Computational Study of the Structure and Dynamics of Androgen Receptor Polyglutamine Tract

Busra Topal

Computational Study of the Structure and Dynamics of Androgen Receptor Polyglutamine Tract

Busra Topal

UNIVERSITAT DE BARCELONA

Universitat de Barcelona
Facultat de Farmàcia i Ciències de L'alimentació

Institute for Research in Biomedicine (IRB Barcelona)

September 2019

Universitat de Barcelona

Facultat de Farmàcia i Ciències de L'alimentació

Programa de Doctorat

Computational Study of the Structure and Dynamics of Androgen Receptor Polyglutamine Tract

Memòria presentada per Busra Topal per optar al títol de doctor per la Universitat de Barcelona

Director i Tutor
Xavier Salvatella

Doctoranda
Busra Topal

Busra Topal
2019

To my family

PUBLISHED AND SUBMITTED CONTENT

1. Busra Topal, Albert Escobedo, Micha B. A. Kunze, Juan Aranda, Giulio Chiesa, Daniele Mungianu, Ganeko Bernardo-Seisdedos, Bahareh Eftekharzadeh, Margarida Gairi, Roberta Pierattelli, Isabella C. Felli, Tammo Diercks, Oscar Millet, Jesus Garcia, Modesto Orozco, Ramon Crehuet, Kresten Lindorff-Larsen & Xavier Salvatella. *Side chain to main chain hydrogen bonds stabilize a polyglutamine helix in a transcription factor.* Nature Communications volume 10, Article number: 2034 (2019). Paper can be accessed from: `https://www.nature.com/articles/s41467-019-09923-2`.

    - The material in this paper is used entirely in Chapter 3 and Chapter 4 of this thesis.

Contents

List of Figures

List of Tables

List of Acronyms

3D	three-dimensional
AR	androgen receptor
BME	Bayesian-Maximum Entropy
CD	circular dichroism
DBD	DNA binding domain
DRPLA	dentatorubropallidoluysian atrophy
ER	estrogen receptor
GR	glucocorticoid receptor
HD	Huntington's disease
HSQC	heteronuclear single quantum correlation
IDP	intrinsically disordered protein
IDR	intrinsically disordered region
LBD	ligand-binding domain
LCR	low-complexity region
LR	Lifson–Roig
MaxEnt	maximum entropy
MD	molecular dynamics
MR	mineralocorticoid receptor

NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
NR	nuclear receptor
NTD	N-terminal domain
PBC	periodic boundary conditions
PDB	Protein Data Bank
polyQ	polyglutamine
PR	progesterone receptor
QM	quantum mechanics
SASA	solvent-accessible solvent area
SBMA	spinal bulbar muscular atrophy
SCA	spinocerebellar ataxias
TF	transcription factor
ZB	Zimm–Bragg

1 Introduction

1.1 Biological Background

1.1.1 Disordered proteins

Despite the general belief that the biological functions of proteins require unique three-dimensional (3D) structures, structure-less intrinsically disordered proteins (IDPs) or regions (IDRs) are functional under physiological conditions, being able to engage in biological activities (Figure 1.1)[1, 2].

The idea that protein function depends on a 3D structure started with Fischer's 'lock and key' model(Figure 1.2)[4] and continued with studies by Mirsky and Pauling, and by Wu, which separately showed loss of activity upon protein denaturation[5, 6]. Since these findings, thousands of structures have been solved and deposited in the Protein Data Bank (PDB)[7]. All these advances supported the protein structure paradigm. However, in most deposited structures, there were some clear indications of disorder, like missing electron density, which were mostly overlooked. Over the years, various proteins that lack a stable structure have been linked to specific functions[8, 9] and the traditional protein structure paradigm has been challenged by the discovery of IDPs.

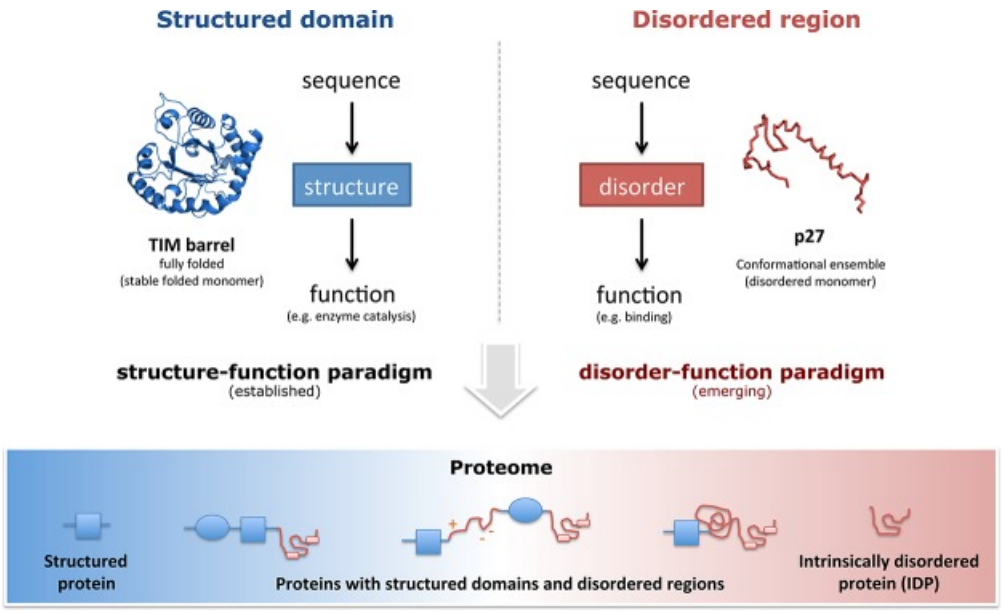

Figure 1.1: IDPs challenge the protein structure paradigm[3].

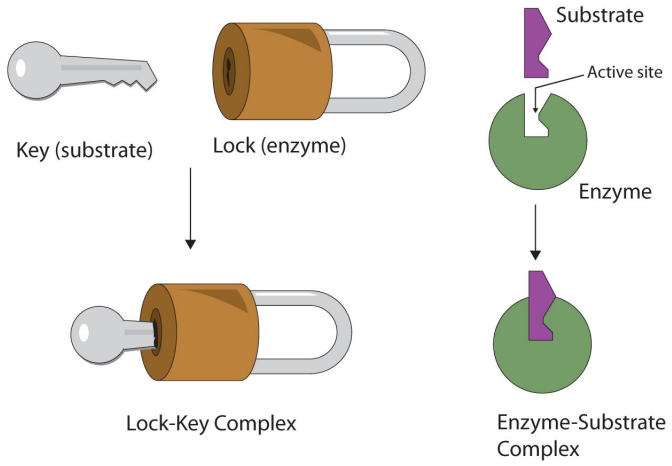

Figure 1.2: The Lock and Key Model assumes substrates fit perfectly to the active site on enzymes as a keys fit into their lock[10].

Sequence characteristics

In addition to experimental studies, computational studies have been performed on IDPs to discover signs of disorder in sequence. As the amino acid sequence of each protein contains the information needed to fold into a specific 3D structure, IDPs present a series of sequence determinants that are key for their lack of tertiary

organization. These studies showed that the amino acid composition IDPs differs significantly from that of ordered proteins[2, 11, 12].

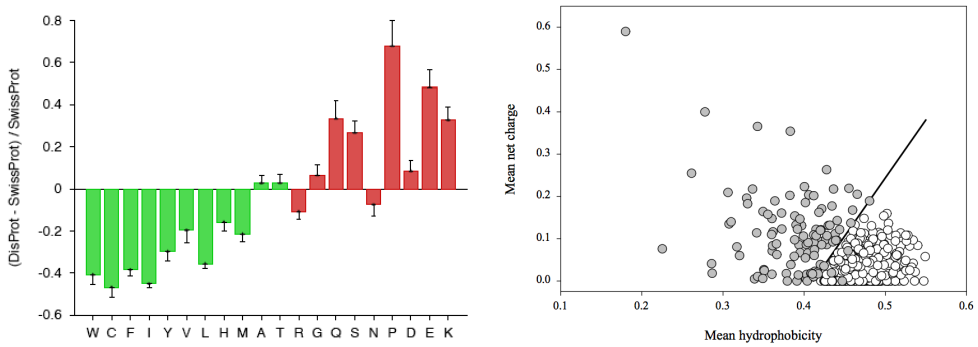

Figure 1.3: a. Amino acid composition profile comparison of DisProt[13] with SwissProt [14]. Plot shows enriched amino acids in disordered proteins in red and depleated ones in green. b. Mean hydrophobicity and charge phase space for proteins. The line represents the border between ordered and disordered proteins. IDPs tend to have high net charge and low mean hydrophobicity.

Compared to the structured proteins, IDPs are depleted in amino acids commonly found in the structural core of globular proteins (I, L, V, W, F, Y), while they are enriched in residues that are characteristic of exposed and flexible regions (E, K, R, S, G, Q, P)[15] (Figure 1.3a). A high net charge and low mean hydrophobicity are two of the main characteristics of IDPs[16](Figure 1.3b). While the former contributes charge-charge repulsion and therefore minimizes aggregation, the latter removes the driving force for protein compaction. IDPs also often present low-complexity regions (LCRs): sequence tracts with high enrichment in a particular type of amino acid, i. e. Gly, Pro or Gln[17, 18]. In this regard, the differences between the sequences of ordered and disordered proteins have allowed the development of IDP predictors[19–23].

Sequence analysis of complete genomes by using computational tools to detect disorder showed that the proportion of IDP/IDRs in proteins increases with the complexity of organisms, with greater representation of IDPs in eukaryotes than in prokaryotes[24]. In this regard, 33.0% of eukaryotic proteins have IDRs with at least 30 residues, while these regions are present in only 4.2% and 2.0% for eubacterial and archaean proteins respectively[25].

Experimental characterization

Both indirect and direct biophysical techniques can be used to detect structural disorder. The former includes X-ray crystallography, and the latter solution state nuclear magnetic resonance (NMR), which is the most powerful technique for the detection of IDPs[26, 27].

Once resonance assignments have been obtained for a given protein, a residue specific dynamic and structural characterization of the conformational ensemble can be obtained by analysis of NMR parameters such chemical shifts, coupling constants and short-range nuclear Overhauser effects (NOEs)[28]. Chemical shifts are the main and most powerful observables to obtain residue-specific information. They are highly sensitive to the environment and to backbone torsion angles -and therefore to structure- and their differences from 'random coil' values can be used to estimate the population of specific secondary structure in the conformational ensemble[29–32].

Circular Dichroism (CD) is another experimental tool through which to obtain structural information for proteins in solution[33]. However, unlike NMR, CD does not provide residue-specific data and it informs only about the global secondary structure.

Characteristics and functions of IDPs

In general, IDPs cannot fold spontaneously into a stable well-defined structure due to an insufficient number of buried hydrophobic residues. However, these proteins can hold stable transient secondary structure elements, without forming the tertiary structure. These regions with transient secondary structure are required for the function of the protein[34–36] and can undergo a disorder-to-order transition upon binding[37–39]. Occasionaly, IDPs carry out their functions and bind other proteins while remaining unstructured[40](Figure 1.4).

Of the structures deposited in the PDB, only around 32% are fully ordered, without any missing residues[41, 42]. Also, many of the eukaryotic proteins have flexible linker regions that separate independently folded globular domains[43].

As a consequence of advances in structure-function studies, it is now clear that structural disorder provides various functional advantages. In this regard, IDPs are characterized by specific but weak binding, frequent regulation by post-translational modification, adaptability in binding, and high functional density -all crucial properties for specific types of cellular functions, including signaling, transcription and translation[1, 44–46].

The presence of IDRs in transcriptional regulatory proteins was identified more

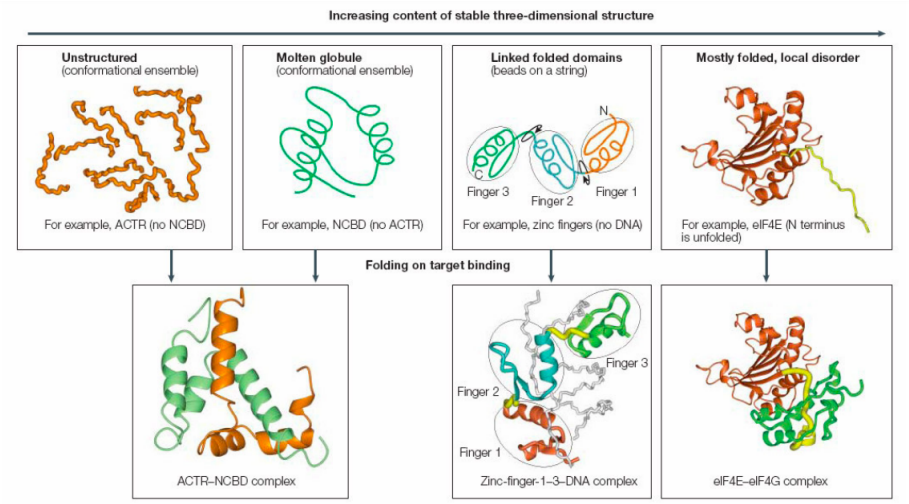

Figure 1.4: Continuum of protein structure: Proteins can have a distinct number of structural types from fully disordered to folded states.

than 30 years ago[47]. It has been shown that transcriptional activation domains are mostly disordered or have disordered regions, and that they can undergo a disorder-to order transition upon binding. One of the common characteristics of transcriptional regulatory proteins is their specific sequence composition in their activation domains, which are enriched in glutamine or proline-rich low-complexity regions[48, 49].

1.1.2 Polyglutamine proteins and polyglutamine diseases

Glutamine (Gln)-rich low-complexity regions, known as polyglutamine (polyQ) tracts, are the most common amino acid repeats in eukaryotic proteins[50]. These tracts are polymorphic in length and typically consist of ten to hundreds of Gln residues[51].

The expansion of polyQ tracts beyond a certain threshold in specific proteins is linked to nine inherited neurodegenerative disorders called polyQ diseases(Table 1.1)[51], namely Huntington's disease (HD), the six spinocerebellar ataxias (SCA 1–3, 6, 7, 17), Dentatorubropallidoluysian atrophy (DRPLA) and spinal bulbar muscular atrophy (SBMA, also known as Kennedy's disease)[52–60].

A common feature of polyQ diseases is the aggregation of protein with expanded polyQ tract in inclusion bodies. Still, the mechanism of the disease has not been fully defined yet. One of the most accepted explanations is that the expanded polyQ tracts decrease protein solubility, cause aggregation, and form

Table 1.1: List of polyQ diseases with healthy and pathogenic Gln thresholds

Disease	Gene	Normal repeat number	Pathogenic repeat number
SBMA	AR	9 - 36	38 - 62
HD	HTT	6 - 35	36 - 250
SCA1	ATXN1	6 - 35	49 - 88
SCA2	ATXN2	14 - 32	33 - 77
SCA3	ATXN3	12 - 40	55 - 86
SCA6	CACNA1A	4 - 18	21 - 30
SCA7	ATXN7	7 - 17	38 - 120
SCA17	TBP	25 - 42	47 - 63
DRPLA	ATN1	6 - 35	49 - 88

fibrillar species that are toxic to cells [61–63]. On the other hand, it has been also proposed that expanded polyQ proteins are inherently neurotoxic[64]. One other alternative explanation is the expanded transcripts themselves are the neurotoxic species due to their propensity to phase separate[65, 66].

Spinal bulbar muscular atrophy (SBMA)

Of the nine polyQ diseases, SBMA was the first to be linked to polyQ tract expansion [67, 68]. SBMA is characterized by late-onset with slow progress. It is not lethal and does not reduce life expectancy. However, it causes dysarthria (speech disorder), dysphagia (swallowing difficulties), wasting and muscle twitch of the tongue, weakness of the proximal muscles and absence of tendon reflexes as a result of the loss of motor neurons[69, 70].

SBMA is a rare and hereditary disease with X-linked heritability due to an expansion in the androgen receptor (AR) gene located on the X chromosome[71], and it affects only men (1 in every 50000 males), while females are carriers of the disease.[69, 72].

To understand the mechanism of the disease, initial research focused on the polyCAG tract in exon 1 of the AR gene. Studies of clinical cases revealed that while individuals with 13 to 34 Gln residues are healthy, SBMA patients have 37 to 66 Gln residues in the polyQ region of AR. This observation showed that the number of repeats is related to the onset of the disease[68, 69, 73].

1.1.3 Androgen receptor

Nuclear receptors (NRs) are a superfamily of transcription factors (TFs) that regulate the expression of the target gene by binding to the steroid or thyroid hormones[74, 75]. AR is one of the NRs activated by binding testosterone and dihydrotestosterone and it plays an important role in the development of the male phenotype[76, 77]. AR has 919 residues and shares a similar domain organization with the other members of the NR family, such as the estrogen receptor (ER), the glucocorticoid receptor (GR), the mineralocorticoid receptor (MR), and the progesterone receptor (PR)(Figure 1.5).

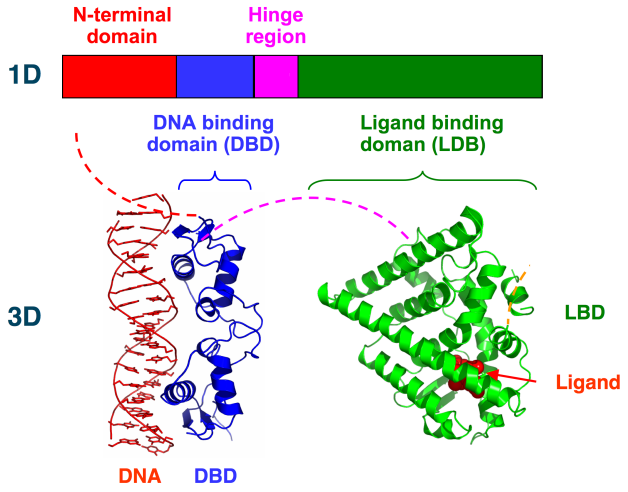

Figure 1.5: Structural organization of NRs. NRs share similar domain organization and contain following domains: NTD, DBD and LBD[78].

Domain organization

NRs have three main domains; an N-terminal domain (NTD), a DNA-binding domain (DBD), and a ligand-binding domain (LBD). The N-terminal domains of the NRs are intrinsically disordered and present one or more transactivation units. Their sequence composition and length differs from one to another, and the AR has one of the longest N-terminal domains with 559 residues. The DBD, which contains two zinc fingers, is the most conserved domain across the members of the NR family. In the AR, the DBD spans the residues 560 to 622. Between the DBD and the LBD, there is a flexible hinge region from residues 623 to 670, which connects these two domains. Finally, the LBD is located between residues 671 and 919. This region is where androgens bind, and it also contains the less potent AF2

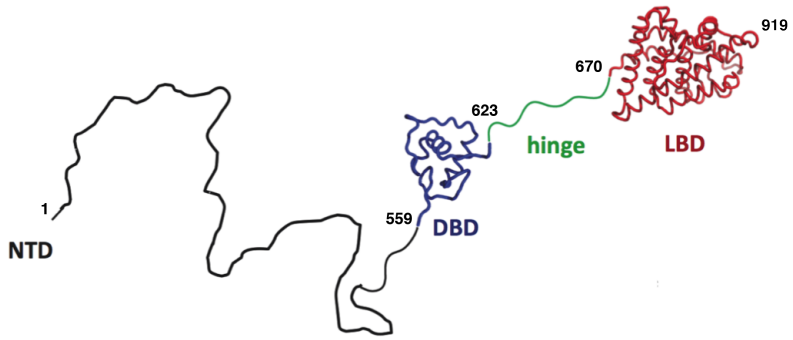

Figure 1.6: Domain structure of AR: NTD and hinge region are representative drawings, DBD is from PDB structure 2AB9, and LBD is from PDB structure 1R4I.

activation function. Among the proteins in NR, the LBD domain has a highly conserved structure and a relatively less conserved sequence(Figure 1.6).

Due to the presence of a polyQ tract and a polyG stretch in the N-terminal domain of the AR, the numbering shows discrepancies in the literature. In this regard, the numbering of residues that defines the domains of the AR in this study comes from the UniProt entity, with 21 Gln between residues 58 and 78, and 24 Gly residues between residues 449 and 472 (UniProt id: P10275).

1.2 Theoretical Background

1.2.1 Molecular dynamic simulations

The microscopic behavior of macromolecules defines their macroscopic properties. Molecular dynamics (MD) simulations bridge the gap between the microscopic and macroscopic scales. By capturing the atomic resolution of a biological system spanning 12 orders of magnitude (Figure 1.7), MD simulations cover the spatiotemporal domain in which experimental characterization is difficult to achieve. MD is a computational simulation method that calculates the evolution of the position of an atom in order to provide information on the dynamic behavior of the system. Researchers from a variety of fields, including physics, chemistry, and biology, use MD to study gases, liquids, and solids, and also organic or inorganic systems of distinct sizes. In biochemistry and biophysics, MD allows researchers to model and understand protein folding, solvation, drug-receptor interactions, and conformational changes of the molecules under a range of conditions.

Understanding the dynamic behavior of the atoms at the microscopic level by

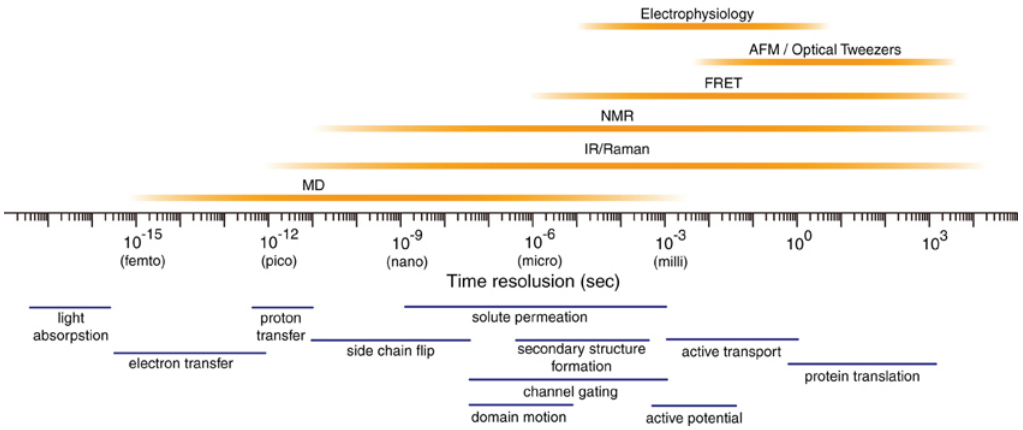

Figure 1.7: Spatiotemporal resolution of various biophysical techniques. Also below, the timescales of some fundamental motions of atom or molecules are shown taken from: [83].

MD can be achieved using the classical equations of motion to calculate the time evolution of a many-body system. By using the basic principles of classical mechanics, we can provide a formal description of a system in equilibrium. The most important ingredient dictates the quality of a molecular dynamics simulation is having the right energy description of the energy in between the atoms. In classical MD simulations, force fields are used to define interactions between particles to calculate energy[79]. In the simplest form, MD simulations integrate Newton's equation of motion to calculate the positions and velocities of the atoms from the forces obtained using force fields[80–82].

1.2.2 Force fields

To calculate forces, MD simulation requires the definition of a potential energy function of a system of atoms. Quantum mechanics (QM) MD or ab initio (quantum Hamiltonian) can be used for this purpose. However, the direct application of quantum mechanics to protein systems is too computationally expensive and not possible due to the large molecular size of proteins. A decrease in accuracy achieved by reducing a fully quantum description to a classical potential has made it possible to have trajectories that are sufficiently accurate for many purposes. This reduction requires two main approximations. The first one is the Born–Oppenheimer approximation, which states that the dynamics of electrons are so fast that they can be considered to react instantaneously to the motion of their nuclei. Consequently, the electrons may be treated separately. The second one treats nuclei, which are much heavier than electrons, as point particles that

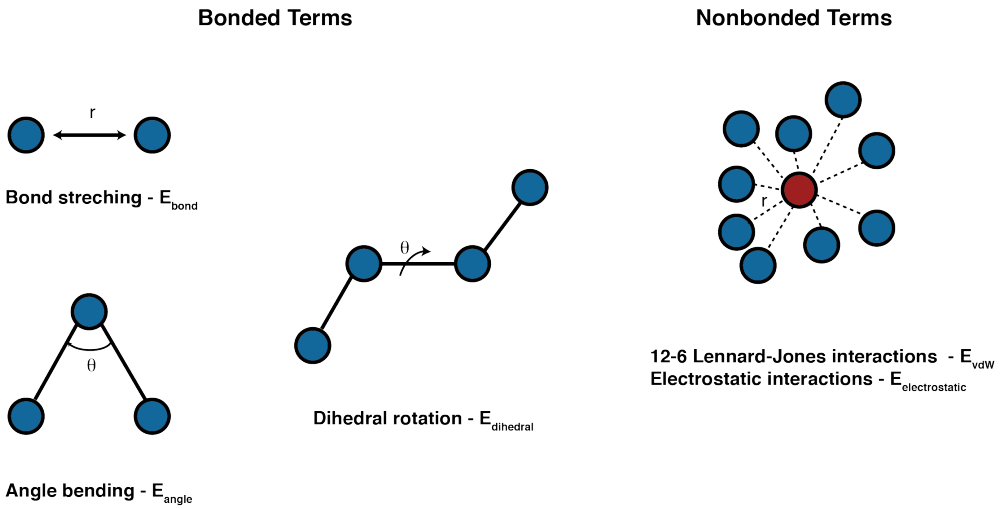

Figure 1.8: Schematic representations of the terms in a classical force field i.e. bond stretching(E_{bond}), angle bending(E_{angle}), dihedral rotation($E_{dihedral}$), van der Waals interactions(E_{vdW}), and electrostatic interactions($E_{electrostatic}$).

follow classical Newtonian dynamics. Still, the quantum mechanical effects are represented implicitly as empirical potentials by functional approximations. Potentials like partial atomic charges, van der Waals parameters, equilibrium bond length, angles, and dihedrals are obtained by experimental physical properties or QM simulations by fitting against detailed electronic calculations. As shown in the Figure 1.8, the basic functional form of the potential energy consists of bonded (interactions of atoms that are linked by covalent bonds) and non-bonded (long-range electrostatic and van der Waals forces) forces (Eq (1.1)–(1.3)).

$$E_{total} = E_{bonded} + E_{non-bonded} \tag{1.1}$$

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral} \tag{1.2}$$

$$E_{non-bonded} = E_{electrostatic} + E_{vdW} \tag{1.3}$$

Bonded potential terms

The mechanical molecular model treats atoms as spheres and bonds as springs. The mathematical description of spring deformation can be used to study the ability of bonds to stretch, bend and twist.

- **Bond stretching:** To calculate the potential energy of a covalent bond between two atoms, we use Hooke's law for a spring.

$$E_{bond} = \sum_{bonds} k_b(r - r_0)^2, \tag{1.4}$$

where k_b is force constant that regulates the stiffness of the bond, r is the length of the bond, and r_0 is the equilibrium bond length, which is a value adopted in a structure with minimum energy. k_b and r_0 values are defined for each pair of atoms on the basis of their type (C-H, C-C, etc.).

- **Angle bending:** The equation that gives the energy of bond angle vibration between three atoms (two consecutive bonds) is also based on Hooke's law.

$$E_{angle} = \sum_{angles} k_\theta(\theta - \theta_0)^2, \tag{1.5}$$

where k_θ is the stiffness parameter that controls the angle, θ is the angle between the three atoms and θ_0 is the equilibrium angle. All the parameters are unique for each bonded triplet of atoms in function of their type (C-O-H, C-C-C, etc.) and are obtained experimentally or theoretically.

- **Dihedral rotation:** If a molecule contains more than four atoms in a row, the inclusion of a dihedral or torsional term is needed. The energy of the dihedral rotation is represented by a simple periodic function.

$$E_{dihedral} = \sum_{dihedrals} \frac{V_n}{2}[1 + cos(n\phi - \delta)], \tag{1.6}$$

where V_n defines the potential barrier height, ϕ is the dihedral angle, δ is the phase angle and n is the number of minima in the energy function. By combining two or more terms with a different n, it is also possible to have a dihedral potential with a different depth and height for each well and barrier respectively. Also, distinct force fields may have alternative representations of the dihedral potential.

Non-bonded potential terms

The non-bonded potential terms contain the pairwise sum of all the energies of all possible interacting non-bonded atoms. Non-bonded interactions are the most expensive part of the MD calculations and their computational cost increases with the number of particles.

- **Electrostatic interactions:** The most accurate way to obtain molecular electron density is using high-level quantum mechanical (QM) calculations. However, this approach cannot be used for large systems like proteins. Neither is reducing these calculations to manageable procedures for MD simulations a straightforward task. This problem is commonly tackled by assigning a partial atomic charge to each nucleus and measuring their total energy contributions by means of Coulomb's law.

$$E_{electrostatic} = \sum_{i,j} \frac{q_i q_j}{\varepsilon_D r_{ij}}, \tag{1.7}$$

  where ε_D is the dielectric constant of the solution, q_i and q_j are charges of particles i and j, and r_{ij} is the distance between them.

- **Van der Waals interactions:** Van der Waals interactions explain attraction and repulsion between two atoms. Repulsive forces appear due to the overlap of the electron clouds when atoms are close to each other. Attraction arises from dispersion forces generated between dipoles as a result of fluctuations in electronic charge distributions. The 12-6 Lennard-Jones potential is the most common way to model the balance between the repulsion and attraction of the atoms.

$$E_{vdW} = \sum_{i,j} 4\varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}] \tag{1.8}$$

  Here, ε defines the depth of the potential well, σ is the finite distance at which the inter-particle potential is zero, and r is the distance between the atoms. While the $(\frac{\sigma_{ij}}{r_{ij}})^{12}$ term corresponds to the repulsion designed to rapidly blow up at close ranges, $-(\frac{\sigma_{ij}}{r_{ij}})^{6}$ term corresponds to the attraction.

1.2.3 Water models

Biomolecules function in the environment of water and ions. Water is a polarizable molecule that can act as both a donor and an acceptor of a hydrogen bond. These properties make water crucial not just for life but also for molecular simulations. Interaction with water has an important effect on the thermodynamics and conformational properties of biomolecules. In simulations, water can be implicitly represented as a continuous medium or can be defined as individual explicit water molecules. Implicit water models are faster to compute because they average

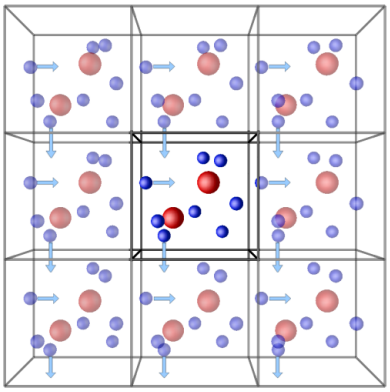

Figure 1.9: Diagram of periodic boundary conditions that show diffusion of the particles[85].

out the behavior of highly dynamic solvent molecules to compute potential mean force. Consequently, key characteristics like hydrogen bond fluctuations and water dipole reorientation are overlooked in these models. Explicit solvent models are therefore the approach of choice in MD simulations. Many water models have been developed to imitate the nature of water molecules. In this regard, TIP3P and TIP4P, which efficiently balance accuracy and computational cost, are widely used[84].

1.2.4 Periodic boundary conditions

Typical MD simulations can involve systems containing thousands of atoms. In particular, the number of atoms in explicit solvent simulations with water molecules can often reach up to 100,000. Consequently, a very large proportion of the atoms would be affected by the surface of the simulation box when MD simulations are performed in a finite size box. Given that water molecules have different behavior near surfaces than the ones inside the box due to Laplace pressure, simulations are done with periodic boundary conditions (PBC) in order to efficiently simulate bulk properties. The use of PBC implies that the simulation box is replicated infinitely in three dimensions of space, as shown in Figure 1.9. If an atom falls out of one side of the simulation box, it re-enters the box from the opposite side. Thus atoms move freely instead of bouncing off the walls.

However, simulation of the system with PBC increases the calculation cost of the non-bonded interactions. Evaluating Lennard-Jones and Coulomb potentials is the most expensive part of the MD simulations since they involve all particles in the system. As PBC create an infinite system, to reduce the number of interactions

and cost of computation at some finite distance, force fields cut these interactions to zero.

1.2.5 Temperature and pressure control

In MD simulations, we usually analyze the motion of a fixed number of particles (N) in a unit cell which has fixed volume (V). In addition to N and V, since there is no external force in the system, the total energy (E) is also conserved. In statistical mechanics ensembles in which where N, V, E are conserved -called microcanonical or NVE ensembles-, these three quantities are controllable parameters of the system. However, experiments are carried out mostly at a constant temperature and constant pressure, and the control of these two parameters is crucial for the compatibility with the experiments.

Many thermostats and barostat algorithms have been improved in order to be able to run simulations at constant temperature (isothermal) or/and constant pressure (isobaric). The most common temperature coupling schemes are Langevin dynamics[86, 87], the Berendsen thermostat[88], the Andersen thermostat[89], the Nose-Hoover thermostat [90, 91] and velocity rescaling[89]. Like the thermostat algorithms, MD methods have been modified to perform isobarically by using barostat algorithms. Some of the widely used techniques to control pressure include the Berendsen barostat, Parrinello-Rahman barostat[92, 93], and Nose-Hoover bath.

1.2.6 Current classical force fields

Since the development of molecular mechanics in the 1960s, many force fields have been developed for a range of purposes. The first force fields were oriented mainly towards predicting the structures, enthalpies, and vibrational spectra of small organic molecules[94]. MM2, one of the first force fields developed, was created to simulate hydrocarbons[95]. Later on, modified versions with the capacity to simulate many other different types of molecules such as alcohols and amides, etc. (MM3[96], MM4[97]), became available. Since then, force fields have now reached a maturity and can deal with much more complex systems under a range of conditions. AMBER (Assisted Model Building and Energy Refinement)[98], CHARMM (Chemistry at HARvard Macromolecular Mechanics)[99], GROMOS (GROningen MOlecular Simulation)[100], and OPSL (Optimized Potential for Liquid Simulations)[101] are the force fields most widely used to simulate biomolecules. These force fields are under continuous improvement and each one has many modified versions.

1.2.7 Molecular dynamics algorithms

The basis of the MD simulation is Newton's second law or the equation of motion, which states that it is possible to calculate the movement of each atom in the system from the force applied to them. By integrating equations of motion, we can generate a trajectory that shows the changes in the position, velocity and acceleration of a particle over time. From this trajectory, we can observe dynamic and equilibrium properties, and the average thermodynamic properties of the system.

Consider a system with N particles moving under the impact of the internal forces defined by the force fields. Each particle will have a spatial position (r_i) and velocity (v_i) that changes with time. The motion of these particles is directly related to the applied forces (F_i) through Newton's second law,

$$F_i = m_i \ddot{r}, \tag{1.9}$$

where m_i is the mass of the particle. Newton's second law allows us to calculate how forces affect the movement of the particles and these forces are derived from interatomic potential functions.

$$F = -\nabla U, \tag{1.10}$$

where U is the potential energy.

By calculating the energy between each pair of atoms as a function of their distances, force -and therefore the motion- can be determined by solving Eq. (5.3). This is the goal of the MD methods which rely on a cycle of calculations iterate on different steps(Figure 1.10). To start this cycle, knowledge of the initial positions and velocities of each atom at t=0 is required, and this set up is called the initial configuration. For the ordered proteins with solved structures, positions of the atoms are available in databases like PDB[7], and in the case of disordered proteins, the atom positions can be generated randomly. The initial velocity of each particle is assigned from a Maxwellian distribution in random directions with a fixed magnitude centered on the selected temperature. Initial velocities are also adjusted to ensure that net velocity results in a total momentum of zero.

The first step of the cycle is calculating the potential energy of each atom pair as a function of their distance from the given model (Eq (1.4)–(1.8)). Since force is equal to the minus gradient of the energy, using Eq. (5.3) it is straightforward to calculate the force for each particle. using the mass and force of the particles, Newton's second law allows us to obtain the acceleration for each atom. By using

information about the velocity and acceleration of the atoms, the next step is predicting how they move after a small increment of time(Δt). This calculation relies on numerical integration. Time discretization is also one of the most important steps of the algorithm, and the accuracy of the numerical solution is determined by the Δt chosen, also referred to as the time step. This time step is selected on the basis of the nature of forces applied to the system. To accurately integrate the fastest motion in the system, which is the bond stretching of a hydrogen atom, the time step has to be smaller than the fastest motion. An increase in the time step adds instability to the system. To solve this problem, several constraining methods are developed. Removing or slowing down the fast motion of the hydrogen atoms by freezing their bonds to the parent-atom is the most common approach used in these constraining algorithms. SHAKE[102], LINCS[103] and RATTLE[104], the most known algorithms, are used in the AMBER, GROMACS and NAMD force fields, respectively. These algorithms increase the time step to 2 fs by imposing holonomic constraints, which depend only on the position of the particles involved.

Another useful tool to accelerate MD simulations is a method called hydrogen mass repartitioning (HMR)[105]. HMR is based on slowing down the high-frequency vibrations of the molecules by increasing the mass of the hydrogen and decreasing the mass of the heavy atom by an equivalent amount, since equilibrium thermodynamics averages do not depend on the exact mass distribution of the system. This method allows time steps up to 4 fs and effectively accelerates the simulation.

1.2.8 Integration algorithms

Since being introduced to solve Newton's equations of motion for the simulated systems, the Verlet Integrator[106, 107] continues to be the most used algorithm. *Leap-frog*[108] and *velocity Verlet*[109] are essentially equivalent variants of the Verlet Integrator, but here we concentrate on the original method. Integrator algorithms have to conserve the Hamiltonian and be time-reversible and computationally efficient. The Verlet Integrator is the simplest algorithm that satisfies these criteria.

The aim is to solve Eq. (1.9) numerically and calculate new positions and velocities of the particles after time Δt.

$$r_i(t_0) \rightarrow r_i(t_0 + \Delta t) \rightarrow r_i(t_0 + 2\Delta t) \rightarrow ...r_i(t_0 + n\Delta t) \tag{1.11}$$

All the integration algorithms are derived from Taylor expansions. The Verlet Integrator uses two Taylor expansions:

$$r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{1}{2}\frac{F(r(t))}{m}\Delta t^2 \tag{1.12}$$

$$r(t - \Delta t) = r(t) - v(t)\Delta t + \frac{1}{2}\frac{F(r(t))}{m}\Delta t^2 \tag{1.13}$$

The Verlet Integrator uses positions at time t and the positions from time t-Δt to calculate new positions at time t+Δt. To sum up (1.12) and (1.13):

$$r(t + \Delta t) = 2r(t) + r(t - \Delta t) + \frac{F(r(t))}{m}\Delta t^2 + O(\Delta t^4) \tag{1.14}$$

cancels the first- and third-order terms from the Taylor expansion and makes the Verlet Integrator one order more accurate than using just one simple Taylor expansion.

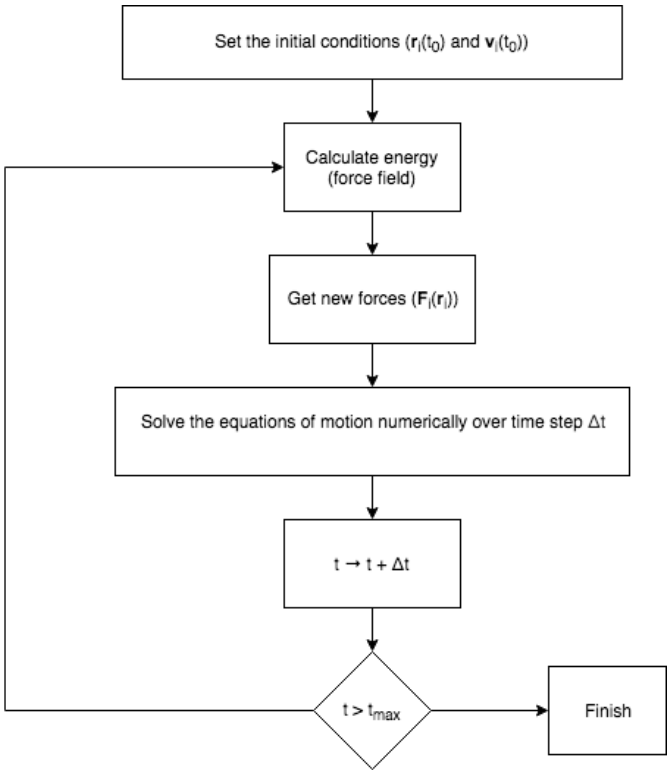

Figure 1.10: Flow chart of basic MD algorithm.

1.2.9 Hydrogen bonds

A hydrogen bond is an interaction between a hydrogen atom covalently bound to a more electronegative atom and another electronegative atom(Figure 1.11) [110]. A hydrogen bond can occur intermolecularly or intramolecularly and it is a dipole-dipole interaction[111, 112]. In this interaction, the molecule that donates the hydrogen atom is called the donor, and the molecule that accepts the hydrogen with the lone pair is called the acceptor. The strength of the hydrogen bond varies in function of the environment, geometry and the nature of the donor and acceptor, and it can be between 1 and 40 kcal/mol[113].

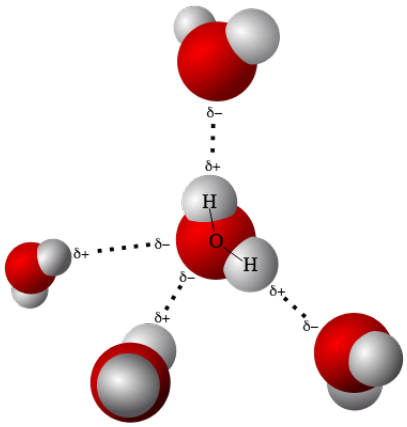

Figure 1.11: Schematic representation of hydrogen bond interaction between four molecules of water. In a water molecule, there is one oxygen atom and two hydrogen atoms. Water can donate two hydrogens and oxygen can form two hydrogen bonds due to its two lone pairs of electrons as an acceptor. Therefore, a water molecule can have four hydrogen bonds[114].

The hydrogen bond is one of the most important interactions in biomolecular and chemical processes. For example, the solving capabilities of water derive from the hydrogen bonds that water forms. Furthermore, these bonds are highly effective stabilizers, and they therefore play a key role in protein folding and DNA structure. The secondary and tertiary structure of proteins is partially determined by hydrogen bonding. Parallel strands of DNA come together with hydrogen bonds to form the double helix. In addition, the networks of hydrogen bonds show cooperativity and the effect of this cooperativity is much stronger than pairwise additivity.

Geometric criteria for hydrogen bonds

Various studies have proposed geometric criteria to identify a hydrogen bond[115, 116, 110]. For the calculations in this thesis, we used the most strict definition of a hydrogen bond. Where A is the acceptor, D is the donor and H is hydrogen, the criteria for a hydrogen bond are (D-H • • • A):

- The distance between H and the acceptor: H-A $< 2.4\mathring{A}$
- The angle between the heavy atom of the donor, hydrogen, and acceptor: D-H-A $> 120^{\circ}$

The shorter the distance, the stronger the hydrogen bonds. Hydrogen bonds are typically stronger than van der Waals interactions and weaker than covalent bonds.

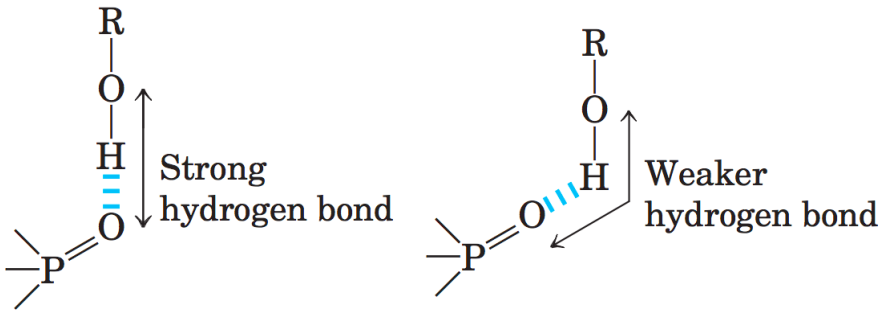

Figure 1.12: Directionality of the hydrogen bond.The attraction between the partial electric charges is greatest when the three atoms involved in the bond (in this case O, H, and O) lie in a straight line[117].

Directionality of the hydrogen bond

The strength of the hydrogen bond depends not only on the distance between the acceptor and the donor but also on the angle of the atoms in the space. To maximize electrostatic interaction between them, atoms need to be oriented in proper alignment. When the unshared electron pair of the acceptor atom is in line with the covalent bond between the donor atom and H, hydrogen bonds are strongest(Figure 1.12). Therefore, hydrogen bonds are highly directional.

However, in high-resolution crystal structures, hydrogen bonds are rarely linear [119, 118]. Even though linear hydrogen bonds are strongest among other configurations, they are not the most stable ones[111]. There is ample evidence

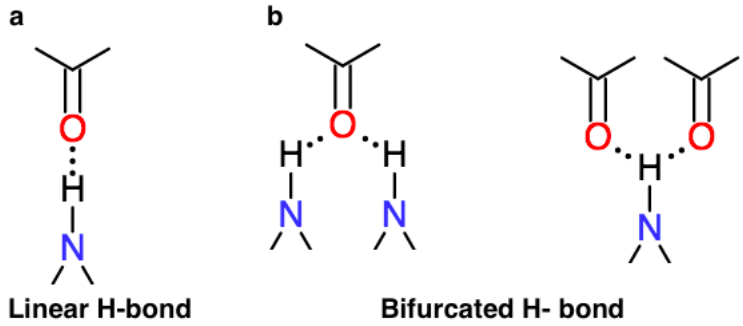

Figure 1.13: Different hydrogend bond configurations a. Linear hydrogen bond proposed by Pauling, b. Three centered bifurcated hydrogen bonds(adapted from[118]).

that hydrogen bonds in protein helices are non-linear. High-resolution structures of proteins show that hydrogen bond interactions are mostly bifurcated (three-centered)[118, 120–125] (Figure 1.12).

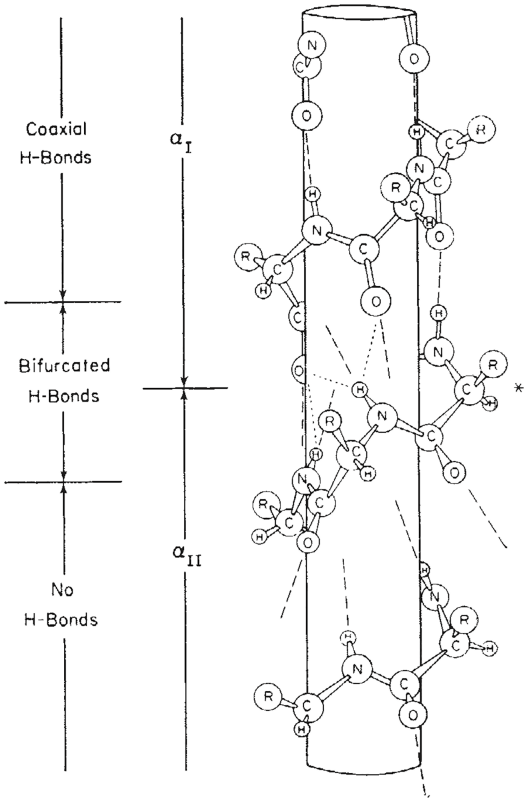

Figure 1.14: Right-handed helical structure with the parameters of the Pauling-Corey α-helix with bifurcated hydrogen bonds[126].

In 1967 a helix model that contains the bifurcated hydrogens bond with the same helical pitch and residues-per-turn as the Pauling α-helix was proposed (Figure 1.14)[127]. And recently hydrogen bonds from 53,040 proteins with 2.0$\mathring{A}$ or higher resolution were analyzed. It has been shown that bifurcated hydrogen bonds are the common feature of the helices in high-resolution structures, and and these helices have the same properties as the model proposed by Nemethy et al.[118].

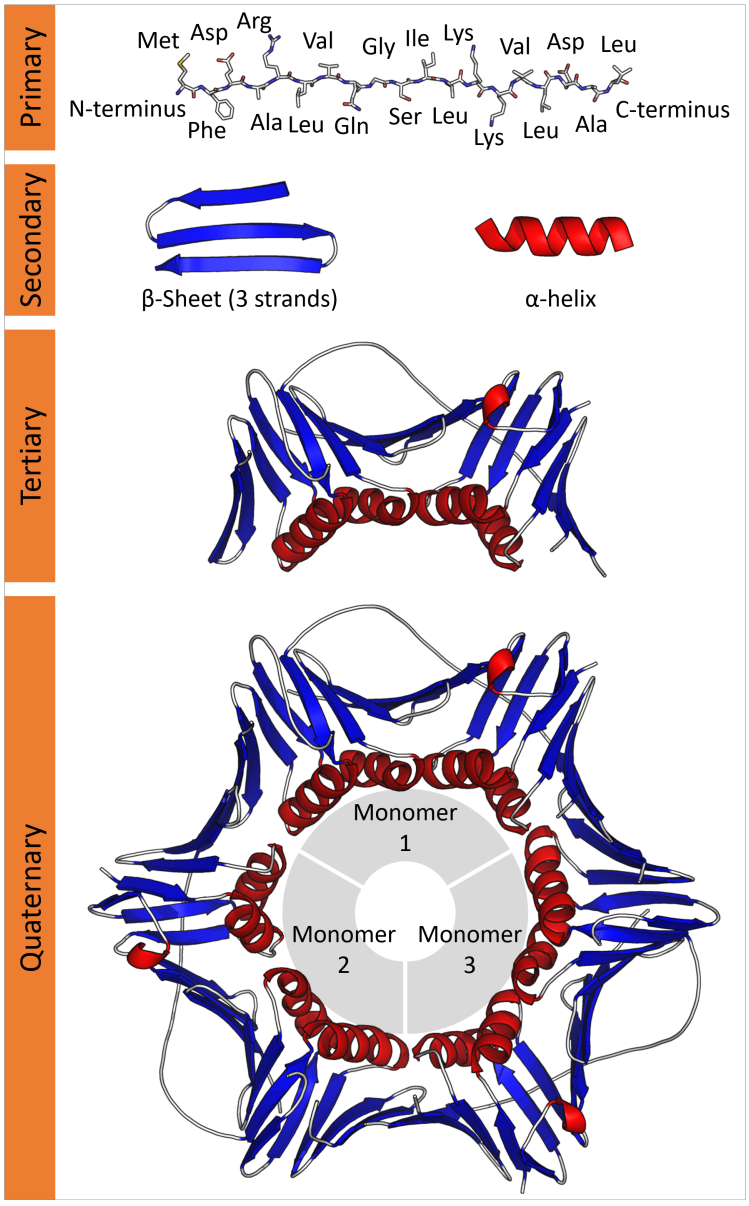

Figure 1.15: Four different levels of protein structure represented by using PCNA as an example (PDB id: 1AXC)[128].

1.2.10 Structure of the α-helix

Proteins can have four different levels of structure (Figure 1.15). These are:

- **Primary structure:** The amino acid sequence of the protein.
- **Secondary structure:** Local interactions between amino groups and carboxyl groups of the protein lead to certain folding patterns such as α-helices and beta sheets.
- **Tertiary structure:** Proteins generally have more than one secondary structures and they fold into a compact globular structure.
- **Quarternary structure:** Defines the arrangement of multiple polypeptide chains or subunits in a protein.

Although IDPs do not fold into stable 3D structures, they may have regions with a propensity to form secondary structures. In this thesis, we will focus mainly on α-helices.

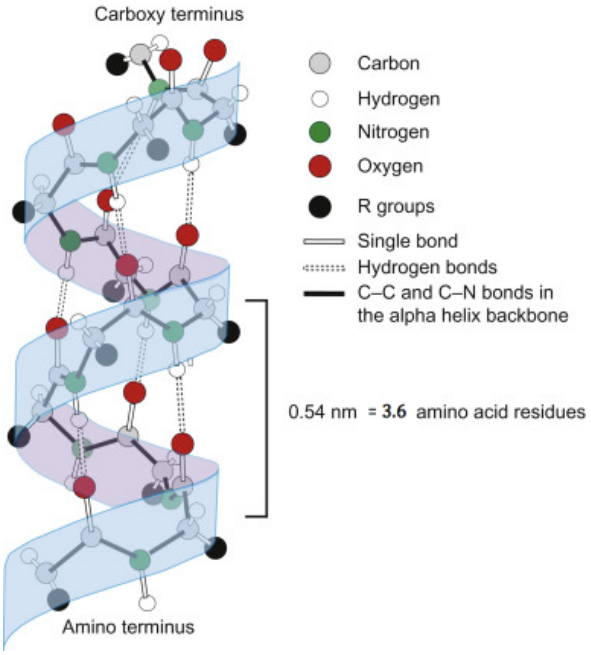

Figure 1.16: The geometry of a right-handed α-helix structure[129].

The α-helix structure was first described by Pauling in 1951[130]. The formation of a hydrogen bond between a backbone carboxyl group and the amino group of the amino acid four residues away give rise to one turn of a right-handed

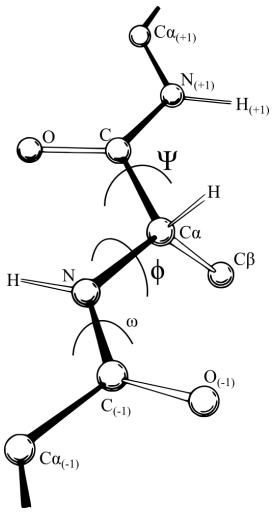

Figure 1.17: Dihedral angles (ϕ, ψ, and ω) of amino acids [131].

α-helix(Figure 1.16). When this behavior is repeated in i→i-4 pattern, it gives the protein a helical shape. In an α-helix, one turn has 3.6 amino acids and each amino acid rotates helix 100°and extends it 1.5$\AA$ along the helical axis. The distance between two turns of the helix is 5.4$\AA$ (1.5$\AA$x3.6) and this distance is called the pitch. Due to the geometry of the amino acids, side chains are close to each other and interaction between them can affect the stability of the helix. However, the helix stability is conferred mainly by the backbone to backbone hydrogen bonds.

Dihedral angles and Ramachandran plot

In proteins, there are three backbone dihedral angles, namely phi (ϕ), psi (ψ), and omega (ω)(Figure 1.17).

- ϕ defines the rotation of the bond between NH and Cα (C-N-Cα-C^{+1}).
- ψ defines the rotation of the bond between Cα and C (N^{-1}-Cα-C-N).
- ω is a torsion angle within the peptide bond. Since it's planar, it's fixed to 180°.

The peptide structure can also be defined by its two backbone dihedral angles, ϕ and ψ. Due to the steric clashes, most of the ϕ and ψ combinations are not allowed. A Ramachandran plot is an illustrative way to show the distribution and allowed combinations of dihedral angles in a protein structure. The plot takes its name from G. N. Ramachandran, who developed it in 1963[132]. The plot shows

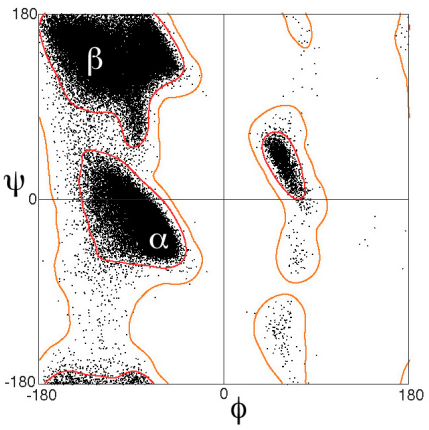

Figure 1.18: Ramachandran plot: generated from phi and psi angle distributions of 100,000 residues obtained from high-resolution structures. Contours show favored and allowed regions[133].

the range of ϕ and ψ angles occupied by each secondary structure. As seen from the Figure 1.18, the x-axis of a Ramachandran plot has ϕ values and the y-axis has the ψ values and both α-helix and β-sheet structures cover only limited areas of the plot.

α-helical propensities of the amino acids

Secondary structure formation depends on the sequence of the peptides. Amino acids differ in their propensity to be in the helix. Analysis of many helices has revealed that Ala has the highest helical propensity, and Pro and Gly the lowest[134]. The α-helical propensities of each amino acid are summarized in the Table 1.2.

Table 1.2: A helix propensity scale of amino acids compared to the Ala[134]

Amino acid	Helix propensity $\Delta(\Delta G)$(kcal mol^{-1})
Ala	0
Glu^0	0.16
Leu	0.21
Arg^+	0.21
Met	0.24
Lys^+	0.26
Gln	0.39
Glu^-	0.4
Ile	0.41
Asp^0	0.43
Trp	0.49
Ser	0.50
Tyr	0.53
Phe	0.54
His^0	0.56
Val	0.61
Asn	0.65
Thr	0.66
His^+	0.66
Cys	0.68
Asp^-	0.69
Gly	1.00
Pro	3.16

1.2.11 Helix-coil theory

The helix-coil transition theory has been studied extensively since the late 1950s[135–141]. Pauling's discovery of the hydrogen-bonded helices[130] was the starting point of the studies to understand the structure and stability of the proteins and led to the development of an elegant and complete area of macromolecular science. The helix-coil theory seeks to analyze the helix–coil equilibrium of the peptides in solution. Instead of being simply 100% unfolded or 100% helical, they form a mixture that has fully helix, fully coil or peptides with a different fraction of helices which also called as helix fraying in solution. Using the helix-coil transition theory,

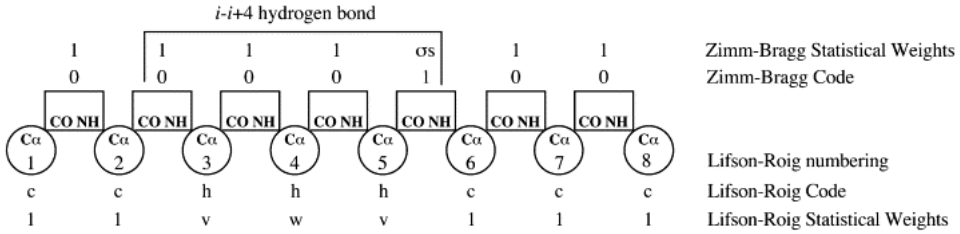

Figure 1.19: Model for Zimm–Bragg and Lifson–Roig and weights for the α-helix[142].

it is possible to predict helical segments from the sequence of the protein. The two major types helix-coil models are Zimm–Bragg (ZB)[136] and Lifson–Roig (LR) [137] models (Figure 1.19), which have a few important differences but share three essential parameters:

- The length of the peptide chain (N)
- The helix nucleation parameter (σs,v)
- The helix propagation parameter (s,w).

Zimm-Brag model

The ZB model defines residues as a helix or a coil on the basis of the participation of their NH group in a hydrogen bond. A peptide group is a unit of the model and, as seen in the Figure, it takes 1 as a unit value where the NH group forms a hydrogen bond and 0 where it does not. Each unit also has a statistical weight. The statistical weight of the unit where the helix starts is σs and the units that come after that are s. The weight of the non-hydrogen bonded units is defined as 1. To form the first hydrogen bond to start the helix, three residues at the beginning of the structure need to be fixed in helical geometry. However, to propagate a helix with an additional hydrogen bond, only one residue needs to be fixed. This explains why nucleation of the helix is more entropically costly than its propagation. To reflect this significant behavior in the model σs is chosen much smaller than s. Propagating the helix with an additional hydrogen bond multiplies its weight by s, but the cost of the nucleation enters the calculations only once. The weight of the helix with an N hydrogen bond can be formulated as σs^{N-1}.

After calculating the statistical weights by dividing it to the partition function (sum of the statistical weights of each conformation), the population of each conformation can be calculated.

Lifson–Roig model

The LR model is a similar but improved version of the ZB model[137, 143]. The basic difference between these two models comes from how they define a helical unit. In the LB model, the definition of helicity depends on the ϕ and ψ angles of the residues. If the ϕ and ψ angles of a residue are in the helical region in the Ramachandran plot, the model assigns helix conformation to that residue. Otherwise, the residue is labeled as coil conformation. As in the ZB model, each residue has statistical weight, and in the LR model, this weight depends also on conformations of neighbor residues. The weight of the coil residues defined as 1, and they are defined as reference residues. Also, the residues that nucleates (v) the helix and propagate (w) the helix have different weights in the model. The calculated statistical weight indicates the stability of the structure. If the weight is higher than 1 it means that it is more stable than a coil, and an increase in the weight implies an increase in stability.

1.2.12 Agadir

Agadir is an algorithm related to both ZM and LR for the prediction of the helical content of peptides from their sequences[144]. Agadir defines the residue centered on the Ca as a unit of the model. In contrast to ZM and LR that defines the minimum length of the helix as three, in the Agadir minimum four residues plus N- and C- caps are required to form a helix[145]. And with this rule, all the helices with single hydrogen bonds are eliminated. N-cap and C-cap residues are flanking residues of a helix at the N- and C-terminal with fixed dihedral angles (ψ for N-cap and ϕ for C-cap). These residues have different statistical weights than the random coil and they are important and required for the stability of the helix. Contrary to ZM and LR that defines the flanking residues as random coils, N- and C-caps are introduced to Agadir.

From the original version to the latest version of Agadir, many modifications added to the classical helix-coil theory to be able to describe α-helix formation in heteropolypeptides. The current version of Agadir introduces terms for side chain-side chain interactions[144, 146–150], interactions between charged groups, electrostatics[151], pH[152], temperature[152], ionic strength[151], helix macrodipole[153, 152, 151], capping motifs[151] such as Schellman motif, and Pro-capping motif.

Free energy of a helical segment that contains contributions of all these parameters is described in:

$$\Delta G_{Hel} = \Delta G_{Int} + \Delta G_{HBond} + \Delta G_{SD} + \Delta G_{nonH} + \Delta G_{Dipole} + \Delta G_{es} \quad (1.15)$$

where ΔG_{Int} is the sum of the intrinsic tendency of the residues to adopt helical dihedral angles, ΔG_{HBond} is sum of the contribution of *i*, *i+4* main chain-main chain hydrogen bonds, ΔG_{SD} is the sum of the net contributions of side chain-side chain interactions in the locations *i*,*i+3* and *i*,*i+4* with respect to the random-coil state, ΔG_{nonH} is the sum of the net contributions of N- and C-cap residues, ΔG_{Dipole} represents the interaction of charged groups with the helix macrodipole, and ΔG_{es} represents all electrostatic interactions between two charged residues inside and outside the helical segment.

2
Objectives

In this project, we focus on the polyQ tract in the intrinsically disordered N-terminal domain of AR. PolyQ tracts longer than 37 repeats are linked to the SBMA which is one of the nine hereditary neurodegenerative diseases linked to the polyQ extension. Transcriptional activity of this receptor is impaired in the case of polyQ elongation. Still, why the length of the polyQ tract affects the function and the aggregation properties is unknown. To understand the structural basis of the effect of length and overcome the limitations of the force field and sampling we will combine molecular simulation with NMR and simulated polyQ peptides of increasing lengths.

3

Side chain to main chain hydrogen bonds in polyQ helices

3.1 Introduction

Recently, our group used high-resolution solution NMR to study a 156-residue proteolytic fragment of AR containing the polyQ tract the production of which has been linked to the onset of SBMA. To investigate how the properties of the polyQ tract depend on its length, two variants containing 4 (AR_{1-156} 4Q) and 25 (AR_{1-156} 25Q) residues were characterized. It was found that the helicity of the four Leu residues preceding the polyQ as well as its N-terminal residues is well over 90% and decays progressively towards the C-terminus. In addition, it was found that the Leu-rich motif preceding the polyQ tract makes it helical[154].

3.2 Experimental Results

3.2.1 Choice of the fragment

In this work we examined polyQ peptides with different tract length to understand the molecular basis of the effect of length on secondary structure. To simplify the

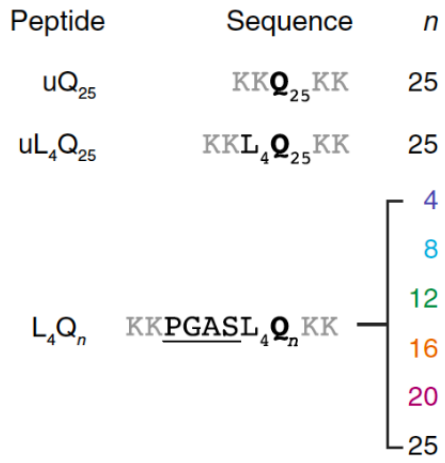

Figure 3.1: Sequences of the uQ_{25}, uL_4Q_{25}, and L_4Q_n peptides used in this project.

acquisition and interpretation of our NMR data and to decrease the computational cost, we decided to work on the shortest and most meaningful AR sequence which contains the polyQ tract. From the previous work[154] in constructs with different tract length, we know that the structural effect is local i.e. we only see structural differences in the eight residues flanking the tract at the N-terminal and the tract itself. This lead to the identification of a short sequence that preserves its helicity (Figure 3.1).

We generated three different sets of synthetic peptides by taking into consideration of helical propensities predicted by Agadir [144] and analyzed them by circular dichroism (CD) (Figure 3.2). To maximize the secondary structure content experiments were performed at 277K and pH 7.4. These three sets of peptides are;

- uQ_{25}: uncapped, polyQ tract with 25 Gln residues
- uL_4Q_{25}: uncapped, four Leu residues at the N-ter of the polyQ
- L_4Q_n: composed of the motif Pro-Gly-Ala-Ser, predicted to be a N-capping sequence according to Agadir (Figure 3.2a), four Leu residues and the polyQ tract.

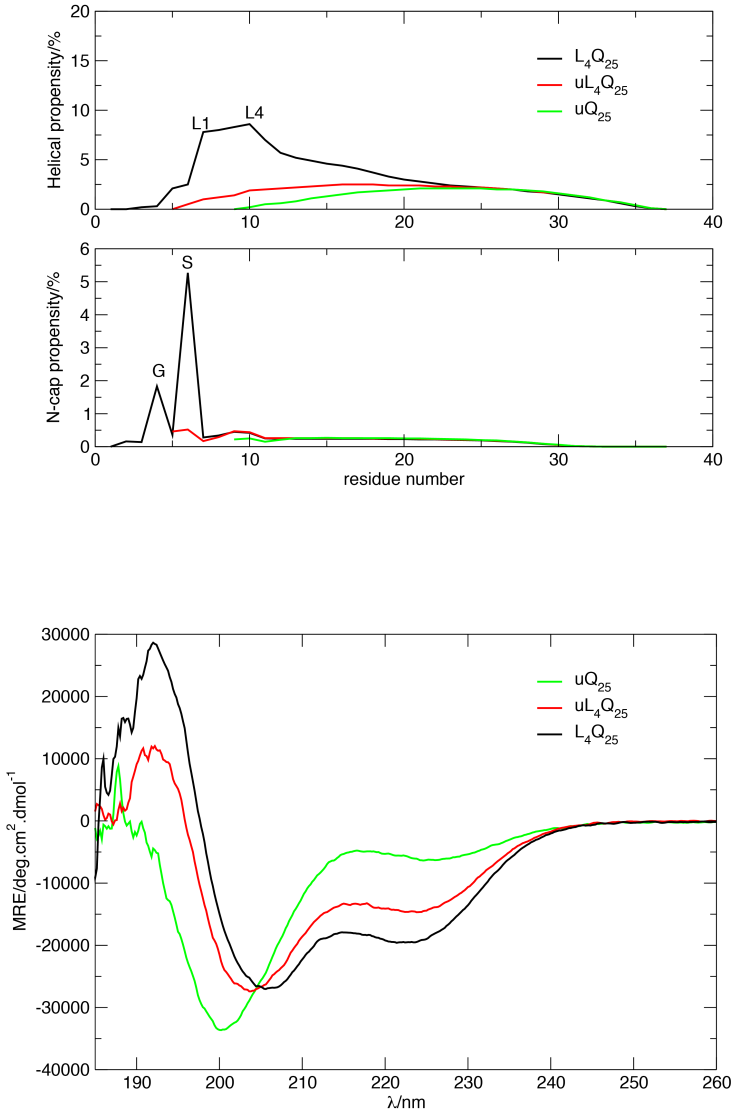

Figure 3.2: a. Prediction of helical and N-cap propensity for peptides uQ25, uL4Q25 and L4Q25 obtained by using Agadir [144]. The peptides are aligned so that the Gln residues, that are common to all of them, have the same residue numbers (Q11-Q35). b. CD spectra of uQ25, uL4Q25 and L4Q25 at 277K and pH 7.4.

Note that, to increase the solubility of peptides at physiological pH, all the peptides are flanked by pair of Lys residues. As can be interpreted from Figure 3.2b, both uL_4Q_{25} and L_4Q_{25} have higher helical content than uQ25. The total helicity

of uQ_{25} was calculated to be 20%, whereas that of uL_4Q_{25} and L_4Q_{25} was 40% and 55% respectively. These results are in agreement with the study performed with constructs AR_{1-156} 4Q-25Q and confirm that eight residues flanking the polyQ region at the N-terminus induce the formation of the helix.

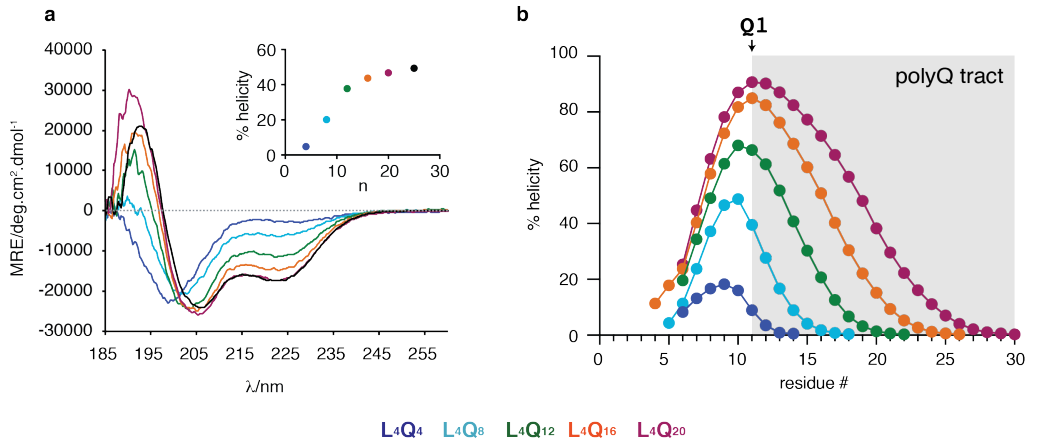

Figure 3.3: The stability of the androgen receptor (AR) polyQ helix increases upon tract expansion: a. CD spectra of peptides L_4Q_4 to L_4Q_{25} and plot of the helicity determined by CD as a function of the size of the polyQ tract length, n (inset, color coded). b. Residue-specific helicity of peptides L_4Q_4 to L_4Q_{20} obtained from an analysis of the backbone chemical shifts by using the algorithm δ2D[155] with an indication of the region of sequence corresponding to the polyQ tract and of its first residue.

3.2.2 Length of the polyQ tract and helicity

To show how helicity correlates with the length of the polyQ tract, we studied L_4Q_n polyQ peptides with n = 4, 8, 12, 16, 20. By CD we showed that helicity increases upon elongation of the polyQ tract. CD provides information about global helicity, but to determine the residue specific helicity, we used NMR and calculated secondary structure propensities with the algorithm δ2D by using the backbone chemical shifts[155]. The results in agreement with CD, showed that helicity increases on the N-terminal and propagates towards the C-terminal of the peptide. A striking result of that adding residues to the C-terminal of the peptides has an important effect on residues up to twenty position away. For example, the helicity of first Gln in L_4Q_{12} is 66% and adding four more Gln residues to the C-terminal of the tract increases the helicity of first Gln, which is fifteen residues away, to 85% in L_4Q_{16}. The maximum helicity goes from 20% for the peptide L_4Q_4 to 80% for the peptide L_4Q_{20} and the residue with maximum helicity moves from the first Leu to the first Gln respectively Figure 3.3. Also both NMR and

CD experiments report that the helicity saturates upon elongation. This means that the difference of helical propensity between L_4Q_4 and L_4Q_8 is much larger than between L_4Q_{16} and L_4Q_{20}.

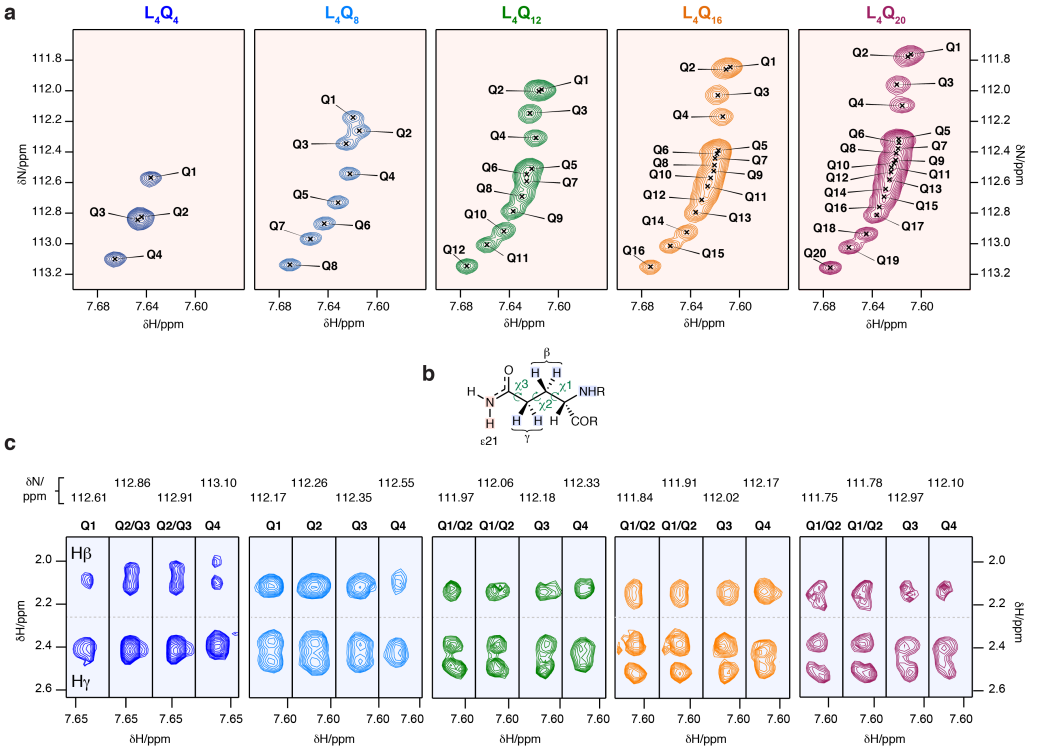

Figure 3.4: The conformations of the Gln side chains are well defined. a. Expanded regions of the ^1H, ^{15}N heteronuclear single quantum correlation (HSQC) spectra of peptides L_4Q_4 to L_4Q_{20} showing the Hϵ_{21} side chain resonances. b. Structure of the Gln side chain with an indication of the nuclei whose resonances are shown in a (red shade) and in c (blue shade). c. Regions of the ^{15}N planes of the H(CC)(CO)NH spectra of peptides L_4Q_4 to L_4Q_{20} measured at pH 7.4 and 278K containing the side chain aliphatic ^1H resonances of the first four residues (Q1–Q4) of the polyQ tract. All nuclear magnetic resonance (NMR) spectra were measured at pH 7.4 and 278K

3.2.3 The conformation of Gln side chains

Since all amino acids share the same backbone atoms and only differ in their side chains except for proline, the effects that lead to the different structural formation comes from the properties of their side chains. To rationalize our findings, we extended the study to the side chains of Gln residues. When we analyzed the NMR results, we found that the ^{15}N side chain resonances of Gln residues are well dispersed, and the ^{15}N chemical shifts of each Gln increase along the polyQ tract where the first four residues have lowest and most dispersed chemical shifts.

This result indicates that in the polyQ tract each Gln side chain has a different chemical environment, and the largest differences can be seen in the first four residues Figure 3.4a.

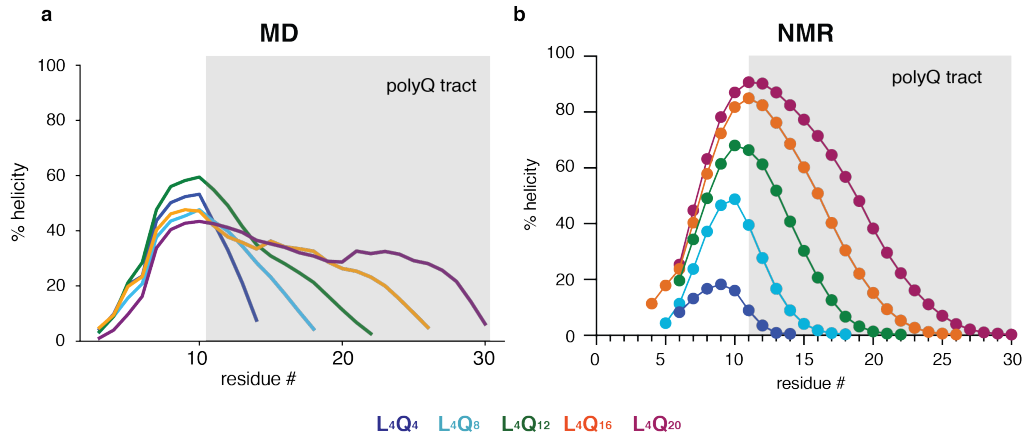

Figure 3.5: Residue specific helicity profiles for peptides L_4Q_4 to L_4Q_{20} obtained from a.MD b.NMR.

Secondly, we analyzed ^1H resonances of the Gln side chains. The analysis of ^1H spectra of these peptides suggested that the side-chain free motion is reduced in the glutamine residues where the helicity is high Figure 3.4c. This effect is strong, especially in the first four residues of the tract but still visible in the rest of it. To conclude, the NMR results indicate that side chain and main chain conformations are coupled, especially first residues of the polyQ tract have a distinct rotameric state and when helicity increases side chains behave more restrained. Since the backbone amide ^{15}N chemical shifts depend on the hydrogen bond status of the HN and CO next to it, we hypothesize that the dispersion of the ^{15}N chemical shift and redistribution of side chain rotameric states is carboxamide group of Gln side chains form hydrogen bonds[156].

3.3 MD Simulations

To rationalize these observations we used molecular dynamics simulations. MD simulations were run in MD simulation software ACEMD[157] using the CHARMM22* [158] force field at 300K. By using the information from the experiments, we generated fully helical conformations for the peptides L_4Q_4 to L_4Q_{20} as starting structures for the simulations. Each of the systems was explicitly solvated by using the TIP3P water potential inside a cubic box of water molecules. We produced 5 μs simulations for each peptide except for L_4Q_{20}, that we simulated for only 3 μs.

3.3.1 Disagreement between simulations and experiments

When we analyzed the results, we observed that even though the residues with the highest helicity were the four Leu residues from the flanking region, the helicity did not increase upon elongation in contrast to the experiments (Figure 3.5). The helicity of L_4Q_{12} and L_4Q_{16} is particularly much lower than the experiments. To be able to compare experiments and simulations one needs a converged simulations and even in this case, their accuracy depends on the quality of the force field. At the present time, nor force fields is perfect, and the computational power needed to produce converged simulation for most of the systems is not available. One of the solutions to these problems is combining experimental data with simulations.

There are many ways to incorporate simulations and experiments priorly but the most efficient way to do it posteriorly is reweighting simulations by using experimental data. By doing this, we generate a weighted ensemble more consistent with experiments without biasing the system.

3.3.2 Reweighting

Reweighting is a way to obtain a conformational ensemble with calculated averages of experimental observables closer to the experiment by combining imperfect simulations with experimental data. For reweighting to be successful, not convergence but sampling all relevant states is needed. The quality of reweighting depends on the existence of all states in the simulations. During the simulations, we observed many unfolding and folding events. Which is why even though we have not reached to the convergence, Figure 3.6 shows that we sampled most of the possible conformations. Having a trajectory sampled enough folded and unfolded state gave us to the possibility to use chemical shifts to reweight trajectories to obtain quantitative agreement between simulations and experiments(Figure 3.6). We used the Bayesian-Maximum Entropy (BME) algorithm to reweight our simulations[159, 160].

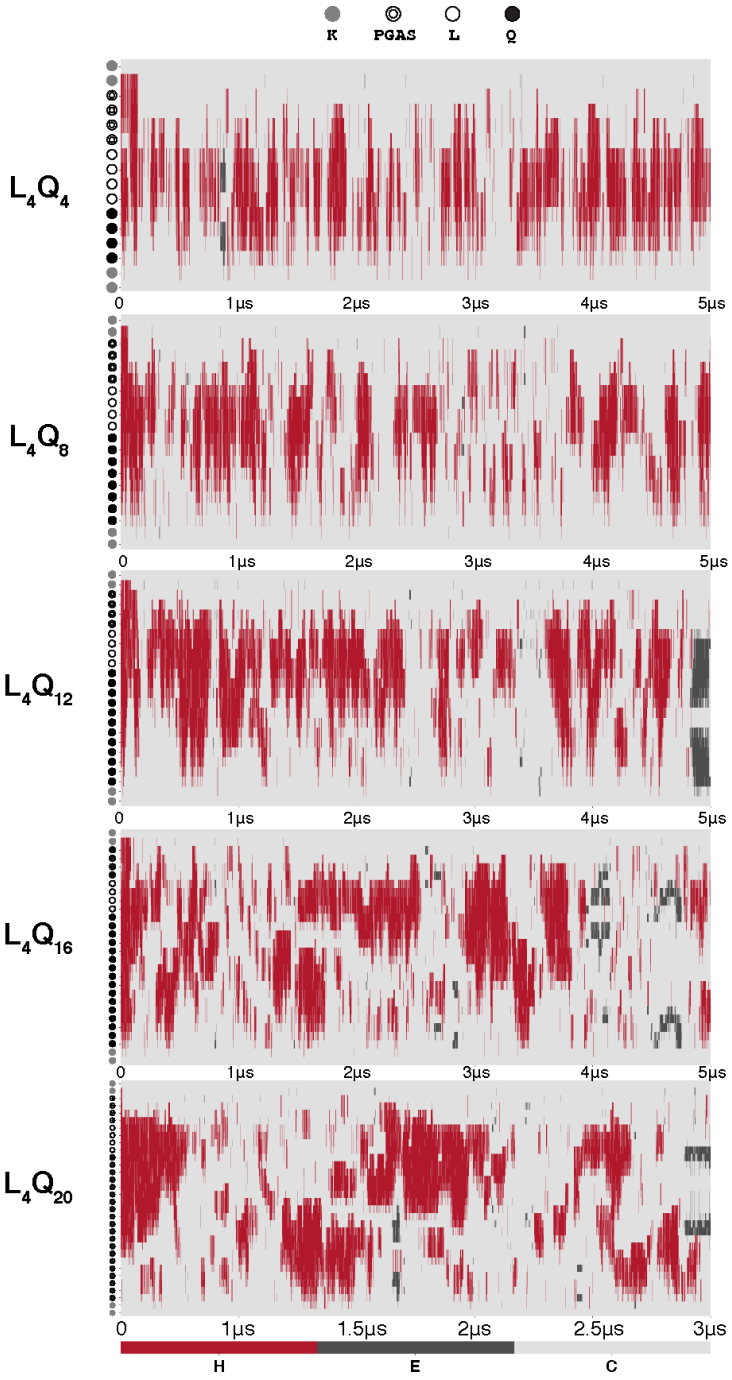

Figure 3.6: Time series of the secondary structure of peptides L_4Q_4 to L_4Q_{20} as obtained by using the algorithm DSSP[161] where H stands for helix, in red; E stands for extended, in dark gray; and C stands for coil, in light gray

Bayesian-Maximum Entropy

Maximum entropy (MaxEnt) methods aim to obtain the most appropriate distribution that fits the empirical data, by perturbating the prior distribution minimally [162, 163]. In our case, the prior distribution is the simulated ensemble and the empirical data is chemical shifts obtained by NMR. According to the MaxEnt principle, the probability distribution that describes the experimental data the most reliable is the one with maximum entropy. In this approach[160], BME maximizes the relative Shannon entropy [163, 164] (or Kullback-Leibler divergence[165] which is equal to the negative of Shannon entropy).

$$S_{REL}(P||P^0) = -\int d\mathbf{x} P(\mathbf{x}) log[\frac{P(\mathbf{x})}{P^0(\mathbf{x})}] \tag{3.1}$$

Here S_{REL} is relative entropy, $\mathbf{x}$ denotes the atomic coordinates, and $P^0(\mathbf{x})$ denotes the prior distribution. In ideal conditions, P^0 is close to the P^{TRUE} which defines the 'true' probability distribution. However, due to the inaccuracies in the model P^0 and P^{TRUE} are generally different than each other. With MaxEnt method, we aim to find a $P(\mathbf{x})$ as close P^{TRUE} as possible by introducing the minimum possible amount of information to system.

From NMR, we have chemical shifts as experimental values F^{exp}. We obtain the ensemble averages (F_i^{calc}) from M observables $F_i(x)$ which are back-calculated chemical shifts from the atomic coordinates by using the formula:

$$< F_i^{calc} >= \sum_{j=1}^{n} \omega_j F_i(\mathbf{x}) \tag{3.2}$$

where n is the number of the frames in the simulation and ω_j^0 is the weight of the j^{th} conformation. Classic MD simulations samples from the Boltzmann distribution, therefore weights are uniform.

$$\omega_j^0 = \frac{1}{n} j = 1, ..., n \tag{3.3}$$

However, in the BME method, we calculate the optimal weights by:

$$\omega_j^* = \frac{1}{Z(\lambda)} \omega_j^0 exp[\sum_{i}^{m} \lambda_i^* F_i(\mathbf{x})] \tag{3.4}$$

where Z is the partition function,

$$Z(\lambda^*) = \sum_{j=1}^{n} \omega_j^0 exp[-\sum_{i}^{m} \lambda_i^* F_i(\mathbf{x}_j)] \tag{3.5}$$

$\lambda^* = \lambda_1^* ... \lambda_m^*$ are Lagrangian multipliers that are used for maximizing Eq. (3.1) and Lagrangian multipliers are determined by minimizing:

$$\Gamma = log(Z(\lambda)) + \sum_{i}^{m} \lambda_i F_i^{exp} + \frac{\theta}{2} \sum_{i}^{m} \lambda_i^2 \sigma_i^2. \tag{3.6}$$

Since from experiments and the calculations, it is possible to introduce to the system both systematic and random errors, trying to force the system for a perfect match between experimental data and simulations could lead to incorrect results. Thus:

$$F_i^{exp} =< F_i^{calc} + \epsilon_i > i = 1, ..., m \tag{3.7}$$

Here in this equation, variable ϵ is included as an error model. So it's important to take into account the derivations when calculating the reweighted ensemble. ϵ does not depend on the coordinates and uncertainties are modeled by independent Gaussian distribution:

$$P(\epsilon_i) \propto exp(-\frac{\epsilon_i^2}{2\theta\sigma_i^2}) \tag{3.8}$$

where σ is the standard deviation. However, these error parameters hard to decide accurately. Therefore, theta a global scaling parameter is introduced to the equation. Large theta increases the uncertainty by multiplying all the sigma with large factor. A large sigma means high experimental error and this reverts to the prior distribution. Thereby, theta=0 corresponds to the perfect match between the experimental and the calculated data.

In Bayesian terms, we can obtain the optimal weights by minimizing the negative log posterior:

$$L(w_1 ... w_N) = \frac{m}{2}\chi^2 + \theta S_{REL} \tag{3.9}$$

where χ^2 shows derivation from the experimental averages:

$$\chi^2 = \sum_i^m (\sum_j^N w_j F_i(x_j) - F_i^{exp})^2 / m\sigma_i^2 \tag{3.10}$$

and S_{REL} is the relative entropy that defines derivation from the prior distribution:

$$S_{REL} = \sum_j^N w_j log(w_j / w_j^0) \tag{3.11}$$

3.3.3 Choosing the parameters

As explained before in the Eq. (3.9), a global scaling parameter, θ, is used fine tuning the weights. As it can be interpreted from the equation, if θ is too small, the fit between experimental data and reweighted simulations will be perfect but it will disturb the system to the levels where it can lead to unrealistic results. By contrast, if θ is chosen very large, results will be similar to the unreweighted simulations. The main aim of the maximum entropy principle is to change the statistical weights by disturbing the system minimally. So, θ has to be chosen in a careful way. To choose the best θ for the reweighting algorithm, we used a variety of θ values for all of the peptides and then compared their effects on the system by calculating: derivation from the experimental averages χ^2, derivation from the prior distribution S_{REL} (relative entropy) and effective fraction of frames N_{eff} that shows how many frames effectively contribute to the calculated averages and this value equals to the exponential of S_{REL}.

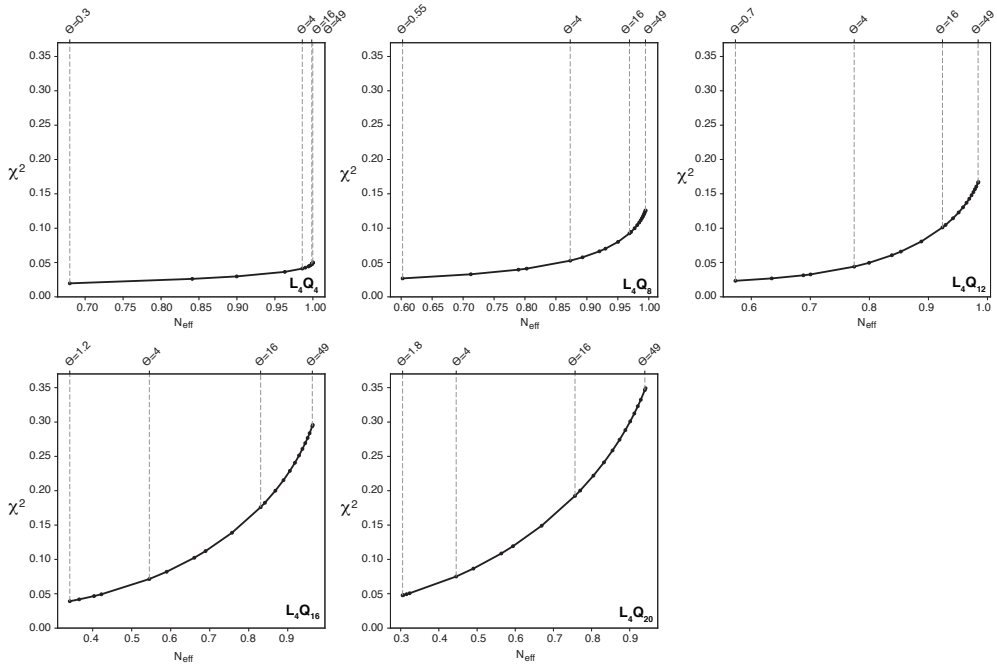

Figure 3.7: Effective fraction of frames after reweighting vs χ^2 for different values of θ. N_{eff} was calculated as $\exp(S_{REL})$ with S_{REL} being the relative entropy term in the BME reweighting approach. In short, N_{eff} quantifies the effective fraction of frames that are left after the trajectory has been reweighted to fit the data to an extent measured by χ^2.

The optimum θ, provides a setting where we obtain sufficient fitting to the experimental data while minimally perturbing the prior distribution. Therefore what we need is a balance between low χ^2 and large N_{eff}. In Figure 3.7, we plotted χ^2 vs large N_{eff} for each peptide for scanned θ values. To decide which value to choose as θ, the best approach is to choose the value at the corner or *'elbow'* of the curve. Thus we decided to use $\theta = 4$ for our calculations. Other than choosing θ, a first observation from the plots is that changing θ for L_4Q_4 has almost no effect on the weights. This indicates that the L_4Q_4 ensemble is already very close to the true ensemble obtained from experimental data. On the other hand, upon expansion of the polyQ tract, the difference between the effects of the low θ and high θ on the weights increases. In Figure 3.5, we showed their helicity is especially underestimated for L_4Q_{16} and L_4Q_{20}, which also agrees with the finding that the longer the tract, the ensembles are more different than the experiments. Secondly, as explained before we showed that when θ approaches 0, χ^2 gets closer to 0 which means that simulations agree with experimental data perfectly. However, the algorithm achieves this by up-weighting a small number

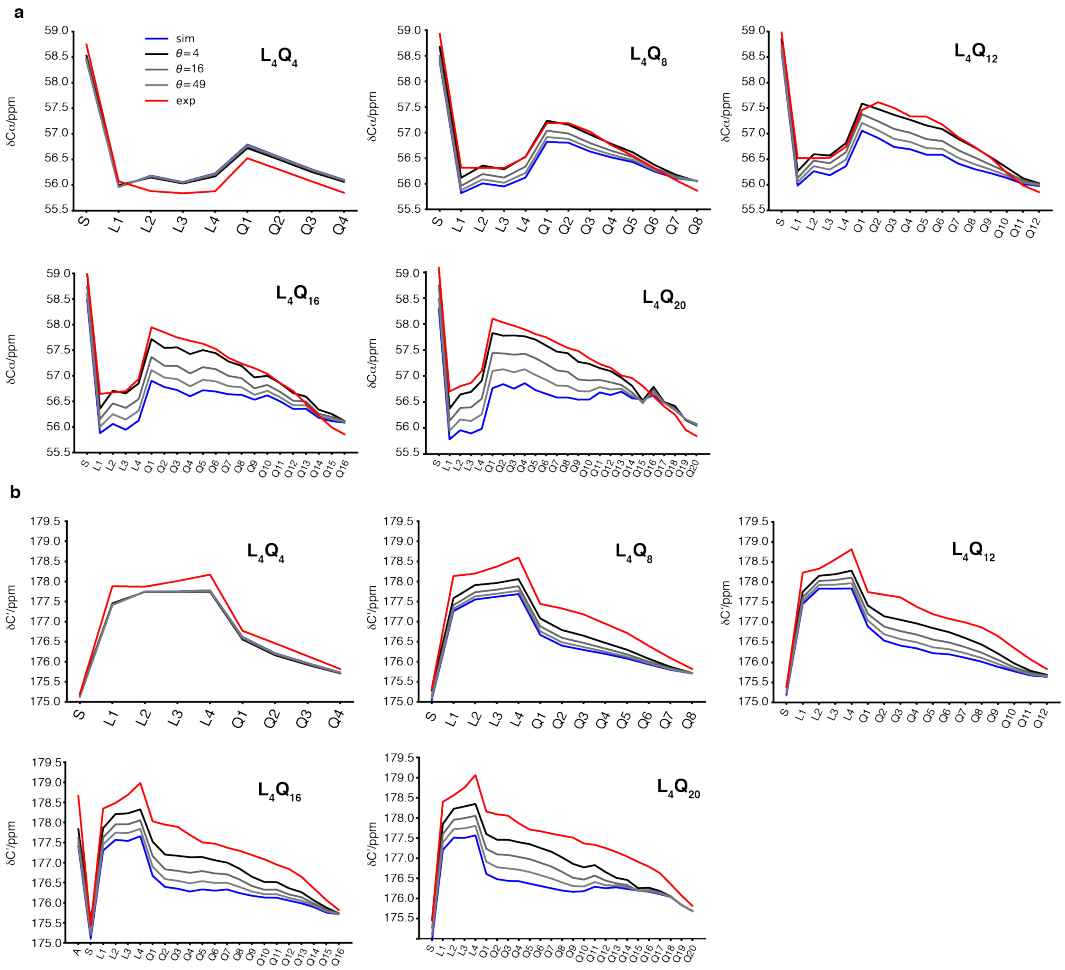

Figure 3.8: Experimental and back-calculated a. Cα and b. C' chemical shifts of peptides L_4Q_4 to L_4Q_{20} from the reweighted trajectories obtained with different values of the parameter θ, that determines the degree of reweighting applied in the BME algorithm.

of frames as can be seen from the small N_{eff} values and therefore disturbing the system excessively.

To see more clearly, how θ affects the ensembles and how much we improved our simulations by reweighting, we plotted the Cα and C' chemical shifts for each residue obtained from experiments, from MD simulations and from the reweighted ensembles (rwMD), as a function of θ, in Figure 3.8a-b. Note that, as stated before, the difference between experimental chemical shifts and back calculated chemical shifts from the simulation increases with the length of the polyQ tract. Also, it can be easily seen that $\theta = 4$ gives very similar calculated averages for chemical shifts to experimental data and increasing θ gives results to unreweighted

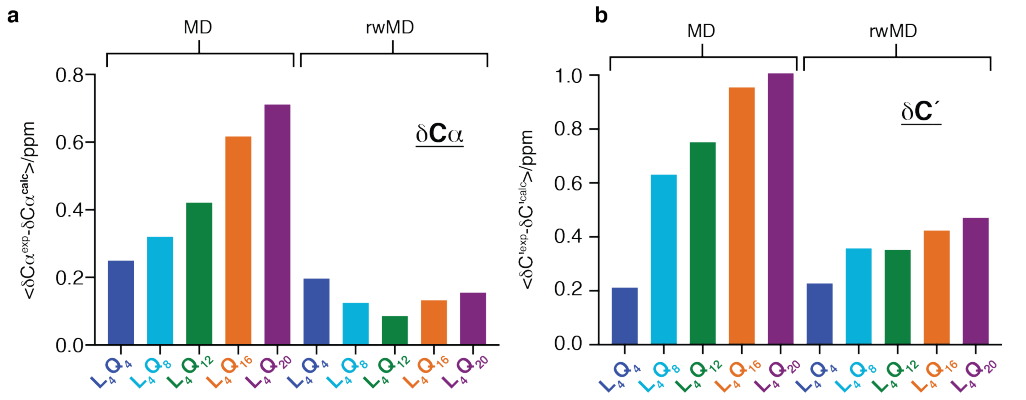

Figure 3.9: Comparison of the difference between the experimental and back-calculated chemical shifts for a. Cα and b. C'. Y-axis shows the error, which is calculated by taking the mean of the difference between calculated and experimental chemical shifts.

MD simulations. To sum up, the calculated errors between experimental and back calculated Cα and C' chemical shifts from MD and rwMD for L_4Q_4 to L_4Q_{20} can be seen in Figure 3.9.

3.3.4 Generating conformational ensemble that agrees with experiments

Using the weights obtained with $\theta = 4$, we calculated the new average residue-specific helicity for the peptides and obtained quantitatively similar helicity profile to experimental data we got from CD and NMR (Figure 3.10). This time in rwMD ensembles we observed an increase in the helicity with the expansion of the tract and helical propensities for the peptides similar to those obtained by δ2D. For example, as shown in the Figure 3.10, helicity of the peptides increased upon the elongation, and the maximum helicity of L_4Q_{20} went from 40% to 80% after reweighting. Having an ensemble with a secondary structure profile closer to the data obtained experimentally gives us the opportunity to use them to calculate other structural properties of the polyQ peptides.

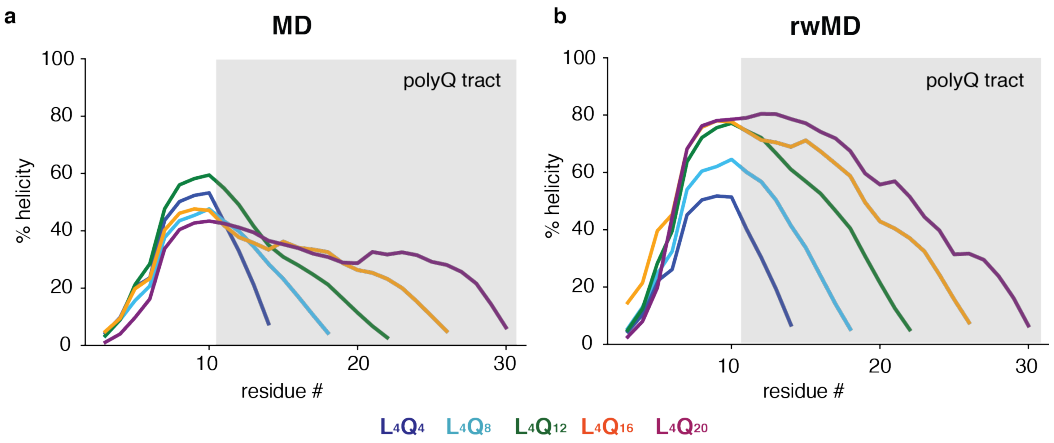

Figure 3.10: Residue specific helicity obtained for peptides before and after reweighting.

To visualize the average conformation of the peptides, we generated 3D models by using the average residue specific helicity profile from reweighted simulations (Figure 3.11). Each residue that has average helicity more than 50%, defined as a helical residue in the model and the rest as a random coil. Representative structures were selected directly from simulations. Coloring represents the secondary structure where it goes from dark blue (0% helicity) to dark red (78% helicity).

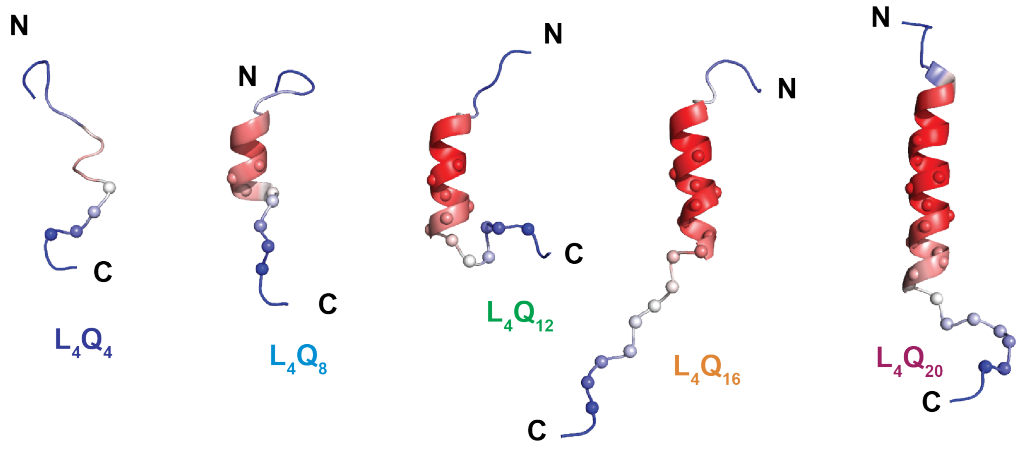

Figure 3.11: Representative structures for peptides L_4Q_4 to L_4Q_{20}, defined as the frame of each trajectory with residue-specific helicity most similar to the ensemble-averaged counterpart. Residues are colored as a function of their average helicity and the $C\alpha$ atoms of Gln residues are shown as spheres.

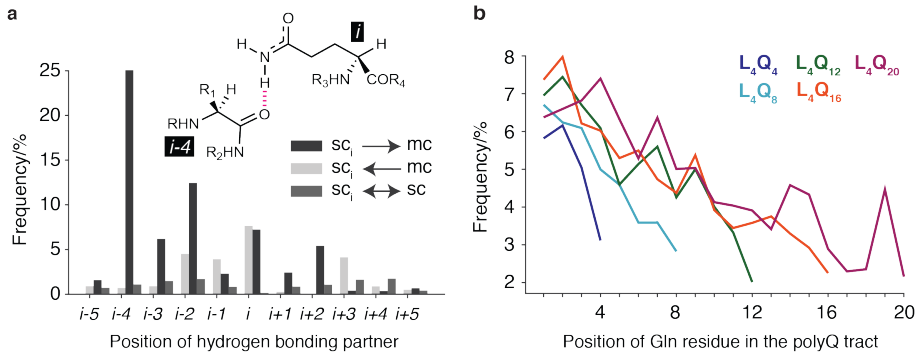

Figure 3.12: The helices formed by polyQ peptides feature $sc_i \rightarrow mc_{i-4}$ hydrogen bonds: a. Populations of the various types of hydrogen bonds involving Gln side chains. b. Populations of such hydrogen bonds in the reweighted ensembles obtained for peptides L_4Q_4 to L_4Q_{20} as a function of residue number.

3.3.5 $sidechain_i \rightarrow mainchain_{i-4}$ hydrogen bonds

From the NMR experiments, we hypothesized that the side chain chemical shifts indicate that the carboxamide group of the Gln side chain form hydrogen bonds. To investigate this we analyzed all possible hydrogen bonds formed by this group, and their populations, in the reweighted trajectories. As can be seen from Figure 3.12a, the most abundant hydrogen bond is formed between the Gln side chain NH_2 group at position i and the CO of the main chain of the residue at position *i-4*.

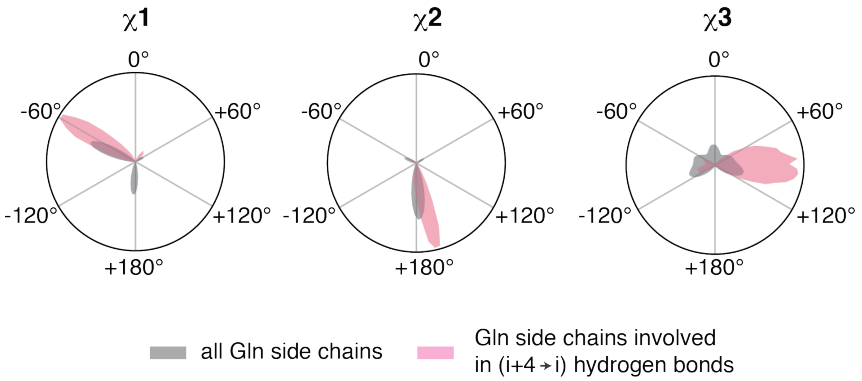

Figure 3.13: Dihedral angle distribution for Gln side-chains. Residues involved in $sc_i \rightarrow mc_{i-4}$ hydrogen bonds display rotameric selection.

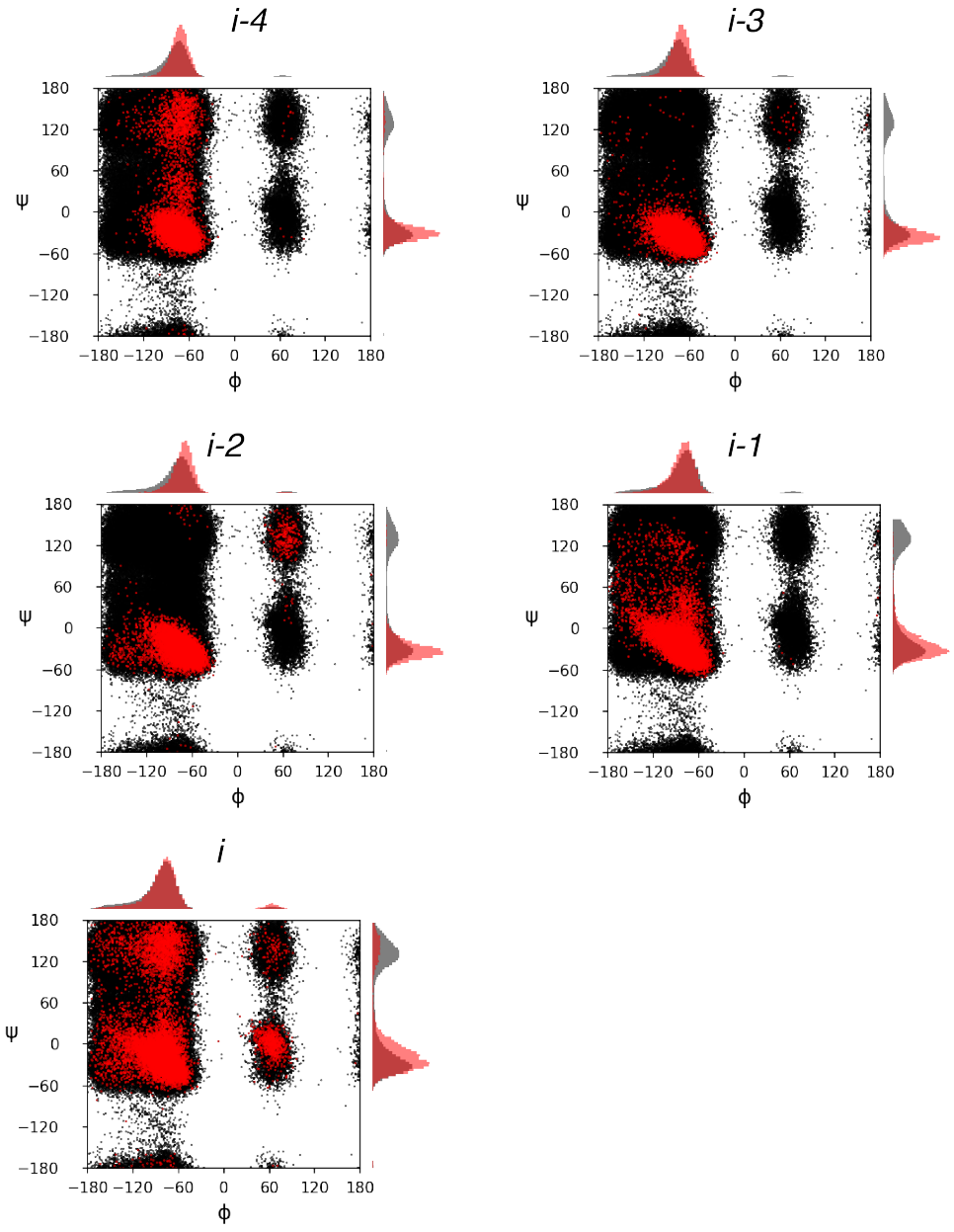

Figure 3.14: 2D histogram, in red, of backbone torsion angles ϕ and ψ for segments of five residues where the first and last residue are involved in a $sc_i \rightarrow mc_{i-4}$ hydrogen bond. In black we show the result obtained when these residues are not involved in such interaction.

We named this specific hydrogen bond as $sidechain_i \rightarrow mainchain_{i-4}(sc_i \rightarrow mc_{i-4})$ hydrogen bond. When we calculated the residue based population of $sc_i \rightarrow mc_{i-4}$ from reweighted simulations, we observed that it decreases along the polyQ tract (Figure 3.12b). This behavior is in agreement with the helicity profile we obtained from NMR experiments.

Since we also observed from NMR that the first residues of the polyQ tract have distinct rotameric state, we analyzed the dihedral angles of Gln side chains through the trajectory and also in the subset defined by side chains involve in $sc_i \rightarrow mc_{i-4}$ hydrogen bond. Gln side chain has three rotameric angles $\chi 1$, $\chi 2$ and $\chi 3$. In the reweighted simulations, while $\chi 1$ is generally bimodal (around -60°or +180°), $\chi 2$ is mostly around +180°and $\chi 3$ varies between -120°and +120°. By contrast, we observed that forming a $sc_i \rightarrow mc_{i-4}$ hydrogen bond constraints the $\chi 1$ and $\chi 3$ values. As NMR results pointed out, side chains adopt a specific conformation where $\chi 1$ is around -60°and $\chi 3$ is round +90°(Figure 3.13).

To check if forming $sc_i \rightarrow mc_{i-4}$ hydrogen bond causes any differences on the backbone torsion angles, we plotted and compared ϕ and ψ angles of residues i to i-4, both in the absence and existence of the $sc_i \rightarrow mc_{i-4}$ hydrogen bond in the Figure 3.14 as Ramachandran plots. In the plots, red dots represent the hydrogen bonded frames and black dots represent the rest of the frames. Since the plot is too crowded due to high number of frames, we added histograms to the axes of the Ramachandran plots. We can summarize that, in both of the cases, the distribution of the angles in the α-helix area of the Ramachandran plots remained the same. Also it can be observed that when a $sc_i \rightarrow mc_{i-4}$ hydrogen bond forms, the residues in between (*i-3, i-2, i-1* became more helical in agreement with the experimental results, that shows a correlation between helicity and restrained behaviour of the side chains.

In Figure 3.15, as an example, we show a frame from the trajectory of L_4Q_{16} which has two of the $sc_i \rightarrow mc_{i-4}$ hydrogen bond occurs simultaneously, shown in the dashed purple lines between Q1-L1 and Q4-L4. One other important result that we observed from the NMR derived structural ensembles is sc-mc hydrogen bonds are bifurcated. As can be seen from the example frame in Figure 3.15, the side chain and the main chain of Gln simultaneously donates a hydrogen atom to the CO of the Leu four residues away. We examined the properties of this bifurcated hydrogen bond in Chapter 4 by using QM/MM simulations.

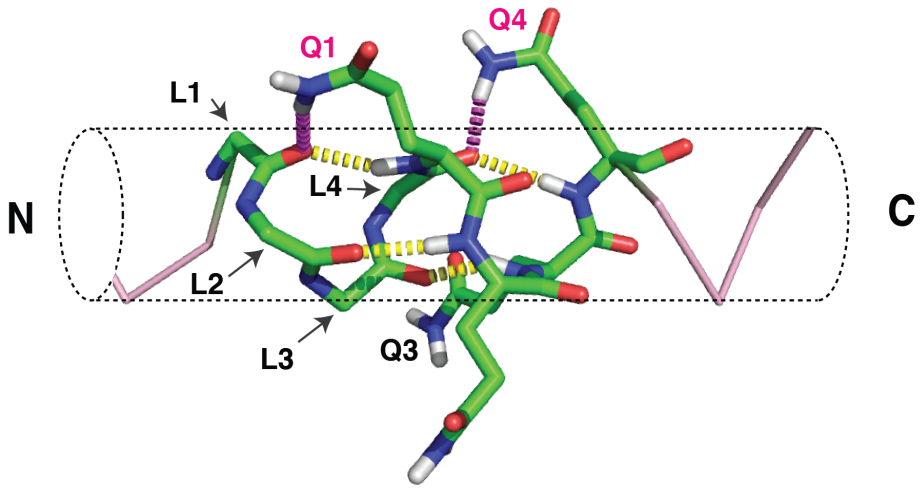

Figure 3.15: Frame of the trajectory obtained for peptide L_4Q_{16} where residues Q1 and Q4 (in purple) but not Q2 and Q3 (in black) are involved in $sc_i \rightarrow mc_{i-4}$ hydrogen bonds with the CO groups of residues L1 and L4 , shown in purple, with an indication of the conventional $mc_i \rightarrow mc_{i-4}$ main chain to main chain hydrogen bonds, shown in yellow. The Leu side chains are not shown, for clarity.

4
QM/MM Simulations

4.1 Introduction

Hydrogen bonds are the most important non-covalent interactions that stabilize secondary structures. Most of the studies focus on hydrogen bonds that involve a single donor and acceptor pair. However, due to the directionality of the lone pairs of the acceptor, a single oxygen atom can participate in two simultaneous hydrogen bonds. Yet, these types of interactions are misrepresented in the atom-centric representation of electrostatic interactions used in current molecular dynamics force fields, which might explain the problems we had in obtaining experimental helicities in classical MD simulations in Chapter 3.

In contrast to conventional molecular simulation force fields, which model hydrogen bonds as purely electrostatic interactions between the partial charges of the donor and acceptor, QM simulations take lone pair directionality and electronic polarization into account explicitly. Despite their accuracy, QM simulations are computationaly demanding and therefore are restricted to systems with a limited number of atoms. Since our focus is hydrogen bond forming and breaking, that occurs in a well-defined region, we decided to use combined quantum-mechanics/molecular-mechanics (QM/MM) method to overcome these limitations. The aim of this methodology is to use the QM method for the region that contains $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond and MM treatment for the rest of the system and

benefit from the accuracy of the QM and the efficiency of MM for the calculations.

4.2 The hydrogen bonds between Gln side chain NH_2 groups and main chain COs are bifurcate

4.2.1 Selecting QM region and the starting structure

We selected the starting structure, which contains a bifurcated hydrogen bond between Q1 and L1, from the classical MD trajectory obtained for L_4Q_{16}. As the QM subsystem, we included all the backbone atoms from Q1 to L1 and the side chains of only Q1 and L1 (make a fig for the seq). The QM subsystem was treated at the BLYP/6-31G* level, including dispersion corrections, while the classical subsystem was described with the CHARMM22* force field. We simulated the system for 150 ps at 300K.

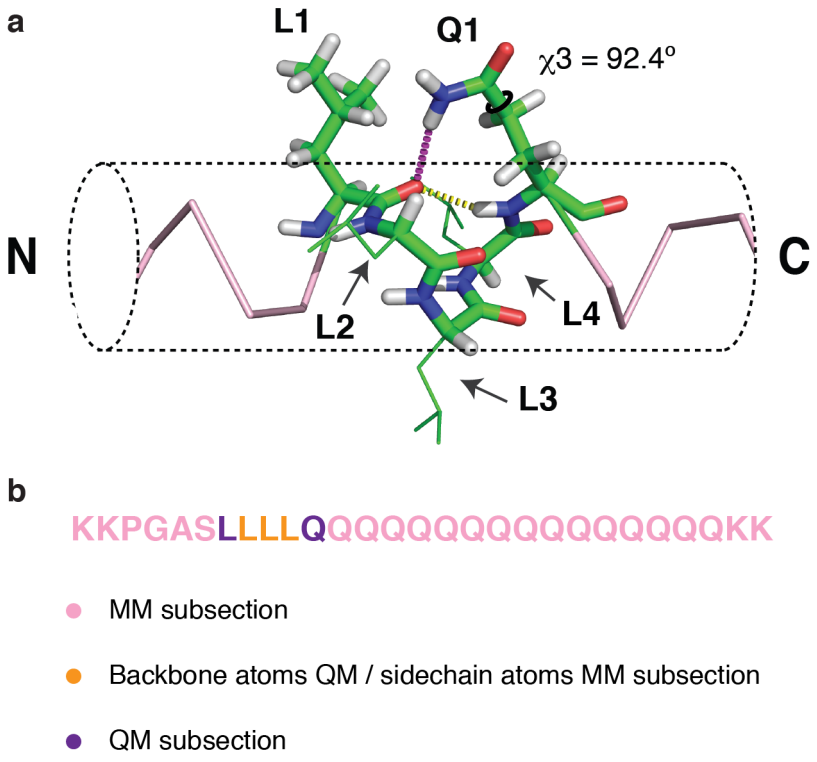

Figure 4.1: a.Starting configuration used in the QM/MM simulation, with the atoms included in the QM subsystem shown as sticks. b. QM subsystem contains atoms of L1,Q1 (Purple) and backbone atoms of L2,L3,L4 (Orange), and MM subsystem contains side chain atoms of L2,L3,L4 (Orange) and rest of the residues(Pink), water molecules are also included in MM region.

4.2.2 Analysis of the bifurcated hydrogen bond

Due to the specific geometry of the hydrogen bonds in proteins, the distance between the donor and the acceptor atoms is a clear indicator of the existence of the hydrogen bond, and we used 2.4$\AA$ as a strict threshold[166]. Therefore we first monitored both $sc_{Q1} \rightarrow mc_{L1}$ and $mc_{Q1} \rightarrow mc_{L1}$ by measuring the distances between L1-O and Q1-Hϵ_{21}, for the $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond, and between L1-O and Q1-HN for the $mc_{Q1} \rightarrow mc_{L1}$ hydrogen bond (Figure 4.2a). We observed that, while the $mc_{Q1} \rightarrow mc_{L1}$ bond is stable, $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond breaks and forms reversibly.

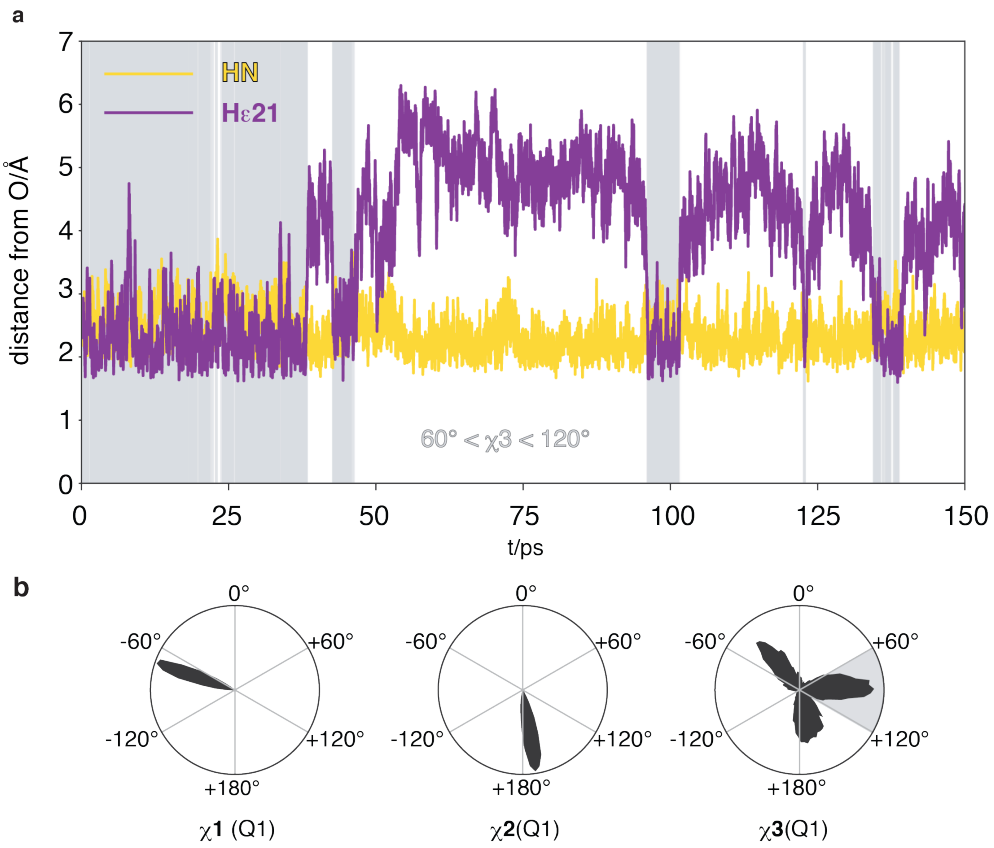

Figure 4.2: The $sc_i \rightarrow mc_{i-4}$ hydrogen bonds bifurcate with $mc_i \rightarrow mc_{i-4}$ hydrogen bonds: a.Time series of the distances between donor and acceptor for the $mc_{Q1} \rightarrow mc_{L1}$ (Yellow) and $sc_{Q1} \rightarrow mc_{L1}$ (Purple) interactions, with an indication, in a gray background, of the frames for which 60°$< \chi^3 <$ 120°b.Distributions of the χ^1, χ^2, χ^3 dihedral angles of the side chains of Q1 with an indication, as a gray shade, of the range of values of χ^3 that are compatible with the $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond.

For further information, we analyzed the rotamer angles of the Q1 side chain.

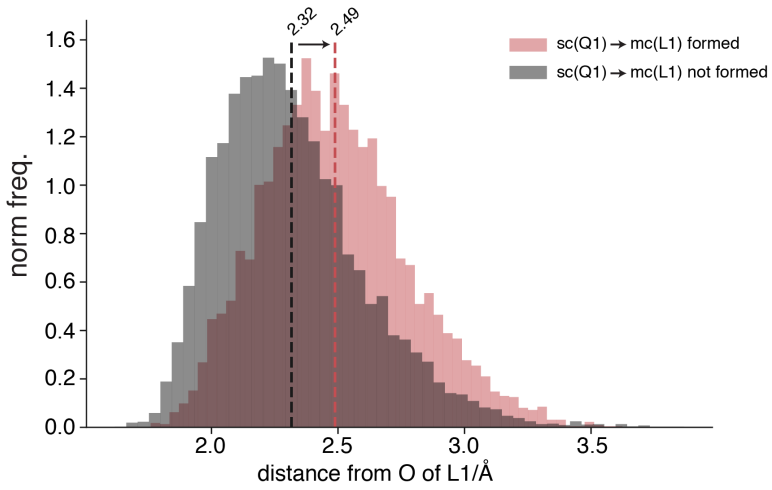

Figure 4.3: Distribution of the distance between the main chain NH of Q1 and the main chain CO of L1 in the absence and in the presence of the $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond.

Due to the time scale of the QM simulations, the $\chi 1$ and $\chi 2$ angles are stable during the simulation, but $\chi 3$ fluctuates and samples three different conformations (Figure 4.2a). In Chapter 3, we showed that the $\chi 1$ and $\chi 3$ angles of the Gln side chain are constrained when it forms a $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond, and $\chi 3$ needs to be between 60°and 120°(mostly around 90°). When we highlighted frames with 60°$< \chi 3 <$ 120°, we observed that forming and breaking of the $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond in QM/MM simulations strongly correlated the $\chi 3$ values.

4.2.3 Strength of the bifurcated hydrogen bond

Since constraining the side chain conformation is entropically costly we investigated the contribution of a $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond on the helix stability. First, we checked the effect of the $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond on the $mc_{Q1} \rightarrow mc_{L1}$ hydrogen bond. When we compared the distribution of the distances of donor to acceptor atoms of the $mc_{Q1} \rightarrow mc_{L1}$ hydrogen bond, we found out that the presence of the $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond shifts distribution to longer distances(Figure 4.3). In other words, the formation of the $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond weakens the $mc_{Q1} \rightarrow mc_{L1}$ hydrogen bond.

One of the benefits of the QM/MM approach is we can calculate the electron density from the simulations. Therefore, we evaluated the specific strength of these hydrogen bonds by exploring the topology of the electron density distribution and taking advantage of the known correlation between the intrinsic strength and the electron density at the interaction's natural bond critical points (ρ(r))[167, 168].

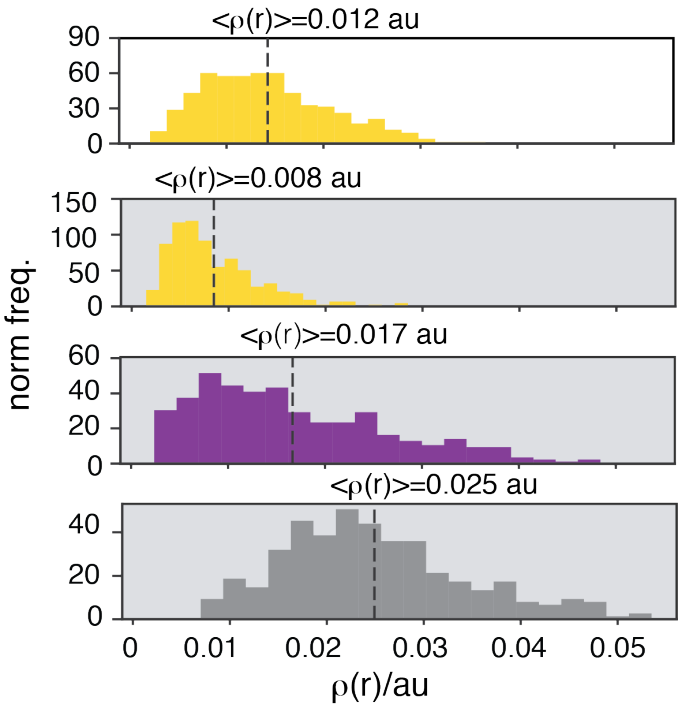

Figure 4.4: Distribution, plotted as a normalized histogram, of the electron density ρ(r) corresponding to the $mc_{Q1} \rightarrow mc_{L1}$ interaction (Yellow) in the absence (white background) and in the presence (gray background) of the $sc_{Q1} \rightarrow mc_{L1}$ interaction and of the electron density corresponding to the $sc_{Q1} \rightarrow mc_{L1}$ interaction (Purple) and bifurcate interactions (Gray).

Our analysis showed that, when the $mc_{Q1} \rightarrow mc_{L1}$ hydrogen bond is formed but the $sc_{Q1} \rightarrow mc_{L1}$ is not, its average ρ(r) value is of 0.014 au. In the presence of the $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond, the average ρ(r) value decreases to 0.008 au. However, the average ρ(r) value for the $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond in the absence of the $mc_{Q1} \rightarrow mc_{L1}$ hydrogen bond was observed as 0.017 au, indicating that the Gln side chain is a better donor than the NH backbone group[169]. More importantly, the average total density associated with the bifurcated hydrogen bonds was calculated as 0.025 au (Figure 4.4) pointing towards a very strong interaction that enhances the stability of the polyQ helices.

5

Replica Averaged Restrained Simulations

5.1 Introduction

While performing their functions, proteins populate different conformations such as open-close states of membrane proteins or disorder to ordered transformations of IDPs[170–172]. Also by definition, in their native states, IDPs adopt large ensembles of heterogeneous conformations[173, 174]. Therefore it's important to describe the proteins as an ensemble of conformations instead of a single structure. In principle, with the right model and sufficient sampling, it is possible to have an exact atomic-level description of conformational ensembles with MD simulations. However, even though there is a great effort in the field to improve the force fields[158, 175–177], they are still not sufficiently accurate[178, 179]. Therefore most of the time, the computations are not in quantitative agreement with experimental data.

To overcome this problem, there has been significant progress in the development of techniques that combines simulations with experimental data [180–185]. Using experimental data in simulations allows us to increase the accuracy of the force fields, therefore, obtain results consistent with experiments such as NMR spectroscopy, small-angle X-ray scattering (SAXS), fluorescence resonance energy transfer (FRET), or cryo-electron microscopy (cryoEM)[186–188]. Among these methods that give information about the dynamics and the structure of proteins,

NMR is especially powerful to provide different parameters that are sensitive to protein dynamics over different timescales, as well as to study the protein structure in solution.

Chemical shifts are highly sensitive to structural changes, and they are the most accurate and fast NMR observables[189, 190]. Recent advances in computational techniques for calculation backbone chemical shifts from structure data fast and reliably led to the development of computational methods that incorporate chemical shifts with simulations[156, 191–195]. There are two main ways to combine simulations with experimental data. The first way is as explained in the Chapter 3, reweighting the trajectory *posteriori* by using the experimental data. The second way is using the chemical shifts (or another structural data) during a simulation.

It has been shown that it is possible to study the dynamics and the structure of the proteins accurately using chemical shifts as structural restraints in MD simulations. The idea behind is very similar to the standard approach that has been used for the refinement of experimentally defined NMR structures by using a penalty function to obtain structures consistent with the experimental results[196–198].

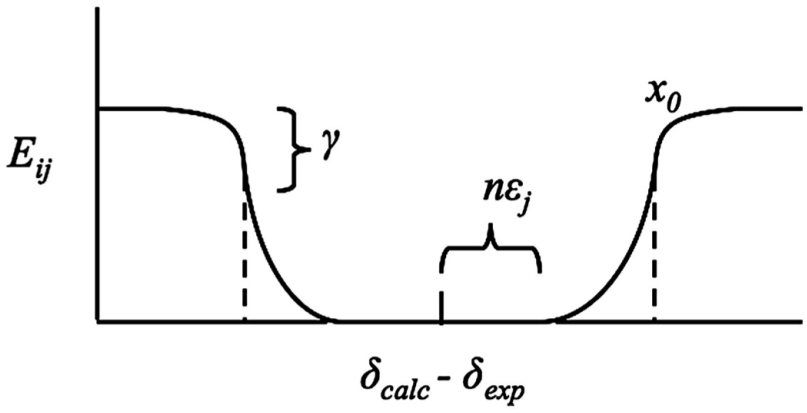

Figure 5.1: Illustration of chemical shift penalty function. E_{ij}, where i is the residue number and j is the chemical shift type, gives the contribution of each chemical shift to the total penalty E_{CS}. δ_{calc} is calculated and δ_{exp} is experimental chemical shifts. $n_{\varepsilon j}$ controls the width of the flat region. The penalty is harmonic until the deviation reaches a cutoff value x_0 . γ is a parameter that controls the growth of the penalty function beyond x_0 [199].

5.1.1 Using chemical shifts as structural restraints in MD simulations

Using the chemical shifts as structural restraints enables us to modify force fields in a system-dependent manner by using the experimental data to overcome its limitations[199–201]. For this purpose, a penalty function is defined to keep the conformational search withing the conformational space that agrees with experimental data[200]. To define this penalty function and convert differences between experimental and calculated backbone chemical shifts (H_α, C_α, C_β, C', H_N, N), mostly a flat-bottomed harmonic potential is being used (Figure 5.1)[199].

The penalty function needs to be calculated at each time step of the simulation. For this reason, computing the chemical shifts rapidly is one of the important steps of this approach. CamShift is a program that predicts the backbone chemical shifts using distance dependent functions of the atomic coordinates[191]. CamShift functions are differentiable, therefore it allows us to calculate forces from chemical shifts. Since the previous chemical shift predictors, such as SHIFTX[194], were using discontinuous functions, it was only possible to use them with Monte Carlo simulations[200]. However, CamShift enabled us to restraint also the MD simulations. The differences between experimental chemical shifts and back-calculated shifts are computed during the MD simulations and then converted to a penalty function:

$$E_{CS} = \alpha \sum_{i=1}^{N} \sum_{j} E_{ij} \tag{5.1}$$

where E_{CS} is the total chemical shift penalty energy and E_{ij} (Figure 5.1) is the contribution of each chemical shift. Here i is the residue number and the j is the chemical shift type (H_α, C_α, C_β, C', H_N, N). α describes the weight of the E_{CS} relative to the force field term E_{FF}. The total energy of the system is calculated by:

$$E_{total} = E_{FF} + E_{CS} \tag{5.2}$$

And the derivative of the Eq. (5.1) gives force between the two atoms in the simulation in each direction:

$$f_{(x,y,z)}(a,b) = -\frac{\partial E_{CS}}{\partial (x,y,z)} \tag{5.3}$$

The size of the force vectors is directly related to the slope of the penalty function E_{ij} which is defined by

$$E_{ij} = \begin{cases} 0, & \text{if } |\delta^{ij}_{calc} - \delta^{ij}_{exp}| < n_{\varepsilon j} \\ \left(\frac{|\delta^{ij}_{calc} - \delta^{ij}_{exp}| - n_{\varepsilon j}}{\beta_j}\right)^2, & \text{if } n\varepsilon j < |\delta^{ij}_{calc} - \delta^{ij}_{exp}| < x_0 \\ \left(\frac{(x_0 - n_{\varepsilon j})}{\beta_j}\right)^2 + \gamma \cdot tanh\left(\frac{2(x_0 - n_{\varepsilon j}(|\delta^{ij}_{calc} - \delta^{ij}_{exp}| - x_0)}{\gamma \beta_j^2}\right), & \text{for } x_0 < |\delta^{ij}_{calc} - \delta^{ij}_{exp}| \end{cases} \tag{5.4}$$

where δ_{calc} is calculated and δ_{exp} is experimental chemical shifts. n is a tolerance parameter that controls the width of the flat region of the penalty function, and ε is the accuracy of predictions for the type j of the chemical shifts. x_0 is the cutoff value that defines the point until where the penalty function is harmonic. γ is a parameter that controls the growth of the penalty function beyond x_0. Prediction algorithms are not completely accurate. Therefore back-calculations of the chemical shifts may contain errors; however, having a flat-bottomed region in the penalty function allows a window where small variations from experimental values do not generate a penalty[199, 200].

5.1.2 Replica averaged molecular dynamics simulations

However, the NMR chemical shifts are the results of time and ensemble-averaged measurements. To reflect the true nature of the experimental data, another option is not enforcing the system with a simple restraint on one simulation, instead, applying the restraint over an averaged experimental observable. Having more than one replicas and restraining their average is also a way to prevent the risk of over restraining. For the replica averaged simulations, the total penalty function Ecs is defined by[184, 202]

$$E_{CS} = \alpha \sum_{N}^{i=1} \sum_{6}^{j=1} \left(\delta^{ij}_{exp} - \frac{1}{M} \sum_{m=1}^{M} \delta^{ij}_{calc}\right)^2 \tag{5.5}$$

where M is the number of the replicas. It has been shown that four replicas are enough to determine the structure and dynamics of the proteins accurately. Increasing the number of replicas will lead to an increase in computational cost[202, 184].

More importantly, it has been proved that by restraining the simulations over the averaged experimental data we obtain conformational ensembles compatible with the maximum entropy principle. Therefore, the incorporation of experimental

data as replica-averaged structural restraints in molecular dynamics simulations provides an accurate representation of the unknown Boltzmann distribution given an approximate force field and a set of experimental data[183, 203, 204].

5.2 Results

In this project, we used four replicas for the peptides from L_4Q_4 to L_4Q_{16}. First, we generated an ensemble by the Monte Carlo method to select our initial conformations. For each replica, we chose four different initial conformations randomly by assuring that they cover different structural states from fully helical to disordered. MD simulations were performed using Gromacs modified with Plumed and CHARMM22* force field at 278K. We used the Cα, C', HN, and N chemical shifts as restraints. During the simulation, the chemical shifts of the simulated structures were back-calculated with Camshift[191].

Briefly, in replica-averaged restrained simulations, we back-calculated chemical shifts for each replica at each step. To determine the replica-averaged values, these back-calculated chemical shifts were linearly averaged and were compared to the experimental data to penalize deviations between them. Therefore, the average of the ensemble is restrained to fit the experimental data, while the fluctuation of individual replicas is allowed.

5.2.1 Secondary structure calculations

We analyzed the secondary structure of the replicas and observed that each replica of the peptide has a different helical profile as expected from replica-averaged simulations(Figure 5.2). Since we restrained the average of the four replicas, we need to interpret the helical profiles of the averaged conformations. To analyze the convergence and time evolution of the simulations, we performed block analysis on the averaged conformations by dividing trajectory to six equal blocks. For each block, we linearly averaged the helical profiles from four replicas. From Figure 5.3, we observed that the helicity of the L_4Q_4 and L_4Q_8 peptides is overestimated compared to the reweighted simulations and secondary structure chemical shift analysis. In addition, L_4Q_{12} and L_4Q_{16} show a less smooth profile compared to L_4Q_4. This indicates that they have not converged yet.

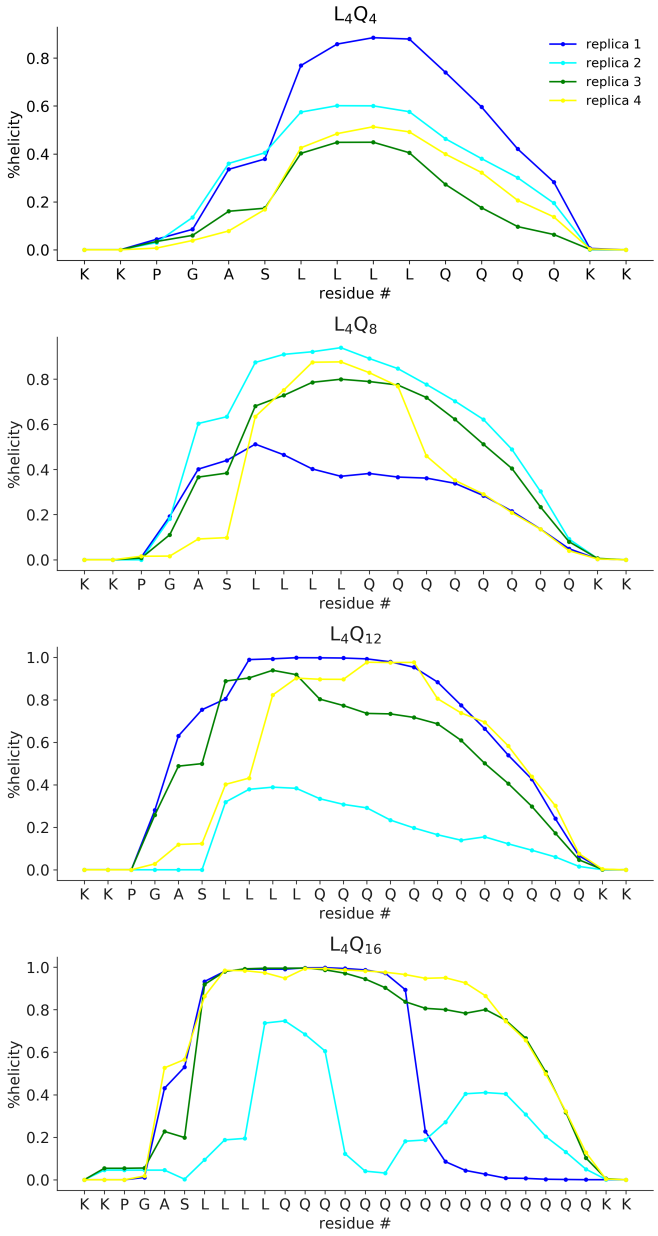

Figure 5.2: Residue specific helicity profiles for peptides L_4Q_4 to L_4Q_{16}.

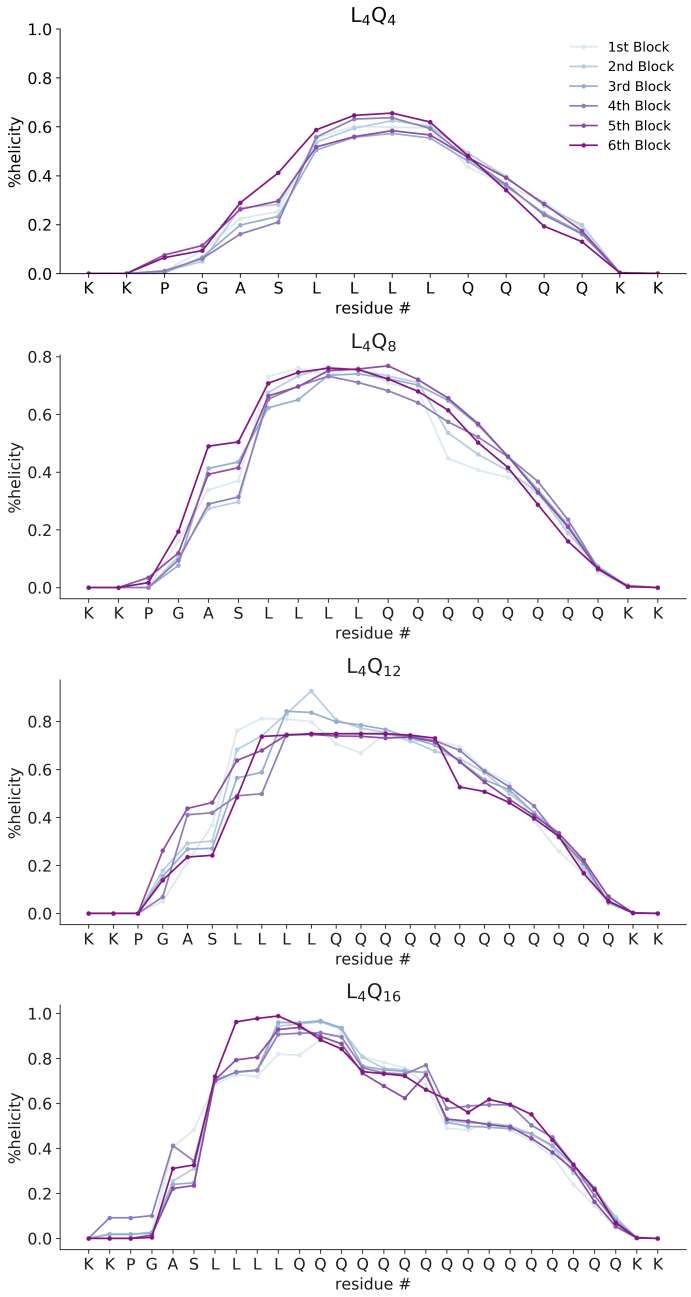

Figure 5.3: Block averaging of helicity profiles for peptides L_4Q_4 to L_4Q_{16}.

5.2.2 $sidechain_i \rightarrow mainchain_{i-4}$ hydrogen bonds

We then analyzed the $sc_i \rightarrow mc_{i-4}$ hydrogen bond populations both in replicas and their averages(Figure 5.5). We observed that the population of the $sc_i \rightarrow mc_{i-4}$ hydrogen bond directly correlated with the helicity of the peptide. In the peptides with higher helical populations have a higher number of frames with $sc_i \rightarrow mc_{i-4}$ hydrogen bonds in their trajectories. The most important result we obtained is that the averaged $sc_i \rightarrow mc_{i-4}$ hydrogen bond population of the peptides is similar to the reweighted simulations and higher than classical MD simulations.

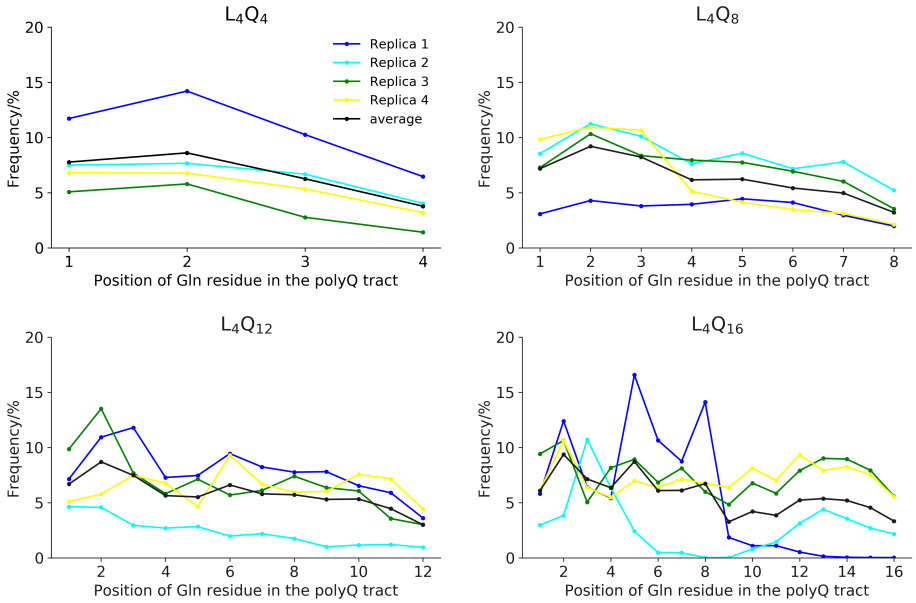

Figure 5.4: Populations of $sc_i \rightarrow mc_{i-4}$ hydrogen bonds hydrogen bonds.

When we analyzed the $\chi 1$, $\chi 2$ and $\chi 3$ distributions of all Gln residues involved in the $sc_i \rightarrow mc_{i-4}$ hydrogen bond, we observed that results are very similar to both reweighted and QM/MM simulations. In this regard, it is clear that the range of values that $\chi 1$, $\chi 2$, and $\chi 3$ angles can adopt are constrained and they are independent of method or force field used for calculations.

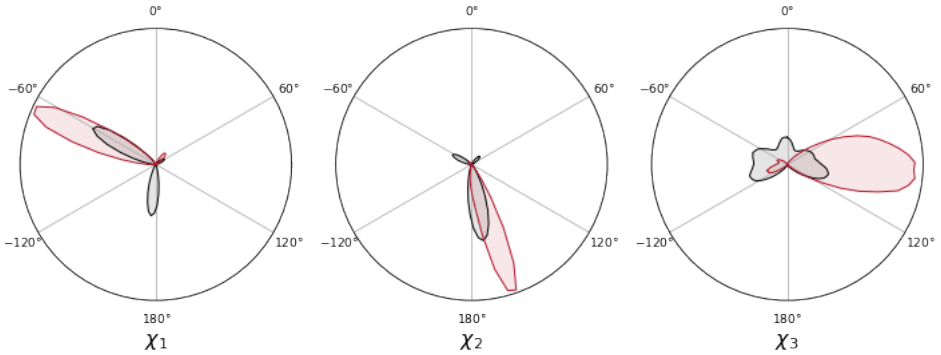

Figure 5.5: Dihedral angle distribution for Gln side chains.

6

Accounting for Side Chain to Main Chain Hydrogen Bonds in PolyQ Helicity Predictions

6.1 Introduction

The secondary structures of polypeptides are stabilized by hydrogen bonding interactions between main chain NH and CO groups. The side chains of many residues can however also act as hydrogen bond donors and acceptors. This is in fact one of the factors that explains why different amino acids have different propensities to adopt certain specific secondary structures [134]. For example asparagine (Asn) has a low propensity to be in the area of the Ramachandran plot corresponding to α-helices due to the ability of the carboxamide group of its side chain to hydrogen bond the main chain in other conformations [134]. We recently showed that Gln side chains, by contrast, can instead stabilize this secondary structure by donating a hydrogen to the main chain CO of the residue at relative position i-4[205]. This hydrogen bond, that we call Gln side chain to main chain hydrogen bond ($sc_i \rightarrow mc_{i-4}$), is bifurcate with the conventional main chain to main chain $i \rightarrow i-4$ hydrogen bond.

Since the ability of this interaction to contribute to helix stability has just been

established[205] algorithms that predict the secondary structure of peptides from their sequences[144] do not yet account for it. In sequences containing few Gln residues this does may not have serious consequences but in sequences where they are frequent, such as prion-like domains and, especially, polyQ tracts, this can lead to severe under-estimations of helicity[205]. This can in turn lead to an under-appreciation of this secondary structure in important processes for biomedicine such as aggregation in neurodegeneration[206]. To remedy this, we introduced a term in agadir to account for this interaction, calibrated its value by using nuclear magnetic resonance (NMR) data obtained for a series of peptides derived from the sequence of the transactivation domain of the androgen receptor (AR). These findings have implications for understanding of the mechanism of transcription activation and challenge the common notion that transcription factors are not druggable due to their disordered nature.

6.2 Introduction of a Gln side chain to main chain hydrogen bond in Agadir

6.2.1 Agadir underestimate AR polyQ helicity

Agadir is an algorithm successfully predicts their helical content from the sequence of peptides based on statistical mechanics[144]. However for the polyQ tract of AR, helicity was underestimated when compared to the secondary structure propensity profile obtained from NMR chemical shifts and the difference between algorithm and experiments were increasing due to expansion of the tract (Figure 6.1). Agadir calculates helical propensity of the peptides from each residue's intrinsic tendency to be in helical conformation, contribution of main chain-main chain hydrogen bonds and side chain-side chain interactions, capping effects of the residues and electrostatic interactions. However, in our previous work we showed that Gln side chains can stabilize secondary structure by donating a hydrogen to the main chain CO of the residue at relative position *i-4* and this interaction was not implemented in Agadir. To improve the prediction, we introduced an additional term in the equation describing the change in energy corresponding to the formation of side chain to main chain helix-stabilizing hydrogen involving Gln residues, $\Delta \mathrm{E}_{i+4,i}$ in the algorithm. From the previous work we know that the strength of this interaction strongly depends on the identity of the hydrogen bond acceptor [205], and Leu is better acceptor than Gln. Since $\Delta \mathrm{E}_{i+4,i}$ has different value for each amino acid (e.g. X) the term is in principle specific for it,$\Delta {E^X}_{i+4,i}$.

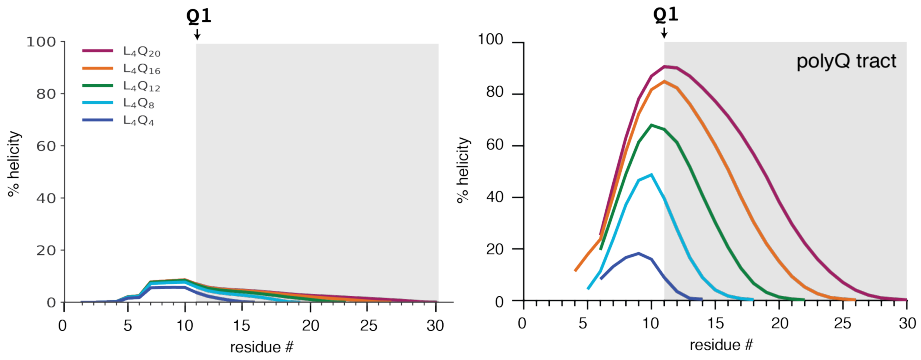

Figure 6.1: a.Helicity predictions for the peptides from current version of Agadir b.Helicity profile of the L_4Q_n peptides as derived from the analysis of the backbone chemical shifts measured at pH 7.4 and 278K by using the algorithm δ2D, with an indication of the region of sequence corresponding to the polyQ tract.

6.2.2 Predicting Helicity of polyQ peptides correctly

Since we already have residue specific helical propensities obtained from NMR experiments for the peptides L_4Q_4 to L_4Q_{20}, to fix the Agadir for the proteins have polyQ regions, we used AR as a starting point. We used L_4Q_n peptides which are composed of the motif Pro-Gly-Ala-Ser, that acts as a N-capping sequence, of four Leu residues that accept side chain to main chain hydrogen bonds donated by the first four Gln residues of the polyQ tract and of n Gln residues[205].

In the context of the AR polyQ region, four Leu residues that preceding the polyQ tract form $sc_i \rightarrow mc_{i-4}$ hydrogen bond with first four Gln residues in the polyQ tract. Thus Leu-Gln was the first amino acid couple we optimized the energy contribution of $sc_i \rightarrow mc_{i-4}$ hydrogen bond in Agadir. To determine the value of $\Delta E^L{}_{i+4,i}$, we used the residue-specific helical propensities calculated by using the algorithm δ2D[207] from the backbone chemical shifts. We then determined the values of $\Delta E^L_{i+4,i}$ that minimize the RMSD between the residue-specific helical propensities determined experimentally and those predicted with the modified version of Agadir by using the equation(6.1) below:

$$RMSD = \sqrt{\frac{1}{length}\sum_{res=1}^{length}(Hel_i^{exp} - Hel_i^{pred}(\Delta E^X_{i,i+4}))^2}. \qquad (6.1)$$

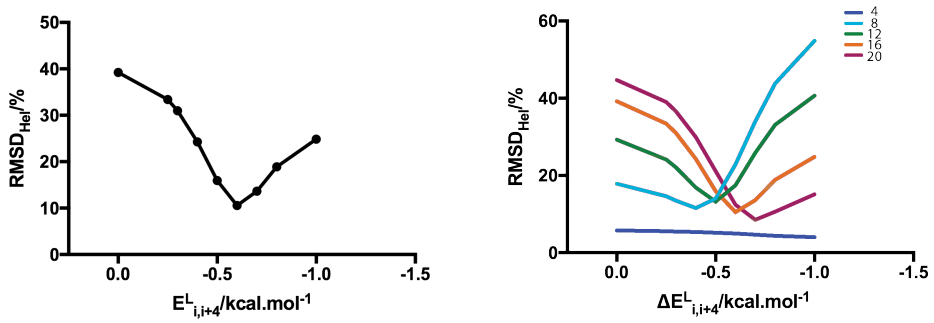

Figure 6.2: a.Optimization of $\Delta E^L_{i+4,i}$ with a X_4 peptide containing 16 Gln residues b.Optimization of $\Delta E^L_{i+4,i}$ for X_4 peptides with polyQ tracts containing 4, 8, 12, 16 and 20 Gln residues.

For example for the L_4Q_4 peptide, we found that $\Delta E^L_{i+4,i} = -0.6$ kcal.mol^{-1}, as shown in (Figure 6.1). As it can be seen from the Figure 6.2a, just increasing the affinity between Gln side chain and Leu we decreased the difference between the predicted helical content and the experimental content from 40% to 10%, and this value is around 7% for the 778 peptides tested for the latest version of Agadir. Therefore we can safely state that, introducing $sc_i \rightarrow mc_{i-4}$ hydrogen bond to Agadir fixes the helicity underestimation problem for the AR polyQ tract[151].

6.3 Cooperativity in side chain to main chain hydrogen bonding

One unexpected property of the peptide sequence derived from the transactivation domain of the AR is that its helical propensity increased monotonically with polyQ tract length in the range 4 to 25[205]. In addition, our NMR experiments indicated that the increase of helicity occurred throughout the sequence i.e. that increasing tract length by adding Gln residues to its C-terminus caused substantial increases of helicity in residues at the N-terminus. This suggests that the formation of the polyQ helix is cooperative i.e. that the effective strength of the side chain to main chain hydrogen bonds is not independent on their number.

To determine whether this is indeed the case we repeated the procedure to optimize $\Delta E^L_{i+4,i}$ with peptides containing 4, 8, 12 and 20 Gln residues. First, we observed that the RMSD between the helicity predicted by the current version of Agadir and that determined experimentally increased from ca 5% to ca 45% upon elongation of the tract (Figure 6.2b). In other words, the quality of the prediction

decreased as the number of Gln increased. We also obtained that the predicted helicity was not sensitive to the strength of the interaction for a peptide containing 4 Gln residues, but that it instead markedly depended on it for peptides containing 8 to 20 Gln residues. Note that the peptide containing 4 Gln residues can possess 4 Gln side chain to Leu main chain hydrogen bonds but no Gln side chain to Gln main chain hydrogen bond. For peptides with more than four Gln residues we found, in agreement with our hypothesis, that the effective strength of the interaction depended on the number of residues found in the tract, ranging from -0.4 kcal.mol^{-1} for the peptide containing 8 Gln residues to -0.7 kcal.mol^{-1} for the peptide containing 20 Gln. In summary, we found that the effective strength of $sc_i \rightarrow mc_{i-4}$ hydrogen bonds depended directly on the number of Gln residues following them.

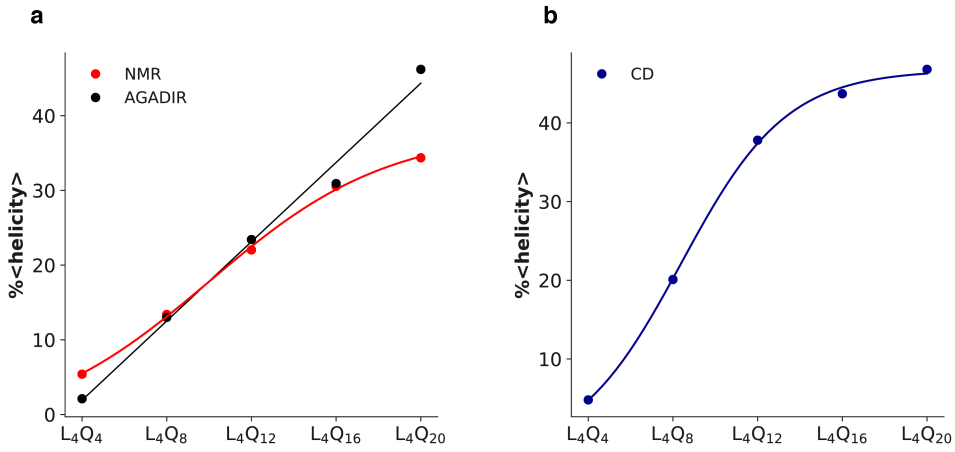

Figure 6.3: Comparison of fractional helicities of the peptides L_4Q_4 to L_4Q_{20}. a. Red dots represent fractional helicity obtained from an analysis of the backbone chemical shifts, black dots fractional helicity obtained from Agadir calculations with optimized $\Delta E^L_{i+4,i}$ value for each length. b. Plot of the helicity determined by CD as a function of the size of the polyQ tract length.

Another indication of cooperativity is sigmoidal behavior[208, 209]. When cooperativity is present, plotting helicity profile as a function of residue number produces a sigmoidal curve. When we plotted fractional helicity of each peptide calculated by optimized $\Delta E^L_{i+4,i}$ values with Agadir, we obtained a linear behavior. Linearity indicates that Agadir does not take the cooperativity of the $sc_i \rightarrow mc_{i-4}$ hydrogen bonds into consideration(Figure 6.3a). However, as it can be seen from Figure 6.3a-b, curves obtained from NMR and CD reproduce the sigmoid features

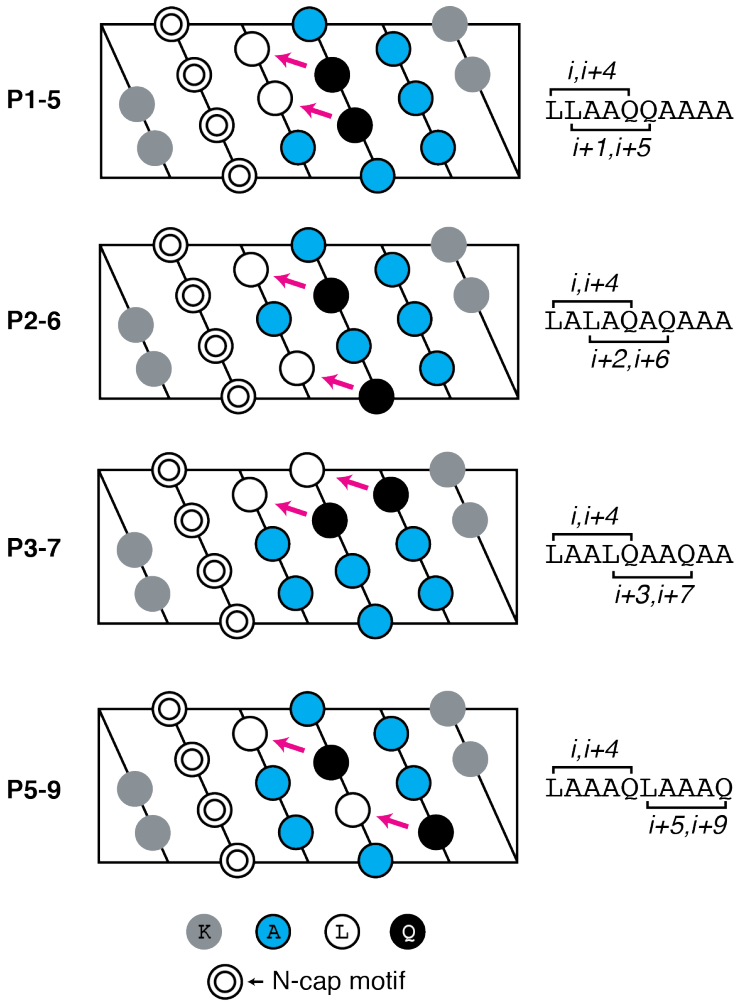

Figure 6.4: Helical projections of four peptides with the same amino acid composition and potential.

of the length dependence of the helicity data. In summary, to be able to fix the Agadir, besides optimizing $\Delta E^{X}_{i+4,i}$ values, the cooperativity effect also has to be added to the calculations.

Since through the C-terminal, the polyQ tract loses its helicity, and Gln residues are not good acceptors as Leu residues, the sigmoid curve is not very steep. Therefore we hypothesize that it is possible to have steeper sigmoidal curves and higher helicity with mutating Gln residues to Leu residues in the positions that promote the cooperativity.

6.3.1 Structural basis for the cooperativity between side chain to main chain hydrogen bonds

We designed peptides able to form two side chain to main chain peptide bonds in an equivalent sequence background but differing in relative positions to investigate the source of the cooperativity, as shown in Figure 6.4. Each peptide has two Leu, two Gln and six Ala residues. By keeping first Leu-Gln pair constant in their positions and we moved second pair one by one.

To quantify how helicity depends on the positions of side chain to main chain hydrogen bonds we first studied these peptides by CD. From the Figure 6.5, we can see that peptide P3-7 has higher helical content whereas P2-6 has prominently lowest and P1-5 and P5-9 are very similar and slightly less helical han 3-7. Which means all peptides have different helical propensities even though they have same amino acids and same number of side chain to main chain hydrogen bond. Therefore, it is clear that both Agadir and CD results point effect of the cooperativity on the formation of helices.

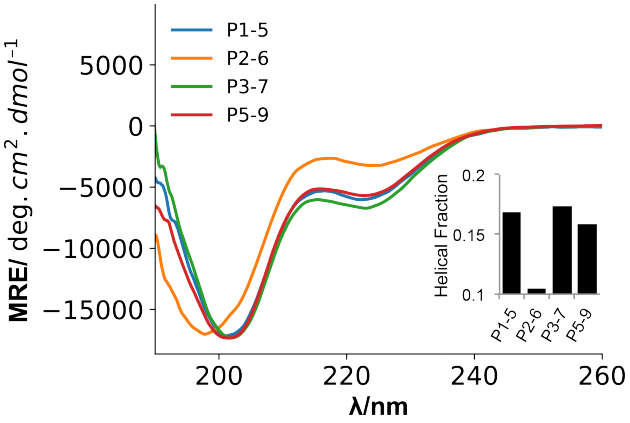

Figure 6.5: CD spectra of the cooperativity peptides measured at pH 7.4 and 278K and plot of the helicity determined by CD.

Since P3-7 has an arrangement that leads to higher helicity, we examined the structure and found out that it is the only peptide two side chain to main chain hydrogen bonds share a peptide plane. It has been proposed that hydrogen bonds between backbone amides and backbone CO groups play a role in folding cooperativity because the two hydrogen bonds that peptide planes establish in proteins are not independent, as shown in Figure 6.6.

They are on the one hand inter-dependent because the changes in electron density caused by each bond render the other more likely as a consequence of

the charge transfer processes. In addition, the planar nature of the peptide bond fixes the relative orientation of the acceptor and donor groups and thus reduces the entropic cost of forming two (or an array of) simultaneous hydrogen bonding interactions. This effect has been invoked to explain the presence of weak but measurable correlations in the backbone motions of residues in adjacent strands in β-sheets, observed both in conformational ensembles determined from NMR data[210] and in equivalent ensembles determined from the overlay of a large number of β-sheet backbones extracted from high resolution X-ray structures[211]. We thus hypothesized that equivalent effects may underlie the cooperativity between two bifurcated side chain/main chain to main hydrogen bonds that share a peptide plane in a polyQ tract.

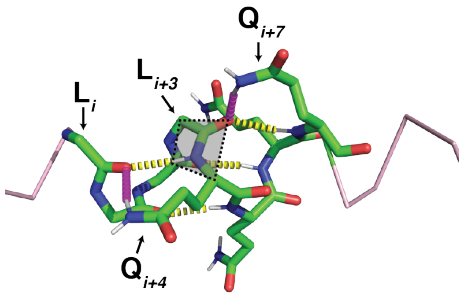

Figure 6.6: Frame of MD simulation of peptide L_4Q_{16} where two side chain to main chain hydrogen bonds in a *i,i+3,i+4* and *i+7* pattern are formed with an indication in grey of the peptide plane connecting residue at positions *i+3* and *i+4*.

7

Bioinformatics analysis of polyQ proteins

7.1 Is it only AR?

In the previous chapters, we showed that four Leu residues preceding the polyQ tract in the androgen receptor (AR) have a critical role in keeping the polyQ in an α-helical conformation by forming an $sc_i \rightarrow mc_{i-4}$ hydrogen bond. To understand if we can generalize findings from AR to other proteins bearing polyQ tracts, we extracted human proteins with polyQ from UniProt[212] to analyze their sequences.

7.2 Definition of polyQ

PolyQ repeats can be in different lengths and most of them can have insertions of amino acids other than Gln. PolyQs with insertions are called imperfect polyQs. To generate a dataset with proteins that contain polyQ regions, first, we need to set a definition for it(Figure 7.1). Main rules are:

- Stretch has to start and end at least 2 Gln residues.
- The total number of Gln residues has to be at least 7.

- 70% of the consecutive stretch of amino acids has to be Gln.
- Insertions cannot be longer than 5 residues.

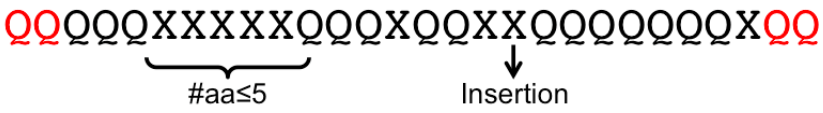

Figure 7.1: Model of the polyQ definition, **X** represents insertion residues.

Especially for the composition analysis of the preceding residues at the N-ter of the polyQ region, it is important to define where does polyQ starts and ends clearly. Other definitions with different criteria were also tested to check the robustness of the criteria and its effect on the results(appendix).

7.3 Dataset

Reviewed human proteome with 20,431 proteins were downloaded from the UniProt [212] database in FASTA format[213, 214]. After applying polyQ criteria to all sequences with the method we developed by using R software[215], we obtained a dataset that contains 153 polyQ stretches. Then each polyQ tract was extracted from the protein with ten extra residues at the terminals in 'X_{10}-QQQXQ..QQ-X_{10}' format. As seen in the Figure 7.2, we named residues from p1 to p10 (position 1 to position 10) to address them. To analyze the characteristics of the N-ter of the polyQs and compare them with the human proteome, we also generated a random dataset. From each protein in the human proteome obtained from the UniProt, we randomly extracted 10 residues. We generated three different random datasets to be sure that our results are random enough to do statistical analysis.

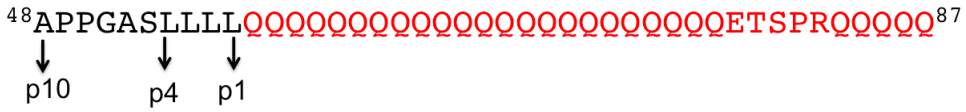

Figure 7.2: Scheme of the relative positions of the studied residues from the N-terminal of the polyQ tract. Here AR polyQ tract has been used as an example. The closest residue to the tract is called p1, and we named them till p10.

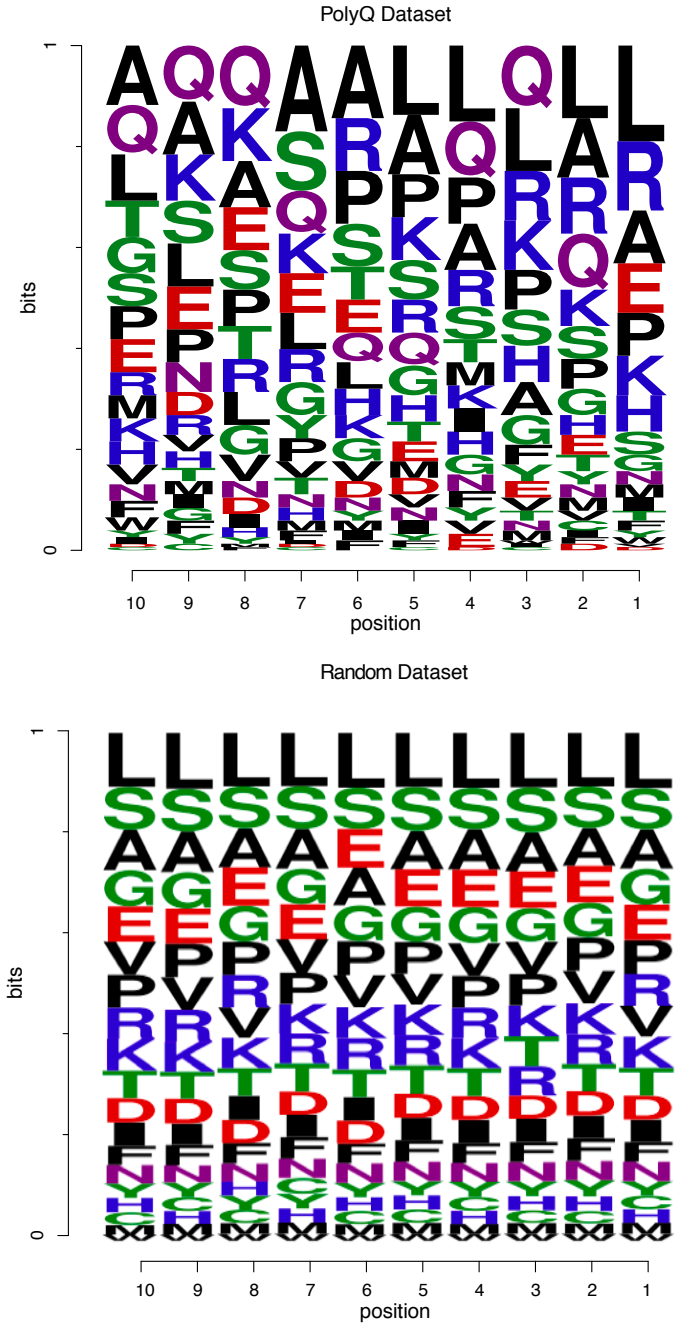

Figure 7.3: Sequence logo representation for polyQ and random datasets. Size of the letters are related to their frequency on that position and colors denote the properties of amino acids (**Black** = Hydrophobic, Green = Hydrophilic, Purple = Amidic, Red = Charged(-), Blue = Charged(+)).

7.4 Amino acid composition

In previous chapters, we showed that four Leu residues preceding the polyQ tract in AR have a critical role in its folding to α-helix by forming $sc_i \rightarrow mc_{i-4}$ hydrogen bonds. Therefore, we first analyzed the amino acid composition of N-ter of polyQs up to 10th residue and compared them with the random dataset.

We calculated the frequency of each amino acid for each position in the N-ter regions of both polyQ and random datasets. In Figure 7.3, we used the sequence logo[216] approach calculated with 'seqLogo'[217] library in R to represent amino acid frequencies in different positions. The sequence logo method displays amino acids on top of each other with different heights proportional to their frequencies. This approach makes it easier to determine the importance of the entities and their relative frequencies. As seen from the Figure 7.3, Leu is already the most abundant amino acid in the human proteome. However, in the polyQ dataset, it clusters around the first four positions just before the polyQ tract and its frequency decreases through the position 10. Also in the random dataset, each amino acid has the same abundancy independent from its position in the sequence. This result confirms that our dataset is random. Even though the sequence logo gives us a general idea about the frequencies, we also analyzed the sequences in more detail.

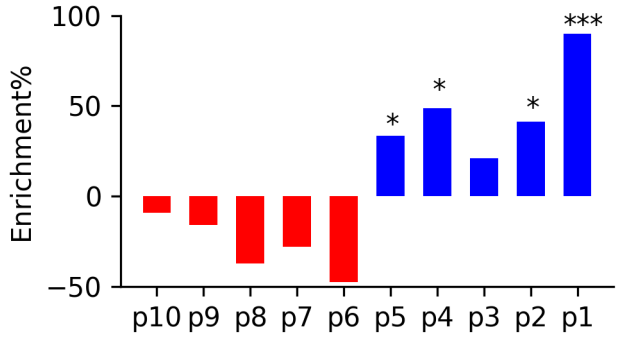

Figure 7.4: Enrichment of Leu residues when compared to random dataset obtained from human proteome. '*'s denote statistical significance.

In the random dataset, Leu has an average occurrence of 10%. However, when we analyze the polyQ dataset, we see that the first five positions preceding the polyQ tract are highly enriched when compared to the random dataset. Especially p1 is with %95.7 score highly enriched. After the fifth position, it starts to deplete.

When we focused on the first position(p1) preceding the polyQ tract, we observed that other than Leu, especially Arg, His are over-represented in the polyQ

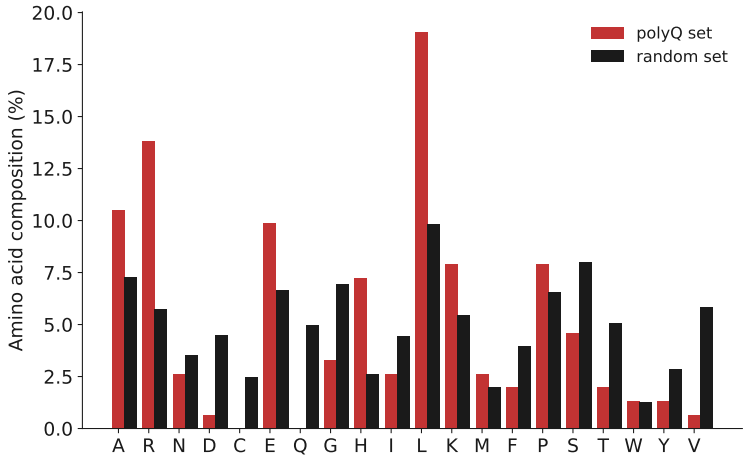

Figure 7.5: Amino acid composition of the first positions of the random and the polyQ datasets.

dataset and Asp, Cys, Val, and Gly are under-represented (Figure 7.5). From previous studies, it is known that Pro and His amino acids are enriched in particularly insertions of imperfect polyQs and depleted in Asp, Cys, and Gly amino acids[218].

7.5 Structural analysis

PolyQ region of AR folds into the alpha-helix by forming sc-mc hydrogen bonds with four Leu residues at the N-ter of the region. One turn of an alpha-helix contains four amino-acids, and their ability to form a helix is important for the nucleation of the helix. Therefore, we predicted the propensities of the residues at the positions 1 to 4 for the fragments in the polyQ dataset and calculated the average helicity for each fragment by using these four residues (Figure 7.6). For the residue-level helicity predictions, we used Agadir. Then we compared the helicity profile of the polyQ dataset and the random dataset.

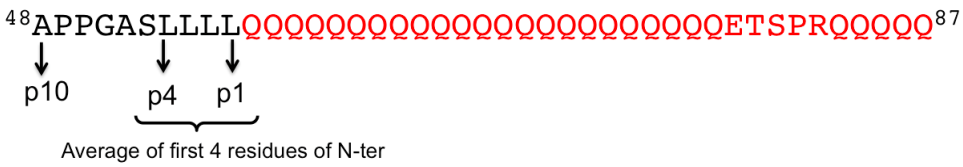

Figure 7.6: Residues p1 to p4 were used for the helicity predictions.

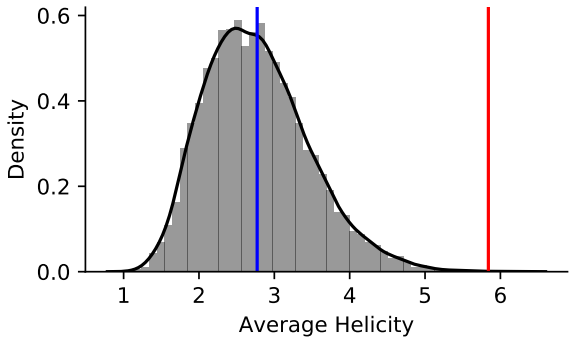

Figure 7.7: Distribution of average helicity of 10,000 subsets from random dataset. Blue line shows the mean of the distribution, red line shows where polyQ dataset loacated in the distribution.

Permutation test

The number of fragments in the polyQ dataset (153) is much smaller than the number of fragments in the random dataset (∼20,000). Therefore a direct comparison of these two datasets will be misleading. To solve this problem, we used the permutation test. The permutation test is a statistical method to measure statistical significance.

First, we generated 10,000 subsets from the random dataset. Each subset contains randomly selected 153 fragments. Then we calculated the average helical propensities for each fragment and each subset. From the mean of the subsets, we generated a distribution and checked where the polyQ dataset located in this distribution. P-value is simply calculated by the formula:

$$P - val = N_{(PolyQ>subset)}/N_{Total} \tag{7.1}$$

where N_{Total} = 10,000.

In previous chapters, we have shown that Agadir underestimates the helicity of polyQ tracts and the four residues preceding the tract nucleates the polyQ helix. Therefore, we examined the helical propensities of the N-terminal of the polyQ tracts. As can be seen from the Figure 7.7, polyQ dataset has remarkably high predicted helical propensity with p-val=2.10^{-4} compared to the random subsets generated from the human proteome. Among 10,000 subsets, only a few of them predicted to be as helical as the polyQ dataset.

8 Methods

8.1 Molecular dynamics simulations

Input coordinates were generated using MacPyMOL[219] in fully helical conformations. All simulations were performed in MD simulation software ACEMD[157] using the CHARMM22* force field[158], which was designed to have an accurate helix-coil balance (See Appendix for all of the parameters used). Each system was explicitly solvated in the TIP3P water model inside cubic boxes from 25 to 40 $\mathring{A}$ distance around the peptides, depending on their length, and neutralized with Cl- and Na+ ions. Initial conformations were minimized and equilibrated under NPT conditions at 1 atm and 300K for 1 ns. Production simulations were performed at 300K in the NVT ensemble using a 4 fs time step for up to 5 μs.

8.2 QM/MM calculations

As a starting structure, a frame from the 5 μs L_4Q_4 simulation that contains a $sc_{Q1} \rightarrow mc_{L1}$ hydrogen bond formed was selected. All the water molecules and ions were preserved from the classical MD simulation. For the QM/MM simulation, we used the AMBER 16 program [220] interfaced to the Terachem 1.9 program (www.petachem.com, access date: June 1, 2017). QM atoms were described at the BLYP/6-31G* level, including a dispersion correction [221]. The Chamber

keyword of the Parmed program from AMBERTOOLS 16[220] is used to describe the classical subsystem with the CHARMM22* force field[158]. To saturate the valence of the frontier atoms, we used the link atoms procedure as implemented in the AMBER package. We used the electrostatic cutoff of 12 $\mathring{A}$ to apply periodic boundary conditions. We first minimized and the equilibrated the structure for 10 ps in the QM/MM run. Production simulations were performed at 300 K for 150 ps with a 1 fs time step. The NBO 6.0 program [222] was used to perform the Natural Bond Critical Point analysis[223, 224].

8.3 Circular dichroism (CD)

CD is a very fast technique to investigate secondary structure, folding and binding properties of proteins. CD is an absorption spectroscopy method that takes advantage of difference in absorption of clockwise (R) and anti-clockwise (L) polarized light in an optically active sample. Amino acids have distinct absorption patterns of L and R polarized light and proteins analyzed by CD will reveal information on their structure[225].

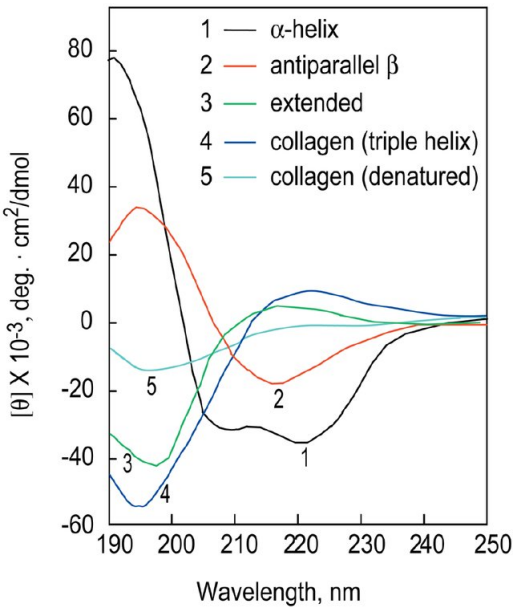

Figure 8.1: Reference CD spectra of protein secondary structures. CD spectra of poly-L-lysine at pH 11.1 in the a-helical (1, black) and antiparallel b-sheet conformations (2, red) and at pH 5.7 in the extended conformations (3, green) and placental collagen in its native triple-helical (4, blue) and denatured (5, cyan) forms taken from[226].

While monitoring the window of wavelengths from 260 nm to 320 nm gives information about the tertiary structure, the window of low wavelengths from 180 nm to 260 nm is used to get information about the secondary structure[227]. Between these wavelengths, each secondary structure type has a distinct CD signal. As seen from the Figure 8.1, typically an alpha-helix has two minima at 208 nm and 220 nm with a maximum at 192 nm, while random coil shows a minimum around 200 nm[226].

Deconvolution of CD spectra to determine secondary structure propensities was performed with the analysis program CONTIN by using reference set 7 hosted at DichroWeb[228–230] (dichroweb.cryst.bbk.ac.uk). DICHROWEB fits the provided CD spectrum to the dataset that contains CD spectra of known secondary structures and calculates the composition of α-helix, β-sheet, and random coils.

8.4 Chemical shift back calculations and Reweighting

We used chemical shift predictor PPM[231] for the back-calculation of the chemical shifts from the trajectories. Frames of the MD ensembles stripped from water molecules and saved as PDB files. These PDB files were used as input in PPM to predict chemical shifts. MD ensembles were reweighted by using the BME method[159] (the code is available at https://github.com/KULL-Centre/ BME).

9 Discussion

9.1 $sc_i \rightarrow mc_{i-4}$ hydrogen bonds

Polyglutamine (polyQ) tracts are low sequence complexity regions present in around 150 human proteins, mostly in the context of intrinsically disordered regions (IDRs). Frequently they are found in the activation domains of transcription factors and transcriptional co-regulators[49]. Due to the problems arising during genome replication, the length of polyQ tracts varies, often resulting in their expansion[51]. These expansions lead to the oligomerization and aggregation of the protein. In nine proteins, polyQ tract expansion beyond a threshold that is specific for each protein causes neurodegenerative polyQ diseases. Besides the thresholds, the risk of having the disease or age of onset, the aggregation propensity of the protein is correlated with the length of the polyQ tract. All these observations hint conformational change in the protein due to the length of the tract.

Recently we have shown that the polyQ tract of AR adopts a helical configuration induced by its N-terminal flanking sequence[154]. In this project, we proved that the overall helical content of the tract positively correlates with its length, with helicity gradually increasing at the N-terminus upon tract elongation, suggesting a cooperative effect in folding.

With our detailed analysis, by combining the simulations with experiments, we found that unconventional $sc_i \rightarrow mc_{i-4}$ hydrogen bonds between the $H\epsilon_{21}$ in the

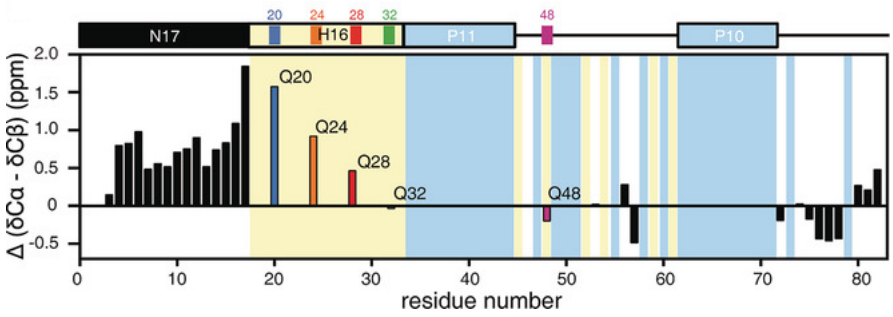

Figure 9.1: Secondary chemical shift analysis on Huntingtin protein using experimental chemical shifts, Q residues are highlighted in yellow taken from:[233].

Gln side chain carboxamide and the oxygen atom of the backbone carbonyl of the residue at relative position $i-4$ can stabilize the α-helices of the AR polyQ tract. We observed that the helicity profile of the polyQ tract is heterogeneous. Tract has the highest helicity at the N-terminal, and helicity decreases gradually towards the C-terminal. This can be explained by the observation that the strength of the $sc_i \rightarrow mc_{i-4}$ hydrogen bond depends on the amino acid type of the acceptor. Leu residues are better acceptors than the Gln residues. Presumably, its geometry, bulkiness, and hydrophobicity helps Leu to shield the hydrogen bond and protects the bond from water competition. Also, while Leu residues can only be acceptors, Gln can act as both an acceptor and a donor. When a Gln at the position i donates its side chain carboxamide $H\epsilon_{21}$ to the residue at the position *i-4*, its backbone oxygen atom is exposed to the water. Therefore, it could also affect the $sc_i \rightarrow mc_{i-4}$ hydrogen bond formation negatively and makes it easier to form a hydrogen bond with a water molecule.

Fiumara and co-workers showed that Leu insertions inside the polyQ tracts increases the helicity and the coiled-coil character of the tract. Our results give a structural interpretation of their results, by explaining how residues with a high propensity to form a $sc_i \rightarrow mc_{i-4}$ hydrogen bond will increase the helicity and if not present, helicity will decay through the C-terminal of the polyQ tract. The fact that different amino acids have different propensities to form $sc_i \rightarrow mc_{i-4}$ hydrogen bonds helps to explain the different properties of the polyQ tracts reported in the literature[154, 232–234].

Huntington's disease is one of the thoroughly studied polyQ disorders, and Baias [232] and Urbanek [233] reported that the polyQ tract of the huntingtin protein adopts helical conformation (Figure 9.1). Besides huntingtin, it has been shown that ataxin 7 also has a helical polyQ tract (Figure 9.2)[234]. The helicity

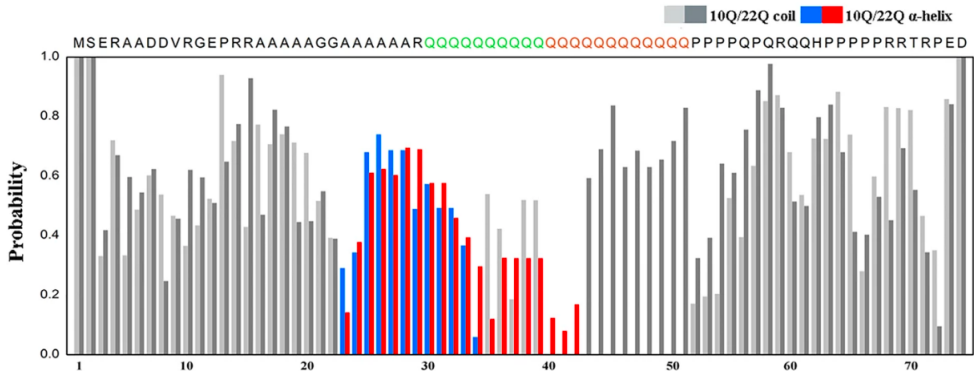

Figure 9.2: Secondary structure probabilities derived from structural ensemble analysis for $Ataxin7_{10Q}$ (blue) and $Ataxin7_{22Q}$ (red)[234].

profile of the polyQ tract of both proteins is similar to the AR polyQ tract. They both have the highest helicity at the N-terminal and beyond helicity decreases through the C-terminal. However, the maximum helicity of the polyQ tract of both proteins is lower than what we observed for AR. We can again explain these behaviors with the $sc_i \rightarrow mc_{i-4}$ forming propensity of the amino acids. Even though both of the proteins have polyQ tracts that are long enough to account for a cooperatively folded helix, their N-terminal compositions are different. Since four residues preceding the polyQ tract forms the first turn of the polyQ helix, they define the helical propensity of the tract. The ability to accept a $sc_i \rightarrow mc_{i-4}$ hydrogen bond of each amino acid has not been determined yet, and having only one Leu residue in the huntingtin protein (LKSF-Q_n) may explain the lower helical content of the tract. Ataxin 7 (AAAR-Q_n), has three Ala N-terminal to the polyQ tract. Despite the fact that Ala has the highest intrinsic helical propensity among all amino acids, the polyQ tract of the ataxin 7 also shows lower helical content than AR[134]. Compared to Leu, Ala is less bulky and hydrophobic, therefore, in light of our results, we can explain the low helical content of ataxin 7 with the lower propensity of Ala residue to accept a $sc_i \rightarrow mc_{i-4}$ hydrogen bond.

Even if the helicity in aqueous solution at physiological temperature is certainly low, it might be enhanced in vivo by the localization of polyQ tract-containing proteins in water depleted environments such as liquid-liquid phase separated protein condensates. PolyQ tracts potentially play a role in such phenomena [235–237]. The absence of the water may lead the unsatisfied hydrogens in the Gln side chains to form the $sc_i \rightarrow mc_{i-4}$ hydrogen bond with the backbone oxygens of the proteins.

9.2 Simulations and reweighting

PolyQ peptides outside of their native sequence context are intrinsically disordered. The goal of our work is to show that, despite this, they can become helical when they are flanked by sequences that can accept up to four of the $sc_i \rightarrow mc_{i-4}$ hydrogen bonds that we reveal in this work involving, as donors, the first residues of the polyQ tract. Yet, in classical MD simulations that we have performed, calculated helicity of the polyQ tract was not as high as experiments.

Since hydrogen bonds play an important role, it is essential to correctly simulate hydrogen bond geometry[238–242]. However, current force fields ignore the partial charges and the directionality of the lone pairs of the acceptor. The simplified description of the hydrogen bond in the force fields may explain the low population of the $sc_i \rightarrow mc_{i-4}$ hydrogen bonds hence the low population of the helical propensity when compared to the experiments. Although QM/MM calculations can help to describe the Hydrogen bond geometry accurately[243], the computational cost makes it impractical to reach convergence. For this reason, we did not compare the frequencies of hydrogen bond formation in the respective simulations. Instead, the goals of these simulations were to qualitatively assess the kinetic stability of the $sc_i \rightarrow mc_{i-4}$ hydrogen bond in the QM/MM force field and, especially, to compute the electron density.

It is also important to mention that even though NMR experiments were performed at 278K, we carried out our MD simulations at 300K. We did not expect the back-calculated chemical shifts computed from the MD trajectories to agree correctly with those measured experimentally because the two procedures are carried out at different temperatures. However, the aim of the simulations was to sample the conformational space to an extent that it allows the reweighting procedure to yield valid conformational ensembles for further analysis. The higher temperature allows a faster and more homogeneous sampling of all configurations and is less sensitive to unbalances in the force field. Still, to check the effect of the temperature on the results, we have carried out a 3 μs simulation of the peptide L_4Q_{16} equivalent to that presented in the Chapter 3 but at the temperature used in the NMR experiments, 278K. As shown in Figure 9.3a the sampling at this temperature is reduced but after reweighting this trajectory on the basis of the main chain chemical shifts obtained at 278K the residue-specific secondary structure is very similar (Figure 9.3b) and, most importantly for our work, the frequencies of the various types of hydrogen bonds involving Gln side chains is also similar to that obtained after reweighting the trajectory obtained at 300K (Figure 9.3c). This indicates that the approach we have used to produce the conformational ensembles

is robust to the temperature used for the MD simulations.

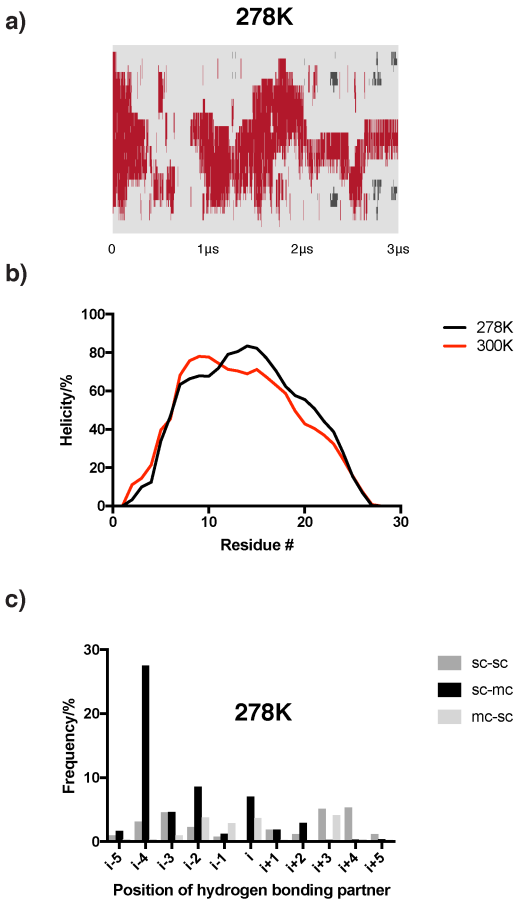

Figure 9.3: a. Time series of the residue-specific secondary structure of peptide L_4Q_{16} in a simulation carried out at 278K as obtained by using the algorithm DSSP where helical residues are shown in red, extended residues are shown in dark gray and disordered residues are shown in light gray b. Comparison of the residue-specific secondary structure obtained by reweighting the MD trajectory of L_4Q_{16} shown in panel a with the main chain chemical shifts obtained at 278K with that obtained by reweighting the trajectory used in the Chapter 3, obtained at 300K c. Frequencies of the various types of hydrogen bonds involving Gln side chains by reweighting the MD trajectory obtained for L_4Q_{16} at 278K. In this figure sc-sc refers to hydrogen bonds between Gln side chains, where one acts as acceptor and the other one as donor; sc-mc refers to hydrogen bond where the Gln side chain acts as donor; and mc-sc to hydrogen bonds where the Gln side chain acts as acceptor.

Finally, regarding convergence, we would like to clarify that, provided that all states of the peptide are sampled, convergence is not necessary to produce

valid conformational ensembles by reweighting. To illustrate that we have divided into two halves the MD trajectory obtained for the largest system, L_4Q_{20}, repeated in each case the reweighting procedure and analyzed the secondary structure (Figure 9.4a) and hydrogen bonds involving Gln side chains (Figure 9.4b) of the resulting reweighted trajectories. As the results illustrate the results are very similar, indicating that the degree of sampling in each half of the original simulation is sufficient to obtain consistent results. In conclusion, reweighting is one of the cheapest solutions to the force field related and convergence problems. It enabled us to obtain a robust model that fits the experimental data to analyze our hypothesis in detail.

9.3 Agadir

Agadir is an algorithm based on the LR-based helix-coil transition theory to predict helical behavior. Still, it underestimates the helicity of the polyQ tract of AR. In our work, we showed that introducing a term that defines the energy contribution, ${\Delta E^L}_{i+4,i}$, of the $sc_i \rightarrow mc_{i-4}$ hydrogen bond between Gln and Leu improved the results for the AR polyQ tract. However, after the optimization of ${\Delta E^X}_{i+4,i}$ for different peptide lengths, we observed that E changes along with the Gln number of the tract. This result suggests that the formation of the polyQ helix is cooperative and having one $sc_i \rightarrow mc_{i-4}$ hydrogen bond makes it easier to form for the second one.

Fixing Agadir requires the optimizing ${\Delta E^X}_{i+4,i}$ values for each type of amino acid and introducing the cooperativity to the algorithm. Each amino acid has a different intrinsic helical propensity, solvent-accessible solvent area (SASA) and hydrophobicity or charge according to their sidechain geometry. Thereby, all these characteristics will have an effect on the ${\Delta E^X}_{i+4,i}$. AR polyQ tract is one of the most convenient systems to calculate the value of ${\Delta E^X}_{i+4,i}$, since it has four Leu residues. Four residues form exactly one turn of a helix, having four of the same amino acid at the N-terminal of the tract amplifies the effect of the amino acid. Right now, our colleagues are working on point mutations in the first position preceding the polyQ tract but the effect may not be as clear to optimize the values. Also, to introduce the cooperativity to the algorithm, first we need to prove our hypothesis and solve the mechanism of cooperativity.

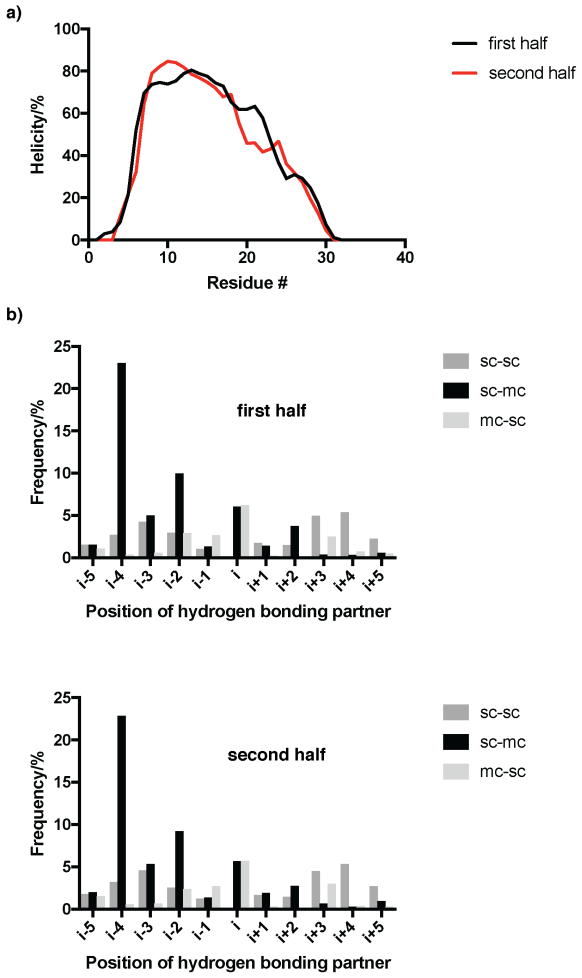

Figure 9.4: a. Comparison of the residue-specific secondary structure obtained by reweighting the first (black) and second (red) halves of the MD trajectory of peptide L_4Q_{20} b. Comparison of the frequencies of the various types of hydrogen bonds involving Gln side chains by reweighting the first (top) and second (bottom) halves of the MD trajectory obtained for L_4Q_{16}. In this figure sc-sc refers to hydrogen bonds between Gln side chains, where one acts as acceptor and the other one as donor; sc-mc refers to hydrogen bond where the Gln side chain acts as donor; and mc-sc to hydrogen bonds where the Gln side chain acts as acceptor.

10
Conclusions

- The stability of AR polyQ helix is positively correlated with its length, and the helical propensity of the residues gradually increases at the N-terminus upon tract elongation.
- Side chain to main chain hydrogen bonds between the H$\epsilon_2$1 of the carboxamide group in the side chain of glutamine at position i and the mainchain carbonyl group of residue at position i-4 ($sc_i \rightarrow mc_{i-4}$)stabilizes the polyQ helix.
- Gln can simultaneously donate hydrogen both from its side chain and main chain to the CO of the residue at relative position *i-4*. These types of hydrogen bonds called bifurcated hydrogen bonds.
- According to QM calculations, the Gln side chain is a better donor than the main chain and the cumulative interaction in the bifurcated hydrogen bond is strong.
- Agadir underestimates the helicity of the AR polyQ tract. Introducing $sc_i \rightarrow mc_{i-4}$ hydrogen bonds to the Agadir improves the results.
- The formation of the polyQ helix is cooperative. The effective strength of the $sc_i \rightarrow mc_{i-4}$ hydrogen bond correlated with the positions and the total number of bonds.

- Bioinformatics analysis shows that the four residues preceding the polyQ tracts in the human proteome enriched in Leu and they are predicted to be more helical when compared to random sets.

11 Appendix

11.1 AR Sequence

```
        10         20         30         40         50         60
MEVQLGLGRV YPRPPSKTYR GAFQNLFQSV REVIQNPGPR HPEAASAAPP GASLLLLQQQ

        70         80         90        100        110        120
QQQQQQQQQQ QQQQQQQQET SPRQQQQQQG EDGSPQAHRR GPTGYLVLDE EQQPSQPQSA

       130        140        150        160        170        180
LECHPERGCV PEPGAAVAAS KGLPQQLPAP PDEDDSAAPS TLSLLGPTFP GLSSCSADLK

       190        200        210        220        230        240
DILSEASTMQ LLQQQQQEAV SEGSSSGRAR EASGAPTSSK DNYLGGTSTI SDNAKELCKA

       250        260        270        280        290        300
VSVSMGLGVE ALEHLSPGEQ LRGDCMYAPL LGVPPAVRPT PCAPLAECKG SLLDDSAGKS

       310        320        330        340        350        360
TEDTAEYSPF KGGYTKGLEG ESLGCSGSAA AGSSGTLELP STLSLYKSGA LDEAAAYQSR

       370        380        390        400        410        420
DYYNFPLALA GPPPPPPPPH PHARIKLENP LDYGSAWAAA AAQCRYGDLA SLHGAGAAGP

       430        440        450        460        470        480
GSGSPSAAAS SSWHTLFTAE EGQLYGPCGG GGGGGGGGGG GGGGGGGGGG GGEAGAVAPY

       490        500        510        520        530        540
GYTRPPQGLA GQESDFTAPD VWYPGGMVSR VPYPSPTCVK SEMGPWMDSY SGPYGDMRLE

       550        560        570        580        590        600
TARDHVLPID YYFPPQKTCL ICGDEASGCH YGALTCGSCK VFFKRAAEGK QKYLCASRND

       610        620        630        640        650        660
CTIDKFRRKN CPSCRLRKCY EAGMTLGARK LKKLGNLKLQ EEGEASSTTS PTEETTQKLT

       670        680        690        700        710        720
VSHIEGYECQ PIFLNVLEAI EPGVVCAGHD NNQPDSFAAL LSSLNELGER QLVHVVKWAK

       730        740        750        760        770        780
ALPGFRNLHV DDQMAVIQYS WMGLMVFAMG WRSFTNVNSR MLYFAPDLVF NEYRMHKSRM

       790        800        810        820        830        840
YSQCVRMRHL SQEFGWLQIT PQEFLCMKAL LLFSIIPVDG LKNQKFFDEL RMNYIKELDR

       850        860        870        880        890        900
IIACKRKNPT SCSRRFYQLT KLLDSVQPIA RELHQFTFDL LIKSHMVSVD FPEMMAEIIS

       910
VQVPKILSGK VKPIYFHTQ
```

polyQ
DBD
Hinge
LBD

11.2 Backbone chemical shifts of L_4Q_4

#	AA	H	N	C	CA	CB
49	Lys	0.0	0.0	0.0	0.0	0.0
50	Lys	0.0	0.0	0.0	0.0	0.0
51	Pro	0.0	0.0	0.0	0.0	0.0
52	Gly	8.68804	109.9799	0.0	0.0	0.0
53	Ala	8.33197	123.9989	0.0	0.0	0.0
54	Ser	8.51705	115.26975	175.20	58.74703	63.36192
55	Leu	8.35150	124.6916	177.88	56.07240	41.87189
56	Leu	8.10257	121.61542	177.87	55.88152	41.86799
57	Leu	8.04165	121.90772	178.01	55.83973	41.91361
58	Leu	8.11173	121.85241	178.17	55.87823	42.05869
59	Glu	8.33104	120.06676	176.77	56.52391	28.97372
60	Gln	8.37547	120.68263	176.46	56.29909	29.19332
61	Gln	8.41722	120.86788	176.14	56.07039	29.38276
62	Gln	8.42887	121.59968	175.82	55.85014	29.39399
63	Lys	8.45195	112.03120	0.0	0.0	0.0
64	Lys	8.18612	117.06059	0.0	0.0	0.0

11.3 Backbone chemical shifts of L_4Q_8

#	AA	H	N	C	CA	CB
49	Lys	0.0	0.0	0.0	0.0	0.0
50	Lys	0.0	0.0	0.0	0.0	0.0
51	Pro	0.0	0.0	0.0	0.0	0.0
52	Gly	0.0	0.0	0.0	0.0	0.0
53	Ala	8.33981	123.9700	0.0	53.23899	19.34415
54	Ser	8.52467	115.24488	175.32405	58.93018	63.31054
55	Leu	8.32830	124.6700	178.13466	56.31548	41.87122
56	Leu	8.07549	121.22485	178.19657	56.30884	41.87791
57	Leu	7.98996	121.49885	178.37086	56.30884	41.87122
58	Leu	8.07122	121.41161	178.59465	56.52222	41.88437
59	Glu	8.30788	119.74579	177.43975	57.18406	28.77146
60	Gln	8.35149	120.42568	177.32777	57.18564	28.96234
61	Gln	8.41158	120.43608	177.17847	57.01944	28.96711
62	Gln	8.39531	120.43344	176.94713	56.75690	28.97072
63	Gln	8.36342	120.56965	176.71627	56.54933	29.17011
64	Gln	8.38387	120.72348	176.39577	56.30800	29.19043
65	Gln	8.40580	120.97214	176.09276	56.07174	29.40070
66	Gln	8.44401	121.71199	175.82115	55.86555	29.41073
67	Lys	8.47072	124.1100	0.0	56.51940	33.01324
68	Lys	8.19155	105.0700	0.0	0.0	0.0

11.4 Backbone chemical shifts of L_4Q_{12}

#	AA	H	N	C	CA	CB
49	Lys	0.0	0.0	0.0	0.0	0.0
50	Lys	0.0	0.0	0.0	0.0	0.0
51	Pro	0.0	0.0	0.0	0.0	0.0
52	Gly	8.67893	121.84785	0.0	0.0	0.0
53	Ala	8.34462	123.9600	0.0	0.0	0.0
54	Ser	8.52441	115.25812	175.39514	58.99273	63.29034
55	Leu	8.32475	124.7000	178.23859	56.52503	41.79823
56	Leu	8.06876	121.07664	178.33371	56.52268	41.87229
57	Leu	7.96937	121.34651	178.56589	56.52905	41.77881
58	Leu	8.05937	121.25263	178.81874	56.74576	41.76248
59	Glu	8.30923	119.67024	177.75455	57.46137	28.72469
60	Gln	8.35082	120.43685	177.68664	57.61583	28.77017
61	Gln	8.43305	120.44370	177.62237	57.50406	28.74594
62	Gln	8.40209	120.40231	177.38821	57.34006	28.73897
63	Gln	8.34600	120.46795	177.20172	57.33535	28.74372
64	Gln	8.34800	120.46752	177.08730	57.17795	28.77306
65	Gln	8.35730	120.55844	176.99493	56.92953	28.87817
66	Gln	8.37600	120.59200	176.87233	56.74359	28.97309
67	Gln	8.40217	120.70270	176.65591	56.53346	29.07394
68	Gln	8.41360	120.84774	176.35721	56.23760	29.18904
69	Gln	8.43011	121.07155	176.07589	55.99267	29.32309
70	Gln	8.46304	121.78703	175.83552	55.85918	29.40708
71	Lys	8.48099	124.6000	175.63987	0.0	0.0
72	Lys	8.19478	105.1200	0.0	0.0	0.0

11.5 Backbone chemical shifts of L_4Q_{16}

#	AA	H	N	C	CA	CB
49	Lys	0.0	0.0	0.0	0.0	0.0
50	Lys	0.0	0.0	0.0	0.0	0.0
51	Pro	0.0	0.0	0.0	0.0	0.0
52	Gly	8.68659	121.89359	174.17642	45.11139	0.0
53	Ala	8.34493	123.9500	178.66270	0.0	0.0
54	Ser	8.52863	115.22656	175.43275	59.03091	63.26770
55	Leu	8.30813	124.6830	178.34616	56.64598	41.76242
56	Leu	8.05183	120.91999	178.48562	56.67580	41.78212
57	Leu	7.94433	121.17432	178.69000	56.70043	41.75799
58	Leu	8.03889	121.09633	178.98269	56.93778	41.79497
59	Glu	8.30071	119.56306	178.02803	57.94841	28.66866
60	Gln	8.34408	120.41775	177.94517	57.84877	28.70385
61	Gln	8.43521	120.41793	177.89145	57.74899	28.66866
62	Gln	8.40107	120.34850	177.68945	57.68478	28.66866
63	Gln	8.32674	120.41201	177.50354	57.62612	28.66450
64	Gln	8.32989	120.41201	177.47247	57.52445	28.66866
65	Gln	8.33775	120.41201	177.37262	57.34848	28.73318
66	Gln	8.34875	120.44555	177.29496	57.23997	28.78923
67	Gln	8.36248	120.50137	177.18956	57.14839	28.82249
68	Gln	8.37209	120.55488	177.08416	57.03958	28.82665
69	Gln	8.38463	120.62177	176.95102	56.85782	28.85991
70	Gln	8.39425	120.66525	176.84007	56.70421	28.94723
71	Gln	8.40653	120.74652	176.64843	56.45668	29.11797
72	Gln	8.42507	120.90956	176.35713	56.21869	29.15884
73	Gln	8.43829	121.12643	176.06968	55.99278	29.35945
74	Gln	8.46860	121.82459	175.82213	55.85581	29.41104
75	Lys	8.48305	124.1700	175.64567	56.45094	32.93116
76	Lys	8.19281	105.0900	0.0	0.0	0.0

11.6 Backbone chemical shifts of L_4Q_{20}

#	AA	H	N	C	CA	CB
49	Lys	0.0	0.0	0.0	0.0	0.0
50	Lys	0.0	0.0	0.0	0.0	0.0
51	Pro	0.0	0.0	0.0	0.0	0.0
52	Gly	0.0	0.0	0.0	0.0	0.0
53	Ala	0.0	0.0	0.0	0.0	0.0
54	Ser	8.53659	115.22902	175.46075	59.08105	63.29383
55	Leu	8.30619	124.68	178.39997	56.70177	41.77524
56	Leu	8.05084	120.83257	178.56275	56.80352	41.87274
57	Leu	7.93784	121.0876	178.75536	56.8665	41.6907
58	Leu	8.03460	121.00796	179.06464	57.10013	41.79801
59	Glu	8.30431	119.51435	178.15838	58.10212	28.60899
60	Gln	8.34776	120.38054	178.08956	58.02622	28.62014
61	Gln	8.44584	120.40928	178.05492	57.96645	28.5458
62	Gln	8.41007	120.33089	177.86569	57.89523	28.55016
63	Gln	8.32674	120.39076	177.71224	57.80495	28.56512
64	Gln	8.32455	120.32389	177.66902	57.74325	28.59095
65	Gln	8.33023	120.39319	177.61031	57.63781	28.69308
66	Gln	8.34350	120.42784	177.55895	57.53978	28.74013
67	Gln	8.35459	120.42784	177.51125	57.47884	28.7507
68	Gln	8.36055	120.48847	177.36512	57.33999	28.81308
69	Gln	8.37307	120.53684	177.33153	57.23037	28.84267
70	Gln	8.37645	120.56391	177.25176	57.1573	28.93298
71	Gln	8.38263	120.60993	177.151	57.01114	28.84162
72	Gln	8.39109	120.60993	177.0409	56.95999	28.95709
73	Gln	8.39842	120.66858	176.90749	56.80653	28.98143
74	Gln	8.41111	120.76558	176.78994	56.60922	28.98006
75	Gln	8.42659	120.82255	176.6304	56.41191	29.14513
76	Gln	8.43846	120.95401	176.35751	56.25571	29.2217
77	Gln	8.45078	121.17525	176.07202	55.9634	29.39091
78	Gln	8.47913	121.85921	175.8181	55.84465	29.41107
79	Lys	0.0	0.0	0.0	0.0	0.0
80	Lys	0.0	0.0	0.0	0.0	0.0

11.7 MD parameters for the minimization and equilibration

```
# VARIABLES DECLARATION

set numMin 5000 ; # Number of steps to minimize
set numNVE 25000 ; # Number of steps for NVE
set numNPT 500000 ; # Number of steps for NPT
set numSteps [expr $numNVE + $numNPT] ; # Total number of steps for the simulation.
set inputname input
set outputname equ
set structure structure
set parameters par_all22star_prot.inp
set temperature 300
set logfreq 1000

# SIMULATION SETTINGS
restart on
restartfreq 5000
restartname $outputname.restart

structure $structure.psf
coordinates $structure.pdb
parameters $parameters

temperature $temperature

#celldimension
extendedsystem $inputname.xsc

timestep 4
rigidbonds all
hydrogenscale 4

switching on
switchdist 7.5
cutoff 9
```

```
exclude scaled1-4
1-4scaling 1.0
fullelectfrequency 2

langevin on
langevintemp $temperature
langevindamping 1

outputname $outputname
dcdfreq 25000
dcdfile $outputname.dcd

pme on
pmegridspacing 1.0

constraints on
consref structure.restrained.pdb
constraintscaling 1.0

berendsenpressure on
berendsenpressuretarget 1.01325
berendsenpressurerelaxationtime 800

energyfreq $logfreq

# TCL FORCES SETTINGS
tclforces on
set cons_off 30000 ; # Turn off constraints after X many steps

proc calcforces_init {} {
# Declaration of this procedure is required, even
# if you don't need to initialize atom groups, etc.
}

proc calcforces {} {
global numNVE cons_off logfreq
# Get Current Step
set step [getstep]
```

```
# Control switch from NVE to NPT
if {$step > $numNVE && $step%$logfreq == 0} {
#puts ‘Turning on barostat’
berendsenpressure on
} else {
if {$step == 0} {
#puts ”Barostat off, step $step”
berendsenpressure off
}
}

    # Turn off restraints
if {$step == $cons_off} {
puts ‘# Constraints set to 0, step $step’
constraintscaling 0
}
}

proc calcforces_endstep { } { }

# Run minimization
minimize $numMin
# Run simulation
run $numSteps
```

11.8 MD parameters for the production run

```
# VARIABLES DECLARATION

set inputname run.restart
set outputname run
set structure structure
set parameters par_all22star_prot.inp
set temperature 300
set logfreq 1000

# SIMULATION SETTINGS
restart on
restartfreq 25000
restartname $outputname.restart
wrap on

structure $structure.psf
coordinates $structure.pdb
bincoordinates $inputname.coor
binvelocities $inputname.vel
parameters $parameters

temperature $temperature

#celldimension
extendedsystem $inputname.xsc

timestep 4
rigidbonds all
hydrogenscale 4

switching on
switchdist 7.5
cutoff 9
exclude scaled1-4
1-4scaling 1.0
fullelectfrequency 2
```

```
langevin on
langevintemp $temperature
langevindamping 0.1

# Configure barostat
berendsenpressure off
berendsenpressuretarget 1.01325
berendsenpressurerelaxationtime 800
useflexiblecell off
useconstantratio off

outputname $outputname
energyfreq 5000
xtcfreq 25000
xtcfile $outputname.xtc

pme on
pmegridspacing 1.0

energyfreq $logfreq

# Run simulation
run 5000ns
```

11.9 Replica Averaged MD parameters for the equilibration run

```
; VARIOUS PREPROCESSING OPTIONS
; Preprocessor information: use cpp syntax.
; e.g.: -I/home/joe/doe -I/home/mary/roe
include =
; e.g.: -DPOSRES -DFLEXIBLE (note these variable names are case sensitive)
define =

; RUN CONTROL PARAMETERS
integrator = steep
; Start time and timestep in ps
tinit = 0
dt = 0.001
nsteps = 10000
; For exact run continuation or redoing part of a run
init-step = 0
; Part index is updated automatically on checkpointing (keeps files separate)
simulation-part = 1
; mode for center of mass motion removal
comm-mode = Linear
; number of steps for center of mass motion removal
nstcomm = 100
; group(s) for center of mass motion removal
comm-grps =

; LANGEVIN DYNAMICS OPTIONS
; Friction coefficient (amu/ps) and random seed
bd-fric = 0
ld-seed = 1993

; ENERGY MINIMIZATION OPTIONS
; Force tolerance and initial step-size
emtol = 1000
emstep = 0.001
; Max number of iterations in relax-shells
niter = 20
```

; Step size (ps^2) for minimization of flexible constraints
fcstep = 0
; Frequency of steepest descents steps when doing CG
nstcgsteep = 100
nbfgscorr = 10

; TEST PARTICLE INSERTION OPTIONS
rtpi = 0.05

; OUTPUT CONTROL OPTIONS
; Output frequency for coords (x), velocities (v) and forces (f)
nstxout = 0
nstvout = 0
nstfout = 0
; Output frequency for energies to log file and energy file
nstlog = 1000
nstcalcenergy = 100
nstenergy = 1000
; Output frequency and precision for .xtc file
nstxtcout = 0
xtc-precision = 1000
; This selects the subset of atoms for the .xtc file. You can
; select multiple groups. By default all atoms will be written.
xtc-grps =
; Selection of energy groups
energygrps =

; NEIGHBORSEARCHING PARAMETERS
; cut-off scheme (group: using charge groups, Verlet: particle based cut-offs)
cutoff-scheme = Verlet
; nblist update frequency
nstlist = 1
; ns algorithm (simple or grid)
ns-type = grid
; Periodic boundary conditions: xyz, no, xy
pbc = xyz
periodic-molecules = no
; Allowed energy drift due to the Verlet buffer in kJ/mol/ps per atom,

```
; a value of -1 means: use rlist
verlet-buffer-drift = 0.005
; nblist cut-off
rlist = 1
; long-range cut-off for switched potentials
rlistlong = -1
nstcalclr = -1

; OPTIONS FOR ELECTROSTATICS AND VDW
; Method for doing electrostatics
coulombtype = PME
coulomb-modifier = Potential-shift-Verlet
rcoulomb-switch = 0
rcoulomb = 1.0
; Relative dielectric constant for the medium and the reaction field
epsilon-r = 1
epsilon-rf = 0
; Method for doing Van der Waals
vdw-type = Cut-off
vdw-modifier = Potential-shift-Verlet
; cut-off lengths
rvdw-switch = 0
rvdw = 1.0
; Apply long range dispersion corrections for Energy and Pressure
DispCorr = No
; Extension of the potential lookup tables beyond the cut-off
table-extension = 1
; Separate tables between energy group pairs
energygrp-table =
; Spacing for the PME/PPPM FFT grid
fourierspacing = 0.12
; FFT grid size, when a value is 0 fourierspacing will be used
fourier-nx = 0
fourier-ny = 0
fourier-nz = 0
; EWALD/PME/PPPM parameters
pme-order = 4
ewald-rtol = 1e-05
```

ewald-geometry = 3d
epsilon-surface = 0
optimize-fft = no

; IMPLICIT SOLVENT ALGORITHM
implicit-solvent = No

; GENERALIZED BORN ELECTROSTATICS
; Algorithm for calculating Born radii
gb-algorithm = Still
; Frequency of calculating the Born radii inside rlist
nstgbradii = 1
; Cutoff for Born radii calculation; the contribution from atoms
; between rlist and rgbradii is updated every nstlist steps
rgbradii = 1
; Dielectric coefficient of the implicit solvent
gb-epsilon-solvent = 80
; Salt concentration in M for Generalized Born models
gb-saltconc = 0
; Scaling factors used in the OBC GB model. Default values are OBC(II)
gb-obc-alpha = 1
gb-obc-beta = 0.8
gb-obc-gamma = 4.85
gb-dielectric-offset = 0.009
sa-algorithm = Ace-approximation
; Surface tension (kJ/mol/nm^2) for the SA (nonpolar surface) part of GBSA
; The value -1 will set default value for Still/HCT/OBC GB-models.
sa-surface-tension = -1

; OPTIONS FOR WEAK COUPLING ALGORITHMS
; Temperature coupling
tcoupl = No
nsttcouple = -1
nh-chain-length = 10
print-nose-hoover-chain-variables = no
; Groups to couple separately
tc-grps =
; Time constant (ps) and reference temperature (K)

```
tau-t =
ref-t =
; pressure coupling
pcoupl = No
pcoupltype = Isotropic
nstpcouple = -1
; Time constant (ps), compressibility (1/bar) and reference P (bar)
tau-p = 1
compressibility =
ref-p =
; Scaling of reference coordinates, No, All or COM
refcoord-scaling = No

; OPTIONS FOR QMMM calculations
QMMM = no
; Groups treated Quantum Mechanically
QMMM-grps =
; QM method
QMmethod =
; QMMM scheme
QMMMscheme = normal
; QM basisset
QMbasis =
; QM charge
QMcharge =
; QM multiplicity
QMmult =
; Surface Hopping
SH =
; CAS space options
CASorbitals =
CASelectrons =
SAon =
SAoff =
SAsteps =
; Scale factor for MM charges
MMChargeScaleFactor = 1
; Optimization of QM subsystem
```

```
bOPT =
bTS =

; SIMULATED ANNEALING
; Type of annealing for each temperature group (no/single/periodic)
annealing =
; Number of time points to use for specifying annealing in each group
annealing-npoints =
; List of times at the annealing points for each group
annealing-time =
; Temp. at each annealing point, for each group.
annealing-temp =

; GENERATE VELOCITIES FOR STARTUP RUN
gen-vel = no
gen-temp = 300
gen-seed = 173529

; OPTIONS FOR BONDS
constraints = none
; Type of constraint algorithm
constraint-algorithm = Lincs
; Do not constrain the start configuration
continuation = no
; Use successive overrelaxation to reduce the number of shake iterations
Shake-SOR = no
; Relative tolerance of shake
shake-tol = 0.0001
; Highest order in the expansion of the constraint coupling matrix
lincs-order = 4
; Number of iterations in the final step of LINCS. 1 is fine for
; normal simulations, but use 2 to conserve energy in NVE runs.
; For energy minimization with constraints it should be 4 to 8.
lincs-iter = 1
; Lincs will write a warning to the stderr if in one step a bond
; rotates over more degrees than
lincs-warnangle = 30
; Convert harmonic bonds to morse potentials
```

```
morse = no

; ENERGY GROUP EXCLUSIONS
; Pairs of energy groups for which all non-bonded interactions are excluded
energygrp-excl =

; WALLS
; Number of walls, type, atom types, densities and box-z scale factor for Ewald
nwall = 0
wall-type = 9-3
wall-r-linpot = -1
wall-atomtype =
wall-density =
wall-ewald-zfac = 3

; COM PULLING
; Pull type: no, umbrella, constraint or constant-force
pull = no

; ENFORCED ROTATION
; Enforced rotation: No or Yes
rotation = no

; NMR refinement stuff
; Distance restraints type: No, Simple or Ensemble
disre = No
; Force weighting of pairs in one distance restraint: Conservative or Equal
disre-weighting = Conservative
; Use sqrt of the time averaged times the instantaneous violation
disre-mixed = no
disre-fc = 1000
disre-tau = 0
; Output frequency for pair distances to energy file
nstdisreout = 100
; Orientation restraints: No or Yes
orire = no
; Orientation restraints force constant and tau for time averaging
orire-fc = 0
```

orire-tau = 0
orire-fitgrp =
; Output frequency for trace(SD) and S to energy file
nstorireout = 100

; Free energy variables
free-energy = no
couple-moltype =
couple-lambda0 = vdw-q
couple-lambda1 = vdw-q
couple-intramol = no
init-lambda = -1
init-lambda-state = -1
delta-lambda = 0
nstdhdl = 50
fep-lambdas =
mass-lambdas =
coul-lambdas =
vdw-lambdas =
bonded-lambdas =
restraint-lambdas =
temperature-lambdas =
calc-lambda-neighbors = 1
init-lambda-weights =
dhdl-print-energy = no
sc-alpha = 0
sc-power = 1
sc-r-power = 6
sc-sigma = 0.3
sc-coul = no
separate-dhdl-file = yes
dhdl-derivatives = yes
dh-hist-size = 0
dh-hist-spacing = 0.1

; Non-equilibrium MD stuff
acc-grps =
accelerate =

```
freezegrps =
freezedim =
cos-acceleration = 0
deform =

; simulated tempering variables
simulated-tempering = no
simulated-tempering-scaling = geometric
sim-temp-low = 300
sim-temp-high = 300

; Electric fields
; Format is number of terms (int) and for all terms an amplitude (real)
; and a phase angle (real)
E-x =
E-xt =
E-y =
E-yt =
E-z =
E-zt =

; AdResS parameters
adress = no

; User defined thingies
user1-grps =
user2-grps =
userint1 = 0
userint2 = 0
userint3 = 0
userint4 = 0
userreal1 = 0
userreal2 = 0
userreal3 = 0
userreal4 = 0
```

11.10 Replica Averaged MD parameters for the production run

; RUN CONTROL PARAMETERS
integrator = md
; Start time and timestep in ps
tinit = 0
dt = 0.002
nsteps = 500000
; For exact run continuation or redoing part of a run
init-step = 0
; Part index is updated automatically on checkpointing (keeps files separate)
simulation-part = 1
; mode for center of mass motion removal
comm-mode = Linear
; number of steps for center of mass motion removal
nstcomm = 100
; group(s) for center of mass motion removal
comm-grps =

; LANGEVIN DYNAMICS OPTIONS
; Friction coefficient (amu/ps) and random seed
bd-fric = 0
ld-seed = -1

; ENERGY MINIMIZATION OPTIONS
; Force tolerance and initial step-size
emtol = 10
emstep = 0.01
; Max number of iterations in relax-shells
niter = 20
; Step size (ps^2) for minimization of flexible constraints
fcstep = 0
; Frequency of steepest descents steps when doing CG
nstcgsteep = 1000
nbfgscorr = 10

; TEST PARTICLE INSERTION OPTIONS

```
rtpi = 0.05

; OUTPUT CONTROL OPTIONS
; Output frequency for coords (x), velocities (v) and forces (f)
nstxout = 0
nstvout = 5000
nstfout = 0
; Output frequency for energies to log file and energy file
nstlog = 1000
nstcalcenergy = 20
nstenergy = 1000
; Output frequency and precision for .xtc file
nstxout-compressed = 5000
compressed-x-precision = 1000
; This selects the subset of atoms for the compressed
; trajectory file. You can select multiple groups. By
; default, all atoms will be written.
compressed-x-grps =
; Selection of energy groups
energygrps =

; NEIGHBORSEARCHING PARAMETERS
; cut-off scheme (Verlet: particle based cut-offs, group: using charge groups)
cutoff-scheme = Verlet
; nblist update frequency
nstlist = 20
; ns algorithm (simple or grid)
ns-type = grid
; Periodic boundary conditions: xyz, no, xy
pbc = xyz
periodic-molecules = no
; Allowed energy error due to the Verlet buffer in kJ/mol/ps per atom,
; a value of -1 means: use rlist
verlet-buffer-tolerance = 0.005
; nblist cut-off
rlist = 1
; long-range cut-off for switched potentials
rlistlong = -1
```

```
nstcalclr = -1

; OPTIONS FOR ELECTROSTATICS AND VDW
; Method for doing electrostatics
coulombtype = PME
coulomb-modifier = Potential-shift-Verlet
rcoulomb-switch = 0.8
rcoulomb = 1.0
; Relative dielectric constant for the medium and the reaction field
epsilon-r = 1
epsilon-rf = 0
; Method for doing Van der Waals
vdw-type = Cut-off
vdw-modifier = Potential-shift-Verlet
; cut-off lengths
rvdw-switch = 0.8
rvdw = 1.0
; Apply long range dispersion corrections for Energy and Pressure
DispCorr = No
; Extension of the potential lookup tables beyond the cut-off
table-extension = 1
; Separate tables between energy group pairs
energygrp-table =
; Spacing for the PME/PPPM FFT grid
fourierspacing = 0.12
; FFT grid size, when a value is 0 fourierspacing will be used
fourier-nx = 0
fourier-ny = 0
fourier-nz = 0
; EWALD/PME/PPPM parameters
pme-order = 4
ewald-rtol = 1e-05
ewald-rtol-lj = 0.001
lj-pme-comb-rule = Geometric
ewald-geometry = 3d
epsilon-surface = 0

; IMPLICIT SOLVENT ALGORITHM
```

implicit-solvent = No

; GENERALIZED BORN ELECTROSTATICS
; Algorithm for calculating Born radii
gb-algorithm = Still
; Frequency of calculating the Born radii inside rlist
nstgbradii = 1
; Cutoff for Born radii calculation; the contribution from atoms
; between rlist and rgbradii is updated every nstlist steps
rgbradii = 1
; Dielectric coefficient of the implicit solvent
gb-epsilon-solvent = 80
; Salt concentration in M for Generalized Born models
gb-saltconc = 0
; Scaling factors used in the OBC GB model. Default values are OBC(II)
gb-obc-alpha = 1
gb-obc-beta = 0.8
gb-obc-gamma = 4.85
gb-dielectric-offset = 0.009
sa-algorithm = Ace-approximation
; Surface tension (kJ/mol/nm^2) for the SA (nonpolar surface) part of GBSA
; The value -1 will set default value for Still/HCT/OBC GB-models.
sa-surface-tension = -1

; OPTIONS FOR WEAK COUPLING ALGORITHMS
; Temperature coupling
Tcoupl = v-rescale
nsttcouple = 5
nh-chain-length = 10
print-nose-hoover-chain-variables = no
; Groups to couple separately
tc-grps = system
; Time constant (ps) and reference temperature (K)
tau-t = 0.2
ref-t = 278
; pressure coupling
Pcoupl = no
Pcoupltype = Isotropic

```
nstpcouple = -1
; Time constant (ps), compressibility (1/bar) and reference P (bar)
tau-p = 1
compressibility = 4.5e-5
ref-p = 1.0
; Scaling of reference coordinates, No, All or COM
refcoord-scaling = No

; OPTIONS FOR QMMM calculations
QMMM = no
; Groups treated Quantum Mechanically
QMMM-grps =
; QM method
QMmethod =
; QMMM scheme
QMMMscheme = normal
; QM basisset
QMbasis =
; QM charge
QMcharge =
; QM multiplicity
QMmult =
; Surface Hopping
SH =
; CAS space options
CASorbitals =
CASelectrons =
SAon =
SAoff =
SAsteps =
; Scale factor for MM charges
MMChargeScaleFactor = 1
; Optimization of QM subsystemv bOPT =
bTS =

; SIMULATED ANNEALING
; Type of annealing for each temperature group (no/single/periodic)
annealing =
```

; Number of time points to use for specifying annealing in each groupv annealing-npoints =
; List of times at the annealing points for each group
annealing-time =
; Temp. at each annealing point, for each group.
annealing-temp =

; GENERATE VELOCITIES FOR STARTUP RUN
gen-vel = no
gen-temp = 300
gen-seed = -1

; OPTIONS FOR BONDS
constraints = all-bonds
; Type of constraint algorithm
constraint-algorithm = lincs
; Do not constrain the start configuration
continuation = no
; Use successive overrelaxation to reduce the number of shake iterations
Shake-SOR = no
; Relative tolerance of shake
shake-tol = 0.0001
; Highest order in the expansion of the constraint coupling matrix
lincs-order = 4
; Number of iterations in the final step of LINCS. 1 is fine for
; normal simulations, but use 2 to conserve energy in NVE runs.
; For energy minimization with constraints it should be 4 to 8.
lincs-iter = 1
; Lincs will write a warning to the stderr if in one step a bond
; rotates over more degrees than
lincs-warnangle = 30
; Convert harmonic bonds to morse potentials
morse = no

; ENERGY GROUP EXCLUSIONS
; Pairs of energy groups for which all non-bonded interactions are excluded
energygrp-excl =

```
; WALLS
; Number of walls, type, atom types, densities and box-z scale factor for Ewald
nwall = 0
wall-type = 9-3
wall-r-linpot = -1
wall-atomtype =
wall-density =
wall-ewald-zfac = 3

; COM PULLING
pull = no

    ; ENFORCED ROTATION
; Enforced rotation: No or Yes
rotation = no

; Group to display and/or manipulate in interactive MD session
IMD-group =

; NMR refinement stuff
; Distance restraints type: No, Simple or Ensemble
disre = No
; Force weighting of pairs in one distance restraint: Conservative or Equal
disre-weighting = Conservative
; Use sqrt of the time averaged times the instantaneous violation
disre-mixed = no
disre-fc = 1000
disre-tau = 0
; Output frequency for pair distances to energy file
nstdisreout = 100
; Orientation restraints: No or Yes
orire = no
; Orientation restraints force constant and tau for time averaging
orire-fc = 0
orire-tau = 0
orire-fitgrp =
; Output frequency for trace(SD) and S to energy file
nstorireout = 100
```

; Free energy variables
free-energy = no
couple-moltype =
couple-lambda0 = vdw-q
couple-lambda1 = vdw-q
couple-intramol = no
init-lambda = -1
init-lambda-state = -1
delta-lambda = 0
nstdhdl = 50
fep-lambdas =
mass-lambdas =
coul-lambdas =
vdw-lambdas =
bonded-lambdas =
restraint-lambdas =
temperature-lambdas =
calc-lambda-neighbors = 1
init-lambda-weights =
dhdl-print-energy = no
sc-alpha = 0
sc-power = 1
sc-r-power = 6
sc-sigma = 0.3
sc-coul = no
separate-dhdl-file = yesv dhdl-derivatives = yes
dh-hist-size = 0
dh-hist-spacing = 0.1

; Non-equilibrium MD stuff
acc-grps =
accelerate =
freezegrps =
freezedim =
cos-acceleration = 0
deform =

```
; simulated tempering variables
simulated-tempering = no
simulated-tempering-scaling = geometric
sim-temp-low = 300
sim-temp-high = 300

; Electric fields
; Format is number of terms (int) and for all terms an amplitude (real)
; and a phase angle (real)
E-x =
; Time dependent (pulsed) electric field. Format is omega, time for pulse
; peak, and sigma (width) for pulse. Sigma = 0 removes pulse, leaving
; the field to be a cosine function.
E-xt =
E-y =
E-yt =
E-z =
E-zt =

; Ion/water position swapping for computational electrophysiology setups
; Swap positions along direction: no, X, Y, Z
swapcoords = no

; AdResS parameters
adress = no

; User defined thingies
user1-grps =
user2-grps =
userint1 = 0
userint2 = 0
userint3 = 0
userint4 = 0
userreal1 = 0
userreal2 = 0
userreal3 = 0
userreal4 = 0
```

References

[1] P E Wright and H J Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, 293(2):321–331, October 1999.

[2] A K Dunker, J D Lawson, C J Brown, R M Williams, P Romero, J S Oh, C J Oldfield, A M Campen, C M Ratliff, K W Hipps, J Ausio, M S Nissen, R Reeves, C Kang, C R Kissinger, R W Bailey, M D Griswold, W Chiu, E C Garner, and Z Obradovic. Intrinsically disordered protein. *J. Mol. Graph. Model.*, 19(1):26–59, 2001.

[3] Robin van der Lee, Marija Buljan, Benjamin Lang, Robert J Weatheritt, Gary W Daughdrill, A Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T Jones, Philip M Kim, Richard W Kriwacki, Christopher J Oldfield, Rohit V Pappu, Peter Tompa, Vladimir N Uversky, Peter E Wright, and M Madan Babu. Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, 114(13):6589–6631, July 2014.

[4] Emil Fischer. Einfluss der configuration auf die wirkung der enzyme. *Ber. Dtsch. Chem. Ges.*, 27(3):2985–2993, October 1894.

[5] A E Mirsky and L Pauling. On the structure of native, denatured, and coagulated proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 22(7):439–447, July 1936.

[6] Hsien Wu. Studies on denaturation of proteins XIII. a theory of denaturation. *Chin J. Physiol.*, 5(4):321–344, 1931.

[7] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1):235–242, January 2000.

[8] A S Manalan and C B Klee. Activation of calcineurin by limited proteolysis. *Proc. Natl. Acad. Sci. U. S. A.*, 80(14):4291–4295, July 1983.

[9] C R Kissinger, H E Parge, D R Knighton, C T Lewis, L A Pelletier, A Tempczyk, V J Kalish, K D Tucker, R E Showalter, and E W Moomaw. Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex. *Nature*, 378(6557):641–644, December 1995.

[10] David Ball, John Hill, Rhonda J Scott. *The Basics of General, Organic, and Biological Chemistry*. Saylor Foundation, 2011.

[11] Peter Tompa. Intrinsically unstructured proteins. *Trends Biochem. Sci.*, 27(10):527–533, October 2002.

[12] Vladimir N Uversky. What does it mean to be natively unfolded? *Eur. J. Biochem.*, 269:2–12, 2002.

[13] Megan Sickmeier, Justin A Hamilton, Tanguy LeGall, Vladimir Vacic, Marc S Cortese, Agnes Tantos, Beata Szabo, Peter Tompa, Jake Chen, Vladimir N Uversky, Zoran Obradovic, and A Keith Dunker. DisProt: the database of disordered proteins. *Nucleic Acids Res.*, 35(Database issue):D786–93, January 2007.

[14] A Bairoch and R Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, 28(1):45–48, January 2000.

[15] Vladimir Vacic, Vladimir N Uversky, A Keith Dunker, and Stefano Lonardi. Composition profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics*, 8:211, June 2007.

[16] V N Uversky, J R Gillespie, and A L Fink. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, 41(3):415–427, November 2000.

[17] P Romero, Z Obradovic, X Li, E C Garner, C J Brown, and A K Dunker. Sequence complexity of disordered protein. *Proteins*, 42(1):38–48, January 2001.

[18] P Romero, Z Obradovic, and A K Dunker. Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Lett.*, 462(3):363–367, December 1999.

[19] Zsuzsanna Dosztányi, Veronika Csizmok, Peter Tompa, and István Simon. IUPred: web server for the prediction of intrinsically unstructured regions

of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, August 2005.

[20] Rune Linding, Lars Juhl Jensen, Francesca Diella, Peer Bork, Toby J Gibson, and Robert B Russell. Protein disorder prediction: implications for structural proteomics. *Structure*, 11(11):1453–1459, November 2003.

[21] Rune Linding, Robert B Russell, Victor Neduva, and Toby J Gibson. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, 31(13):3701–3708, July 2003.

[22] P Romero, Z Obradovic, C Kissinger, J E Villafranca, and A K Dunker. Identifying disordered regions in proteins from amino acid sequence. In *Proceedings of International Conference on Neural Networks (ICNN'97)*, volume 1, pages 90–95 vol.1, June 1997.

[23] J J Ward, J S Sodhi, L J McGuffin, B F Buxton, and D T Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, 337(3):635–645, March 2004.

[24] Rita Pancsa and Peter Tompa. Structural disorder in eukaryotes. *PLoS One*, 7(4):e34687, April 2012.

[25] J J Ward, J S Sodhi, L J McGuffin, B F Buxton, and D T Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, 337(3):635–645, March 2004.

[26] H J Dyson and P E Wright. Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states. *Methods Enzymol.*, 339:258–270, 2001.

[27] H Jane Dyson and Peter E Wright. Unfolded proteins and protein folding studied by NMR. *Chem. Rev.*, 104(8):3607–3622, August 2004.

[28] David Eliezer. Biophysical characterization of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, 19(1):23–30, February 2009.

[29] D S Wishart, B D Sykes, and F M Richards. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.*, 222(2):311–333, November 1991.

[30] D S Wishart and B D Sykes. Chemical shifts as a tool for structure determination. *Methods Enzymol.*, 239:363–392, 1994.

[31] Gene Merutka, H Jane Dyson, and Peter E Wright. 'random coil' 1H chemical shifts obtained as a function of temperature and trifluoroethanol concentration for the peptide series GGXGG. *J. Biomol. NMR*, 5(1):14–24, January 1995.

[32] Daniel Braun, Gerhard Wider, and Kurt Wuethrich. Sequence-Corrected 15N "random coil" chemical shifts. *J. Am. Chem. Soc.*, 116(19):8466–8469, September 1994.

[33] G D Fasman, editor. *Circular Dichroism and the Conformational Analysis of Biomolecules*. Springer, 1996.

[34] Vladimir N Uversky, Christopher J Oldfield, and A Keith Dunker. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.*, 18(5):343–384, September 2005.

[35] Alan Fersht Peter Tompa. *Structure and Function of Intrinsically Disordered Proteins*. Chapman and Hall/CRC, November 2009.

[36] Vladimir N Uversky. Unusual biophysics of intrinsically disordered proteins. *Biochim. Biophys. Acta*, 1834(5):932–951, May 2013.

[37] R S Spolar and M T Record, Jr. Coupling of local folding to site-specific binding of proteins to DNA. *Science*, 263(5148):777–784, February 1994.

[38] H Jane Dyson and Peter E Wright. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.*, 12(1):54–60, February 2002.

[39] A P Demchenko. Recognition between flexible protein molecules: induced and assisted folding. *J. Mol. Recognit.*, 14(1):42–61, January 2001.

[40] Alessandro Borgia, Madeleine B Borgia, Katrine Bugge, Vera M Kissling, Pétur O Heidarsson, Catarina B Fernandes, Andrea Sottini, Andrea Soranno, Karin J Buholzer, Daniel Nettels, Birthe B Kragelund, Robert B Best, and Benjamin Schuler. Extreme disorder in an ultrahigh-affinity protein complex. *Nature*, 555(7694):61–66, March 2018.

[41] Vladimir N Uversky, Vrushank Davé, Lilia M Iakoucheva, Prerna Malaney, Steven J Metallo, Ravi Ramesh Pathak, and Andreas C Joerger. Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. *Chem. Rev.*, 114(13):6844–6879, July 2014.

[42] Zoran Obradovic, Kang Peng, Slobodan Vucetic, Predrag Radivojac, Celeste J Brown, and A Keith Dunker. Predicting intrinsic disorder from amino acid sequence. *Proteins*, 53 Suppl 6:566–572, 2003.

[43] A K Dunker, Z Obradovic, P Romero, E C Garner, and C J Brown. Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.*, 11:161–171, 2000.

[44] Lilia M Iakoucheva, Celeste J Brown, J David Lawson, Zoran Obradović, and A Keith Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, 323(3):573–584, October 2002.

[45] Lilia M Iakoucheva, Predrag Radivojac, Celeste J Brown, Timothy R O'Connor, Jason G Sikes, Zoran Obradovic, and A Keith Dunker. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, 32(3):1037–1049, February 2004.

[46] Slobodan Vucetic, Celeste J Brown, A Keith Dunker, and Zoran Obradovic. Flavors of protein disorder. *Proteins*, 52(4):573–584, September 2003.

[47] P J Mitchell and R Tjian. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, 245(4916):371–378, July 1989.

[48] Matthieu Legendre, Nathalie Pochet, Theodore Pak, and Kevin J Verstrepen. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.*, 17(12):1787–1796, December 2007.

[49] Rita Gemayel, Marcelo D Vinces, Matthieu Legendre, and Kevin J Verstrepen. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.*, 44:445–477, 2010.

[50] Noel G Faux, Stephen P Bottomley, Arthur M Lesk, James A Irving, John R Morrison, Maria Garcia de la Banda, and James C Whisstock. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res.*, 15(4):537–551, April 2005.

[51] Sergei M Mirkin. Expandable DNA repeats and human disease. *Nature*, 447(7147):932–940, June 2007.

[52] K Nakamura, S Y Jeong, T Uchihara, M Anno, K Nagashima, T Nagashima, S Ikeda, S Tsuji, and I Kanazawa. SCA17, a novel autosomal dominant

cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. *Hum. Mol. Genet.*, 10(14):1441–1448, July 2001.

[53] F A Nahhas, J Garbern, K M Krajewski, B B Roa, and G L Feldman. Juvenile onset huntington disease resulting from a very large maternal expansion. *Am. J. Med. Genet. A*, 137A(3):328–331, September 2005.

[54] T Matilla, V Volpini, D Genís, J Rosell, J Corral, A Dávalos, A Molins, and X Estivill. Presymptomatic analysis of spinocerebellar ataxia type 1 (SCA1) via the expansion of the SCA1 CAG-repeat in a large pedigree displaying anticipation and parental male bias. *Hum. Mol. Genet.*, 2(12):2123–2128, December 1993.

[55] R Koide, S Kobayashi, T Shimohata, T Ikeuchi, M Maruyama, M Saito, M Yamada, H Takahashi, and S Tsuji. A neurological disease caused by an expanded CAG trinucleotide repeat in the TATA-binding protein gene: a new polyglutamine disease? *Hum. Mol. Genet.*, 8(11):2047–2053, October 1999.

[56] O Komure, A Sano, N Nishino, N Yamauchi, S Ueno, K Kondoh, N Sano, M Takahashi, N Murayama, and I Kondo. DNA analysis in hereditary dentatorubral-pallidoluysian atrophy: correlation between CAG repeat length and phenotypic variation and the molecular basis of anticipation. *Neurology*, 45(1):143–149, January 1995.

[57] C S Benton, R de Silva, S L Rutledge, S Bohlega, T Ashizawa, and H Y Zoghbi. Molecular and clinical studies in SCA-7 define a broad clinical spectrum and the infantile phenotype. *Neurology*, 51(4):1081–1086, October 1998.

[58] Y Kawaguchi, T Okamoto, M Taniwaki, M Aizawa, M Inoue, S Katayama, H Kawakami, S Nakamura, M Nishimura, and I Akiguchi. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat. Genet.*, 8(3):221–228, November 1994.

[59] Peter S Harper Marcy E. MacDonald. A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. the huntington's disease collaborative research group. *Cell*, 72(6):971–983, March 1993.

[60] H T Orr, M Y Chung, S Banfi, T J Kwiatkowski, Jr, A Servadio, A L Beaudet, A E McCall, L A Duvick, L P Ranum, and H Y Zoghbi. Expansion

of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat. Genet.*, 4(3):221–226, July 1993.

[61] Steffen J Sahl, Lucien E Weiss, Whitney C Duim, Judith Frydman, and W E Moerner. Cellular inclusion bodies of mutant huntingtin exon 1 obscure small fibrillar aggregate species. *Sci. Rep.*, 2:895, November 2012.

[62] Gillian P Bates, Ray Dorsey, James F Gusella, Michael R Hayden, Chris Kay, Blair R Leavitt, Martha Nance, Christopher A Ross, Rachael I Scahill, Ronald Wetzel, Edward J Wild, and Sarah J Tabrizi. Huntington disease. *Nat Rev Dis Primers*, 1:15005, April 2015.

[63] Kuan-Yu Liu, Yu-Chiau Shyu, Brett A Barbaro, Yuan-Ta Lin, Yijuang Chern, Leslie Michels Thompson, Che-Kun James Shen, and J Lawrence Marsh. Disruption of the nuclear membrane by perinuclear inclusions of mutant huntingtin causes cell-cycle re-entry and striatal cell death in mouse and cell models of huntington's disease. *Hum. Mol. Genet.*, 24(6):1602–1616, March 2015.

[64] Yoshitaka Nagai, Takashi Inui, H Akiko Popiel, Nobuhiro Fujikake, Kazuhiro Hasegawa, Yoshihiro Urade, Yuji Goto, Hironobu Naiki, and Tatsushi Toda. A toxic monomeric conformer of the polyglutamine protein. *Nat. Struct. Mol. Biol.*, 14(4):332–340, April 2007.

[65] Ankur Jain and Ronald D Vale. RNA phase transitions in repeat expansion disorders. *Nature*, 546(7657):243–247, June 2017.

[66] R Nalavade, N Griesche, D P Ryan, S Hildebrand, and S Krauss. Mechanisms of RNA-induced toxicity in CAG repeat disorders. *Cell Death Dis.*, 4:e752, August 2013.

[67] C Stefanis, T h Papapetropoulos, S Scarpalezos, G Lygidakis, and C P Panayiotopoulos. X-linked spinal and bulbar muscular atrophy of late onset. a separate type of motor neuron disease? *J. Neurol. Sci.*, 24(4):493–403, April 1975.

[68] Anna Sułek, Dorota Hoffman-Zacharska, Wioletta Krysa, Walentyna Szirkowiec, Elzbieta Fidziańska, and Jacek Zaremba. CAG repeat polymorphism in the androgen receptor (AR) gene of SBMA patients and a control group. *J. Appl. Genet.*, 46(2):237–239, 2005.

[69] A R La Spada, E M Wilson, D B Lubahn, A E Harding, and K H Fischbeck. Androgen receptor gene mutations in x-linked spinal and bulbar muscular atrophy. *Nature*, 352(6330):77–79, July 1991.

[70] Lindsay E Rhodes, Brandi K Freeman, Sungyoung Auh, Angela D Kokkinis, Alison La Pean, Cheunju Chen, Tanya J Lehky, Joseph A Shrader, Ellen W Levy, Michael Harris-Love, Nicholas A Di Prospero, and Kenneth H Fischbeck. Clinical features of spinal and bulbar muscular atrophy. *Brain*, 132(Pt 12):3242–3251, December 2009.

[71] Albert R La Spada and J Paul Taylor. Polyglutamines placed into context. *Neuron*, 38(5):681–684, June 2003.

[72] Masahisa Katsuno, Hiroaki Adachi, Masahiro Waza, Haruhiko Banno, Keisuke Suzuki, Fumiaki Tanaka, Manabu Doyu, and Gen Sobue. Pathogenesis, animal models and therapeutics in spinal and bulbar muscular atrophy (SBMA). *Exp. Neurol.*, 200(1):8–18, July 2006.

[73] M Katsuno, H Adachi, A Inukai, and G Sobue. Transgenic mouse models of spinal and bulbar muscular atrophy (SBMA). *Cytogenet. Genome Res.*, 100(1-4):243–251, 2003.

[74] Iain J McEwan. Nuclear receptors: One big family. In Iain J McEwan, editor, *The Nuclear Receptor Superfamily: Methods and Protocols*, pages 3–18. Humana Press, Totowa, NJ, 2009.

[75] Marc Robinson-Rechavi, Hector Escriva Garcia, and Vincent Laudet. The nuclear receptor superfamily. *J. Cell Sci.*, 116(Pt 4):585–586, February 2003.

[76] C A Quigley, A De Bellis, K B Marschke, M K el Awady, E M Wilson, and F S French. Androgen receptor defects: historical, clinical, and molecular perspectives. *Endocr. Rev.*, 16(3):271–321, June 1995.

[77] Edward P Gelmann. Molecular biology of the androgen receptor. *J. Clin. Oncol.*, 20(13):3001–3015, July 2002.

[78] Wikipedia contributors. Nuclear receptor - Wikipedia, the free encyclopedia, 2019.

[79] Andrew R Leach and Addison Wesley. *Molecular Modeling: Principles and Applications*. Pearson Education Ltd, September 1998.

[80] S Lifson and A Warshel. Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules. *J. Chem. Phys.*, 49(11):5116–5129, December 1968.

[81] A Rahman. Correlations in the motion of atoms in liquid argon. *Phys. Rev.*, 136(2A):A405–A411, October 1964.

[82] B J Alder and T E Wainwright. Studies in molecular dynamics. i. general method. *J. Chem. Phys.*, 31(2):459–466, August 1959.

[83] Hirotaka Ode, Masaaki Nakashima, Shingo Kitamura, Wataru Sugiura, and Hironori Sato. Molecular dynamics simulation in virus research. *Front. Microbiol.*, 3:258, July 2012.

[84] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, July 1983.

[85] Sébastien Le Roux and Valeri Petkov. ISAACS – interactive structure analysis of amorphous and crystalline systems. *J. Appl. Crystallogr.*, 43(1):181–185, February 2010.

[86] Axel Brünger, Charles L Brooks, and Martin Karplus. Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chem. Phys. Lett.*, 105(5):495–500, March 1984.

[87] Niels Grønbech-Jensen and Oded Farago. A simple and effective verlet-type algorithm for simulating langevin dynamics. *Mol. Phys.*, 111(8):983–991, April 2013.

[88] H J C Berendsen, J P M Postma, W F van Gunsteren, A DiNola, and J R Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, October 1984.

[89] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, January 2007.

[90] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.*, 81(1):511–519, July 1984.

[91] W G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A Gen. Phys.*, 31(3):1695–1697, March 1985.

[92] M Parrinello and A Rahman. Crystal structure and pair potentials: A Molecular-Dynamics study. *Phys. Rev. Lett.*, 45(14):1196–1199, October 1980.

[93] M Parrinello and A Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, December 1981.

[94] Angelo Gavezzotti. *Structure Analysis and Molecular Simulation of Crystals and Liquids*. Oxford University Press, 2007.

[95] Norman L Allinger. Conformational analysis. 130. MM2. a hydrocarbon force field utilizing V1 and V2 torsional terms. *J. Am. Chem. Soc.*, 99(25):8127–8134, December 1977.

[96] Norman L Allinger, Young H Yuh, and Jenn Huei Lii. Molecular mechanics. the MM3 force field for hydrocarbons. 1. *J. Am. Chem. Soc.*, 111(23):8551–8566, November 1989.

[97] N L Allinger, K Chen, and J H Lii. An improved force field (MM4) for saturated hydrocarbons. *Journal of Computational*, 1996.

[98] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules j. am. chem. soc. 1995, 117, 5179-5197. *J. Am. Chem. Soc.*, 118(9):2309–2309, January 1996.

[99] A D MacKerell, D Bashford, M Bellott, R L Dunbrack, J D Evanseck, M J Field, S Fischer, J Gao, H Guo, S Ha, D Joseph-McCarthy, L Kuchnir, K Kuczera, F T Lau, C Mattos, S Michnick, T Ngo, D T Nguyen, B Prodhom, W E Reiher, B Roux, M Schlenkrich, J C Smith, R Stote, J Straub, M Watanabe, J Wiórkiewicz-Kuczera, D Yin, and M Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616, April 1998.

[100] Chris Oostenbrink, Alessandra Villa, Alan E Mark, and Wilfred F van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.*, 25(13):1656–1676, October 2004.

[101] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and testing of the OPLS All-Atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, November 1996.

[102] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23(3):327–341, March 1977.

[103] B Hess, H Bekker, H J C Berendsen, and others. LINCS: a linear constraint solver for molecular simulations. *Journal of*, 1997.

[104] Hans C Andersen. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.*, 52(1):24–34, October 1983.

[105] Chad W Hopkins, Scott Le Grand, Ross C Walker, and Adrian E Roitberg. Long-Time-Step molecular dynamics through hydrogen mass repartitioning. *J. Chem. Theory Comput.*, 11(4):1864–1874, April 2015.

[106] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, 159(1):98–103, July 1967.

[107] Loup Verlet. Computer "experiments" on classical fluids. II. equilibrium correlation functions. *Phys. Rev.*, 165(1):201–214, January 1968.

[108] R W Hockney and J W Eastwood. *Computer simulation using particles*. Bristol: Hilger. January 1988.

[109] William C Swope, Hans C Andersen, Peter H Berens, and Kent R Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76(1):637–649, January 1982.

[110] Elangannan Arunan, Gautam R Desiraju, Roger A Klein, Joanna Sadlej, Steve Scheiner, Ibon Alkorta, David C Clary, Robert H Crabtree, Joseph J Dannenberg, Pavel Hobza, Henrik G Kjaergaard, Anthony C Legon, Benedetta Mennucci, and David J Nesbitt. Definition of the hydrogen bond (IUPAC recommendations 2011). *Pure Appl. Chem.*, 83(8):1637–1641, 2011.

[111] George A Jeffrey. *An introduction to hydrogen bonding*. Oxford University Press, New York, 1997.

[112] George A Jeffrey and Wolfram Saenger. *Hydrogen Bonding in Biological Structures*. Springer, Berlin, Heidelberg, 1991.

[113] Thomas Steiner. The hydrogen bond in the solid state. *Angew. Chem. Int. Ed Engl.*, 41(1):49–76, January 2002.

[114] A M Sweetman, S P Jarvis, Hongqian Sang, I Lekkas, P Rahe, Yu Wang, Jianbo Wang, N R Champness, L Kantorovich, and P Moriarty. Mapping the force field of a hydrogen-bonded assembly. *Nat. Commun.*, 5:3931, May 2014.

[115] Zahra Shahbazi, Horea T Ilies, and Kazem Kazerounian. Hydrogen bonds and kinematic mobility of protein molecules. *J. Mech. Robot.*, 2(2):021009, 2010.

[116] Ivan Y Torshin, Irene T Weber, and Robert W Harrison. Geometric criteria of hydrogen bonds in proteins and identification of "bifurcated" hydrogen bonds. *Protein Eng.*, 15(5):359–363, May 2002.

[117] David L Nelson, Albert L Lehninger, and Michael M Cox. *Lehninger principles of biochemistry*. Macmillan, 2008.

[118] Daniel J Kuster, Chengyu Liu, Zheng Fang, Jay W Ponder, and Garland R Marshall. High-resolution crystal structures of protein helices reconciled with three-centered hydrogen bonds and multipole electrostatics. *PLoS One*, 10(4):e0123146, April 2015.

[119] Gail J Bartlett and Derek N Woolfson. On the satisfaction of backbone-carbonyl lone pairs of electrons in protein structures. *Protein Sci.*, 25(4):887–897, April 2016.

[120] G A Jeffrey and Hanna Maluszynska. A survey of hydrogen bond geometries in the crystal structures of amino acids. *Int. J. Biol. Macromol.*, 4(3):173–185, April 1982.

[121] R Taylor, O Kennard, and W Versichel. The geometry of the N–H...O=C hydrogen bond. 3. hydrogen-bond distances and angles. *Acta Crystallogr. B*, 40(3):280–288, June 1984.

[122] G A Jeffrey, H Maluszynska, and J Mitra. Hydrogen bonding in nucleosides and nucleotides. *Int. J. Biol. Macromol.*, 7(6):336–348, December 1985.

[123] George A Jeffrey and Shozo Takagi. Hydrogen-bond structure in carbohydrate crystals. *Acc. Chem. Res.*, 11(7):264–270, July 1978.

[124] G A Jeffrey and J Mitra. The hydrogen-bonding patterns in the pyranose and pyranoside crystal structures. *Acta Crystallogr. B*, 39(4):469–480, August 1983.

[125] G A Jeffrey and H Maluszynska. A survey of the geometry of hydrogen bonds in the crystal structures of barbiturates, purines and pyrimidines. *J. Mol. Struct.*, 147(1):127–142, September 1986.

[126] G Némethy, D C Phillips, S J Leach, and H A Scheraga. A second right-handed helical structure with the parameters of the Pauling-Corey alpha-helix. *Nature*, 214(5086):363–365, April 1967.

[127] Kenneth B Wiberg, Manuel Marquez, and Henry Castejon. Lone pairs in carbonyl compounds and ethers. *J. Org. Chem.*, 59(22):6817–6822, November 1994.

[128] Wikipedia contributors. Protein structure — Wikipedia, the free encyclopedia, 2019.

[129] Joseph Feher. 2.3 - protein structure. In Joseph Feher, editor, *Quantitative Human Physiology (Second Edition)*, pages 130–141. Academic Press, Boston, January 2017.

[130] L Pauling, R B Corey, and H R Branson. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.*, 37(4):205–211, April 1951.

[131] Wikipedia contributors. Dihedral angle - Wikipedia, the free encyclopedia, 2019. [Online; accessed 25-August-2019].

[132] G N Ramachandran, C Ramakrishnan, and V Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7:95–99, July 1963.

[133] Simon C Lovell, Ian W Davis, W Bryan Arendall, 3rd, Paul I W de Bakker, J Michael Word, Michael G Prisant, Jane S Richardson, and David C Richardson. Structure validation by calpha geometry: phi,psi and cbeta deviation. *Proteins*, 50(3):437–450, February 2003.

[134] C N Pace and J M Scholtz. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.*, 75(1):422–427, July 1998.

[135] J A Schellman. The stability of hydrogen-bonded peptide structures in aqueous solution. *C. R. Trav. Lab. Carlsberg Chim.*, 29(14-15):230–259, 1955.

[136] B H Zimm and J K Bragg. Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.*, 31(2):526–535, August 1959.

[137] Shneior Lifson and A Roig. On the theory of Helix—Coil transition in polypeptides. *J. Chem. Phys.*, 34(6):1963–1974, June 1961.

[138] S Lifson and N Lotan. On the statistical mechanical theories of Helix-Coil transition in polypeptides, and the structure and structural stability of proteins. *Isr. J. Chem.*, 12(1-2):201–206, November 1974.

[139] Gauri P Misra and Chung F Wong. Predicting helical segments in proteins by a helix-coil transition theory with parameters derived from a structural database of proteins. *Proteins: Struct. Funct. Bioinf.*, 28(3):344–359, July 1997.

[140] B E Eichinger and M Fixman. Helix-coil transition in heterogeneous chains. II. DNA model. *Biopolymers*, 9(2):205–221, February 1970.

[141] Jeffrey Skolnick. Effect of loop entropy on the helix-coil transition of α-helical, two-chain, coiled coils. 2. supermatrix formulation of the perfect-matching model. *Macromolecules*, 16(11):1763–1770, November 1983.

[142] Andrew J Doig. Recent advances in helix-coil theory. *Biophys. Chem.*, 101-102:281–293, December 2002.

[143] Andreas Vitalis and Amedeo Caflisch. 50 years of Lifson-Roig models: Application to molecular simulation data. *J. Chem. Theory Comput.*, 8(1):363–373, January 2012.

[144] V Muñoz and L Serrano. Elucidating the folding problem of helical peptides using empirical parameters. *Nat. Struct. Biol.*, 1(6):399–409, June 1994.

[145] V Muñoz and L Serrano. Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers*, 41(5):495–509, April 1997.

[146] M Vásquez and H A Scheraga. Effect of sequence-specific interactions on the stability of helical conformations in polypeptides. *Biopolymers*, 27(1):41–58, January 1988.

[147] P J Gans, P C Lyu, M C Manning, R W Woody, and N R Kallenbach. The helix-coil transition in heterogeneous peptides with specific side-chain interactions: theory and comparison with CD spectral data. *Biopolymers*, 31(13):1605–1614, November 1991.

[148] Benjamin J Stapley, Carol A Rohl, and Andrew J Doig. Addition of side chain interactions to modified Lifson-Roig helix-coil theory: Application to energetics of phenylalanine-methionine interactions. *Protein Sci.*, 4(11):2383–2391, 1995.

[149] William Shalongo and Earle Stellwagen. Incorporation of pairwise interactions into the Lifson-Roig model for helix prediction. *Protein Sci.*, 4(6):1161–1166, 1995.

[150] Andrew J Doig, Avijit Chakrabartty, Tod M Klingler, and Robert L Baldwin. Determination of free energies of N-Capping in .alpha.-helixes by modification of the Lifson-Roig Helix-Coil theory to include N- and C-Capping. *Biochemistry*, 33(11):3396–3403, March 1994.

[151] Ana Rosa Viguera Emmanuel Lacroix, Luis Serrano. Elucidating the folding problem of aaa-helices: Local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J. Mol. Biol.*, 284:173–191, 1998.

[152] V Muñoz and L Serrano. Elucidating the folding problem of helical peptides using empirical parameters. II. helix macrodipole effects and rational modification of the helical content of natural peptides. *J. Mol. Biol.*, 245(3):275–296, January 1995.

[153] J M Scholtz, H Qian, V H Robbins, and R L Baldwin. The energetics of ion-pair and hydrogen-bonding interactions in a helical peptide. *Biochemistry*, 32(37):9668–9676, September 1993.

[154] Bahareh Eftekharzadeh, Alessandro Piai, Giulio Chiesa, Daniele Mungianu, Jesús García, Roberta Pierattelli, Isabella C Felli, and Xavier Salvatella. Sequence context influences the structure and aggregation behavior of a PolyQ tract. *Biophys. J.*, 110(11):2361–2366, June 2016.

[155] Ramya Rangan, Massimiliano Bonomi, Gabriella T Heller, Andrea Cesari, Giovanni Bussi, and Michele Vendruscolo. Determination of structural ensembles of proteins: Restraining vs reweighting. *J. Chem. Theory Comput.*, November 2018.

[156] Xiao-Ping Xu and David A Case. Probing multiple effects on 15n, 13C alpha, 13C beta, and 13c' chemical shifts in peptides using density functional theory. *Biopolymers*, 65(6):408–423, December 2002.

[157] M J Harvey, G Giupponi, and G De Fabritiis. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.*, 5(6):1632–1639, June 2009.

[158] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.*, 100(9):L47–9, May 2011.

[159] Sandro Bottaro, Giovanni Bussi, Scott D Kennedy, Douglas H Turner, and Kresten Lindorff-Larsen. Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. *Sci Adv*, 4(5):eaar8521, May 2018.

[160] Sandro Bottaro, Tone Bengtsen, and Kresten Lindorff-Larsen. Integrating molecular simulation and experimental data: A Bayesian/Maximum entropy reweighting approach. October 2018.

[161] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

[162] E T Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106(4):620–630, May 1957.

[163] C E Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, July 1948.

[164] Ariel Caticha. Relative entropy and inductive inference. November 2003.

[165] S Kullback and R A Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22(1):79–86, March 1951.

[166] I K McDonald and J M Thornton. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, 238(5):777–793, May 1994.

[167] E Cubero, M Orozco, and F J Luque. Electron density topological analysis of the c–h...o anti-hydrogen bond in the fluoroform–oxirane complex. *Chem. Phys. Lett.*, 310(5):445–450, September 1999.

[168] E Cubero, M Orozco, P Hobza, and F J Luque. Hydrogen bond versus Anti-Hydrogen bond: A comparative analysis based on the electron density topology. *J. Phys. Chem. A*, 103(32):6394–6401, August 1999.

[169] Eric S Eberhardt and Ronald T Raines. Amide-Amide and Amide-Water hydrogen bonds: Implications for protein folding and stability. *J. Am. Chem. Soc.*, 116(5):2149–2150, March 1994.

[170] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, December 2007.

[171] Henry van den Bedem and James S Fraser. Integrative, dynamic structural biology at atomic resolution–it's about time. *Nat. Methods*, 12(4):307–318, April 2015.

[172] Anthony Mittermaier and Lewis E Kay. New tools provide new insights in NMR studies of protein dynamics. *Science*, 312(5771):224–228, April 2006.

[173] Johnny Habchi, Peter Tompa, Sonia Longhi, and Vladimir N Uversky. Introducing protein intrinsic disorder. *Chem. Rev.*, 114(13):6561–6588, July 2014.

[174] Pietro Sormanni, Damiano Piovesan, Gabriella T Heller, Massimiliano Bonomi, Predrag Kukic, Carlo Camilloni, Monika Fuxreiter, Zsuzsanna Dosztanyi, Rohit V Pappu, M Madan Babu, Sonia Longhi, Peter Tompa, A Keith Dunker, Vladimir N Uversky, Silvio C E Tosatto, and Michele Vendruscolo. Simultaneous quantification of protein order and disorder. *Nat. Chem. Biol.*, 13(4):339–342, March 2017.

[175] Robert B Best, Nicolae-Viorel Buchete, and Gerhard Hummer. Are current molecular dynamics force fields too helical? *Biophys. J.*, 95(1):L07–9, July 2008.

[176] Robert B Best, Xiao Zhu, Jihyun Shim, Pedro E M Lopes, Jeetain Mittal, Michael Feig, and Alexander D Mackerell, Jr. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J. Chem. Theory Comput.*, 8(9):3257–3273, September 2012.

[177] Da-Wei Li and Rafael Brüschweiler. Iterative optimization of molecular mechanics force fields from NMR data of Full-Length proteins. *J. Chem. Theory Comput.*, 7(6):1773–1782, June 2011.

[178] David L Mobley. Let's get honest about sampling. *J. Comput. Aided Mol. Des.*, 26(1):93–95, January 2012.

[179] Santiago Esteban-Martín, Robert Bryn Fenwick, and Xavier Salvatella. Synergistic use of NMR and MD simulations to study the structural heterogeneity of proteins. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2(3):466–478, May 2012.

[180] Kresten Lindorff-Larsen, Robert B Best, Mark A Depristo, Christopher M Dobson, and Michele Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128–132, January 2005.

[181] Carlo Camilloni, Paul Robustelli, Alfonso De Simone, Andrea Cavalli, and Michele Vendruscolo. Characterization of the conformational equilibrium between the two major substates of RNase a using NMR chemical shifts. *J. Am. Chem. Soc.*, 134(9):3968–3971, March 2012.

[182] Juuso Lehtivarjo, Kari Tuppurainen, Tommi Hassinen, Reino Laatikainen, and Mikael Peräkylä. Combining NMR ensembles and molecular dynamics simulations provides more realistic models of protein structures in solution and leads to better chemical shift prediction. *J. Biomol. NMR*, 52(3):257–267, March 2012.

[183] Jed W Pitera and John D Chodera. On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.*, 8(10):3445–3451, October 2012.

[184] Carlo Camilloni and Michele Vendruscolo. Statistical mechanics of the denatured state of a protein using replica-averaged metadynamics. *J. Am. Chem. Soc.*, 136(25):8982–8991, June 2014.

[185] Enrico Ravera, Luca Sgheri, Giacomo Parigi, and Claudio Luchinat. A critical assessment of methods to recover information from averaged data. *Phys. Chem. Chem. Phys.*, 18(8):5686–5701, February 2016.

[186] Andrew B Ward, Andrej Sali, and Ian A Wilson. Biochemistry. integrative structural biology. *Science*, 339(6122):913–915, February 2013.

[187] Troy Cellmer, Marco Buscaglia, Eric R Henry, James Hofrichter, and William A Eaton. Making connections between ultrafast protein folding kinetics and molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.*, 108(15):6103–6108, April 2011.

[188] Anton Arkhipov, Yibing Shan, Rahul Das, Nicholas F Endres, Michael P Eastwood, David E Wemmer, John Kuriyan, and David E Shaw. Architecture and membrane interactions of the EGF receptor. *Cell*, 152(3):557–569, January 2013.

[189] G Cornilescu, F Delaglio, and A Bax. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, 13(3):289–302, March 1999.

[190] D S Wishart and D A Case. Use of chemical shifts in macromolecular structure determination. *Methods Enzymol.*, 338:3–34, 2001.

[191] Kai J Kohlhoff, Paul Robustelli, Andrea Cavalli, Xavier Salvatella, and Michele Vendruscolo. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J. Am. Chem. Soc.*, 131(39):13894–13895, October 2009.

[192] Juuso Lehtivarjo, Tommi Hassinen, Samuli-Petrus Korhonen, Mikael Peräkylä, and Reino Laatikainen. 4D prediction of protein (1)h chemical shifts. *J. Biomol. NMR*, 45(4):413–426, December 2009.

[193] Jens Meiler. PROSHIFT: protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR*, 26(1):25–37, May 2003.

[194] Stephen Neal, Alex M Nip, Haiyan Zhang, and David S Wishart. Rapid and accurate calculation of protein 1h, 13C and 15N chemical shifts. *J. Biomol. NMR*, 26(3):215–240, July 2003.

[195] Yang Shen and Ad Bax. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR*, 38(4):289–302, August 2007.

[196] Charles D Schwieters, John J Kuszewski, and G Marius Clore. Using Xplor–NIH for NMR molecular structure determination. *Prog. Nucl. Magn. Reson. Spectrosc.*, 48(1):47–62, March 2006.

[197] J Kuszewski, A M Gronenborn, and G M Clore. The impact of direct refinement against proton chemical shifts on protein structure determination by NMR. *J. Magn. Reson. B*, 107(3):293–297, June 1995.

[198] G M Clore and A M Gronenborn. New methods of structure refinement for macromolecular structure determination by NMR. *Proc. Natl. Acad. Sci. U. S. A.*, 95(11):5891–5898, May 1998.

[199] Paul Robustelli, Kai Kohlhoff, Andrea Cavalli, and Michele Vendruscolo. Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. *Structure*, 18(8):923–933, August 2010.

[200] Paul Robustelli, Andrea Cavalli, Christopher M Dobson, Michele Vendruscolo, and Xavier Salvatella. Folding of small proteins by monte carlo simulations with chemical shift restraints without the use of molecular fragment replacement or structural homology. *J. Phys. Chem. B*, 113(22):7890–7896, June 2009.

[201] Falk Hoffmann and Birgit Strodel. Protein structure prediction using global optimization by basin-hopping with NMR shift restraints. *J. Chem. Phys.*, 138(2):025102, January 2013.

[202] Carlo Camilloni, Andrea Cavalli, and Michele Vendruscolo. Replica-Averaged metadynamics. *J. Chem. Theory Comput.*, 9(12):5610–5617, December 2013.

[203] Andrea Cavalli, Carlo Camilloni, and Michele Vendruscolo. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.*, 138(9):094112, March 2013.

[204] Benoît Roux and Jonathan Weare. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J. Chem. Phys.*, 138(8):084107, February 2013.

[205] Albert Escobedo, Busra Topal, Micha B A Kunze, Juan Aranda, Giulio Chiesa, Daniele Mungianu, Ganeko Bernardo-Seisdedos, Bahareh Eftekharzadeh, Margarida Gairí, Roberta Pierattelli, Isabella C Felli, Tammo Diercks, Oscar Millet, Jesús García, Modesto Orozco, Ramon Crehuet, Kresten Lindorff-Larsen, and Xavier Salvatella. Side chain to main chain hydrogen bonds stabilize a polyglutamine helix in a transcription factor. *Nat. Commun.*, 10(1):2034, May 2019.

[206] Ferdinando Fiumara, Luana Fioriti, Eric R Kandel, and Wayne A Hendrickson. Essential role of coiled coils for aggregation and activity of Q/N-rich prions and PolyQ proteins. *Cell*, 143(7):1121–1135, December 2010.

[207] Carlo Camilloni, Alfonso De Simone, Wim F Vranken, and Michele Vendruscolo. Determination of secondary structure populations in disordered states

of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry*, 51(11):2224–2231, March 2012.

[208] Justin S Miller, Robert J Kennedy, and Daniel S Kemp. Solubilized, spaced polyalanines: a context-free system for determining amino acid alpha-helix propensities. *J. Am. Chem. Soc.*, 124(6):945–962, February 2002.

[209] David Whitford. *Proteins: Structure and Function*. Wiley, May 2005.

[210] R Bryn Fenwick, Santi Esteban-Martín, Barbara Richter, Donghan Lee, Korvin F A Walter, Dragomir Milovanovic, Stefan Becker, Nils A Lakomek, Christian Griesinger, and Xavier Salvatella. Weak long-range correlated motions in a surface patch of ubiquitin involved in molecular recognition. *J. Am. Chem. Soc.*, 133(27):10336–10339, July 2011.

[211] R Bryn Fenwick, Laura Orellana, Santi Esteban-Martín, Modesto Orozco, and Xavier Salvatella. Correlated motions are a fundamental property of β-sheets. *Nat. Commun.*, 5:4070, June 2014.

[212] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, 47(D1):D506–D515, January 2019.

[213] D J Lipman and W R Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, March 1985.

[214] W R Pearson and D J Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.*, 85(8):2444–2448, April 1988.

[215] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[216] T D Schneider and R M Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18(20):6097–6100, October 1990.

[217] Oliver Bembom. *seqLogo: Sequence logos for DNA sequence alignments*, 2016. R package version 1.40.0.

[218] Matteo Ramazzotti, Elodie Monsellier, Choumouss Kamoun, Donatella Degl'Innocenti, and Ronald Melki. Polyglutamine repeats are associated to specific sequence biases that are conserved among eukaryotes. *PLoS One*, 7(2):e30824, February 2012.

[219] Warren L DeLano. PyMOL: An Open-Source molecular graphics tool. *DeLano Scientific*, 2002.

[220] D A Case. Amber16.pdf. *University of California, San Francisco*, 2016.

[221] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.*, 32(7):1456–1465, May 2011.

[222] Eric D Glendening, Clark R Landis, and Frank Weinhold. NBO 6.0: natural bond orbital analysis program. *J. Comput. Chem.*, 34(16):1429–1437, June 2013.

[223] Werner Kutzelnigg. Atoms in molecules. a quantum theory. (reihe: International series of monographs on chemistry, vol. 22.) von R.F.W. bader. clarendon press, oxford, 1990. XVIII, 438 s., geb. £ 50.00. – ISBN 0-19-855168-1. *Angew. Chem. Int. Ed Engl.*, 104(10):1423–1423, October 1992.

[224] Frank Weinhold. Natural bond critical point analysis: quantitative relationships between natural bond orbital-based and QTAIM-based topological descriptors of chemical bonding. *J. Comput. Chem.*, 33(30):2440–2449, November 2012.

[225] G D Fasman. *Circular Dichroism and the Conformational Analysis of Biomolecules*. Springer Science & Business Media, November 2013.

[226] Norma J Greenfield. Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.*, 1(6):2876–2890, 2006.

[227] Sharon M Kelly, Thomas J Jess, and Nicholas C Price. How to study proteins by circular dichroism. *Biochim. Biophys. Acta*, 1751(2):119–139, August 2005.

[228] A Lobley, L Whitmore, and B A Wallace. DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra. *Bioinformatics*, 18(1):211–212, January 2002.

[229] Lee Whitmore and B A Wallace. Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers*, 89(5):392–400, May 2008.

[230] Lee Whitmore and B A Wallace. DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Res.*, 32(Web Server issue):W668–73, July 2004.

[231] Da-Wei Li and Rafael Brüschweiler. PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *J. Biomol. NMR*, 54(3):257–265, November 2012.

[232] Maria Baias, Pieter E S Smith, Koning Shen, Lukasz A Joachimiak, Szymon Żerko, Wiktor Koźmiński, Judith Frydman, and Lucio Frydman. Structure and dynamics of the huntingtin exon-1 N-Terminus: A solution NMR perspective. *J. Am. Chem. Soc.*, 139(3):1168–1176, January 2017.

[233] Annika Urbanek, Anna Morató, Frédéric Allemand, Elise Delaforge, Aurélie Fournet, Matija Popovic, Stephane Delbecq, Nathalie Sibille, and Pau Bernadó. A general strategy to access structural information at atomic resolution in polyglutamine homorepeats. *Angew. Chem. Int. Ed Engl.*, 57(14):3598–3601, March 2018.

[234] Jun-Ye Hong, Dong-Dong Wang, Wei Xue, Hong-Wei Yue, Hui Yang, Lei-Lei Jiang, Wen-Ning Wang, and Hong-Yu Hu. Structural and dynamic studies reveal that the ala-rich region of ataxin-7 initiates α-helix formation of the polyq tract but suppresses its aggregation. *Sci. Rep.*, 9(1):7481, May 2019.

[235] Thomas R Peskett, Frédérique Rau, Jonathan O'Driscoll, Rickie Patani, Alan R Lowe, and Helen R Saibil. A liquid to solid phase transition underlying pathological huntingtin exon1 aggregation. *Mol. Cell*, 70(4):588–601.e6, May 2018.

[236] Erin M Langdon, Yupeng Qiu, Amirhossein Ghanbari Niaki, Grace A McLaughlin, Chase A Weidmann, Therese M Gerbich, Jean A Smith, John M Crutchley, Christina M Termini, Kevin M Weeks, Sua Myong, and Amy S Gladfelter. mRNA structure determines specificity of a polyq-driven phase separation. *Science*, 360(6391):922–927, May 2018.

[237] Ammon E Posey, Kiersten M Ruff, Tyler S Harmon, Scott L Crick, Aimin Li, Marc I Diamond, and Rohit V Pappu. Profilin reduces aggregation and phase separation of huntingtin n-terminal fragments by preferentially binding to soluble monomers and oligomers. *J. Biol. Chem.*, 293(10):3734–3746, March 2018.

[238] E N Baker and R E Hubbard. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.*, 44(2):97–179, 1984.

[239] J A Ippolito, R S Alexander, and D W Christianson. Hydrogen bond stereochemistry in protein structure and function. *J. Mol. Biol.*, 215(3):457–471, October 1990.

[240] Alexandre V Morozov, Tanja Kortemme, Kiril Tsemekhman, and David Baker. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. U. S. A.*, 101(18):6946–6951, May 2004.

[241] Tanja Kortemme, Alexandre V Morozov, and David Baker. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.*, 326(4):1239–1259, February 2003.

[242] C H Görbitz. Hydrogen-bond distances and angles in the structures of amino acids and peptides. *Acta Crystallogr. B*, 45(4):390–395, August 1989.

[243] Alexandre V Morozov, Tanja Kortemme, Kiril Tsemekhman, and David Baker. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. U. S. A.*, 101(18):6946–6951, May 2004.