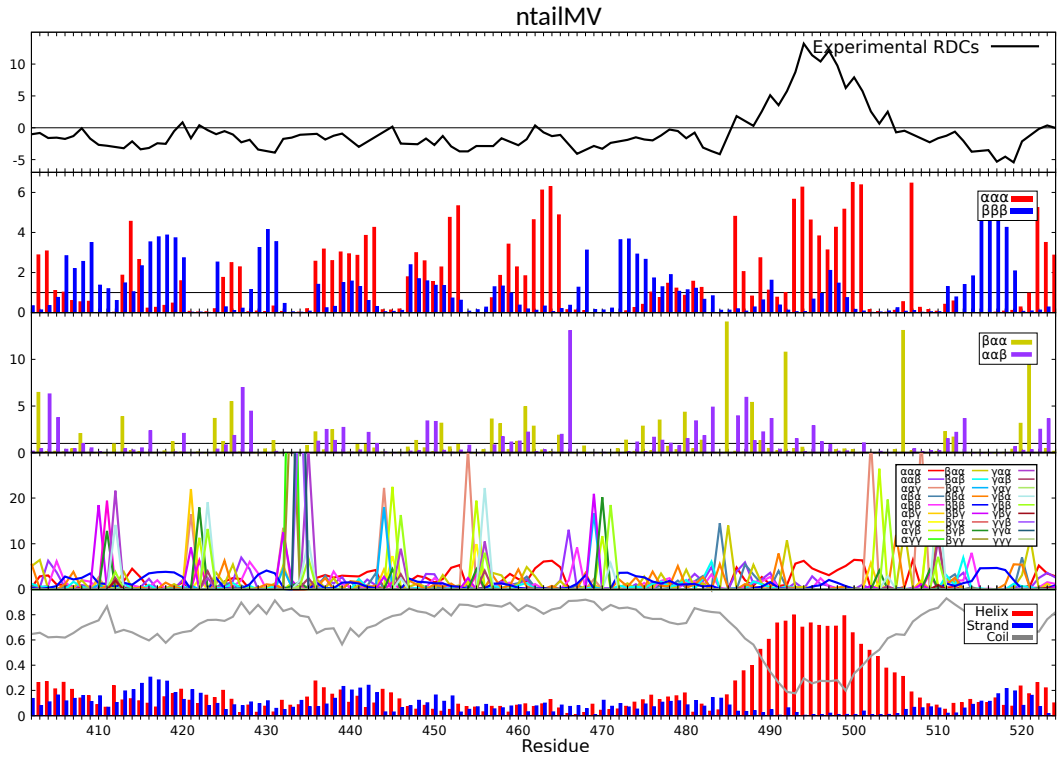


Figure 4.8. 5 panels with results for ntailMV. From top to bottom: Experimental RDCs; $\alpha\alpha\alpha$ and $\beta\beta\beta$ propensity useful to identify helical and extended regions; $\beta\alpha\alpha$ and $\alpha\alpha\beta$ propensity useful to identify turns; The 27 structural classes propensities; s2D result.

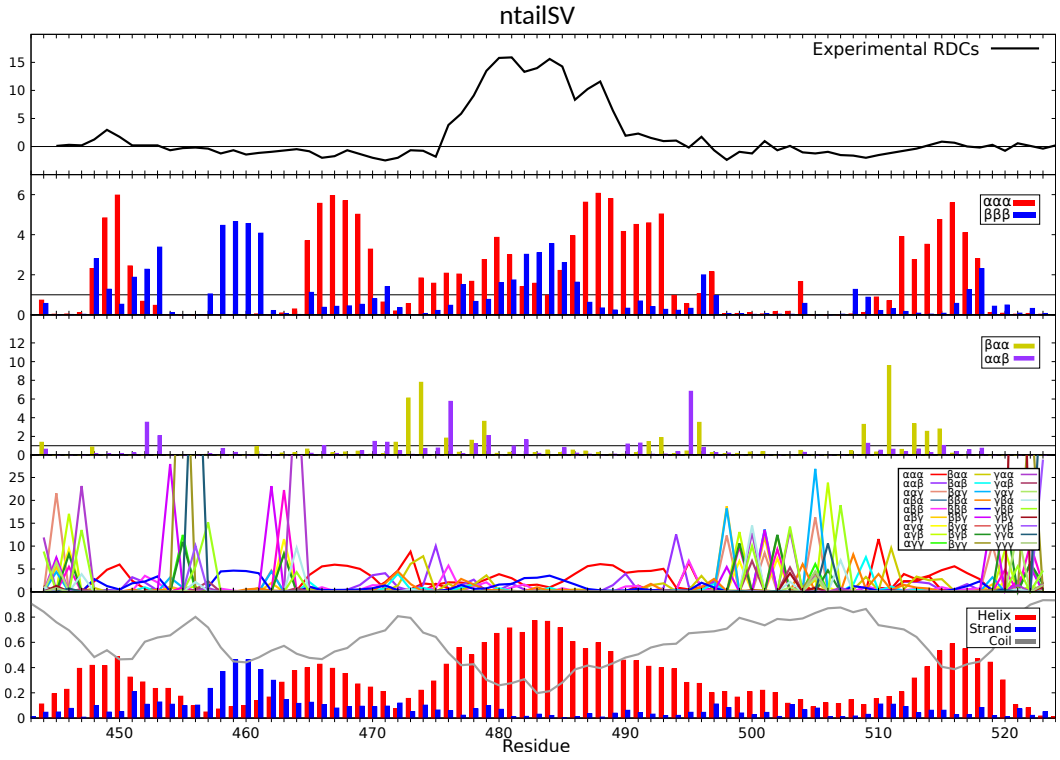


Figure 4.9. 5 panels with results for ntailSV. From top to bottom: Experimental RDCs; $\alpha\alpha\alpha$ and $\beta\beta\beta$ propensity useful to identify helical and extended regions; $\beta\alpha\alpha$ and $\alpha\alpha\beta$ propensity useful to identify turns; The 27 structural classes propensities; s2D result.

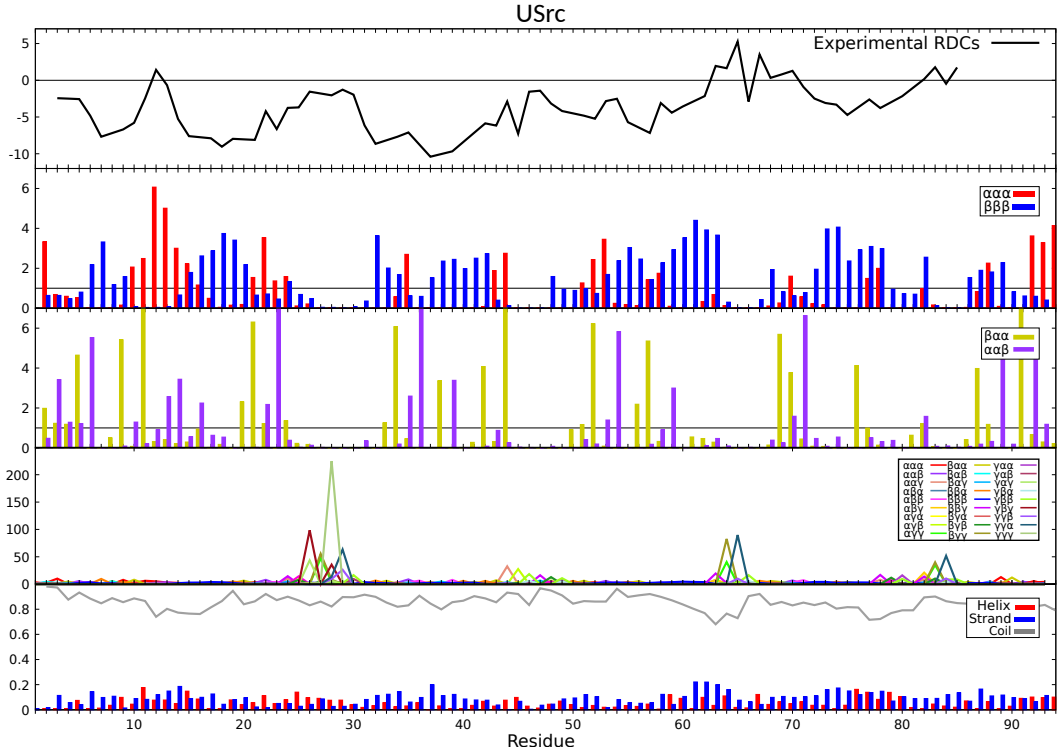


Figure 4.10. 5 panels with results for USrc. From top to bottom: Experimental RDCs; $\alpha\alpha\alpha$ and $\beta\beta\beta$ propensity useful to identify helical and extended regions; $\beta\alpha\alpha$ and $\alpha\alpha\beta$ propensity useful to identify turns; The 27 structural classes propensities; s2D result.

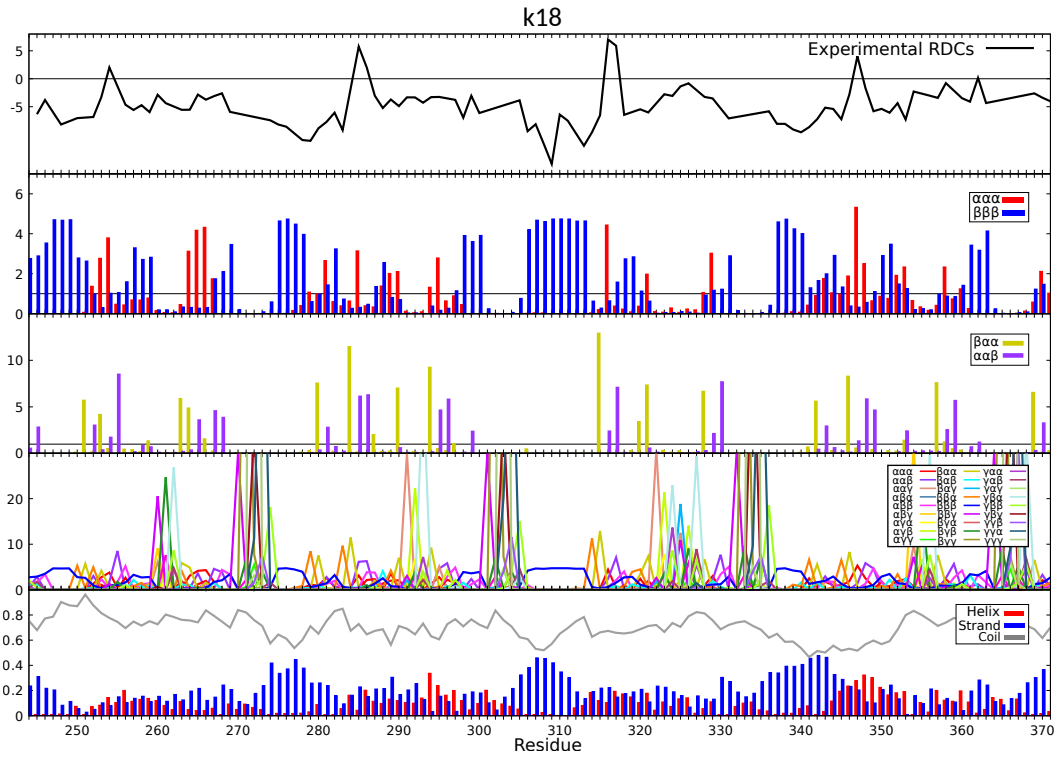


Figure 4.11. 5 panels with results for K18. From top to bottom: Experimental RDCs; $\alpha\alpha\alpha$ and $\beta\beta\beta$ propensity useful to identify helical and extended regions; $\beta\alpha\alpha$ and $\alpha\alpha\beta$ propensity useful to identify turns; The 27 structural classes propensities; s2D result.

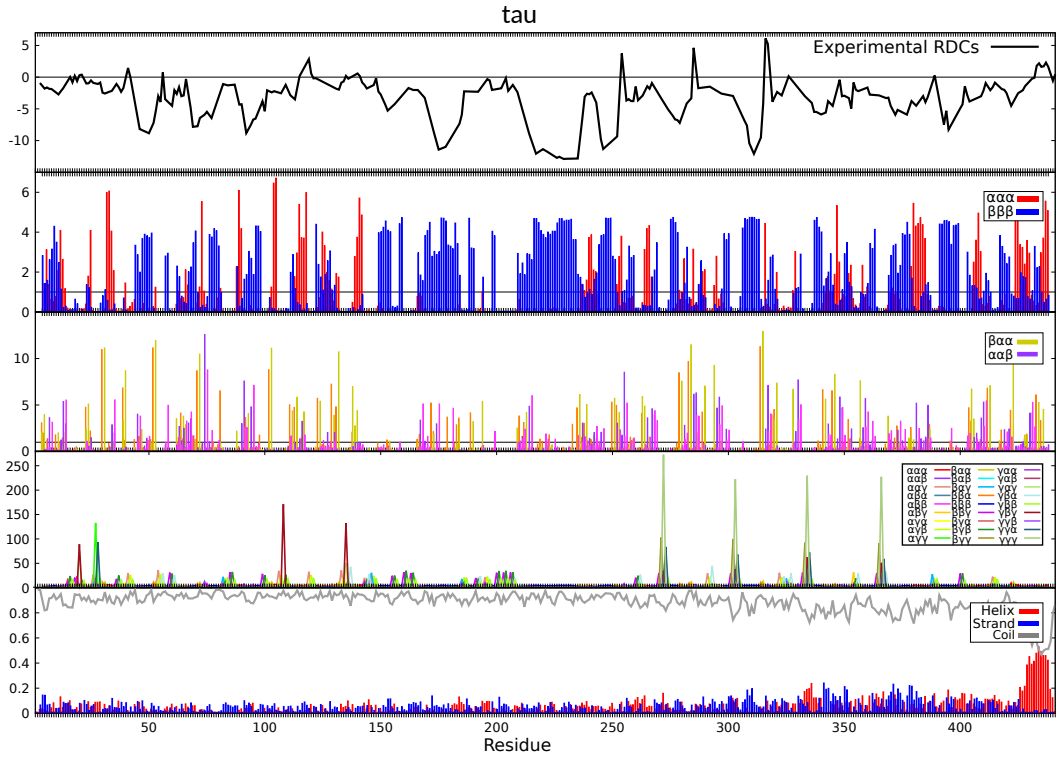


Figure 4.12. 5 panels with results for Tau. From top to bottom: Experimental RDCs; $\alpha\alpha\alpha$ and $\beta\beta\beta$ propensity useful to identify helical and extended regions; $\beta\alpha\alpha$ and $\alpha\alpha\beta$ propensity useful to identify turns; The 27 structural classes propensities; s2D result.

Chapter 5

An algorithm to build realistic ensemble models of IDPs

Contents

5.1 Introduction

Multiple computational tools using distinct levels of description have been developed to characterize IDPs when no or limited experimental information is available. As explained in Chapter 4, current disorder prediction tools, which are based on the statistical analysis of protein sequences, provide rough estimations of partly structured regions in IDPs [47], although the exact secondary structure classes are poorly defined. In principle, a more accurate characterization can be provided by MD-based methods. However, despite significant advances in the extension of MD methods to IDPs [173, 83], their applicability to exhaustively explore the conformational space of these proteins is still limited. Knowledge-based approaches have emerged as an alternative to overcome some of these limitations. These approaches

usually describe the conformational properties of individual residues using the so-called coil libraries, which contain residue-specific $\{\phi, \psi\}$ angles from fragments of experimentally determined protein structures that do not form secondary structural elements [209, 103, 15, 66, 223, 203]. Despite their simplicity, coil models provide an accurate description of NMR parameters such as J-couplings [209, 203] and RDCs [15, 99], and SAXS curves [18] for flexible peptides and disordered proteins. Nevertheless, these approaches fail to identify secondary structural elements in IDPs. This limitation is caused by the chain building strategy, which sequentially appends individual residues accounting for the amino acid type and overlooks the sequence and structural context [103, 15]. Consequently approaches such as Flexible-Meccano provide excellent models for the random-coil but do not capture structural features involving multiple consecutive residues. The omission of coordinated effects precludes the capacity of current approaches to predict structural classes and their populations, and hamper their application for advanced purposes.

Here we present a new approach to build atomistic models of IDPs that uses an extensive coil library of three-residue fragments (presented in Chapter 3), which are the minimal fragments containing structural information [88]. The exploitation of the structural information encoded in the library provides accurate descriptions of RDCs and SAXS datasets for multiple disordered proteins presenting distinct secondary structural motifs. This observation suggests that, by capturing conformational restrictions in turns, α-helices, and β-strands inserted in IDPs, our structural ensembles are realistic models of these proteins. The relative population, the internal coordination that transiently stabilizes these secondary structural elements, and the fluctuating behavior of these elements naturally emerge from our strategy. Our study seeks to extend structure prediction approaches to disordered chains, thereby enabling the identification of the structural perturbations that deleterious point mutations or alternative splicing exert on IDPs and IDRs.

5.2 Materials and Methods

5.2.1 Sampling method

The sampling algorithms builds conformations incrementally from N- to C-termini in a residue-by-residue manner. When placing a new residue, its backbone angles $\{\phi, \psi, \omega\}$ are extracted from the coil database. An all-atom model is used for the backbone, whereas a simplified model was used for the side chains, considering a pseudo-atom placed at the Cβ position for each residue, as previously proposed [127, 15, 163]. When placing a new residue, collisions with the previously built residues are tested. In case of collision, a new configuration of the residue is sampled and tested. This is repeated until a valid configuration is found or a maximum number trials of 100 ($n_{fail}^{col} = 100$) is reached. In these cases, a backtracking search process is applied, which consists of removing the last three residues and restarting sampling from this point. When the backtracking process results unsuccessful, the chain construction is restarted from the beginning.

Different strategies can be used within this method:

Single-residue-based sampling (SRS): This strategy is similar to the one used in Flexible-Meccano [15, 163]. The backbone angles of each residue are sampled disregarding the neighboring residues. In this strategy, when the residue type is alanine, the angles are randomly selected among all tripeptide conformations of type X-Ala-Z, X and Z being any of the 20 amino acid types (i.e. 400 tripeptide sequence types). The process is slightly different when the Z residue is a proline. In this case, the conformation is selected from sequences X-Ala-Pro.

Three-residue-based sampling (TRS): This strategy takes into account the sequence of the neighboring residues $i-1$ and $i+1$ when sampling the conformation of residue i. In other words, when the amino acid types of residues $i-1$, i, $i+1$ are X, Y, Z, respectively, the conformation of residue i is sampled from the corresponding class X-Y-Z in the tripeptide database. In addition, the conformation of these two neighbors is considered in order to restrict sampling to the most structurally probable regions. For this purpose, sampling of residue i is constrained to a subset of conformations of the tripeptide class X-Y-Z, such that the backbone angles of residue $i-1$ are within a given angular range ($\pm 20^\circ$) around its current conformation, which was built in the previous step. Since the conformation of residue $i+1$ is not sampled in this building step, the structural restriction requires a back-step test. Once the conformation of residue i has been built, the conformation of the tripeptide formed by residues $i-2$, $i-1$, and i is checked to be present in the database of the corresponding sequence, considering the aforementioned angular tolerance. As for collision tests, this structural test can also fail. In this case, a backtracking process is also applied, with $n_{fail}^{str} = 250$.

Hybrid Sampling: The two sampling strategies SRS and TRS can be combined in the hybrid strategy. Based on experimental RDCs and on additional information from previous studies, TRS is applied to sample partially-structured regions while SRS is used for the disordered regions. Note that in the absence of experimental information, the predictor presented in Chapter 4 could be used to identify partially-structured regions.

5.2.2 Computation of experimental properties from ensembles

Alignment properties and associated RDCs for each conformation can be computed by exploiting the similarity between the radius of gyration and the alignment tensors as previously described in section 2.4.5.2 [3, 15]. In the results presented below, reported RDCs correspond to averages over 100,000 conformations of each ensemble. Computational RDCs were homogeneously scaled to minimize discrepancy with the experimental ones. The agreement of the resulting RDCs with the experimental ones was evaluated using the Q-factor [34]: $Q = \texttt{rms}(D_{\text{meas}} - D_{\text{calc}})/\texttt{rms}(D_{\text{meas}})$, where D_{meas} and D_{calc} are the experimental and computed RDCs, respectively. Ensemble-averaged SAXS data were computed from 2,000 randomly selected conformations from the ensembles generated with the hybrid sampling strategy. Side chains for each conformation were introduced with SCWRL4 [118] before compu-

Table 5.1: References to the articles from which the experimental SAXS curves for p15, USrc and Tau where obtained.

IDP	SAXS curves
p15	[40]
USrc	[5]
Tau	[155]

tation of its associated theoretical SAXS profile with CRYSOL [217] using default parameters. The ensemble-averaged curve was compared with the experimental one by optimizing a scaling and a shift parameter, using χ^2 as a figure of merit. Averaged Cα, Cβ, CO and NH chemical shifts were computed from ensembles of 5,000 conformations with SPARTA+ [201]. Side chains for each conformation were introduced with SCWRL4 [118] before the calculation. Random coil chemical shifts were computed using POTENCI [158] and subtracted from the computed ones to facilitate the interpretation.

5.3 Results

5.3.1 Computational models

We generated ensembles of 100,000 conformations for several IDPs using the different building strategies explained above. We used the same benchmark set of structurally-characterized IDPs than in the previous chapter, listed in Table 4.1. N-HN RDCs and SAXS curves were computed from the resulting ensembles using standard methods explained in the previous section, and were compared with the experimental RDCs for MAPK Kinase 7 (MKK7) [117], the fragment 955-1097 of the Erythrocyte binding antigen 181 (eba181) [22], p15 [40], sic1 [147], Measles virus ntail (ntailMV) [97], Sendai virus ntail (ntailSV) [98], the unique domain of the src kinase (USrc) [170], K18 fragment of Tau protein (K18) [154], and full-length Tau protein [194] were used to probe the residue-specific sampling of the models, including the presence of partially-formed secondary structural elements. The agreement of the different building strategies with the experimental data was quantified using Q-factors [34] (Table 5.2). Moreover, SAXS curves for p15 [40], USrc [5], and Tau [155] were used to probe the overall size and shape of the ensembles constructed (see Table 5.1).

5.3.2 The coil model describes disordered regions in IDPs

As a first approach, we built the conformations by randomly selecting $\{\phi, \psi\}$ values from the database in a residue-specific manner without taking into account the neighboring residues. Only residues preceding prolines were specifically selected from the database, since the Ramachandran distributions of these residues differ

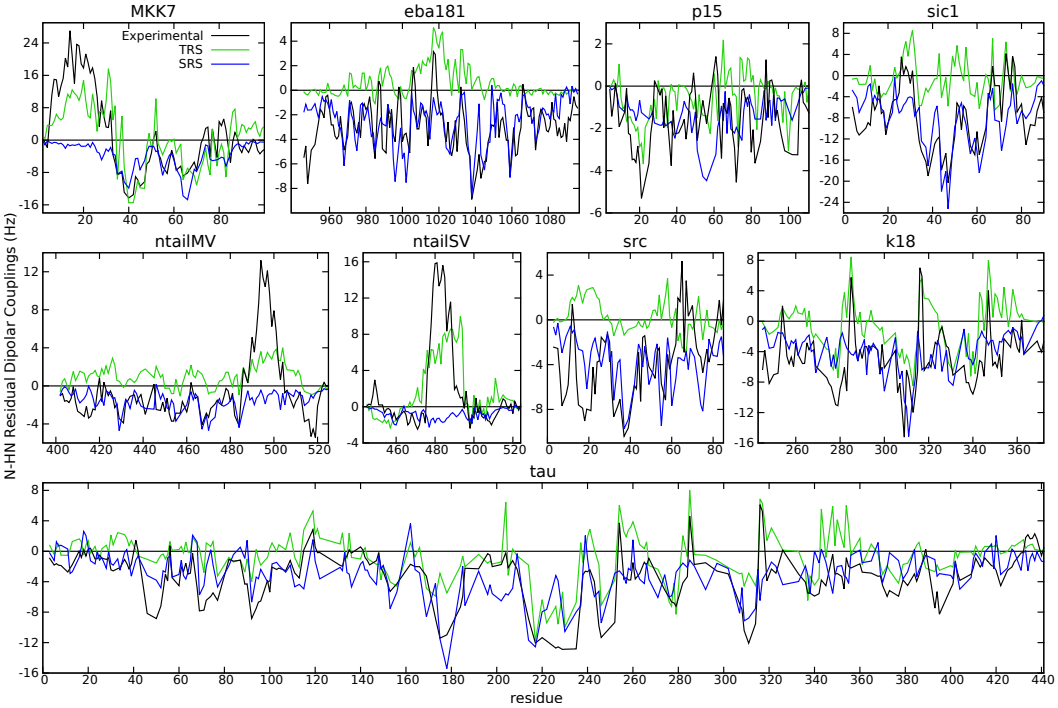


Figure 5.1. Experimental N-HN RDCs (black solid lines) for the nine proteins analyzed compared with the theoretical RDCs computed using the SRS (blue solid line) and TRS (green solid line) sampling strategies. To facilitate visual analysis, RDCs from the SRS method were scaled considering only the regions defined as random coil in the hybrid approach

considerably [139, 223]. This building mode, which we call single-residue-based sampling (SRS), can be considered a Flory model since the sequence context of the building units is not used. The RDC profiles computed using the SRS strategy nicely reproduced the experimental ones for large sections of all the proteins (Fig. 5.1, blue lines). Conversely, other regions displaying large (positive or negative) RDCs were not properly reproduced by SRS ensembles. Not surprisingly, this lack of agreement was observed in known α-helical regions with positive RDCs (ntailMV, ntailSV and MKK7), extended regions with strongly negative N-HN RDCs (p15), and turns displaying sharp positive peaks (eba181, K18 and Tau). Note that inaccuracies in the representation of partially structured regions have also been observed when using similar building strategies, such as Flexible-Meccano [15, 163]. The proteins with highly populated secondary structural elements, such as ntailMV, ntailSV and MKK7 present large Q-factors (around 100).

5.3.3 Structural information encoded in the tripeptide database identifies partially formed secondary structural elements

We generated large conformational ensembles using a three-residue-based sampling strategy (TRS) that selects $\{\phi, \psi\}$ values for each residue i, taking into account the

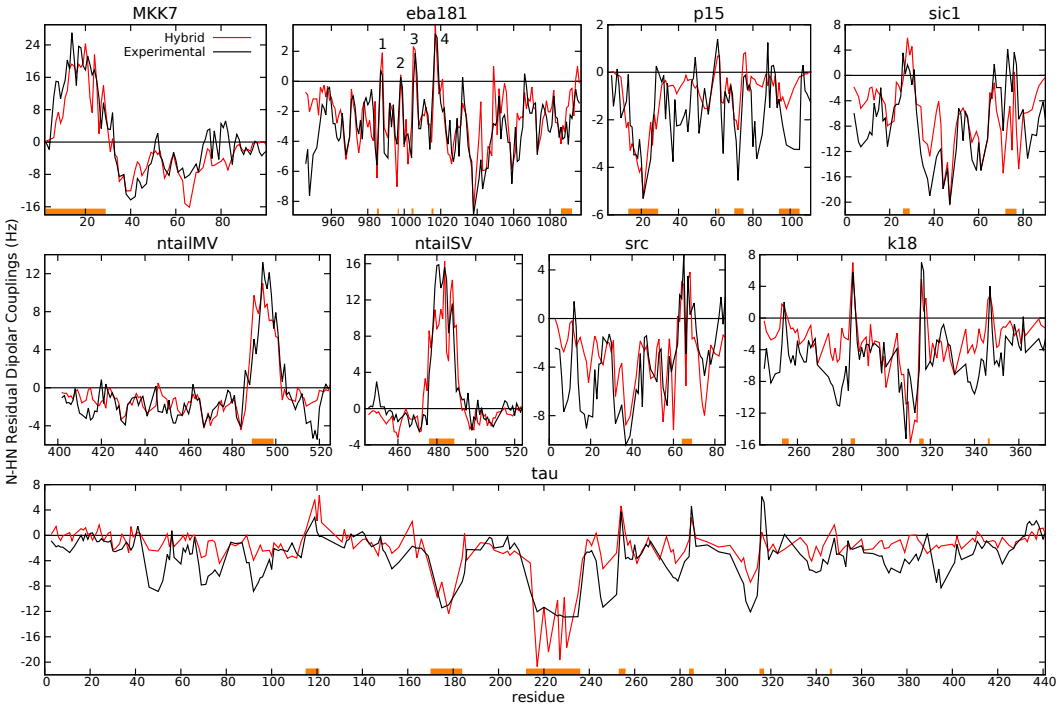


Figure 5.2. Experimental N-HN RDCs (black solid lines) for the nine IDPs studied compared with those computed using the hybrid SRS-TRS sampling strategy (red solid lines). Fragments highlighted in orange correspond to regions considered partially structured, for which the TRS was applied (see Table 5.3 for details).

amino acid type and the conformation of the neighboring residues $i-1$ and $i+1$ (see Method Details). In general, RDCs derived from the TRS strategy adopted less negative or even positive values compared to those obtained from the SRS strategy (Fig. 5.1, green lines). In some cases, such as for eba181 and ntailMV, almost the entire RDC profile remained positive. We attribute this systematic deviation towards positive values to an overpopulation of α-helical conformations in the tripeptide database, as previously observed when using coil libraries derived from globular proteins [103, 197]. Interestingly, some local features observed in the experimental profiles, which were not reproduced by the SRS strategy, were captured by the TRS strategy. Theoretical RDCs for α-helical regions in ntailMV, ntailSV and MKK7 were systematically more positive than those corresponding to their flanking regions. In fact, these were the only three cases for which the Q-factor for the TRS was better than that of the SRS. Moreover, turns in K18 and Tau were naturally pinpointed by the TRS strategy, producing sharp peaks in the RDC profile. Note that more negative RDC values were also observed in some cases, such as the N-terminus of p15. These observations indicate that some tripeptide sequences in the database are enriched in particular conformational classes that are present in solution.

Table 5.2: Q-factor and χ^2 values obtained from the comparison of the experimental and computational data for each of the proteins studied.

Protein	SRS		TRS		Hybrid		
	Q-factor RDCs	χ^2 SAXS	Q-factor RDCs	χ^2 SAXS	Q-factor RDCs	χ^2 SAXS	% Structure
MKK7	100.36		67.98		45.20		29.00
eba181	56.01		114.33		46.90		8.18
p15	79.61	1.03	88.11	1.01	63.12	1.04	30.90
sic1	50.67		92.16		53.29		9.78
ntailMV	98.98		91.07		47.23		8.33
ntailSV	110.61		68.97		43.97		15.85
USrc	68.88	2.70	114.16	2.58	60.57	1.93	5.26
K18	63.15		83.61		59.32		8.34
Tau	63.32	2.02	77.87	2.15	60.37	1.52	18.55

5.3.4 A hybrid sampling strategy simultaneously describes structural properties of disordered and partially ordered regions

The satisfactory description of disordered and partially structured regions achieved with the SRS and TRS strategies, respectively, prompted us to apply a hybrid building approach. In this approach, residues belonging to a partially structured region defined *a priori* were incorporated into the model using the TRS strategy, while the rest of the chain was built with the SRS strategy. For the nine proteins tested, we defined the partially structured regions on the basis of the experimental N-HN RDCs and previously reported structural analyses (see Table 5.3). In this regard, SRS-derived RDCs were compared with the experimental ones, and those regions presenting a systematic deviation were initially assigned as partially structured. The exact borders of these regions were subsequently refined by testing multiple alternatives. The Q-factors, revealed excellent agreement between the simulated and the experimental RDC profiles for all the proteins tested (Fig. 5.2 and Table 5.2). This metric thereby indicates that the hybrid strategy, which simultaneously describes disordered and partially structured regions, notably improved the SRS and TRS chain building approaches. However, the level of Q-factor improvement depended on the percentage of the sequence involved in secondary structural elements. In highly disordered proteins such as eba181, the improvement of the hybrid method with respect to the SRS approach was modest, with Q-factors of 56.01 and 46.90 for the SRS and hybrid strategies respectively. Conversely, a considerable improvement in the Q-factor was observed in proteins with long and highly populated α-helices, such as MKK7, ntailMV and ntailSV, whose Q-factors decreased from 100.36, 98.89 and 110.62 for SRS to 45.20, 47.23 and 43.97 with the hybrid strategy, respectively.

Computed RDCs for the α-helical regions of MKK7, ntailMV and ntailSV nicely

Table 5.3: Regions defined as partially structured in the hybrid strategy for the studied proteins.

IDPs	TRS regions
MKK7	[A2-D29]
eba181	[D985-P986], [P997], [D1004-A1005], [D1015-D1016], [K1086-G1092]
p15	[Y13-S29], [W61-Q62], [R70-D75], [A94-H105]
sic1	[S26-L29], [P72-T77]
ntailMV	[R489-Q499]
ntailSV	[V476-E489]
USrc	[F64-S69]
K18	[L253-V256], [L284-N286], [L315-K317], [F346-K347]
Tau	[E115-H121], [R170-S184], [T212-P236]
	[L253-V256], [L284-N286], [L315-K317], [F346-K347]

reproduced the experimentally observed bell-shape and the saw-teeth. Importantly, the description of the positive RDCs did not compromise that of the disordered regions as the model captured their relative intensity. Other characteristic features observed in the experimental RDC profiles, such as turns in eba181, K18 and Tau (see below), the broken helix in the 60-75 fragment of USrc caused by two consecutive glycine residues [170], and the sharp inverse γ-turn of W61 of p15 [40], naturally emerged when using the hybrid approach. Remarkably, this building method did not require the specification of either the type or the population of secondary structures. Protein Tau is a particularly challenging example due to its size and the presence of multiple structural features, which have been extensively studied by NMR [154, 164, 194]. Seven regions of Tau were defined as structured using the hybrid approach, four of them being the well described turns found in the repeat region corresponding to the K18 construct [154, 164]. The presence of highly positive RDC values found in these four turns were captured by the hybrid approach in both proteins (Fig. 5.2), thereby indicating the accurate conformational representation of their sub-sequences in the database.

Chemical Shifts (CSs) were used to further validate the conformational ensembles built with the hybrid SRS-TRS strategy. In this regard, averaged Cα, Cβ, CO and NH CSs for ntailMV were computed from the ensembles using the program SPARTA+ [201] and then compared with the experimental ones (Fig. 5.3). The simulated CSs were in good agreement with the experimental ones, and they clearly captured deviations from the purely random coil behavior represented by the SRS ensemble. These observations substantiate the results obtained when using RDCs.

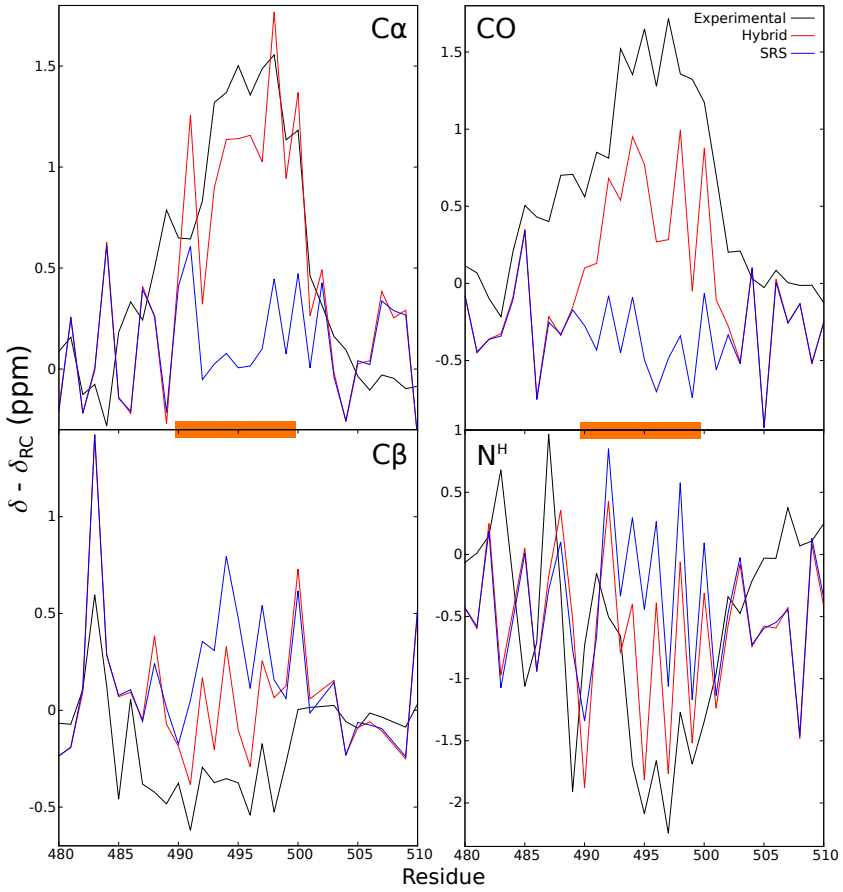


Figure 5.3. Comparison of the experimental (black) Cα, Cβ, CO and NH chemical shifts of ntailMV with these obtained from the ensembles built with the SRS (blue) and hybrid (red) strategies. Orange bars indicate these regions defined as structured in the hybrid modeling.

5.3.5 Comparison to SAXS data

SAXS accurately probes the overall properties of conformational ensembles in solution, thus complementing the residue-specific information provided by RDCs and CSs [33, 207]. Simulated SAXS profiles were computed from the ensembles using standard procedures (see Method Details). Overall, excellent agreement between experimental and simulated profiles was observed for the three proteins, with χ^2 of 1.93, 1.04, and 1.52 for USrc, p15 and Tau, respectively (Fig. 5.4). For USrc and Tau, these values were notably better than those obtained with the SRS (χ^2 of 2.70 and 2.02) and the TRS (χ^2 of 2.58 and 2.15) sampling approaches. For p15, the profiles achieved the three sampling strategies showed an excellent correlation with the experimental profile, with χ^2 near 1.0. These results strongly suggest that the ensembles built with the hybrid approach properly describe the overall properties of IDPs.

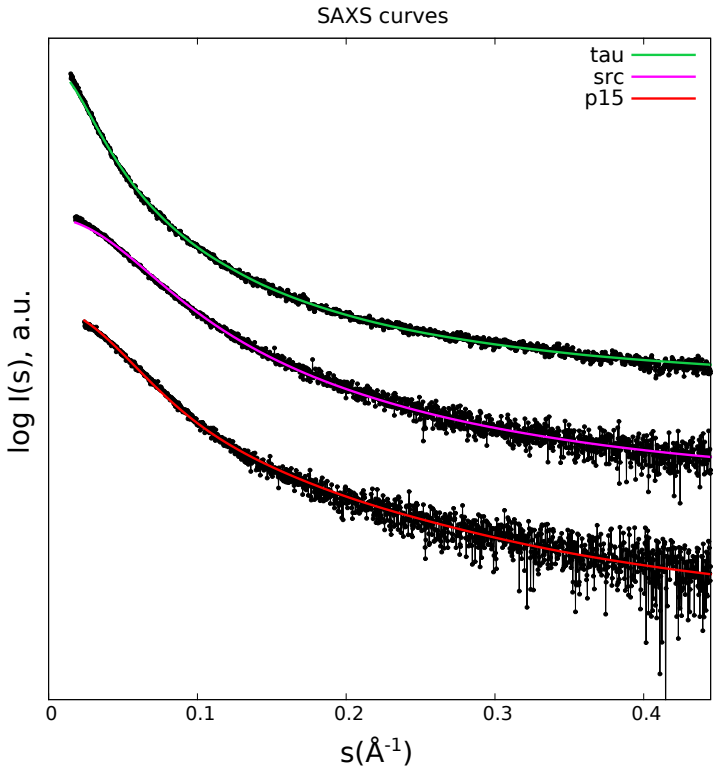


Figure 5.4. Experimental scattering intensity (black dots) in logarithmic scale as a function of the momentum transfer, s, compared with the averaged profiles computed from the hybrid ensemble models for tau (green), USrc (pink), and p15 (red). The profiles have been displaced along the y-axis for a better inspection.

5.3.6 Prediction of local conformations and secondary structural elements

The previous sections demonstrate that the ensembles built with the hybrid approach are excellent representations of IDPs in solution. Next, we explored the structural features of the resulting models using the helical region in ntailMV, the extended region at the N-terminus of p15, and the turns in eba181 and K18 as examples.

For ntailMV, the hybrid strategy notably enriched the structured region in α-helical conformations while it was depleted in extended (β-S) and polyproline-II (β-P) (Fig. 5.5a). This structural enrichment in helical conformations induced positive RDC values in this region. The conformational analysis of the ensemble built for the N-terminus of p15 indicated a strong enrichment in extended conformations, β-S and β-P, whereas α-helical ones were depleted (Fig. 5.5b). Interestingly, neither β-S nor β-P were homogeneously populated along the segment, and either one or the other became dominant depending on the specific sequence.

A highly relevant feature of the hybrid strategy is its ability to identify turns from sequences. Four turns have been localized in eba181 based on their positive

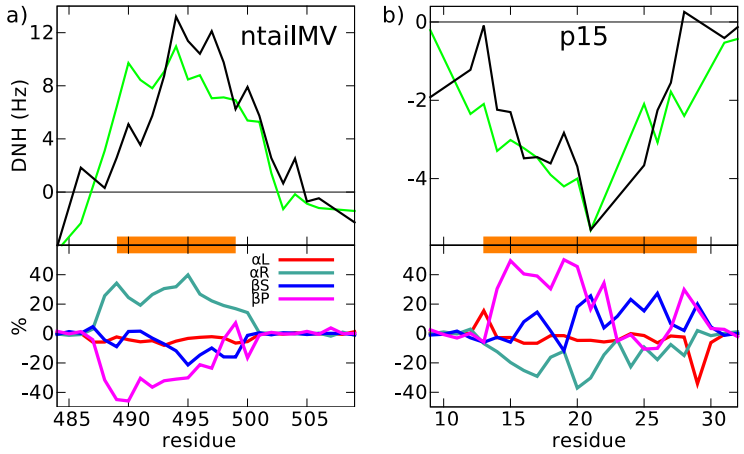


Figure 5.5. Experimental (black) and hybrid building model (green) N-HN RDCs for two fragments of (a top) ntailMV and (b top) p15. Fragments highlighted in orange were considered partially structured and built using the TRS strategy. In bottom panels, the percentage of enrichment of secondary structure classes present in the ensemble built with the hybrid strategy compared with that built with the SRS strategy. Secondary structure classes were identified using definitions in related work [164]. Concretely, $[\beta S : -100 > \phi; -120 > \psi > 50]$, $[\beta P : 0 > \phi > -100; -120 > \psi > 50]$, $[\alpha R : 0 > \phi; 50 > \psi > -120]$, $[\alpha L : \phi > 0]$.

RDCs [22], however the sizes of these RDCs differed (Fig. 5.2). While turns 3 (DASL) and 4 (DDAK) presented highly positive values, turns 1 (DPEK) and 2 (DPNT) were only slightly positive thereby suggesting distinct structural features. Fig. 5.6 shows the conformations adopted by the residues involved in the four turns. In all turns, residue $i+1$ adopted an α-helical conformation. However, while residue i in turns 1 and 2 was mainly extended due to the following proline, it was α-helical in turns 3 and 4. This structural difference most probably explains the distinct RDC values of the four turns. According to current definitions [41], the four turns can be considered β-turns, types *I* and *VIII* being compatible with the conformation of the residue $i+1$. Nevertheless, the sequence composition clearly suggests that turns 1 and 2 with D and P in positions i and $i+1$, respectively, are type *I* β-turns [41]. In another example, the four turns identified in K18 were enriched in α-helical conformations in their two central residues (Fig.5.7), an observation that is in line with the original study [154]. However, residues in position $i+1$ (L253, L248, L315 and F346) sampled the region $\{\phi = -90, \psi = 0\}$ whereas residues $i+2$ (K254, N285, S316, and K347) adopted mainly an α-helical conformation with $\{\phi = -60, \psi = -30\}$. Although resembling type *I* β-turns, they did not adopt the canonical conformation [41].

5.3.7 Coordinated formation of structural elements

We further studied how secondary structural elements are formed within the conformational ensembles using the helical region in MKK7 as an example (Fig. 5.8a).

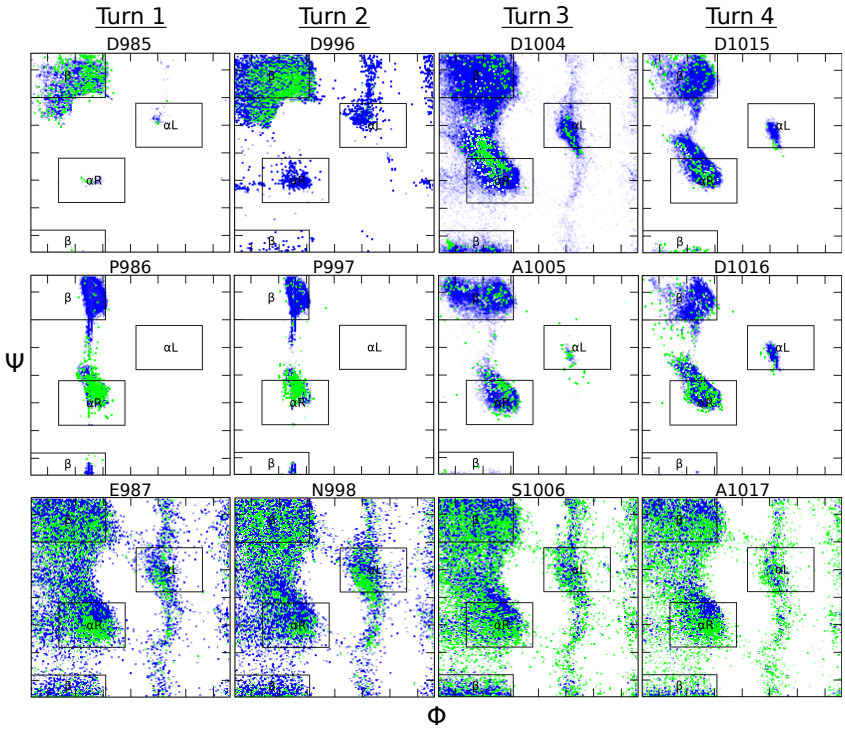


Figure 5.6. Conformational sampling for the four turns identified in eba181. Each column displays the Ramachandran plots for the three first residues in turns 1 to 4 when using the SRS (blue) or the hybrid (green) sampling approaches.

The Secondary Structure-map (SS-map) [91], which allows the quantification of multiple structured elements within conformational ensembles, was used for this analysis. According to the SS-map, the ensemble of the N-terminal region of MKK7 presented scarcely populated helical regions of virtually all sizes from 4 up to 28 residues. Although the helix encompassing the whole 28-residue-long region was found in the ensemble, its population was extremely low, and shorter α-helices were preferred. In this regard the most populated helices (around 5%) involved eight and nine residues in non-overlapping segments of the protein. Interestingly, the N-terminal region of this fragment seemed more prone to form long α-helices expanding up to 15 residues. The continuum of multiple overlapping helical sections observed in the ensemble of MKK7, which induces the bell-shape of the resulting RDC profile, highlights the conformational complexity of helical regions in IDPs.

We tested two alternative procedures to introduce helicity into ensembles generated using a Flory model (i.e. the SRS strategy in our implementation) that are frequently used to describe NMR data [40, 164, 170, 238, 13]. Firstly, a 25% increase in α-helical conformations was imposed for each of the residues within the region, but no structural coordination between residues was forced (Fig. 5.8b). Secondly, a canonical α-helix spanning the 28-residue-long region was introduced in 25% of the conformations (Fig. 5.8c). When the helical tendency was increased at the residue level, the resulting ensemble displayed multiple short helices spanning

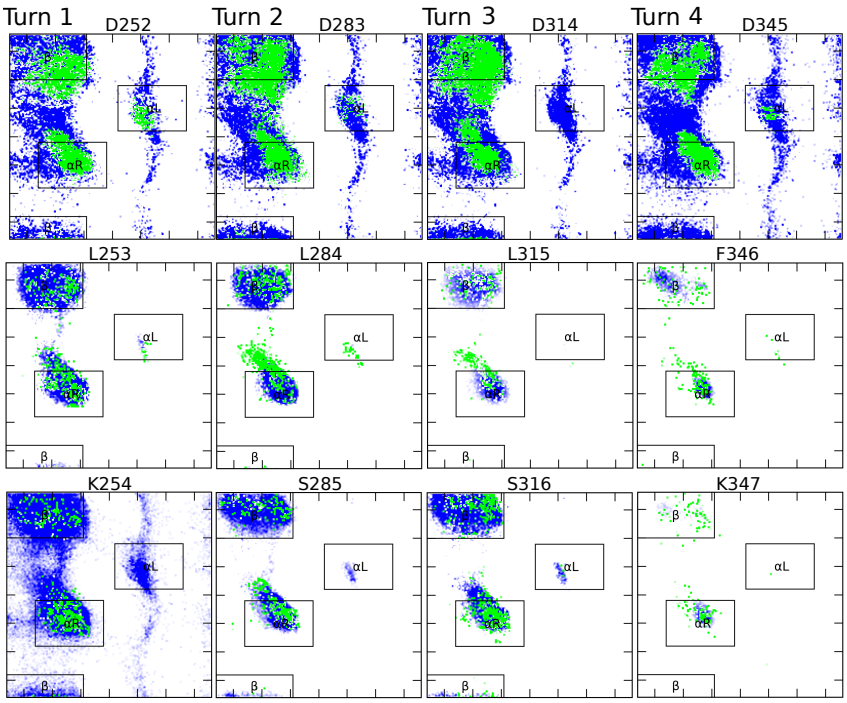


Figure 5.7. Conformational sampling for the four turns identified in K18. Each column displays the Ramachandran plots for the first three residues belonging to turns 1 to 4 when using the SRS (blue) or the hybrid (green) sampling approaches.

the whole region. However, the population of longer helices decreased dramatically. Consequently, resulting RDCs were positive but with values close to zero and they did not display residue-specific features. When a canonical α-helix was forced within the complete region no shorter helices spontaneously formed in the remaining 75% of the ensemble. As a result of this conformational homogeneity, the RDC profile adopted large positive values with the saw-teeth shape induced by the continuous α-helix. However, RDCs did not present the overall bell-shape observed experimentally.

To further evaluate the ensembles generated with the aforementioned procedures, we also used two-dimensional plots that display the deviation with respect to a canonical α-helix (see Fig.5.9 with the associated explanations). In these plots, each conformation is represented by a point with the x coordinate corresponding to the distance between the first N and last C backbone atoms of the 28-residue fragment of MKK7, and the y coordinate corresponding to the average distance between H-bond donor and acceptor atoms within this fragment. The proposed hybrid sampling strategy (Figure 5.9, left image), using TRS for the selected 28-residue region, produces a wide variety of conformations, and several regions of the plot are more densely sampled. In particular, we can observe more conformations near the coordinates corresponding to the canonical helix, but slightly shorter (in terms of end-to-end distance). This can be explained by the propensity to form shorter helices, as observed in the the SS-map, giving the possibility to flanking

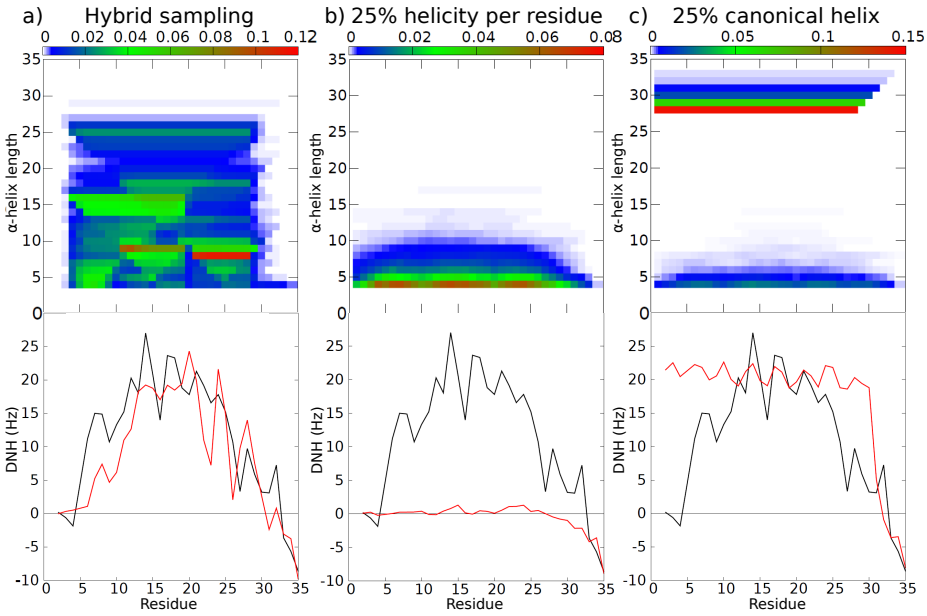


Figure 5.8. Structural analysis of the helical region in MKK7. (Top panels) Length and encompassing residues of the α-helices found in ensembles computed using (a) the hybrid sampling and two theoretical models imposing (b) 25% of enhanced helicity per residue, and (c) 25% population of a canonical α-helix in the 28-residue long segment. Colors from white to red indicate the population of helical segments found in the ensembles. (Bottom panels) Theoretical RDCs calculated from the above described ensembles (red lines) compared with the experimental ones (black lines).

regions to fold back. Using the procedure that imposes helicity at the residue level (Figure 5.9, middle image), the distribution of the sampled conformations is much more compact, and the region near the canonical helix is not sampled at all. This is in line with the SS-map representation, showing that only very short helices are formed. Finally, the plot corresponding to the other procedure clearly shows the imposed 25% populated canonical helix (Figure 5.9, right image), being the rest of the conformations far from it. Such a discontinuity in the conformational space is unrealistic. Indeed, although this last procedure can in some cases yield a good agreement between computationally generated ensembles and experimental data, these ensembles are inaccurate representations of the conformational heterogeneity of partially structured regions in IDPs.

A SS-map analysis was also performed in the helical regions of ntailSV and ntailMV (Fig. 5.10). As in the case of MKK7, the co-existence of multiple overlapping short α-helices was observed. However, in contrast to MKK7, these two proteins displayed a triangular shape in the SS-map, in agreement with their similar amino acid sequence and function. This shape arises from the presence at the N-terminus of a motif of multiple residues with a strong tendency to trigger the formation of α-helical segments. The most prevalent initial residues of the detected helices in our ensembles were aspactic acid and serine. These two amino acids have

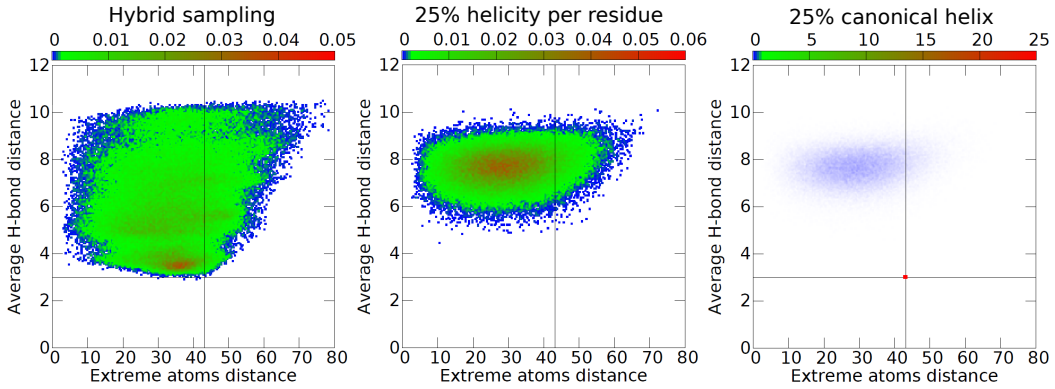


Figure 5.9. Deviation with respect to a canonical helix of conformations of the N-terminal region of MKK7 generated using different sampling strategies. The abscissa axis corresponds to the distance between the first N and last C backbone atoms of the 28-residue long structured region of MKK7. The ordinate axis corresponds to the average distance between H-bond donor and acceptor atoms. For a canonical 28-residue long helix these values are 3 Å and 43.5 Å, respectively.

been identified as helix N-capping amino acids, which stabilize α-helices with their side chain by forming a hydrogen bond at positions 2 or 3 in the helix [98, 137]. This observation suggests that the N-capping properties of these amino acids are encoded in the tripeptide database and that their capacity to initiate helical motifs naturally emerges in ensembles built with the hybrid strategy.

5.4 Conclusion

Partially structured motifs are key elements to trigger signaling events and to regulate transcription and metabolic pathways [227]. The localization and characterization of these motifs inserted within fully disordered fragments have been the focus of intense research [227, 232, 150]. Here we present an approach that exploits the structural information encoded in tripeptide fragments extracted from coil regions of experimentally determined protein structures to build accurate structural ensembles of IDPs/IDRs, including scarcely populated structured motifs. Although Flory models, which do not consider the sequence context, generate conformational ensembles with the capacity to reproduce diverse experimental data for disordered chains, they fail to predict and model partially structured elements. Our results demonstrate that the tripeptide database, which accounts for this sequence context, contains structural features that are subsequently found experimentally in solution. Whereas libraries involving larger fragments have been shown to be powerful tools for the prediction of probable (stable) conformations of globular proteins and peptides [79, 116, 184, 8, 202, 140], our results highlight that our extensive database of three-residue fragments is enough to accurately represent the conformational variability and local structural propensities in IDPs. Moreover, representing the conformational variability of disordered chains requires a broad sampling of struc-

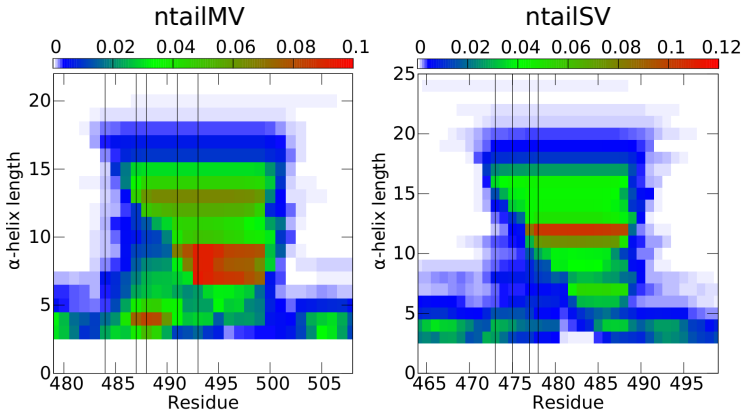


Figure 5.10. SS-map analysis for the helical regions in ntailMV and ntailSV displaying the length and the composing residues of the α-helices found in ensembles generated with the hybrid sampling strategy for both proteins. Color from white to red indicates the population of these helices. Vertical lines indicate aspartic acids and serines in the sequence that act as helix N-capping residues. Concretely, D484, D487, S488, S491, and D493 are highlighted for ntailMV, and D473, D475, S477, and D478 and highlighted for ntailSV.

tures, which would not be guaranteed using databases of larger fragments. In this regard our tripeptide emerges as optimal for this purpose.

The general agreement between experimental and simulated RDCs implies that the residue-specific structural information encoded in our tripeptide database is coherent with the conformational behavior of IDPs in solution. This is a remarkable observation as the database has been derived from coil regions of crystallographic structures, which are susceptible to experience packing contacts and/or reduced mobility. Therefore, the sequence context is a major determinant of structural propensities, regardless of the state (globular/disordered) or the environment (crystal/solution). However, for some sequences, a less accurate agreement between the experimental and simulated RDC profiles has been observed. We attribute this punctual lack of agreement to the limited conformational coverage of these sequences in our database. With the increasing number of experimentally determined high-resolution protein structures, we expect that more extensive and higher quality tripeptide databases will be built in the future, which will further improve the accuracy of conformational ensembles generated with our method.

Our approach relies on the discrimination between disordered and partially structured regions to subsequently apply the SRS and TRS sampling strategies, respectively. Here we have used the experimental RDCs and previous studies of the considered proteins to define both regions. In the absence of RDCs, other experimental data and bioinformatics predictions can be used to identify partially structured motifs. CSs, which are the primary information derived from NMR, are also very sensitive to small conformational bias at the residue level [221, 196]. Partially structured motifs can also be discriminated from fully disordered regions by their faster NMR transverse relaxation rates [97, 40]. In the absence of experimental information or to complement it, bioinformatics tools, such as the one

presented in Chapter 4, can be applied identify regions prone to forming secondary structure elements. Another interesting source to distinguish structured elements is sequence conservation analysis. In IDPs, motifs involved in protein-protein interactions present slower mutational rates when compared to non-functional regions [160].

Partially structured motifs are not permanently folded in IDPs. They can be seen as an equilibrium between conformations hosting distinct smaller structured elements that are in continuous exchange driven by their extension or shortening. In other words, these sequences lack the internal coordination to form permanent secondary structural motifs and, as a consequence, are susceptible to partial unfolding events. Recognition processes exploit this structural heterogeneity to efficiently achieve the desired biological tasks. Binding affinities of the co-existing conformers are modulated by the entropic penalty caused by the folding of the recognition motif fragment that remains disordered in the unbound state [167]. Moreover, recognition kinetics studies have demonstrated the existence of transiently populated encounter complexes, and different conformational states of the recognition element most probably present distinct energy barriers to achieve the final bound form [192, 215, 46]. In the context of RDCs, the coexistence of multiple partially folded helical elements in the same region leads to the bell-shaped RDC profile and the saw-teeth, which report on the prevalence of the different helical fragments. Importantly, this structural heterogeneity is nicely captured by our hybrid sampling strategy, thereby highlighting the correspondence between the information encoded in the database and the conformational sampling of IDPs in solution. This feature is exemplified by the helix N-capping properties that we observed in the ensembles of ntailMV and ntailSV.

In summary, we have developed a method to build realistic conformational ensembles of IDPs and IDRs that describes scarcely populated secondary structural elements embedded in otherwise fully disordered regions. Our strategy is based on an extensive database of tripeptide structures and on the separation between disordered and conformationally biased regions within the chain. This approach detects binding motifs involved in partner recognition that are, in most cases, linked to biological tasks. Our approach has the potential to anticipate structural effects caused by point mutations with an eventual role in disease, and the insertion or deletion of disordered fragments originating from alternative splicing processes. In this regard, we believe that our approach is the first step towards extending structural bioinformatics and protein design to disordered proteins.

Chapter 6

A heuristic search algorithm to investigate the formation of structural elements

Contents

6.1 Introduction

As mentioned in previous chapters, some regions in IDPs are partially structured, meaning that secondary structure elements constantly form and vanish. These transient structural elements, usually called Molecular Recognition Elements (MOREs), are functionally important in many cases, since they are involved in the recognition of molecular partners [167, 226]. MOREs recognize their globular partners with high specificity while displaying a moderate affinity, explaining their fundamental role in signalling, metabolic regulation and homeostasis [227]. In this chapter, we present a method to investigate the formation of these structural motifs. In addition to the interest for the study of IDPs, the proposed method has a wider range of applications, such as the investigation of folding mechanisms in proteins.

Understanding the mechanisms of protein folding and unfolding as a function of the amino acid sequence is of paramount importance, giving their relevance in biological processes [233]. Furthermore, numerous diseases are related to the inability of proteins to fold correctly or to form insoluble amyloidogenic aggregates due to mutations or metabolic deregulation [230, 112].

Intensive research efforts over several decades, using both experimental and computational approaches, have yielded important bricks of knowledge on the underlying mechanisms of protein folding, unfolding and other conformational transitions [9, 241, 52, 186, 132, 19]. Nevertheless, we still lack of a complete understanding of these mechanisms. Some theories about protein folding give more importance to interactions between the protein side chains, whereas others consider that the propensity of protein backbone fragments to form secondary structural elements, such as α-helices, β-sheets and turns, is the most important mechanism for protein folding.

We believe that local, sequence-dependent structural preferences are essential to drive the formation of structural elements, while other phenomena such as hydrophobic effects or electrostatic forces help stabilizing the overall structure. Following this hypothesis, we propose a theoretical approach to compute conformational transitions using local structural information extracted from experimental data. Interactions between distant residues are (explicitly) neglected for the exploration of transition paths, with the exception of collisions that would lead to unrealistic conformations. However, as further explained below, non-bonded interactions associated with local structural preferences are implicitly considered, and can be propagated along the sequence thanks to the application of constrains within the path search algorithm.

Information extracted from experimentally determined protein structures is frequently used in computational biology. The usual usage is the prediction of the conformation of the protein side chains, using the so-called *rotamer* libraries [55], which encode the most frequent values of the side chain dihedral angles for each amino acid type. The construction of protein backbone structural databases is less straightforward than for the side chains as it requires to subdivide proteins into fragments. The length of the fragments and considerations regarding the amino acid sequence may depend on the specific application. As also explained in previous chapters, statistics about the most frequent values of the backbone dihedral angles of amino acid types have been frequently used to explore the conformational sampling of highly-flexible proteins or regions [209, 103, 15]. However, such minimalistic single-residue fragments neglect the effects exerted by neighboring residues. Structural libraries involving larger fragments (usually, from 3 to 14 residues) have been shown to be powerful tools for the prediction of probable (stable) conformations of globular proteins and peptides [116, 184, 8, 143]. Fragment libraries can also be used to investigate conformational transitions in proteins. In a recent work, local moves using a fragment library were combined with other types of structural perturbations to compute transitions between several folded states of a protein [152]. Since the aforementioned fragment libraries were mainly conceived for protein structure prediction, they are focused on the most probable conformations of small and medium-sized fragments. As a consequence, they are not exhaustive enough for the study of conformational transitions. This limitation is more evident when the length of the fragments increases. Fragments involving three consecutive amino acid residues (*tripeptides*) represent a good trade-off between sequence-dependent struc-

tural preferences and exhaustiveness. Indeed, tripeptides contain relevant structural information [88] and are sufficiently small to capture the conformational variability of the 20 proteinogenic amino acids in their sequence context. In Chapter 5, we showed that an extensive database of tripeptides allows to accurately sample the conformational variability of IDPs. Here, we exploit the combination of this type of local structural information with a path search algorithm to compute conformational transitions in small proteins and protein fragments corresponding to relevant structural elements.

A protein cannot exhaustively explore its huge conformational space to seek transition pathways. This idea, referred to as the Levinthal's paradox [126, 185], is widely accepted. Indeed, a protein performs some search process to find the most efficient folding and transition pathways. We can say that the protein follows a powerful *heuristic* to avoid exploring an astronomically large number of possible pathways. This heuristic is not well understood yet, but, as mentioned above, we believe that local sequence-dependent structural preferences play an important role in it. Our contribution investigates this open question, and proposes a simple, heuristically-guided search algorithm, inspired from Artificial Intelligence (AI) and Robotics, to compute conformational transitions. AI and Robotics planning representations and techniques have been found valuable for solving several computational biology problems [2, 71, 200]. This chapter illustrates through an original approach their effectiveness in modeling folding mechanisms of structural elements in proteins.

The approach presented herein is very different from the ones in related works. First, the structural information is collected and used in a different way, and secondly, the algorithmic approach is totally different. Concretely, we use a heuristically guided depth-first algorithm, adapted from search techniques in constraint satisfaction problems over finite sets (CSP) and in automated task planning [70]. In our case, the state variables are the protein tripeptides, which range over finite sets of conformations extracted from a global database. The equivalent of an *action* is a constrained local change in a state variable. The algorithm relies on *adjacency graphs* of the state variables [183], which are computed at preprocessing time and are essential for efficiently testing the feasibility of transitions and for calculating the heuristic, which is based on statistical physics considerations. Our approach tends to favor paths going through high-density states, which are the most probable ones according to experimental observations recorded in the structural database. In other words, if we assume that the probability of the observed states for each tripeptide follows a Boltzmann distribution, we can say that the path search tends to follow the valleys of the free-energy landscape [236]. The search process also gives priority to short paths, which should correspond to faster transitions. The structural preferences for a tripeptide (*i.e.* at the state variable level) tend to be propagated along the sequence due to constraints imposed on the bond angles in the state transition validation, which reinforces neighbor-dependent structural preferences encoded in the database (see Section S2 in supplementary material for details). Thus, the path search process incorporates in an implicit way non-local

interactions along the sequence such as backbone hydrogen bonds in α-helices.

We applied our approach to two synthetic mini-proteins, Chignolin [85] and DS119 [129], which were particularly designed to fold into well-defined structural motifs present in natural proteins. These two molecules have been investigated in recent years using different methods [61, 178]. The results reported in this chapter are consistent with respect to those described in related literature, and already show the interest of the proposed approach, which is extremely fast when compared with currently-used computational methods based on molecular dynamics (MD) simulations [181]. Indeed, MD simulations of large-amplitude protein motions require *ad-hoc* computer architectures [132] or massively-distributed computing [210]. The efficiency of our approach allows to widely investigate, with modest computational resources, the effect of mutations on protein folding and unfolding, or on other functionally-important conformational transitions.

6.2 Materials and Methods

The proposed approach relies on a large database of protein structures, represented as sequences of partially overlapping tripeptides (the construction of the database was explained in Chapter 3). As stressed above, tripeptides are the minimal structurally-relevant units in proteins. The conformational transition problem is formalized as a search in a space of tripeptide conformations for a feasible path from an initial state to a target state of a protein. The state variables correspond to tripeptides; their values are the conformations of tripeptides actually observed and recorded in the database. A state variable in the sequence describing a protein shares its first two residues with its predecessor and its last two with its successor state variables in the sequence (see Figure 3.2). A transition between two values of a state variable is feasible if it meets a consistency constraint with respect to the predecessor and successor state variables, and if the corresponding conformation of the protein is collision free. The search algorithm seeks a feasible path using a heuristically-guided depth-first search schema. The heuristic function is a weighted sum of the distance between two conformations, an estimate of the distance to the target and a density term to advantage energetically favorable states.

We present next how the information in the tripeptide database is used in the present context. Then the statement of the conformational transition problem as a discrete path search problem is presented. Finally, we detail the proposed algorithm and the heuristics used to solve this problem.

6.2.1 Use of the structural database

Let $\mathscr{X}$ be the set of all $8,000$ tripeptides. An element $x_i \in \mathscr{X}$ is a state variable in our representation. Let D_i be the set of all the conformations of x_i recorded in our database. The conformation of x_i is characterized by the six backbone dihedral angles of the three residues in the tripeptide, denoted $\phi_{i,j}$ and $\psi_{i,j}$, for $1 \leq j \leq 3$. Although a conformation is characterized by an angular vector of 6 real numbers,

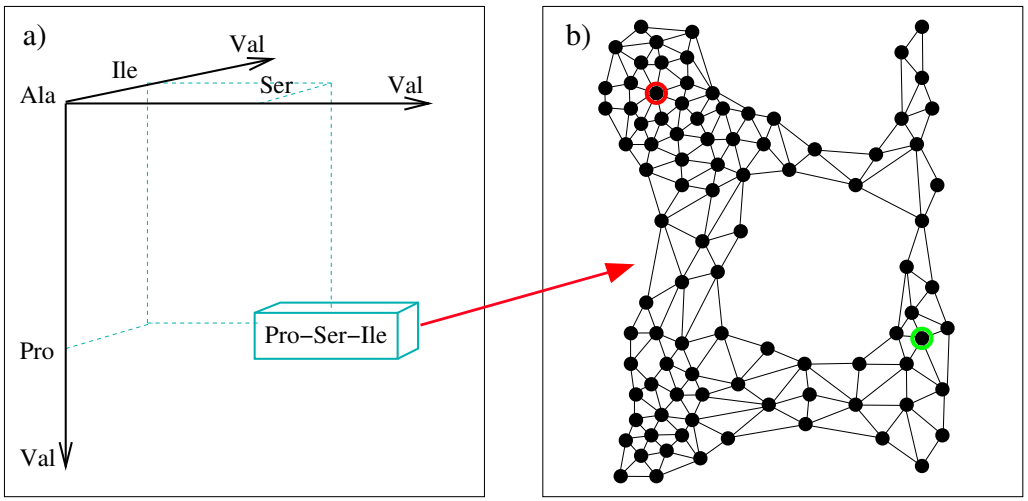


Figure 6.1. (a) Database containing one record for each tripeptide (8,000 in total). (b) For each tripeptide, the conformations recorded in the database are related with a proximity criterion and structured into an adjacency graph (the figure shows a simplified representation of this graph for tripeptide Pro-Ser-Ile).

for the purpose of our search algorithm over biologically observed conformations, we consider that the range of each state variable x_i is the finite set D_i of the recorded conformations in the database. We write $x_i = v_i$ for some $v_i \in D_i$.

The distance $d(v_i, v_i')$ between two values v_i and v_i' is defined as the angular root-mean-square deviation (RMSD) between the two corresponding angular vectors. More precisely:

$$d(v_i, v_i') = \sqrt{1/6 \sum_{j=1}^{3} \left((\phi_{i,j} - \phi'_{i,j})^2 + (\psi_{i,j} - \psi'_{i,j})^2 \right)}$$

We also define the central distance $d_c(v_i, v_i')$ with an identical formula for $j = 2$ solely, *i.e.*, restricted to the central amino acid residue of x_i. The idea is to compute a feasible path in the conformations of a protein as a sequence of elementary transitions focused on the central residue of each tripeptide.

These distances d and d_c allows us to structure the finite range D_i of each state variable as an *adjacency graph*, as illustrated in Figure 6.1.b. Its vertices are the elements in D_i. There is an edge (v_i, v_i') when $d_c(v_i, v_i') < \theta$ and $d(v_i, v_i') < \theta + \xi$, where θ is a variable adjacency threshold and ξ is a small constant tolerance margin. The adjacency threshold θ represents a tradeoff between a fully connected graph (no transition constraints between conformations) and an unconnected one (unreachable conformations), both cases being unrealistic. We set the threshold such that the adjacency graph of each tripeptide has a single connected component with moderate edge connectivity. This threshold θ is slightly different for different tripeptides, with an average value around 1.0 radian. The value of ξ was set to 0.35 radians in all

the cases.

The vertices are also characterized by a density function defined as follows:

$$\rho(v_i) = 1 + |\{v'_i \mid v'_i \text{ connected to } v_i \text{ and } d(v_i, v'_i) < \zeta\}|.$$

The threshold ζ has to be smaller than the adjacency threshold θ. Here, we set $\zeta = 0.2$ radians for all the tripeptides. The density ρ is related to the probability of existence of the corresponding conformation of the tripeptide. Considering basic principles in statistical physics (*i.e.*, the Boltzmann distribution), this probability depends on the energy of the state of the molecule. Thus, the most dense regions in the adjacency graph are also the most energetically-favorable ones.

6.2.2 Formal statement of the conformation path finding problem

A protein (or protein region) of interest is defined by a sequence of state variables $\langle x_1, \ldots, x_i, \ldots, x_n \rangle$, with overlaps. For example, the mini-protein Chignolin is a sequence of 10 amino acid residues: ⟨Gly-Tyr-Asp-Pro-Glu-Thr-Gly-Thr-Trp-Gly⟩; it is defined with 8 state variables x_1 =Gly-Tyr-Asp, x_2 =Tyr-Asp-Pro, ... x_8 =Thr-Trp-Gly. Hence, the state variables are not independent: a transition in a state variable may or may not be consistent with another transition in the previous or following state variables in the sequence.

For a given conformational state of the protein $s = \langle (x_1 = v_1), \ldots, (x_i = v_i) \ldots, , (x_n = v_n) \rangle$, the overlap between consecutive state variables means that a tripeptide x_i shares its first two residues with its predecessors in the sequence and its last two with its successors; that is:

$$\phi_{i,1} = \phi_{i-1,2} = \phi_{i-2,3}, \quad \phi_{i,2} = \phi_{i-1,3} = \phi_{i+1,1}, \text{ and } \phi_{i,3} = \phi_{i+1,2} = \phi_{i+2,1} \ , \quad (6.1)$$

and similarly for the ψ angles.

An elementary state transition with respect to x_i, from the value v_i to an adjacent value v'_i, involves a conformational change mainly in the central residue of x_i (by construction of the adjacency graph). This entails constraints on x_{i-1} and x_{i+1} with respect to their current values in state s. We express these constraints as inequalities with a tolerance margin as follows:

$$\begin{aligned} &|\phi'_{i,2} - \phi_{i-1,3}| < \varepsilon, \quad |\phi'_{i,2} - \phi_{i+1,1}| < \varepsilon, \\ &|\psi'_{i,2} - \psi_{i-1,3}| < \varepsilon, \quad |\psi'_{i,2} - \psi_{i+1,1}| < \varepsilon. \end{aligned} \quad (6.2)$$

where the angles for the last and first residues of x_{i-1} and x_{i+1} correspond to their current values v_{i-1} and v_{i+1}. These constraints can be relaxed during the search by dynamically adjusting the value of ε, as explained below. Here, we set initially $\varepsilon = 0.35$ radians.

Définition 1 (Feasible transition) *A transition in the conformation of a protein from a state s where $x_i = v_i$ to a state s' where $x_i = v'_i$ is said to be a feasible transition if and only if:*

(i) the values v_{i-1} and v_{i+1} meet the constraints of Equation 6.2, and

(ii) there are no collisions between the atoms of the protein in the state s'.
A feasible path is a sequence of feasible transitions.

Let $\gamma(s, (v_i \rightarrow v_i'))$ denotes the state s' corresponding to this transition when it is feasible, otherwise γ is undefined.

The conformation path finding problem can be formally stated as follows: given $\mathscr{X}$ and the adjacency graphs of all the state variables in a protein, and given an initial state s_0 and a goal state s_g, the problem is to find a feasible path that transforms the protein conformation from s_0 into s_g, if there exists such a path.

6.2.3 Search algorithm

To generate a feasible path from s_0 to s_g, we rely on a heuristically-guided depth-first search in the space $\prod_i D_i$, over all state variables x_i in the protein. To ease the presentation, the algorithm is stated in the pseudo-code of Figure 6.2 as a simple recursive nondeterministic search procedure called HDFS. The initial call is HDFS$(s_0, \langle s_0 \rangle)$. The *nondeterministic choice* (step labelled $\triangleleft$) is a convenient notation meaning that the algorithm makes at this point a branching decision; it explores potentially all possible options, expressed here as the set $\mathscr{E}$; it stops on the first path which succeeds or it returns failure if all paths fail.[1] The deterministic implementation of HDFS makes at this step a heuristic choice over which it backtracks in case of failure; if needed, this is repeated as long as an option in $\mathscr{E}$ remains unexplored. The heuristic driving this choice is detailed below.

The algorithm iterates over all tripeptides in the protein to find their feasible transitions. For a given state variable $x_i = v_i$ in s, procedure Transition-Filter checks the values adjacent to v_i in graph D_i. Unfeasible transitions are disregarded, as well as transitions that loop back into a circuit of the search space. The set $\mathscr{E}$ is the union of all retained transitions $(v_i \rightarrow v_i')$ over all state variables. When $\mathscr{E}$ is empty, then s is a dead end; a backtracking is performed.

In our more efficient and deterministic implementation of the algorithm, $\mathscr{E}$ is computed incrementally. $\mathscr{E}$ starts with the transitions of a single state variable, which has feasible transitions. $\mathscr{E}$ is augmented with respect to new state variables when backtracking requires alternative options. In our current code, the ordering of the state variables in the HDFS loop is not heuristically guided. The effects of state variable ordering heuristics, such as the proximity to the goal or the average density in the adjacency graph, remain to be investigated.

Heuristic guidance function For the results presented in this chapter, the search is guided though the ordering in procedure Transition-Filter of the set $\mathscr{A}$

[1] The metaphor to help explain a nondeterministic specification of an algorithm is that of a machine able to multiply itself at each branching point into identical copies, each copy pursuing the search in parallel until one finds a solution or all fail.

```
HDFS(s, Path)
  if s = s_g then return(Path · s)
  E ← ∅
  for each state variable x_i in s do
    E ← E ∪ Transition-Filter(s, x_i, Path)
  if E = ∅ then return(failure)
  else do
    Nondeterministically choose in E a transition (v_i → v'_i)        ◁
    s' ← γ(s, (v_i → v'_i))
    HDFS(s', Path · s)

Transition-Filter(s, x_i, Path)
  v_i ← value of x_i in s
  A ← set of values adjacent to v_i in adjacency graph D_i
  for each v'_i ∈ A do
    if γ(s, (v_i → v'_i)) is undefined or
    if it is a state already in Path
      then remove v'_i from A
  return(A)
```

Figure 6.2. Main procedure as a recursive nondeterministic best-first search. The choice (in step ◁) is guided with the heuristic *cost* function used to order the set $\mathscr{A}$. In the case of failure, backtracking is performed at this step to other remaining options in the set $\mathscr{E}$, which is computed incrementally.

of feasible values. $\mathscr{A}$ is ordered with the following cost function:

$$cost(v_i, v'_i) = d(v_i, v'_i) + w_1 \times h(v'_i, v^g_i) + w_2/\rho(v'_i),$$

where d and ρ are the distance and density functions defined earlier, v^g_i is the value of x_i in the goal state s_g, h is the shortest path in the transition graph to the goal, and w_1 and w_2 are weight parameters. The first term seeks to minimize the distance between consecutive states along the path (*i.e.*, to maximize the continuity of the path). The second term is the sum of the distances of a minimal path from v'_i to the goal. The third term intends to maximize the density of the states along the path, which, as explained earlier, are the most energetically favorable ones. The weights w_1 and w_2 permit a tuning of the three components; their proper setting remains to be investigated. Here, we simply set $w_1 = w_2 = 1$. Note that h is a lower bound for the remaining *cost* from v' to v^g, since a path in the transition graph, minimal with respect to the distance d, relaxes the feasibility constraints of Definition 1 and cannot be longer than a feasible path.

In order to speedup the search, a preprocessing of the adjacency graphs labels

edges with their distance d and computes for every vertex the shortest path to the goal as well as the density of every node in each graph. This is done with a standard graph search algorithm.

The test of collision-free states is computed using a variant of the classical Cell Linked-List (CLL) algorithm [189]. A pair of non-bonded (pseudo-)atoms is considered to be in collision if their distance is less than 65% of the sum of their radii. In this work, we considered the radii values proposed by Bondi [23] for the backbone atoms, and those proposed by Levitt [127] for the side chains pseudo-atoms.

Note that the feasibility constraints in Equation 6.2 are too conservative. A more flexible definition would also accept as feasible the transitions for which either the current values of x_{i-1} and x_{i+1}, or some of their respectively adjacent values v'_{i-1} and v'_{i+1}, meet these constraints. In that case, the state $s' = \gamma(s, (v_i \rightarrow v'_i))$ involves changes in x_i but also in its predecessor and successor state variables. The cost function driving the search would naturally be extended to $cost(s, s')$ over entire states. Instead, we have implemented a simpler mechanism to locally relax this constraint if the search process gets blocked : if state transitions fail f consecutive times ($f = 5$ in our implementation), the tolerance value ε is increased to 0.7 radians. ε is reset to 0.35 radians after a successful transition. The next section shows that, even with such a simplified implementation, the proposed approach already gives meaningful results.

Properties of HDFS The algorithm is *sound*; that is, it returns a path which is feasible, in case of success. This is because each transition meets Definition 1. HDFS is also *complete*; that is, it finds a feasible path if one exists with respect to the transitions in the adjacency graphs of the state variables. This is the case since in each search state s, $\mathscr{E}$ is the entire set of feasible transitions over all state variables, loops are avoided, and backtracking is systematic.

As for any backtrack search algorithm, the worst case complexity is exponential, in $O(\prod_i |D_i|)$.[2] A more useful complexity model is in $O(d^b)$, where d is the depth of the search (i.e., the length of the found path), and b is the branching factor. An upper bound on the branching factor is $n \times p$, where n is the length of the protein and p is the maximum degree of vertices over all adjacency graphs. However, thanks to the search guidance of its heuristics, we observed a manageable complexity growth. Our experiments with seven proteins, ranging in length $10 \leq n \leq 67$ residues, show that b does not grow with n; it is constant and very small, about $b \simeq 1.04$. The overall search complexity has a low polynomial growth in n. Furthermore, we confirmed that, as expected for a local propagation mechanism, the computation time required for each search state is not a function of n, but a quite small constant, of about 0.9 ms per state on a standard CPU. The Section S1 in the supplementary

[2] It is possible to compute the total size of the search space for each given problem (using Dynamic Programming and taking into account state variable dependencies); but this information is not very useful since in practice the algorithm explores a very small fraction of the search space.

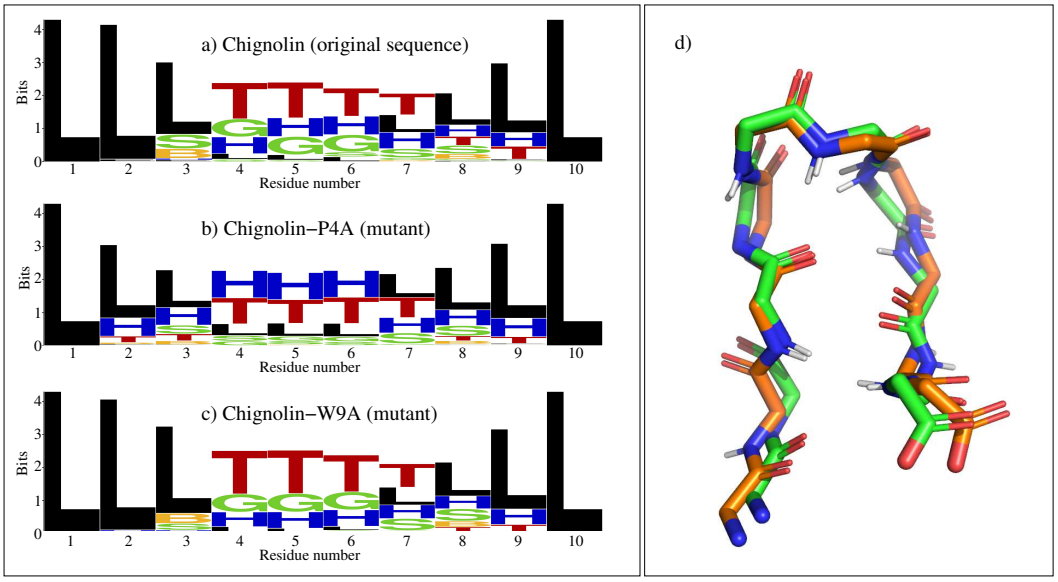


Figure 6.3. The left side panel represents the structural propensities at the residue level observed from a set of 1,000 conformations randomly generated from the structural database. Each plot displays the DSSP structural classes using the WebLogo format for (a) Chignolin, and two mutants: (b) Chignolin-P4A, and (c) Chignolin-W9A. (d) Structural representation of Chignolin: superposition of an experimentally determined structure (with carbon atoms in green) and the closest one in the set of 1,000 sampled conformations (with carbon atoms in orange). For clarity, only the protein backbone is represented, using PyMOL [193].

material details this analysis as well as a discussion contrasting the scalability of our approach with that of MD methods.

6.3 Results and Discussion

This section presents results obtained with the proposed approach for the analysis of the folding process of two synthetic mini-proteins, Chignolin and DS119, which were designed to fold into structural motifs present in natural proteins. First, we present a deeper analysis for Chignolin and two point mutants. Then, results presented for DS119 show that the approach is general and can be applied to the investigation of different structural elements.

6.3.1 Chignolin

Chignolin is a synthetic polypeptide consisting of 10 residues [85]. Despite its small size, Chignolin behaves as a macromolecular protein from structural and thermodynamic points of view: it folds into a well-defined structure in water, and shows a cooperative thermal transition between unfolded and folded states [191]. The folded conformation of Chignolin corresponds to a β-hairpin motif, which can

be found in many natural proteins (Figure 6.3.d). Therefore, elucidating the folding mechanism of Chignolin helps to understand the folding patterns of more complex proteins. This has motivated several experimental and computational studies on Chignolin in recent years. Here, we compare our results with those of Enemark et al. [61], which are based on extensive molecular dynamics simulations, and provide detailed information at the single-residue level.

Table 6.1 provides the number of conformations (*i.e.*, number of values of state variables) contained in our database for the eight overlapping tripeptides composing Chignolin. The search space size is upper-bounded by $\prod_i |D_i| \approx 4 \times 10^{23}$, which is huge when compared to the extremely focused explorations performed by our algorithm. Thanks to the search guidance of its heuristics, we observed a manageable complexity growth, as explained in Section 6.2.3 and in the supplementary material.

In a first experiment, we assessed the ability to obtain realistic conformations of Chignolin using the structural information encoded in our tripeptide database. We generated an ensemble of $1,000$ Chignolin states by randomly sampling values of the state variables one by one, in an incremental manner, enforcing the consistency with neighbor state variables, and rejecting those leading to collisions between atoms. Interestingly, several states in this relatively small ensemble are close to the folded conformation of Chignolin [85]. Indeed, 240 over the $1,000$ sampled states have an angular RMSD distance to the folded conformation below 0.5 radian, the closest one being around 0.2 radians (see Figure 6.3.d). This confirms that the most important regions of the conformational space can be sampled by building states from the tripeptide database.

In order to better characterize the conformational ensemble, secondary structure types for each state were identified at the single residue level using DSSP [105]. DSSP distinguishes eight types of structural classes, labeled with a letter: H for α-helix, B for β-bridge, E for strand, G for helix-3, I for helix-5, T for turn, S for bend, and "blank" (here labeled as L) for coil/loop. We used the WebLogo tool [37]

Table 6.1: Number of conformations (i.e. number of values of state variables) for the eight overlapping tripeptides composing Chignolin.

Tripeptide sequence	Nb conformations
Gly-Tyr-Asp	994
Tyr-Asp-Pro	710
Asp-Pro-Glu	1541
Pro-Glu-Thr	1030
Glu-Thr-Gly	1446
Thr-Gly-Thr	1779
Gly-Thr-Trp	545
Thr-Trp-Gly	240

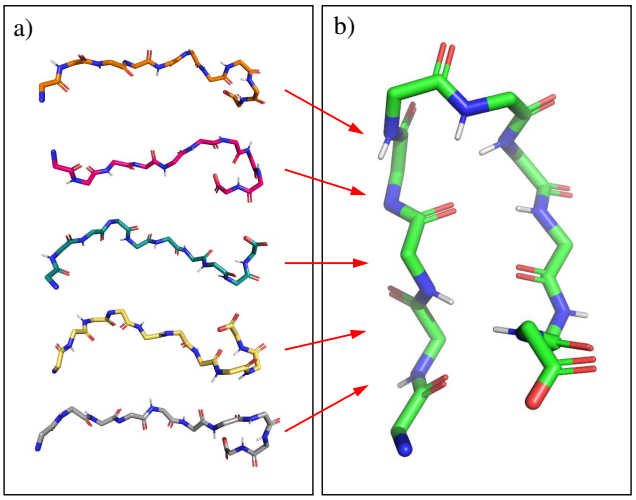


Figure 6.4. Structural representation of Chignolin. (a) A set of extended conformations involving the initial turn at the C-terminal side. (b) Folded conformation. Only the protein backbone is represented, using PyMOL [193].

to display the structural propensities in the ensemble. WebLogo is usually applied to analyze results of multiple sequence alignment, but it can be used in a different context, as we did. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the conservation of the DSSP structural class at that position, while the height of symbols within the stack indicates the relative frequency of each class at that position. The results in Figure 6.3.a clearly show the propensity of the central residues to adopt a turn conformation. The rest of the molecule tends to be more extended, although turns are also formed in the C-terminal region. As discussed in detail below, these turns in residues 8 and 9 play a key role in the folding mechanism of Chignolin. Conversely, turns are not observed in the N-terminal side. These observations are fully consistent with the original study [61], and show that the states sampled using the tripeptide database are structurally relevant.

We repeated the experiment for two mutants of Chignolin: Chignolin-P4A (Pro4 replaced by Ala) and Chignolin-W9A (Trp9 replaced by Ala). Figure 6.3.b shows that, for Chignolin-P4A, the turn propensity slightly decreases in the central region, and that it increases in the N-terminal side. For Chignolin-W9A, Figure 6.3.c shows that the propensity to form turns in the central region is similar to that of the native Chignolin molecule. However, it decreases in the C-terminal region, which may have consequences for the efficiency of the folding process. Overall, these observations are very similar to the results reported in [61], which use computationally expensive molecular dynamics simulations; they show the strong influence of single modifications in the sequence on the conformational preferences of the molecule, and that our approach captures these perturbations.

It has been suggested that the turn in Chignolin originates in the C-terminal region, and then propagates along the chain until reaching the middle residues [61].

This has been called the "roll-up" mechanism. To investigate this mechanism, we selected (among the set of 1,000 conformations) 15 conformations of Chignolin presenting turns in residues 8 and 9, and with a relatively extended conformation for the rest of the chain. These conformations were used as initial states to compute folding paths, as illustrated in Figure 6.4. The goal state was defined as the closest conformation to the experimental structure of Chignolin built from values contained in the tripeptide database. These two conformations are very similar, with an angular RMSD of 0.1 radians. The HDFS algorithm was applied 20 times to solve each of these 15 problems (i.e. 300 runs in total). On average, the algorithm required around 10 seconds to find folding pathways (1^{st} column in Table 6.2), which is extremely fast.[3] Intermediate states along each path were selected with a step-size corresponding to $1/10^{th}$ of its total length. The left side panel in Figure 6.5 shows the structural propensities at the residue level for these intermediate states. It can be observed that the turns in the C-terminal residues tend to disappear, while these structural elements appear in the middle residues. This "roll-up" mechanism can also be observed in the right side panel in Figure 6.5, which represents several intermediate states along one of the folding paths. The first frames (starting from the top) show that the curvature of the molecule, initially involving residues 8 and 9, rapidly propagates to residues 6 and 7. Then, residues 5 and 4 also bend successively, and the molecule tends to form a hairpin-like structure. Finally, the two terminal parts adopt a relatively extended conformation.

As explained in related work [191], the folding process of Chignolin may lead to misfolded states, which are characterized by interactions between residue pairs Tyr2-Thr8 and Asp3-Gly7, rather than Tyr2-Trp9 and Asp3-Thr8, as in the correctly folded structure. We generated a representative model of a misfolded state, and we computed conformational transitions from initial conformations with the C-terminal turn (C-ter T) to this state. We also computed transitions from fully-extended conformations to folded and misfolded states. The results are summarized in the top part of Table 6.2. This table provides average values (over 300 runs) for: the computing time required by the HDFS algorithm to find a path; the number of recursions and backtracks; the number of steps in the solution path; the length of the solution path, computed as the sum of the lengths associated to edges in the adjacency graphs; the density of the solution paths, computed as the average of the density of all the state variables along the path. The most meaningful numbers in this table are those associated with the density, since they reflect the probability of existence of each pathway. Compared to the extended→folded pathway, the C-ter T→folded pathway goes across more dense and probable regions. This may explain why Chignolin efficiently folds from unfolded states involving this structural feature. In both cases, starting from C-ter T or fully-extended states, the transitions to misfolded states seem to be much less probable. This may explain why the misfolded state of Chignolin is much less frequently observed than the correctly folded state [120].

[3]CPU time was measured with an Intel® Core™ i7 processor at 2.8 GHz, using a single core.

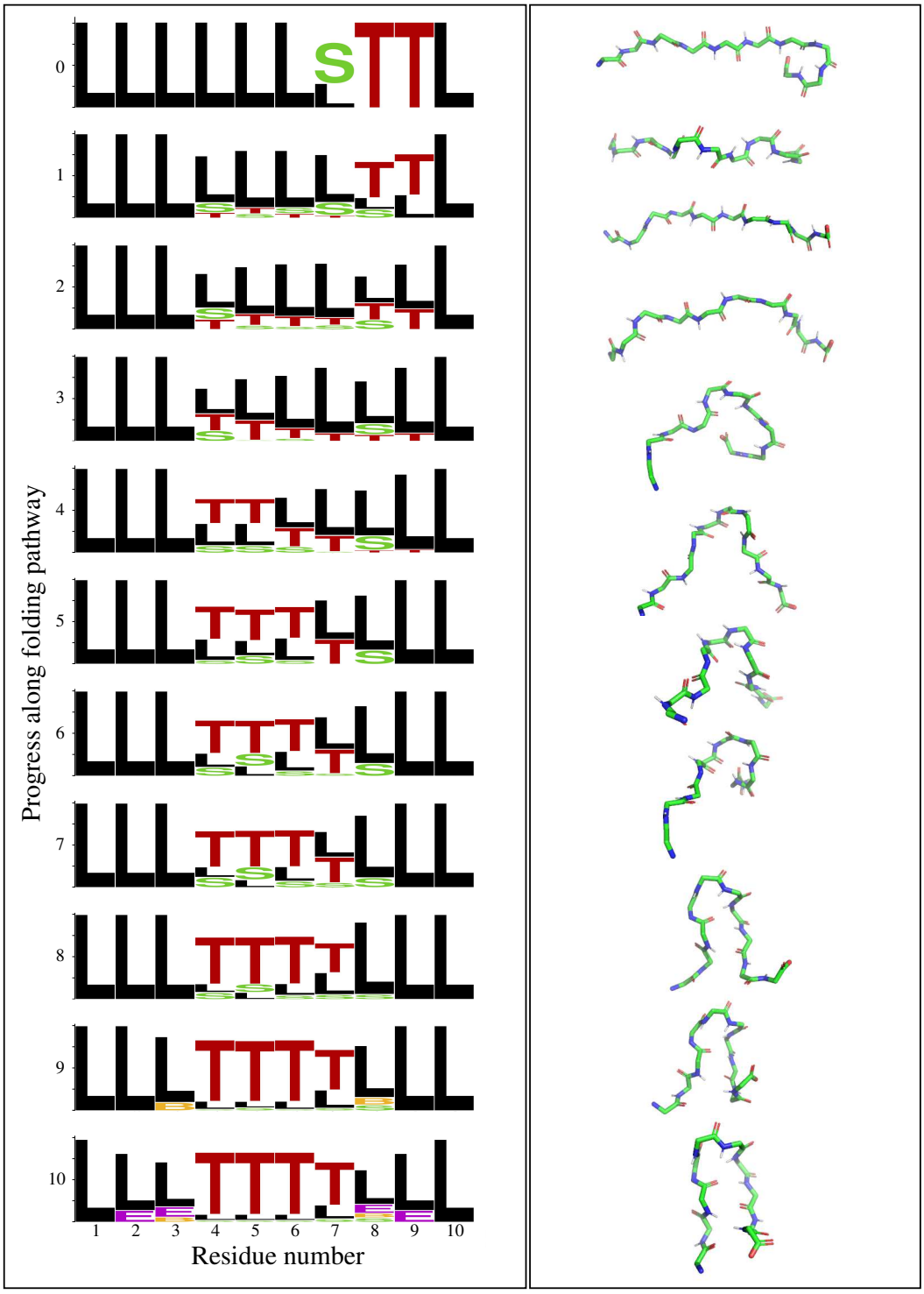


Figure 6.5. The left side panel represents the evolution of the structural propensities at the residue level along Chignolin folding pathway (see Figure 6.3 and the associated comments for additional explanations about this representation). The right side panel shows some intermediate states along one of the computed folding paths. Only the protein backbone is represented, using PyMOL [193].

Table 6.2: Performance indicators of the HDFS algorithm to compute different conformational transitions of Chignolin (top) and the mutant Chignolin-W9A (bottom). CPU time was measured with an Intel® Core™ i7 processor at 2.8 GHz, using a single core.

	chignolin (original sequence)			
	C-ter T→folded	C-ter T→misfolded	extended→folded	extended→misfolded
CPU time (s)	11.1	8.7	5.2	3.5
# states	5416.4	2587.7	2800.1	849.5
# backtracks	234.6	136.6	124.6	39.2
Path length (# steps)	133.8	54.5	106.3	48.7
Path distance (rad)	8.8	5.1	6.0	7.0
Path density	**31.9**	5.5	23.3	**4.5**

	chignolin-W9A (mutant)			
	C-ter T→folded	C-ter T→misfolded	extended→folded	extended→misfolded
CPU time (s)	12.2	8.8	5.6	5.1
# states	4943.6	2567.8	2317.0	2946.0
# backtracks	219.6	139.0	101.3	126.3
Path length (# steps)	140.3	51.3	103.0	125.7
Path distance (rad)	8.2	9.0	5.8	8.2
Path density	**31.2**	4.6	23.4	**23.8**

We repeated the experiments for the mutant Chignolin-W9A. The results are summarized in the bottom part of Table 6.2. As mentioned above, the set of conformations generated for these two molecules look structurally similar (see Figure 6.3 and the associated comments). The figures in Table 6.2 also show a very similar behavior of the HDFS algorithm when computing transition paths for this mutant compared to the original Chignolin. Interestingly, the main difference is observed for the density of the path extended→misfolded. This path is significantly more favorable in the case of the mutant. Our results complement the study of Enemark et al. [61], which suggested that the replacement of Trp9 by Ala facilitates a "roll-back" mechanism, acting against the "roll-up" mechanism, hindering the formation of the native turn in the middle residues. We show another possible effect of this mutation, favoring the formation of misfolded states in competition with the native structure.

6.3.2 DS119

DS119 is another synthetic polypeptide, consisting of 36 amino acid residues, which was designed to fold into a $\beta\alpha\beta$ motif [129] (see last frame in Figure 6.6). The folding process of DS119 has been studied using molecular dynamics simulations [178]. This previous work showed that the N-terminal side of the central helix tends to form very quickly. Then, the C-terminal side of the helix starts to form, and the full helix is finally stabilized. The relatively extended fragments at the two ends of the molecule tend to come together at the end of the folding process.

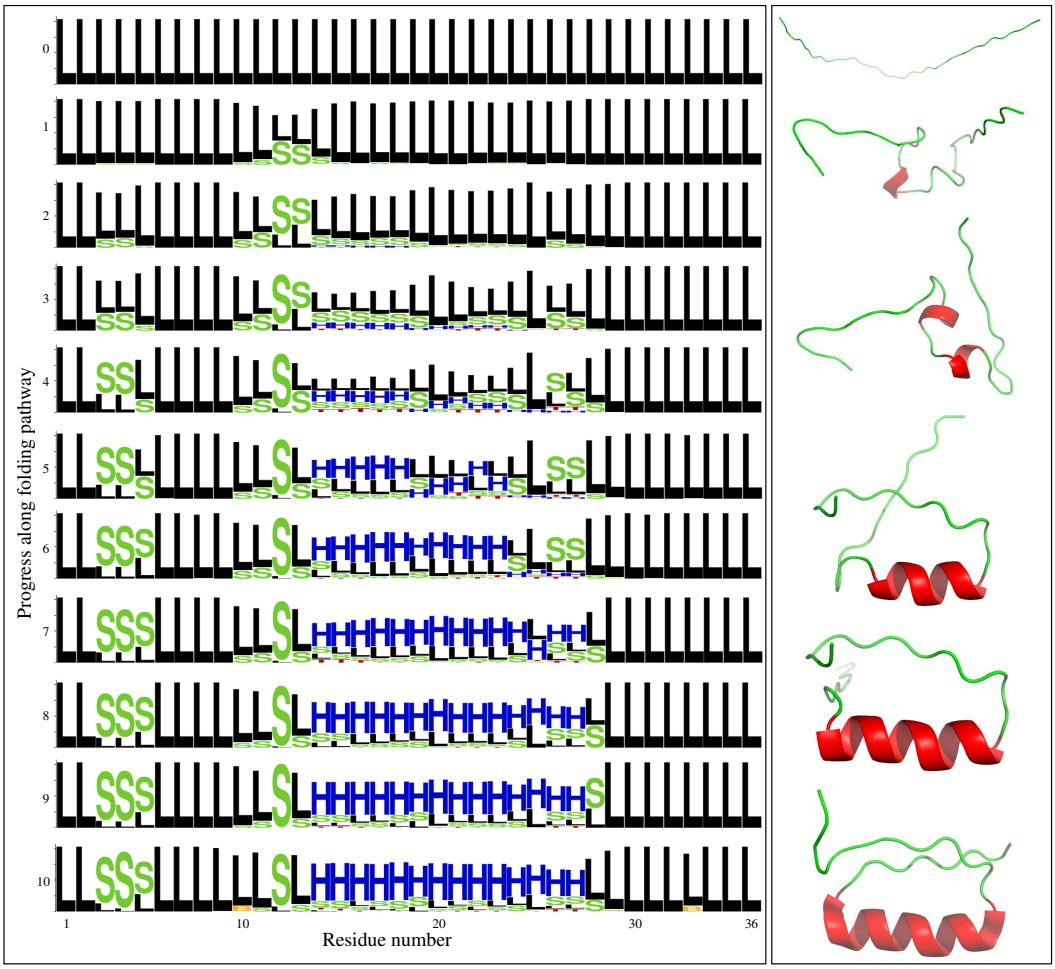


Figure 6.6. The left side panel represents the evolution of the structural propensities at the residue level along DS119 folding pathway. The right side panel shows some intermediate states along one of the computed folding paths. The "cartoon" representation clearly shows the formation of the helix. PyMOL [193] was used for the structural visualization.

To investigate the folding mechanism of DS119, we applied a similar procedure as for Chignolin. In this case, we selected 15 relatively extended conformations, involving only the L DSSP structural class for all the residues, from a set of $1,000$ randomly generated conformations using the tripeptide database. These conformations were used as initial states for the HDFS algorithm. As final state, we used the closest conformation to the experimentally solved structure of DS119 (PDB ID: 2KI0) built from values contained in the tripeptide database. These two conformations are very similar, with an angular RMSD of 0.06 rad. The algorithm was applied 20 times to solve each of these 15 problems (i.e. 300 runs in total).

Figure 6.6 illustrates the results obtained by the HDFS algorithm. The left side panel shows the evolution of the structural propensities along the folding path, using logos based on DSSP classes. The right side panel represents several intermediate states along one of the solution paths. For clarity purposes, only a few intermediate

Table 6.3: Performance indicators of the HDFS algorithm on DS119.

	DS119 : extended→folded
CPU time (s)	25.2
# states	70558.2
# backtracks	8210.4
Path length (# steps)	158.2
Path distance (rad)	11.3
Path density	124.4

states are shown using a "cartoon" representation of the backbone, where the helical fragments can be easily identified. It can be observed that, starting from an extended conformation, the protein backbone rapidly starts to bend around residues 12-13. Recall that the S letter, for "bend", corresponds to a highly curved protein backbone. Hydrogen bonds required to stabilize the helical conformation are not yet identified by DSSP at this early stage. Next, curved/helical fragments start to appear in all central residues (from residue 14 until residue 27), as well as in three residues in the N-terminal side (residues 3-5). The central helix continues to fold, and it is almost completely formed at the 7^{th} intermediate frame. In the final part of the path, the extended fragments at the two ends get close to each other, nearly forming a parallel β-sheet. This description of the folding process strongly resembles the one reported in the literature, based on computationally-expensive simulations [178].

Table 6.3 presents numbers (averaged over the 300 runs) concerning the performance of the HDFS algorithm to compute folding paths of DS119. The required CPU time (and the number of recursions) is only about three times the one requited to compute folding paths for Chignolin. This shows that, despite the theoretical (worst-case) exponential complexity, in practice, the computing time scales approximately linearly with the number of variables. This tendency has been confirmed by preliminary tests for larger molecules (not presented in this chapter). Once again, we insist that computing time is orders of magnitude faster that traditional molecular dynamics simulation methods. The higher density of the path compared to Chignolin can be explained by the higher number of conformations for some of the tripeptides, particularly for those composing the middle helix. Table 6.4 provides the numbers of conformations (*i.e.*, number of values of state variables) contained in our database for the 34 overlapping tripeptides composing DS119.

Scalability analysis

We applied the proposed method to five other proteins with increasing size, from 20 to 67 amino acid residues: Trp-cage, WW-domain, BBL, CENP-B, and Villin (whose respective PDB IDs are: 1L2Y, 1QQV, 1E0M, 2WXC, and 1BW6). The folded states of the proteins are represented in Figure S6.7. The HDFS algorithm

Table 6.4: Number of conformations (i.e. number of values of state variables) for the eight overlapping tripeptides composing DS119.

Tripeptide sequence	Nb conformations	Tripeptide sequence	Nb conformations
Gly-Ser-Gly	3727	Lys-Lys-Leu	2286
Ser-Gly-Gln	1118	Lys-Leu-Lys	1996
Gly-Gln-Val	1294	Leu-Lys-Glu	3100
Gln-Val-Arg	607	Leu-Glu-Glu	1631
Val-Arg-Thr	970	Glu-Glu-Ala	2591
Arg-Thr-Ile	757	Glu-Ala-Lys	1514
Thr-Ile-Trp	181	Ala-Lys-Lys	1714
Ile-Trp-Val	180	Lys-Lys-Ala	1629
Trp-Val-Gly	279	Lys-Ala-Asn	1009
Val-Gly-Gly	2443	Ala-Asn-Ile	1010
Gly-Gly-Thr	2510	Asn-Ile-Arg	647
Gly-Thr-Pro	1428	Ile-Arg-Val	998
Thr-Pro-Glu	1738	Arg-Val-Thr	1351
Pro-Glu-Glu	1752	Val-Thr-Phe	888
Glu-Glu-Leu	3433	Thr-Phe-Trp	151
Glu-Leu-Lys	2378	Phe-Trp-Gly	192
Leu-Lys-Lys	2528	Trp-Gly-Asp	257

was applied to find paths between a fully-extended all-trans conformation and the folded state of each of these proteins. These experiments were not aimed to provide insights into the folding mechanisms of these proteins, but only to analyze the scalability of our method.

The performance indicators of the algorithm are summarized in Table S6.5, which gives the average values over 5 runs for these additional five proteins, as well as the values presented in this chapter for Chignolin and DS119. We denote n the length of the protein, t the time it takes to find a path, m the number of states explored by the search, and d the depth of the search, i.e., the length of the found path. The analysis of our result can be summarized in the following points:

- As expected for a local propagation mechanism, the complexity of each search step is not a function of n. This is clearly shown by the ratio t/m which does not increase with n; its average is about 0.94 ms per search step. In comparison, each simulation step with usual MD approaches has a complexity in $O(n^2)$.

- As a backtrack search algorithm, HDFS is exponential with respect to d, the depth of the search, but not with respect to n, the size of the protein. The

number of search states m grows as $m = d^b$, where b is the *branching factor*. Thanks to the heuristics guidance of the search, in our case b is very small, in average $b \simeq 1.03$. Again, b is not a function of n (we even observe smaller values of b for the larger proteins than for the smaller ones), but d grows with n.

- The overall complexity, in time or in the number of steps, increases with n, but with a quite reasonable polynomial growth, as illustrated with the three parameters α_1, α_2, and α_3 in Table 6.5. Their average values provide the following approximate growth: $t = n^{1.3}$, $m = n^{3.4}$, or $m = K \times n^{1.4}$, for K=1000 (this last function is more adequate given the constant value of t/m of about 1s for 1000 states). Note again that a simulation with MD would involve a number of steps growing with d (hence indirectly with n), each step being quadratic in n.

- As for any heuristics search algorithm, the performance figures are not smooth. Much more data would be needed to support precise average complexity models. However the above results give the main trends for the scalability of the approach: a quasi-constant t/m, a very small branching factor b, and a reasonable polynomial growth of the global complexity in the size of the protein. Clearly, the approach is scalable: for the largest system in our test set, Villin (with $n = 67$), the search algorithm explores about 4.5×10^6 states, requiring 35' of a single core standard processor. In contrast, results reported in the literature, reference [132], indicate that in order to find folding pathways for a fast-folding protein such as Villin, MD simulations would require in the order of 10^9 steps, each of which being of quadratic complexity in n.

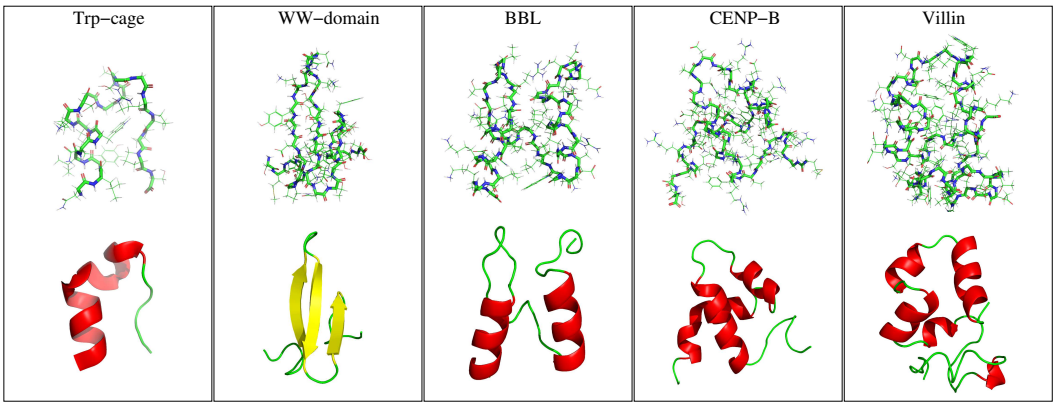


Figure 6.7. Structural representation of five proteins with increasing size used to analyze the scalability of the algorithm (in addition to Chignolin and DS119). The images at the top are detailed representations, in which thinker lines correspond to the protein backbone and thinner lines are used for the side chains. The images at the bottom are "cartoon" representations that highlight the main structural elements: α-helices in red, and β-sheets in yellow.

The results also show that the performance of the method depends on the structural elements in the protein. This can be clearly illustrated with the WW-domain. The folded structure of this protein is mainly composed of β-sheets, whereas the other four proteins mainly involve α-helices. Since the backbone of proteins has a natural propensity to twist, helical fragments are much more frequent that extended fragments, which lead to the formation of β-sheets. This explains the lower density of the states along the folding pathway for the WW-domain compared with the other four proteins. On the other hand, the presence of β-sheets in the folded structure significantly facilitates the search of folding paths from extended conformations, since these structural elements already correspond to extended fragments. This explains why the algorithm is faster on the WW-domain compared with the other proteins.

Table 6.5: Performance indicators of the HDFS algorithm on seven proteins with increasing size. CPU time was measured on an Intel® Core™ i7 processor at 2.8 GHz, using a single core. The last four parameters are defined as follow: $b = e^{(\log m)/d}$, $\alpha_1 = \log t/\log n$, $\alpha_2 = \log m/\log n$, and $\alpha_3 = (\log m - \log K)/\log n$. The average values are $t/m = 0.95$, $b = 1.03$, $\alpha_1 = 1.34$, $\alpha_2 = 3.46$, and $\alpha_3 = 1.42$.

	Chignolin	Trp-cage	DS119	WW-domain	BBL	CENP-B	Villin
$n =$ Nb residues	**10**	**20**	**36**	**37**	**47**	**56**	**67**
$t =$ CPU time (s)	**7.5**	**158.3**	**25.2**	**8.9**	**735**	**1096.5**	**2182.3**
$m =$ # states ($\times 10^3$)	3.1	235.1	70.6	30.6	1730.8	549.3	4507.7
$d =$ Path length (# steps)	96	508.0	158	165.5	2024.8	5041.6	3241.3
# backtracks ($\times 10^3$)	0.1	3.8	8.2	0.3	15.4	1.9	26.5
Path distance (rad) Å	7.3	16.0	11.3	9.4	33.3	139.7	45.7
Path density Å	18.5	39.0	124.4	6.3	147.9	52.4	45.8
t/m (in ms)	2.42	0.67	0.36	0.29	0.42	1.99	0.48
b	1.08	1.024	1.07	1.06	1.007	1.002	1.004
α_1	0.87	1.69	0.90	0.60	1.71	1.73	1.83
α_2	3.49	4.13	3.11	2.86	3.73	3.28	3.64
α_3	0.49	1.82	1.19	0.94	1.93	1.56	2.0

6.4 Conclusion

Despite the simplicity of both the algorithm and the heuristic, the results presented in this chapter show that the proposed approach constitutes a promising new research direction towards the identification of relevant protein folding pathways. The structural analysis of the folding mechanisms of Chignolin and DS119 are consistent with respect to descriptions provided in the literature. Note however that a more detailed and quantitative comparison between the paths obtained with other methods and trajectories obtained from MD simulations would not be very meaningful, since the aims of both methods are different: The paths provided by our algorithm are an approximation, from which interesting information about folding mechanisms can already be obtained, but that should be refined (using other methods and models) to get access to accurate information at the atomic level (as provided by MD simulations). On the other hand, our algorithm is orders of magnitude faster than atomistic MD simulations.

Overall, the results highlight the importance of local structural preferences, which are encoded in our tripeptide database. They also suggest that interactions between distant residues in the sequence, even though they can be essential for stabilization of the final fold, are less important at an earlier stage to drive the formation of structural elements.

The good results obtained with the implementation presented in this chapter motivate us to continue in this research direction. Several points remain to be further investigated. One important question is about the possibility to include non-local interactions in the heuristic cost function. Although this does not seem to be necessary for structural elements or small proteins, interactions between distant residues in the sequence can be essential to study folding processes of larger molecules, or aspects related to stability. We also plan to implement and evaluate transitions over several state variables, as well as different heuristics for variable ordering. More sophisticated, tree-based search algorithms [70] can improve the quality and the diversity of the solutions, particularly for large proteins. Finally, let us mention the limitations imposed by the information contained in the structural database. Structural information is very limited in some regions of the conformational space corresponding to states of low probability, but which may be relevant for an accurate modeling of conformational transitions. With the increasing number of experimentally-determined high-resolution protein structures, we expect that more extensive and higher-quality tripeptide databases will be constructed in the future. Alternatively, these sparsely populated transition regions can be identified using our approach and subsequently explored using physics-based molecular models and (continuous) motion planning algorithms [48].

In this work, we have used two well-characterized synthetic mini-proteins to evaluate the performance of the proposed algorithm. Nevertheless, our main interest in the future is to apply this method for the investigation of MOREs in IDPs.

Chapter 7

Hybrid-multiTRRT algorithm to explore the energy landscape

Contents

7.1 Introduction

In Chapter 6, we presented a heuristic method to compute conformational transition paths using structural information in the tripeptide database (presented in Chapter 3). Here, we present a more general approach to explore the conformational space of highly-flexible molecules such as IDPs. Starting from a set of conformations, the method globally explores the conformational space aiming to find the most likely transition paths between them. In the case of IDPs, this initial set of conformations can be generated using the method presented in Chapter 5. To perform the exploration in a very short time, we have developed a parallel version of

an efficient algorithm called MultiTRRT, originating from robotics. The resulting set of conformations and transition paths provides a simplified representation of the energy landscape.

In recent years, algorithms originally developed for path planning in robotics have been proposed as an alternative to classic approaches [2, 71, 199]. This work focuses on one of these algorithms, the Transition-based Rapidly-exploring Random Tree (TRRT) [94], which performs a randomized exploration of the conformational space aiming at finding probable transition paths between stable states of a molecule. More precisely, we present a parallel implementation of a multi-tree variant of TRRT [49, 48], which we will call Multi-TRRT hereafter (the principle TRRT and Multi-TRRT will be reminded in Section 7.3). TRRT is based on the Rapidly-exploring Random Tree (RRT) algorithm [122], a popular path-planning algorithm that can tackle complex problems in high-dimensional spaces. Although path-planning problems in robotics are not very computationally expensive in general, several approaches have been proposed for the parallelisation of RRT-like algorithms aiming at reducing computing time. A brief survey on parallel path planning algorithms will be provided in Section 7.2.2.

Today, the majority of high-performance computing (HPC) systems are clusters of multi-core processors [1]. To take advantage of this architecture, we propose a hybrid parallelization strategy of the Multi-TRRT algorithm (Section 7.4): shared memory parallelization (OpenMP) within each multi-core processor and distributed memory parallelization (MPI) between processors. Such a hybrid strategy is well suited to the Multi-TRRT algorithm, since extension operations of each tree can be performed efficiently inside each processor using multi-threading, and several trees can be constructed almost independently from each other (with few communication requirements).

The performance of the Hybrid Multi-TRRT algorithm is evaluated using several molecules of different sizes. Indeed, the size of the molecule is directly related to the complexity of the problem. Results presented in Section 7.5 show a very good performance of the parallel implementation, which can provide near-linear speedup for large molecules.

7.2 Background on parallel computing

Modeling molecular systems is computationally intensive. This has motivated numerous initiatives to introduce parallelization within this domain. This section first provides a brief survey on the parallelization of molecular dynamics simulation methods, which are the most frequently used computational techniques to study biomolecules. Then, more closely related to our contribution, we present a concise review of previous work on parallel path planning algorithms originating from robotics.

[1] To avoid ambiguity between *nodes* of the exploration tree and *nodes* of the computer cluster, in this chapter we will refer to each component of a cluster as a *computer* or as a *processor*.

7.2.1 Parallel molecular simulation methods

As mentioned in Chapter 2, Molecular Dynamics (MD) [107] is the most widely-used technique for studying molecular movements. This method explores the conformational space of a molecular system by numerically solving the Newton's equations of motion. In MD, roughly 90% of the computing time is spent in the calculation of the forces between non-bonded atoms. Therefore, efforts have been focused on the parallelization of this part of the algorithm aiming to speed-up MD simulations since the 1980s [157]. Nowadays, OpenMP (for shared memory architectures) and MPI (for distributed memory architectures) are widely used for parallelizing MD algorithms, and are included within many software packages such as GROMACS [130] and NAMD [172].

More sophisticated variants of MD methods enable parallelization at a higher level. Replica Exchange Molecular Dynamics (REMD) [216] is probably the most clear representative of these advanced methods. REMD consists of running n instances of the same molecular dynamics problem with different parameter settings across the different replicas. After several iterations, the configurations of different replicas are exchanged if an stochastic transition test succeeds. The n simulations can be run in parallel, only requiring communication for replica exchange. REMD has been proved to sample the conformational space much more efficiently than a basic MD thanks to its ability to scape from local minima traps [180].

As an alternative to MD, Monte Carlo (MC) methods sample the conformational space of a molecular system by generating states according to a Boltzmann distribution [68]. Although the stochastic sampling process performed by a basic MC method is sequential, several parallelization strategies have been developed [214]. The most simple approach is to run multiple independent MC in parallel [187], without any communication requirement, aiming to provide a more exhaustive sampling of the conformational space. However, more sophisticated parallelization strategies can be more efficient [214, 74]. In particular, the method proposed by Strid [214] improves the sequential MC by obtaining several draws from the posterior distribution doing multiple evaluations in parallel. Another advanced variant, the Parallel Tempering Monte Carlo scheme, similar to REMD, enhances sampling by performing exchanges between replicas running in parallel [220].

7.2.2 Parallel path planning algorithms

Path planning algorithms have been developed to automatically compute robot movements [121]. In the last decades, sampling-based algorithms [123] have become very popular thanks to their efficiency, generality and conceptual simplicity. In addition to robotics, they have been applied to problems in other areas such as computational biology [2, 71, 199]. The basic principle of these algorithms is to construct a graph or a tree that captures the topology of the admissible (e.g. collision-free or energetically-feasible) regions of the search-scape by randomly sampling configurations and attempting local connections. The Multi-TRRT algorithm

addressed in this work belongs to this family of methods.

When applied to simple robot systems, path planning algorithms are computationally fast, and parallelized implementations are not required. However, parallel computing may provide significant performance gain when dealing with complex systems. Starting from seminal work on the parallelization of classical path-planning algorithms [82], most efforts have been focused on shared-memory parallelization strategies. This is particularly relevant when using current multi-core central processing units (CPUs) [90] or many-core graphics processing units (GPUs) [21, 166]. Several approaches have also been proposed for shared-memory systems with the aim to enable a more general and large-scale parallelization (e.g. [4, 175, 51]). In this work, we propose to combine both shared-memory and distributed-memory strategies. To the best of our knowledge, this is the first work proposing such a hybrid approach in the context of path-planning algorithms.

As for molecular simulation methods, parallelization can be done at different levels of the algorithm. In the case of path planning algorithms for robotics applications, collision detection is the most computationally expensive operation, and parallelization can be focused in this part [21, 166]. Nevertheless, the algorithms can also be parallelized at a higher level using diverse strategies, as explained below. This chapter focus on high-level parallelization of the Multi-TRRT algorithm, but combining it with lower-level parallelization for collision detection and/or energy evaluation could be an interesting direction for future work.

The best parallelization strategy to be adopted depends on the characteristics of each path-planning algorithm. Some algorithm present an inherent parallelism whereby they can be easily subdivided into *embarrassingly parallel* processes. This is the case of the Probabilistic Road-Map (PRM) algorithm [108]. Using a basic parallelization strategy, based on a shared-memory programming paradigm, significant speed-up can be achieved by building the roadmap cooperatively across processes [4]. Other algorithms, such as RRT and its variants, are more difficult to parallelize due to the intrinsic sequentiality of the tree construction process.

The simplest parallelization strategy, known as the *OR Parallel paradigm*, can be applied to all types of radomized algorithms. The idea is to run in parallel several independent instances of the same sequential processes using different seeds for the initialization of the random sampling process. The first instance to reach the solution reports it and the other processes stop. This method reduces the execution time by multiplying the chances to finding a solution rapidly. It can significantly speed-up the resolution of problems that have a huge variability on running time. The strategy has been applied successfully to randomized algorithms using distributed-memory architecture [27, 26] and in shared-memory machines [25, 1].

A recent work [51] presents and compares the performance of three distributed-memory parallel version of RRT: OR parallel RRT, Manager-Worker RRT and Distributed RRT. The Manager-Worker RRT approach uses a single processor to manage the tree construction, whereas the other processors take in charge the calculation of the most computationally expensive part of the tree expansion. Thus,

a single copy of the tree is maintained in the memory of the master processor. In the Distributed RRT strategy, all the processes build together the same tree. Each processor needs to update its own copy with information provided by the others. Results presented in the referred paper show that the Manager-Worker RRT and Distributed RRT present a good speed-up when solving constrained problems. However, both approaches present drawbacks that hinder a good scalability using a large number of processors. For the Manager-Worker RRT approach, the manager process can rapidly become a bottleneck. For the Distributed RRT strategy, communication time increases with the number of processors, and can rapidly become a detriment of performance. Several strategies can be applied to improve the performance of simple distributed-memory approaches mentioned above, such as the subdivision of the search-space among processes and the implementation of efficient nearest neighbor search methods [92, 90].

More intricate strategies can be applied for the parallelization of algorithms that simultaneously construct several exploration trees. For instance, a *master-client scheme* was proposed to solve large-scale problems using a forest of random trees [175]. In a first stage, all the processes are run to build in parallel a number of trees that can be RRTs [122, 124] or Expansive Space Trees (ESTs) [87]. Then, a scheduler-processor distributes work to link the trees among client processes. Another strategy consist of exchanging information between several processors running instances of the same problem (each processor builds its own tree) with different random seeds, aiming to find the shortest path more efficiently [161].

The work presented in this section shares ideas with some of the aforementioned approaches. As in [175], the construction of several (independent) trees is parallelized. As in [92, 90], the space is subdivided to improve computational efficiency. Note however that, in our approach, space subdivision is implicit and evolves during execution, compared to an explicit and constant subdivision proposed in related work.

7.3 The Multi-TRRT algorithm

This section presents the main principles of the Multi-TRRT algorithm [49, 48] that we have parallelized. The pseudo-code of the overall algorithm is presented in Algorithm 2. It incorporates the parallelization explained in next section. In all the pseudo-code presented in this chapter, brown lines correspond to OpenMP commands for shared-memory parallelization, and blue lines indicate functions involving MPI calls for distributed-memory parallelization. Red text is used to highlight parts of the code involving only the master processor.

Multi-TRRT is a multiple-tree variant of the TRRT algorithm [95, 94]. It explores a continuous cost space $\mathscr{C}$ (i.e. the conformational space of a molecule, in the present application context) by iteratively expanding several trees rooted at a given set of m initial configurations $Q_{\text{init}} = \{q^1_{\text{init}} \dots q^m_{\text{init}}\}$.

At each iteration, a tree $\mathscr{T}_i$ is selected for expansion (line 4 in Algorithm 2).

Algorithm 2: Parallel Multi-TRRT

input : the configuration space $\mathscr{C}$; the extension step-size δ;
the energy function $E : \mathscr{C} \to \mathbb{R}$; the set of initial configurations Q_{init}
output: the tree(s) $\mathscr{T}$

```
1  (T, S) ← initProcesses(Q_init)
2  #pragma omp parallel (NumThreads)
3  while not stoppingCriteria () do
4      T_i ← chooseNextTreeToExpand()
5      q_rand ← sampleRandomConf(S_i)
6      q^i_near ← findNearestNeighbor(T_i, q_rand)
7      q_new ← extend(q^i_near, q_rand, E, δ)
8      if q_new ≠ null and
9      transitionTest(T_i, E(q^i_near), E(q_new)) then
10         #pragma omp critical (addNodeAndEdge)
11         addNewNodeAndEdge(T_i, q^i_near, q_new)
12         (T_j, q^j_near) ← findNearestNeighbor(T, q_new)
13         if distance(q_new, q^j_near) ≤ δ then
14             T_i ← merge(T_i, T_j, q_new, q^j_near)
15             #pragma omp critical (Communication)
16             MPI_send(Master, q_new, i, j)
17     updateStructures(q_new, S_i)
18 return(T)
```

Following the principle of TRRT, the tree expansion involves three main stages, corresponding to lines 5-7 in Algorithm 2: 1) A configuration q_{rand} is randomly sampled[2]. 2) The nearest neighbor $q^i_{\text{near}} \in \mathscr{T}_i$ to q_{rand} is selected for expansion. 3) A new node q_{new} (and its corresponding edge) is created by extending/perturbing q_{near} in the direction of q_{rand} using a given step size. The principle is illustrated in Figure 7.1.a. This simple exploration strategy favors a rapid growth of the tree towards unexplored regions and guarantees convergence towards a uniform coverage of the reachable regions of $\mathscr{C}$ from q_{init} [95].

A stochastic transition test (lines 9 in Algorithm 2) is then applied to q_{new} aiming to favor the exploration of low-cost/high-quality regions of $\mathscr{C}$. This important component of the TRRT algorithm will be further explained below. If the transition test succeeds, the new configuration q_{new} is added to $\mathscr{T}_i$ and connected to q^i_{near} (lines 10-11 in Algorithm 2). Then, the algorithm searches for the closest configuration q^j_{near} to q_{new} in all the other trees $\mathscr{T}_j \neq \mathscr{T}_i$. If q^j_{near} and q_{new} are close enough and the transition test is accepted in at least one direction, the two trees are connected (lines 12-14 in Algorithm 2). The exploration continues until all trees are merged or another stop condition (e.g. maximum number of iterations, timeout, ...) is reached. Figure 7.2 illustrates the behavior of TRRT (the single-tree variant is illustrated here for clarity purposes) exploring the conformational energy landscape

[2]In this pseudo-code, corresponding to the parallel implementation, q_{rand} is not sampled in the whole space $\mathscr{C}$, as is the case for the basic Multi-TRRT algorithm, but in a subset S_i. This will be further explained in Section 7.4.

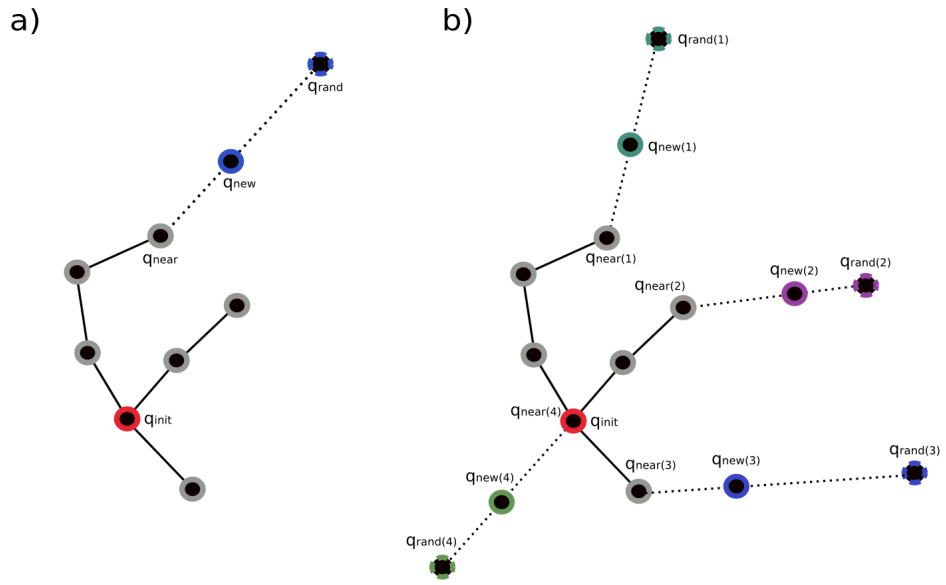


Figure 7.1. Illustration of the TRRT expansion process (for a single tree). The red node is the initial configuration and grey nodes have been generated in previous iterations. A new node q_{new} is created by the extension of the nearest node in the tree q_{near} toward a randomly sampled configuration q_{rand}. a) Mono-thread extension. b) Multi-thread extension, where k new nodes are generated in parallel.

of a very simple molecular system, alanine dipeptide [94]. The figure shows that the tree tends to explore low energy regions aiming to find the most favorable transition pathway connecting two given configurations.

The stochastic transition test applied to q_{new} (lines 9 in Algorithm 2) is based on the evaluation of a function $E : \mathscr{C} \to \mathbb{R}$ that associates a real-value cost with each configuration q.

In the case of the molecular models considered in this work, the cost/quality function E correspond to the potential energy computed from a classical molecular mechanics forcefield [115], as generally used in molecular simulations. More detailed explanations about this cost/energy function are out of the scope of this chapter, and the choice of a more suitable function to investigate IDPs remains for future work (see Conclusions section). However, it is important to note that this energy function is computationally expensive, and typically scales quadratically with the size of the molecule being evaluated. The transition test within TRRT is inspired by the Metropolis criterion commonly used in MC methods, and involves a self-adaptive parameter T that we call temperature by analogy with methods in statistical physics. The pseudo-code is provided in Algorithm 3. Moves to lower-cost configurations are always accepted, and the temperature is unchanged in that case. Uphill moves are accepted with a probability that decreases exponentially with the local energy variation. After each accepted uphill move, T is decreased to avoid over-exploring high-cost regions. After each rejected uphill move, T is increased to facilitate the exploration and to avoid being trapped in local minima. The greedi-

Algorithm 3: transitionTest ($\mathscr{T}$, E_i, E_j)

input : the current temperature T; the temperature increase rate T_rate; the Boltzmann constant K
output: *True* if the transition is accepted, *False* otherwise

```
1  if E_j ≤ E_i then return True
2  if e^{-(E_j - E_i) / (K·T)} > 0.5 then
3      #pragma omp critical (Temperature)
4      T ← T / 2^{(E_j - E_i) / energyRange(T)}
5      return True
6  else
7      #pragma omp critical (Temperature)
8      T ← T · 2^{T_rate}; return False
```

Algorithm 4: updateStructures (q_new, S_i)

```
1  if q_new ≠ null and nodeNearBoundary(q_new, S_i) then
2      #pragma omp critical (updateNPolytope)
3      S_i ← updateNPolytope(q_new, S_i)
4      #pragma omp critical (Communication)
5      MPI_send(Master, S_i)
6      forall (I_{i,j}) do
7          if nodeInsideIntersection(q_new, I_{i,j}) then
8              #pragma omp critical (Communication)
9              MPI_send(Processor(j), q_new)
10 while MPI_received(q_new) do
11     (T_j, q^j_near) ← findNearestNeighbor(T, q_new)
12     if distance(q_new, q^j_near) ≤ δ then
13         MPI_send(Master, q_new, i, j)
14 while MPI_received(I_{i,j}) do
15     addToIntersectionList(I_{i,j})
16 if Master then
17     while MPI_received(S_j) do
18         if intersection(S_i, S_j) then
19             MPI_send(Processor(i,j), I_{i,j})
20     while MPI_received(q_new, i, j) do
21         updateGraphOfTrees(q_new, i, j)
22     if numberCC(GraphOfTrees) = 1 then
23         MPI_broadcast(endMessage)
```

ness of the algorithm depends on the parameter T_rate, which is a real value in the interval $(0, 1]$ provided as input. In general, $T_\mathrm{rate} = 0.1$ has been empirically shown to be a good tradeoff between path quality and computing time [50].

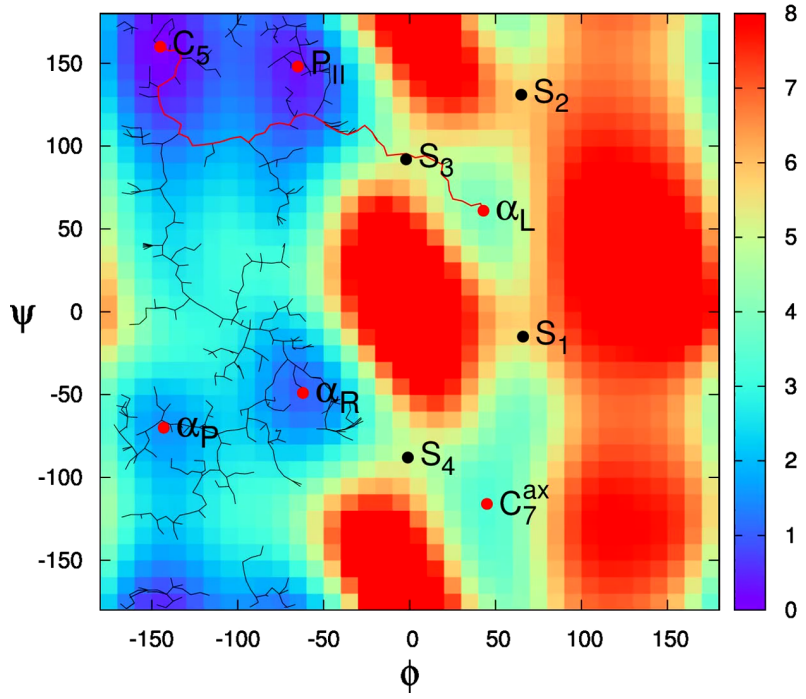


Figure 7.2. Illustration of TRRT exploring an energy landscape. The background image represents a two-dimensional projection of the conformational space of a small peptide. The red dots represent the local energy minima (C_5, P_{II}, α_R, α_P, α_L, C_7^{ax}), and the black dots represent the main transition states (S_1, S_2, S_3, S_4). The background colors represent energy values with respect to the global energy minium, C_5. TRRT is applied here to find a path from C_5 to one of the local energy minima, α_L. The TRRT search tree rooted at C_5 grows on low-energy regions and explores other basins of the landscape before finding a higher-energy saddle region (around S_3) from which α_L can be easily reached.

7.4 Materials and methods

This section presents first a global overview of the approach, and then provides details of the parallel implementation.

7.4.1 General principle

The hybrid parallelization of Multi-TRRT presented below is aimed to better exploit current (multi-core) computer clusters. The idea is to minimize inter-processor communication overhead while taking advantages of shared-memory operations. In addition, this kind of parallelization permits to easily adapt the algorithm to all types of computer architectures. For the implementation, we use the standard and widely-used Message Passing Interface (MPI) for inter-processor communication and Open Multi-Processing (OpenMP) for intra-processor work. Some rules have to be respected between these two portable APIs in order to perform an efficient global interaction, as will be explained below.

A logical view of the architecture consists of p processes and k threads within

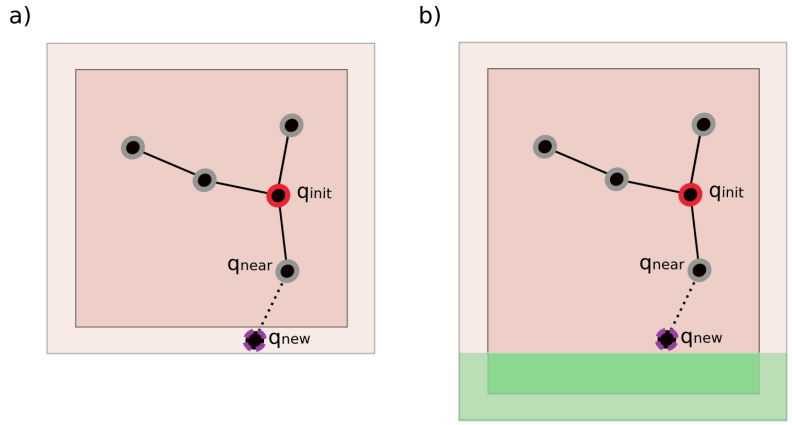


Figure 7.3. Illustration of the concept of a bounding n-polytope of a tree and its growth process. a) A new node is created near the boundary of the n-polytope, represented by the light brown region. b) The n-polytope is extended around the node (new green region).

each process. We consider one process per processor and one thread per real core. To be clear, we do not use the available hyper-threading technology that allows to run two threads per core, since this can have some drawbacks in our case. One of the processors is chosen as a master processor, which will take in charge some specific tasks while also performing the other tasks as the rest of the processors. In a first stage of the algorithm (function `initProcesses` in Algorithm 2), the master processor distributes the m initial configurations Q_{init} among the processes. We consider $m \geqslant p$, so that each process builds m/p trees in average, almost independently of the other processes. The distribution takes into account the distances between the initial configurations, in such a way that neighboring configurations are assigned to each processor. This is important to reduce inter-processor communication, as will be better understood later on.

An adaptive space subdivision approach is applied to avoid redundancies in the exploration performed by the different processes, and to reduce the communication between processors. The idea is to associate a bounding volume that we call n-dimensional polytope (n-polytope) S_i to each tree $\mathscr{T}_i$. The shape of the n-polytope evolves as the tree grows. If the n-polytopes S_i and S_j associated to two trees $\mathscr{T}_i$ and $\mathscr{T}_j$ do not intersect, this means that the exploration is taking place in different regions of the space. Thus, the trees can grow independently from each other, with no communication requirements between the corresponding processes. On the other hand, when two n-polytopes intersect, communication is required for attempting the connection of the associated trees. As will be further explained below, in our implementation, the master processor manages the information concerning the n-polytopes of all the trees, computes the intersections, and informs the other processors when they need to exchange data. The master process also detects when all the trees are connected, and informs that the exploration can be stopped.

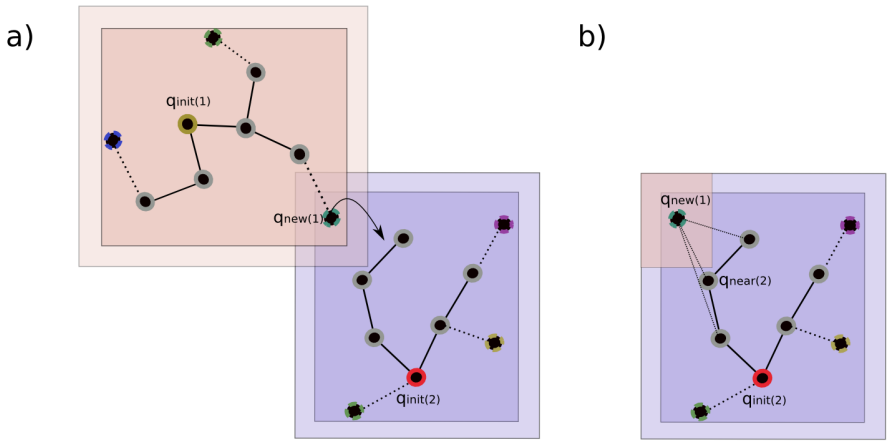


Figure 7.4. Illustration of the tree junction process. a) The extension of $\mathscr{T}_1$ generates a new node $q_{\text{new}(1)}$ in the intersection of bounding n-polytopes $\mathscr{I}_{1,2}$. b) The processor in charge of the construction of $\mathscr{T}_2$ tries to connect $q_{\text{new}(1)}$ to the nearest neighbor.

7.4.2 Cooperative construction of trees inside each process (OpenMP)

We explain next how a group of trees is built using multiple threads in a shared-memory multi-core processor. The idea is to parallelize the whole exploration loop. From the directive `#pragma omp parallel` (line 2 in Algorithm 2), several threads are created. All the threads building cooperatively the set of trees assigned to a processor (see Figure 7.1.b for an illustration). Each thread performs independently of the others the tree selection, sampling and extension operations. The only operations that require to be inside critical sections (i.e. that cannot be executed in parallel) involve the modification of shared variables: the trees $\mathscr{T}_i$, the bounding n-polytopes S_i, and the temperatures T_i. Thanks to a careful implementation, the workload in critical sections is very small compared to the rest of the process, which is essential for a good performance of the algorithm. The most important implementation details are provided next.

Copies of the molecular system Most of the operations concerning sampling and tree expansion require handling a model of the molecular system. Thus, a copy of this model is provided to each thread in order to avoid race conditions. The drawback of this solution is that it requires more memory space. Nevertheless, this is not an important issue, since the memory space required by the copies of the molecular system is small compared to the space requited by the exploration trees being constructed.

Multiple temperatures In the basic Multi-TRRT algorithm, the temperature T is a global variable parameter. In other words, a single T parameter is used for the construction of all the trees. However, as each tree is exploring a different

region of the space, it seems reasonable to assign different temperatures to different trees. This has several advantages. First, since different threads in a process are often working with different trees (especially when $m \gg p$), they are rarely blocked because of the critical section for updating the value of T. Besides, when using several processors, they do not need to communicate about variations of T, since it remains as a local variable for each tree. Furthermore, considering multiple temperatures improves the quality of the overall exploration (experimental results showing this are not presented in this chapter).

Multi-threading nearest neighbor search The Hierarchical K-means Tree method of the Fast Library for Approximate Nearest Neighbors (FLANN) [153] has been applied to perform the search of the nearest neighbor in the tree (function `findNearestNeighbor` in Algorithm 2). The search can be executed simultaneously by multiple threads. However, the addition of a node is critical since the data-structure used for the rapid search has to be updated, which cannot be done while other thread are writing or reading in it. Thus, node insertion must be in a critical section that also affects node search. To avoid a possible bottleneck, all threads write the nodes to be inserted in a shared container. The insertion is done when none of the other threads is performing the search. Only if the container reaches a limit size, one of the threads enforces node insertion, blocking the access to the other threads. This is required to ensure a regular updating of the data-structure used by the search algorithm.

Merging trees When two trees constructed by the same process can be connected, they are merged in a single tree (lines 14-16 in Algorithm 2). The merging operation is performed by the thread that created the connecting node. To avoid race conditions during this operation, merging has to be delayed until all the other threads finish ongoing operations within the two trees concerned.

7.4.3 Limiting communication between processes (MPI)

The efficiency of a parallel algorithm strongly depends on the computational cost associated to inter-process communication. To reduce this cost, we have implemented an adaptive space subdivision approach as explained below. The idea is to delay the communication between processes until is it really necessary, and to use a master process to manage information exchange. The main operations requiring communication are performed inside the `updateStructures` function (line 17 in Algorithm 2), which is detailed in Algorithm 4.

Adaptive space subdivision At the initialization (line 1 in Algorithm 2), a small n-dimensional polytope (n-polytope) S_i is associated to each $q^i_{\text{init}} \in Q_{\text{init}}$, where n is the dimension of the space being explored. In this work, we use hyperrectangles because of the simplicity to update their shape and to compute intersections. Nevertheless, other more accurate representations could be used. At

the beginning, S_i is symmetric and centered on q^i_{init}. Random sampling for the expansion of each tree is performed in its corresponding n-polytope. When a new node is created near the boundary, the n-polytope grows in the direction of the tree expansion (lines 1-2 in Algorithm 4). The process is illustrated in Figure 7.3. Every time that a n-polytope is updated, the information is sent to the master processor (lines 3-4 in Algorithm 4), which will compute possible intersections with n-polytopes associated to tress built by other processors. Note that communication must be in an OpenMP critical section.

N-polytope intersection and trees junction The master processor computes the intersection between the n-polytope S_i sent by a processor i and the n-polytopes S_j associated to trees managed by other processors (lines 16 in Algorithm 4). If the intersection $\mathscr{I}_{i,j}$ is not empty, the information is sent to the corresponding processors (lines 17-18 in Algorithm 4), and they add $\mathscr{I}_{i,j}$ to their n-polytope intersections lists (lines 13-14 in Algorithm 4).

When a processor i creates a new node q_{new} lying inside an intersection between n-polytopes $\mathscr{I}_{i,j}$, the node is sent to the corresponding processor j (lines 5-8 in Algorithm 4). Then, the other processor will try to connect the two trees (lines 9-12 in Algorithm 4), as in the basic Multi-TRRT. The process is illustrated in Figure 7.4.

Stopping condition and path extraction In addition to performing space intersections, the master processor maintains a graph data structure to represent the connectivity between all the trees (lines 19-20 in Algorithm 4) using information sent by all the processors. The exploration process can stop when this graph contains a single connected component (lines 21-22 in Algorithm 4). Then, the solution path connecting all the initial configurations Q_{init} can be extracted. As the solution path is distributed among the processors, each processor i extracts the part of the path that connects q^i_{init} with the nodes that served as connectors with trees constructed by other processors.

7.4.4 Implementation framework

To implement the two aforementioned levels of parallelization, we use a combination of MPI (for distributed-memory parallelization) and OpenMP (for shared-memory parallelization). In such a hybrid framework, multiple threads may concurrently call MPI functions, requiring the MPI implementation to be thread-safe. Our algorithm makes MPI calls sporadically, but requires all the threads to be available to make them. Two levels of thread safety (among the four available) can be used in our case: MPI-THREAD-SERIALIZED and MPI-THREAD-MULTIPLE. For simplicity purposes, we use the MPI-THREAD-SERIALIZED safety level, where multiple threads can make MPI calls but not simultaneously, as is the case for MPI-THREAD-MULTIPLE. This implies that all the communications have to be performed inside

OpenMP critical sections. Non-blocking receive operations (`MPI_irecv`) are used to reduce the time spent inside these critical sections.

Since our software is written in C++ and MPI is a library of routines to create parallel programs in C or Fortran, we had to use a C++ binding of MPI. In addition, our application messages may contain instances of high-level classes, whose attributes can be pointers or Standard Template Library (STL) containers, which is incompatible with low-level MPI communication, requiring the programmer to explicitly specify the size of each message. Therefore, we exploit the higher-level abstraction provided by the Boost.MPI library (http://www.boost.org/). Coupled with the Boost.Serialization library, it enables processes to exchange class instances, making the tasks of gathering, packing and unpacking the underlying data transparent to the programmer. Given that OpenMP (http://openmp.org/) supports C++ language, no extra library was required.

7.5 Results and discussion

This section presents an empirical performance analysis of the proposed algorithm. First, we present the problems considered for this analysis, as well as the specifications of the computers we used. Then, the performance of the sequential algorithm, running on a single core, is analyzed to identify the most computationally expensive operations. The performance of the parallelized algorithm is then analyzed on a multi-core processor and on a cluster of processors, showing the interest of the hybrid approach.

7.5.1 Problem studied

We have evaluated the Hybrid-MultiTRRT algorithm on several energy landscape exploration problems involving flexible biomolecules of different sizes. The number of degrees of freedom (DOF) of the molecule defines the dimension n of the space being explored. In theory, the complexity of the problem grows exponentially with n. We have used two relatively small peptides, met-enkephalin [165] and chignolin [191], and intrinsically disordered regions of the vasopressin 2 receptor [206] and a β-2 adrenergic receptor [73]. Hereafter, we will refer to these four molecules as MNK, CHGN, V2R and B2AR, respectively. The conformational exploration was performed using an internal-coordinates representation of the molecules, assuming constant bond lengths and bond angles. The number of DOF for MNK, CHGN, V2R and B2AR are 24, 46, 173 and 425, respectively. For all four molecules, the energy landscape exploration was started from 32 randomly sampled configurations. More precisely, we generated 32 configurations using random sampling followed by local energy minimisation in order to obtain acceptable structures. In addition, a minimum distance was imposed between these initial configurations in order to maximize space coverage. The problem then consisted of building exploration trees rooted at these 32 initial states, and eventually to find paths connecting all of them.

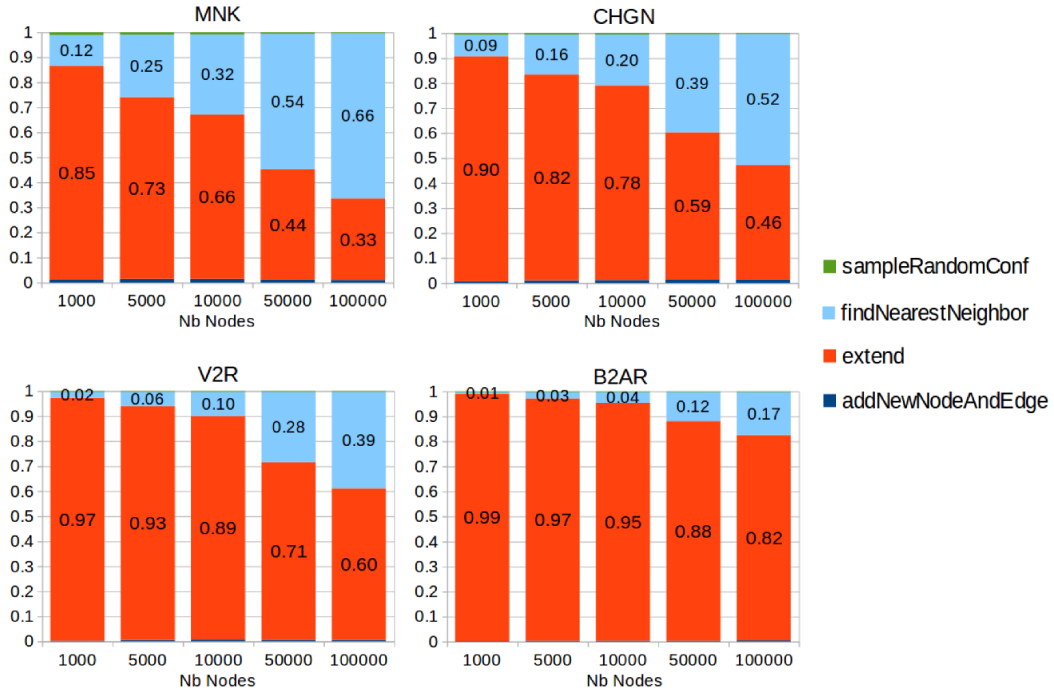


Figure 7.5. Computation time decomposition depending on the number of nodes in the trees for the four molecules under study.

The same problem was solved for all the instances we tested in terms of number of threads and processors.

7.5.2 Computer architecture

For the evaluation of the sequential algorithm and the multi-threaded implementation, we used a server with the following features: Intel® Core™ i7 processor at 2.8 GHz, 16-cores, 32 GB RAM.

The evaluation of the hybrid algorithm was performed on the EOS supercomputer at CALMIP (Centre de Calcule Midi-Pyrénées). EOS is a Bull supercomputer with 612 Intel® IVYBRIDGE processors at 2.8 GHz, with 20 cores and 64 GB RAM per processor, and a InfiniBand Full Data Rate with 6.89 GB/s bandwidth for inter-processors communication. For our experiments, we used up to 32 processors (i.e. up to 640 cores).

7.5.3 Analysis of the sequential algorithm

Figure 7.5 shows the percentage of time that the algorithm is expending in the most computationally-expensive operations. For each molecule, the plot shows the time decomposition depending on the number of nodes in the set of exploration trees. The sum of `extend` and `findNearestNeighbor` methods requires for all the cases more than 99% of the time. Note that the energy computation is actually

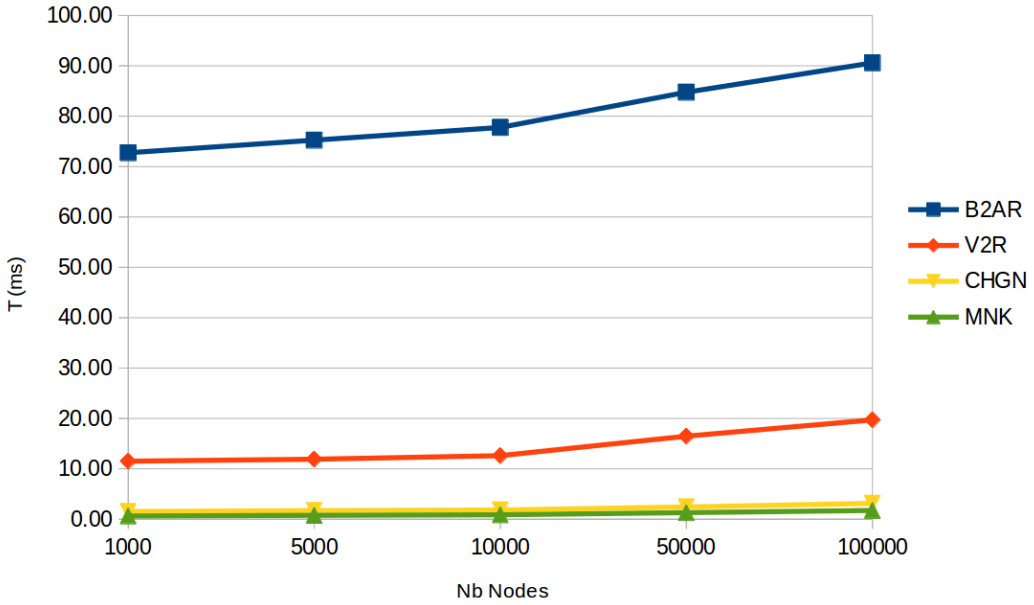


Figure 7.6. Evolution of the time per iteration with respect to the total number of nodes in the tree.

performed inside `extend`, and the value is stored for subsequent use. Therefore, the parallelization of these two operations is essential for performance improvements. It can be clearly seen in the figure that the time consumed by `extend` increases with the size of the molecule, being largely dominant for the two biggest molecules because of the computational cost of the energy computation. As expected, the time required by `findNearestNeighbor` becomes really significant when the trees grow above several thousands of nodes. `sampleRandomConf` and `addNewNodeAndEdge` operations are very fast in all the cases, representing less than 0.1% of the total time. Therefore, introducing an OpenMP critical section to protect memory writing inside `addNewNodeAndEdge` will not represent an important overhead.

Figure 7.6 show the average time per iteration depending on the number of nodes in the tree. Since the cost of the most time-consuming operation, `extend`, is independent of the trees size, the time per iteration increases slowly for all the molecules. The time varies significantly with the size of the molecule. Compared to the smallest molecule, MNK, the time per iteration increases by one order of magnitude for V2R, and by almost 2 orders of magnitude for B2AR.

7.5.4 Analysis of the multi-threaded algorithm running on a single processor

As a preamble, before the implementation and analysis of the hybrid parallelization approach, we analyzed the performance of the Multi-TRRT algorithm running on a multi-core processor. The goal was to evaluate the potential interest of a larger-scale parallelization.

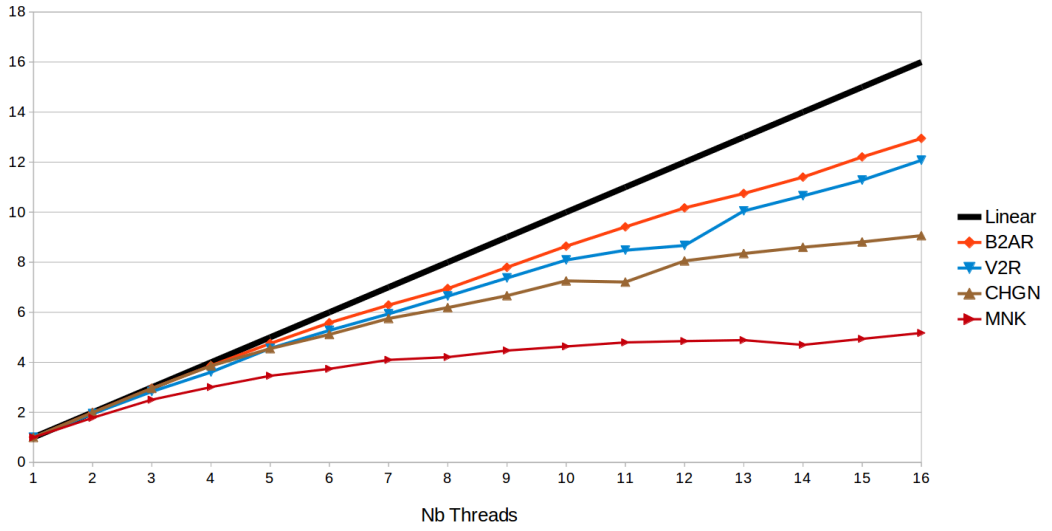


Figure 7.7. Evolution of the speed-up of the parallel algorithm running on a single (multi-core) processor for the four molecules. As a reference, the black line represent the linear speed-up.

The usual metric to evaluate the performance of a parallel algorithm over its sequential counterpart is the speed-up $S(k) = t_s/t_p(k)$. Where t_s is the time needed to solve the problem working with one thread: $t_s = t_p(1)$, and $t_p(k)$ is the running time when k threads are used. The results presented in Figure 7.7 show that the performance strongly depends on the size of the molecule. The plot corresponds to CPU times averaged over 5 executions for each instance. For a small system such as MNK, a maximum speed-up of around 4.5 is reached for 11 threads. Then, the performance does not improve with a larger number of threads. For CHGN, the speed-up curve also tends to show an asymptotic shape, as for MNK, but with a maximum value approaching 8 for 16 threads. The reason of this limited performance gain is that the time per iteration of the main loop of the Multi-TRRT algorithm is very short for small systems (see Figure 7.6). Therefore, when the number of threads increases, the time spent waiting for access to critical sections starts to be significant, eventually becoming a bottleneck. Indeed, large-scale parallelization is useless for small molecules. For large systems such as V2R and B2AR, however, the speed-up increases almost linearly from 1 to 16 threads, with maxima values around 12 and 13, respectively. This shows that, in principle, better performance gain can be obtained with a larger number of threads using a computer cluster.

In many cases, the speed-up is not the only criterion to assess the quality of a parallel algorithm. For instance, in our case, it is important to verify that the quality of the exploration and of the solution paths do not degrade with an increasing number of threads.

A side effect of the parallelization of RRT-based algorithms is that some redundancy in the exploration can be introduced by multiple threads simultaneously

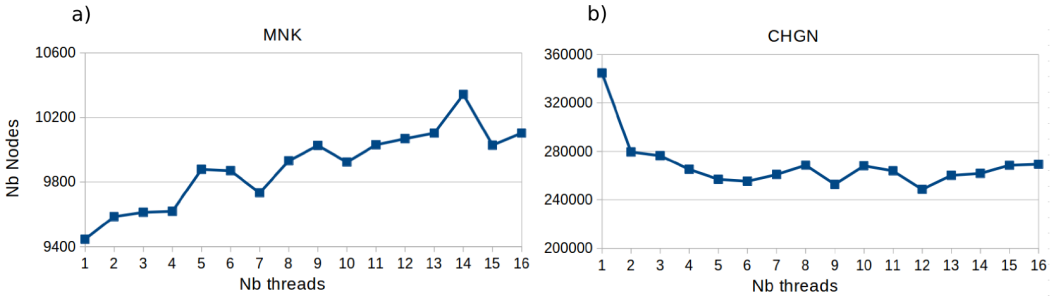


Figure 7.8. Evolution of the total number of nodes needed to solve the transition path-finding problem depending on the number of threads for the two peptides: a) MNK, b) CHGN.

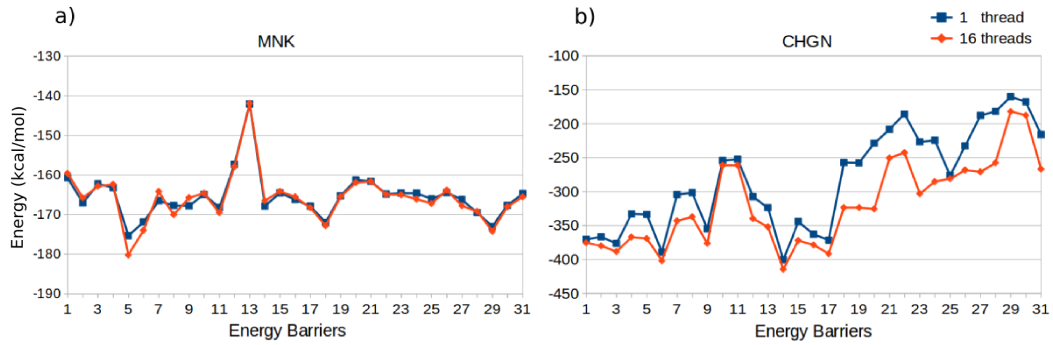


Figure 7.9. Energy profiles along the solution paths obtained using 1 thread and 16 threads for the two peptides: a) MNK, b) CHGN.

creating nodes in nearby regions of the space. A simple test to detect if this happens is to look at the size of the trees (i.e. number of nodes) requires to solve the same problem with different numbers of threads. An increasing number of nodes with the number of threads would mean a degradation of the exploration quality. As shown in Figure 7.8.a, the total number (averaged over 100 runs of the algorithm) of nodes in the trees required to find paths connecting the 32 initial configurations for MNK increases slightly with the number of processors (up to 10%). This shows that, for problems in relatively low dimension, there is some redundancy in the exploration performed by the parallel version of Multi-TRRT. However, such an undesired behaviour disappears when the dimension increases. This can be clearly seen in Figure 7.8.b for CHGN, which shows that the number of nodes in the trees required to solve te problem is almost constant, independently on the number of threads. Surprisingly, in this case, the number of nodes is larger for a single thread (i.e. for the sequential algorithm), which actually shows that the parallel algorithm performs a more efficient exploration in high-dimensional spaces, due to what we call the “OR parallel effect" [51]: As each thread performs its own sampling of the space, when multiple threads are involved, the parallel algorithm reaches smaller-size solutions than the sequential one, on average. This phenomenon is more important in problems containing “narrow passages", corresponding to saddle regions in the

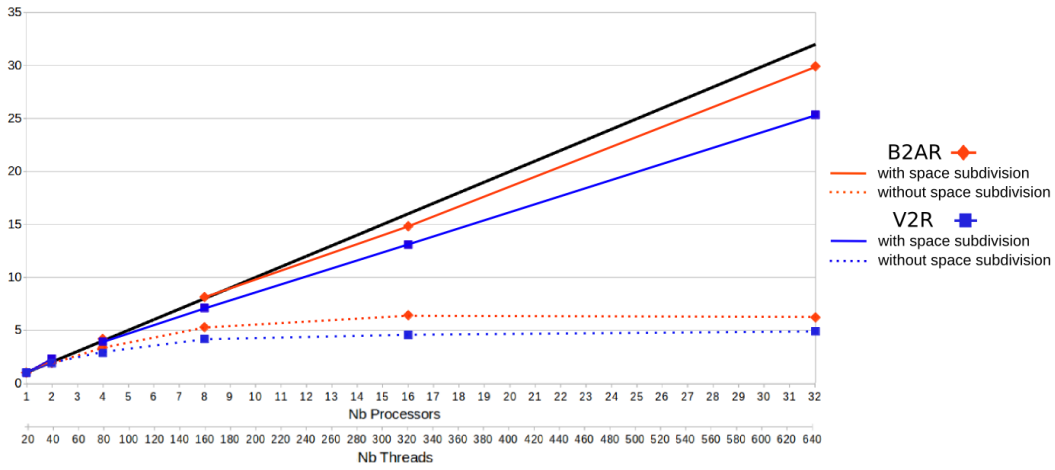


Figure 7.10. Evolution of the speed-up of the parallel algorithm with respect to the number of processors (working with 20 threads per processor). As a reference, the black line represents the linear speed-up.

energy landscape of the molecules, which require intensive sampling to be found.

To evaluate the quality of the paths obtained by the parallel implementation of Multi-TRRT compared to the sequential one, we can compare the corresponding energy profiles. Remind that the solution provided by the algorithm is a set of connected trees from which a path connecting the 32 initial configurations can be extracted. We can obtain a simplified representation of the energy profile by identifying the highest-energy configuration (i.e. the transition state) between each pair of initial configurations directly connected along the path. Figure 7.9.a shows that the energy profiles of the solutions obtained with the sequential and the parallel versions (using 16 cores) of Multi-TRRT are very similar in the case of MNK, thus demonstrating that the quality of the solutions is preserved for problems in relatively low dimension. In higher dimension, as shown in Figure 7.9.b for CHGN, the quality of the paths are better for the solutions obtained by parallel algorithm running on 16 threads than for the sequential algorithm. This is also a consequence of the “OR parallel effect", which yields a better sampling of high-dimensional spaces with Multi-TRRT when several processes run in parallel.

7.5.5 Analysis of hybrid algorithm

We have evaluated the performance of the hybrid parallelization of Multi-TRRT for the two large molecules: V2R and B2AR. For this, instead of measuring the time required to connect the 32 initial configurations, we have measured the time required to generate 200,000 nodes. The reason is that the former involves a larger variance than the latter, therefore requiring a larger number of runs in order to obtain a meaningful average value. As we have seen in the previous subsection, the number of nodes needed to solve a problem is almost constant independently on the number of threads/processes, and the variance is very low. Therefore, the time

to generate a given number of nodes is a good indicator of the performance that requires only a few runs (we performed 10 runs) per instance.

Figure 7.10 shows how the speed-up of the parallel algorithm increases with the number of processors. The speed-up increase is linear at the beginning, when using up to 4 processors for V2R and up to 8 processors for B2AR. Then, it slightly decreases, this performance degradation being a bit more significant for V2R. Nevertheless, the speed-up using 32 processors is around 30 for the largest molecule, B2AR. This shows that the proposed parallelization strategy is very efficient, particularly thanks to the space subdivision approach to reduce inter-processor communication.

7.6 Conclusions

Nowadays, most HPC systems are clusters of multi-core processors. We have presented a parallel implementation of Multi-TRRT to efficiently exploit this type of architectures, combining distributed-memory parallelization using MPI with shared memory-parallelization using OpenMP. Such a hybrid parallelization strategy clearly outperforms our previous fully-distributed implementations of RRT-like algorithms [51], significantly reducing communication overhead and memory needs. This implementation is also very flexible, since the algorithms can be run on a single multi-core processor (without communication requirements) or on a large computer cluster without any modification in the code. The adaptive space subdivision approach is a key component of the proposed parallelization strategy. It drastically reduces computational cost associated to inter-processor communication and nearest neighbor search.

Although the work presented here is focused on the application of the proposed parallel algorithm to highly-flexible biomolecules, the explanations concerning the methods can be easily extracted from the application context. The principle is very general, and could be applied to other sampling-based path search/planning algorithms applied in different domains. Indeed, we expect that our work will be a source of inspiration for the parallelization of related methods.

As mentioned above, in this work, we have used a classical molecular mechanics forcefield to evaluate the potential energy of the molecule. This choice of an AMBER-like potential was motivated by the generality of this type of energy function, and because our first goal was the development of the parallel version of the conformational exploration algorithm rather than a particular application. Nevertheless, AMBER-like potentials are probably not the best choice for the investigation of IDPs. The analysis of different energy functions in order to select the most suitable one for IDPs remains for future work.

Chapter 8

Conclusions and Perspectives

This manuscript has presented several algorithmic contributions aiming at better understanding the structural and dynamic behaviour of Intrinsically Disordered Proteins (IDPs). A fundamental pillar of the thesis is a tripeptide database built from a large set of experimentally-determined high-resolution protein structures. The results derived from the different algorithms based on this database have demonstrated the importance of local sequence-dependent structural properties to determine IDP conformations.

The first algorithm, a secondary structure predictor specially designed for IDPs, aims at distinguishing partially ordered fragments from disorder within IDPs. It shows that we can reliably extract secondary structure propensity in IDPs by a classification of the tripeptide conformations based on the Ramachandran angles and a simple statistical approach accounting for the neighboring residues. Our algorithm provides better results that other published methods. The main advantage is the simplicity that allows the connection between sequence and specific local structural features.

The second algorithmic contribution is a structural ensemble builder for IDPs that uses the tripeptides of the database as building-blocks. The algorithm uses a combination of two sampling strategies: one for random coil regions and the other for regions with a tendency to form secondary structures. To validate the model, we built ensembles for a benchmark set of nine well-characterized IDPs. The excellent agreement between the RDCs computed from the ensembles and the experimental ones demonstrates the accuracy of the method. Building realistic ensembles allowed us to better understand the distribution of secondary structural types within IDPs. One of the difficulties of the method is the selection of the optimal sampling strategy to be used in each section of the protein. This can be done by selecting the regions according to the experimental data, as presented in Chapter 5, but we envision the application of the secondary structure predictor described in Chapter 4 as a valuable tool for this purpose.

The third algorithm presented in this thesis uses the tripeptide database to compute likely transitions between two conformational states of a protein (or protein region). The proposed heuristic search algorithm is able to identify relevant protein folding pathways of small structured motifs. It can also be used to better understand the transitions between order and disorder in partially structured regions of IDPs. Indeed, IDPs undergo these structural transitions continuously. Our approach is orders of magnitude faster than MD simulations. It is worth mentioning that even if the resulting folding pathway is an approximation, relevant information about

the main steps of the folding mechanism can be derived.

Finally, the fourth contribution is a parallel version of a global exploration algorithm called Multi-TRRT. Multi-TRRT is an efficient algorithm, originating from robotics, applied to compute the transition between multiple states by exploring the energy landscape of a given system. Due to the high dimensional space of IDPs, the calculation of transition paths between different states becomes computationally intractable in a reasonable time. To overcome this problem, we designed the parallelization of the code to be executed in a HPC cluster. The results of our hybrid parallelization, combining OpenMP and MPI, show an almost-linear speedup, meaning that the execution time can be approximately divided by the number of processor cores used. Thanks to this, computing time to globally explore the conformational space of IDPs can be easily reduced from years to days.

Future Work

The good results obtained during this thesis are encouraging to continue with the development of computational methods to investigate IDPs. We envision several steps and directions to pursue this work.

First, we plan to make the methods accessible to the scientific community, in particular through user-friendly web servers. Indeed, the method to generate IDP ensemble models has already been used by others to model disordered regions in proteins such as linkers and tails. The possibility of sampling disordered regions can be interesting for a broad research community in different contexts of application. Moreover, making our methods available will also allow us to get feedback from the users, which can be very valuable to improve our methods.

One interesting possible usage of our algorithms is the study of structural and dynamic perturbations of point mutations. Our methods can be used to compare the result of the native sequence with that of a given mutated sequence: with the secondary structure predictor the resulting secondary structure propensities can be compared; the ensemble generation algorithm could produce different conformations for the mutated sequence and the differences in terms of structure can be analysed; the resulting transitions paths from the two algorithms presented (heuristically guided algorithm and Hybrid Multi-TRRT) could also highlight interesting differences between the mutants and the native sequence.

Another point for future work is the evolution of the tripeptide database. The PDB is continuously growing through new three-dimensional structures deposited. As a consequence, our tripeptide database will also become larger and richer with time. A more complete database will automatically improve the results of our algorithms. Another possibility to increase the number of structures in the tripeptide database would be the use of all the structures deposited in the PDB, instead of using SCOP. Our database can be subsequently filtered according to specific needs. A complementary direction of improvement of the database, which is already being explored in our group, is to select the structures according to the solvent exposure.

In other words, choosing tripeptides according to their position in the protein, from the core of the protein to the surface. Preliminary work (not presented in this manuscript) shows that tripeptides extracted from the core of the protein are more likely to form secondary structures than tripeptides extracted from the surface. Using experimental data, such as RDCs or chemical shifts, we can evaluate which subset of the tripeptide database is more accurate for the different usages proposed in my thesis.

There are multiple ways to improve the algorithms presented in Chapters 6 and 7. For the heuristic database-assisted path search algorithm, more sophisticated techniques, such as Monte Carlo tree search, could be implemented. Such improved implementation would provide higher quality paths in terms of energy (paths passing through higher-density regions) and in terms of accuracy (more intermediate steps could be found).

Concerning the Hybrid Multi-TRRT, the significance of the results provided by the method is strongly dependent on the applied energy function. The integration of more appropriate energy models for IDPs instead of the generic AMBER-like potential used here for the evaluation of the computational performance of the algorithm is an interesting direction for future work. Besides, Multi-TRRT can also be improved by choosing more appropriate distance metrics for IDPs. The metric used in this work, the RMSD over all the dihedral angles, is not very suitable for large conformational changes, and is computationally expensive. In the case of IDPs, other simplified metrics, such as the Cα-Cα euclidean distance over some selected residues could be a better-suited metrics.

Finally, we can envision future research topics, such as the rational design of IDPs with tailored structural properties. Traditionally, protein design consists of finding a sequence that folds into a given three-dimensional structure. This paradigm is limited to the design of rigid systems, and cannot be easily extended to the design of highly-flexible proteins or regions. The acquired knowledge during this thesis on the sequence/structure relationships paves the way to the computational design of IDPs for a broad range of applications in biotechnology.This manuscript has presents several algorithmic contributions aiming at better understanding the structural and dynamic behaviour of Intrinsically Disordered Proteins (IDPs). A fundamental pillar of the thesis is a tripeptide database built from a large set of experimentally-determined high-resolution protein structures. The results derived from the different algorithms based on this database have demo

Bibliography

[1] I. Aguinaga, D. Borro, and L. Matey. Parallel RRT-based path planning for selective disassembly planning. *The International Journal of Advanced Manufacturing Technology*, 36(11-12):1221–1233, 2008. (Cited in page 118.)

[2] I. Al-Bluwi, T. Siméon, and J. Cortés. Motion planning algorithms for molecular simulations: A survey. *Computer Science Review*, 6(4):125–143, 2012. (Cited in pages 37, 95, 116, and 117.)

[3] A. Almond and J. B. Axelsen. Physical interpretation of residual dipolar couplings in neutral aligned media. *Journal of the American Chemical Society*, 124(34):9986–9987, 2002. (Cited in page 77.)

[4] N. M. Amato and L. K. Dale. Probabilistic roadmap methods are embarrassingly parallel. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 688–694, 1999. (Cited in page 118.)

[5] M. Arbesú, M. Maffei, T. N. Cordeiro, J. M.C. Teixeira, Y. Pérez, P. Bernadó, S. Roche, and M. Pons. The unique domain forms a fuzzy intramolecular complex in Src family kinases. *Structure*, 25(4):630 – 640.e4, 2017. (Cited in page 78.)

[6] M. Aslam, J. M. Guthridge, B. K. Hack, R. J. Quigg, V. M. Holers, and S. J. Perkins. The extended multidomain solution structures of the complement protein crry and its chimeric conjugate crry-ig by scattering, analytical ultracentrifugation and constrained modelling: Implications for function and therapy. *Journal of Molecular Biology*, 329(3):525 – 550, 2003. (Cited in page 32.)

[7] M. M. Babu, R. van der Lee, N. Sanchez de Groot, and J. Gsponer. Intrinsically disordered proteins: Regulation and disease. *Current Opinion in Structural Biology*, 21(3):432 – 440, 2011. (Cited in page 13.)

[8] L. Baeten, J. Reumers, V. Tur, F. Stricher, T. Lenaerts, L. Serrano, F. Rousseau, and J. Schymkowitz. Reconstruction of protein backbones from the brix collection of canonical protein fragments. *PLOS Computational Biology*, 4(5):1–11, 05 2008. (Cited in pages 41, 89, and 94.)

[9] R. L. Baldwin. Protein folding: Matching speed and stability. *Nature*, 369:183–184, 1994. (Cited in page 94.)

[10] A. Bax, G. Kontaxis, and N. Tjandra. Dipolar couplings in macromolecular structure determination. In *Methods in Enzymology*, pages 127–174. Elsevier, 2001. (Cited in page 27.)

[11] A. Bax and N. Tjandra. High-resolution heteronuclear nmr of human ubiquitin in an aqueous liquid crystalline medium. *Journal of Biomolecular NMR*, 10(3):289–292, 1997. (Cited in page 27.)

[12] P. Bernadó, C. W. Bertoncini, C. Griesinger, M. Zweckstetter, and M. Blackledge. Defining long-range order and local disorder in native -synuclein using residual dipolar couplings. *Journal of the American Chemical Society*, 127(51):17968–17969, 2005. (Cited in page 36.)

[13] P. Bernadó and M. Blackledge. A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering. *Biophysical Journal*, 97(10):2839–2845, Nov 2009. (Cited in pages 35 and 86.)

[14] P. Bernado and M. Blackledge. Structural biology: Proteins in dynamic equilibrium. *Nature*, 468(7327):1046–1048, Dec 2010. (Cited in pages 32 and 34.)

[15] P. Bernadó, L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok, and M. Blackledge. A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering. *Proceedings of the National Academy of Sciences of the U.S.A.*, 102(47):17002–17007, 2005. (Cited in pages 14, 27, 35, 38, 41, 46, 76, 77, 79, and 94.)

[16] P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun. Structural characterization of flexible proteins using small-angle X-ray scattering. *Journal of the American Chemical Society*, 129(17):5656–5664, 2007. (Cited in page 13.)

[17] P. Bernado and D. I. Svergun. Analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Methods Molecular Biology*, 896:107–122, 2012. (Cited in page 32.)

[18] P. Bernadó and D. I. Svergun. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Molecular BioSystems*, 8:151–167, 2012. (Cited in pages 13, 14, 32, and 76.)

[19] R. B. Best. Atomistic molecular simulations of protein folding. *Current Opinion in Structural Biology*, 22(1):52–61, 2012. (Cited in page 94.)

[20] R. B. Best, W. Zheng, and J. Mittal. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *Journal of Chemical Theory and Computation*, 10(11):5113–5124, 2014. (Cited in page 37.)

[21] J. Bialkowski, S. Karaman, and E. Frazzoli. Massively parallelizing the RRT and the RRT*. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3513–3518, 2011. (Cited in page 118.)

[22] M. Blanc, T. L. Coetzer, M. Blackledge, M. Haertlein, E. P. Mitchell, V. T. Forsyth, and M. R. Jensen. Intrinsic disorder within the erythrocyte binding-like proteins from plasmodium falciparum. *Biochimica et Biophysica Acta*, 1844(12):2306 – 2314, 2014. (Cited in pages 56, 58, 78, and 85.)

[23] A. Bondi. Van der Waals volumes and radii. *Journal of Physical Chemistry*, 68:441–451, 1964. (Cited in page 101.)

[24] P. Calmettes, D. Durand, M. Desmadril, P. Minard, V. Receveur, and J. C. Smith. How random is a highly denatured protein? *Biophysical Chemistry*, 53(1-2):105–113, Dec 1994. (Cited in page 34.)

[25] S. Carpin and E. Pagello. On parallel RRTs for multi-robot systems. In *Proc. International Conference of the Italian Association for Artificial Intelligence (AI*IA)*, pages 834–841, 2002. (Cited in page 118.)

[26] S. Caselli and M. Reggiani. ERPP: An experience-based randomized path planner. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 1002–1008, 2000. (Cited in page 118.)

[27] D. Challou, D. Boley, M. Gini, V. Kumar, and C. Olson. Parallel search algorithms for robot motion planning. In K. Gupta and A. P. del Pobil, editors, *Practical Motion Planning in Robotics: Current Approaches and Future Directions*, pages 115–131. John Wiley & Sons Ltd., 1998. (Cited in page 118.)

[28] Y. Chebaro, A. J. Ballard, D. Chakraborty, and D. J. Wales. Intrinsically disordered energy landscapes. *Scientific Reports*, 5:10386, May 2015. (Cited in page 36.)

[29] M. K. Cho, H. Y. Kim, P. Bernadó, C. O. Fernandez, M. Blackledge, and M. Zweckstetter. Amino acid bulkiness defines the local conformations and dynamics of natively unfolded α-synuclein and tau. *Journal of the American Chemical Society*, 129(11):3032–3033, 2007. (Cited in pages 60 and 61.)

[30] J. J. Chou, S. Gaemers, B. Howder, J. M. Louis, and A. Bax. A simple apparatus for generating stretched polyacrylamide gels, yielding uniform alignment of proteins and detergent micelles*. *Journal of Biomolecular NMR*, 21(4):377–382, 2001. (Cited in page 27.)

[31] G. M. Clore, M. R. Starich, and A. M. 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, 1998. (Cited in page 27.)

[32] T. N. Cordeiro, F. Herranz-Trillo, A. Urbanek, A. Estaña, J. Cortés, N. Sibille, and P. Bernadó. Structural characterization of highly flexible proteins by small-angle scattering. In *Biological Small Angle Scattering: Techniques,*

Strategies and Tips, pages 107–129. Springer Singapore, 2017. (Cited in page 33.)

[33] T. N. Cordeiro, F. Herranz-Trillo, A. N Urbanek, A. N Estaña, J. Cortés, N. Sibille, and P. Bernadó. Small-angle scattering studies of intrinsically disordered proteins and their complexes. *Current Opinion in Structural Biology*, 42:pp.15 – 23, 2017. (Cited in pages 14 and 83.)

[34] G. Cornilescu, J. L. Marquardt, M. Ottiger, and A. Bax. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *Journal of the American Chemical Society*, 120(27):6836–6837, 1998. (Cited in pages 77 and 78.)

[35] J. Cortés, L. Jaillet, and T. Siméon. Molecular disassembly with rrt-like algorithms. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3301–3306, 2007. (Cited in page 37.)

[36] T. E Creighton. *Proteins: structures and molecular properties*. Macmillan, 1993. (Cited in page 18.)

[37] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Research*, 14:1188–1190, 2004. (Cited in page 103.)

[38] V. Csizmok, A. V. Follis, R. W. Kriwacki, and J. D. Forman-Kay. Dynamic protein interaction networks and new structural paradigms in signaling. *Chemical Reviews*, 116(11):6424–6462, 2016. (Cited in page 13.)

[39] N. E Davey. The functional importance of structure in unstructured protein regions. *Current Opinion in Structural Biology*, 56:155 – 163, 2019. (Cited in page 54.)

[40] A. De Biasio, A. Ibáñez de Opakua, T. N. Cordeiro, M. Villate, N. Merino, N. Sibille, M. Lelli, T. Diercks, P. Bernadó, and F. J. Blanco. p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins. *Biophysical Journal*, 106(4):865 – 874, 2014. (Cited in pages 32, 33, 36, 38, 56, 60, 62, 78, 82, 86, and 90.)

[41] A. G. de Brevern. Extension of the classical classification of beta-turns. *Scientific Reports*, 6:33191, 2016. (Cited in page 85.)

[42] P. Debye. Zerstreuung von röntgenstrahlen. *Annalen der Physik*, 351(6):809–823, 1915. (Cited in page 30.)

[43] P. Debye and A. M. Bueche. Scattering by an inhomogeneous solid. *Journal of Applied Physics*, 20(6):518–525, 1949. (Cited in page 29.)

[44] M. M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, and Christopher M. Dobson. Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *Journal of the American Chemical Society*, 127(2):476–477, 2005. (Cited in pages 13 and 36.)

[45] S. DeForte and V. N. Uversky. Order, disorder, and everything in between. *Molecules*, 21(8), 2016. (Cited in page 54.)

[46] E. Delaforge, J. Kragelj, L. Tengo, A. Palencia, S. Milles, G. Bouvignies, N. Salvi, M. Blackledge, and M. R. Jensen. Deciphering the dynamic interaction profile of an intrinsically disordered protein by nmr exchange spectroscopy. *Journal of the American Chemical Society*, 140(3):1148–1158, 2018. (Cited in page 91.)

[47] X. Deng, J. Gumm, S. Karki, J. Eickholt, and J. Cheng. An overview of practical applications of protein disorder prediction and drive for faster, more accurate predictions. *International Journal of Molecular Sciences*, 16(7):15384–15404, 2015. (Cited in pages 14 and 75.)

[48] D. Devaurs, K. Molloy, M. Vaisset, A. Shehu, T. Siméon, and J. Cortés. Characterizing Energy Landscapes of Peptides using a Combination of Stochastic Algorithms. *IEEE Transactions on NanoBioscience*, 14(5):545–552, 2015. (Cited in pages 114, 116, and 119.)

[49] D. Devaurs, A. Shehu, T. Siméon, and J. Cortés. A multi-tree extension of the Transition-based RRT: Application to ordering-and-pathfinding problems in continuous cost spaces. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014. (Cited in pages 116 and 119.)

[50] D. Devaurs, T. Siméon, and J. Cortés. Enhancing the Transition-based RRT to deal with complex cost spaces. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 4105–4110, 2013. (Cited in page 122.)

[51] D. Devaurs, T. Siméon, and J. Cortés. Parallelizing RRT on large-scale distributed-memory architectures. *IEEE Transactions on Robotics*, 29(2):571–579, 2013. (Cited in pages 118, 132, and 134.)

[52] C. M. Dobson. Protein folding and misfolding. *Nature*, 426:884–890, 2003. (Cited in page 94.)

[53] S. Doniach. Changes in biomolecular conformation seen by small angle X-ray scattering. *Chemical Reviews*, 101(6):1763–1778, Jun 2001. (Cited in pages 32 and 34.)

[54] S. M. Douglas, J. J. Chou, and W. M. Shih. DNA-nanotube-induced alignment of membrane proteins for NMR structure determination. *Proceedings*

of the National Academy of Sciences, 104(16):6644–6648, 2007. (Cited in page 27.)

[55] R. Dunbrack. Rotamer libraries in the 21st century. *Current Opinion in Structural Biology*, 12:431–440, 2002. (Cited in page 94.)

[56] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–6582, 2002. (Cited in page 20.)

[57] A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva, and V. N. Uversky. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.*, 272(20):5129–5148, Oct 2005. (Cited in page 20.)

[58] H. J. Dyson and P. E. Wright. Unfolded proteins and protein folding studied by NMR. *Chemical Reviews*, 104(8):3607–3622, 2004. (Cited in pages 13 and 23.)

[59] D. Eliezer. Biophysical characterization of intrinsically disordered proteins. *Current Opinion in Structural Biology*, 19(1):23 – 30, 2009. (Cited in pages 13 and 22.)

[60] A. Emperador, P. Sfriso, M. A. Villarreal, J. L. Gelpi, and M. Orozco. PACSAB: Coarse-Grained Force Field for the Study of Protein-Protein Interactions and Conformational Sampling in Multiprotein Systems. *Journal of Chemical Theory and Computation*, 11(12):5929–5938, Dec 2015. (Cited in page 37.)

[61] S. Enemark, N. A. Kurniawan, and R. Rajagopalan. β-hairpin forms by rolling up from C-terminal: Topological guidance of early folding dynamics. *Scientific Reports*, 2(649), 2012. (Cited in pages 96, 103, 104, and 107.)

[62] H. J. Feldman and C. W. Hogue. Probabilistic sampling of protein conformations: new hope for brute force? *Proteins: Structure, Function, and Bioinformatics*, 46(1):8–23, 2002. (Cited in page 39.)

[63] H. J. Feldman and C. W.V. Hogue. A fast method to sample real protein conformational space. *Proteins: Structure, Function, and Bioinformatics*, 39(2):112–131, 2000. (Cited in page 39.)

[64] H.-P. Fink. Structure analysis by small-angle x-ray and neutron scattering. *Acta Polymerica*, 40(3):224–224, 1989. (Cited in pages 28 and 34.)

[65] C. K. Fisher, A. Huang, and C. M. Stultz. Modeling intrinsically disordered proteins with bayesian statistics. *Journal of the American Chemical Society*, 132(42):14919–14927, 2010. (Cited in page 13.)

[66] N. C. Fitzkee, P. J. Fleming, and G. D. Rose. The protein coil library: A structural database of nonhelix, nonstrand fragments derived from the pdb. *Proteins*, 58(4):852–854, 2005. (Cited in pages 14 and 76.)

[67] N. K. Fox, S. E. Brenner, and J.-M. Chandonia. SCOPe: Structural classification of proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, 42(D1):D304–D309, 2014. (Cited in page 42.)

[68] D. Frenkel and B. Smit. *Understanding Molecular Simulations: From Algorithms to Applications*, volume 1 of *Computational Science Series*. Academic Press, 2nd edition, 2001. (Cited in page 117.)

[69] J. C. Freudenberger, P. Spiteller, R. Bauer, H. Kessler, and B. Luy. Stretched poly(dimethylsiloxane) gels as NMR alignment media for apolar and weakly polar organic solvents: an ideal tool for measuring RDCs at low molecular concentrations. *Journal of the American Chemical Society*, 126(45):14690–14691, 2004. (Cited in page 26.)

[70] M. Ghallab, D. Nau, and P. Traverso. *Automated Planning: Theory and Practice*. Morgan Kaufmann Publishers, Elsevier, 2004. (Cited in pages 95 and 114.)

[71] B. Gipson, D. Hsu, L. Kavraki, and J. C. Latombe. Computational models of protein kinematics and dynamics: Beyond simulation. *Annual Review of Analytical Chemistry*, 5:273–91, 2012. (Cited in pages 37, 95, 116, and 117.)

[72] M. A. Graewert and D. I. Svergun. Impact and progress in small and wide angle x-ray scattering (saxs and waxs). *Current Opinion in Structural Biology*, 23(5):748 – 754, 2013. (Cited in page 28.)

[73] S. Granier, S. Kim, A. M. Shafer, V. R. Ratnala, J. J. Fung, R.N. Zare, and B. Kobilka. Structure and conformational changes in the c-terminal domain of the beta2-adrenoceptor: insights from fluorescence resonance energy transfer studies. *Journal of Biological Chemistry*, 282:13895–13905, 2007. (Cited in page 128.)

[74] D. Gront and A. Kolinski. Efficient scheme for optimization of parallel tempering Monte Carlo method. *Journal of Physics: Condensed Matter*, 19:036225, 2007. (Cited in page 117.)

[75] J. Gu and P. E. Bourne. *Structural bioinformatics*, volume 44. John Wiley & Sons, 2009. (Cited in page 18.)

[76] J. C. Le Guillou and J. Zinn-Justin. Critical exponents for the n-vector model in three dimensions from field theory. In J. C. LE GUILLOU and J. ZINN-JUSTIN, editors, *Large-Order Behaviour of Perturbation Theory*, volume 7

of *Current Physics–Sources and Comments*, pages 527 – 530. Elsevier, 1990. (Cited in page 34.)

[77] A. Guinier. La diffraction des rayons x aux très petits angles : application à l'étude de phénomènes ultramicroscopiques. *Annals of Physics*, 11(12):161–237, 1939. (Cited in page 31.)

[78] F. Bovey P. Mirau H. S. Gutowsky. *Nuclear Magnetic Resonance Spectroscopy*. Academic Press, 1988. (Cited in page 22.)

[79] K. F. Han and D. Baker. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proceedings of the National Academy of Sciences of the U.S.A.*, 93(12):5814–5818, 1996. (Cited in pages 41 and 89.)

[80] M. R. Hansen, L. Mueller, and A. Pardi. Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nature Structural Biology*, 5(12):1065–1074, 1998. (Cited in page 27.)

[81] A. R. Hawkins and H. K. Lamb. The molecular biology of multidomain proteins selected examples. *European Journal of Biochemistry*, 232(1):7–18, 1995. (Cited in page 20.)

[82] D. Henrich. Fast motion planning by parallel processing – a review. *Journal of Intelligent and Robotic Systems*, 20(1):45–69, 1997. (Cited in page 118.)

[83] J. Henriques, C. Cragnell, and M. Skepö. Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. *Journal of Chemical Theory and Computation*, 11(7):3420–3431, 2015. (Cited in pages 14, 37, and 75.)

[84] K. Henzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, 450:964–972, 2007. (Cited in page 20.)

[85] S. Honda, K. Yamasaki, Y. Sawada, and H. Morii. 10 residue folded peptide designed by segment statistics. *Structure*, 12(8):1507–1518, 2004. (Cited in pages 96, 102, and 103.)

[86] P. Hore. *Nuclear Magnetic Resonance*. Oxford Chemistry Primers, 2015. (Cited in page 22.)

[87] D. Hsu, R. Kindel, J. C. Latombe, and S. M. Rock. Randomized kinodynamic motion planning with moving obstacles. *The International Journal of Robotics Research*, 21(3):233–255, 2002. (Cited in page 119.)

[88] J. R. Huang, V. Ozenne, M. R. Jensen, and M. Blackledge. Direct prediction of NMR residual dipolar couplings from the primary sequence of unfolded proteins. *Angewandte Chemie International Edition*, 52(2):687–690, 2013. (Cited in pages 44, 76, and 95.)

[89] J. C. Hus, D. Marion, and M. Blackledge. Determination of protein backbone structure using only residual dipolar couplings. *Journal of chemical information and modeling*, 58:1541–1542, 2018. (Cited in page 27.)

[90] J. Ichnowski and R. Alterovitz. Scalable multicore motion planning using lock-free concurrency. *IEEE Transactions on Robotics*, 30(5):1123–1136, 2014. (Cited in pages 118 and 119.)

[91] J. Iglesias, M. Sanchez-Martínez, and R. Crehuet. SS-map: Visualizing cooperative secondary structure elements in protein ensembles. *Intrinsically Disordered Proteins*, 1(1):e25323, 2013. (Cited in page 86.)

[92] S. A. Jacobs, N. Stradford, C. Rodriguez, S. Thomas, and N. M. Amato. A scalable distributed rrt for motion planning. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 5088–5095, 2013. (Cited in page 119.)

[93] D. A. Jacques and J. Trewhella. Small-angle scattering for structural biology–expanding the frontier while avoiding the pitfalls. *Protein Science*, 19(4):642–657, Apr 2010. (Cited in page 28.)

[94] L. Jaillet, F. J. Corcho, J. J. Pérez, and J. Cortés. Randomized tree construction algorithm to explore energy landscapes. *Journal of Computational Chemistry*, 32(16):3464–3474, 2011. (Cited in pages 116, 119, and 121.)

[95] L. Jaillet, J. Cortés, and T. Siméon. Sampling-based path planning on configuration-space costmaps. *IEEE Transactions on Robotics*, 26(4):635–646, 2010. (Cited in pages 119 and 120.)

[96] S. Jaswinder, H. Jack, H. Rhys, P. Kuldip, Y. Yuedong, and Z. Yaoqi. Detecting proline and non-proline cis isomers in protein structures from sequences using deep residual ensemble learning. *Journal of the American Chemical Society*, 123:2033–2042, 2001. (Cited in page 45.)

[97] M. R. Jensen, G. Communie, E. A. Ribeiro, N. Martinez, A. Desfosses, L. Salmon, L. Mollica, F. Gabel, M. Jamin, S. Longhi, R. W. H. Ruigrok, and M. Blackledge. Intrinsic disorder in measles virus nucleocapsids. *Proceedings of the National Academy of Sciences of the U.S.A.*, 108(24):9839–9844, 2011. (Cited in pages 46, 56, 59, 78, and 90.)

[98] M. R. Jensen, K. Houben, E. Lescop, L. Blanchard, R. W. H. Ruigrok, and M. Blackledge. Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: Application to the molecular recognition element of sendai virus nucleoprotein. *Journal of the American Chemical Society*, 130(25):8055–8061, 2008. (Cited in pages 27, 39, 56, 59, 78, and 89.)

[99] M. R. Jensen, P. R. L. Markwick, S. Meier, C. Griesinger, M. Zweckstetter, S. Grzesiek, P. Bernadó, and M. Blackledge. Quantitative determination of

the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Structure*, 17(9):1169 – 1185, 2009. (Cited in pages 13, 14, 57, and 76.)

[100] M. R. Jensen, M. Zweckstetter, J. Huang, and M. Blackledge. Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using nmr spectroscopy. *Chemical Reviews*, 114(13):6632–6660, 2014. (Cited in pages 20 and 23.)

[101] M. Ringkjøbing Jensen, P. R.L. Markwick, S. Meier, C. Griesinger, M. Zweckstetter, S. Grzesiek, P. Bernadó, and M. Blackledge. Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Structure*, 17(9):1169 – 1185, 2009. (Cited in pages 23 and 27.)

[102] M. Ringkjøbing Jensen, R. WH Ruigrok, and M. Blackledge. Describing intrinsically disordered proteins at atomic resolution by NMR. *Current Opinion in Structural Biology*, 23(3):426 – 435, 2013. (Cited in page 23.)

[103] A. K. Jha, A. Colubri, K. F. Freed, and T. R. Sosnick. Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proceedings of the National Academy of Sciences of the U.S.A.*, 102(37):13099–13104, 2005. (Cited in pages 14, 27, 35, 76, 80, and 94.)

[104] Q. Jiang, X. Jin, S.J. Lee, and S. Yao. Protein secondary structure prediction: A survey of the state of the art. *Journal of Molecular Graphics and Modelling*, 76:379–402, 2017. (Cited in page 53.)

[105] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. (Cited in pages 42 and 103.)

[106] M. Kachala, E. Valentini, and D. I. Svergun. Application of SAXS for the Structural Characterization of IDPs. *Advances in Experimental Medicine and Biology*, 870:261–289, 2015. (Cited in page 32.)

[107] M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9:646–652, 2002. (Cited in pages 36 and 117.)

[108] L. E. Kavraki, P. Švestka, J. C. Latombe, and M. H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Transactions on Robotics and Automation*, 12(4):566–580, 1996. (Cited in page 118.)

[109] A. G. Kikhney and D. I. Svergun. A practical guide to small angle x-ray scattering (saxs) of flexible and intrinsically disordered proteins. *FEBS Letters*, 589(19, Part A):2570 – 2577, 2015. (Cited in page 32.)

[110] P. M. Kim, A. Sboner, Y. Xia, and M. Gerstein. The role of disorder in interaction networks: a structural analysis. *Molecular Systems Biology*, 4:179, 2008. (Cited in page 20.)

[111] M. Kjaergaard, S. Brander, and F. M. Poulsen. Random coil chemical shift for intrinsically disordered proteins: effects of temperature and pH. *Journal of Biomolecular NMR*, 49(2):139–149, Feb 2011. (Cited in page 26.)

[112] T. P. Knowles, M Vendruscolo, and C. M. Dobson. The amyloid state and its association with protein misfolding diseases. *Structure*, 15(6):384–396, 2014. (Cited in page 93.)

[113] M. H. Koch, P. Vachette, and D. I. Svergun. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Quarterly Reviews of Biophysics*, 36(2):147–227, May 2003. (Cited in pages 28 and 32.)

[114] 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. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proceedings of the National Academy of Sciences of the U.S.A.*, 101(34):12491–12496, Aug 2004. (Cited in page 34.)

[115] P. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot, and A. Pohorille. The development/application of a “minimalist" organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data. In WilfredF. van Gunsteren, PaulK. Weiner, and AnthonyJ. Wilkinson, editors, *Computer Simulation of Biomolecular Systems*, volume 3 of *Computer Simulations of Biomolecular Systems*, pages 83–96. Springer Netherlands, 1997. (Cited in page 121.)

[116] R. Kolodny, P. Koehl, L. Guibas, and M. Levitt. Small libraries of protein fragments model native protein structures accurately. *Journal of Molecular Biology*, 323(2):297 – 307, 2002. (Cited in pages 41, 89, and 94.)

[117] J. Kragelj, A. Palencia, M. H. Nanao, D. Maurin, G. Bouvignies, M. Blackledge, and M. R. Jensen. Structure and dynamics of the MKK7-JNK signaling complex. *Proceedings of the National Academy of Sciences of the U.S.A.*, 112(11):3409–3414, 2015. (Cited in pages 56, 58, and 78.)

[118] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, 77(4):778–795, 2009. (Cited in pages 77 and 78.)

[119] M. Krzeminski, Joseph A. Marsh, Chris Neale, Wing-Yiu Choy, and Julie D. Forman-Kay. Characterization of disordered proteins with ensemble. *Bioinformatics*, 29(3):398–399, 2013. (Cited in pages 13 and 39.)

[120] P. Kührová, A. De Simone, M. Otyepka, and R. B. Best. Force-field dependence of chignolin folding and misfolding: Comparison with experiment and redesign. *Biophysical Journal*, 102(8):1897–1906, 2012. (Cited in page 105.)

[121] J. C. Latombe. *Robot Motion Planning*. Kluwer Academic Publishers, 1991. (Cited in page 117.)

[122] S. M. LaValle. Rapidly-exploring Random Trees: a new tool for path planning. Technical Report TR: 98-11, Computer Science Dept., Iowa State University, 1998. (Cited in pages 37, 116, and 119.)

[123] S. M. LaValle. *Planning Algorithms*. Cambridge University Press, Cambridge, U.K., 2006. (Cited in page 117.)

[124] S. M. LaValle and J. J. Kuffner. Rapidly-exploring random trees: progress and prospects. In B. R. Donald, K. M. Lynch, and D. Rus, editors, *Algorithmic and Computational Robotics: New Directions*, pages 293–308. A. K. Peters, Wellesley, MA, 2001. (Cited in page 119.)

[125] K. H. Lee and J. Chen. Multiscale enhanced sampling of intrinsically disordered protein conformations. *Journal of Computational Chemistry*, 37(6):550–557, Mar 2016. (Cited in page 36.)

[126] C. Levinthal. How to fold graciously. *Mössbauer Spectroscopy in Biological Systems Proceedings*, pages 22–24, 1969. (Cited in page 95.)

[127] M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology*, 104(1):59 – 107, 1976. (Cited in pages 76 and 101.)

[128] M. Levitt. Nature of the protein universe. *Proceedings of the National Academy of Sciences*, 106(27):11079–11084, 2009. (Cited in page 20.)

[129] H. Liang, H. Chen, K. Fan, P. Wei, X. Guo, C. Jin, C. Zeng, C. Tang, and L. Lai. De novo design of a $\beta\alpha\beta$ motif. *Angewandte Chemie International Edition*, 48(18):3301–3303, 2009. (Cited in pages 96 and 107.)

[130] E. Lindahl, B. Hess, and D. van der Spoel. Gromacs 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual*, 7(8):306–317, 2001. (Cited in page 117.)

[131] K. Lindorff-Larsen, S. Kristjansdottir, K. Teilum, W. Fieber, C. M. Dobson, F. M. Poulsen, and M. Vendruscolo. Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. *Journal of the American Chemical Society*, 126(10):3291–3299, Mar 2004. (Cited in page 36.)

[132] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011. (Cited in pages 94, 96, and 111.)

[133] R. S. Lipsitz and N. Tjandra. Residual dipolar couplings in NMR structure analysis. *Annual Review of Biophysics and Biomolecular Structure*, 33(1):387–413, 2004. (Cited in page 26.)

[134] Y. Liu, X. Wang, and B. Liu. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Briefings in Bioinformatics*, 20(1):330–346, 2017. (Cited in page 53.)

[135] J. Lorieau, L. Yao, and A. Bax. Liquid crystalline phase of g-tetrad DNA for NMR study of detergent-solubilized proteins. *Journal of the American Chemical Society*, 130(24):7536–7537, 2008. (Cited in page 27.)

[136] M. Louhivuori, K. Pääkkönen, K. Fredriksson, P. Permi, J. Lounila, and A. Annila. On the origin of residual dipolar couplings from denatured proteins. *Journal of the American Chemical Society*, 125(50):15647–15650, 2003. (Cited in page 27.)

[137] S. C. Lovell, I. W. Davis, W. B. Arendall, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins*, 50(3):437–450, 2003. (Cited in page 89.)

[138] J. Ma, G. I. Goldberg, and N. Tjandra. Weak alignment of biomacromolecules in collagen gels: An alternative way to yield residual dipolar couplings for NMR measurements. *Journal of the American Chemical Society*, 130(48):16148–16149, 2008. (Cited in page 27.)

[139] M. W. MacArthur and J. M. Thornton. Influence of proline residues on protein conformation. *Journal of Molecular Biology*, 218(2):397 – 412, 1991. (Cited in page 79.)

[140] C. O. Mackenzie, J. Zhou, and G. Grigoryan. Tertiary alphabet for the observable protein structural universe. *Proceedings of the National Academy of Sciences of the U.S.A.*, 113(47):E7438–E7447, 2016. (Cited in pages 41 and 89.)

[141] J. A. Marsh, J. M. R. Baker, M. Tollinger, and J. D. Forman-Kay. Calculation of residual dipolar couplings from disordered state ensembles using local alignment. *Journal of the American Chemical Society*, 130(25):7804–7805, 2008. (Cited in page 27.)

[142] J. A. Marsh, C. Neale, F. E. Jack, W. Y. Choy, A. Y. Lee, K. A. Crowhurst, and J. D. Forman-Kay. Improved structural characterizations of the drkN

SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure. *Journal of Molecular Biology*, 367(5):1494–1510, 2007. (Cited in page 39.)

[143] J. Maupetit, P. Derreumaux, and P. Tufféry. A fast method for large-scale de novo peptide and miniprotein structure prediction. *Journal of Computational Chemistry*, 31(4):726–738, 2010. (Cited in page 94.)

[144] S. Meier, S. Grzesiek, and M. Blackledge. Mapping the conformational landscape of urea-denatured ubiquitin using residual dipolar couplings. *Journal of the American Chemical Society*, 129(31):9799–9807, 2007. (Cited in page 35.)

[145] H. D. Mertens and D. I. Svergun. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *Journal of Structural Biology*, 172(1):128–141, Oct 2010. (Cited in page 28.)

[146] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E.Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953. (Cited in page 36.)

[147] T. Mittag, J. Marsh, A. Grishaev, S. Orlicky, H. Lin, F. Sicheri, M. Tyers, and J. D. Forman-Kay. Structure/function implications in a dynamic complex of the intrinsically disordered sic1 with the cdc4 subunit of an {SCF} ubiquitin ligase. *Structure*, 18(4):494 – 506, 2010. (Cited in pages 60, 62, and 78.)

[148] T. Mittag, S. Orlicky, W.-Y. Choy, X. Tang, H. Lin, F. Sicheri, L. E. Kay, M. Tyers, and J. D. Forman-Kay. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proceedings of the National Academy of Sciences of the U.S.A.*, 105(46):17772–17777, 2008. (Cited in pages 39, 56, and 60.)

[149] A. Mittermaier and L. E. Kay. New tools provide new insights in nmr studies of protein dynamics. *Science*, 312(5771):224–228, 2006. (Cited in page 22.)

[150] A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker, and V. N. Uversky. Analysis of molecular recognition features (MoRFs). *Journal of Molecular Biology*, 362(5):1043 – 1059, 2006. (Cited in pages 13 and 89.)

[151] R. Mohana-Borges, N. K. Goto, G. J. Kroon, H. J. Dyson, and P. E. Wright. Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. *Journal of Molecular Biology*, 340(5):1131–1142, Jul 2004. (Cited in page 27.)

[152] K. Molloy and A. Shehu. A general, adaptive, roadmap-based algorithm for protein motion computation. *IEEE Transactions on NanoBioscience*, 15(2):158–165, 2016. (Cited in page 94.)

[153] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *In VISAPP International Conference on Computer Vision Theory and Applications*, pages 331–340, 2009. (Cited in page 126.)

[154] M. D. Mukrasch, P. Markwick, J. Biernat, M. von Bergen, P. Bernadó, C. Griesinger, E. Mandelkow, M. Zweckstetter, and M. Blackledge. Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. *Journal of the American Chemical Society*, 129(16):5235–5243, 2007. (Cited in pages 36, 38, 56, 57, 60, 61, 78, 82, and 85.)

[155] E. Mylonas, A. Hascher, P. Bernado, M. Blackledge, E. Mandelkow, and D. I. Svergun. Domain conformation of tau protein studied by solution small-angle X-ray scattering. *Biochemistry*, 47(39):10345–10353, Sep 2008. (Cited in pages 38 and 78.)

[156] R. L. Narayanan, U. H. N. Dürr, S. Bibow, J. Biernat, E. Mandelkow, and M. Zweckstetter. Automatic assignment of the intrinsically disordered protein tau with 441-residues. *Journal of the American Chemical Society*, 132(34):11906–11907, 2010. (Cited in page 23.)

[157] H. L. Nguyen, H. Khanmohammadbaigi, and E. Clementi. A parallel molecular dynamics strategy. *Journal of Computational Chemistry*, 6:634, 1985. (Cited in page 117.)

[158] J. T. Nielsen and Frans A. A. Mulder. POTENCI: prediction of temperature, neighbor and ph-corrected chemical shifts for intrinsically disordered proteins. *Journal of Biomolecular NMR*, 70(3):141–165, 2018. (Cited in page 78.)

[159] O. Kratky O. Glatter. *Small Angle X-ray Scattering*. Academic Press, 1982. (Cited in page 28.)

[160] H. Ota and S. Fukuchi. Sequence conservation of protein binding segments in intrinsically disordered regions. *Biochemical and Biophysical Research Communications*, 494(3):602 – 607, 2017. (Cited in page 91.)

[161] M. Otte and N. Correll. C-FOREST: Parallel shortest-path planning with super linear speedup. *IEEE Transactions on Robotics*, 29:798–806, 2013. (Cited in page 119.)

[162] M. Ottiger and A. Bax. *Journal of Biomolecular NMR*, 12(3):361–372, 1998. (Cited in page 27.)

[163] V. Ozenne, F. Bauer, L. Salmon, J. R. Huang, M. R. Jensen, S. Segard, P. Bernado, C. Charavay, and M. 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, 2012. (Cited in pages 14, 35, 38, 46, 76, 77, and 79.)

[164] V. Ozenne, R. Schneider, M. Yao, J. Huang, L. Salmon, M. Zweckstetter, M. R. Jensen, and M. Blackledge. Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. *Journal of the American Chemical Society*, 134(36):15138–15148, 2012. (Cited in pages 13, 43, 44, 82, 85, and 86.)

[165] G. H. Paine and H. A. Scheraga. Prediction of the native conformation of a polypeptide by a statistical-mechanical procedure. III. Probable and average conformations of enkephalin. *Biopolymers*, 26(7):1125–1162, 1987. (Cited in page 128.)

[166] J. Pan and D. Manocha. Gpu-based parallel collision detection for fast motion planning. *International Journal of Robotics Research*, 31(2):187–200, 2012. (Cited in page 118.)

[167] R. Pancsa and M. Fuxreiter. Interactions via intrinsically disordered regions: What kind of motifs? *IUBMB Life*, 64(6):513–520, 2012. (Cited in pages 13, 91, and 93.)

[168] L. Pauling and R. B. Corey. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. *Proceedings of the National Academy of Sciences*, 37(11):729–740, November 1951. (Cited in page 53.)

[169] J. Pérez and Y. Nishino. Advances in x-ray scattering: from solution saxs to achievements with coherent beams. *Current Opinion in Structural Biology*, 22(5):670 – 678, 2012. (Cited in page 28.)

[170] Y. Pérez, M. Gairí, M. Pons, and P. Bernadó. Structural characterization of the natively unfolded N-terminal domain of human c-Src kinase: Insights into the role of phosphorylation of the unique domain. *Journal of Molecular Biology*, 391(1):136 – 148, 2009. (Cited in pages 56, 64, 78, 82, and 86.)

[171] M. V. Petoukhov and D. I. Svergun. Analysis of X-ray and neutron scattering from biomacromolecular solutions. *Current Opinion in Structural Biology*, 17(5):562–571, Oct 2007. (Cited in page 28.)

[172] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005. (Cited in page 117.)

[173] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw. Water dispersion interactions strongly influence simulated structural properties of disordered

protein states. *The Journal of Physical Chemistry B*, 119(16):5113–5123, 2015. (Cited in pages 14 and 75.)

[174] S. Piana and D. E. Shaw J. L. Klepeis. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. (Cited in page 36.)

[175] E. Plaku, K. E. Bekris, B. Y. Chen, A. M. Ladd, and L. E. Kavraki. Sampling-based roadmap of trees for parallel motion planning. *IEEE Transactions on Robotics*, 21(4):597–608, 2005. (Cited in pages 118 and 119.)

[176] J. H. Prestegard, C. M. Bougault, and A. I. Kishore. Residual dipolar couplings in structure determination of biomolecules. *Chemical Reviews*, 104(8):3519–3540, 2004. (Cited in page 26.)

[177] C. D. Putnam, M. Hammel, G. L. Hura, and J. A. Tainer. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Quarterly Reviews of Biophysics*, 40(3):191–285, Aug 2007. (Cited in pages 28 and 29.)

[178] Y. Qi, Y. Huang, H. Liang, Z. Liu, and L. Lai. Folding simulations of a de novo designed protein with a $\beta\alpha\beta$ fold. *Biophysical Journal*, 98(2):321 – 329, 2010. (Cited in pages 96, 107, and 109.)

[179] R. P. Rambo and J. A. Tainer. Super-resolution in solution X-ray scattering and its applications to structural systems biology. *Annual Review of Biophysics*, 42:415–441, 2013. (Cited in page 28.)

[180] F. Rao and A. Caflisch. Replica exchange molecular dynamics simulations of reversible folding. *Journal of Chemical Physics*, 119:4035–4042, 2003. (Cited in page 117.)

[181] D. C. Rapaport. *The art of molecular dynamics simulation.* Academic Press, 2007. (Cited in page 96.)

[182] V. Receveur-Brechot and D. Durand. How random are intrinsically disordered proteins? A small angle scattering perspective. *Current Protein and Peptide Science*, 13(1):55–75, Feb 2012. (Cited in pages 13 and 32.)

[183] S. Richter and M. Westphal. The LAMA planner: Guiding cost-based anytime planning with landmarks. *Journal of Artificial Intelligence Research*, 39(1):127–177, 2010. (Cited in page 95.)

[184] C. A. Rohl, C. E.M. Strauss, K. M.S. Misura, and D. Baker. Protein structure prediction using Rosetta. In *Numerical Computer Methods, Part D*, volume 383 of *Method. Enzymol.*, pages 66 – 93. Academic Press, 2004. (Cited in pages 41, 89, and 94.)

[185] M. Rooman, Y. Dehouck, J. Kwasigroch, C. Biot, and D. Gilis. What is paradoxical about Levinthal paradox? *Journal of Biomolecular Structure & Dynamics*, 20:327–9, 01 2003. (Cited in page 95.)

[186] G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan. A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences of the U.S.A.*, 103(45):16623–16633, 2006. (Cited in page 94.)

[187] J. S. Rosenthal. Parallel computing and monte carlo algorithms. *Far East Journal of Theoretical Statistics*, 4:207–236, 2000. (Cited in page 117.)

[188] M. Rückert and G. Otting. Alignment of biological macromolecules in novel nonionic liquid crystalline media for nmr experiments. 2000. (Cited in page 27.)

[189] V. Ruiz de Angulo, J. Cortés, and J. M. Porta. Rigid-CLL: Avoiding constant-distance computations in cell linked-lists algorithms. *Journal of Computational Chemistry*, 33(3):294–300, 2012. (Cited in page 101.)

[190] H.J. Sass. *Journal of Biomolecular NMR*, 18(4):303–309, 2000. (Cited in page 27.)

[191] D. Satoh, K. Shimizu, S. Nakamura, and T. Terada. Folding free-energy landscape of a 10-residue mini-protein, chignolin. *FEBS Letters*, 580(14), 2006. (Cited in pages 102, 105, and 128.)

[192] R. Schneider, D. Maurin, G. Communie, J. Kragelj, D. F. Hansen, R. W. H. Ruigrok, Malene R. J., and M. Blackledge. Visualizing the molecular recognition trajectory of an intrinsically disordered protein using multinuclear relaxation dispersion NMR. *Journal of the American Chemical Society*, 137(3):1220–1229, 2015. (Cited in pages 13 and 91.)

[193] LLC Schrödinger. The PyMOL molecular graphics system, version 1.8. November 2015. (Cited in pages 102, 104, 106, and 108.)

[194] M. Schwalbe, V. Ozenne, S. Bibow, M. Jaremko, L. Jaremko, M. Gajda, M. R. Jensen, J. Biernat, S. Becker, E. Mandelkow, M. Zweckstetter, and M. Blackledge. Predictive atomic resolution descriptions of intrinsically disordered hTau40 and α-synuclein in solution from NMR and small angle scattering. *Structure*, 22(2):238–249, 2014. (Cited in pages 39, 56, 57, 60, 78, and 82.)

[195] S. Schwarzinger, G. J. Kroon, T. R. Foss, P. E. Wright, and H. J. Dyson. Random coil chemical shifts in acidic 8 M urea: implementation of random coil shift data in NMRView. *Journal of Biomolecular NMR*, 18(1):43–48, Sep 2000. (Cited in page 26.)

[196] S. Schwarzinger, G. J. A. Kroon, T. R. Foss, J. Chung, P. E. Wright, and H. J. Dyson. Sequence-dependent correction of random coil nmr chemical

shifts. *Journal of the American Chemical Society*, 123(13):2970–2978, 2001. (Cited in page 90.)

[197] R. Schweitzer-Stenner and S. E. Toal. Construction and comparison of the statistical coil states of unfolded and intrinsically disordered proteins from nearest-neighbor corrected conformational propensities of short peptides. *Molecular BioSystems*, 12:3294–3306, 2016. (Cited in page 80.)

[198] R. A. Scott and H. A Scheraga. Conformational analysis of macromolecules. III. Helical structures of polyglycine and poly-L-alanine. *The Journal of Chemical Physics*, 45(6):2091–2101, 1966. (Cited in page 42.)

[199] A. Shehu. Probabilistic search and optimization for protein energy landscapes. In S. Aluru and M. Singh, editors, *Handbook of Computational Molecular Biology*, Computer & Information Science Series. Chapman & Hall/CRC, 2nd edition, 2013. in press. (Cited in pages 37, 116, and 117.)

[200] A. Shehu and E. Plaku. A survey of computational treatments of biomolecules by robotics-inspired methods modeling equilibrium structure and dynamic. *Journal of Artificial Intelligence Research*, 57:509–572, 2016. (Cited in page 95.)

[201] Y. Shen and A. Bax. SPARTA+: a modest improvement in empirical nmr chemical shift prediction by means of an artificial neural network. *Journal of Biomolecular NMR*, 48(1):13–22, 2010. (Cited in pages 45, 78, and 82.)

[202] Y. Shen, J. Maupetit, P. Derreumaux, and P. Tufféry. Improved PEP-FOLD approach for peptide and miniprotein structure prediction. *Journal of Chemical Theory and Computation*, 10(10):4745–4758, 2014. (Cited in pages 41 and 89.)

[203] Y. Shen, J. Roche, A. Grishaev, and A. Bax. Prediction of nearest neighbor effects on backbone torsion angles and nmr scalar coupling constants in disordered proteins. *Protein Science*, 27(1):146–158, 2018. (Cited in pages 14 and 76.)

[204] Y. Shen, R. Vernon, D. Baker, and A. Bax. De novo protein structure generation from incomplete chemical shift assignments. *Journal of biomolecular NMR*, 43(2), 2009. (Cited in page 26.)

[205] G. N. Shilstone, C. Zannoni, and C. A. Veracini. Solute alignment in liquid crystal solvents the saupe ordering matrix for perylene and pyrene. *Liquid Crystals*, 6(3):303–317, 1989. (Cited in page 26.)

[206] A. K. Shukla, G. H. Westfield, K. Xiao, et al. Visualization of arrestin recruitment by a G protein-coupled receptor. *Nature*, 512:218–222, 2014. (Cited in page 128.)

[207] N. Sibille and P. Bernadó. Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS. *Biochemical Society Transactions*, 40(5):955–962, 2012. (Cited in pages 14 and 83.)

[208] J. Silvestre-Ryan, C. W. Bertoncini, R. Fenwick, S. Esteban-Martin, and X. Salvatella. Average conformations determined from pre data provide high-resolution maps of transient tertiary interactions in disordered proteins. *Biophysical Journal*, 104(8):1740 – 1751, 2013. (Cited in page 13.)

[209] 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. *Journal of Molecular Biology*, 255(3):494 – 506, 1996. (Cited in pages 14, 76, and 94.)

[210] C. Snow, B. Zagrovic, and V. Pande. The Trp-cage: Folding kinetics and unfolded state topology via molecular dynamics simulations. *Journal of the American Chemical Society*, 124:14548–9, 01 2003. (Cited in page 96.)

[211] P. Sormanni, C. Camilloni, P. Fariselli, and M. Vendruscolo. The s2D method: Simultaneous sequence-based prediction of the statistical populations of ordered and disordered regions in proteins. *Journal of Molecular Biology*, 427(4):982–996, 2015. (Cited in pages 53 and 62.)

[212] P. Sormanni, D. Piovesan, G. T. Heller, M. Bonomi, P. Kukic, C. Camilloni, M. Fuxreiter, Z. Dosztanyi, R. V. Pappu, M. Madan Babu, S. Longhi, P. Tompa, A. D., V. N. Uversky, S. C. E. Tosatto, and M. Vendruscolo. Simultaneous quantification of protein order and disorder. *Nature Chemical Biology*, 13(4):339–342, 2017. (Cited in pages 54 and 62.)

[213] Y. G. Sterckx, A. N. Volkov, W. F. Vranken, J. Kragelj, M. R. Jensen, L. Buts, A. Garcia-Pino, T. Jove, L. Van Melderen, M. Blackledge, N. A. van Nuland, and R. Loris. Small-angle X-ray scattering- and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin PaaA2. *Structure*, 22(6):854–865, 2014. (Cited in pages 36 and 38.)

[214] I. Strid. Efficient parallelisation of metropolis-hastings algorithms using a prefetching approach. *Computational Statistics & Data Analysis*, 54:2814–2835, 2009. (Cited in page 117.)

[215] K. Sugase, H. J. Dyson, and P. E. Wright. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*, 447:1021, 2007. (Cited in page 91.)

[216] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letter*, 314:141–151, 1999. (Cited in page 117.)

[217] D. Svergun, C. Barberato, and M. H. J. Koch. CRYSOL – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of Applied Crystallography*, 28(6):768–773, 1995. (Cited in page 78.)

[218] D. I. Svergun and L. A. Feigin. *Structure Analysis by Small-Angle X-Ray and Neutron Scattering*. Springer Science Business Media, 1987. (Cited in page 28.)

[219] D. I. Svergun and M. H. J. Koch. Small-angle scattering studies of biological macromolecules in solution. *Reports on Progress in Physics*, 2003. (Cited in page 28.)

[220] R. H. Swendsen and J. Wang. Replica monte carlo simulation of spin-glasses. *Physical Review Letters*, 57:2607–2609, 1986. (Cited in page 117.)

[221] K. Tamiola, B. Acar, and F. A. Mulder. Sequence-specific random coil chemical shifts of intrinsically disordered proteins. *Journal of the American Chemical Society*, 132(51):18000–18003, Dec 2010. (Cited in pages 26 and 90.)

[222] T. Terakawa and S. Takada. Multiscale ensemble modeling of intrinsically disordered proteins: p53 N-terminal domain. *Biophysical Journal*, 101(6):1450–1458, Sep 2011. (Cited in page 36.)

[223] D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, and Roland L. Dunbrack, Jr. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. *PLOS Computational Biology*, 6(4):1–21, 04 2010. (Cited in pages 14, 76, and 79.)

[224] N. Tjandra, A. Szabo, and A. Bax. Protein backbone dynamics and 15n chemical shift anisotropy from quantitative measurement of relaxation interference effects. *Journal of the American Chemical Society*, 118(29):6986–6991, 1996. (Cited in page 26.)

[225] J. R. Tolman and K. Ruan. NMR residual dipolar couplings as probes of biomolecular dynamics. *Chemical Reviews*, 106(5):1720–1736, 2006. (Cited in page 27.)

[226] P. Tompa, N. E. Davey, T. J Gibson, and M M. Babu. A million peptide motifs for the molecular biologist. *Molecular Cell*, 55(2):161–169, 2014. (Cited in page 93.)

[227] P. Tompa, E. Schad, A. Tantos, and L. Kalmar. Intrinsically disordered proteins: emerging interaction specialists. *Current Opinion in Structural Biology*, 35:49 – 59, 2015. (Cited in pages 13, 20, 89, and 93.)

[228] R. Tycko, F. J. Blanco, and Y. Ishii. Alignment of biopolymers in strained gels: a new way to create detectable dipole-dipole couplings in high-resolution

biomolecular NMR. *Journal of the American Chemical Society*, 122(38):9340–9341, 2000. (Cited in page 27.)

[229] V. N. Uversky, C. J. Oldfield, and A. K. Dunker. Intrinsically disordered proteins in human diseases: Introducing the D2 concept. *Annual Review of Biophysics*, 37(1):215–246, 2008. (Cited in pages 13 and 22.)

[230] J. S. Valastyan and S. Lindquist. Mechanisms of protein-folding diseases at a glance. *Disease Models & Mechanisms*, 7(1):9–14, 2014. (Cited in page 93.)

[231] R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. Keith Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, and M. Madan Babu. Classification of intrinsically disordered regions and proteins. *Chemical Reviews*, 114(13):6589–6631, 2014. (Cited in page 21.)

[232] K. Van Roey, B. Uyar, R. J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T. J. Gibson, and N. E. Davey. Short linear motifs: Ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chemical Reviews*, 114(13):6733–6778, 2014. (Cited in pages 13 and 89.)

[233] M. Vendruscolo, J. Zurdo, C. Macphee, and C. M Dobson. Protein folding and misfolding: A paradigm of self-assembly and regulation in complex biological systems. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 361:1205–22, 07 2003. (Cited in page 93.)

[234] A. Vitalis and R. V. Pappu. ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions. *Journal of Computational Chemistry*, 30(5):673–699, 2009. (Cited in page 37.)

[235] A. Vitalis and R. V. Pappu. Methods for Monte Carlo simulations of biomacromolecules. *Annual Reports in Computational Chemistry*, 5:49–76, Jan 2009. (Cited in page 36.)

[236] D. Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge University Press, 2003. (Cited in page 95.)

[237] F. T. Wall. Principles of polymer chemistry. *Science*, 119(3095):555–556, 1954. (Cited in page 34.)

[238] M. Wells, H. Tidow, T. J. Rutherford, P. Markwick, M. R. Jensen, E. Mylonas, D. I. Svergun, M. Blackledge, and A. R. Fersht. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proceedings of the National Academy of Sciences of the U.S.A.*, 105(15):5762–5767, Apr 2008. (Cited in pages 36, 38, and 86.)

[239] D. S. Wishart, C. G. Bigam, A. Holm, R. S. Hodges, and B. D. Sykes. 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *Journal of Biomolecular NMR*, 5(1):67–81, Jan 1995. (Cited in page 26.)

[240] D. S. Wishart and B. D. Sykes. The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. *Journal of Biomolecular NMR*, 4(2):171–180, Mar 1994. (Cited in page 26.)

[241] P. Wolynes, J. N. Onuchic, and D Thirumalai. Navigating the folding routes. *Science*, 267:1619–1620, 1995. (Cited in page 94.)

[242] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293(2):321 – 331, 1999. (Cited in pages 19 and 20.)

[243] P. E. Wright and H. J. Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, 16(1):18–29, Jan 2015. (Cited in pages 20 and 23.)

[244] K. P. Wu, D. S. Weinstock, C. Narayanan, R. M. Levy, and J. Baum. Structural reorganization of alpha-synuclein at low pH observed by NMR and REMD simulations. *Journal of Molecular Biology*, 391(4):784–796, Aug 2009. (Cited in page 36.)

[245] H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky, and Z. Obradovic. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *Journal of Proteome Research*, 6(5):1882–1898, 2007. (Cited in page 13.)

[246] G. H. Zerze, C. M. Miller, D. Granata, and J. Mittal. Free energy surface of an intrinsically disordered protein: Comparison between temperature replica exchange molecular dynamics and bias-exchange metadynamics. *Journal of Chemical Theory and Computation*, 11(6):2776–2782, 2015. (Cited in page 36.)

[247] H. X. Zhou. Polymer models of protein stability, folding, and interactions. *Biochemistry*, 43(8):2141–2154, 2004. (Cited in page 35.)

[248] J. M. Zimmerman, N. Eliezer, and R. Simha. Characterization of amino acid sequences in proteins by statistical methods. *Journal of Theoretical Biology*, 21(X):170 – 201, 1968. (Cited in page 61.)

Chapter 9

Annex-1

Articles produced during the PhD

- Disordered tether from viral oncoprotein organizes molecular ambush of eukaryotic cell cycle. *(In preparation)*

  Nicolas S. Gonzalez-Foutel, Wade M. Borcherds, Alex S. Holehouse, Juliana Glavina, Susana Barrera-Vilarmau, Amin Sagar, Alejandro Estaña, Amelie Barozet, Gregorio Fernandez Ballester, Gonzalo de Prat-Gay, Ignacio E. Sanchez, Juan Cortes, Pau Bernado, Rohit V. Pappu, Gary W. Daughdrill and Lucía B. Chemes

- Predicting secondary structure propensities in IDPs using simple statistics from three-residue fragments. *(In preparation)*

  Alejandro Estaña, Amelie Barozet, Assia Mouhan, Nathalie Sibille, Pau Bernadó, Juan Cortés.

- Atomistic evidence of the reduced abundance of proline cis conformation in protein poly-proline tracts. *(In preparation)*

  Annika Urbanek, Matija Popovic, Carlos A. Elena-Real, Anna Morató, Alejandro Estaña, Aurélie Fournet, Frédéric Allemand, Ana M. Gil, Carlos Cativiela, Juan Cortés, Ana I. Jiménez, Nathalie Sibille, Pau Bernadó.

- Flanking regions define the conformation of the poly-glutamine homorepeat in huntingtin through opposite structural mechanisms. *(Submitted)*

  Annika Urbanek, Matija Popovic, Anna Morató, Alejandro Estaña, Carlos A. Elena-Real, Pablo Mier, Aurélie Fournet, Frédéric Allemand, Stephane Delbecq, Miguel A. Andrade-Navarro, Juan Cortés, Nathalie Sibille, Pau Bernadó.

- Realistic Ensemble Models of Intrinsically Disordered Proteins Using a Structure-Encoding Coil Database. *Structure*, 27 (2), 381–391, 2018.

  Alejandro Estaña, Nathalie Sibille, Elise Delaforge, Marc Vaisset, Juan Cortés, Pau Bernadó.

- Investigating the Formation of Structural Elements in Proteins Using Local Sequence-Dependent Information and a Heuristic Search Algorithm. *Molecules*, 24(6), 1150, 2019.

  Alejandro Estaña, Malik Ghallab, Pau Bernadó, Juan Cortés.

- Hybrid parallelization of a multi-tree path search algorithm: Application to highly-flexible biomolecules. *Parallel Computing*, 77, 84–100, 2018.

  Alejandro Estaña, Kevin Molloy, Marc Vaisset, Nathalie Sibille, Thierry Simeon, Pau Bernadó, Juan Cortés.

- Structural Characterization of Highly Flexible Proteins by Small-Angle Scattering. In *Biological Small Angle Scattering: Techniques, Strategies and Tips*, pages 107–129. Springer Singapore, 2017.

  Tiago Cordeiro, Fatima Herranz-Trillo, Annika Urbanek, Alejandro Estaña, Juan Cortés, Nathalie Sibille, Pau Bernadó.

- Small-angle scattering studies of intrinsically disordered proteins and their complexes. *Current Opinion in Structural Biology*, 42:pp.15 – 23, 2017.

  Tiago Cordeiro, Fátima Herranz-Trillo, Annika Urbanek, Alejandro Estaña, Juan Cortés, Nathalie Sibille and Pau Bernadó.

Chapter 10

Annex-2

Articles included in the manuscript

- Flanking regions define the conformation of the poly-glutamine homorepeat in huntingtin through opposite structural mechanisms. *Submitted*

  Annika Urbanek, Matija Popovic, Anna Morató, Alejandro Estaña, Carlos A. Elena-Real, Pablo Mier, Aurélie Fournet, Frédéric Allemand, Stephane Delbecq, Miguel A. Andrade-Navarro, Juan Cortés, Nathalie Sibille, Pau Bernadó.

- Small-angle scattering studies of intrinsically disordered proteins and their complexes. *Current Opinion in Structural Biology*, 42:pp.15 – 23, 2017.

  Tiago Cordeiro, Fátima Herranz-Trillo, Annika Urbanek, Alejandro Estaña, Juan Cortés, Nathalie Sibille and Pau Bernadó.

Flanking regions define the conformation of the poly-glutamine homo-repeat in huntingtin through opposite structural mechanisms

Annika Urbanek[1,#], Matija Popovic[1,#], Anna Morató[1], Alejandro Estaña[1,2], Carlos A. Elena-Real[1], Pablo Mier[3], Aurélie Fournet[1], Frédéric Allemand[1], Stephane Delbecq[4], Miguel A. Andrade-Navarro[3], Juan Cortés[2], Nathalie Sibille[1], Pau Bernadó[1,*]

[1] Centre de Biochimie Structurale (CBS), INSERM, CNRS, Université de Montpellier. 29, rue de Navacelles, 34090 Montpellier. France.

[2] LAAS-CNRS, Université de Toulouse, CNRS, 31400 Toulouse, France.

[3] Institute of Organismic and Molecular Evolution, Johannes Gutenberg University of Mainz, Mainz, Germany.

[4] Laboratoire de Biologie Cellulaire et Moléculaire (LBCM-EA4558 Vaccination Antiparasitaire), UFR Pharmacie, Université de Montpellier, Montpellier, France.

# These authors contributed equally to this work

Corresponding Author: Pau Bernadó (pau.bernado@cbs.cnrs.fr)

Abstract

The poly-Q homo-repeat in the N-terminal region of huntingtin (httex1) is the causative agent of Huntington's disease, a neurodegenerative pathology arising when the number of consecutive glutamines exceeds 35. Httex1 poly-Q is flanked by a 17-residue-long fragment (N17) and a proline-rich region (PRR), which have been shown to promote and inhibit the aggregation propensity of the protein, respectively, by poorly understood mechanisms. Based on experimental data obtained from site-specifically labeled NMR samples, we derived an ensemble model of httex1 containing 16 consecutive glutamines that presents an equilibrium of lowly populated α-helices of different lengths extending towards the poly-Q tract. The model identified both flanking regions as opposing poly-Q secondary structure promoters. While N17 triggers helicity through a promiscuous hydrogen bond network involving the side chains of the first glutamines in the poly-Q tract, the PRR promotes extended conformations in its neighboring glutamines. Computational analyses confirmed that these opposed conformational influences dictate the structure of the poly-Q tract in a position-dependent manner. Furthermore, a bioinformatics analysis of the human proteome showed that these structural traits are present in many human glutamine-rich proteins and that they are more prevalent in proteins with longer poly-Q tracts. Taken together, these observations provide the structural bases to understand previous biophysical and functional data on httex1.

Introduction

Huntington's disease (HD) is one of nine hereditary neurodegenerative disorders caused by an expansion of CAG triplet repeats beyond a pathological threshold. For HD, this expansion is located in the first exon of the huntingtin gene and results in an abnormally long poly-glutamine (poly-Q) tract within the N-terminus of the huntingtin protein (httex1)[1]. When the number of consecutive glutamines exceeds 35, the resulting mutant protein forms large cytoplasmic and nuclear aggregates, a hallmark of HD, and causes neuronal degeneration, especially affecting the neurons of the striatum[2–5]. Aggregation, disease risk and age of onset correlate with the length of the poly-Q tract[1,2]. Interestingly, the aggregates predominantly contain mutant httex1 fragments, instead of the full-length protein, which comprises 3,142 amino acids in the non-pathogenic form. Indeed, it has been shown that the httex1 fragment alone is enough to reproduce the HD symptoms in mice[6].

While the httex1 aggregation mechanism and the resulting β-sheet amyloid fibrils have been thoroughly characterized[7–12], the structural bases of the pathological threshold and the mechanisms by which the native form of mutant httex1 give rise to toxicity and cell death are still poorly understood. Some clues regarding aggregation and pathogenicity of mutant httex1 have been found in the flanking regions of the poly-Q tract. The N-terminal domain, composed of 17 residues (N17) (Figure 1a), enhances aggregation of longer poly-Q tracts *in vitro* and *in vivo* and has been shown to form an amphipathic helix that interacts with membranes and chaperones[11–17]. Moreover, post-translational modifications of N17 modulate huntingtin function, translocation, aggregation, and toxicity[18–23]. The poly-Q region is followed by a poly-proline (poly-P) tract of 11 consecutive prolines, which is part of the proline-rich region (PRR) containing 31 prolines in total (Figure 1a). In contrast to N17, the poly-P tract has a protective effect against aggregation *in vitro* and *in vivo,* but is necessary for the formation of visible aggregates in cells[12,18,24,25]. This effect is directional, as N-terminal poly-P tracts do not attenuate the aggregation of poly-Q peptides[24]. It has also been shown that the flanking regions differently shape the aggregation pathways of pathological httex1, define the structure and stability of fibrils, and modulate its neuronal toxicity[12].

Two models linking poly-Q abnormal expansion and cytotoxicity have been proposed[26]. The 'toxic structure' model proposes the appearance of a distinct toxic conformation when the tract expands beyond the pathological threshold[27–29]. The second model, the so-called 'linear lattice' model, suggests that even short poly-Qs are inherently toxic and httex1 toxicity systematically increases with the tract length[30–32]. Evidence for both models has been obtained using monoclonal antibodies in cells expressing httex1 of different lengths[29–33]. However, this approach provides a very indirect perspective on httex1 conformations, and higher resolution information is required to discriminate between both hypotheses[26].

In a recent study, combining single-molecule FRET (smFRET) data with atomistic simulations, no sharp conformational change of monomeric httex1 around the pathological threshold could be observed, but rather a continuous global compaction with increasing poly-Q length induced by the

interaction between N17 and the poly-Q tract was suggested[34,35]. Recent circular dichroism (CD) and electronic paramagnetic resonance (EPR) experiments report on a systematic increase of the helical propensity and rigidity in httex1 when the poly-Q tract length increases[36,37]. Observations from these *in vitro* studies are in coherence with the 'linear lattice' model. However, they only focused on the overall properties of the protein and could not probe httex1 at atomic resolution. Nuclear magnetic resonance (NMR) is the most suitable technique to provide a high-resolution picture of the conformational preferences of flexible proteins and structural characteristics of subpopulations of toxic conformers[38]. However, NMR studies of httex1 are inherently challenging due to its strong compositional bias, which impedes residue-specific assignment and the measurements of structural constraints. Due to this challenge, only incomplete observations regarding the conformational preferences of the poly-Q and the flanking regions have been reported[15,35,36]. All these NMR studies, independently of the poly-Q length, indicate a transient helical propensity encompassing N17 and the homo-repeat. Current structural models of httex1 suggest a compact overall arrangement in which N17 and the poly-Q tract interact through fuzzy contacts while the PRR sticks out. These tadpole-like structures display a systematic increase of the surface area with the length of the tract, also in line with the 'linear lattice' toxicity model[34,35]. However, these models are based on sparse data or single conformation structural modeling.

In order to overcome the previously mentioned challenges, we have recently developed a methodology to site-specifically incorporate a single [^{15}N, ^{13}C]-labeled glutamine into proteins, and thereby obtain simplified NMR spectra[39]. By systematically applying this site-specific isotopic labeling (SSIL) strategy, which combines cell-free protein expression[40] and nonsense suppression[41], we have obtained the NMR assignment at nearly physiological conditions of all non-proline residues in a httex1 construct containing 16 consecutive glutamines (H16). The ensemble modeling of the resulting chemical shifts demonstrated the presence of multiple, partially formed α-helical regions initiated in N17 and involving fragments of the poly-Q tract of different lengths. The application of SSIL to N17 and PRR mutants demonstrated that the distinct conformational features of both flanking regions are propagated into the poly-Q tract, which acts as a conformationally versatile polypeptide. These observations provide the structural determinants underlying the key role of flanking regions in modulating the aggregation properties of httex1[9,24].

Results

Glutamine NMR scanning of H16

The monomeric httex1 that we characterized, H16, contained 16 glutamines in the poly-Q tract and another six in the PRR (Figure 1a). We produced H16 samples with glutamine-specific isotopic labeling using the SSIL strategy previously developed in our group[39]. To streamline the preparation of the 22 H16 NMR samples, we first made sure that all samples could be prepared with similar

efficiency by scanning all the TAG-mutated H16-sfGFP plasmids in a 96-well plate after addition of 10 µM glutamine loaded $tRNA_{CUA}$ (Figure 1b). All positions showed fluorescence intensities of ~30% of the positive control (H16 without amber stop codon), indicating that the efficiency of the incorporation of the labeled glutamine is independent of the specific sequence and the yield is similar to those achieved in other studies[42,43]. Once the suppression efficiency was verified at a small scale, the CF reaction volume was increased to 5 mL to produce the NMR samples.

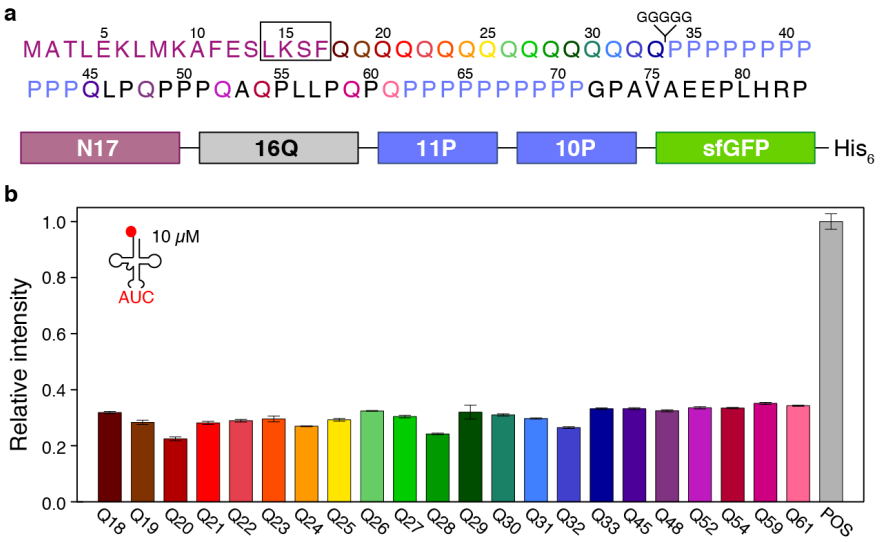


Figure 1. Glutamine SSIL scanning of H16. (**a**) Sequence of H16 and scheme of the sfGFP-fused construct used in this study. The color code identifies the individual glutamines throughout the study. The box encompassing residues ^{14}LKSF^{17} identifies the residues mutated to probe the structural connection between N17 and the poly-Q tract. The position of the insertion of glycines between the poly-Q and the PPR to structurally disconnect both regions is also displayed. (**b**) A scan probing the suppression efficiency using 10 µM loaded $tRNA_{CUA}$ showed no strong position-specific effects. The experiments were repeated three times.

The ^{15}N-HSQC of H16 displayed the typical features of poly-Q-containing proteins[35,36,44,45]. While peaks from N17 and the PRR are well dispersed, a large density of unresolved peaks corresponding to glutamine residues was observed (Figure 2a). In order to disentangle this massive overlap we measured ^{15}N- and ^{13}C-HSQCs of the SSIL H16 samples containing a single [^{15}N, ^{13}C]-labeled glutamine. As observed in Figure 2b, the glutamines adjacent to N17 (Q18-Q21) appear in the upfield region of the poly-Q density without any specific trend. The following glutamines (Q22-Q28) display a consistent ^{1}H and ^{15}N downfield shift, indicating a systematic structural change along the homo-repeat. A large deshielding effect is subsequently observed for Q29, Q30 and Q31, which are strongly overlapped. Finally, the last two glutamines of the tract, Q32 and Q33, display isolated peaks induced by the proximity of the downstream poly-P. The chemical shifts of glutamines in the PRR are more dispersed due to their different neighboring residues. Cα-Hα correlations measured in the same SSIL samples follow similar trends along the poly-Q tract (Figure 2c).

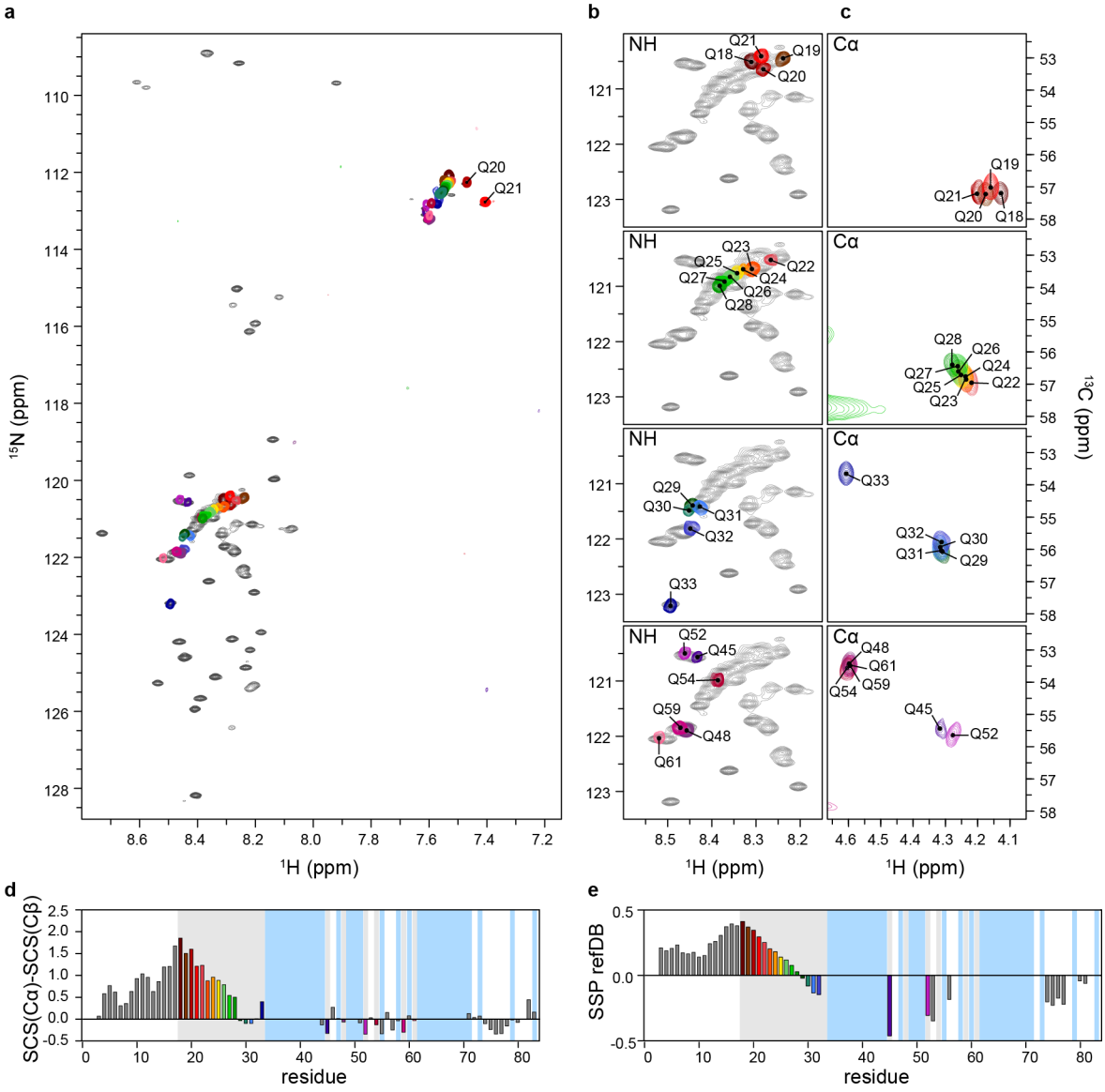


Figure 2. NMR analysis of H16. (**a**) Overlay of fully labeled H16 (grey) with individually colored SSIL ^{15}N-HSQC spectra. (**b**) Zoomed ^{15}N-HSQC overlay showing the poly-Q region with different glutamine clusters (Q18-Q21; Q22-Q28; Q29-Q33; and PRR glutamines). (**c**) Zoomed ^{13}C-HSQC overlay showing the poly-Q region with the same glutamine clusters as in (**b**). (**d**) Secondary chemical shift analysis of H16 using experimental Cα and Cβ chemical shifts and a neighbor-corrected random-coil library[46] and (**e**) secondary structure propensity plot[47,48]. The positions of glutamine and proline residues in the primary sequence are highlighted in grey and blue, respectively. Prolines and residues followed by prolines were not considered in the SSP refDB analysis.

α-helical propensity in N17 and the poly-Q tract

The Cα and Cβ chemical shifts measured for all glutamines in this study and the previously reported assignment of H16[39] allowed the determination of the structural propensities of H16. The secondary

chemical shift (SCS) analysis using a neighbor-corrected random coil database[46] indicates that both N17 and the poly-Q tract are enriched in α-helical conformations, although this propensity is not homogeneous (Figure 2d). Helicity increases along N17, reaching its maximum at the first glutamine, Q18, and subsequently decreases smoothly. A transition is observed at Q29, which adopts a small and negative SCS value. This extends to the following three glutamines, indicating the presence of random coil or slightly extended conformations. This conformational transition is pinpointed in the secondary structure propensity (SSP) analysis[47,48] (Figure 2e). Note that the helical propensity of the N-terminal part of H16 remains below 40%, in agreement with similar analyses using an httex1 fragment with 17 glutamines and the partially assigned httex1 with 25 glutamines[35,44]. The C-terminal region of H16 presents negative SCS values, probably reflecting the enrichment in polyproline-II conformations induced by the large number of prolines[7].

The ensemble model of H16 reveals a conformational equilibrium involving multiple α-helices

The ensemble structure of H16 was investigated by combining the backbone NMR chemical shifts and a recently developed approach to build realistic ensemble models of intrinsically disordered proteins[49]. Briefly, our method appends residues, which are considered to be either fully disordered or partially structured, to build the complete chain without steric clashes. For fully disordered residues, amino acid specific ϕ/ψ angles defining the residue conformation are randomly selected from the database, disregarding their flanking residues. For partially structured residues, the nature and the conformation of the flanking residues are taken into account when selecting the conformation of the incorporated residue (see detailed explanation of the algorithm in the original publication[49]). Two families of ensembles were built to investigate the conformational influence of both flanking regions of H16. For the first family (N→C ensembles), starting with the ^{10}AFESLKSF^{17} region of N17 as partially structured, multiple ensembles of 5,000 conformations were built by successively including an increasing number of glutamines in the poly-Q tract (from Q18 to Q33) as partially structured, while the rest of the chain was considered to be fully disordered. Note that in the partially structured building strategy secondary structural elements are propagated due to the neighboring effects. An equivalent strategy was followed for the second family of ensembles (N←C ensembles) for which glutamines were considered successively as partially structured from the poly-P tract (from Q33 to Q18). For the resulting 17 ensembles of each family, and after building the side chains with the program SCWRL4[50], averaged Cα and Cβ chemical shifts were computed with SPARTA+[51] and compared with the experimental ones (Figure S1).

Theoretical Cα chemical shifts for the poly-Q tract present different values for regions built as partially structured (influenced by the flanking regions) or disordered. Three Cα CS plateaus are observed corresponding to α-helical, extended and random coil conformations, and transitions are observed between regions built as disordered and influenced by the flanking regions (Figure S1). Not

surprisingly, the Cβ chemical shifts turned out to be less sensitive to the presence of structured regions in the homo-repeat region (Figure S1). These simulations indicate that flanking regions induce a distinct conformational bias to the neighboring glutamines. While N17 induces helical conformations with Cα chemical shift values larger than those usually observed for a random coil (N→C ensembles), the poly-P tract enriches the ensemble with extended conformations with smaller Cα chemical shift values compared to a random coil (N←C ensembles). However, the simulated conformational ensembles fail to reproduce the chemical shifts measured in H16, indicating that our simple sampling strategy cannot simultaneously describe the structural influence exerted by both flanking regions.

A third ensemble model of H16 was built by reweighting the populations of the pre-computed ensembles, using the experimental Cα and Cβ chemical shifts as constraints. In order to capture the influence of the flanking regions, glutamines within the tract were divided into two groups: those influenced by N17 and those influenced by the poly-P tract, whose chemical shifts were fitted with the N→C and N←C ensembles, respectively. The limit between both families was systematically explored, reaching an optimal description of the experimental chemical shifts when Q28 was chosen as the last residue structurally connected with N17 (Figure S1). Importantly, the optimization, which was performed through a Monte-Carlo procedure, was repeated multiple times always yielded equivalent populations. The resulting ensemble nicely described the complete Cα and Cβ CS profiles for H16 (Figure S2). Importantly, the systematic decrease of the Cα chemical shifts along the poly-Q tract and the flat profile observed for the Cβ chemical shifts were well reproduced, indicating that the refined ensemble captures the structural features of the homo-repeat and the distinct conformational perturbations exerted by both flanking regions.

The conformational properties of the optimized ensemble were subsequently investigated in detail. First, we explored the conformational preferences of individual glutamines using Ramachandran plots (Figures 3 and S3). While the first four glutamines of the tract (Q18-Q21) displayed a strong enrichment in helical conformations (Figure 3a), the last four (Q30-Q33) preferred extended ones (Figure 3b). The conformational preferences along the tract, calculated from the derived ensemble, indicate a systematic decrease in the helical population from ~65% (Q18) to ~50% (Q28) (Figure 3c). In line with the NMR measurements (Figure 2), a sharp conformational transition is observed for Q29, which is the first residue displaying a preference for extended conformations.

The cooperativity between the residue-specific conformations to form stable α-helices was analyzed using the secondary structure map (SS-map) tool[53]. The fragment encompassing N17 and the poly-Q tract can be described as a complex equilibrium of multiple co-existing α-helices of variable length (Figure 3d). The core of this family of helical structures includes the last four residues of N17 and the first two glutamines of the homo-repeat. The last residues of N17 act as nucleation points for the helices that afterward extend to include a variable number of glutamines of the tract, giving a triangular shape to the SS-map. According to our analysis, no α-helices are nucleated within the poly-Q tract and, as a consequence, helices involving inner glutamines belong only to lowly populated long

helical elements. This is shown in Figures 3e and S4, which display representative conformations and the α-helical fragments of the four sub-ensembles selected to describe the NMR CSs. Three of these ensembles present α-helices that encompass the last residues of N17 and the first residues of the poly-Q. No persistent turns in the residues connecting both domains are observed, which would otherwise yield a strong signature in the chemical shift profile. As a consequence, H16 should be considered as an elongated flexible particle, in contrast to the previously proposed compact tadpole-like model[34,35].

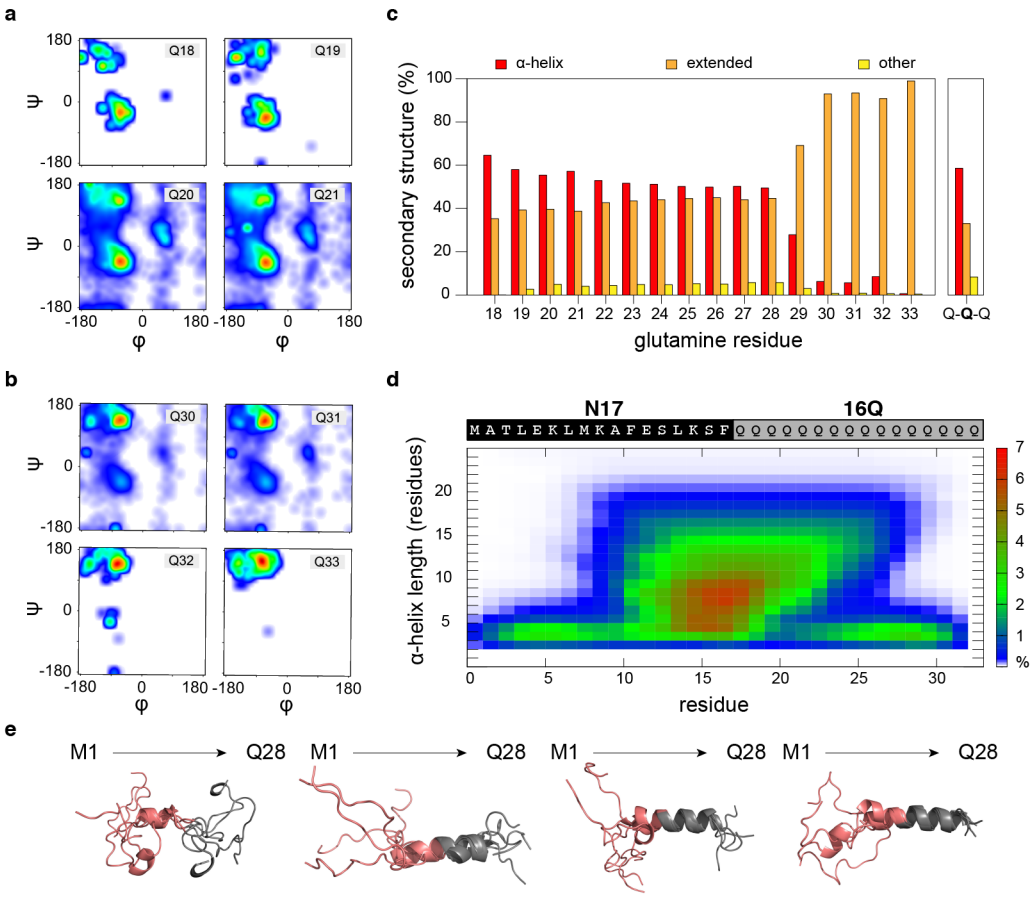


Figure 3. NMR-derived ensemble model of H16. Residue-specific Ramachandran plots for (**a**) Q18-Q21 and (**b**) Q30-Q33 obtained from the optimized ensemble. (**c**) Population of α-helix, extended and *other* conformations calculated from the Ramachandran plot for all glutamines in H16 (see Figure S3). The side panel Q-**Q**-Q shows these populations for glutamine tri-peptides present in a coil database[49]. (**d**) Secondary structure map (SS-map) displaying the length and the residues encompassing the α-helices found in the N-terminal region of the optimized ensemble model of H16. The color code (right) indicates the population of the α-helices. (**e**) Representative conformations of the four ensembles used to describe the NMR CSs measured for H16. Only the region from M1 to Q28, optimized with the N→C ensembles, is displayed. The SS-maps for these ensembles are displayed in Figure S4.

Glutamine side chains indicate a structural coupling of N17 and the poly-Q tract

According to our ensemble model, the last four residues of N17 are strongly linked to the first two

glutamines of the poly-Q tract. However, the model, which is based on backbone CSs, does not unveil the structural bases of this structural connection. Benefitting from the lack of signal overlap in the ^{13}C-HSQC of the SSIL samples, glutamine-specific Cβ-H_2 and Cγ-H_2 correlations could be analyzed (Figures 4 and S5). As expected for a flexible protein, the majority of glutamines in H16 display two correlation peaks for Cβ-H_2 and a single one for Cγ-H_2, indicating increased mobility along the side chain (Figure S5). Interestingly, the first four glutamines, Q18-Q21, present different spectroscopic features. While Q18 and Q19 display a single peak for Cβ-H_2 and Cγ-H_2, these correlations are split in two for Q20 and Q21 (Figure 4a). Most probably, the splitting of Cγ-H_2 is caused by the rigidification of the glutamine side chains, which results in a different chemical environment for the two diastereotopic Hγ atoms. This rigidification likely originates from the formation of a hydrogen bond between the side chain amide group and the backbone of a neighboring residue. Notice that similar spectroscopic features were observed in a recent characterization of the androgen receptor (AR) N-terminal domain fragments hosting poly-Q tracts of different lengths[55].

In order to substantiate this hypothesis and profiting that Q20 and Q21 Nε-H_{21} displayed isolated peaks (see below), we determined the temperature coefficients ($\sigma H^N/T$) for these two atoms in a ^{15}N-labeled H16 sample (Figure S5). We derived $\sigma H^N/T$ values of -4.1 and -3.5 ppb/K for Q20 and Q21 $H^N\varepsilon_{21}$, respectively. These values are less negative than the threshold value, −4.5 ppb/K, suggesting their participation in a hydrogen bond[54]. Conversely, we obtained $\sigma H^N/T$ values of -5.8 and -6.2 for Q32 and Q54 $H^N\varepsilon_{21}$, respectively, confirming the singularity of the first glutamines of the tract.

Multiple α-helical N-capping hydrogen bonding networks involving glutamine side chains have been described[56–59]. In the AR study, the authors propose a bifurcated hydrogen bond where the amide backbone of residue *i*-4 simultaneously forms hydrogen bonds with the backbone and the side chain of glutamine in position *i*[55]. Indeed, in this novel mechanism, the side chain hydrogen bond further stabilizes the canonical (*i-4* → *i*) backbone helical hydrogen network. This interaction would be protected by the hydrophobic side chain of residue *i-4*, a leucine in AR[60]. According to this model and in the context of huntingtin, the side chain amide groups of Q20 and Q21 would form hydrogen bonds with S16 and F17, respectively, the latter one being the most stable interaction according to the extent of the Cγ-H_2 splitting. The Nε-H_{21} peaks for Q20 and Q21, which appear clearly shifted from the other side-chain peaks, further substantiate this feature (Figure 4b). The frequency shift for these two peaks cannot be only attributed to the involvement of these two atoms in an α-helical hydrogen bond, whose signature is a ^{15}N upfield shift[55]. An alternative explanation is the ring current effects exerted by F17 that, upon formation of the canonical hydrogen bond with Q21, places its side chain in the proximity of Q21 Nε-H_{21} and to lesser extent to Q20 Nε-H_{21}. Note that the magnitude of the ring current shift is difficult to anticipate as it depends on persistence and the orientation of the aromatic side chain with respect to the shifted atom. Conversely, Q18 Nε-H_{21}, which is adjacent to F17 in the sequence, is not affected by the presence of the aromatic side chain. This last observation, which is in line with the protective role of the phenylalanine hydrophobic side chain, underpins the structural coupling between

the N17 and the poly-Q tract through a hydrogen bonds network.

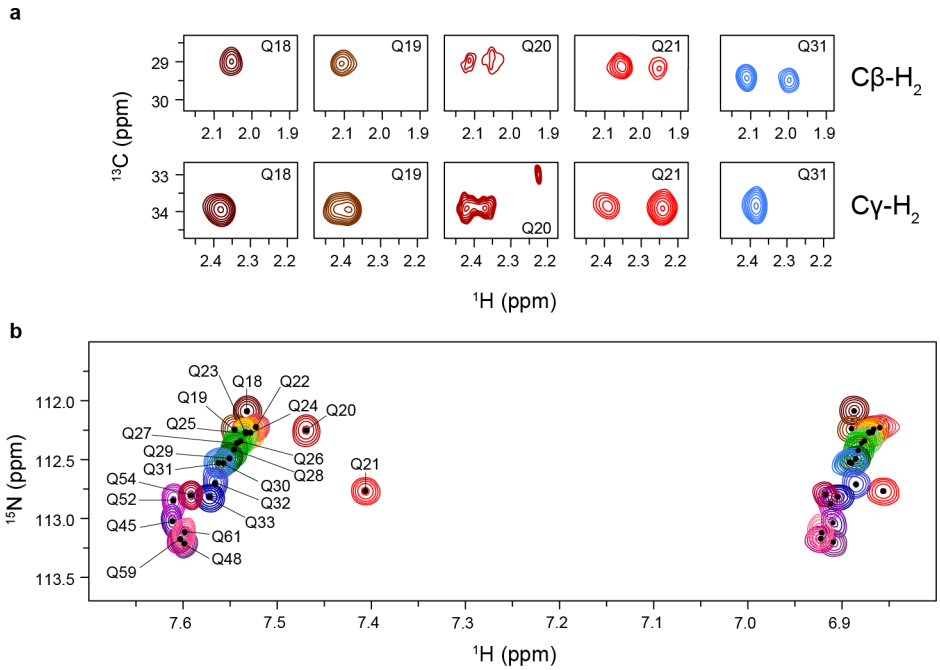


Figure 4. Side chain NMR scanning. (**a**) Cβ-H_2 and Cγ-H_2 regions of the ^{13}C-NMR spectra of glutamines Q18, Q19, Q20, Q21, and Q31. The spectra of Q31 display the standard behavior of disordered glutamines with a doublet and singlet for Cβ-H_2 and Cγ-H_2, respectively. (**b**) Zoom on the Nε2-Hε2 side chain region of the ^{15}N-HSQC spectra measured for all glutamines in H16. Q20 and Q21 do not follow the trend displayed by the other glutamines due to their implication in the formation of hydrogen bonds.

Mutants reveal the effects of N17 side chains on structural coupling

In order to further investigate the structural bases of the connection between the N17 and the poly-Q domains, we designed three H16 mutants in which the last residues of N17 (^{14}LKSF^{17}) were mutated to ^{14}LKGG^{17}, ^{14}LLLF^{17} and ^{14}LKAA^{17} (Figures 1a and 5). The LKGG and LKAA mutants were designed to weaken to different extents the hydrogen bond network found in wild-type httex1, while the LLLF would strengthen the network. The ^{15}N-HSQC spectrum of the LKGG mutant presented very clear differences with respect to the wild-type one, especially in the glutamine region (Figures 5a and S6). The relatively disperse glutamine peaks of wild-type H16 coalesced in a broad, high-intensity downfield-shifted peak. Furthermore, the dispersion of the Nε2-H_2 side chain signals in LKGG-H16 was dramatically reduced (Figure S6). These observations demonstrated that the helical nature of the poly-Q tract is lost when mutating the last two residues of N17 to glycine. The origin of the dramatic structural changes was investigated using the SSIL strategy by isotopically labeling residues Q18, Q20 and Q21 of LKGG-H16 (Figure 5b). Compared to H16, the three residues present very different peak positions in both spectra. While the N-H correlation of Q18 was strongly influenced by the

neighboring glycines, Q20 and Q21 appeared shifted downfield, in the same position as the broad glutamine peak (Figure 5a). Cα-Hα correlation peaks for these three residues were strongly shifted towards a less helical region of the spectrum (Figure 5b). The SCS analysis of these three residues indicated that the helicity was severely reduced compared to wild-type H16 but not completely abolished, indicating that the poly-Q tract is slightly helical for this mutant (Figure 5l). For the three glutamines, Cβ-H_2 and Cγ-H_2 correlations presented a doublet and a singlet, respectively (Figure 5c), indicating the loss of the hydrogen bonds connecting the N17 to the poly-Q. However, it was unclear whether the absence of this structural coupling affected the inherent helical tendency of N17. To resolve this point we assigned the N17 region of LKGG-H16, using traditional 3D-NMR experiments, and computed the SCSs (Figure S7). Comparison of the wild-type and LKGG-H16 SCS analyses showed that the double point mutation is resulting in a bidirectional loss of helicity, impacting the last six residues of N17 as well as the following glutamines.

The mutant LLLF-H16 was designed to provide new sites to the first glutamines of the tract to form side chain hydrogen bonds and thus strengthen the helical tendency of the homo-repeat. Glutamine peaks of the LLLF-H16 ^{15}N-HSQC spectrum presented an additional upfield density that was attributed to an increased helical content in this mutant (Figure 5d). SSIL samples for Q18, Q20 and Q21 displayed important chemical shift changes in both the N-H and the Cα-Hα correlations (Figure 5e). In fact, Q20 and Q21 N-H and Cα-Hα peaks appear shifted towards more helical conformations with respect to the wild-type. Unfortunately, the Cα-Hα peak for Q18 could not be observed, most probably due to a folding/unfolding process in the μs to ms dynamic regime that broadens the peak beyond detection. The SCS analysis showed a strong α-helical increase for Q20 and Q21, substantiating the above-mentioned qualitative observations regarding the helical increase for this mutant (Figure 5l). Despite their overall low intensity, the Cβ-H_2 and Cγ-H_2 peaks demonstrate a stronger structural coupling between N17 and the poly-Q tract. The Cγ-H_2 splitting of Q20 and Q21 is larger than that observed in the wild-type. Cβ-H_2 presents a single peak for Q18 and Q20, something occurring only for Q18 and Q19 in the wild-type (Figure 5f, 4a), indicating a stronger hydrogen bond network involving additional residues. Therefore, the LLLF-H16 mutant unambiguously links the strength of the hydrogen bond network between N17 and the first glutamines of the homo-repeat with the persistence and stability of the resulting α-helices.

The third mutant, LKAA-H16, was designed to display an intermediate behavior with respect to the other two. Alanine is a helical promoter amino acid but its side chain is smaller than those of leucine and phenylalanine. The LKAA-H16 ^{15}N-HSQC spectrum was similar to the wild-type one, although less density was observed in the upfield part of the glutamine spectral region (Figure 5g and S6). The Cα-Hα peaks for Q18, Q20 and Q21 were shifted downfield in the ^{1}H dimension with respect to those of the wild-type (Figure 5h). This feature was quantified in the SCS analysis, which indicates a decrease in the helical tendency for Q18 and Q20, while Q21 remained almost unchanged. Exploration of the side chains of these three residues suggested some clues to this observation. Interestingly, only

Q18 presented two Cγ-H_2 peaks, indicating a hydrogen bond between the side chain of this residue and the backbone of L14. Therefore, the structural connectivity is modified in LKAA-H16 by exchanging the two side chain hydrogen bonds present in the wild-type by a new one involving the first glutamine of the tract and concomitantly a decrease of the helical tendency for this mutant.

The inspection of the Nε-H_2 peaks of the suppressed samples further substantiates the structural model of the hydrogen bond connection (Figure S6c). Q21 Nε-H_{21} peak of LLLF-H16 displays a stronger upfield shift in the ^{1}H dimension than in the wild type, suggesting more persistent ring current effect by F17 aromatic ring caused by the formation of a more stable hydrogen bond. This enhanced stabilization of the α-helix is also manifested in the Q18 Nε-H_{21} peak that now appears strongly upfield shifted in the ^{15}N dimension. In LKAA-H16, where F17 is mutated by an alanine, the Nε-H_{21} peaks of Q18, Q20 and Q21 are not displaced in the ^{1}H dimension despite the fact that they are involved in an α-helix, demonstrating that the ring current effects are at the origin of the unusual frequencies of Nε-H_{21} atoms in httex1.

The poly-P C-terminal flanking region breaks the helical tendency of the glutamine homo-repeat

In order to explore the structural connection between the poly-Q and the poly-P homo-repeats, we designed a mutant with five glycines between these tracts (H16-5G), aiming to structurally uncouple them (Figure 1a)[24]. This mutant yielded a very similar ^{15}N-HSQC spectrum to that of H16, with glutamine peaks displaying an equivalent level of dispersion (Figure 5j and S6). No relevant differences were observed in the backbone or side chain correlations between both spectra, suggesting that the presence of the five glycines does not perturb the overall structure of H16. Nevertheless, we prepared an SSIL H16-5G sample with [^{15}N, ^{13}C]-glutamine in position Q30, which lies in the non-helical part of the poly-Q tract of H16, to investigate structural changes resulting from uncoupling both homo-repeats at residue level. In comparison with the wild-type, the N-H correlations were shifted upfield, whereas the Cα-Hα correlations were shifted downfield in the ^{1}H and upfield in the ^{13}C dimension (Figure 5k). This observation suggested an increase in the helical tendency of this residue in the new context, which was quantitatively proven by SCS analysis. Q30 adopts a positive SCS value in H16-5G while in the wild-type this residue has a slightly negative value (Figure 5l). This observation demonstrates that the poly-P tract in httex1 exerts a strong conformational perturbation on the neighboring glutamines by enriching the ensemble with extended conformations, which break the inherent helical propensity of the poly-Q.

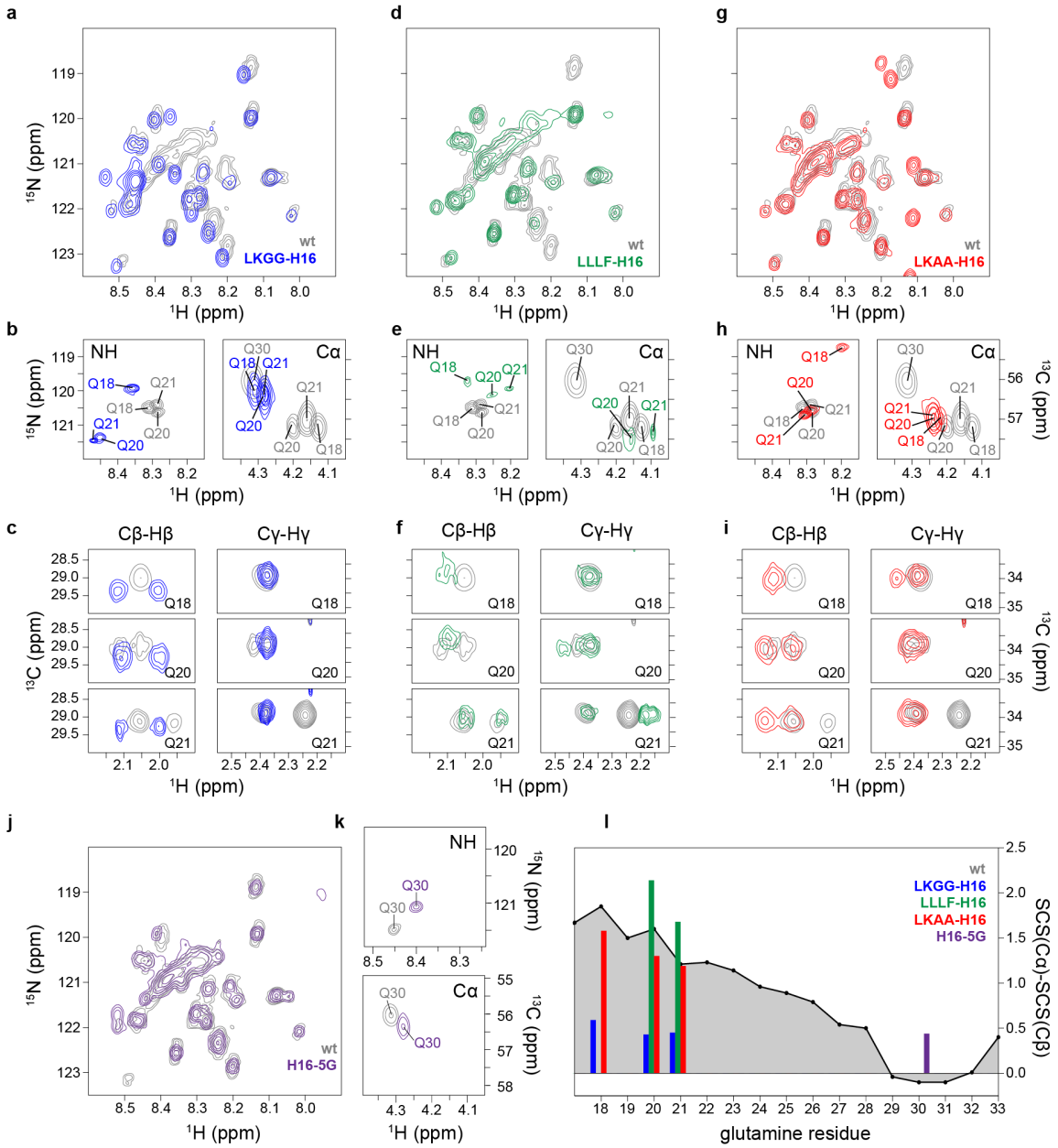


Figure 5. SSIL analyses of the structural effects of the flanking regions on the poly-Q tract of H16. Overlay of the glutamine region of the ^{15}N-HSQC spectra of fully labeled wild-type H16 (grey) with the N17 mutants LKGG-H16 (**a, blue**), LLLF-H16 (**d, green**), and LKAA-H16 (**g, red**). The same color-code was used throughout the figure. Zoomed overlays of the ^{15}N- and ^{13}C-HSQCs for site-specifically labeled Q18, Q20 and Q21 of wild-type H16 (grey) with LKGG-H16 (**b**), LLLF-H16 (**e**) and LKAA-H16 (**h**). Cβ-H_2 and Cγ-H_2 NMR peaks of the Q18, Q20 and Q21 glutamine side chains of the three N17 mutants LKGG-H16 (**c**), LLLF-H16 (**f**) and LKAA-H16 (**i**) compared with those obtained for the wild-type (grey). Zoomed ^{15}N- and ^{13}C-HSQC spectra for the H16-5G mutant, which probes the structural perturbation exerted by the poly-P tract, displaying the N-H glutamine region (**j, purple**) overlaid with the wild-type (grey), and the SSIL spectra measured for Q30 (**k**). (**l**) Histogram of the SCS analyses for the different SSIL samples of the structural mutants measured: Q18, Q20 and Q21 for the LKGG-H16, LLLF-H16 and LKAA-H16 mutants, and Q30 for the H16-5G mutant. The SSIL-derived SCS values are compared to those obtained for the wild-type (grey area). Note that no SCS value was derived for Q18 in the LLLF-H16 mutant due to the absence of the Cα-Hα peak in the ^{13}C-HSQC.

Sequence analyses of poly-Q flanking regions in human proteins

In a previous bioinformatics analysis it was shown that leucines, prolines and histidines were especially enriched in the flanking regions of human poly-Q tracts[61]. While leucine and histidine were similarly enriched on both sides, proline displayed a preference for the C-flanking region. We complemented this study by exploring whether the compositional bias in the flanking regions was poly-Q length dependent. For that, four hundred fragments with ten or more glutamine residues and containing a maximum of two non-glutamine residues were collected from 309 different human proteins, and the ten preceding (-10 to -1) and succeeding (+1 to +10) residues were compositionally analyzed. Figure S8 shows that using our poly-Q definition (maximum of 2 non-glutamine residues in fragments of 10 or more glutamine residues), we obtain similar results as those derived by Ramazzotti *et al.*[61], with leucine, proline and to a lesser extent histidine and alanine being enriched in poly-Q flanking regions, as well as the positional asymmetry of proline. Interestingly, using our poly-Q definition we identify an enhanced enrichment of leucines in the N-flanking region compared with the C-flanking one. Note that a less restrictive definition of the homo-repeat to include larger glutamine-rich regions was used in the previous study and this could lead to changes in the enrichment levels.

We then analyzed the effect of the length of the glutamine homo-repeats on the above-described compositional biases by selecting pure glutamine stretches. The leucine population in position -1 increases with the length of the poly-Q tract, reaching a maximum of 30.0% when the number of consecutive glutamines in the tract is seven or more, and it is slightly reduced for longer homo-repeats (Figure 6a). Interestingly, positions from -2 to -4 also display a similar length dependency, although the enrichment is less prominent than in position -1. The population of prolines in the C-flanking region systematically increases with the length of the poly-Q tract. The maximum of the enrichment occurs at position +1 that extends over the complete region, while it remains close to the background in the N-flanking region (Figure 6b).

Next, we explored the secondary structure propensity in the N-flanking region of long human poly-Q tracts with a recently developed approach (manuscript in preparation) based on the previously mentioned large database of three-residue fragments[49]. Briefly, the residue-specific conformational bias was evaluated accounting for the effects exerted by the preceding and succeeding amino acids. Then, the percentage of α-helical, extended or other conformations was derived. The position-specific percentages obtained for each family were averaged in increasing sections of the N-flanking regions and reported as notched box plots in Figure 6c. For each fragment, the α-helical conformation was preferred with median values ranging from 50.2% to 70.6%, while the preference for extended or other conformations was always lower than 25%. Interestingly, the α-helical tendency presents its largest percentage when close to the poly-Q homo-repeat (residues -1 and -2), and systematically decreases when more residues of the N-flanking region are incorporated in the analysis.

In summary, these sequence analyses indicate that the structural and compositional characteristics observed in httex1 flanking regions are shared by a large number of other human poly-Q-containing proteins. This observation suggests that the structure-mediated functional mechanisms found for httex1 in the present study are common to many other human glutamine-rich proteins.

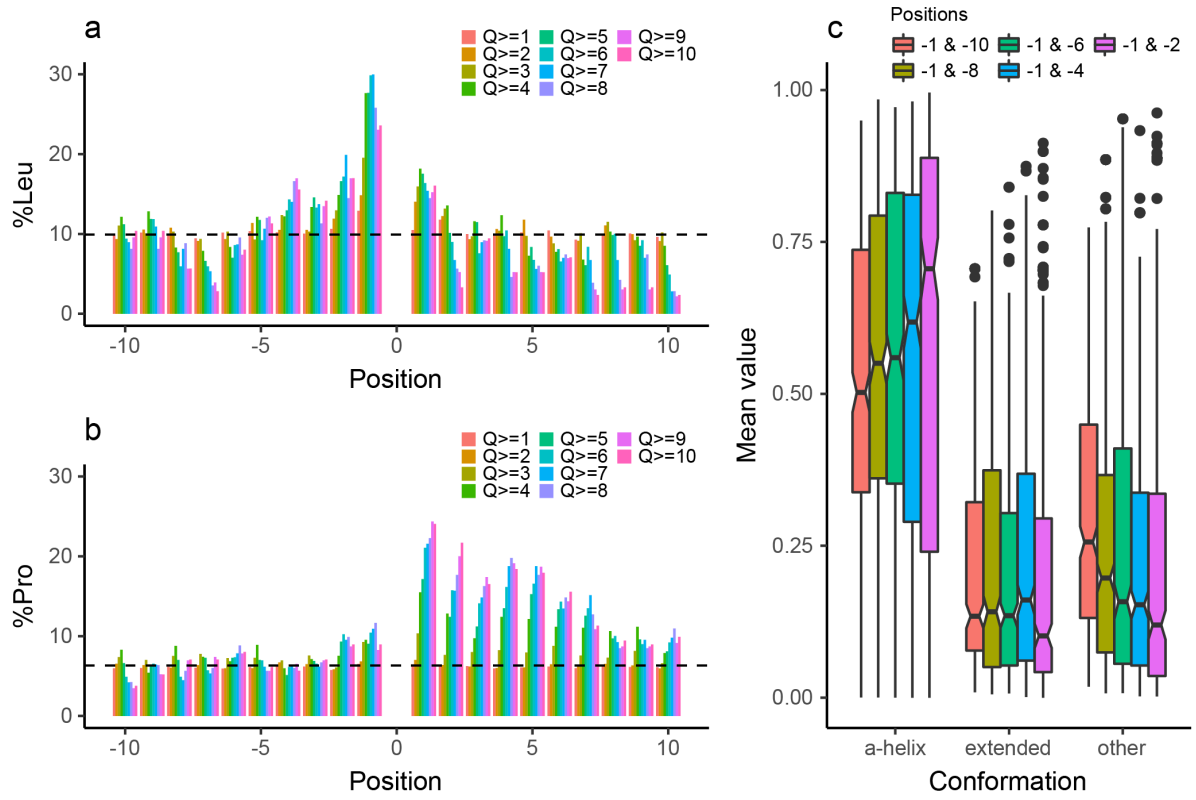


Figure 6. Primary and secondary structure context of human glutamine-rich proteins. (**a**) Leucine and (**b**) proline abundance per position in region -10 to +10 of poly-Q regions in the context of pure glutamine stretches of variable length. Horizontal dashed lines correspond to the percentage of leucines (9.9%) and prolines (6.3%) found in the human proteome. An analysis of all 20 natural amino acids is displayed in Figure S8. (**c**) Secondary structural prediction (α-helix, extended and others) per two-residue block in the N-terminal flanking regions of glutamine-rich fragments.

Discussion

In this study, we demonstrate that the previously developed SSIL strategy[39] can be systematically applied to investigate poly-Q tracts, one of the most abundant homo-repeats in eukaryotes[62–64], and to connect their structural features with their specialized biological functions. The NMR analysis of the SSIL samples demonstrates that H16 is disordered, but hosts an important level of helicity that is initiated in N17, reaching the maximum at the beginning of the poly-Q tract and smoothly vanishing along the homo-repeat. A conformational ensemble model refined from experimental data recapitulates this non-uniform helical propensity as an equilibrium of multiple canonical helices of

different lengths. All these helices start in N17 and extend towards the poly-Q tract, comprising an increasing number of glutamines. Q28 is the last glutamine influenced by the α-helical tendency, and subsequent glutamines present random coil or slightly extended conformations. The enrichment in α-helical conformations in httex1 is in agreement with crystallographic structures[65,66] and NMR data[35,44]. However, the non-homogeneous helicity can only be captured when an ensemble representation is used, as done in the present study.

Our NMR measurements demonstrate that N17 has an inherent α-helical tendency that is transferred to the glutamine homo-repeat through a hydrogen bond network involving glutamine side chains. Although the structure of this network cannot be unambiguously resolved with our NMR data, a recent study on the poly-Q tract of the AR demonstrates that glutamine side chains form hydrogen bonds with hydrophobic residues in the *i-4* position, reinforcing the canonical $CO_{i-4} \rightarrow H_{N,i}$ backbone hydrogen bond[55]. In this study it was suggested that the large and hydrophobic residues in the *i-4* position were key for the formation of the bifurcated hydrogen bond by protecting it from water molecules. In the context of H16, the last two residues of N17, ^{16}SF^{17}, would play the main role in stabilizing and propagating the helix within the poly-Q tract. We have validated this model by monitoring the side chain CSs of three mutants in which we modified the last residues in N17 and profiting the chemical shift changes induced by the ring current effects of F17 to spatially close atoms. While the LLLF-H16 mutant strengthens the structural coupling between N17 and the poly-Q tract, the LKGG-H16 mutant is unable to form the hydrogen bond network. Interestingly, LKAA-H16 provides evidence of the malleability of this helical propagation. For this mutant, hydrogen bonds involving ^{20}QQ^{21} are hampered by the absence of large hydrophobic amino acids in positions *i-4* and, instead, this mutant utilizes L14 and Q18 to trigger the structural coupling between both regions. In addition, these results highlight that the conformational nature of the residues involved in the hydrogen bond network is important. In that sense, despite not forming bifurcate hydrogen bonds, the inherent helical propensity of alanines is required to connect N17 with the poly-Q tract, a phenomenon that is not observed in the LKGG-H16 mutant. These observations suggest that the residue preceding the poly-Q tract (position -1 according to our nomenclature) is the preferred one to trigger helicity in the homo-repeat. Consequently, the large population of leucines in this position found here and in a previous bioinformatics analysis of eukaryotic proteomes strongly suggests the generality of helical induction in poly-Q tracts through side chain hydrogen bonds[61]. Interestingly, this enrichment increases for poly-Q tracts with seven or more consecutive glutamines (Figure 6a). Altogether, these observations point towards a general structure/function relationship for poly-Q fragments involving long α-helices of variable length and stability, depending on the residues preceding the tract. This observation is in line with the recurrent presence of coiled-coils in protein fragments containing poly-Q tracts as well as in their corresponding partners[10].

Multiple post-translational modifications have been described for N17, including phosphorylation, acetylation, ubiquitination and SUMOylation, and it has been shown that their presence perturbs the

function, aggregation properties and toxicity of huntingtin[19,23]. According to our observations, modifications that decrease the helical propensity of N17 or break the hydrogen-bond network will induce an increase in disorder in the poly-Q tract. In a recent study, it was demonstrated that mono-phosphorylation on S13 or S16 and di-phosphorylation strongly disrupt N17 helicity. Interestingly, these post-translationally modified forms of httex1 are less prone to aggregation than the unmodified form[67]. These observations can now be rationalized in the light of our results, indicating a strong link between the level of structure, aggregation and modulation through post-translational modifications.

It is well known that due to the limited conformational variability and the inability to form hydrogen bonds, proline is considered to be a structure-breaking residue with the capacity to extend its structural influence towards neighboring residues[68]. Previous CD experiments on httex1-mimicking peptides demonstrated the enrichment of polyproline-II conformations in poly-Q tracts preceding poly-P[69]. Here, we could demonstrate this effect at residue level through the NMR-driven molecular modeling of httex1 and by monitoring the CS changes in the H16-5G mutant. Moreover, our NMR analysis enables the assessment of the extent of structural perturbation exerted by the poly-P over the poly-Q tract. The last five glutamines of the tract preferentially adopt random coil or slightly extended conformations due to the influence of the proline tract[52]. However, this influence extends much further and causes the smooth decay of the helicity along most of the poly-Q tract in H16. Indeed, recent CD experiments as well as partial NMR assignments of httex1 variants with longer homo-repeats show that the helical content of httex1 systematically increases with the length of the poly-Q[35–37]. The ensemble of these observations suggests that the perturbation exerted by the poly-P tract has a defined range of influence and, therefore, the poly-Q homo-repeat remains helical in the region preceding the perturbed segment. According to the ensemble of these studies, we can estimate that the conformational influence of the poly-P tract extends to the last 13 glutamines of httex1. Glutamines lying in this perturbed region sense a distinct structural influence from both sides, the helical propagation from the N-terminus and the helix-breaking tendency from the C-terminus. These opposing influences are captured in a different balance between α-helix and extended conformations in the individual Ramachandran plots displayed in Figures 3a,b and S3.

Sequence analyses also demonstrate that the presence of prolines at the C-terminal flanking region of glutamine-rich segments is common in eukaryotic proteins and especially significant in the positions immediately adjacent to poly-Q tracts[61]. Here we show that in human proteins the extent of this proline compositional bias is poly-Q length dependent, meaning that proteins having longer poly-Q tracts have a higher probability to be followed by prolines. Interestingly, an examination of huntingtin orthologs shows that the poly-P occurs only in species with four or more consecutive glutamines, suggesting that these two homo-repeats have coevolved[70,71]. The consecutive presence of glutamine and proline repeats is also observed in ataxin-2 and ataxin-7, two proteins whose abnormal poly-Q expansion causes spinocerebellar ataxias SCA2 and SCA7, respectively[72]. This concatenation of glutamine- and proline-rich regions in unrelated proteins from different organisms suggests a strong selective pressure

at the molecular level and a common structure/function mechanism[61]. For many of these proteins this mechanism might be the protection from aggregation of the expanded poly-Q tracts that arises from the conformational influence exerted by proline-rich regions. Prolines at the C-terminus shorten the length of the helical fragments of the poly-Q tract, reducing the stability of the intermolecular interactions and the subsequent aggregation.

Our results point to an overall extended structure of httex1 that is in contrast to the tadpole-like model where N17 and the poly-Q tract form a compact structure stabilized by fuzzy contacts from which the semi-rigid PRR sticks out[34,35]. The compact httex1 structure has been derived from computational studies and sparse distance restrains derived from smFRET[34,73]. Although our experimental data do not report on long-range contacts, the hydrogen network involving N17 and the poly-Q tract, as well as the absence of the spectroscopic features of a turn in the interphase between both domains, strongly privileges the extended model over the compact one. Despite the overall extendedness, our data show that httex1 remains highly disordered, especially the last glutamines of the poly-Q tract and the PRR region. This flexibility would allow transient contacts between remote parts of the protein that could be at the origin of the long-range contacts observed in smFRET experiments[34]. This extended structure supports the 'linear lattice' model of toxicity in which the number of exposed glutamines increases with the length of the tract. However, the emergence of a toxic conformation, appearing after the formation of soluble oligomers as previously suggested[12], is also compatible with our model, which focuses in the monomeric form of httex1.

From a practical point of view, our observations warn about the use of isolated poly-Q peptides disregarding the sequence context to predict the biophysical/structural behavior and the aggregation propensity of glutamine-rich proteins[74,75]. We demonstrate that the chemical and structural features of poly-Q flanking regions govern the conformational behavior of the homo-repeat. Therefore, biophysical studies on poly-Q containing proteins must be performed with fragments including the relevant neighboring elements. With the SSIL approach these protein-specific properties can be now addressed at high resolution in order to unveil among other features the origin of the different pathological thresholds observed in poly-Q related diseases[76].

Altogether, our data demonstrates that the poly-Q tract in httex1 is exposed to opposing structural effects from both flanking regions. Notably, the enrichment in hydrophobic residues and the α-helical conformations in the N-flanking region, as well as the downstream enrichment in prolines, are shared by many eukaryotic glutamine-rich proteins. This suggests that many proteins exploit these structural properties, which are centered on the structural flexibility and versatility of poly-Q tracts, in order to perform specific biological functions while avoiding aggregation and toxicity.

Materials and Methods

Huntingtin exon1 constructs

All plasmids were prepared as previously described[39]. Synthetic genes of wild-type huntingtin exon1 with 16 consecutive glutamines (H16) or H16 carrying the amber codon (TAG) instead of the glutamine codon, e.g. Q18 (H16Q18), were ordered from Integrated DNA Technologies (IDT). Following this strategy, 22 amber mutants were ordered: 16 within the poly-Q tract and six outside. Synthetic genes of the structural mutants (LKGG-H16, LKAA-H16, LLLF-H16 and H16-5G) and their corresponding amber codon mutants (Q18, Q20, Q21 and Q30) were ordered from GeneArt®. All genes were cloned into pIVEX 2.3d, giving rise to pIVEX-H16-3C-sfGFP-His_6 and mutants. The sequence of all plasmids was confirmed by sequencing by GENEWIZ®.

Preparation and aminoacylation of suppressor $tRNA_{CUA}$

A $tRNA_{CUA}$/tRNA synthetase pair based on the Gln2 tRNA[77] and glutamine ligase GLN4 from *Saccharomyces cerevisiae* was prepared in house as previously described[39]. Briefly, the artificial suppressor $tRNA_{CUA}$ was transcribed *in vitro* and purified by phenol-chloroform extraction. Prior to use, the suppressor $tRNA_{CUA}$ was refolded in 100 mM HEPES-KOH pH 7.5, 10 mM KCl at 70°C for 5 min and a final concentration of 5 mM $MgCl_2$ was added just before the reaction was placed on ice. The refolded $tRNA_{CUA}$ was then aminoacylated with [^{15}N, ^{13}C]-glutamine (CortecNet) in a standard aminoacylation reaction: 20 µM $tRNA_{CUA}$, 0.5 µM GLN4, 0.1 mM [^{15}N, ^{13}C]-Gln in 100 mM HEPES-KOH pH 7.5, 10 mM KCl, 20 mM $MgCl_2$, 1 mM DTT and 10 mM ATP[78]. After incubation at 37°C for 1 hour GLN4 was removed by addition of glutathione beads and loaded suppressor $tRNA_{CUA}$ was precipitated with 300 mM sodium acetate pH 5.2 and 2.5 volumes of 96% EtOH at -80°C and stored as dry pellets at -20°C. Successful loading was confirmed by urea-PAGE (6.5% acrylamide 19:1, 8 M urea, 100 mM sodium acetate pH 5.2)[78].

Standard cell-free expression conditions

Lysate was prepared as previously described[39] and based on the *Escherichia coli* strain BL21 Star (DE3)::RF1-CBD_3, a gift from Gottfried Otting (Australian National University, Canberra, Australia)[79]. Cell-free protein expression was performed in batch mode as described by Apponyi *et al.*[80]. The standard reaction mixture consisted of the following components: 55 mM HEPES-KOH (pH 7.5), 1.2 mM ATP, 0.8 mM each of CTP, GTP and UTP, 1.7 mM DTT, 0.175 mg/mL *E. coli* total tRNA mixture (from strain MRE600), 0.64 mM cAMP, 27.5 mM ammonium acetate, 68 µM 1-5-formyl-5,6,7,8-tetrahydrofolic acid (folinic acid), 1 mM of each of the 20 amino acids, 80 mM creatine phosphate (CP), 250 µg/mL creatine kinase (CK), plasmid (16 µg/mL) and 22.5% (v/v) S30 extract. The concentrations of magnesium acetate (5 - 20 mM) and potassium glutamate (60 - 200 mM) were adjusted for each new batch of S30 extract. A titration of both compounds was performed to obtain the maximum yield.

Cell-free H16Qx position screen

Plasmids of all 22 amber mutants of wild-type H16 were tested for possible position specific effects of the amber codon placement on the suppression efficiency at a final concentration of 10 µM $tRNA_{CUA}$. The time-course of H16 protein synthesis was monitored using a fluorescence read-out (sfGFP) and a plate reader/incubator (Gen5, BioTek Instruments, 485 nm (excitation), 528 nm (emission)). Assays were carried out as triplicates in a reaction volume of 50 µL dispensed in 96-well plates. The reactions were incubated at 23°C for 5 hours.

Preparation of NMR samples

Samples for NMR studies were produced at 5-15 mL scale and incubated at 23°C and 750 rpm in a thermomixer for 5 hours. Uniformly labeled NMR samples were obtained by substituting the standard amino acid mix with 3 mg/mL [^{15}N, ^{13}C]-labeled ISOGRO®[40] (an algal extract lacking four amino acids: Asn, Cys, Gln and Trp) and additionally supplying [^{15}N, ^{13}C]-labeled Asn, Cys, Gln and Trp (1 mM each). Furthermore, potassium glutamate was substituted by 80 mM potassium acetate to enable the labeling of glutamates. To produce site-specifically labeled samples, 10 µM of [^{15}N, ^{13}C]-Gln suppressor $tRNA_{CUA}$ were added to the standard reaction mixture (see above).

Protein purification

The cell-free reaction was thawed on ice and diluted 2-3 fold with buffer A (50 mM Tris-HCl pH 7.5, 500 mM NaCl, 5 mM imidazole) before loading onto a Ni gravity-flow column of 1 mL bed volume (cOmplete™ His-Tag Purification Resin, Sigma Aldrich). The column was washed with buffer B (50 mM Tris-HCl pH 7.5, 1000 mM NaCl, 5 mM imidazole) and the target protein was eluted with buffer C (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 250 mM imidazole). Elution fractions were checked under UV light and fluorescent fractions were pooled and dialyzed against NMR buffer (20 mM BisTris-HCl pH 6.5, 150 mM NaCl) at 4°C using SpectraPor 1 MWCO 6-8 kDa dialysis tubing (Spectrum Labs). Dialyzed protein was then concentrated with 10 kDa MWCO Vivaspin centrifugal concentrators (3500 x g, 4°C) (Sartorius). Protein concentrations were determined by means of fluorescence using an sfGFP calibration curve. Final NMR sample concentrations ranged from 4 to 11 µM. Protein integrity was analyzed by SDS-PAGE.

NMR experiments and data analysis

All NMR samples contained final concentrations of 10% D_2O and 0.5 mM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS). Experiments were performed at 293 K on a Bruker Avance III spectrometer equipped with a cryogenic triple resonance probe and Z gradient coil, operating at a ^{1}H frequency of 700 MHz or 800 MHz. ^{15}N-HSQC and ^{13}C-HSQC were acquired for each sample in order to determine amide ($^{1}H_N$ and ^{15}N) and aliphatic ($^{1}H_{aliphatic}$ and $^{13}C_{aliphatic}$) chemical shifts,

respectively. Spectra acquisition parameters were set up depending on the sample concentration and the magnet strength. ^{15}N-HSQC spectra were acquired for 8 to 20 hours using 256-512 scans, 88-128 increments and a spectral width of 21 ppm in the indirect dimension. ^{13}C-HSQC spectra were acquired for 10 to 24 hours using 256-512 scans, 96-128 increments and a spectral width of 60 ppm in the indirect dimension. All spectra were processed with TopSpin v3.5 (Bruker Biospin) and analyzed using CCPN-Analysis software[81]. Chemical shifts were referenced with respect to the H_2O signal relative to DSS using the ^{1}H/X frequency ratio of the zero point according to Markley *et al*.[82].

Random coil chemical shifts were predicted using POTENCI, a pH, temperature and neighbor corrected IDP library (http://nmr.chem.rug.nl/potenci/)[46]. Secondary chemical shifts (SCS) were obtained by subtracting the predicted value from the experimental one ($SCS=\delta_{exp}-\delta_{pred}$). For better reliability of the results regarding possible referencing errors, we used the combined C_α and C_β secondary chemical shifts ($SCS(C_\alpha)-SCS(C_\beta)$). In addition, secondary structure propensities (SSPs) were calculated using the script developed by Marsh *et al*.[47] and the refDB database[48].

Model building and experimental ensemble optimization

Ensemble models for the two families capturing the conformational influences of the flanking regions, N→C and N←C, were constructed with the algorithm described in reference[49], which uses a curated database of three-residue fragments extracted from high-resolution protein structures. The averaged Cα and Cβ CSs for the 34 ensembles, 17 for each family, were computed with SPARTA+[51] and used to refine a final ensemble in agreement with the experimental data. Concretely, the optimized ensemble model of H16 was built by reweighting the populations of the pre-computed ensembles, minimizing the error with respect to the experimental Cα and Cβ CSs. In order to capture the influence of the flanking regions, glutamines within the tract were divided into two groups: those influenced by N17 and those influenced by the poly-P tract, whose chemical shifts were fitted with the N→C and N←C ensembles, respectively. The limit between both families was systematically explored by computing the agreement between the experimental and optimized CSs through a χ_i^2 value. An optimal description of the complete CS profile was obtained when Q28 was chosen as the last residue structurally connected with N17. Finally, an ensemble of 50,000 structures was built using the optimized weights and it was used to analyze the residue-specific Ramachandran propensities and the secondary structure population using SS-map[53].

Acknowledgements

The authors thank Gottfried Otting for providing the BL21 (DE3) Star::RF1-CBD3 strain and Grayson Gerlich for reading the manuscript. This work was supported by the European Research Council under the European Union's H2020 Framework Programme (2014-2020) / ERC Grant agreement n° [648030], and Labex EpiGenMed, an « Investissements d'avenir » program (ANR-10-LABX-12-01) awarded to PB. The CBS is a member of France-BioImaging (FBI) and the French Infrastructure for

Integrated Structural Biology (FRISBI), 2 national infrastructures supported by the French National Research Agency (ANR-10-INBS-04-01 and ANR-10-INBS-05, respectively). AU is supported by a grant from the Fondation pour la Recherche Médicale (SPF20150934061). The authors thank Lionel Imbert, IBS cell-free facility, for his technical help and valuable advice. This work used the Cell-Free facility at the Grenoble Instruct Centre (ISBG; UMS 3518 CNRS-CEA-UJF-EMBL) with support from Instruct (PID: 1552) within the Grenoble Partnership for Structural Biology (PSB). This work benefited from the HPC resources of the CALMIP supercomputing center under the allocation 2016-P16032.

References

(1) Walker, F. O. Huntington's Disease. *Lancet (London, England)* **2007**, *369* (9557), 218–228.

(2) Wanker, E. E. Protein Aggregation and Pathogenesis of Huntington's Disease: Mechanisms and Correlations. *Biol. Chem.* **2000**, *381* (9–10), 937–942.

(3) DiFiglia, M.; Sapp, E.; Chase, K. O.; Davies, S. W.; Bates, G. P.; Vonsattel, J. P.; Aronin, N. Aggregation of Huntingtin in Neuronal Intranuclear Inclusions and Dystrophic Neurites in Brain. *Science* **1997**, *277* (5334), 1990–1993.

(4) Orr, H. T. Beyond the Qs in the Polyglutamine Diseases. *Genes Dev.* **2001**, *15* (8), 925–932.

(5) Hosp, F.; Gutiérrez-Ángel, S.; Schaefer, M. H.; Cox, J.; Meissner, F.; Hipp, M. S.; Hartl, F.-U.; Klein, R.; Dudanova, I.; Mann, M. Spatiotemporal Proteomic Profiling of Huntington's Disease Inclusions Reveals Widespread Loss of Protein Function. *Cell Rep.* **2017**, *21* (8), 2291–2303.

(6) Mangiarini, L.; Sathasivam, K.; Seller, M.; Cozens, B.; Harper, A.; Hetherington, C.; Lawton, M.; Trottier, Y.; Lehrach, H.; Davies, S. W.; et al. Exon I of the HD Gene with an Expanded CAG Repeat Is Sufficient to Cause a Progressive Neurological Phenotype in Transgenic Mice. *Cell* **1996**, *87* (3), 493–506.

(7) Isas, J. M.; Langen, R.; Siemer, A. B. Solid-State Nuclear Magnetic Resonance on the Static and Dynamic Domains of Huntingtin Exon-1 Fibrils. *Biochemistry* **2015**, *54* (25), 3942–3949.

(8) Hoop, C. L.; Lin, H.-K.; Kar, K.; Magyarfalvi, G.; Lamley, J. M.; Boatz, J. C.; Mandal, A.; Lewandowski, J. R.; Wetzel, R.; van der Wel, P. C. A. Huntingtin Exon 1 Fibrils Feature an Interdigitated β-Hairpin-Based Polyglutamine Core. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (6), 1546–1551.

(9) Jayaraman, M.; Kodali, R.; Sahoo, B.; Thakur, A. K.; Mayasundari, A.; Mishra, R.; Peterson, C. B.; Wetzel, R. Slow Amyloid Nucleation via α-Helix-Rich Oligomeric Intermediates in Short Polyglutamine-Containing Huntingtin Fragments. *J. Mol. Biol.* **2012**, *415* (5), 881–899.

(10) Fiumara, F.; Fioriti, L.; Kandel, E. R.; Hendrickson, W. A. Essential Role of Coiled Coils for Aggregation and Activity of Q/N-Rich Prions and PolyQ Proteins. *Cell* **2010**, *143* (7), 1121–1135.

(11) Scherzinger, E.; Lurz, R.; Turmaine, M.; Mangiarini, L.; Hollenbach, B.; Hasenbank, R.; Bates, G. P.; Davies, S. W.; Lehrach, H.; Wanker, E. E. Huntingtin-Encoded Polyglutamine Expansions Form Amyloid-like Protein Aggregates in Vitro and in Vivo. *Cell* **1997**, *90* (3), 549–558.

(12) Shen, K.; Calamini, B.; Fauerbach, J. A.; Ma, B.; Shahmoradian, S. H.; Serrano Lachapel, I. L.; Chiu, W.; Lo, D. C.; Frydman, J. Control of the Structural Landscape and Neuronal Proteotoxicity of Mutant Huntingtin by Domains Flanking the PolyQ Tract. *Elife* **2016**, *5* (OCTOBER2016), 1–29.

(13) Michalek, M.; Salnikov, E. S.; Bechinger, B. Structure and Topology of the Huntingtin 1-17

Membrane Anchor by a Combined Solution and Solid-State NMR Approach. *Biophys. J.* **2013**, *105* (3), 699–710.

(14) Ceccon, A.; Schmidt, T.; Tugarinov, V.; Kotler, S. A.; Schwieters, C. D.; Clore, G. M. Interaction of Huntingtin Exon-1 Peptides with Lipid-Based Micellar Nanoparticles Probed by Solution NMR and Q-Band Pulsed EPR. *J. Am. Chem. Soc.* **2018**, *140* (20), 6199–6202.

(15) Thakur, A. K.; Jayaraman, M.; Mishra, R.; Thakur, M.; Chellgren, V. M.; Byeon, I.-J. L.; Anjum, D. H.; Kodali, R.; Creamer, T. P.; Conway, J. F.; et al. Polyglutamine Disruption of the Huntingtin Exon 1 N Terminus Triggers a Complex Aggregation Mechanism. *Nat. Struct. Mol. Biol.* **2009**, *16* (4), 380–389.

(16) Tam, S.; Spiess, C.; Auyeung, W.; Joachimiak, L.; Chen, B.; Poirier, M. A.; Frydman, J. The Chaperonin TRiC Blocks a Huntingtin Sequence Element That Promotes the Conformational Switch to Aggregation. *Nat. Struct. Mol. Biol.* **2009**, *16* (12), 1279–1285.

(17) Kotler, S. A.; Tugarinov, V.; Schmidt, T.; Ceccon, A.; Libich, D. S.; Ghirlando, R.; Schwieters, C. D.; Clore, G. M. Probing Initial Transient Oligomerization Events Facilitating Huntingtin Fibril Nucleation at Atomic Resolution by Relaxation-Based NMR. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (9), 3562–3571.

(18) Steffan, J. S.; Agrawal, N.; Pallos, J.; Rockabrand, E.; Trotman, L. C.; Slepko, N.; Illes, K.; Lukacsovich, T.; Zhu, Y.-Z.; Cattaneo, E.; et al. SUMO Modification of Huntingtin and Huntington's Disease Pathology. *Science* **2004**, *304* (5667), 100–104.

(19) Ehrnhoefer, D. E.; Sutton, L.; Hayden, M. R. Small Changes, Big Impact: Posttranslational Modifications and Function of Huntingtin in Huntington Disease. *Neuroscientist* **2011**, *17* (5), 475–492.

(20) Atwal, R. S.; Desmond, C. R.; Caron, N.; Maiuri, T.; Xia, J.; Sipione, S.; Truant, R. Kinase Inhibitors Modulate Huntingtin Cell Localization and Toxicity. *Nat. Chem. Biol.* **2011**, *7* (7), 453–460.

(21) Mishra, R.; Hoop, C. L.; Kodali, R.; Sahoo, B.; van der Wel, P. C. A.; Wetzel, R. Serine Phosphorylation Suppresses Huntingtin Amyloid Accumulation by Altering Protein Aggregation Properties. *J. Mol. Biol.* **2012**, *424* (1–2), 1–14.

(22) Ansaloni, A.; Wang, Z.-M.; Jeong, J. S.; Ruggeri, F. S.; Dietler, G.; Lashuel, H. A. One-Pot Semisynthesis of Exon 1 of the Huntingtin Protein: New Tools for Elucidating the Role of Posttranslational Modifications in the Pathogenesis of Huntington's Disease. *Angew. Chem. Int. Ed. Engl.* **2014**, *53* (7), 1928–1933.

(23) Chiki, A.; DeGuire, S. M.; Ruggeri, F. S.; Sanfelice, D.; Ansaloni, A.; Wang, Z.-M.; Cendrowska, U.; Burai, R.; Vieweg, S.; Pastore, A.; et al. Mutant Exon1 Huntingtin Aggregation Is Regulated by T3 Phosphorylation-Induced Structural Changes and Crosstalk between T3 Phosphorylation and Acetylation at K6. *Angew. Chem. Int. Ed. Engl.* **2017**, *56* (19), 5202–5207.

(24) Bhattacharyya, A.; Thakur, A. K.; Chellgren, V. M.; Thiagarajan, G.; Williams, A. D.; Chellgren, B. W.; Creamer, T. P.; Wetzel, R. Oligoproline Effects on Polyglutamine Conformation and Aggregation. *J. Mol. Biol.* **2006**, *355* (3), 524–535.

(25) Dehay, B.; Bertolotti, A. Critical Role of the Proline-Rich Region in Huntingtin for Aggregation and Cytotoxicity in Yeast. *J. Biol. Chem.* **2006**, *281* (47), 35608–35615.

(26) Feng, X.; Luo, S.; Lu, B. Conformation Polymorphism of Polyglutamine Proteins. *Trends Biochem. Sci.* **2018**, *43* (6), 424–435.

(27) Miller, J.; Arrasate, M.; Brooks, E.; Libeu, C. P.; Legleiter, J.; Hatters, D.; Curtis, J.; Cheung, K.; Krishnan, P.; Mitra, S.; et al. Identifying Polyglutamine Protein Species in Situ That Best Predict Neurodegeneration. *Nat. Chem. Biol.* **2011**, *7* (12), 925–934.

(28) Nucifora, L. G.; Burke, K. A.; Feng, X.; Arbez, N.; Zhu, S.; Miller, J.; Yang, G.; Ratovitski, T.; Delannoy, M.; Muchowski, P. J.; et al. Identification of Novel Potentially Toxic Oligomers Formed in Vitro from Mammalian-Derived Expanded Huntingtin Exon-1 Protein. *J. Biol. Chem.* **2012**, *287* (19), 16017–16028.

(29) Peters-Libeu, C.; Miller, J.; Rutenber, E.; Newhouse, Y.; Krishnan, P.; Cheung, K.; Hatters, D.; Brooks, E.; Widjaja, K.; Tran, T.; et al. Disease-Associated Polyglutamine Stretches in Monomeric Huntingtin Adopt a Compact Structure. *J. Mol. Biol.* **2012**, *421* (4–5), 587–600.

(30) Li, P.; Huey-Tubman, K. E.; Gao, T.; Li, X.; West, A. P.; Bennett, M. J.; Bjorkman, P. J. The Structure of a PolyQ-Anti-PolyQ Complex Reveals Binding According to a Linear Lattice Model. *Nat. Struct. Mol. Biol.* **2007**, *14* (5), 381–387.

(31) Klein, F. A. C.; Zeder-Lutz, G.; Cousido-Siah, A.; Mitschler, A.; Katz, A.; Eberling, P.; Mandel, J.-L.; Podjarny, A.; Trottier, Y. Linear and Extended: A Common Polyglutamine Conformation Recognized by the Three Antibodies MW1, 1C2 and 3B5H10. *Hum. Mol. Genet.* **2013**, *22* (20), 4215–4223.

(32) Owens, G. E.; New, D. M.; West, A. P.; Bjorkman, P. J. Anti-PolyQ Antibodies Recognize a Short PolyQ Stretch in Both Normal and Mutant Huntingtin Exon 1. *J. Mol. Biol.* **2015**, *427* (15), 2507–2519.

(33) Bennett, M. J.; Huey-Tubman, K. E.; Herr, A. B.; West, A. P.; Ross, S. A.; Bjorkman, P. J. A Linear Lattice Model for Polyglutamine in CAG-Expansion Diseases. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (18), 11634–11639.

(34) Warner, J. B.; Ruff, K. M.; Tan, P. S.; Lemke, E. A.; Pappu, R. V; Lashuel, H. A. Monomeric Huntingtin Exon 1 Has Similar Overall Structural Features for Wild-Type and Pathological Polyglutamine Lengths. *J. Am. Chem. Soc.* **2017**, *139* (41), 14456–14469.

(35) Newcombe, E. A.; Ruff, K. M.; Sethi, A.; Ormsby, A. R.; Ramdzan, Y. M.; Fox, A.; Purcell, A. W.; Gooley, P. R.; Pappu, R. V; Hatters, D. M. Tadpole-like Conformations of Huntingtin Exon 1 Are Characterized by Conformational Heterogeneity That Persists Regardless of Polyglutamine Length. *J. Mol. Biol.* **2018**, *430* (10), 1442–1458.

(36) Bravo-Arredondo, J. M.; Kegulian, N. C.; Schmidt, T.; Pandey, N. K.; Situ, A. J.; Ulmer, T. S.; Langen, R. The Folding Equilibrium of Huntingtin Exon 1 Monomer Depends on Its Polyglutamine Tract. *J. Biol. Chem.* **2018**, *293* (51), 19613–19623.

(37) Fodale, V.; Kegulian, N. C.; Verani, M.; Cariulo, C.; Azzollini, L.; Petricca, L.; Daldin, M.; Boggio, R.; Padova, A.; Kuhn, R.; et al. Polyglutamine- and Temperature-Dependent Conformational Rigidity in Mutant Huntingtin Revealed by Immunoassays and Circular Dichroism Spectroscopy. *PLoS One* **2014**, *9* (12), e112262.

(38) Milles, S.; Salvi, N.; Blackledge, M.; Jensen, M. R. Characterization of Intrinsically Disordered Proteins and Their Dynamic Complexes: From in Vitro to Cell-like Environments. *Prog. Nucl. Magn. Reson. Spectrosc.* **2018**, *109*, 79–100.

(39) Urbanek, A.; Morató, A.; Allemand, F.; Delaforge, E.; Fournet, A.; Popovic, M.; Delbecq, S.; Sibille, N.; Bernadó, P. A General Strategy to Access Structural Information at Atomic Resolution in Polyglutamine Homorepeats. *Angew. Chem. Int. Ed. Engl.* **2018**, *57* (14), 3598–3601.

(40) Kigawa, T.; Yabuki, T.; Yoshida, Y.; Tsutsui, M.; Ito, Y.; Shibata, T.; Yokoyama, S. Cell-Free Production and Stable-Isotope Labeling of Milligram Quantities of Proteins. *FEBS Lett.* **1999**, *442* (1), 15–19.

(41) Wang, L.; Xie, J.; Schultz, P. G. Expanding the Genetic Code. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 225–249.

(42) Ellman, J. A.; Volkman, B. F.; Mendel, D.; Schultz, P. G.; Wemmer, D. E. Site-Specific Isotopic Labeling of Proteins for NMR Studies. *J. Am. Chem. Soc.* **1992**, *114* (20), 7959–7961.

(43) Peuker, S.; Andersson, H.; Gustavsson, E.; Maiti, K. S.; Kania, R.; Karim, A.; Niebling, S.; Pedersen, A.; Erdelyi, M.; Westenhoff, S. Efficient Isotope Editing of Proteins for Site-Directed Vibrational Spectroscopy. *J. Am. Chem. Soc.* **2016**, *138* (7), 2312–2318.

(44) Baias, M.; Smith, P. E. S.; Shen, K.; Joachimiak, L. A.; Żerko, S.; Koźmiński, W.; Frydman, J.; Frydman, L. Structure and Dynamics of the Huntingtin Exon-1 N-Terminus: A Solution NMR Perspective. *J. Am. Chem. Soc.* **2017**, *139* (3), 1168–1176.

(45) Eftekharzadeh, B.; Piai, A.; Chiesa, G.; Mungianu, D.; García, J.; Pierattelli, R.; Felli, I. C.; Salvatella, X. Sequence Context Influences the Structure and Aggregation Behavior of a PolyQ Tract. *Biophys. J.* **2016**, *110* (11), 2361–2366.

(46) Nielsen, J. T.; Mulder, F. A. A. POTENCI: Prediction of Temperature, Neighbor and PH-Corrected Chemical Shifts for Intrinsically Disordered Proteins. *J. Biomol. NMR* **2018**, *70* (3), 141–165.

(47) Marsh, J. A.; Singh, V. K.; Jia, Z.; Forman-Kay, J. D. Sensitivity of Secondary Structure Propensities to Sequence Differences between Alpha- and Gamma-Synuclein: Implications for Fibrillation. *Protein Sci.* **2006**, *15* (12), 2795–2804.

(48) Zhang, H.; Neal, S.; Wishart, D. S. RefDB: A Database of Uniformly Referenced Protein

Chemical Shifts. *J. Biomol. NMR* **2003**, *25* (3), 173–195.

(49) Estaña, A.; Sibille, N.; Delaforge, E.; Vaisset, M.; Cortés, J.; Bernadó, P. Realistic Ensemble Models of Intrinsically Disordered Proteins Using a Structure-Encoding Coil Database. *Structure* **2019**, *27* (2), 381–391.

(50) Krivov, G. G.; Shapovalov, M. V; Dunbrack, R. L. Improved Prediction of Protein Side-Chain Conformations with SCWRL4. *Proteins* **2009**, *77* (4), 778–795.

(51) Shen, Y.; Bax, A. SPARTA+: A Modest Improvement in Empirical NMR Chemical Shift Prediction by Means of an Artificial Neural Network. *J. Biomol. NMR* **2010**, *48* (1), 13–22.

(52) MacArthur, M. W.; Thornton, J. M. Influence of Proline Residues on Protein Conformation. *J. Mol. Biol.* **1991**, *218* (2), 397–412.

(53) Iglesias, J.; Sanchez-Martínez, M.; Crehuet, R. SS-Map: Visualizing Cooperative Secondary Structure Elements in Protein Ensembles. *Intrinsically Disord. Proteins* **2013**, *1* (1), e25323.

(54) Baxter, N. J.; Williamson, M. P. Temperature Dependence of 1H Chemical Shifts in Proteins. *J. Biomol. NMR* **1997**, *9* (4), 359–369.

(55) Escobedo, A.; Topal, B.; Kunze, M. B. A.; Aranda, J.; Chiesa, G.; Mungianu, D.; Bernardo-Seisdedos, G.; Eftekharzadeh, B.; Gairí, M.; Pierattelli, R.; et al. Side Chain to Main Chain Hydrogen Bonds Stabilize a Polyglutamine Helix in a Transcription Factor. *Nat. Commun.* **2019**, *10* (1), 2034.

(56) Dasgupta, S.; Bell, J. A. Design of Helix Ends. Amino Acid Preferences, Hydrogen Bonding and Electrostatic Interactions. *Int. J. Pept. Protein Res.* **1993**, *41* (5), 499–511.

(57) Richardson, J. S.; Richardson, D. C. Amino Acid Preferences for Specific Locations at the Ends of Alpha Helices. *Science* **1988**, *240* (4859), 1648–1652.

(58) Seale, J. W.; Srinivasan, R.; Rose, G. D. Sequence Determinants of the Capping Box, a Stabilizing Motif at the N-Termini of Alpha-Helices. *Protein Sci.* **1994**, *3* (10), 1741–1745.

(59) Newell, N. E. Mapping Side Chain Interactions at Protein Helix Termini. *BMC Bioinformatics* **2015**, *16* (1), 231.

(60) Gao, J.; Bosco, D. A.; Powers, E. T.; Kelly, J. W. Localized Thermodynamic Coupling between Hydrogen Bonding and Microenvironment Polarity Substantially Stabilizes Proteins. *Nat. Struct. Mol. Biol.* **2009**, *16* (7), 684–690.

(61) Ramazzotti, M.; Monsellier, E.; Kamoun, C.; Degl'Innocenti, D.; Melki, R. Polyglutamine Repeats Are Associated to Specific Sequence Biases That Are Conserved among Eukaryotes. *PLoS One* **2012**, *7* (2), e30824.

(62) Jorda, J.; Kajava, A. V. Protein Homorepeats. *Adv. Protein Chem. Struct. Biol.* **2010**, *79*, 59–88.

(63) Lobanov, M. Y.; Galzitskaya, O. V. Occurrence of Disordered Patterns and Homorepeats in Eukaryotic and Bacterial Proteomes. *Mol. Biosyst.* **2012**, *8* (1), 327–337.

(64) Mier, P.; Alanis-Lobato, G.; Andrade-Navarro, M. A. Context Characterization of Amino Acid

Homorepeats Using Evolution, Position, and Order. *Proteins* **2017**, *85* (4), 709–719.

(65) Kim, M. W.; Chelliah, Y.; Kim, S. W.; Otwinowski, Z.; Bezprozvanny, I. Secondary Structure of Huntingtin Amino-Terminal Region. *Structure* **2009**, *17* (9), 1205–1212.

(66) De Genst, E.; Chirgadze, D. Y.; Klein, F. A. C.; Butler, D. C.; Matak-Vinković, D.; Trottier, Y.; Huston, J. S.; Messer, A.; Dobson, C. M. Structure of a Single-Chain Fv Bound to the 17 N-Terminal Residues of Huntingtin Provides Insights into Pathogenic Amyloid Formation and Suppression. *J. Mol. Biol.* **2015**, *427* (12), 2166–2178.

(67) DeGuire, S. M.; Ruggeri, F. S.; Fares, M.-B.; Chiki, A.; Cendrowska, U.; Dietler, G.; Lashuel, H. A. N-Terminal Huntingtin (Htt) Phosphorylation Is a Molecular Switch Regulating Htt Aggregation, Helical Conformation, Internalization, and Nuclear Targeting. *J. Biol. Chem.* **2018**, *293* (48), 18540–18558.

(68) Theillet, F.-X.; Kalmar, L.; Tompa, P.; Han, K.-H.; Selenko, P.; Dunker, A. K.; Daughdrill, G. W.; Uversky, V. N. The Alphabet of Intrinsic Disorder: I. Act like a Pro: On the Abundance and Roles of Proline Residues in Intrinsically Disordered Proteins. *Intrinsically Disord. proteins* **2013**, *1* (1), e24360.

(69) Darnell, G.; Orgel, J. P. R. O.; Pahl, R.; Meredith, S. C. Flanking Polyproline Sequences Inhibit Beta-Sheet Structure in Polyglutamine Segments by Inducing PPII-like Helix Structure. *J. Mol. Biol.* **2007**, *374* (3), 688–704.

(70) Schaefer, M. H.; Wanker, E. E.; Andrade-Navarro, M. A. Evolution and Function of CAG/Polyglutamine Repeats in Protein-Protein Interaction Networks. *Nucleic Acids Res.* **2012**, *40* (10), 4273–4287.

(71) Tartari, M.; Gissi, C.; Lo Sardo, V.; Zuccato, C.; Picardi, E.; Pesole, G.; Cattaneo, E. Phylogenetic Comparison of Huntingtin Homologues Reveals the Appearance of a Primitive PolyQ in Sea Urchin. *Mol. Biol. Evol.* **2008**, *25* (2), 330–338.

(72) Darling, A. L.; Uversky, V. N. Intrinsic Disorder in Proteins with Pathogenic Repeat Expansions. *Molecules* **2017**, *22* (12), 2027.

(73) Williamson, T. E.; Vitalis, A.; Crick, S. L.; Pappu, R. V. Modulation of Polyglutamine Conformations and Dimer Formation by the N-Terminus of Huntingtin. *J. Mol. Biol.* **2010**, *396* (5), 1295–1309.

(74) Walters, R. H.; Murphy, R. M. Examining Polyglutamine Peptide Length: A Connection between Collapsed Conformations and Increased Aggregation. *J. Mol. Biol.* **2009**, *393* (4), 978–992.

(75) Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. Fluorescence Correlation Spectroscopy Shows That Monomeric Polyglutamine Molecules Form Collapsed Structures in Aqueous Solutions. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (45), 16764–16769.

(76) Zoghbi, H. Y.; Orr, H. T. Glutamine Repeats and Neurodegeneration. *Annu. Rev. Neurosci.* **2000**, *23*, 217–247.

(77) Whelihan, E. F.; Schimmel, P. Rescuing an Essential Enzyme-RNA Complex with a Non-Essential Appended Domain. *EMBO J.* **1997**, *16* (10), 2968–2974.

(78) Walker, S. E.; Fredrick, K. Preparation and Evaluation of Acylated TRNAs. *Methods* **2008**, *44* (2), 81–86.

(79) Loscha, K. V; Herlt, A. J.; Qi, R.; Huber, T.; Ozawa, K.; Otting, G. Multiple-Site Labeling of Proteins with Unnatural Amino Acids. *Angew. Chem. Int. Ed. Engl.* **2012**, *51* (9), 2243–2246.

(80) Apponyi, M. A.; Ozawa, K.; Dixon, N. E.; Otting, G. Cell-Free Protein Synthesis for Analysis by NMR Spectroscopy. *Methods Mol. Biol.* **2008**, *426* (15), 257–268.

(81) Vranken, W. F.; Boucher, W.; Stevens, T. J.; Fogh, R. H.; Pajon, A.; Llinas, M.; Ulrich, E. L.; Markley, J. L.; Ionides, J.; Laue, E. D. The CCPN Data Model for NMR Spectroscopy: Development of a Software Pipeline. *Proteins* **2005**, *59* (4), 687–696.

(82) Markley, J. L.; Bax, A.; Arata, Y.; Hilbers, C. W.; Kaptein, R.; Sykes, B. D.; Wright, P. E.; Wüthrich, K. Recommendations for the Presentation of NMR Structures of Proteins and Nucleic Acids. *J. Mol. Biol.* **1998**, *280* (5), 933–952.

SUPPLEMENTARY INFORMATION

Flanking regions define the conformation of the poly-glutamine homo-repeat in huntingtin through opposite structural mechanisms

Annika Urbanek[1,#], Matija Popovic[1,#], Anna Morató[1], Alejandro Estaña[1,2], Carlos A. Elena-Real[1], Pablo Mier[3], Aurélie Fournet[1], Frédéric Allemand[1], Stephane Delbecq[4], Miguel A. Andrade-Navarro[3], Juan Cortés[2], Nathalie Sibille[1], Pau Bernadó[1,*]

[1] Centre de Biochimie Structurale (CBS), INSERM, CNRS, Université de Montpellier. 29, rue de Navacelles, 34090 Montpellier. France.

[2] LAAS-CNRS, Université de Toulouse, CNRS, 31400 Toulouse, France.

[3] Institute of Organismic and Molecular Evolution, Johannes Gutenberg University of Mainz, Mainz, Germany.

[4] Laboratoire de Biologie Cellulaire et Moléculaire (LBCM-EA4558 Vaccination Antiparasitaire), UFR Pharmacie, Université de Montpellier, Montpellier, France.

# These authors contributed equally to this work

Corresponding Author: Pau Bernadó (pau.bernado@cbs.cnrs.fr)

Figure S1

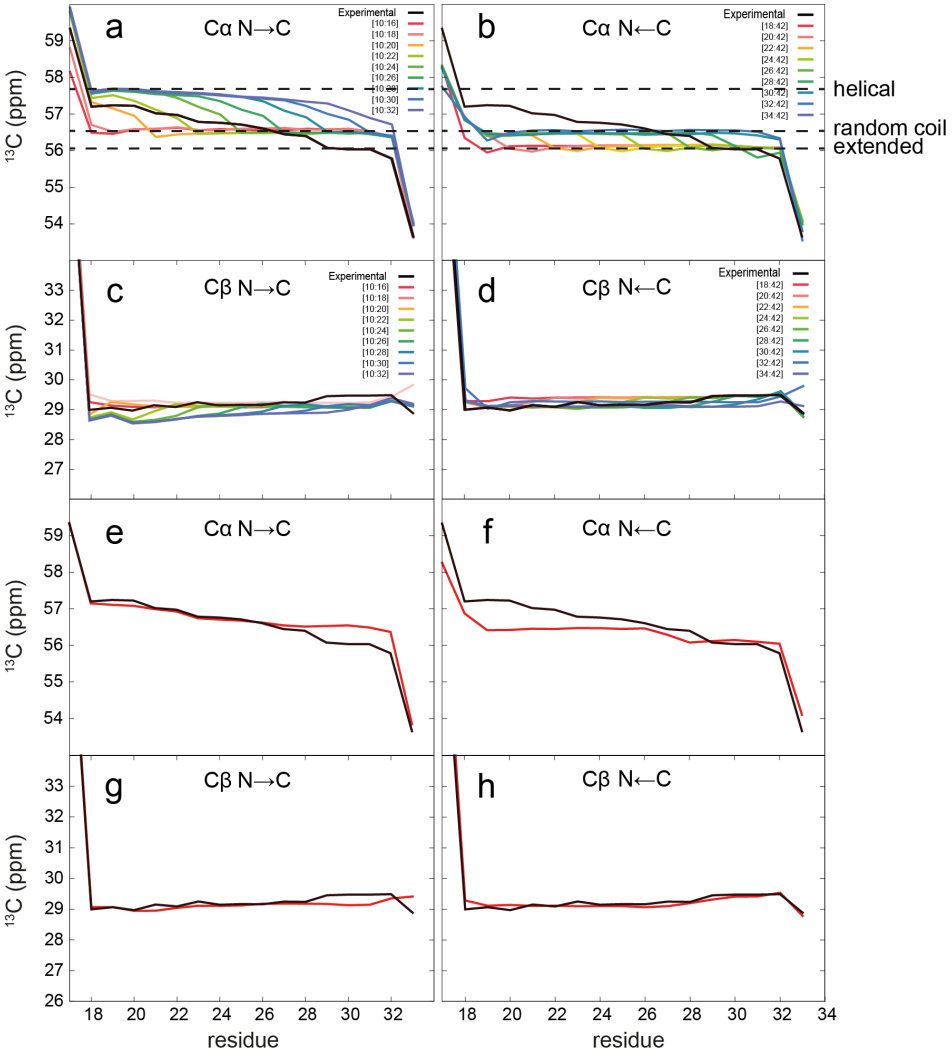


Figure S1. Chemical Shift (CS) based ensemble refinement of H16. (a-d) Overlay of the experimental (black) and theoretical (color) Cα **(a and b)** and Cβ **(c and d)** CSs of the poly-Q tract computed for the N→C **(a and c)** and the N←C **(b and d)** families of ensembles. The ensembles were built by incrementally incorporating glutamines in the partially structured region from Q18-Q33 (N→C) and from Q33 to Q18 (N←C). The boundaries chosen for the partially structured regions for each of the 5,000 conformation ensembles are indicated in the panel. Horizontal dashed lines indicate the three plateaus corresponding to helical, random coil and extended averaged conformations. Experimental (black) *vs* ensemble-optimized (red) Cα **(e and f)** and Cβ **(g and h)** CSs exclusively fitted with the N→C **(e and g)** or the N←C **(f and h)** ensembles. For the N→C only Cα and Cβ experimental CSs from residues Q18-Q28 were used in the optimization. The reweighted model consisted of [10:17; 46.17%], [10:22; 26.93%], [10:23; 5.92%], and [10:26; 20.98%], where [X:Y] refers to the first and last residues considered as partially structured in the model. For the N←C only Cα and Cβ experimental CSs from residues Q29-Q33 were used in the optimization. The reweighted model consisted of [22:42; 46.05%], [28:42, 0.8%], [29:42, 42.86%, and [34:42, 10.28%].

Figure S2

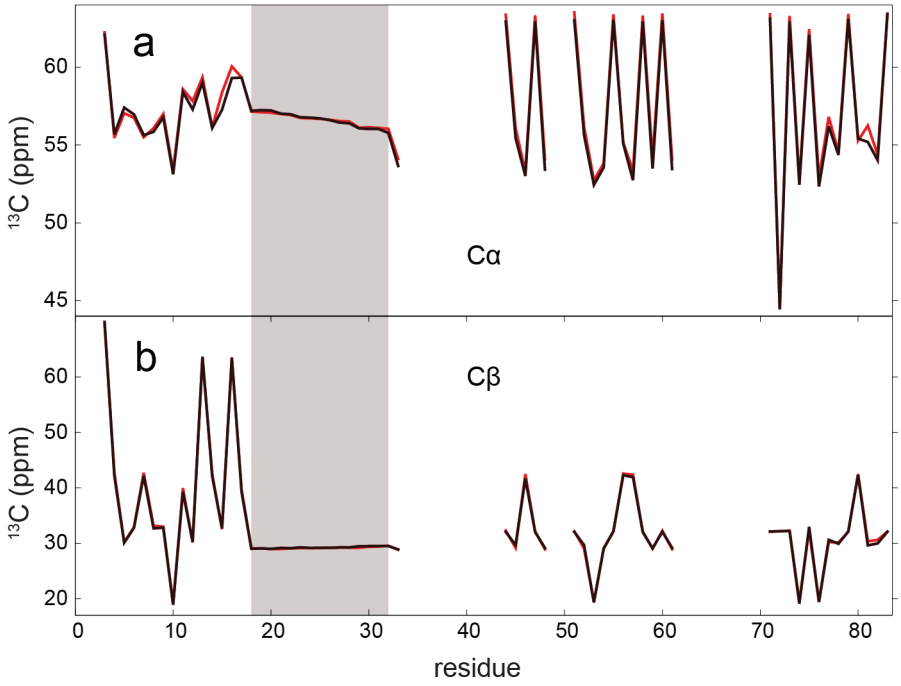


Figure S2. Chemical Shift (CS) based ensemble refinement of H16. Experimental (black) *vs.* ensemble-optimized (red) for all (**a**) Cα and (**b**) Cβ CSs measured for H16. The M1-Q28 and the Q29-P83 fragments of the optimized profile were built from those optimized using the N→C and N←C ensembles, respectively. The poly-Q tract is shaded in gray; gaps are due to proline residues for which we do not have experimental data.

Figure S3

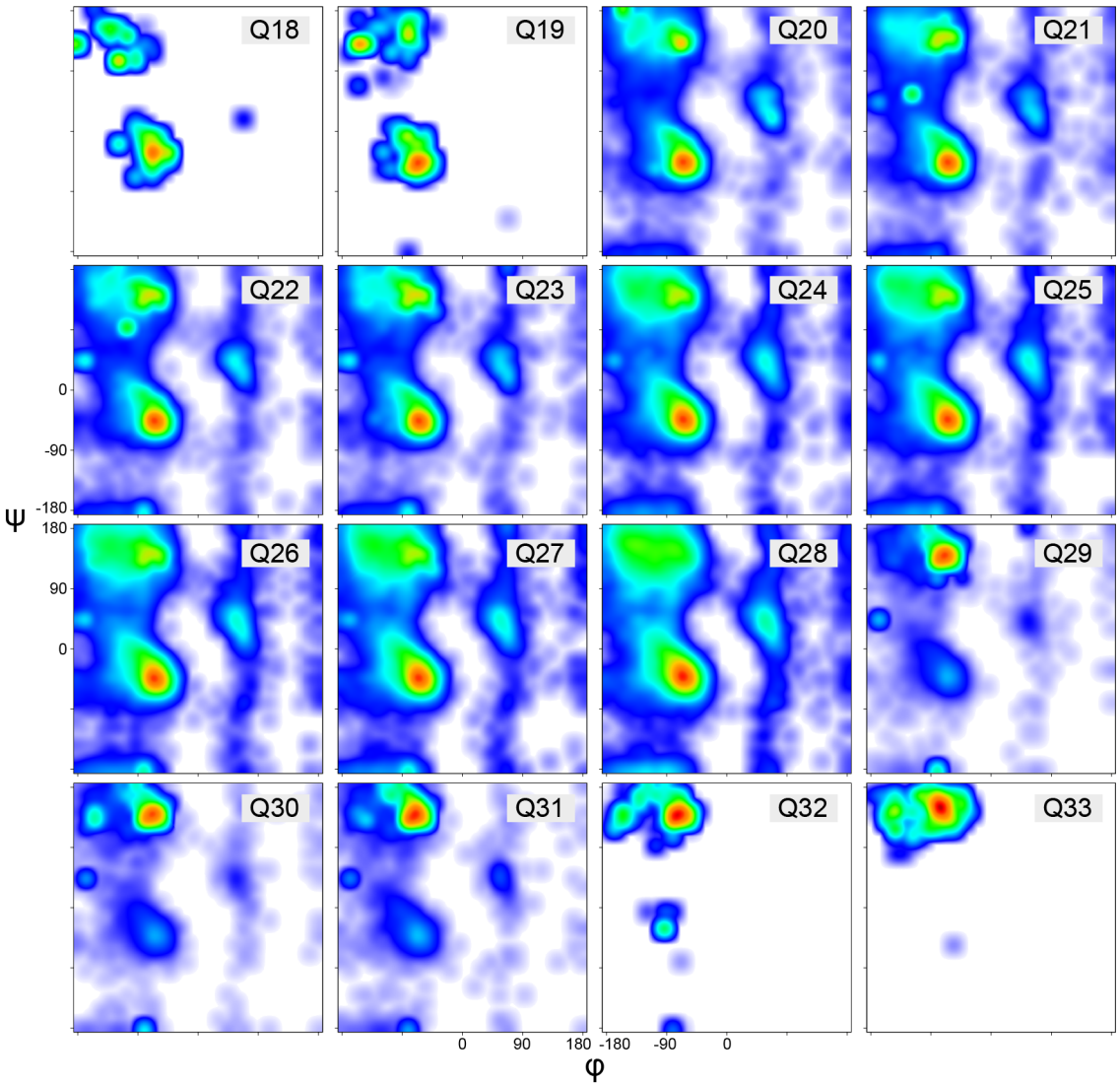


Figure S3. Ramachandran distribution for the glutamine residues of the poly-Q tract of H16 derived from the CS-optimized ensemble of H16. The populations of α-helical ($0 > \varphi$; $50 > \psi > -120$), extended ($0 > \varphi$; $-120 > \psi > 50$), and *other* ($0 < \varphi$) conformations are reported in Figure 3c in the main text. The red color indicates a higher density of conformations.

Figure S4

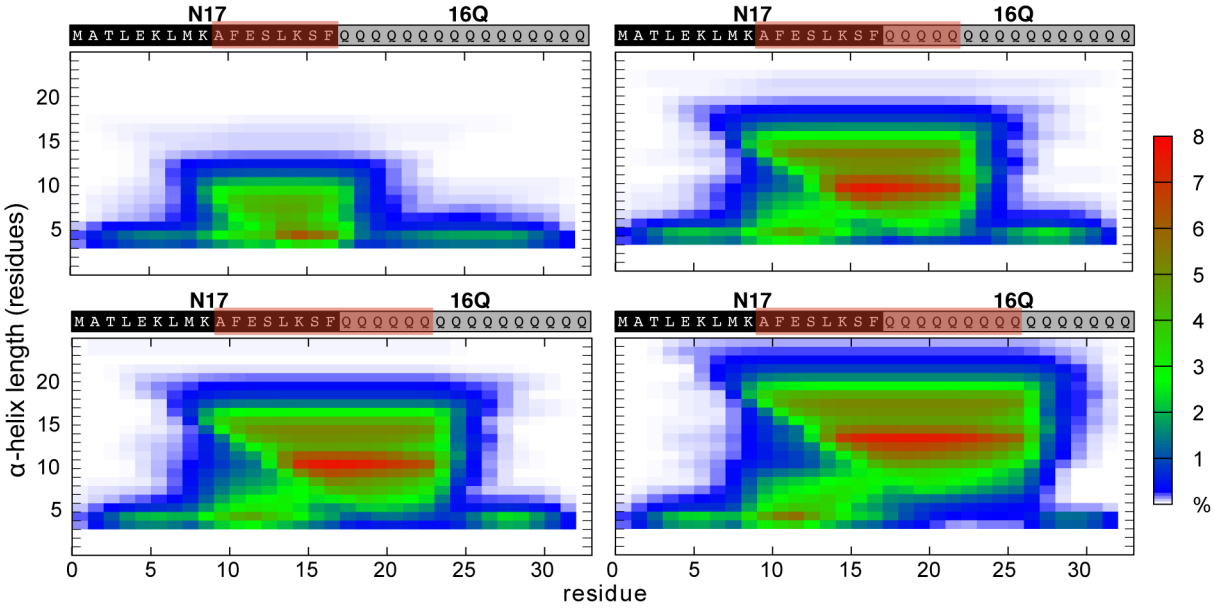


Figure S4. SS-maps for the four ensembles selected using the experimental CSs. Fragments considered as partially structured are highlighted in red on the top of each panel. The percentage of the α-helical fragments is indicated in a color scale from high (red) to low (blue) populations.

Figure S5

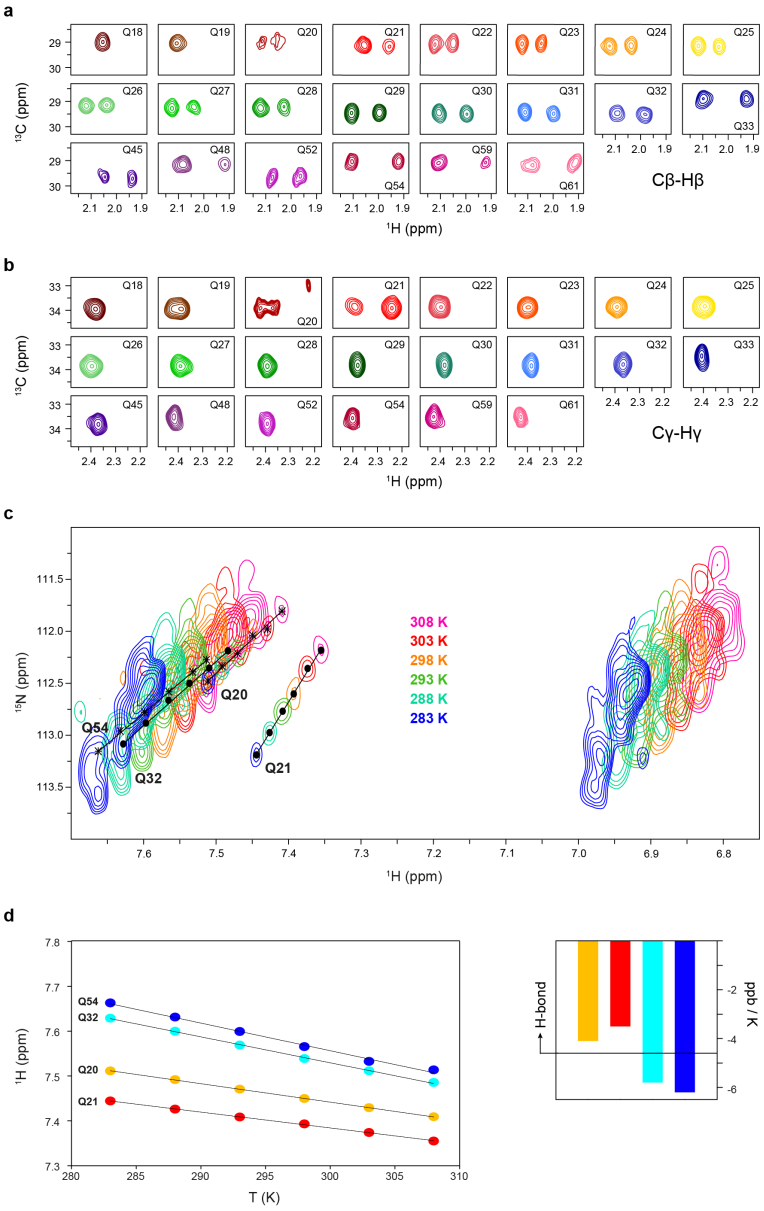


Figure S5: Side chain NMR scanning and temperature factors. (**a**) Cβ and (**b**) Cγ ^{13}C-HSQC spectra for all glutamines in H16. With the exception of the first four glutamines (Q18-Q21), both families of spectra display a canonical behavior where Cβ-H_2 and Cγ-H_2 are doublets and singlets, respectively. The color code is equivalent to the one used in Figure 1 in the main text. (**c**) Zoom of the Nε-H_2 region of the ^{15}N-HSQC of fully labeled H16 measured at different temperatures. Solid lines connect the centers of the peaks for Q20, Q21, Q32 and Q54 at the different temperatures. Other peaks could not be unambiguously identified and therefore were not used in the analysis. (**d, left panel**) Linear fit of the ^{1}H frequency of the Nε-H_2 peaks of the four residues plotted against the temperature. The slope of the linear fit (**d, right panel**) reports on the temperature coefficient and the probability of the atom to be involved in a hydrogen bond. According to these slopes, Q20 and Q21 side chains form a hydrogen bond, while Q32 and Q54 side chains do not.

Figure S6

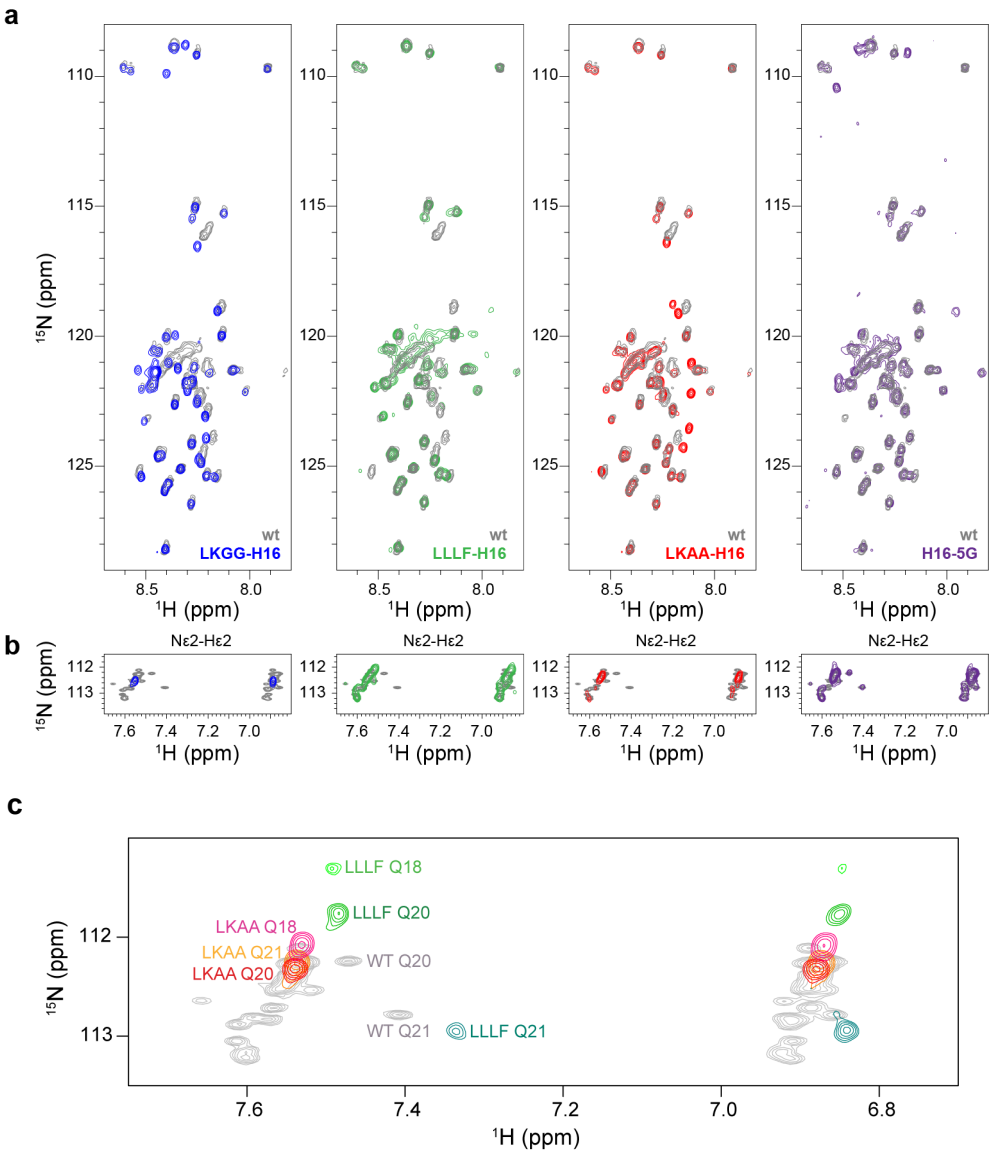


Figure S6: ^{15}N-HSQC spectra of the structural mutants in comparison with the wild-type. (**a**) Backbone spectra of wild-type H16 (grey) overlaid with LKGG-H16 (blue), LLLF-H16 (green), LKAA-H16 (red) and H16-5G (purple). (**b**) Side chain spectra corresponding to the same mutants using the same color code. Note that peaks corresponding to the first glutamines for the LLLF-H16 and LKAA-H16 are not displayed due to the lower intensity (see below). (**c**) Zoom of the Nε-H_2 peaks for residues Q18, Q20 and Q21 of mutants LLLF-H16 and LKAA-H16 obtained from SSIL samples overlaid to the same region for the wild type protein in gray. Their chemical shifts substantiate the hydrogen bond network involving F17 in httex1 (see main text).

Figure S7

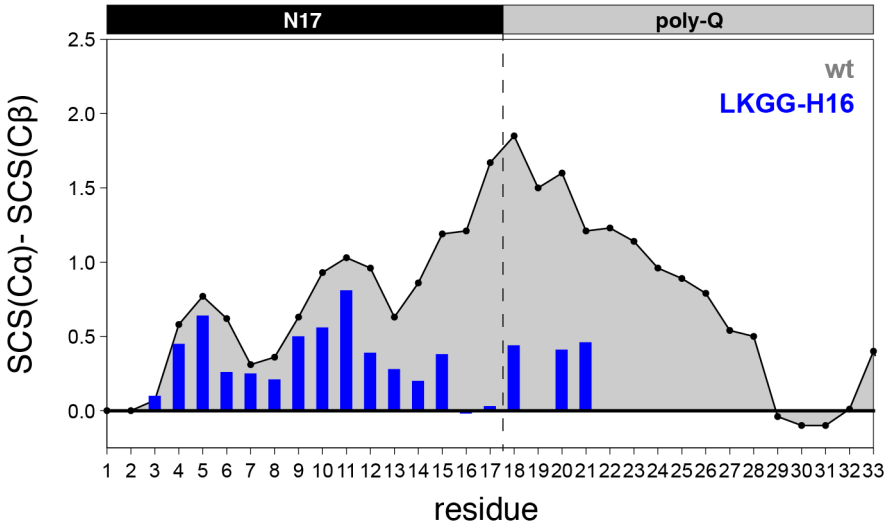


Figure S7. SCS analysis of LKGG-H16 in comparison with the wild-type. Secondary chemical shift analysis using experimental Cα and Cβ chemical shifts and a neighbor-corrected random-coil library POTENCI[46] for the wild-type (connected black points) and the LKGG-H16 mutant (blue histogram). Only residues belonging to N17 and the poly-Q tract are displayed. Data from Q18, Q20 and Q21 was obtained from SSIL samples; the other glutamines were not investigated.

Figure S8

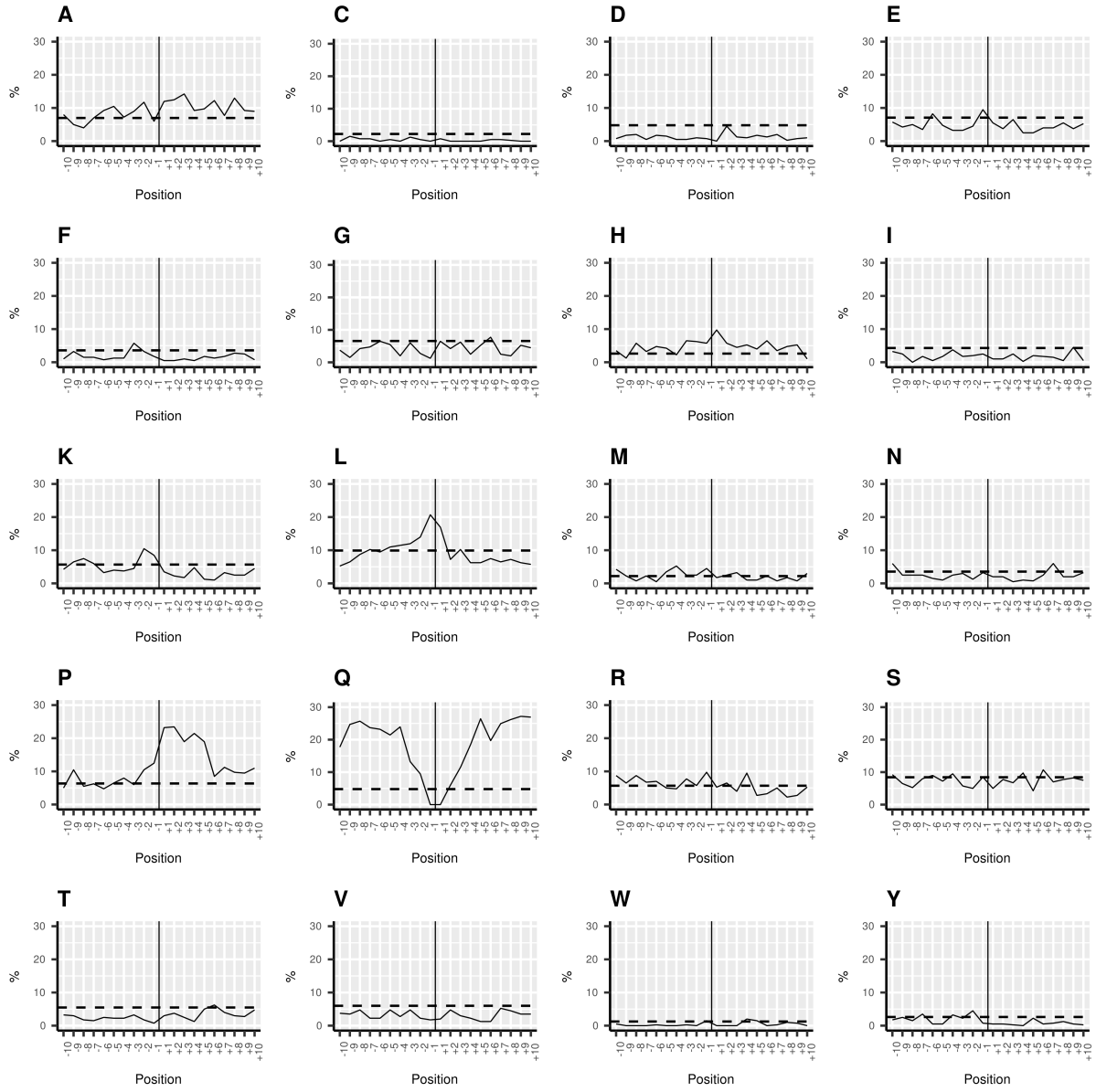


Figure S8. Compositional analysis of the poly-Q flanking regions in human proteins. Percentage for each one of the 20 natural amino acids in the positions preceding (-10 to -1) and following (+1 to +10) the poly-Q tracts in human proteins. The solid vertical line corresponds to the position of the poly-Q tract. Poly-Q tracts were defined as having a maximum of 2 non-glutamine residues in fragments of 10 or more glutamine residues. Horizontal dashed lines define the percentage for each amino acid type in the human proteome.

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Structural Biology

Small-angle scattering studies of intrinsically disordered proteins and their complexes

Tiago N Cordeiro[1], Fátima Herranz-Trillo[1,3], Annika Urbanek[1], Alejandro Estaña[1,2], Juan Cortés[2], Nathalie Sibille[1] and Pau Bernadó[1]

Intrinsically Disordered Proteins (IDPs) perform a broad range of biological functions. Their relevance has motivated intense research activity seeking to characterize their sequence/structure/function relationships. However, the conformational plasticity of these molecules hampers the application of traditional structural approaches, and new tools and concepts are being developed to address the challenges they pose. Small-Angle Scattering (SAS) is a structural biology technique that probes the size and shape of disordered proteins and their complexes with other biomolecules. The low-resolution nature of SAS can be compensated with specially designed computational tools and its combined interpretation with complementary structural information. In this review, we describe recent advances in the application of SAS to disordered proteins and highly flexible complexes and discuss current challenges.

Addresses
[1] Centre de Biochimie Structurale, INSERM U1054, CNRS UMR 5048, Université de Montpellier, 29, rue de Navacelles, 34090 Montpellier, France
[2] LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France
[3] Department of Drug Design and Pharmacology, University of Copenhagen, Universitetsparken 2, 2100 Copenhagen, Denmark

Corresponding author: Bernadó, Pau (pau.bernado@cbs.cnrs.fr)

Current Opinion in Structural Biology 2016, **42**:15–23

This review comes from a themed issue on **Proteins: bridging theory and experiment**

Edited by **Igor N Berezovsky** and **Ugo Bastolla**

For a complete overview see the Issue and the Editorial

Available online 26th October 2016

http://dx.doi.org/10.1016/j.sbi.2016.10.011

Introduction

In the last two decades, Intrinsically Disordered Proteins or Regions (IDPs/IDRs) have emerged as fundamental molecules in a broad range of crucial biological functions such as cell signaling, regulation, and homeostasis [1,2,3••]. Due to their lack of a permanent secondary and tertiary structure, IDPs and IDRs are highly plastic and have the capacity to perform specialized functions that complement those of their globular (folded) counterparts [4]. Disordered regions, which can finely adapt to the structural and chemical features of their partners, are very well suited for protein–protein interactions and are thus abundant in hub positions of interactomes [5–7].

The importance of disordered proteins in a multitude of biological processes has fostered intense research efforts that seek to unravel the structural bases of their function. Nuclear Magnetic Resonance (NMR) has been the main structural biology technique used to characterize the conformational preferences at residue level, and, therefore, to localize partially structured elements [8,9]. However, a number of structural features related to the overall size and shape of IDPs or their complexes remain elusive to NMR. To study these properties, thereby complementing NMR residue-specific information, Small-Angle Scattering (SAS) of X-rays (SAXS) or Neutrons (SANS) is the most appropriate technique [10–12]. Although SAS is a low-resolution technique, the data obtained is sensitive to large-scale protein fluctuations and the presence of multiple species and/or conformations in solution [13–15]. However, the conversion of SAS properties into structural restraints is challenging due to the enormous conformational variability of IDPs and the ensemble-averaged nature of the experimental data [16]. The quantitative analysis of these data in terms of structure has prompted the development of computational approaches to both model disordered proteins and to use ensembles of conformations to describe the experimental data. Here we highlight the most relevant developments and applications of SAS to IDPs and IDRs, with a special emphasis on the computational strategies required to fully exploit the data in order to achieve biologically insightful information.

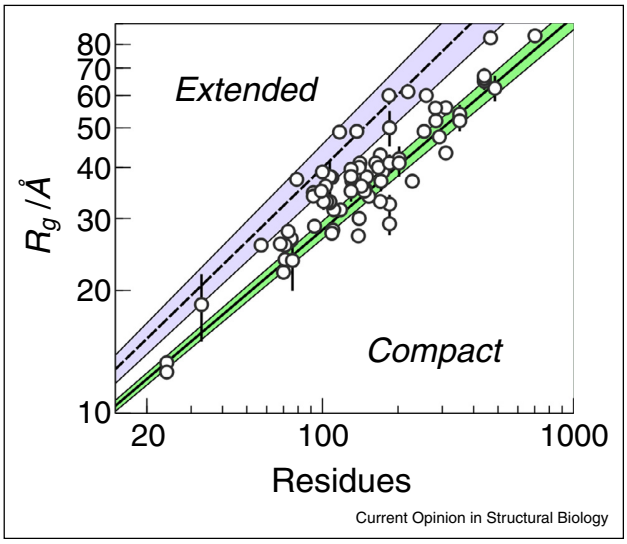
Structural models of IDPs and their experimental validation

For disordered proteins, the structural insights gained from overall SAS parameters, such as the radius of gyration, R_g, the pairwise intramolecular distance distribution, $p(r)$, and the maximum intramolecular distance, D_{max}, are limited. Neither these parameters nor the traditional Kratky representation, $I(s)s^2$ versus s where $I(s)$ represents the scattering intensity and s the momentum transfer, which qualitatively report on the compactness of biomolecules in solution, directly account for the ensemble

nature of disordered proteins. In order to fully exploit the structural and dynamic information encoded in SAS data, it is necessary to use realistic three-dimensional (3D) models. However, the generation of conformational ensembles of disordered proteins is extremely challenging, mainly because of the flat energy landscape and the large number of local minima separated by low-energy barriers [17]. The most popular methods to generate 3D models of IDPs are based on residue-specific conformational landscapes derived from large databases of crystallographic structures [18,19,20•]. However, the main limitation of these approaches is the absence of sequence context information, thereby precluding the prediction of transiently formed secondary structure elements or the presence of long-range interactions between distant regions of the protein. Accurate energy models (force-fields) accounting for the interactions within the chain and with the solvent are required to describe these features. The development of specific force-fields to study conformational fluctuations in disordered proteins is a very active field of research [21–24]. Molecular Dynamics (MD) or Monte-Carlo (MC) simulations, when an appropriate energy description is provided, are suitable methods to correctly sample the conformational space of IDPs. However, the high-dimensionality and the breadth of the energy landscape hamper exhaustive exploration of this space. Replica Exchange MD (REMD) [25,26], which exchanges conformations between parallel simulations running at multiple temperatures, or Multiscale Enhanced Sampling (MSES) [27], which couples temperature and Hamiltonian replica exchange, have been proposed to enhance the conformational exploration of MD methods. The performance of MD-based methods can also be improved by the inclusion of experimental data to delimit the exploration to the most relevant regions of the conformational space [28–30].

The quality of computational models of disordered proteins is normally validated using experimental data. The R_g derived from the low-angle region of SAXS curves or from the $p(r)$ function is an excellent probe of the overall size of a particle in solution. R_g compilations have been extensively used to validate models of denatured and natively disordered proteins through Flory's relationship, which correlates the R_g observed with the number residues of the chain [31,14]. The compilation of the R_gs from 76 IDPs (Figure 1) reveals that these proteins are more compact than chemically denatured ones. It has been shown that denatured proteins present an enhanced sampling of extended conformations, probably due to the interaction of the protein with chemical agents [32]. Importantly, deviations from the expected R_g values for canonical random-coil behavior, which is represented by the green line in Figure 1, indicate the presence of structural features that modify the overall size of the particle in solution towards more extended or more compact (Figure 1). The extendedness detected using this analysis for several Tau protein constructs has been linked to the presence of secondary structural elements probed by NMR [29]. These structural properties can be more thoroughly examined when the complete SAXS curve is used to validate the ensemble models of peptides [33] or proteins [19,34,35].

Figure 1



R_g values from 76 IDPs as a function of the number of residues of the protein are plotted in Log–Log scale. Only proteins lacking a permanent secondary or tertiary structure were considered for the compilation. Proteins with ordered domains, molten globules, or denatured proteins were not considered. Straight lines correspond to Flory's relationships parametrized for denatured proteins using experimental data (purple-dashed) [31] and IDPs using computational ensembles calculated with Flexible-Meccano (green-solid) [32]. Colored bands correspond to uncertainty of the parametrization for both models. Some IDPs contain local structural features and consequently they are globally more extended or more compact than expected for a random coil. These structural features, even if transient, can be manifested in the experimental R_g.

Ensemble approaches

In the last decade, ensemble methods have become highly popular to structurally characterize disordered proteins. Guided by experimental data, these methods aim to derive accurate ensemble models of flexible proteins. Several strategies that apply these methods to SAS data have been reported: Ensemble Optimization Method (EOM) [36,37]; Minimal Ensemble Search (MES) [38]; Basis-Set Supported SAXS (BSS-SAXS) [39]; Maximum Occurrence (MAX-Occ) [40]; Ensemble Refinement of SAXS (EROS) [41]; Broad Ensemble Generator with Re-weighting (BEGR) [42]; and Bayesian Ensemble SAXS (BE-SAXS) [43]. These methods share a common strategy that consists of the following three consecutive steps: (i) computational generation of a large ensemble that describes the conformational landscape of

the protein; (ii) calculation of the theoretical SAXS curves from the individual conformations; and (iii) use of a multiparametric optimization method to select a sub-ensemble of conformations that collectively describe the experimental profile. Despite the common strategy, these approaches present distinct features in the three steps. Readers are referred to the original articles for detailed descriptions. The availability of ensemble methods has transformed the study of flexible proteins by SAS. Ensemble methods provide a description in terms of the statistical distributions of structural parameters or conformations that is revolutionary with respect to traditional analyses based on averaged parameters extracted from raw data. Using this power, structural perturbations exerted by temperature [44•,45], buffer composition [46], or mutations [47] have been monitored in terms of ensembles of conformations.

Despite the popularity of ensemble methods, several aspects are still under debate. The most relevant ones are the use of discrete descriptions for entities that probe an astronomical number of conformations, and the statistical significance of ensembles derived from data containing a very limited amount of information. The strategies described use distinct philosophies to address these issues, including the search for the minimum number of conformations to describe the data [37,38], the representation of the optimal solution as a distribution of low-resolution structural parameters such as R_g or D_{max} [36], and the application of Bayesian statistics [39,43] or maximum entropy approaches [41]. Regardless of the strategy used to derive an ensemble of conformations compatible with the experimental data, one must be careful on the structural interpretation of the final solution. The optimized ensemble is a representation of the behavior of the protein in solution and not the exact enumeration of the conformations adopted by the protein. Consequently, the final ensemble can only be used to derive structural features that describe the protein. Importantly, the nature of these features depends on the experimental data used to derive the model. If only SAS data have been used, then an assessment of the degree of flexibility, and the size and shape distributions sampled by the protein can be obtained from the ensemble. Conversely, conformational preferences at residue level can be extracted if NMR information probing structure in a residue-specific manner is used along the refinement.

Enriching the definition of conformational ensembles of IDPs with complementary information

The definition of protein ensembles derived from SAS data using ensemble methods is limited to the overall structure and the space sampled by the protein in solution. Although this is an important improvement with respect to classical approaches, several crucial features, such as the localization of secondary structural elements or compact regions, remain elusive using this approach. Considerable research efforts have been channeled into enriching the resolution of the resulting ensemble with complementary information.

NMR is the only technique that can provide atomic-resolution information on IDPs and, consequently, it is the most common method applied in combination with SAS [48]. NMR is highly versatile and can measure multiple observables reporting on protein structure and dynamics [49]. Concretely, information reporting on the backbone conformational preferences at residue level can be probed by means of time-averaged and ensemble-averaged chemical-shifts (CSs), J-couplings and Residual-Dipolar Couplings (RDCs). NMR can also probe long-range interactions within a protein chain or in protein complexes through Paramagnetic Relaxation Enhancement (PRE) experiments. In these experiments, a stable radical or a paramagnetic metal is introduced in a specific position of the chain, and the spatially close atoms can be identified by a decrease in their signal intensity that is proportional to the distance.

The best manner to exploit the complementarity between NMR and SAS is to integrate the experimental data into the same refinement protocol. The programs ENSEMBLE [50,51] and ASTEROIDS [52] derive ensembles of disordered proteins by collectively describing SAXS curves, in addition to several NMR observables. These powerful approaches seek to find the appropriate way to combine data with very different information content while avoiding overfitting. In a pioneering study, ensembles of Tau and α-synuclein were determined by combining SAXS with multiple backbone CS, RDC, and PRE datasets [53••]. Those authors addressed the optimal combination of experimental data and the overfitting problem with extensive cross-validation tests that substantiated conformational bias in the aggregation-nucleation regions for both proteins.

Other structural techniques such as single molecule Fluorescence Resonance Energy Transfer (smFRET) [54] and Electron Paramagnetic Resonance (EPR) [55,56••] have been combined with SAXS to study large and flexible complexes. Recent developments in Mass Spectrometry (MS) offer novel sources of structural information [57]. Ion Mobility Spectrometry (IMS) can capture, in a similar way to SAS, the overall properties of conformational ensembles of disordered proteins. However, a recent study comparing IMS and SAXS data for some IDPs suggests that the conformations sampled in solution and in gas-phase are not equivalent [58]. Hydrogen/Deuterium Exchange MS (HDX/MS) probes structural elements in proteins by identifying regions that are protected from the exchange with solvent protons [57]. The availability of fast HDX/MS methods enables the

exploration of secondary structural elements in IDPs and localizing their interaction sites with globular partners [59]. In a recent study HDX/MS information was combined with SAXS to study the calcium-induced structure formation in RD, a protein hosting repeated regions able to bind this cation [60].

The structural definition of a SAXS derived ensemble model can also be enriched by the simultaneous analysis of curves measured for multiple deletion mutants of the same IDP [36]. When applied to two different isoforms of Tau protein, this approach identified the repeat region of the protein as the origin of distinct global rearrangements of its flanking regions [61].

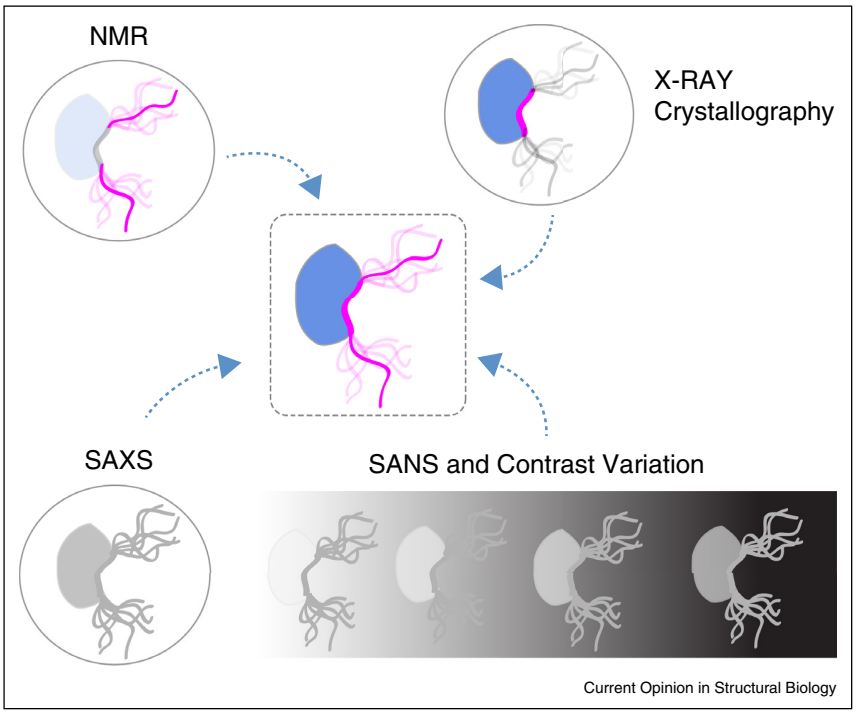
The large toolbox of structural techniques that can probe distinct structural features of IDPs will result in a better understanding on their structure–function relationship. In this regard, the future development of robust and reliable ways to integrate biophysical measurements in ensemble approaches is imperative when addressing complex biomolecular entities such as IDPs and their complexes.

Disordered proteins in complexes

The biological function of many IDPs is manifested when they recognize their biological folded partners [5]. This recognition frequently involves linear motifs of the disordered chain, which, upon binding, adopt relatively fixed conformations while the rest of the IDP remains flexible [62].

The relevance of protein–protein complexes involving disordered partners has promoted growing interest in unraveling their structural characterization, with the aim to understand the bases of their biological activity. This structural characterization is complex and poses multiple challenges to traditional structural biology methods. SAXS has emerged as a valuable alternative. However, overall structural parameters or *ab initio* reconstructions derived from SAXS curves cannot capture the inherent plasticity of these complexes [63,64•,65]. Hybrid (or integrative) methods that combine information from multiple techniques, thus exploiting their individual strengths, are the most appropriate approaches to study highly flexible complexes [66]. In this context, it is important to describe how different structural biology

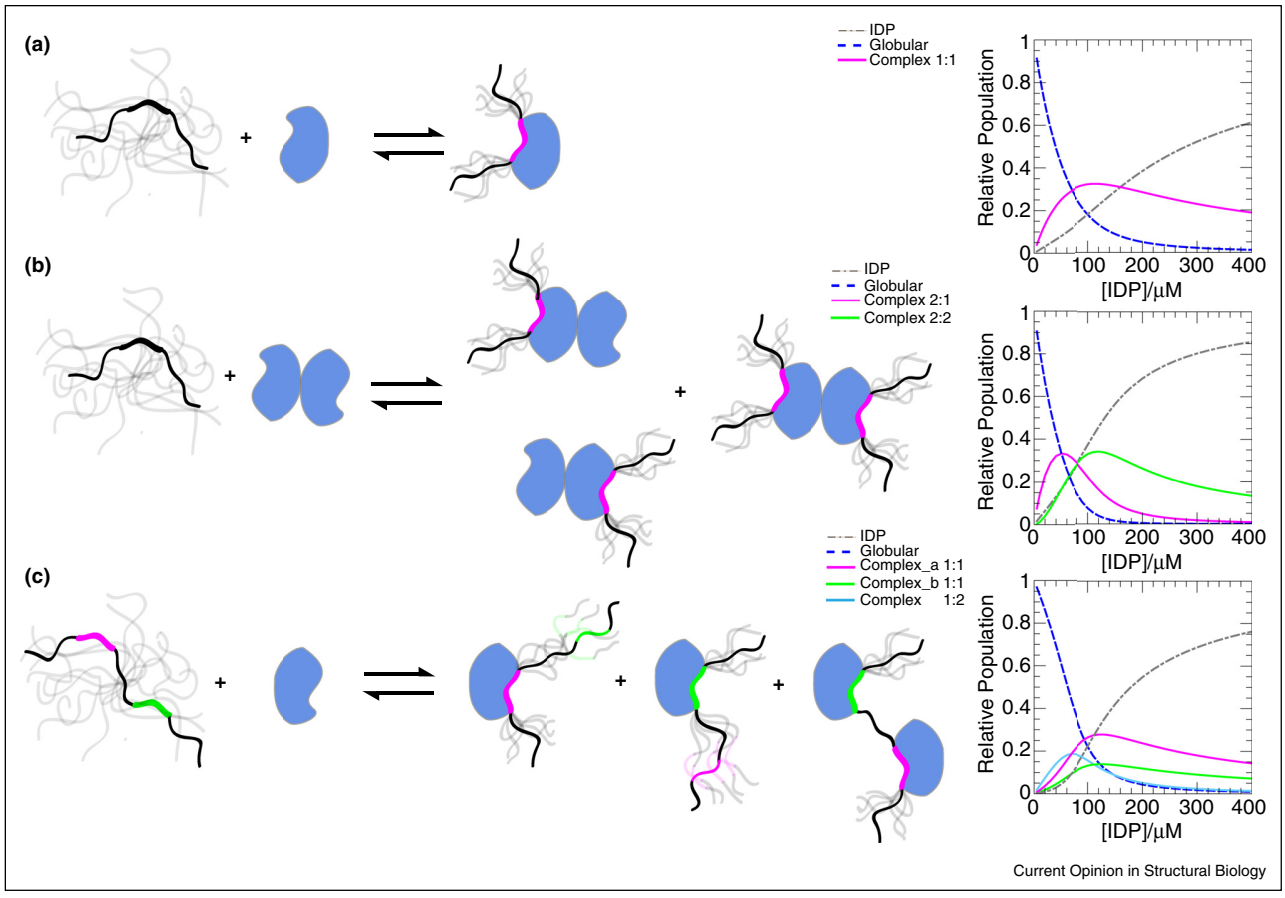
Figure 2



Cartoons representing the structural sensitivity of NMR, X-ray crystallography, and SAS for a complex involving a disordered protein (central cartoon). NMR normally probes the flexible regions of these complexes while the globular partner and the interacting region remain invisible. Crystallography provides detailed information of the interacting region of the complex but not for the flexible parts. SAXS probes the complete ensemble, although the details cannot be assessed due to its inherent low-resolution. SANS, through contrast variation experiments, can probe independently both partners in the context of the complex depending on the deuteration level of the partners and the D_2O/H_2O of the buffer. SAS is an ideal tool to integrate NMR and crystallographic information to build complete structural and dynamic models of disordered biomolecular complexes.

techniques probe complexes involving IDPs (Figure 2). Due to the dynamic nature of the interaction and the distinct hydrodynamic properties of the globular and disordered parts of the complex, NMR generally detects only those regions that remain flexible upon binding. Although not general, it is sometimes possible to crystallize the globular partner in the presence of a small peptide corresponding to the interacting region of the IDP. Therefore, X-ray crystallography provides an atomic resolution picture of the interacting regions that is complementary to NMR since the two techniques probe non-overlapping parts of the same entity [67]. Conversely, SAXS probes the complete assembly and can be used to integrate the information from both NMR and X-ray crystallography. If one of the partners is deuterated, contrast variation SANS experiments can be performed and the individual components of the assembly can be alternatively highlighted depending on the D_2O/H_2O ratio of the buffer. The power of combining multiple techniques is exemplified in the study of the interaction of the Vesicular Stomatitis Virus (VSV) nucleoprotein (N^0) and the dimeric phosphoprotein (P), a high-affinity complex that precludes the oligomerization of N^0 *in vivo* [68••]. Using EOM, the authors simultaneously fitted one SAXS curve and four SANS curves measured at different contrast levels for the complex of N^0 with deuterated P

Figure 3



Examples of polydisperse scenarios that can occur in low-affinity complexes involving an IDP and a globular partner. **(a)** Both proteins have a single binding site. The complex is in equilibrium with the free forms of both proteins. **(b)** The globular partner is a dimer and has two identical binding sites. The free forms are in equilibrium with three possible complexes recognizing one or two binding sites of the globular partner. Due to the symmetry of the dimer, the two singly bound complexes are however indistinguishable by SAS. **(c)** The IDP presents two similar binding sites (pink and green). The free forms are in equilibrium with two 1:1 complexes using a distinct IDP interacting site to bind the globular partner, and a complex where the IDP simultaneously interacts with two globular partners. On the right part of the figure, three panels are displayed representing the molar fraction of each species along a simulated titration experiment for each scenario. These populations were computed assuming a fixed concentration of the globular partner, [globular] = 100 μM, and increasing concentrations of IDP, [IDP], from 1 μM to 400 μM. A common dissociation constant K_d = 20 μM was used for scenarios A and B, in panel C the two IDP binding sites, pink and green, display a K_d = 20 μM and 40 μM, respectively. These panels exemplify the inherent polydispersity of moderate affinity complexes, and how multiple titration experiments will probe differently the species present and their relative populations.

protein. The additional information provided by the distinct contribution of the two proteins in the SANS experiments notably improved the description of the conformational properties of the complex.

In many cases, the conformational mobility of the interacting region of the IDP is reduced (or frozen) upon binding to the biological partner. There is an entropic cost associated with this rigidification that often leads to low-affinity to moderate-affinity complexes ($K_d > 1$ μM) [62]. The structural modulation of the affinity is key to achieving tunable responses to external signals, thereby explaining the prevalent role of disordered proteins in signaling processes [2,3]. In the concentration range normally used in SAXS experiments, the complex is in equilibrium with the free forms of the two partners, thereby giving rise to population-weighted averaged SAXS curves (Figure 3a). This scenario can be even more complex if one or both of the partners have multiple equivalent or similar binding sites (Figure 3b,c). In this case, the polydispersity of the mixture increases as a result of the presence of several complexes with distinct stoichiometries.

The interpretation of SAS data from polydisperse samples is challenging [69]. Although the coupling of SAXS to Size-Exclusion Chromatography (SEC-SAXS) can, in some instances, separate the components of the mixture, there are multiple examples where the coexistence of multiple species is unavoidable. In these circumstances and with the aim to isolate the contribution of the individual species within complex mixtures, analytical approaches have been developed to decompose large SAXS titration datasets [70,71]. This decomposition is easier when prior structural knowledge of the species is used for the analysis [69]. However, to apply this strategy to low-affinity flexible complexes, accurate conformational descriptions of all species in the free and bound forms are mandatory. The analysis of SAS data measured in samples with different relative concentrations of both partners seems the most appropriate strategy to enrich the information content in order to structurally characterize these extremely challenging scenarios (Figure 3).

Conclusions and outlook

During the last decade, SAS has been added to the toolbox of techniques used to study conformational fluctuations in proteins. This dynamic revolution of SAS is linked to the development of computational tools able to describe the conformational landscape of biomolecules and ensemble approaches with the capacity to interpret SAS data in terms of structural variability. These computational tools, which use chemical and structural knowledge of biomolecules, partially compensate for the limited amount of information coded in a SAS curve. Therefore, the capacity to fully exploit the structural information held in SAS data will necessarily be linked to the development of more advanced and precise computational approaches with specially developed force-fields. This notion is especially applicable to IDPs and IDRs, which populate a huge number of conformational states. For these proteins, SAS can be enriched with complementary information obtained by NMR, smFRET, EPR, or MS, and integrated into a common ensemble model embedding structure and dynamics. A particularly challenging subclass of IDPs is that containing Low-Complexity Regions (LCRs), which are involved in multitude of biological processes and are related to severe pathologies. LCRs are unusually simple protein sequences with a strong amino acid composition bias. The resulting similarity of chemical environments within their sequence hampers their structural characterization by NMR. SAS can be a valuable alternative through which to study this important but structurally neglected family of proteins [72–74].

The function of multitude of IDPs is determined by their interaction with biomolecular partners to form assemblies, which, in many cases, are of low to moderate affinity. The capacity of SAS to probe the size and shape of particles in solution places this technique in a unique position to address these polydisperse scenarios. A case in point is the fibrillation process that several IDPs undergo to form amyloids, which are linked to severe diseases. The decomposition of time-dependent SAXS datasets has been successfully used to characterize intermediate oligomeric forms [75•,76], thereby validating SAXS as a practical tool for this purpose.

The need to understand the mechanisms underlying complex cellular processes and recent technical and conceptual advances in structural biology techniques across the board have prompted researchers to tackle challenging systems that were inaccessible some years ago. Many of these systems are inherently dynamic and/or polydisperse and can be exquisitely probed by SAS. As a consequence, we anticipate that SAS will take on greater relevance in hybrid approaches where its unique information will be synergistically integrated with data from multiple sources to deliver accurate structural and dynamic models of disordered proteins and their complexes.

Conflict of interest statement

The authors declare no conflict of interest.

Acknowledgements

This work was supported by the ERC-CoG chemREPEAT, SPIN-HD–Chaires d'Excellence 2011 from the *Agence National de Recherche* (ANR), ATIP-Avenir, and the French Infrastructure for Integrated Structural Biology (FRISBI – ANR-10-INSB-05-01) to PB. FHT is supported by INSERM and the Sapere Aude Programme SAFIR of the University of Copenhagen. AU is supported by a grant from the *Fondation pour la Recherche Médicale*.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z: **Intrinsic disorder and protein function**. *Biochemistry* 2002, **41**:6573-6582.

2. Wright PE, Dyson HJ: **Intrinsically disordered proteins in cellular signalling and regulation**. *Nat Rev Mol Cell Biol* 2015, **16**:18-29.

3. •• Csizmok V, Follis AV, Kriwacki RW, Forman-Kay JD: **Dynamic protein interaction networks and new structural paradigms in signaling**. *Chem Rev* 2016, **116**:6424-6462.
Excellent review with a very complete list of references on the unique mechanisms used by disordered proteins to perform very specific functions in signaling processes. A description is provided of the present knowledge on the emerging field of phase separation induced by the interaction of IDRs with RNA.

4. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z: **Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions**. *J Proteome Res* 2007, **6**:1882-1898.

5. Tompa P, Schad E, Tantos A, Kalmar L: **Intrinsically disordered proteins: emerging interaction specialists**. *Curr Opin Struct Biol* 2015, **35**:49-59.

6. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN: **Flexible nets. The roles of intrinsic disorder in protein interaction networks**. *FEBS J* 2005, **272**:5129-5148.

7. Kim PM, Sboner A, Xia Y, Gerstein M: **The role of disorder in interaction networks: a structural analysis**. *Mol Systems Biol* 2008, **4**:179.

8. Dyson HJ, Wright PE: **Unfolded proteins and protein folding studied by NMR**. *Chem Rev* 2004, **104**:3607-3622.

9. Jensen MR, Ruigrok RWH, Blackledge M: **Describing intrinsically disordered proteins at atomic resolution by NMR**. *Curr Opin Struct Biol* 2013, **23**:426-435.

10. Feigin LA, Svergun DI: *Structure Analysis by Small-angle X-ray and Neutron Scattering*. Plenum Press; 1987.

11. Putnam CD, Hammel M, Hura GL, Tainer JA: **X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution**. *Quart Rev Biophys* 2007, **40**:191-285.

12. Jacques DA, Trewhella J: **Small-angle scattering for structural for structural biology-expanding the frontier while avoiding the pitfalls**. *Protein Sci* 2010, **19**:642-657.

13. Doniach S: **Changes in biomolecular conformation seen by small angle X-ray scattering**. *Chem Rev* 2001, **101**:1763-1778.

14. Bernadó P, Svergun DI: **Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering**. *Mol Biosyst* 2012, **8**:151-167.

15. Receveur-Brechot V, Durand D: **How random are intrinsically disordered proteins? A small angle scattering perspective**. *Curr Protein Pept Sci* 2012, **13**:55-75.

16. Bernadó P, Blackledge M: **Structural biology: proteins in dynamic equilibrium**. *Nature* 2010, **468**:1046-1048.

17. Zhou H-X: **Polymer models of protein stability, folding, and interactions**. *Biochemistry* 2004, **43**:2141-2154.

18. Jha AK, Colubri A, Freed KF, Sosnick TR: **Statistical coil model of the unfolded state: resolving the reconciliation problem**. *Proc Natl Acad Sci USA* 2005, **102**:13099-13104.

19. Bernadó P, Blanchard L, Timmins P, Marion D, Ruigrok RW, Blackledge M: **A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering**. *Proc Natl Acad Sci USA* 2005, **102**:17002-17007.

20. • Ozenne V, Bauer F, Salmon L, Huang J, 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.
Description of Flexible-Meccano. This software computes ensembles of IDPs based on the conformational sampling found in coil regions of crystallographic structures. The program computes averaged RDCs and PREs from the ensembles, and provides scripts to add side chains, and to compute CSs and SAXS data.

21. Vitalis A, Pappu RV: **ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions**. *J Comput Chem* 2009, **30**:673-699.

22. Best RB, Zheng W, Mittal J: **Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association**. *J Chem Theory Comput* 2014, **10**:5113-5124.

23. Henriques J, Cragnell C, Skepö M: **Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment**. *J Chem Theory Comput* 2015, **11**:3420-3431.

24. Mercadante D, Milles S, Fuertes G, Svergun DI, Lemke EA, Gräter F: **Kirkwood-buff approach rescues overcollapse of a disordered protein in canonical protein force fields**. *J Phys Chem B* 2015, **119**:7975-7984.

25. Chebaro Y, Ballard AJ, Chakraborty D, Wales DJ: **Intrinsically disordered energy landscapes**. *Sci Rep* 2015, **5**:10386.

26. Zerze GH, Miller CM, Granata D, Mittal J: **Free energy surface of an intrinsically disordered protein: comparison between temperature replica exchange molecular dynamics and bias-exchange metadynamics**. *J Chem Theory Comput* 2015, **11**:2776-2782.

27. Lee KH, Chen J: **Multiscale enhanced sampling of intrinsically disordered protein conformations**. *J Comput Chem* 2016, **37**:550-557.

28. Dedmon M, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson CM: **Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations**. *J Am Chem Soc* 2005, **127**:476-477.

29. Mukrasch MD, Markwick P, Biernat J, Bergen Mv, Bernadó P, Griesinger C, Mandelkow E, Zweckstetter M, Blackledge M: **Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation**. *J Am Chem Soc* 2007, **129**:5235-5243.

30. Wu K-P, Weinstock DS, Narayanan C, Levy RM, Baum J: **Structural reorganization of α-synuclein at low pH observed by NMR and REMD simulations**. *J Mol Biol* 2009, **391**:784-796.

31. Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, Cingel N, Dothager RS, Seifert S, Thiyagarajan P, Sosnick TR *et al.*: **Random-coil behavior and the dimensions of chemically unfolded proteins**. *Proc Natl Acad Sci USA* 2004, **101**:12491-12496.

32. Bernadó P, Blackledge M: **A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering**. *Biophys J* 2009, **97**:2839-2845.

33. Zagrovic B, Lipfert J, Sorin EJ, Millett IS, van Gunsteren WF, Doniach S, Pande VS: **Unusual compactness of a polyproline type II structure**. *Proc Natl Acad Sci USA* 2005, **102**:11698-11703.

34. Wells M, Tidow H, Rutherford TJ, Markwick P, Jensen MR, Mylonas E, Svergun DI, Blackledge M, Fersht AR: **Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain**. *Proc Natl Acad Sci USA* 2008, **105**:5762-5767.

35. De Biasio A, Ibáñez de Opakua A, Cordeiro TN, Villate M, Merino N, Sibille N, Lelli M, Diercks T, Bernadó P, Blanco FJ: **p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins**. *Biophys J* 2014, **106**:865-874.

36. Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI: **Structural characterization of flexible proteins using small-angle X-ray scattering**. *J Am Chem Soc* 2007, **129**:5656-5664.

37. Tria G, Mertens HD, Kachala M, Svergun DI: **Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering**. *IUCrJ* 2015, **2**:207-217.

38. Pelikan M, Hura GL, Hammel M: **Structure and flexibility within proteins as identified through small angle X-ray scattering**. *Gen Physiol Biophys* 2009, **28**:174-189.

39. Yang S, Blachowicz L, Makowski L, Roux B: **Multidomain assembled states of Hck tyrosine kinase in solution**. *Proc Natl Acad Sci USA* 2010, **107**:15757-15762.

40. Bertini I, Giachetti A, Luchinat C, Parigi G, Petoukhov MV, Pierattelli R, Ravera E, Svergun DI: **Conformational space of flexible biological macromolecules from average data**. *J Am Chem Soc* 2010, **132**:13553-13558.

41. Rozycki B, Kim YC, Hummer G: **SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions**. *Structure* 2011, **19**:109-116.

42. Daughdrill GW, Kashtanov S, Stancik A, Hill SE, Helms G, Muschol M, Receveur-Bréchot V, Ytreberg FM: **Understanding the structural ensembles of a highly extended disordered protein**. *Mol Biosyst* 2012, **8**:308-319.

43. Antonov LD, Olsson S, Boomsma W, Hamelryck T: **Bayesian inference of protein ensembles from SAXS data**. *Phys Chem Chem Phys* 2016, **18**:5832-5838.

44. • Shkumatov AV, Chinnathambi S, Mandelkow E, Svergun DI: **Structural memory of natively unfolded tau protein detected by small-angle X-ray scattering**. *Proteins* 2011, **79**:2122-2131.
Interesting article reporting on the conformational changes experienced by Tau protein when submitted to temperature jumps. Although the authors do not provide a precise explanation on the origin of this phenomenon, it reflects that there are probably structural features in IDPs that have not been unveiled yet.

45. Kjaergaard M, Nørholm AB, Hendus-Altenburger R, Pedersen SF, Poulsen FM, Kragelund BB: **Temperature-dependent structural changes in intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II?** *Protein Sci* 2010, **19**:1555-1564.

46. Leyrat C, Jensen MR, Ribeiro EA Jr, Gérard FC, Ruigrok RW, Blackledge M, Jamin M: **The N(0)-binding region of the vesicular stomatitis virus phosphoprotein is globally disordered but contains transient α-helices**. *Protein Sci* 2011, **20**:542-556.

47. Stott K, Watson M, Howe FS, Grossmann JG, Thomas JO: **Tail-mediated collapse of HMGB1 is dynamic and occurs via differential binding of the acidic tail to the A and B domains**. *J Mol Biol* 2010, **403**:706-722.

48. Sibille N, Bernadó P: **Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS**. *Biochem Soc Trans* 2012, **40**:955-962.

49. Jensen MR, Zweckstetter M, Huang JR, Blackledge M: **Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy**. *Chem Rev* 2014, **114**:6632-6660.

50. Marsh JA, Neale C, Jack FE, Choy WY, Lee AY, Crowhurst KA, Forman-Kay JD: **Improved structural characterizations of the drkN SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure**. *J Mol Biol* 2007, **367**:1494-1510.

51. Krzeminski M, Marsh JA, Neale C, Choy WY, Forman-Kay JD: **Characterization of disordered proteins with ENSEMBLE**. *Bioinformatics* 2013, **29**:398-399.

52. Jensen MR, Houben K, Lescop E, Blanchard L, Ruigrok RW, Blackledge M: **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* 2008, **130**:8055-8061.

53. •• Schwalbe M, Ozenne V, Bibow S, Jaremko M, Jaremko L, Gajda M, Jensen MR, Biernat J, Becker S, Mandelkow E *et al.*: **Predictive atomic resolution descriptions of intrinsically disordered hTau40 and α-synuclein in solution from NMR and small angle scattering**. *Structure* 2014, **22**:238-249.
Seminal study on the structural properties of α-synuclein and Tau. Using ASTEROIDS the authors interpreted complete CS, RDC and PRE datasets in combination with SAXS data to deliver conformational ensembles of both proteins. The most relevant part part of the article is the extensive cross-validation analyses that demonstrate the accuracy of the structural models.

54. Delaforge E, Milles S, Bouvignies G, Bouvier D, Boivin S, Salvi N, Maurin D, Martel A, Round A, Lemke EA *et al.*: **Large-scale conformational dynamics control H5N1 influenza polymerase PB2 binding to importin α**. *J Am Chem Soc* 2015, **137**:15122-15134.

55. Boura E, Rózycki B, Herrick DZ, Chung HS, Vecer J, Eaton WA, Cafiso DS, Hummer G, Hurley JH: **Solution structure of the ESCRT-I complex by small-angle X-ray scattering, EPR, and FRET spectroscopy**. *Proc Natl Acad Sci USA* 2011, **108**:9437-9442.

56. •• Boura E, Różycki B, Chung HS, Herrick DZ, Canagarajah B, Cafiso DS, Eaton WA, Hummer G, Hurley JH: **Solution structure of the ESCRT-I and -II supercomplex: implications for membrane budding and scission**. *Structure* 2012, **20**:874-886.
In this study the ensemble description of the flexible complex formed by ESCRT-I and II is obtained. The authors integrate SAXS, smFRET and EPR data to define the complex. A protocol is introduced in order to find the minimal number of components in the ensemble with the capacity to properly describe the three sources of data.

57. Konermann L, Vahidi S, Sowole MA: **Mass spectrometry methods for studying structure and dynamics of biological macromolecules**. *Anal Chem* 2014, **86**:213-232.

58. Borysik AJ, Kovacs D, Guharoy M, Tompa P: **Ensemble methods enable a new definition for the solution to gas-phase transfer of intrinsically disordered proteins**. *J Am Chem Soc* 2015, **137**:13807-13817.

59. Keppel TR, Weis DD: **Mapping residual structure in intrinsically disordered proteins at residue resolution using millisecond hydrogen/deuterium exchange and residue averaging**. *J Am Soc Mass Spectrom* 2015, **26**:547-554.

60. O'Brien DP, Hernandez B, Durand D, Hourdel V, Sotomayor-Pérez AC, Vachette P, Ghomi M, Chamot-Rooke J, Ladant D, Brier S, Chenal A: **Structural models of intrinsically disordered and calcium-bound folded states of a protein adapted for secretion**. *Sci Rep* 2015, **5**:14223.

61. Mylonas E, Hascher A, Bernadó P, Blackledge M, Mandelkow E, Svergun DI: **Domain conformation of tau protein studied by solution small-angle X-ray scattering**. *Biochemistry* 2008, **47**:10345-10353.

62. Sharma R, Raduly Z, Miskei M, Fuxreiter M: **Fuzzy complexes: specific binding without complete folding**. *FEBS Lett* 2015, **589**:2533-2542.

63. Shell SS, Putnam CD, Kolodner RD: **The N terminus of *Saccharomyces cerevisiae* Msh6 is an unstructured tether to PCNA**. *Mol Cell* 2007, **26**:565-578.

64. • Rochel N, Ciesielski F, Godet J, Moman E, Roessle M, Peluso-Iltis C, Moulin M, Haertlein M, Callow P, Mély Y *et al.*: **Common architecture of nuclear receptor heterodimers on DNA direct repeat elements with different spacings**. *Nat Struct Mol Biol* 2011, **18**:564-570.
Using a combination of SAXS, SANS and smFRET the authors studied the structure of three hormonal nuclear receptor heterodimers in complex with cognate dsDNA and intrinsically disordered co-activators. Although using *ab initio* reconstructions, the asymmetric singly-bound nature of the complex with the co-activator is demonstrated. Excellent study that highlights the power of SAXS/SANS to characterize complex biomolecular entities.

65. Devarakonda S, Gupta K, Chalmers MJ, Hunt JF, Griffin PR, Van Duyne GD, Spiegelman BM: **Disorder-to-order transition underlies the structural basis for the assembly of a transcriptionally active PGC-1α/ERRγ complex**. *Proc Natl Acad Sci USA* 2011, **108**:18678-18683.

66. Różycki B, Boura E: **Large, dynamic, multi-protein complexes: a challenge for structural biology**. *J Phys Condens Matter* 2014, **26**:463103.

67. De Biasio A, de Opakua AI, Mortuza GB, Molina R, Cordeiro TN, Castillo F, Villate M, Merino N, Delgado S, Gil-Cartón D *et al.*: **Structure of p15(PAF)-PCNA complex and implications for clamp sliding during DNA replication and repair**. *Nat Commun* 2015, **6**:6439.

68. •• Yabukarski F, Leyrat C, Martinez N, Communie G, Ivanov I, Ribeiro EA Jr, Buisson M, Gerard FC, Bourhis JM, Jensen MR *et al.*: **Ensemble structure of the highly flexible complex formed between vesicular stomatitis virus unassembled nucleoprotein and its phosphoprotein chaperone**. *J Mol Biol* 2016, **428**:2671-2694.
In this study the ensemble structure of the viral complex between the nucleocapsid protein N^0 and the phosphoprotein P is determined. The ensemble description is performed using the EOM approach. The novelty of the study is the simultaneous description of the SAXS data of the complex with SANS curves measured at four different contrast levels. This is the first study that profits from the rich information from contrast variation in the context of highly flexible biomolecular complexes using ensemble approaches.

69. Tuukkanen AT, Svergun DI: **Weak protein–ligand interactions studied by small-angle X-ray scattering**. *FEBS J* 2014, **281**:1974-1987.

70. Blobel J, Bernadó P, Svergun DI, Tauler R, Pons M: **Low-resolution structures of transient protein–protein complexes using small-angle X-ray scattering**. *J Am Chem Soc* 2009, **131**:4378-4386.

71. Chandola H, Williamson TE, Craig BA, Friedman AM, Bailey-Kellogg C: **Stoichiometries and affinities of interacting proteins from concentration series of solution scattering data: decomposition by least squares and quadratic optimization**. *J Appl Crystallogr* 2014, **47**:899-914.

72. Greving I, Dicko C, Terry A, Callow P, Vollrath F: **Small angle neutron scattering of native and reconstituted silk fibroin**. *Soft Matter* 2010, **6**:4389.

73. Boze H, Marlin T, Durand D, Pérez J, Vernhet A, Canon F, Sarni-Manchado P, Cheynier V, Cabane B: **Proline-rich salivary proteins have extended conformations**. *Biophys J* 2010, **99**:656-665.

74. Owens GE, New DM, West AP, Bjorkman PJ: **Anti-PolyQ antibodies recognize a short PolyQ stretch in both normal and mutant huntingtin exon 1**. *J Mol Biol* 2015, **427**:2507-2519.

75. • Vestergaard B, Groenning M, Roessle M, Kastrup JS, van de Weert M, Flink JM, Frokjaer S, Gajhede M, Svergun DI: **A helical structural nucleus is the primary elongating unit of insulin amyloid fibrils**. *PLoS Biol* 2007, **5**:1089-1097.
Pioneering study on the characterization of fibrillating proteins using SAXS. The fibrillation of insulin is monitored by SAXS in a time-dependent manner. The resulting curves are the population-weigted averages of all species co-existing in solution. In an arduous procedure, the species-pure curves for the three main components of the mixtures were decomposed allowing their structural characterization including their molecular weight, oligomerization state, and 3D arrangement.

76. Giehm L, Svergun DI, Otzen DE, Vestergaard B: **Low-resolution structure of a vesicle disrupting alpha-synuclein oligomer that accumulates during fibrillation**. *Proc Natl Acad Sci USA* 2011, **108**:3246-3251.