Università degli Studi di Bologna

Dipartimento di Chimica Fisica ed Inorganica

Dottorato di ricerca in Scienze Chimiche

XX Ciclo

Coordinatore: Prof. Vincenzo Balzani

ADRIANA PIETROPAOLO

STRUCTURE DETERMINATION OF PROTEINS AND PEPTIDES IN SOLUTION: SIMULATION, CHIRALITY AND NMR STUDIES

Tesi di Dottorato

CHIM/02

Relatore: Prof. Claudio Zannoni

Correlatore: Dr. Luca Muccioli

Anno 2008

Courage is not the absence of fear, but rather the judgement that something else is more important than fear.

Ambrose Redmoon

A mio padre, a mia madre, a mio fratello e a Francesco.

A boat, beneath a sunny sky
Lingering onward dreamily
In an evening of July –

Children three that nestle near,
Eager eye and willing ear,
Pleased a simple tale to hear –

Long has paled that sunny sky:
Echoes fade and memories die:
Autumn frosts have slain July.

Still she haunts me, phantomwise,
Alice moving under skies
Never seen by waking eyes.

Children yet, the tale to hear,
Eager eye and willing ear,
Lovingly shall nestle near.

In a Wonderland they lie,
Dreaming as the days go by,
Dreaming as the summers die:

Ever drifting down the stream –
Lingering in the golden gleam –
Life, what is it but a dream?

Lewis Carroll, from Alice's Adventures in Wonderland, chapter XII

Abstract

The study of protein fold is a central problem in life science, leading in the last years to several attempts for improving our knowledge of the protein structures. Here, this challenging problem is tackled by means of molecular dynamics, chirality and NMR studies.

In the last decades, many algorithms were designed for the protein secondary structure assignment, which reveals the local protein shape adopted by segments of amino acids. In this regard, the use of local chirality for the protein secondary structure assignment was demonstreted, trying to correlate as well the propensity of a given amino acid for a particular secondary structure.

The protein fold can be studied also by Nuclear Magnetic Resonance (NMR) investigations, finding the average structure adopted from a protein. In this context, the effect of Residual Dipolar Couplings (RDCs) in the structure refinement was shown, revealing a strong improvement of structure resolution.

A wide extent of this thesis is devoted to the study of avian prion protein. Prion protein is the main responsible of a vast class of neurodegenerative diseases, known as Bovine Spongiform Encephalopathy (BSE), present in mammals, but not in avian species and it is caused from the conversion of cellular prion protein to the pathogenic *misfolded* isoform, accumulating in the brain in form of amiloyd plaques. In particular, the N-terminal region, namely the initial part of the protein, is quite different between mammal and avian species but both of them contain multimeric sequences called *Repeats*, octameric in mammals and hexameric in avians. However, such repeat regions show differences in the contained amino acids, in particular only avian hexarepeats contain tyrosine residues. The chirality analysis of avian prion protein configurations obtained from molecular dynamics reveals a high stiffness of the avian protein, which tends to preserve its regular secondary structure. This is due to the presence of prolines, histidines and especially tyrosines, which form a hydrogen bond network in the hexarepeat region, only possible in the avian protein, and thus probably hampering the aggregation.

Contents

Aim and results of the research

An overview on the universe of Proteins

In the last years, the studies of proteins and of particular regions of them, have been bringing to a number of papers actually increasing.

Proteins are *biopolymers*, macromolecules built prevalently by carbon, nitrogen, hydrogen, oxygen and in less degree sulphur. The elementary units of proteins, the *bricks*, are the 20-L natural *amino acids*, shown in Figure 1. A different combination of them lead to different proteins, with a different structure, often with a different function. For many proteins, the correct tridimensional structure is essential for their function, failure to fold into the intended shape usually affords inactive proteins with different properties (an example will be shown in Chapter 4). However, sometimes, for unknown reasons, the protein inactivation does not happen, thus producing *misfolded* (incorrectly folded) structures, not able to explicate the function dictaded from the genetic code: a common opinion is actually that several neurodegenerative and other diseases result from the accumulation of *misfolded* proteins. *Proteins* (Figure 2) thus constitute an essential component in living systems partecipating in every process within cells. Many proteins are *enzymes* that catalyze biochemical reactions and are vital for metabolism. Enzymes are usually highly specific catalysts that accelerate only one or a few chemical reactions. At the same time, enzymes affect most of the reactions involved in metabolism and catabolism as well as DNA replication, DNA repair, and RNA synthesis. Actually, the study of proteins is a fundamental branch in life science and understanding their function and how they fold into the structure giving them the proper function is the purpose of *proteomics*.

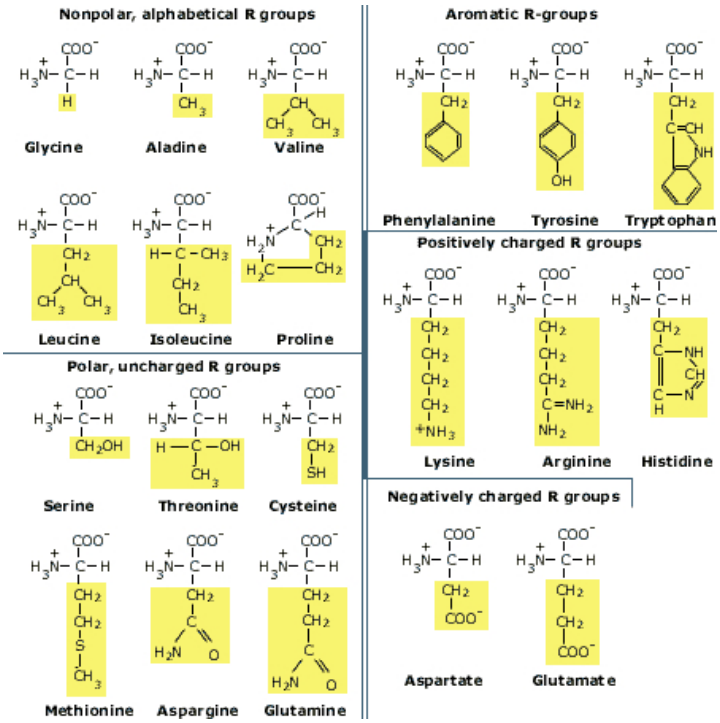


Figure 1: The twenty L-amino acids, classified depending on the polarity of their side chain.

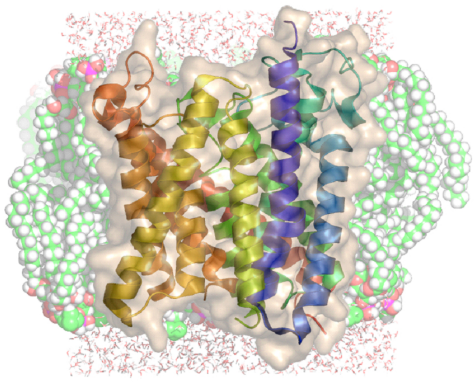


Figure 2: A membrane protein, typically adopting an α helix structure, with helices perpendicular to the membrane surface.

The protein structure: a dynamical entity

Generally, proteins adopt different structures, called secondary, for segments of amino acids. In particular, the right handed α helix, discovered by Linus Pauling, normally involves more than four residues, having a pitch of 5.4 Å. It is featured by hydrogen bonds (weak chemical interactions that becoming high in number give a relative stability to the molecule, Figure 3), between residue i and residue $i+4$; β sheets are flat structures, usually involving parallel or antiparallel hydrogen bonds depending on their orientation; bulges are isolated β sheets; β turns show a $i-i+3$ hydrogen bond pattern, becoming a right handed 3_{10} helix if more than three residues are involved in a turn region and finally Poly-L-proline II is a left-handed helix, with 9.3 Å pitch. If no secondary structure is shared by a group of amino acids, the region is called *coil*. Therefore, coils are unstructured regions, which sometimes may be misconfused with Poly-L-proline II, as we will see in Chapter 2.

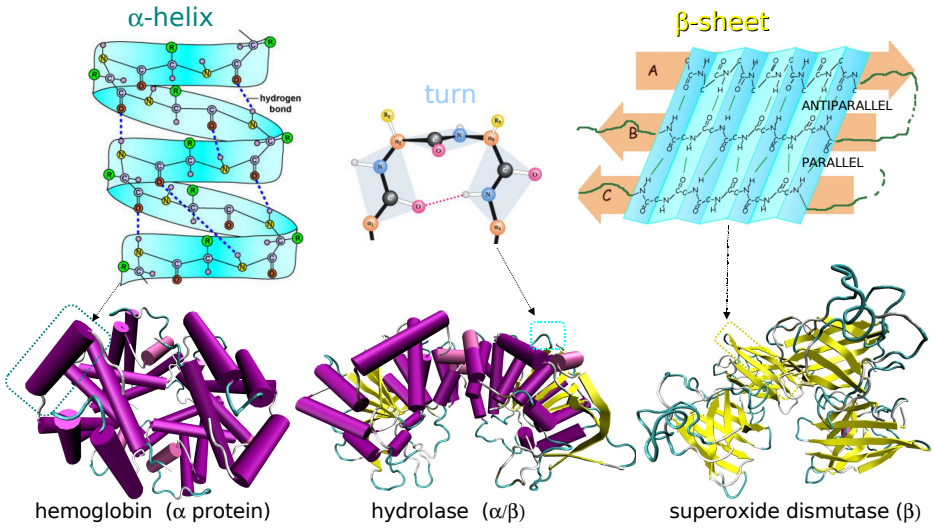


Figure 3: The most common protein secondary structures: α helix is shown in violet, β sheets in yellow, turn in green and coil in white. H-bonds are shown by dotted lines.

How proteins can be studied by a chemist

Nature designed biomolecules as *asymmetric*, thus not superimposable to their specular images; in chemistry such molecules are termed as *chiral*, from the greek, *cheir*, hand, as the two hands are an example of not superimposition. All the amino acids, but glycine, are chiral. Proteins are intrinsically chiral entities, and thus a section of this PhD work is related to the study of protein fold, analyzing the chirality of the different protein secondary structures. This study has permitted to analyze in deep detail the folding and misfolding of proteins, in terms of detection of the different fold of proteins and of relating the propensity of an amino acid for a given secondary structure.

Generally, proteins are studied experimentally by means of Nuclear Magnetic Resonance (NMR) (see chapter 3), which detects the average structure adopted, thus finding the ensemble of conformations populated at given experimental conditions; by X-ray diffraction, finding the equilibrium geometry in a crystal; by Circular Dichroism spectroscopy (CD), i.e. the absorption of circularly polarized light, which distinguishes between the chirality of two specular molecules and in the case of proteins gives information about the average structure adopted in solution, although without being able to access to protein geometry, as NMR or X-ray do.

From a computational viewpoint, molecular dynamics is the most used approach for studying proteins. Molecular dynamics is able to mimick the time evolution of a given protein structure in water for a defined time scale, so that the union of single structures, which constitutes the ensemble, in principle reflects the data found by the NMR technique.

The study of avian prion protein

A wide extent of this thesis is devoted to the study of avian prion protein.

Prion protein is the main responsible of a vast class of neurodegenerative diseases, known as Bovine Spongiform Encephalopathy (BSE), present in mammals, but not in avian species and it is caused from the conversion of cellular prion protein to the pathogenic *misfolded* isoform, accumulating in the brain in form of amiloyd plaques.
In particular, the N-terminal region, namely the initial part of the protein, is quite different between mammal and avian species but both of them contain multimeric sequences called *Repeats*, octameric in mammals and hexameric in avians. However, such

repeat regions show differences in the contained amino acids (see Chapter 4 for details), as only avian repeats contain tyrosine (see Figure 1).

The chirality analysis of avian prion protein configurations obtained from molecular dynamics reveals a high stiffness of the avian protein, which tends to preserve its regular secondary structure. This is due to the presence of proline, histidine and especially tyrosine, whose absence in the primary sequence of mammal prion could be possibly determining for the peculiar aggregation observed in mammal species.

Chapter 1

Computational methods for studying the protein conformation

1.1 A brief introduction on molecular dynamics simulations of proteins

From a thermodynamics point of view, each protein secondary structure can be thought to correspond to a local minima in the free energy of a given protein segment. The resulting collections of local minima can be easily accessible, depending on the temperature of the system. In the light of this, proteins are dynamical entity, able to have local conformational changes, at a given temperature. Therefore, experimentally determining the three dimensional structure of a protein is often very difficult and expensive, especially for highly dynamical structures, such as flexible regions of proteins, i.e. loop or turns. To overcome this difficulty, a growing interest in simulating the dynamics of proteins derives from its application to many properties of them, such as the possibility of studying, at least in principle, the processess of folding and unfolding, the role of dynamics in biological function, the refining of X-ray and nuclear magnetic resonance (NMR) structures (see Chapter 3), and protein-protein and protein-ligand interactions. The advantage of simulation approach is that, within the accuracy of the underlying potential energy functions, it provides information about the folding and unfolding pathways, the final folded (native) structure, the time dependence of these events, and the inter-residue interactions that underline these processes.

In the theoretical approach, based on empirical potential energy functions, Newton's

or Lagrange's equations are solved to obtain coordinates and momenta of the particles along the folding and unfolding trajectories. Alternative approaches are based on solving Langevin's equations when the solvent is not treated explicitly. Both approaches are time-consuming and require extensive computer power to solve these equations. In fact, it is only the development of such computing power that has made possible to solve physical problems by MD calculations. The modern era of MD calculations with electronic computers begun with the work of Alder and Wainwright [1,2], who calculated the nonequilibrium and equilibrium properties of a collection of several hundred hard-sphere particles. By providing an exact solution of the simultaneous classical equations of motion, they were able to obtain the equation of state (pressure and volume) and the Maxwell-Boltzmann velocity distribution. Rahman [3] carried out the first MD simulations of a real system when he studied the dynamics of liquid argon at 94.4 K. Later, Rahman and Stillinger [4] applied the MD technique to explore the physical properties of liquid water. Treating the water molecule as a rigid asymmetric rotor with an effective Ben-Naim and Stillinger pair potential version of the Hamiltonian, they computed the structural properties and kinetic behavior, demonstrating that the liquid water structure consists of a highly strained random hydrogen-bond network, with the diffusion process proceeding continuously by the cooperative interaction of neighbors.

Karplus and coworkers [5] carried out the first application of MD to proteins. However, this study did not deal with the protein folding problem, but instead, they investigated the dynamics of the folded globular protein bovine pancreatic trypsin inhibitor. As in the work of Rahman and Stillinger, Karplus and coworkers [5] solved the classical equations of motion for all the atoms of the protein simultaneously with an empirical potential energy function, starting with the X-ray structure and with initial velocities set equal to zero. Their results provided the magnitude, correlations, and decay of fluctuations about the average structure, and suggested that the protein interior is fluid-like in that the local atomic motions have a diffusional character. Researchers have applied this technique extensively in the refinement of X-ray and NMR structures, but because of the need to take small (femtosecond) time steps along the evolving trajectory to keep the numerical algorithm stable, it has not been successful in treating the real long-time folding of a globular protein, except for very small ones. However, many of the MD applications to globular proteins have been made considering the initial unfolding steps, followed by refolding. In applying the MD technique, one must consider numerous trajectories, rather than a single one, in order to cover the large

multidimensional, conformational potential energy space and to obtain proper statistical mechanical averages of the folding/unfolding properties. Since the first papers from the Karplus lab, numerous MD calculations have been carried out in the laboratories of Brooks [6], van Gunsteren [7], Levitt [8], Jorgensen [9], Daggett [10, 11], Kollman [12], Pande [13], Berendsen [14], Baker [15], McCammon [16], and others.

1.2 Molecular Dynamics

Molecular Dynamics (MD) is a computer simulation technique where the time evolution of a set of interacting particles (generally atoms or molecules) is followed step by step by integrating their equations of motion. Therefore, in contrast with the stochastic Monte Carlo simulations, molecular dynamics is a deterministic technique if we use a deterministic dynamics: given an initial set of positions and velocities, the subsequent time evolution is completely determined and in principle reversible. The forces are usually obtained as the gradient of a potential energy function, depending on the positions and possibly on the orientations of the particles. The realism of the simulation therefore depends on the ability of the potential chosen to reproduce the potential experienced by the real system under the conditions at which the simulation is run, and on the numerical accuracy of the integration of the equations of motions.

In a classical MD simulation the forces are derived from a classical potential, i.e. an interaction potential that is a function of the atoms (molecules) positions, and does not take into account the electrons positions. A quantum MD simulation is one in which the forces can be calculated from both a classical potential and the electronic Schrödinger equation. While evolving in space through time, the system explores a region of *phase space*, the collection of all the configurations or states which a system could assume if there were no constraints on it. However, in reality it is only possible to consider systems under some forms of constraints, in which case only a region of phase space, called *ensemble*, is accessible. As the system moves through phase space, thermodynamic properties can be obtained by taking their average value throughout the ensemble: this technique is analogous to obtaining ensemble averages based on probability distribution functions and can be rationalized with the help of statistical mechanics theory. The simulations usually need extensive computer power, and even with the most powerful computers available today it is not possible to calculate the evolution of more than perhaps 10^6 atoms at a time. This is a very small fraction, considering that a sample

employed to measure experimentally a macroscopic property has a dimension of $O(10^{20})$ atoms. Also, depending on system size, it is not possible to simulate processes that last more than some nanoseconds. In spite of these limitations, molecular dynamics simulations can be used to examine and describe numerous problems in physics and chemistry.

1.2.1 Hamiltonian Dynamics

In this section, the equations of motion used in classical MD and the algorithms for integrating these equations are described [16–18].
The trajectory of a system can be followed with the help of Hamiltonian dynamics. Hamiltonian dynamics was introduced in 1834 as a generalization of Newton's equations for a point particle in a force field; virtually all of the fundamental models in physics are described by such dynamics.
The Lagrangian of a system is defined as

$$\mathcal{L} = \mathcal{T} - \mathcal{V} \tag{1.1}$$

where $\mathcal{T}$ is the total kinetic energy and $\mathcal{V}$ is the total potential energy. Given a Lagrangian $\mathcal{L}$, it is possible to define the Hamiltonian, $\mathcal{H}$, as

$$\mathcal{H}(\mathbf{q}, \dot{\mathbf{q}}, t) = \sum_{i=1}^{n} (\dot{q}_i p_i) - \mathcal{L}(\mathbf{q}, \dot{\mathbf{q}}, t) \tag{1.2}$$

where q_i is a generalized coordinate, p_i is a generalized momentum, which for most of the systems studied correspond to position r_i and momentum $p_i = m_i v_i$. If $\mathcal{L}$ is a sum of homogeneous functions (i.e., no products of different degrees) in generalized velocities of degrees 0, 1, 2 and the equations defining the generalized coordinates are not functions of time, then the Hamiltonian can be expressed as follows:

$$\mathcal{H} = \mathcal{T} + \mathcal{V} = \mathcal{E} \tag{1.3}$$

where $\mathcal{T}$ is the kinetic energy, $\mathcal{V}$ is the potential energy, and $\mathcal{E}$ is the total energy of the system. As p_i and q_i are conjugate variables, an Hamiltonian system has always an even number of dimensions $2N$, therefore N integrals are necessary to specify a trajectory, following Hamilton's equations:

$$\dot{q}_i = \frac{\partial \mathcal{H}}{\partial p_i} \tag{1.4}$$

$$\dot{p}_i = -\frac{\partial \mathcal{H}}{\partial q_i} \tag{1.5}$$

$$\dot{\mathcal{H}} = -\frac{\partial \mathcal{L}}{\partial t} \tag{1.6}$$

These equations have fixed points when

$$\dot{q}_i = \frac{\partial \mathcal{H}}{\partial p_i} = 0 \tag{1.7}$$

$$\dot{p}_i = -\frac{\partial \mathcal{H}}{\partial q_i} = 0 \tag{1.8}$$

In other words an equilibrium point is found when $\nabla \mathcal{H} = 0$, i.e. when the system reaches a critical point of the total energy function $\mathcal{H}$.

A Hamiltonian system is *conservative*, as the energy is invariant along the trajectories:

$$\begin{aligned} \frac{d\mathcal{H}}{dt} &= \sum_{i=1}^{n} \left(\frac{\partial \mathcal{H}}{\partial q_i} \frac{\partial q_i}{\partial t} + \frac{\partial \mathcal{H}}{\partial p_i} \frac{\partial p_i}{\partial t} \right) \\ &= \sum_{i=1}^{n} \left(\frac{\partial \mathcal{H}}{\partial q_i} \frac{\partial \mathcal{H}}{\partial p_i} - \frac{\partial \mathcal{H}}{\partial p_i} \frac{\partial \mathcal{H}}{\partial q_i} \right) \\ &= 0 \end{aligned} \tag{1.9}$$

It can also be proved that Hamiltonian flows are volume preserving. From these properties of the Hamiltonian systems, it follows that the trajectories obtained belongs to the microcanonical (NVE) ensemble.

1.2.2 Integration of the equations of motion

A system of equations given by Equation 1.2, together with initial coordinates and velocities, constitutes an initial-value problem. Consequently, one can use a variety of algorithms for numerical solution of the initial-value problem to integrate the equations of motion; the predictor-corrector Gear method [19] is often applied as a general purpose algorithm in this field. An undesirable feature of the general purpose algorithms is that they usually require high order time derivatives to work with good accuracy. Because of

the demand for low computational cost and high accuracy, a variety of specific integrators have been designed for MD algorithms. Of these, the Verlet-type algorithms (the Verlet, velocity-Verlet, and the leap-frog algorithm) are the most common [18]; all three of these algorithms are mathematically equivalent. Their most important property is the conservation of a slightly perturbed original Hamiltonian (the shadow Hamiltonian); in other words, when the nonconservative forces are not present, the total energy oscillates about a value close to the initial energy and does not drift from the initial value, being the magnitude of the oscillations increased with increasing the time step Δt.

Solving the equations of motion requires a numerical integration of the differential equations. The integration is typically done discretizing the variable t in small timesteps dt using finite difference methods. These are explicit methods, based on a Taylor expansion of the positions and momenta at a time $t + dt$ (eq. 1.10), that use the state of the system at a time t to predict the state at a time $t + dt$:

$$\begin{aligned} \mathbf{r}(t+dt) &= \mathbf{r}(t) + \dot{\mathbf{r}}(t)dt \ + \ \frac{\ddot{\mathbf{r}}(t)}{2}dt^2 \ + \ ... \\ &= \mathbf{r}(t) + \mathbf{v}(t)dt \ + \ \frac{\mathbf{f}(t)}{2m}dt^2 \ + \ ... \end{aligned} \tag{1.10}$$

The most common integration algorithm in Molecular Dynamics is the *Verlet integrator* [20], which is based on the addition of two Taylor expansions in time, one forward and one backward:

$$\mathbf{r}(t+dt) = \mathbf{r}(t) + \mathbf{v}(t)dt \ + \ \frac{\mathbf{f}(t)}{2m}dt^2 \ + \ ... \tag{1.11}$$

$$\mathbf{r}(t-dt) = \mathbf{r}(t) - \mathbf{v}(t)dt \ + \ \frac{\mathbf{f}(t)}{2m}dt^2 \ + \ ... \tag{1.12}$$

$$\mathbf{r}(t+dt) = 2\mathbf{r}(t) + \mathbf{r}(t-dt) \ + \ \frac{\mathbf{f}(t)}{m}dt^2 \ + \ O(dt^4) \tag{1.13}$$

This integration does not require the velocities. These are nevertheless required for the calculation of the energy and can be extimated with the formula obtained subtracting the expansion reported above:

$$\mathbf{v}(t) \ = \ [\mathbf{r}(t+dt) + \mathbf{r}(t-dt)]/(2dt) \tag{1.14}$$

Furthermore, only a single evaluation of forces is required at each time step, the formulation is time reversible, but numerical errors are rather large. This is due to the

addition of an $O(dt^0)$ term $[2\mathbf{r}(t) + \mathbf{r}(t - dt)]$ to an $O(dt^2)$ term $[\frac{\mathbf{f}(t)}{m} dt^2]$.
In order to be stable the integration algorithm, the value of the time step t must be an order of magnitude smaller than the fastest motions of the system. Typically, this motion is the vibration of a bond that involves a hydrogen atom with a period of the order of 10 fs, and consequently the time step is of the order of 1 fs when explicit solvent is used. When implicit solvent is used, the time step can be larger, from 2 to 5 fs. This is much less than the timescale of the fastest biochemically important motions such as helix formation, which takes a fraction of a microsecond, or folding of the fastest α helical proteins, which takes several microseconds [21]. In one option, known as the variable step method [22, 23], the time step is reduced when hot events result in occasional significant variation of forces, but this violates time reversibility and energy conservation. The correct procedure is to use the time-split algorithms [24], which are an extension of the basic Verlet-type algorithms. In these algorithms, the forces are divided into fast-varying ones that are local (as, e.g., the bond-stretching forces) and, consequently, inexpensive to evaluate and slow-varying forces that are nonlocal forces and expensive to evaluate. Integration is carried out with a large time step for the slow forces and an integer fraction of the large time step for the fast-varying forces. Such a procedure enables the use of up to a large 20 fs time step at only a moderate increase of the computational cost [22]. One can achieve further effective increase of the timescale by constraining the valence geometry of the solvent molecules (the SHAKE [25], RATTLE [26], and LINCS [27] algorithms) and, yet further, by using torsional angle dynamics [23] and rigid-body dynamics [18] in which elements of structure (e.g., α helical segments) are considered fixed. The use of simplified protein models enables one to increase the timescale further because of averaging out fast motions that are not present at the coarse-grained level [22].

1.2.3 Constant Temperature Molecular Dynamics

As seen before, Hamilton equations lead to a trajectory in the microcanonical (NVE) ensemble. In reality, protein folding occurs in systems coupled to a temperature bath, and consequently the solution of the equations of motion to give the canonical (NVT) or isothermal-isobaric (NPT) ensembles is required. This remark pertains to all simulations regardless of whether the solvent is considered explicitly or implicitly. To run simulations in other ensembles, some tricks of the trade, or some modification of the Lagrangian are needed. Simulating a system at costant temperature has the thermodynamical meaning

of bringing the system into thermal contact with a large heat bath. In any case the simulation temperature can be calculated from the average kinetic energy of the system $\langle K \rangle$:

$$\frac{3}{2}NkT = \langle K \rangle \tag{1.15}$$

$$T = \frac{2}{3kN}\langle K \rangle$$

$$= \frac{1}{3kN}\langle \sum m_i v_i^2 \rangle \tag{1.16}$$

The simplest way to simulate at constant temperature is to rescale all the velocities to keep kinetic energy constant. It is a very crude approach that consists in a periodic scaling of all the particle velocities of a factor $(\frac{T_{ext}}{T})^{\frac{1}{2}}$, where T is the instantaneous system temperature, calculated from equation 1.16, and T_{ext} is the temperature of the thermal bath. This technique is also often used to equilibrate the system during the first few hundred MD steps before the production run starts and data are collected.

A more gentle way, known as Berendsen or weak-coupling thermostat [28], is to use a factor that depends on the deviation of the instantaneous temperature from the average value T_0. At each time step velocities are scaled by the factor λ:

$$\lambda^2 = 1 + \frac{dt}{\tau_T}(\frac{T}{T_0} - 1) \tag{1.17}$$

where dt is the MD time step, and τ_T is a parameter that defines the strength of the coupling with the thermostat and has the dimension of a time. Both methods do not reproduce canonical ensemble, as the condition of constant average kinetic energy does not correspond to the condition of constant temperature, i.e. the fluctuations of the temperature and kinetic energy follow different laws. Therefore these methods lead to trajectories whose average values correspond to the ones of the canonical enseble, but whose fluctuations do not [18,29]. On the contrary, Nosé Hoover method [30,31] actually generates the canonical ensemble making use of the extendend Lagragian technique: the coupling with an external degree of freedom is performed by adding additional coordinates to the classical Lagrangian (eq. 1.1). The idea is to introduce an additional degree of freedom η, describing the external bath, and a corresponding velocity $\xi = \dot{\eta}$. The additional kinetic and potential energy terms coupled to the particles momenta, respectively, $Q\eta^2/2$, $\eta \sum (\frac{p_i^2}{2m_i} - 3k_b T_0)$, where the quantity Q is the thermostat mass, are

added to the Hamiltonian. Using Hamilton equations the following equations of motion is obtained:

$$\dot{r}_i = \frac{p_i}{m_i} \tag{1.18}$$

$$\dot{p}_i = F_i - \xi p_i \tag{1.19}$$

$$\dot{\xi} = 1/Q \sum (\frac{p_i^2}{2m_i} - 3k_b T_0) \tag{1.20}$$

The whole system, that contains all "real" degrees of freedom plus η, is conservative and obeys Liouville equation. It can be shown by direct substitution that the canonical distribution $p = \exp(-\beta(K+U))$, being K the Boltzmann constant, T the temperature of the system and β equal to $\frac{1}{KT}$, is a stationary, time independent solution. Therefore, configurations sampled by this algorithm represent canonical ensemble. In contrast to the former thermostats, this is an integral thermostat, with the instantaneous values of η and ξ depending on all previous states of the system. This thermostat may be preferable even when explicit solvent is considered, because it results in more uniform distribution of temperature between the solute and the solvent.

1.2.4 Constant Pressure Molecular Dynamics

It is also possible to run simulations at constant pressure, in the NPT and NPH ensembles. The system pressure tensor $\mathbf{\Pi}$ is measured as sum of the kinetic energy contribution (ideal gas contribution, always positive) plus the interparticle energy contribution (the so called virial tensor,$\mathbf{W}$). The pressure P is then calculated from the trace of the pressure tensor:

$$\mathbf{W} = \sum_{i=1}^{N} \mathbf{r}_i \otimes \mathbf{f}_i \tag{1.21}$$

$$\mathbf{\Pi} = \frac{1}{V} \left[\sum_{i}^{N} m_i (v_i \otimes v_i) + \mathbf{W} \right] \tag{1.22}$$

$$P = \frac{1}{3} Tr(\mathbf{\Pi}) \tag{1.23}$$

If a cutoff scheme is used, the virial must be calculated from pairwise forces instead of being calculated from the total force acting on each particle (see, for example reference [32]):

$$\mathbf{W} = \sum_{i=1}^{N} \sum_{j>i} \mathbf{r}_{ij} \otimes \mathbf{f}_{ij} \tag{1.24}$$

The barostat formulations generally mimic the ones derived for thermostats: in particular, the most used barostats are again the weak-coupling barostat and the more elegant Parrinello–Rahman [33].

1.3 Force fields for molecular simulations

1.3.1 Molecular Mechanics

Theoretical investigations of molecules permit the study of the relationships between structure, function and dynamics at atomic level. Since the majority of problems that one would like to address in complex chemical systems involves systems composed by many atoms, it is not yet feasible to treat these systems using quantum mechanics. The answer to the need of high detail at low computational cost is Molecular Mechanics (MM), a technique which uses classical mechanics to analyze the structure and dynamics of molecular systems. Within this approximation, the molecule is treated at the atomic level, i.e. the electrons are not treated explicitly. The energy and the forces are calculated through a certain potential energy function, or *force field* (FF), which is translationally and rotationally invariant and depends on the relative positions of the atoms and on a small number of parameters that have been determined either experimentally or via quantum mechanical calculations. In this way, given a particular conformation or configuration, the energy of the system can be calculated straightforwardly. The interatomic interactions are generally described by simple two- and more rarely three- and four-body potential energy functions. This classical force field-based approach is a good simplification over quantum chemistry, which describes systems in terms of nuclei, electrons and orbitals. This simplicity allows molecular mechanics to be applied to much larger systems than those can be studied by ab initio methods. Current generation force fields provide a reasonably good compromise between accuracy and computational efficiency. They are often calibrated to experimental results and quan-

tum mechanical calculations of small model compounds. The development of parameter sets is a very laborious task, requiring extensive optimization. This is an area of continuing research and many groups have been working over the past two decades to derive functional forms and parameters for potential energy functions of general applicability to biological molecules. Traditionally, the potential forces are calculated using empirical all atom potential functions. Among them, the most commonly used potential energy functions for molecular dynamics simulations, especially of biological systems, are the CHARMM (Chemistry at Harvard Molecular Mechanics) [34], AMBER (assisted model building with energy refinement) [35–37], OPLS-AA (mixed Amber and OPLS) [38], GROMOS (Groningen molecular simulation) [39], and CVFF (consistent valence force field) [40], which include the solvent either explicitly or as a continuum (implicit solvent treatment).

Explicit inclusion of water molecules provides, as realistically as possible, the kinetic and thermodynamic properties of the protein folding process and should be preferred to the implicit solvent model because of the key role played by structured water around the protein. Simulations with explicit water are carried out in a periodic box scheme; the box is usually rectangular, but other shapes are also possible [16]. A less common treatment is to perform simulations in a thin layer of water around a protein molecule restrained with a weak harmonic potential [16]. There are currently a number of water models used in MD simulations. These include the ST2 model of Stillinger [41], the SPC model of Berendsen et al. [42], and Jorgensen's TIP3P, TIP4P, and TIP5P models [43]. These models were parameterized assuming that a cutoff is applied to nonbonded interactions, but they are often used with Ewald summation to treat long-range electrostatics. Horn et al. [44] recently developed an extension of the TIP4P model to be used with Ewald summation termed TIP4P-Ew. All these models treat water as a rigid molecule. Although bond stretching and bond-angle bending [45], or polarization effects and many-body interactions [46], have been introduced into water models, they involve a large increase of computational expense, which has limited their use as widely as the SPC or TIP models. The water models are usually parameterized at a single temperature (298 K) and therefore do not correctly capture the temperature dependence of properties such as the solvent density or diffusion coefficients [44]. The presence of water molecules in the system dramatically increases the number of degrees of freedom (typically by more than 1000). Because of this limitation, along with the small values of the time step in integrating the equations of motion (of the order of femtoseconds),

explicit-solvent all-atom MD algorithms can simulate events in the range of 10^{-9} s to 10^{-8} s for typical proteins and 10^{-6} s for very small proteins [10,21]. These timescales are at least one order of magnitude smaller than the folding times of proteins [47]. The most impressive and the longest explicit-solvent ab initio canonical MD simulation starting from unfolded conformations is one by Duan and Kollman [12] on the villin headpiece. They observed conformations with significant resemblance to the native state in a 1μs run. However, their simulation fell significantly short of the folding time for this protein, which is $\sim 5\mu$s. Therefore, at present, explicit solvent MD by itself is not capable of simulating the folding pathways of proteins in real time, except for very small proteins. However, it has been combined successfully with other search methods in some interesting and ingenious algorithms to study energy landscapes and folding pathways (see section 1.3).

1.3.2 The potential

The typical potential energy function is a sum of diverse bonded and non–bonded contributions, each of them containing a sum over the atoms or groups of atoms. As example, the expression of the AMBER force field is reported; the variables comparing here are distances r_{ij}, angles θ_{ijk} and dihedral angles ϕ_{ijkl}; all the other terms are the force field parameters.

$$U_{\text{total}} = U_{\text{bonds}} + U_{\text{angle}} + U_{\text{dihed}} + U_{\text{LJ}} + U_{\text{charge}} \tag{1.25}$$

$$U_{\text{bonds}} = \sum_{\text{bonds}} K_r^{t_i t_j}(r_{ij} - r_{eq}^{t_i t_j})^2 \tag{1.26}$$

$$U_{\text{angles}} = \sum_{\text{angle}} K_\theta^{t_i t_j t_k}(\theta_{ijk} - \theta_{eq}^{t_i t_j t_k})^2 \tag{1.27}$$

$$U_{\text{dihed}} = \sum_{\text{dihed}} V_\phi^{t_i t_j t_k t_l}[1 + \cos(n^{t_i t_j t_k t_l}\phi_{ijkl} - \gamma^{t_i t_j t_k t_l})] \tag{1.28}$$

$$U_{\text{LJ}} = 4\sum_{i<j} f_{LJ}^{1,4}\ \epsilon_{t_i t_j}\left[\left(\frac{\sigma_{t_i t_j}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{t_i t_j}}{r_{ij}}\right)^{6}\right] \tag{1.29}$$

$$\text{where} \quad \epsilon_{t_i t_j} = (\epsilon_{t_i}\epsilon_{t_j})^{\frac{1}{2}}\ ,\ \sigma^{t_i t_j} = \frac{\sigma_{t_i} + \sigma_{t_j}}{2} \tag{1.30}$$

$$U_{\text{charge}} = \sum_{i<j} f_q^{1,4}\ \frac{q_i q_j}{r_{ij}} \tag{1.31}$$

The first *bonds* sum is over bonds between atom pairs; the second sum is over bond angles defined by three atoms; the third sum is over the four atom sets defining each dihedral angle. In the *non-bonded* interactions (LJ and electrostatics), the summation is over atoms couples i and j, where $i < j$ simply ensures that each interaction is counted only once; generally, atoms separated by one or two bonds are excluded from the non-bonded sum, and those separated by three bonds, *1-4 interactions*, may have non-bonded interactions reduced by a multiplicative scale factor ($f_{LJ}^{1,4}$, $f_q^{1,4}$), which for Amber force field are respectively 1/2 and 5/6. The force fields are based on the concept of atom types (t_i), i.e. a set of parameters defined for a chemical type of atom that can possibly be used in the MM description of a class of molecules, rather than for a single molecular species (e.g. methylene carbon or aromatic carbon are typical atom types, see Table 1.1 for sake of clarity).

1.3.3 Bonded Interactions: Bonds and Angles

This type of interactions has the purpose of describing correctly first the equilibrium geometry of the molecule. As a convention, the bonded energy minimum is set to zero, so that the bonded energy is always positive.

The standard way to approximate the potential energy for a bond in molecular mechanics is to use a Hooke's law term:

$$U_{\text{bond,Hooke}} = K^{t_i,t_j}(r_{ij} - r_{eq}^{t_i,t_j})^2 \tag{1.32}$$

where r_{ij} is the distance between the two bonded atoms i, j; r_{eq} is the equilibrium bond length and K is a force constant. This kind of approach does not attempt to reflect the energy of the bond formation, it only seeks to reflect the energy difference on a small motion about the equilibrium value. A much more accurate representation is based on the application of the Morse potential which has an anharmonic potential energy well (Figure 1.1).

$$U_{\text{bond,Morse}} = D_e^{t_i,t_j}[1 - e^{-a(r_{ij} - r_{eq}^{t_i,t_j})}]^2 \tag{1.33}$$

where D_e is the "equilibrium" dissociation energy of the molecule (measured from the potential minumum) and a is a parameter controlling the width of the potential well. This is not widely used for applications in which the intention is to look at structural details but it is necessary if one is interested in spectroscopic applications. A bond

angle among atoms A-B-C is defined as the angle between the bonds A-B and B-C. As bond angles, in a similar manner to bond lenghts, are found, experimentally and theoretically, to vary around a single value, it is sufficient in most applications to use a harmonic representation for them:

$$U_{\text{angle}} = K_{\theta}^{t_i t_j t_k}(\theta_{ijk} - \theta_{eq}^{t_i t_j t_k})^2 \tag{1.34}$$

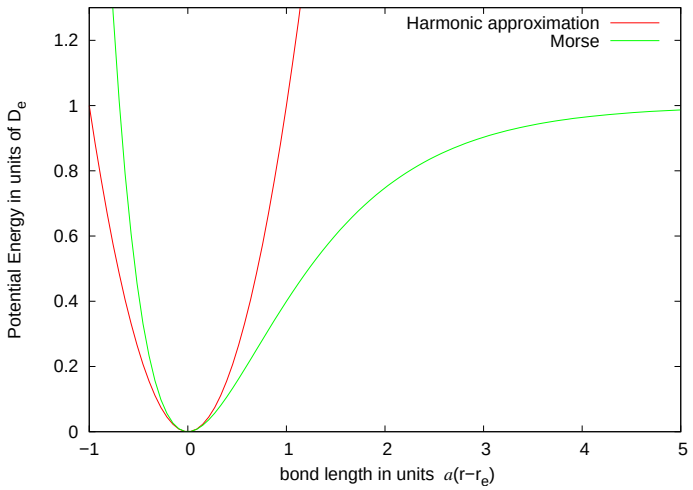


Figure 1.1: Harmonic fit (red line) to the Morse potential (green line) around the equilibrium position. Hooke's law can be used as a useful approximation around the minimum of the Morse curve, namely when equilibrium positions are reached.

1.3.4 Torsion angles

The torsion angles are distinguished in two brands: the dihedral or proper torsion angles and the improper torsion angles. Formally the dihedral angle (also known as a torsion angle) among four atoms A-B-C-D is defined as the angle between the planes ABC and BCD. The angle can vary from -180 to 180 degrees, and its sign is taken as the sign of the scalar product $(\mathbf{n}_{ABC} \times \mathbf{n}_{BCD}) \cdot \mathbf{r}_{BC}$, where the $\mathbf{n}$ are the normal to the planes. The standard functional form to represent the potential energy for a torsional rotation was introduced by Pitzer [48]:

$$U_{\text{dihed}} = V_{\phi}[1 + \cos(n\ \phi_{ijkl} - \gamma)] \tag{1.35}$$

where V_ϕ is the half energy barrier to rotation, n the number of maxima (or minima) in one full rotation and γ determines the angular phase. Barriers for dihedral angle rotation can be attributed to the exchange interaction of electrons in adjacent bonds and to steric effects. The Pitzer potential is insufficient to give a full represention of the energy barriers of dihedral angle change. Modern potential energy functions normally model the dependence of the energy on dihedral angle change by a combination of truncated Fourier series or a sum of Pitzer terms with different non-bonded effects. Improper torsions are so named because the atoms involved are not serially bonded; rather they are branched. Improper dihedral potentials are sometimes necessary to reproduce out-of-plane bending frequencies, i.e. they keep four atoms properly trigonal planar for a two-fold torsional potential. They are additionally used in the united-atom force field model when a carbon with an implicit hydrogen is a chiral center, thus preventing an unphysical inversion of the chirality.

1.3.5 Charges

Electrostatic interactions are of fundamental importance in determining the intermolecular interactions. The most common approach to include their contribution in a simulation is to place a charge at each atomic centre (nucleus). The charge can take a fraction of an electron and can be positive or negative. The electrostatic attraction or repulsion between two charges is described by Coulomb's law:

$$U_{\text{charge}} = \frac{1}{4\pi\epsilon_0\epsilon_r}\frac{q_i q_j}{r_{ij}} \tag{1.36}$$

where q_i and q_j are the atoms partial charges, r_{ij} is the distance separating the atom centres, ϵ_0 is the permittivity of free space and ϵ_r is the relative dielectric coefficient of the medium between the charges (often taken as one). Using partial charges at nuclear centres is the crudest effective abstraction. To obtain a more accurate representation two approaches are commonly used: the first is to add dipole, quadrupole and higher moments to the nuclear centres; the second is to introduce further non-nuclear centres. This is commonly done to represent the anisotropy in potential caused by lone pairs on oxygen atoms [49]. In many respects, electrostatic interactions provided the biggest problems to computational studies of protein behavior, as, by their nature, they are long range and dependent on the properties of the surrounding medium.

1.3.6 Lennard–Jones

The equilibrium distance between two proximal atomic centres is determined by a trade off between an attractive dispersion force and a core-repulsion force that reflects electrostatic repulsion.

The Lennard-Jones potential represents a successful effort in reproducing this balance with a simple expression:

$$U_{\mathrm{LJ}} \;=\; 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right] \;=\; \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}} \tag{1.37}$$

where σ is the contact distance (where $U_{\mathrm{LJ}}(\sigma) = 0$) and ϵ is the well depth (where $\partial U_{\mathrm{LJ}}/\partial r_{ij} = 0$). For simplicity, the Lennard-Jones forces are typically modeled as effectively pair-wise additive, and the rules to calculate the mixing parameters for couples of different atom types, are simple as well:

$$\begin{aligned} \epsilon_{ij} &= (\epsilon_i \epsilon_j)^{\frac{1}{2}} \\ \sigma^{ij} &= \frac{(\sigma_{t_i} + \sigma_{t_j})}{2} \end{aligned} \tag{1.38}$$

The term r_{ij}^{-12}, dominating at short distance, models the repulsion between atoms when they are brought very close to each other. Its physical origin is related to the Pauli principle: when the electronic clouds surrounding the atoms starts to overlap, the energy of the system increases abruptly. The exponent 12 was chosen exclusively on a practical basis, as it is particularly easy to compute.

1.3.7 Finite size effects

The finite–size of the simulated sample introduces systematic deviations from bulk (infinite) behavior. In order to reduce these finite size effects, it is usually employed the common artefact of periodic boundary conditions (PBC). In PBC the simulation box is replicated in all directions to form an infinite lattice; in this way, the volume of interaction around each particle has the same geometry as the sample cell. In the course of the simulation, as a molecule moves in the original box, its periodic images in each of the neighbouring boxes move in the same way. Thus, as a molecule leaves the central box, one of its images will enter through the opposite face. In this way, the system does

not present free surfaces, even if an additional spurious periodic correlation between the particles was introduced. In the case of a short range intermolecular potential, this does not constitute a problem; indeed if the range of the molecular interaction is less than half side length, the central box comprises all interactions and the *minimum image convention* (MIC) is often used, which is the distance between two different particles i and j and it is taken as the distance between i and the nearest image of j [50]. Thus, every particle i interacts only with the image of another molecule j which is the nearest. Basically, most simulations evaluate potentials using some cutoff scheme for computational efficiency: each particle does not interact with all the nearest images of the other $N-1$ particles, but only with those minimum images contained on a sphere of radius R_c centered at a given particle. It is therefore assumed that the interactions are negligible outside that volume.

1.3.8 The Amber Force Field

As seen before, in the Amber Force Field [35] bond and angles employed are represented by a simple diagonal harmonic expression, the Van der Waals (VDW) interaction are represented by a 6-12 potential, electrostatic interactions are modeled by a Coulombic interactions and dihedral species are represented, in most cases, with a simple set of parameters, often only specified by the two central atoms. Electrostatic and VDW interactions are only calculated between atoms in different molecules or for atoms in the same molecule separated by at least three bonds. Those non-bonded interactions separated by exactly three bonds (*1-4 interactions*) are reduced by the application of a scale factor.

Concerning the dihedral parameters, a 3-fold Fourier component (V_3) for dihedral around C-C is employed, with the exception of the ϕ and ψ dihedrals, for which an additional Fourier components is used to try to reproduce as well as possible the relative energies of the alanyl and glycyl dipeptides.

The VDW parameters are the same for a given atom and hybridization, except the oxygen sp^3, where oxygens in water (OW), alcohol (OH) and ether (OS) have slightly different parameters, due to the fact of a zero VDW radius on hydrogens bound to the oxygen. This implies that an effectively larger σ_i is required for a water oxygen than alcohol than ether. The charges developed for Amber force field are called *RESP charges* for Restrained ElectroStatic Potential fit. The basic idea of electrostatic potential fit

charges is a least squares fitting algorithm, which is used to derive a set of atom-centered point charges, reproducing well the quantum mechanical electrostatic potential of the molecule. In the AMBER charge fitting programs, the potential is evaluated at a large number of points defined by 4 shells of surfaces at 1.4, 1.6, 1.8 and 2.0 times the VDW radii. These distances have been shown to be appropriate for deriving charges reproducing typical intermolecular interactions (energies and distances) and also the dipole moment of the molecule is well represented. The value of the electrostatic potential at each grid point is calculated from the quantum mechanical wavefunction, so that the charges derived using this procedure are basis set dependent, being the 6-31G* mainly used.

Table 1.1: List of AMBER 94 atom types.

atom	type	description
carbon	CT	any sp^3 carbon
	C	any carbonyl sp^2 carbon
	CA	any aromatic sp^2 carbon and C_ϵ of Arg
	CM	any sp^2 carbon, double bonded
	CC	sp^2 aromatic in 5-membered ring with one substituent + next to nitrogen (C_γ in His)
	CV	sp^2 aromatic in 5-membered ring next to carbon and lone pair nitrogen (e.g. C_δ in Hisδ)
	CW	sp^2 aromatic in 5-membered ring next to carbon and NH (e.g. C_δ in Hisϵ and in Trp)
	CR	sp^2 aromatic in 5-membered ring next to to hydrogens (C_γ and C_ϵ in His)
	CB	sp^2 aromatic at junction of 5- and 6-membered rings (C_δ in Trp) and both junction atoms in Ade and Gua
	C^*	sp^2 aromatic in 5-membered ring next to two carbons (e.g. C_γ in Trp)
	CN	sp^2 junction between 5- and 6- membered rings and bonded to CH and NH (C_ϵ in Trp)
nitrogen	N	$sp2$ nitrogen in amides
	NA	sp^2 nitrogen in aromatic rings with hydrogen attached (e.g. protonated His, Gua, Trp)
	NB	sp^2 nitrogen in 5-membered ring with lone pair
	NC	sp^2 nitrogen in 6-membered ring with lone pair (e.g. N3 in purines)
	N2	sp^2 nitrogen of aromatic amines and guanidinium ions
	N3	sp^3 nitrogen
oxygen	OW	sp^3 oxygen in TIP3P water
	OH	sp^3 oxygen in alcohols, tyrosine, and protonated carboxylic acids
	OS	sp^3 oxygen in ethers
	O	sp^2 oxygen in amides
	O2	sp^2 oxygen in anionic acids
hydrogen	H	H attached to N
	HW	H in TIP3P water
	HO	H in alcohols and acids
	HS	H attached to sulfur
	HA	H attached to aromatic carbon
	HC	H attached to aliphatic carbon with no electron-withdrawing substituent
	H1	H attached to aliphatic carbon with one electron-withdrawing substituent
	H2	H attached to aliphatic carbon with two electron-withdrawing substituent
	H3	H attached to aliphatic carbon with three electron-withdrawing substituent
	HP	H attached to carbon directly bonded to formally positive atoms (e.g. C next to NH_3^+ of lysine)

1.4 Some aspects and extensions of Molecular Dynamics

1.4.1 The conformational space sampled in MD simulations

In MD, one usually generates a statistical ensemble and the quantity of interest is the average ensemble (e.g., the average structure of the native protein) that better represents the NMR derived models and therefore a good comprehension of the single structures with respect to their average conformation may be achieved. To obtain a good average ensemble, many trajectories have to be simulated so that the statistical errors owing to insufficient sampling are minimized. More trajectories can be run in a given amount of time and consequently, more reliable folding statistics can be collected when using simplified models of polypeptide chains. For example, using their coarse-grained potential biased toward native secondary structure, Brown and Head-Gordon [51] have calculated the folding pathways, the folding temperature, thermodynamic characteristics of folding, kinetic rate, denatured-state ensemble and transition-state ensemble of protein L by a reduced representation of proteins. Pande and coworkers [52] designed a method based on simulating multiple trajectories at the all-atom level that enables one not only to study folding pathways but also to estimate rate constants. The method is based on the observation that, with the assumption that crossing of a single barrier obeys a single-exponential kinetics, the probability for a system to cross the free-energy barrier for the first time is increased M times if M parallel trajectories are simulated. Sampling the conformational space, often involves to search a good reaction coordinate which follows the evolution of the system. A reaction coordinate is an abstract one-dimensional coordinate that represents progress along a reaction pathway. For more complex reactions (e.g., protein folding), the choice of such a coordinate can be difficult.

A good alternative approach is to use *metadynamics* [53]. This technique consists essentially of a modification of a standard MD simulation in which harmonic restraints are imposed on appropriately selected collective coordinates of the system along with a history-dependent potential. The time-dependent restraint is evolved using the extended Lagrangian method [54]. The history-dependent potential, by summing up Gaussian functions at regular time intervals along the trajectory of the auxiliary variables, disfavors configurations in the space of the reaction coordinates that have already been visited, while at the same time reconstructing the negative of free energy surface (FES)

as a function of the reaction coordinates.

Free energy is often plotted against the corresponding reaction coordinates to illustrate the energy landscape or potential energy surface associated schematically with the reaction. Projecting the trajectories onto one or several reaction coordinates, such as the fraction of native contacts, can produce a landscape that shows a clear difference between the native and the unfolded states. But in general, the folding transitions cannot be projected onto two dimensions without overlap of kinetically distinct conformations. Researchers have achieved accurate projections of simulations onto appropriate reaction coordinates, which agreed with the experiment. For example, Onuchic and colleagues [55] used the Gō model for reversible folding and their results matched experiment. Radhakrishnan and Schlick [56] developed the transition-path sampling method for all-atom MD in which a number of MD trajectories are focused near the conformational-transition path, and they applied it to map out the entire closing conformational profile of RNA polymerase. They found that there is a sequence of conformational checkpoints involving subtle protein-residue motion that may regulate fidelity of the polymerase repair or replication process, thus showing the capability in following protein folding with this approach.

1.4.2 Ab-initio Quantum mechanical Molecular Dynamics

The most reliable method for Ab-initio Quantum mechanical Molecular Dynamics, is the Car Parrinello approach.

The Car-Parrinello Molecular Dynamics [57], better known as *CPMD*, is a type of ab initio (first principles) molecular dynamics, usually employing periodic boundary conditions, planewave basis sets, and DFT. In contrast to classical molecular dynamics wherein the nuclear degrees of freedom are propagated using forces which are calculated at each iteration by approximately solving the classical equations of dynamics, the Car-Parrinello method explicitly introduces the electronic degrees of freedom as (fictitious) dynamical variables, writing an extended Lagrangian for the system which leads to a system of coupled equations of motion for both nucleus and electrons. In this way an explicit electronic minimization at each iteration is not needed: after an initial standard electronic minimization, the fictitious dynamics of the electrons keep them on the electronic ground state corresponding to each new nuclear configuration visited along the dynamics, thus yielding accurate nuclear forces. In order to maintain this adiabatic-

ity condition, it is necessary that the fictitious mass of the electrons is chosen small enough to avoid a significant energy transfer from the nucleus to the electronic degrees of freedom. This small fictitious mass in turn requires that the equations of motion are integrated using a smaller time step than the ones (1-10 fs) commonly used in classical molecular dynamics.

1.4.3 Quantum-Classical Molecular Dynamics

The classical equations of motion are valid when chemical reactions are not involved because the typical amplitudes of motions are much smaller than the cor- responding thermal De Broglie wavelengths. Furthermore, some biological processes (such as oxygen binding to hemoglobin, enzymatic reactions, and the light-induced charge transfer in the photosynthetic reaction centres) involve quantum effects such as a change in chemical bonding, noncovalent intermediates, tunneling of proton and electron and dynamics on electronically excited states that cannot be modeled with the classical formulas. One can handle processes involving proton transfer by introducing a special potential function for the protons exchanged between the proton-acceptor atoms [58]. For a general purpose, a hybrid approach known as QM/MM has been designed [59], in which the system is partitioned into a small core (within which the actual chemical reaction occurs) and the surroundings. The core is treated at the quantum-mechanical level, whereas the surroundings are treated at the classical level. The electrostatic potential from the surroundings contributes to the Hamiltonian of the core part. This approach is particularly suitable for the studying of the active site of an enzyme, which normally involves the coordination of a metal centre and the simultaneous breaking and formation of chemical bonds.

Bibliography

[1] B. J Alder, and T. Wainwright T, *J. Chem. Phys.*, **1957**, *27*, 1208–1209.

[2] B. J. Alder and T. Wainwright, *Molecular dynamics by electronic computer;* I. Prigogine, Ed., New York: Intersciences, 1958.

[3] A Rahman, *Phys. Rev. A.*, **1964**, *2*, 405–411.

[4] A. Rahman and F. H. Stillinger, *J. Chem. Phys.*, **1971**, *55*, 3336–3359.

[5] J. A. McCammon, B. R. Gelin, and M. Karplus, *Nature*, **1977**, *267*, 585–590.

[6] C. L. Brooks III, *J. Mol. Biol.*, **1992**, *227*, 375–380.

[7] A. E. Mark and W. F. van Gunsteren, *Biochemistry*, **1992**, *31*, 7745–7748.

[8] V. Daggett and M. Levitt M., *J. Mol. Biol.*, **1993**, *232*, 600–619.

[9] J. Tirado-Rives and W. L. Jorgensen, *Biochemistry*, **1993**, *32*, 4175–4184.

[10] R. Day and V. Dagget, *Adv. Prot. Chem.*, **2003**, *66*, 373–383.

[11] V. Dagget, *Chem. Rev.*, **2006**, *106*, 1898–1916.

[12] V. Duan and P. A. Kollman, *Science*, **1998**, *282*, 740–744.

[13] V. S. Pande and D. S. Rokhsar, *Proc. Natl. Acad. Sci. USA*, **1999**, *96*, 9062–9067.

[14] D. Roccatano, A. Amadei, A. Di Nola, and H. J. C. Berendsen, *Protein Sci.*, **1999**, *8*, 2130–2143.

[15] J. Tsai, M. Levitt, and D. Baker, *J. Mol. Biol.*, **1999**, *291*, 215–225.

[16] S. A. Adock and J. A. McCammon, *Chem. Rev.*, **2006**, *106*, 1589–1615.

[17] H. A. Scheraga, M. Khalili, and A. Liwo, *Annu. Rev. Phys. Chem.*, **2007**, *58*, 57–83.

[18] D. Frenkel and B. Smit, *Understanding Molecular Simulations: From Algorithms to Applications;* New York: Academic, 2000.

[19] C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations;* Cliffs, N. J., Prentice Hall, 1971.

[20] L. Verlet, *Phys. Rev.*, **1967**, *159*, 98–103.

[21] J. Kubelka, J. Hofrichter, and W. A. Eaton, *Curr. Opin. Struct. Biol.*, **2004**, *14*, 76–88.

[22] M. Khalili, A. Liwo, F. Rakowski, P. Grochowski, and H. A. Scheraga, *J. Phys. Chem. B*, **2005**, *109*, 13785–13797.

[23] K. D. Gibson and H. A. Scheraga, *J. Comput. Chem.*, **1990**, *11*, 468–497.

[24] G. J. Martyna, M. E. Tuckerman, D. J. Tobias, and M. L. Klein, *Mol. Phys.*, **1996**, *87*, 1117–1157.

[25] J. Ryckaert, G. Ciccotti, and H. Berendsen, *J. Comput. Phys.*, **1977**, *27*, 327–341.

[26] H. Andersen, *J. Comput. Phys.*, **1983**, *52*, 24–34.

[27] B. Hess, H. Bekker, H. J. C. Berendsen, and J. Fraajie, *J. Comput. Chem.*, **1997**, *18*, 1463–1472.

[28] H. J. C. Berendsen, J. P. M. Postma, A. Di Nola, and J. R. Haak, *J. Chem. Phys.*, **1984**, *81*, 3684–3690.

[29] D. Fincham and D. M. Heyes, *Adv. Chem. Phys.*, **1985**, *63*, 493–575.

[30] S. A. Nosé, *J. Chem. Phys.*, 1984, *81*, 511–519.

[31] W. G. Hoover, *Phys. Rev. A.*, **1985**, *31*, 1695–1697.

[32] E. Paci and M. Marchi, *J. Phys. Chem.*, **1996**, *104*, 3003–3012.

[33] M. Parrinello and A. Rahman, *J. App. Phys.*, **1981**, *52*, 7182–7190.

[34] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.*, **1983**, *4*, 187–217.

[35] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *J. Amer. Chem. Soc.*, **1995**, *117*, 5179–5197.

[36] J. Wang, P. Cieplak, and P. A. Kollman, *J. Comput. Chem.*, **2000**, *21*, 1049–1074.

[37] J. W. Ponder and D. A. Case, *Adv. Prot. Chem.*, **2003**, *66*, 27–85.

[38] W. L. Jorgensen and N. A. McDonald, *J. Mol. Str. Theochem.*, **1998**, *424*, 145–155.

[39] W. F. van Gunsteren and H. J. C. Berendsen; *Groningen Molecular Simulation (GROMOS) Library Manual;* Biomos, Groningen, **1987**.

[40] C. S. Ewig, T. S. Thacher, and A. T. Hagler, *J. Phys. Chem. B*, **1999**, *103*, 6998–7014.

[41] F. H. Stillinger and A. Rahman, *J. Chem. Phys.*, **1974**, *60*, 1545–1567.

[42] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, *Intermolecular Forces;* B. Pullman, editor, Reidel, Dordrecht, 1981.

[43] W. L. Jorgensen , J. Chandrasekhar, J. D. Madura, R. W. Impy, and M. L. Klein, *J. Chem. Phys.*, **1983**, *79*, 926–935.

[44] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, and T. J. Dick et al., *J. Chem. Phys.*, **2004**, *120*, 9665–9678.

[45] D. M. Ferguson, *J. Comput. Chem.*, **1995**, *16*, 501–511.

[46] M. Sprik and M. L. Klein, *J. Chem. Phys.*, **1988**, *89*, 7556–7560.

[47] J. C. Lee, H. B. Gray, I. J. Chang, and J. R. Winkler, *J. Mol. Biol.*, **2002**, *320*, 159–164.

[48] K. S. Pitzer, *Disc. Faraday Soc.*, **1951**, *107*, 4519–4529.

[49] P. Cieplak, W. D. Cornell, C. Bayly, and P. A. Kollmann, *J. Comput. Chem.*, **1995**, *16*, 1347–1377.

[50] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids;* Oxford University Press, Walton Street, Oxford OX2 6DP, 1989.

[51] S. Brown and T. Head-Gordon, *Protein Sci.*, **2004**, *13*, 958–970.

[52] V. S. Pande, I. Baker, J. Chapman, S. Elmer, and S. Kaliq et al., *Biopolymers*, **2003**, *68*, 91–109.

[53] A. Laio, A., M. Parrinello, *Proc. Natl. Acad. Sci. USA*, **2002**, 99, 12562–12566.

[54] A. Laio, A. Rodriguez-Fortea, F. L. Gervasio, M. Ceccarelli, M. Parrinello, *J. Phys. Chem. B*, **2005**, 109, 6714–6721.

[55] J. N. Onuchic, C. Clementi and H. Nymeyer, *J. Mol. Biol.*, **2000**, *298*, 937–953.

[56] T. Schlick and R. Radhakrishnan, *Proc. Natl. Acad. Sci. USA*, **2004**, *101.*

[57] R. Car and M. Parrinello, *Phys. Rev. Lett.*, **1985**, *55*, 2471-2474.

[58] P. Bala, P. Grochowski, K. Nowinski, B. Lesyng, and J. A. McCammon, *Biochem. J.*, **2000**, *79*, 1253–1262.

[59] M. H. M. Olsson, W. W. Parson, and A. Warshel, *Chem. Rev.*, **2006**, *106*, 1737–1756.

Chapter 2

Secondary structure determination of proteins using local chirality

2.1 An overview on the secondary structure assignment

The analysis and assignment of the secondary structure of proteins (Figure 2.1) is a central problem in biophysics and algorithms designed for this purpose are indispensable tools not only for the assignment and classification of newly derived native protein structures but also for all the computational techniques that aim at structural predictions on the basis of the primary sequences, multiple sequence alignment and related statistical studies of local properties like solvent accessibility and native contacts. The first key contribution in the field was probably that of Ramachandran [1,2], correlating the native distribution of the -N–C_α- and -C_α–C- dihedral angles (ϕ, ψ) (Figure 2.2) of constituent amino acids in a given sequence to the protein secondary structure. The Ramachandran map for a sequence of amino acids is indeed very helpful in the individuation of the existent secondary elements, but it may fail in the case of high conformational flexibility, which leads to non standard backbone angles, in particular for peptides [3, 4]. To date, the most commonly used and authoritative secondary structure determination program is the "Dictionary of Protein Secondary Structures" (DSSP) [5]. The DSSP relies on an algorithm based on hydrogen bond patterns involving the C=O and N-H backbone atoms, neglecting the ϕ and ψ dihedral angles and classifying qualitatively the structure in eight classes. Despite its effectiveness, this choice does not help in the detection of

small deviations of the backbone dihedral angles from the ideal structure, which may be important for the biological function of a protein. Furthermore, the DSSP analysis is known to be error–prone in the exact detection of the edges of a given motif [6]. An improvement which tries to address some of these limitations, such as the absence of description of thermal fluctuations present in experimental structures, is DSSPcont [7], that performs a continuous assignment of secondary structure by calculating weighted averages with different hydrogen bond thresholds.

More effectively, STRIDE [8] considers, in classifying secondary structures, both hydrogen bond patterns and backbone dihedrals. Many other variants and different criteria, like α carbon distances and angles, have been proposed over the years (see e.g. [9–12]), all achieving a high global agreement among them and with the PDB classification (higher than 80%). Notwithstanding this success, further improvement would be important in the case of non-standard conformations strongly departing from ideal backbones (the "twilight zone" [9]), like the ones obtained by NMR experiments, and in particular polyproline II structures [13, 14].

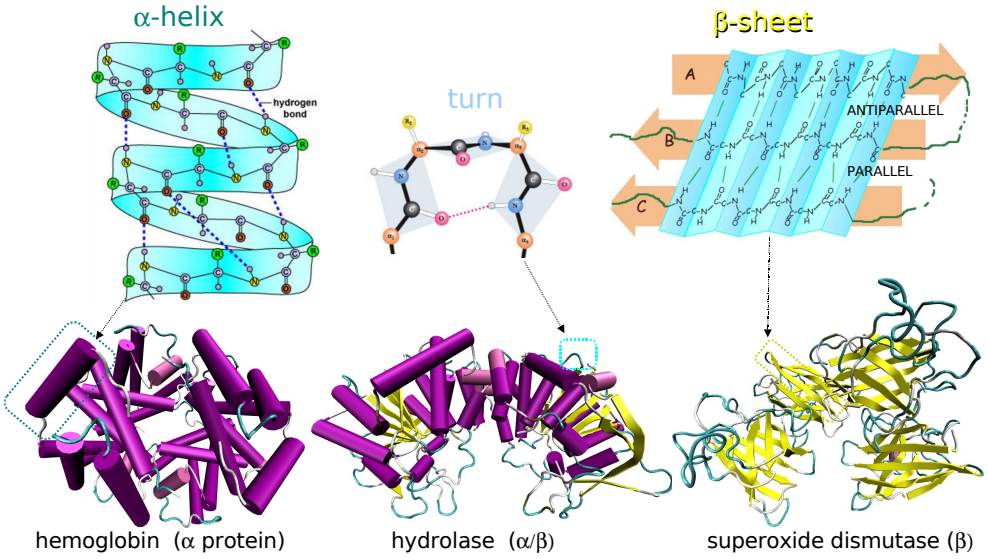


Figure 2.1: VMD visualization [42] of the different secondary structures: α helix (violet), β sheets (yellow) and turn (green)

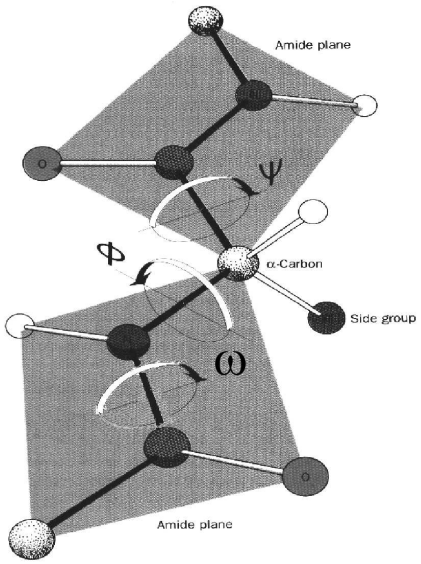


Figure 2.2: Proteins typical backbone dihedrals

2.2 A chirality index for investigating protein secondary structure and their time evolution

Despite the intrinsic chiral nature of amino acids and of many motifs, quantitative measurements of chirality have never been proposed so far as criteria in the field of protein structure analysis, with the notable exception of a variant of Gō–like folding models [15]. The present chapter reports a study as a wish to fill this gap, suggesting the use of a local chirality index that varies continuously as the conformation changes and that aims to provide a quantitative answer to the question "how chiral is a molecule?" [16]. Such an index must be invariant under similarity transformation, change sign upon reflection and be null for symmetric objects [17]. In particular chirality indexes, derived from the disposition in space of the atoms of a given molecule [18–21], have proved to be useful in the attempt of relating molecular structure with macroscopic properties, such as helical pitch [22], helical twisting power [23, 24] and facial diastereoselectivity [25]. In the following, the scaled chiral index of Solymosi et al. [21] is adopted for the analysis of the conformation of ideal backbones and real proteins, showing that local symmetry measurements can actually give reliable information of protein secondary structure.

2.2.1 Chirality calculation on ideal structures

A simple indicator of the conformational chirality of a molecule can be written down as a pseudoscalar combination of three molecule fixed vectors, analogously to the calculation of a dihedral angle. The idea of calculating molecular chirality from atomic coordinates is akin to a generalization of simple models of optical activity, in which a third rank tensor based on dipolar interaction products appears. In that case (see ref. [18] for details) the vectors are related to electronic transitions, but defining the tensor as a purely geometrical entity and reducing it to second-rank on the basis of symmetry arguments, Osipov et al. [18] identified in its trace a pseudoscalar quantity useful for determining molecular handedness. A scaled version of this index was subsequently introduced to facilitate the comparison between molecules of different size [21], leading to the following expression:

$$G = \frac{4!}{3N^4} \sum_{\substack{\text{all permutations of} \\ i,j,k,l=1\ldots N}} w_i w_j w_k w_l \frac{[(\mathbf{r}_{ij} \times \mathbf{r}_{kl}) \cdot \mathbf{r}_{il}](\mathbf{r}_{ij} \cdot \mathbf{r}_{jk})(\mathbf{r}_{jk} \cdot \mathbf{r}_{kl})}{(r_{ij} r_{jk} r_{kl})^n \; r_{il}^m}, \tag{2.1}$$

where i, j, k, l are four of the N atoms belonging to the molecule, $\mathbf{r}_{ab}$ are interatomic distance vectors, w_i, w_j, w_k, w_l are suitably chosen weights for each atom, and n and m are arbitrary integers. This index is commonly employed in a dilatation–invariant form with $n = 2$ and $m = 1$, [21–25], while the weights are set to unity (dimensionless form) or to atomic masses (recalling the Cahn–Ingold–Prelog rules).

To apply this index to the analysis of protein secondary structures, some adjustments are necessary. First, since the structural motifs represent a local property of a small group of amino acids, it is not very meaningful to consider in the calculation the chirality between all possible sets, getting a single value for the whole protein as in equation 2.1. Thus, it was decided to focus only on backbone atoms (N, C_α and C) and to calculate the chirality index for sequences of connected atoms of length N_a (see Figure 2.3). Secondly, a cutoff radius was introduced in eq. 2.2, to avoid the computation of unnecessary long-range terms, that give a negligible contribution to the overall chirality.

$$G^{\mathrm{a,N_a}} = \frac{4!}{3\mathrm{N_a^4}} \sum_{\substack{\text{all permutations} \\ \text{of } i,j,k,l}} \begin{cases} \dfrac{[(\mathbf{r}_{ij}\times\mathbf{r}_{kl})\cdot\mathbf{r}_{il}](\mathbf{r}_{ij}\cdot\mathbf{r}_{jk})(\mathbf{r}_{jk}\cdot\mathbf{r}_{kl})}{(r_{ij}r_{jk}r_{kl})^2 r_{il}} & \text{if } r_{ij},r_{kl},r_{il},r_{jk}<r_c, \text{ and } a\leq i,j,k,l\leq \mathrm{N_a}+\mathrm{a}-1 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

Considering the secondary structure a local geometry–dependent property of a small number of connected amino acids, the variation of the average $G^{\mathrm{N_a}} = \langle G^{a,N_a} \rangle$, was studied as function of the number of backbone atoms N_a, choosing the values of this parameter and of the cutoff distance r_c that maximize the local sensitivity for ideal backbones composed of 40 residues. For this purpose, the cutoff distance was increased until the stability of $G^{\mathrm{N_a}}$ values was achieved, as obtained for r_c greater than 10 Å (see Figure 2.4). In practice a cutoff of 12 Å was chosen, that is appropriate for an extension of the analysis to side chain atoms and should comprise all the possible amino acid native contacts [26]. The value of N_a which allows the best differentiation of the secondary structures is 15, corresponding to five consecutive residues, as noticed from Figure 2.4.

In building ideal secondary structures, it has to be taken into account binary or quaternary periodicity on the backbone angles (ϕ,ψ): ($-67°$, $-41°$) for α helix [27], ($-49°$, $-26°$) for 3_{10} helix [28], ($-67°$, $-59°$) for π helix [27], ($-60°$, $-30°$, $-90°$, $0°$) for type I β turns [29], ($-75°$, $147°$) for PPII helix [30] and ($-130°$, $130°$) for sheets regions [31], while the ω angles can be fixed to the *trans* value of $180°$. Although Type I β turn

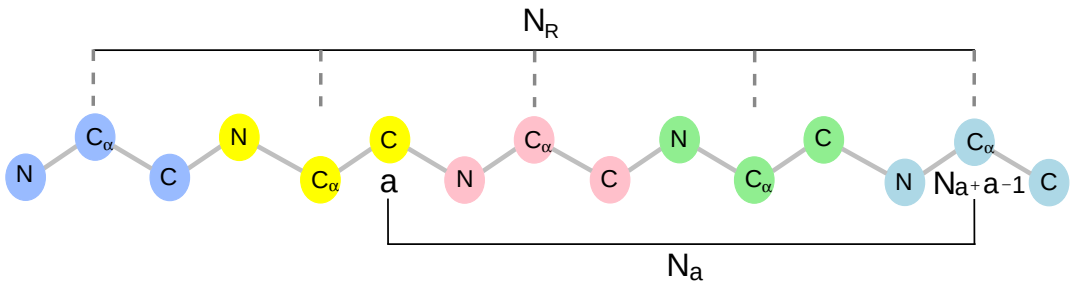


Figure 2.3: A backbone composed of 15 atoms (N_R=5 residues). For N_a=9, the sequence of atoms contributing to the calculation of G^{a,N_a} is indicated, starting from atom $a = 6$.

conformation is not periodic in proteins, involving generally only 4 consecutive residues, it was considered periodic for ease of comparison with the other motifs.

In Figure 2.5 the behavior of the G index for ideal structures along the backbone, calculated with $N_a = 15$ and $r_c = 12$ Å, is reported. Different patterns are clearly distinguishable: in particular, the right handed α helix, type I β turn and 3_{10} helix possess negative chirality index values, which exhibit the correct periodicity when moving along the backbone. Furthermore, the left handed helix of poly-L-proline II shows a positive sign of chirality index, in accord with its opposite handedness with respect to the other helices. The β sheets structure, having a flat shape and symmetric ϕ and ψ dihedrals, shows a chirality index close to zero, as well as the π helix, which shows negative values approaching zero (its chirality is low, as it possesses ϕ and ψ angles respectively $-67°$, $-59°$). In summary, the various important motifs can be all assigned and differentiated on the basis of their intrinsic chirality.

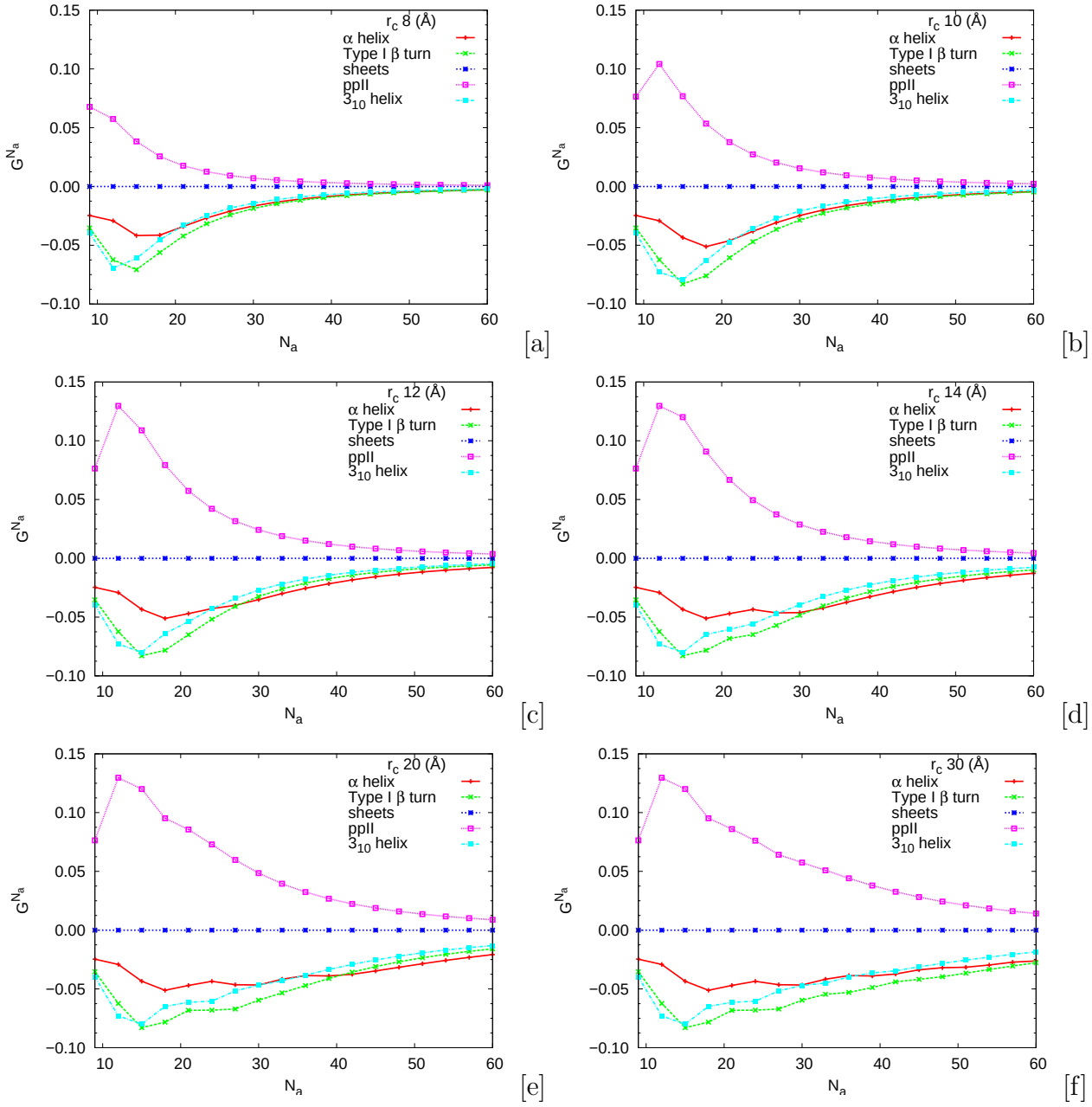


Figure 2.4: Average chiral index on the overall backbone as a function of number of fragments considered for the chirality calculation and using different distances cutoff.

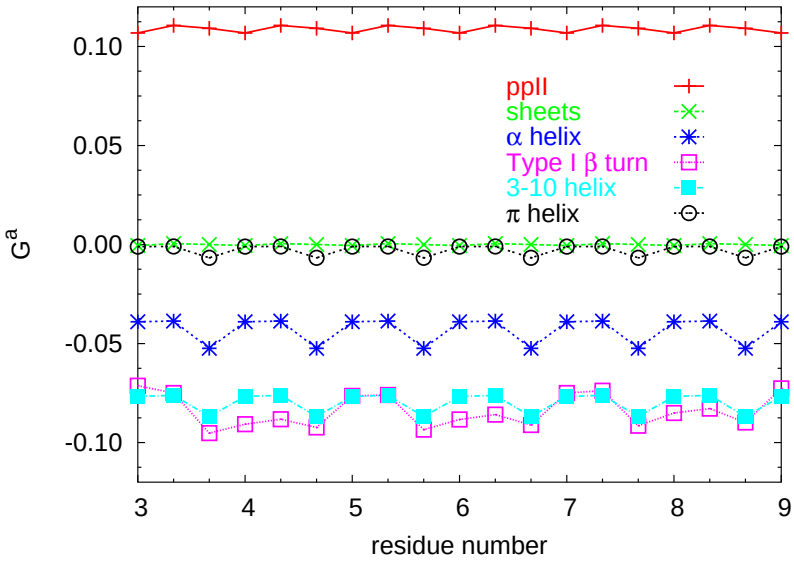


Figure 2.5: Chirality index, G along the backbone for different secondary ideal structures. The cutoff used is r_C=12 Å and N_a=15 atoms was considered.

2.2.2 Chirality of crystalline protein structures

After this preliminary study, a set of seven real protein structures, collected from the Protein Data Bank, containing the most important structural motifs, were analyzed. The chain A of hemoglobin (pdb code 2MHB), a globin representative α protein, and again for helix structures, the avian prion globular domain (pdb code 1U3M), which contains three α helices [32] and ubiquitin (pdb code 1D3Z), with one α helix, were analyzed. Concerning turn and sheet regions, the chain A of immunoglobulin antigen (pdb code 1REI), previously included in the DSSP data set [5], and serine protease, a turn rich protein (pdb code 1HPJ), were studied. Model peptide systems for 3_{10} helix (pdb code 1LB0) and poly-L-proline (pdb code 1JMQ 51-60) were also taken from the protein databank.

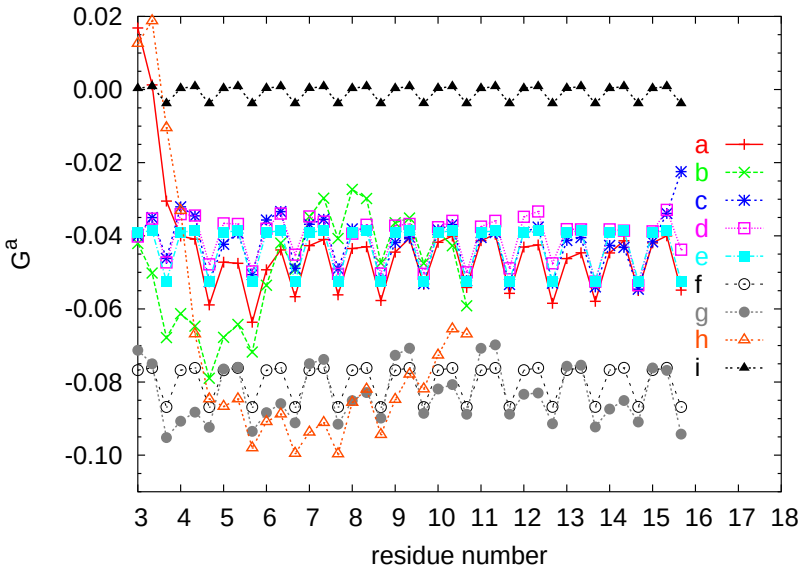


Figure 2.6: Chirality index, G, along the backbone, for α helices belonging to different proteins: [a]: hemoglobin 1-18 helix, [b]: ChPrP helix1, [c]: ChPrP helix2, [d]: ChPrP helix3, [e]: ideal α helix, [f]: ideal 3_{10} helix, [g]: ideal Type I β turn, [h]: 1LB0 3_{10} helix model peptide, [i]: polyalanine π helix. Type I β turn, 3_{10} and π helices are shown as comparison for ChPrP helix 1, which shows imperfections in the N-terminal region.

The helix of hemoglobin (5-18) and the helices 2 and 3 of avian prion protein (Figure 2.6), show the G pattern typical of ideal α helices, while for avian prion protein helix 1, G values reveal imperfections in the helix backbone, as also suggested by secondary

structure prediction algorithms [33]. In fact, the index shows irregularities in the first few residues, assuming the values typical of an α helix only after the sixth (Figure 2.6). The abundance of α helices can also be visually noticed looking at the G values along the backbone of hemoglobin (Figure 2.7 [a]), with the motifs helix-turn-helix and a high positive peak due to the presence of residues with ϕ and ψ values typical of poly-L-proline II, in the region after residue 90. Like hemoglobin, also for avian prion globular domain (Figure 2.7 [b]), it is easy to distinguish the different secondary structures along the backbone, like the three helices followed by turns, the positive peaks around residues 140 and 175 and after residue 200, due to at least one residue adopting poly-L-proline conformation. The zero G values suggest the presence of β sheets, quite evident in the plateau region centered at residue 169. Ubiquitin (Figure 2.7 [c]) has only one helix, and in fact only one region with negative periodic fluctuations of G is present, while at least four β-sheets can be identified, while serine protease (Figure 2.7 [d]) possesses a high number of turn regions, detectable from the sudden alternation of negative and positive peaks, which are instead only negative for 3_{10} helices (cf Figure 2.6), being constituted by at least three residues.

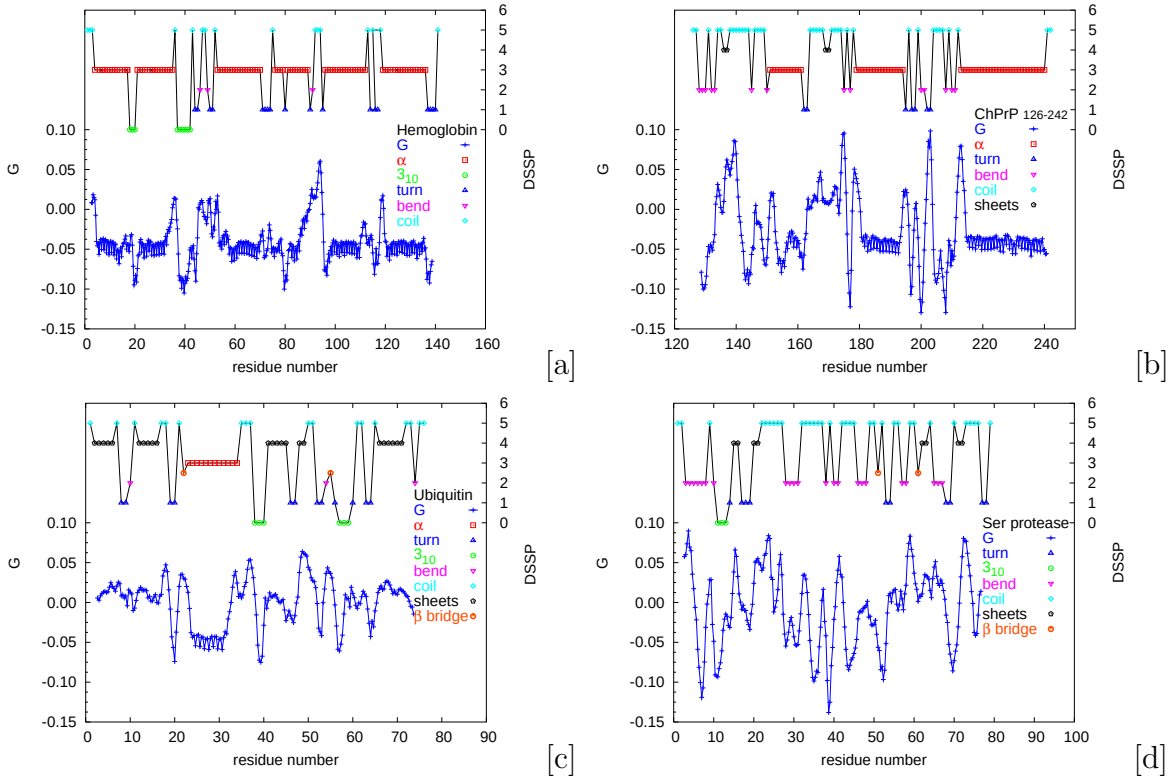


Figure 2.7: Chirality index, G, along the backbone for different crystalline proteins. Typical secondary structures, with the negative periodicity concerning the α helices, and with the G typical values for the other secondary structures (cf figure 2.5) are easily identified. The DSSP assignment is also plotted as the numeric code: 3_{10}=0, turn=1, bend=2, bridge=2.5, α=3, sheets=4, coil=5.

Concerning the β sheet-containing peptides (Figure 2.8 [a-b]), the analysis appears to be more difficult, because these structures occur in proteins with parallel or antiparallel regions formed by groups of residues far away in the protein sequence. Consequently further investigations, like hydrogen bonds screening, should be carried out in these cases to match the sequences. However, the plateaus at zero values of G generally help in identifying such structures.

As previously said, the chirality index is very sensitive to poly-L-proline dihedrals: a positive peak underlines in fact that at least one amino acid with PPII structural motif is present in a given protein region. Concerning the G of the model poly-L-proline peptide (fragment 51-60 of 1JMQ) reported in Figure 2.8 [b], a good overlap between the PPII ideal structure and the PPII model peptide results in the 3-5 region. After residue 5 the G values of 1JMQ drop as they take into account residues 7 and 8 which are not in PPII conformation. A full detection of PPII structure using DSSP-like algorithms is hampered because prolines do not form hydrogen bonds and although this structure is adopted also by other amino acids, its extended conformation (9.3 Å pitch) does not allow hydrogen bond pattern; therefore PPII regions are usually misclassified as loop or coils [14]. The sensitivity of G to PPII chirality seems important for a better identification of this class of structures. For a visual comparison with DSSP classification, in Figures 2.7 and 2.8 the DSSP sequence assignment is reported: in all cases the qualitative agreement between the two indexes is good, confirming the ability of the G index for discriminating secondary structures of real proteins.

To summarize the relation between chirality values and secondary structure, in Figure 2.9 the cumulative G distributions among all the structures analyzed in this section are reported. The histogram shows clearly four maxima, corresponding to Type I β turn/3_{10} helix, α helix, β sheets and PPII respectively, which all present G values close to the ones of the ideal structures (blue dots in Figure 2.9), and reveals the approximate content of these motifs in the data set. Even if necessarily limited by our particular choice of proteins and peptides, this finding suggests the possible use of such distributions for a quick similarity check between two protein structures or data sets.

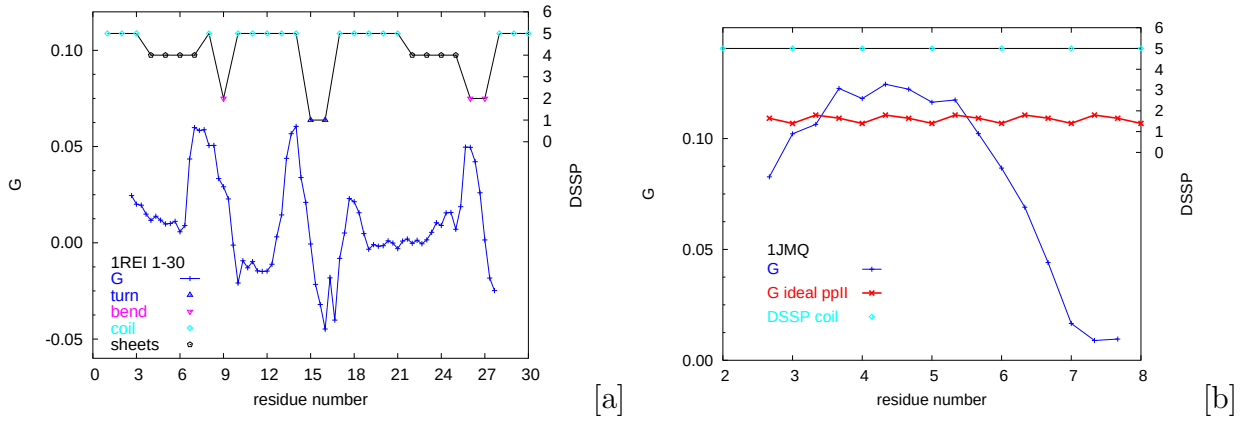


Figure 2.8: Chirality index behavior for model peptides: [a] Immunoglobulin antigen 1-30 1REI (β sheet 4-7 and 22-25); [b] 1JMQ (poly-L-proline II between residues 3-6). The DSSP assignment is plotted according to a number code which mimics the variation of G (3_{10}=0, turn=1, bend=2, bridge=2.5, α=3, sheets=4, coil=5).

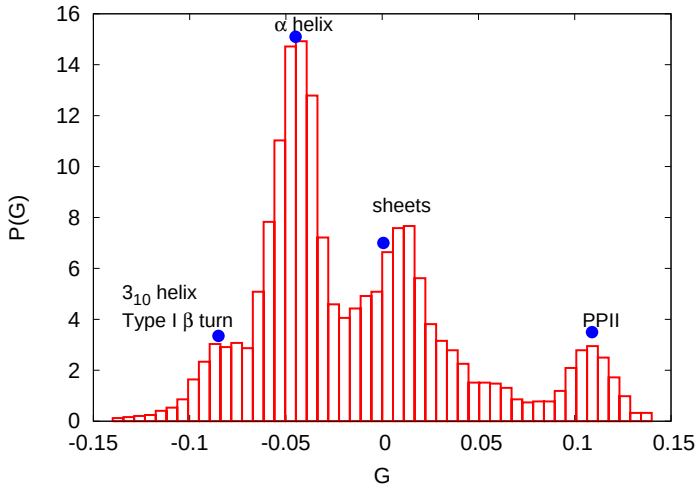


Figure 2.9: G distributions among the proteins and peptides analyzed in this work. The typical G values of the ideal structures are shown with blue dots.

2.2.3 Stability of the chirality index

In the previous section, the behavior of the chirality index for single structures of ideal and real proteins was investigated. Here, instead, the aim is to test the performance of the method in response to random thermal fluctuations and to conformational changes. This is important as in analyzing real structures, particularly in solution, fluctuations are unavoidable. To mimic this condition, the effects of a gaussian noise in the backbone angles values were studied, altering the secondary structure periodicity. To this end, 2000 configurations were randomly built for each type of secondary structure, extracting their ϕ and ψ from a gaussian distribution centered on the ideal ϕ and ψ values (see section 2.1) and from these, the overall average standard deviation of the chirality index was evaluated as a function of the gaussian half-height amplitude η (Table 2.1). In Table 2.1 such standard deviations are reported with the α helix and poly-L-proline II structures showing the highest ones, thus the chirality of these two structures is more sensitive to backbone variations with respect to others. In general, the index does not seem overly sensitive to a random perturbation of the dihedral angles and thus appears to be sufficiently robust to follow the fluctuations of the protein structure during a computer simulation, without being disrupted by thermal noise.

As a final inspection of the applicability of the chirality index analysis, the purpose is to test secondary structure assignment in the more realistic situation of a protein in water, where the geometry fluctuations or possibly, conformation changes, are also caused by the interaction with the solvent at certain thermodynamic conditions. Thus two rather long molecular dynamics runs were performed [34]: a 110 ns simulation of a fragment of hemoglobin, and a 50 ns simulation of a fragment of immunoglobulin antigen, in which a sheet-turn-sheet motif is present. Both simulations were run in water using ORAC 4.0 code [35] and the Amber94 force field (FF) [36]. Cubic boxes containing the protein chain and 484 water molecules for hemoglobin, and 1359 for the immunoglobulin fragment, were used with periodic boundaries and isothermal-isobaric conditions [37] (P=1 atm, T=300 K). Temperature was controlled using a Nosé-Hoover thermostat [38, 39] and the SPC model [40] was used for water. An r-RESPA multiple time-step algorithm with a potential subdivision specifically tuned for proteins [41] was used for integrating the equations of motion, using an overall time step equal to 10 fs. As it is possible to see from Figure 2.10 [a], the negative periodic pattern of the chirality index is retained during the simulation of hemoglobin, reflecting the fact that its

Table 2.1: Average G values and the relative standard deviations σ_G of the secondary structures and as a function of the extent of the gaussian noise amplitude η introduced on the value of the dihedral angles ϕ, ψ.

	α helix		**3_{10} helix**		**turn**		**sheets**		**PPII**		**π helix**	
η/deg	$\langle G \rangle$	σ_G	$\langle G \rangle$	σ_G	$\langle G \rangle$	σ_G	$\langle G \rangle$	σ_G	$\langle G \rangle$	σ_G	$\langle G \rangle$	σ_G
0	-0.043	-	-0.079	-	-0.083	-	0.0	-	0.11	-	-0.003	-
5	-0.04	0.01	-0.079	0.003	-0.082	0.004	0.0	0.003	0.10	0.01	-0.004	0.006
10	-0.04	0.02	-0.075	0.007	-0.080	0.009	0.0	0.007	0.10	0.02	-0.01	0.01
15	-0.04	0.02	-0.07	0.01	-0.07	0.01	0.0	0.01	0.10	0.03	-0.01	0.02
20	-0.04	0.03	-0.06	0.02	-0.07	0.02	0.0	0.01	0.10	0.03	-0.02	0.02

helical structure is not disrupted. Indeed the time behavior during the simulations of G for selected residues (Figure 2.11 [a,d]), shows that the chirality index is stable in an ensemble of configurations fluctuating around the same secondary structures, i.e. that folding/unfolding does not happen. More interestingly, in the case that major conformational changes occur, as for the 1REI immunoglobulin antigen 1-30 fragment (Figure 2.10 [b]), the chirality index gives precious indications about the different conformational states that the fragment explores, detectable from the different values adopted by G during the time evolution (Figure 2.12 [a,d]). The comparison with the instantaneous DSSP classification in Figures 2.11 and 2.12 confirms the qualitative agreement between the two indexes and the greater capabiliy of G in quantifying even small structural changes in time.

2.2.4 Chirality index dynamics and folding

Having established the link between chirality index and motif of a certain fragment, it is important to make full use of the fact that, differently from DSSP, the chirality index is a continuous dynamical quantity that can be employed to assess average structural changes during the simulation rather than just visually examine them along an individual trajectory. To this end, a time correlation function between the chirality index of two fragment a, b, was introduced and it is expressed as follows:

$$\chi^{a,b}(t) = \langle G^a(0) G^b(t) \rangle, \tag{2.3}$$

and a normalized version:

$$\chi_N^{a,b}(t) = \frac{\langle G^a(0)G^b(t)\rangle}{\langle G^a(0)G^b(0)\rangle}. \tag{2.4}$$

The normalized correlation has the advantage of bringing all the various fragment correlation in the same range, facilitating the comparison of time evolutions: values that remain close to one and slowly decaying indicate strong correlation, while functions reaching rapidly zero are proof of fast, uncorrelated motions. However the initial value is of course important, as it allows to distinguish the type of secondary structure. These equations were used for the calculation of auto–correlation functions, namely with $a = b$ in equations 2.3 and 2.4, reported in Figures 2.13 and 2.14 for hemoglobin and immunoglobulin respectively. Cross–correlation functions $\chi^{a,b}(t)$ and $\chi_1^{a,b}(t)$, were also calculated for selected residues, and their time behavior is shown in Figures 2.15 [a], [b] and 2.16 [a], [b]. In particular, examining the helix of hemoglobin it was found that both the auto and cross–correlations functions have high values in the internal core of the helix structure while in the N-terminal domain the memory of the initial configuration is rapidly lost (Figures 2.13 [a], [b] and 2.15 [a], [b]). This can be noticed from the asymptotic trend towards 1 of the functions $\chi_N^{22,22}$, $\chi_N^{37,37}$ and $\chi_N^{22,37}$, centered on residues 10 and 15 respectively, which correspond to the internal core (Figure 2.13 [a], [b] and 2.15 [a], [b]). More interestingly, a transition between α and 3_{10} helix is also observed in residue 5, corresponding to $\chi_N^{7,7}$. This is shown both by the decrease of the auto–correlation functions (Figures 2.13 [a], [b]) and by the variation of the index during the time, which exhibits evidently the transition approximately after 40 ns for G^7 (Figure 2.11 [c]). The N-terminal region alternatively is unstructured, or assumes turn conformations, as seen from the negative peaks in the G value reported in Figures 2.11 [a],[b]; this is also shown by overlapping the structures obtained from MD simulations (Figure 2.13 [c]).

Concerning the 1REI immunoglobulin fragment, high flexible and unstructured regions are present. A multiple transition between coil–sheets–coil–turn–3_{10} and rarely α helix, occurs in residue 6, understandable from the variation of G index during the time (Figure 2.12 [a]) and from the auto–correlation functions in Figure 2.14 [a],[b], where three minima and one shoulder for the sheets–coil transition at 20 ns, are present. The other sheet region, centered at residue 23, becomes unstructured (Figure 2.12 [c]), consequently the functions show a fluctuating behaviour, as underlined from Figure 2.14 [a] and more evidently in Figure 2.14 [b]. Residue 10 (Figure 2.12[b]) shows values of

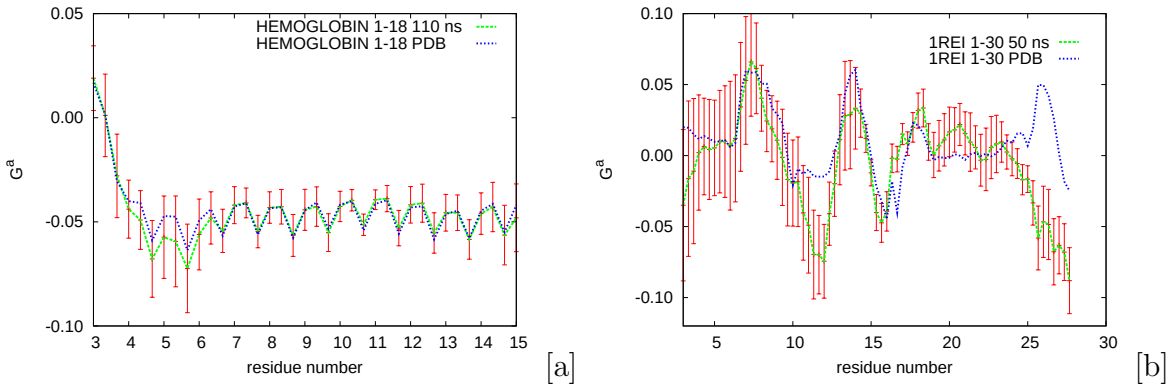


Figure 2.10: Standard deviations of G among hemoglobin 1-18 [a] and 1REI 1-30 [b] configurations. It is worth to note the persistence of the chirality index inside the average configurations for hemoglobin, while in the 1REI immunoglobulin antigen 1-30 fragment the crystal structure is not retained during the simulation. As comparison the G from PDB and from the trajectories is shown.

G typical of turn, coil and interestingly, of polyproline II at 35 ns, whose presence was confirmed with a check of backbone dihedral values, while residue 26 is in a less flexible region of the peptide (cf Figure 2.12 [d]). This is also confirmed by the auto correlation functions reported in Figure 2.14 [a],[b] showing both a flat shape. The cross–correlation functions of immunoglobulin fragment reported in Figure 2.16 [a],[b] show uncorrelated regions, thus pointing to high dynamical states. Even if the time scale of the simulations performed does not allow a complete exploration of the conformational space of these long peptides, and only a few exchanges between the most probable structures are sampled, the functions introduced here seem to be able to effectively quantify the time correlation between the different structures and between different regions of a given protein.

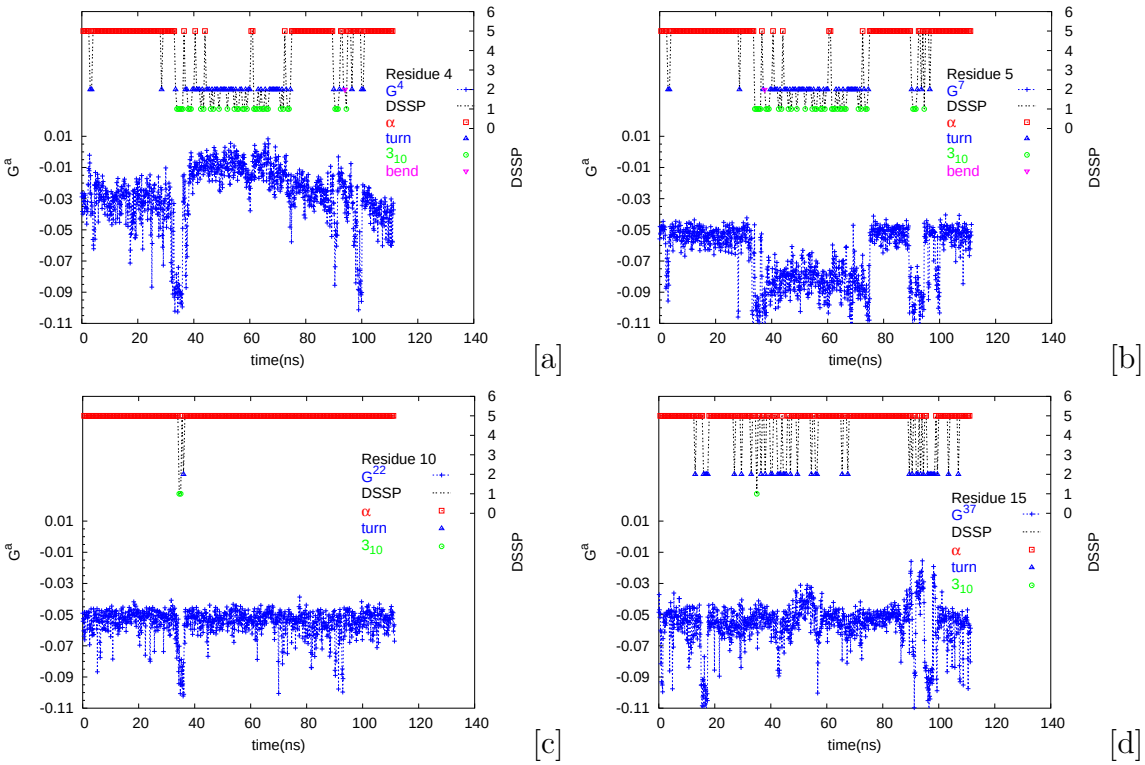


Figure 2.11: Time dependence of G for the fragments 4, 5, 10, 15 of hemoglobin 1-18 helix. The conversion from α helix to 3_{10} helix is underlined from the lowering of the G index (residue 5, G^7 [b]); the coil-turn transition is evident from the conversion to negative peaks (residue 4, G^4)[a]; the rigid core could be noticed from the constant G values [c],[d]. The DSSP assignment is plotted according to a number code which mimics the variation of G (3_{10}=0, turn=1, bend=2, bridge=2.5, α=3, sheets=4, coil=5).

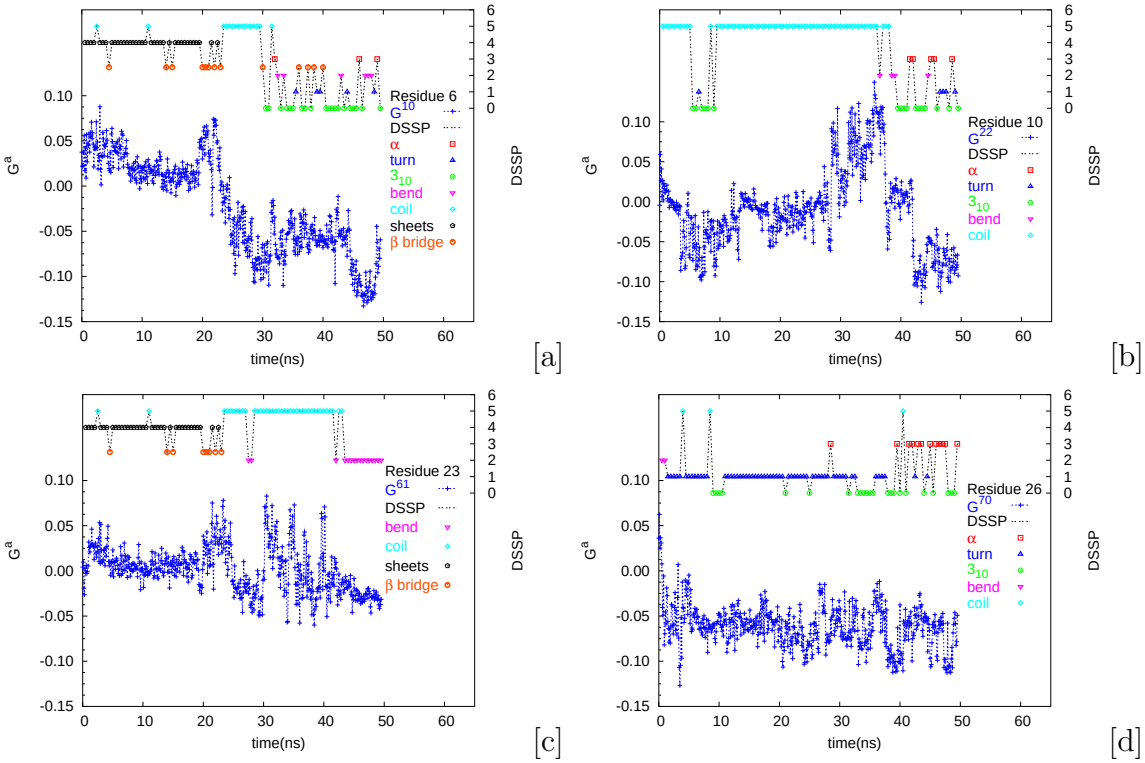


Figure 2.12: Time dependence of G for the fragments 6, 10, 23, 26 of immunoglobulin antigen 1-30 fragment. The conversion from sheets to coil, turn and 3_{10} helix is underlined from the lowering of the G index and few typical values of α helix are also detected at around 31 ns (residue 6, G^{10} [a]); PPII values can be individuated by the high positive peak present at 35 ns (residue 10, G^{22})[b]; the sheets conformation (residue 23, G^{61} [c]) can be distinguished from the values approaching zero and the transition to unordered states can be detected from the oscillations to positive and negative values near zero. The less flexible core could be noticed from the constant G values (residue 26, G^{70}) [d]). The DSSP assignment is plotted according to a number code which mimics the variation of G (3_{10}=0, turn=1, bend=2, bridge=2.5, α=3, sheets=4, coil=5).

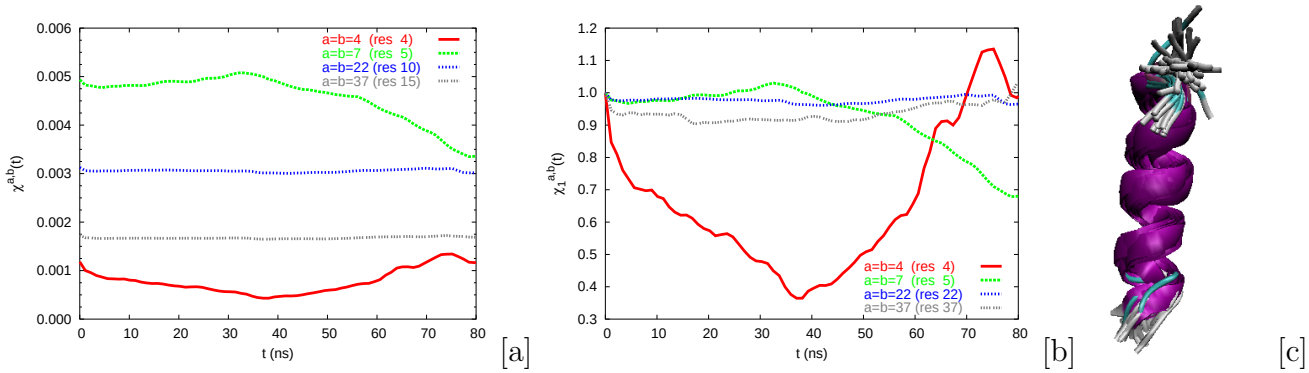


Figure 2.13: [a] Auto–correlation functions of the chirality index for residues 4, 5, 10, 15 of hemoglobin 1-18. [b] Auto–correlation functions scaled with the square of $G^{\mathrm{a}}(0)$. $\chi^{4,4}$ (residue 4) shows the unordered N-terminal region; $\chi^{7,7}$ (residue 5) shows clearly the transition between α helix and 3_{10} helix; $\chi^{22,22}$, $\chi^{37,37}$ (residue 10 and 22 respectively) underline a rigid core in α helix for the fragment 10-18. [c]: VMD visualization [42] of the 1-18 hemoglobin helix, which underlines a flexible N-terminal region (turns are shown in cyan and coils in gray) and a rigid core structure adopting α helix, shown in violet.

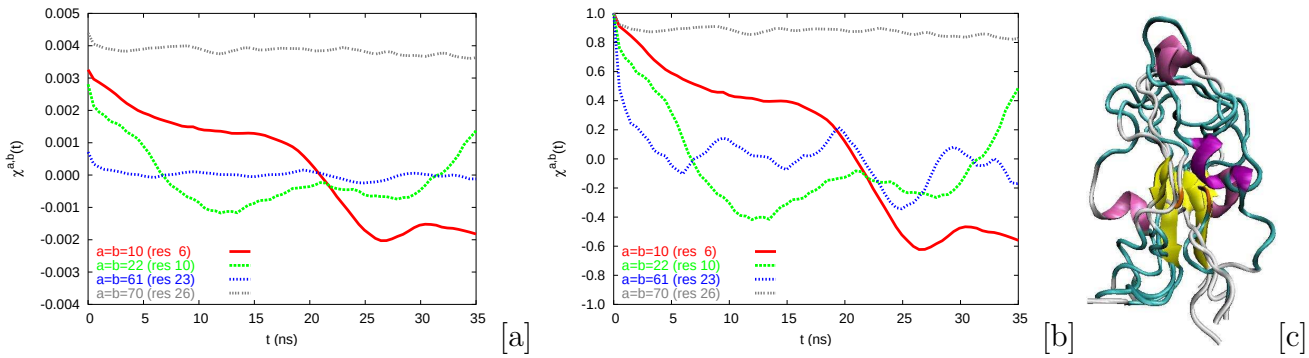


Figure 2.14: [a] Auto–correlation functions of the chirality index for residues 6, 10, 23, 26 of immunoglobulin antigen 1-30 peptide. [b] Auto–correlation functions scaled with the square of $G^{\mathrm{a}}(0)$. $\chi^{10,10}$ (residue 6) shows two minima for the coil–sheets, coil–turn transitions and one shoulder for the sheets–coil transition; $\chi^{22,22}$ (residue 10) shows clearly the two transitions between coil and 3_{10} helix; $\chi^{61,61}$ (residue 23) underlines a less correlation in the trajectories, due to unordered dihedrals, $\chi^{70,70}$ (residue 26), underlines a less flexible core. [c]: VMD visualization [42] of the 1-30 immunoglobulin antigen fragment, which underlines a flexible structure. For residue 6 it can be noticed the transition from sheets to 3_{10} helix; (turns are shown in cyan, coils in gray, sheets in yellow, α helix in violet and 3_{10} in pink).

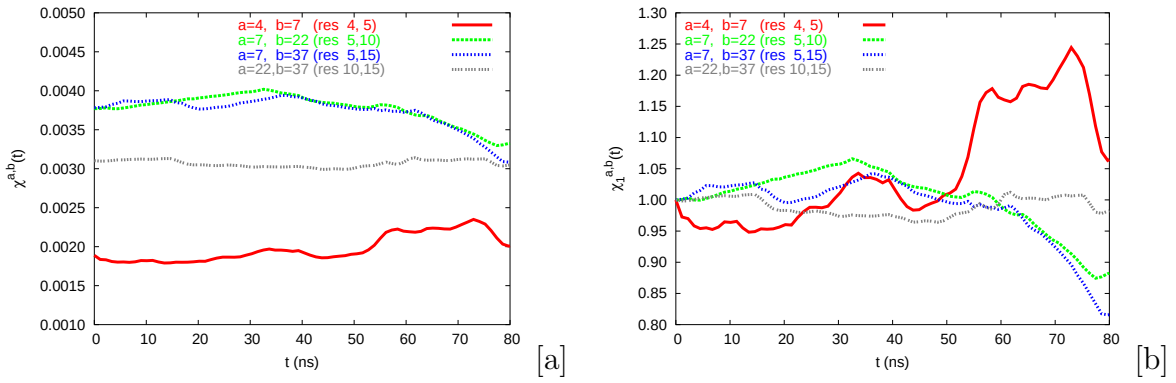


Figure 2.15: [a] Cross–correlation functions of the chirality index for residues 4, 5, 10, 15 of hemoglobin 1-18. [b] Cross–correlation functions scaled with the square of $G^{a}(0)$. The cross correlation clearly shows the presence of the rigid core for residue 10-15, namely the internal core of the helix and a less correlation between the central core and the N-terminal region.

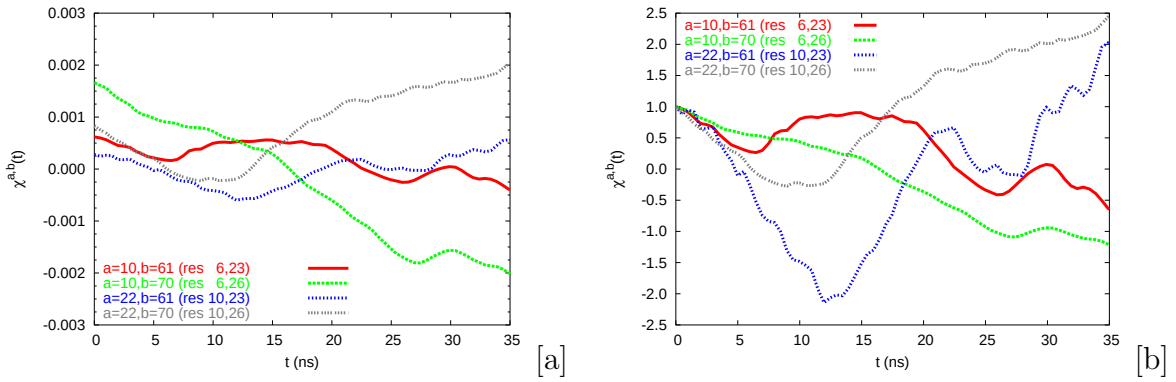


Figure 2.16: [a] Cross–correlation functions of the chirality index for residues 6, 10, 23, 26 of immunoglobulin antigen 1-30 fragment. [b] Cross–correlation functions scaled with the square of $G^{a}(0)$. The cross correlations reveal a strong loss of correlation in the secondary structures adopted by the immunoglobulin antigen 1-30 fragment, thus showing a high flexibility explored by the 1REI peptide.

2.2.5 Brief summary of the section

A geometrical chirality index G (see eq. 2.2), that can be easily calculated from the instantaneous conformation of a certain protein fragment [43], was introduced.
This index assumes well defined values for the typical secondary structure elements and, differently from other methods, is particularly effective in detecting polyproline II motifs. It has been shown that the index is robust towards random perturbations of the structures and that it is stable for long molecular dynamics trajectories that conserve the motif. On the other hand, following the evolution of fragments chirality in time and its correlation offers a direct possibility of monitoring protein conformational changes, showing this analyzing 110 ns and 50 ns–long runs for selected hemoglobin and immunoglobulin segments.

The index proposed here can be a powerful tool in complementing existing structure assignment algorithms, in following folding and misfolding processes for proteins in solution [44, 45], and in particular in capturing the early stages of these extremely important processes.

2.3 Local chirality of proteins: a new tool for structural bioinformatics

In the first part of this chapter, the capability of the chirality index in detecting the protein secondary structures and in following their evolution during the dynamics process, was verified. In this section, the natural chirality of proteins from the investigation of the PDB database is reported, trying as well to introduce some useful quantities to predict the secondary structure of a given protein, once known its primary sequence.

2.3.1 Chirality in native protein structures

For sake of simplicity, a unique value of G was defined for each amino acid i of a protein (see Table 2.4 for the G values) as the average of the $G^{\mathrm{a},15}$ (eq. 2.2) values whose 15 atoms window is centered on the the N, C_α, C atoms of amino acid i:

$$G_i \quad = \quad (G^{3(i-3)+1,\ 15} + G^{3(i-3)+2,\ 15} + G^{3(i-3)+3,\ 15})/3. \tag{2.5}$$

This definition is valid only for $i >= 3$ and $i < N - 2$ where N is the number of amino acids forming the protein.

To try to improve our knowledge on the chirality of native proteins, the Protein Data Bank (PDB) was analyzed. Therefore, all the structures were downloaded via FTP from the protein databank[1] web site. Subsequently, the DSSP (database of secondary structure proteins) assignments [5] were obtained, by downloading them via FTP from the DSSP web site[2]. Then, the secondary structures were classified for all the proteins present in the DSSP database also with the STRIDE algorithm [8]. In Table 2.2 the percentage of the secondary structures according to both the two methods is reported, concerning the X-ray and NMR protein structures. α helix is the more adopted conformation, due to the presence of a wide abundance of globular proteins; a few extent of 3_{10} helix is present according to both DSSP and STRIDE classifications; concerning β sheets, they are almost equally populated from both the two algorithms, while the turn classification according to STRIDE likely takes into account the bend structure present only in the DSSP classification.

For all the structures, whose DSSP assignment is available (10504 X-ray and 2340 NMR), the chirality index was also calculated, with equation 2.5, looking for a deeper analysis of the protein native fold, using the local chirality of amino acids.

Table 2.2: Percentage of the different protein secondary structures sampled from the protein databank, for the X-ray and NMR structures, using DSSP and STRIDE classification. All the structures are labeled according to the DSSP and STRIDE notation: H: α helix, E: β sheets, C: Coil, T: Turn, S: Bend, G: 3_{10}, B: Bridge, I: π helix.

	Xray		**NMR**	
SS	DSSP	STRIDE	DSSP	STRIDE
H	31.76	33.75	29.95	32.72
E	22.41	22.47	22.45	20.23
C	19.36	16.92	19.19	17.49
T	11.87	21.47	13.92	25.93
S	9.26	-	11.07	-
G	3.96	4.05	2.08	2.35
B	1.35	1.30	1.31	1.25
I	0.03	0.02	0.03	0.01

2.3.2 Fingerprint of evolutionary information

An additional aim is to better characterize the edges of secondary structures, namely the variation of the chirality index where a particular structure begins and where it ends. With this purpose, in Figures 2.17 and 2.18 the chirality index respectively of the X-ray and NMR structures is reported, for every residues classified as α helix using both DSSP [a] and STRIDE [b]. Were considered as *edges* the initial and the final residues of a particular structure; when subtracting the distributions of such edges to the full ones (see Figures 2.17, 2.18), it is possible to notice the absence of a broadened peak, present around values of G index approaching zero, typical of coil regions. This is shown both for Xray and NMR structures (see Figures 2.17 and 2.18, respectively), assigned using DSSP [a] and STRIDE [b] classifications. As expected, the G index distribution for residues classified as α helix, belonging to NMR structures, is broader if compared to the X-ray one (see Figures 2.17, 2.18). All these considerations are more evident for the chirality index of residues, classified as 3_{10} helix (Figure 2.19, 2.20) belonging to X-ray structures, in which the second maximum around values of G approaching zero is more exalted if compared to those ones of Figures 2.17 and 2.18. Here, the subtraction of the edges gives a more defined main peak for the G index distribution of residues classified as 3_{10} helix using DSSP (Figure 2.19 [a]) and also, despite in less extent, for the STRIDE one (2.19[b]). Again, the G index distributions (Figure 2.20) of residues classified as 3_{10} helix belonging to NMR structures, appear to be broader when compared to the X-ray database (Figure 2.19).

The β sheet distributions are reported in Figures 2.21, 2.22. Such secondary structure and its edges show zero approaching chirality, shared also from coil regions and reported for sake of clarity in Figures 2.21, 2.22. However, the subtraction of the edges gives a better defined distribution, almost eliminating the positive values of G, similar to those ones belonging to residues classified as coil. In Figure 2.23 the G index distributions are reported for residues belonging to X-ray [a] and NMR structures [b] classified as Turn both with DSSP and STRIDE. Interestingly, the main peak shows zero chirality, although ideal type I β turn structure occurs with negative G index chirality [46], possibly due to a wide range of dihedral ϕ and ψ angles. For completeness, the G index distributions of residues classified as bend are reported (Figure 2.23). These are only present in DSSP classifications and likely included inside the Turn regions in the STRIDE classification.

To deeply rationalize the chirality indexes adopted by turn regions, we used the STRIDE classification of turns. In table 2.3 the percentages of the different turns, according to STRIDE, are reported. Notably, type IV β turn, namely every turn which fades off from the typical ϕ and ψ dihedrals, shows the highest percentage.
In Figures 2.24, 2.25 the chirality indexes for the different β turns are reported. Notably, all the turns show a main peak with chirality indexes value approaching zero, possibly due to a wide range of ϕ and ψ dihedrals, adopted by these structures. By the way, a second peak, with negative chirality indexes values, is present especially for Type I β turn of the NMR dataset, consistent with the negative chirality indexes, found in the ideal type I β turn.

As a final investigation, the chirality of proline amino acid was studied, because of its importance in stabilizing turns and polyproline II (PPII), this latter structure being not recognized by STRIDE and DSSP. In Figure 2.26, the chirality of proline, alanine and histidine residues, all classified as coil with DSSP and belonging to X-ray [a] and NMR structures [b], together with the chirality of the coil distribution according to DSSP, is reported. It is evident that the distribution of G for prolines (see Figure 2.26) is significantly different from the one typical of coils, revealing a broad shoulder at G $\sim$ 0.1 value, typical of Poly-L-proline II structures. Such positive chirality values indicate that coil regions for prolines, actually contain also left-handed helices, usually neglected in the study of protein fold. On the whole, the chirality among the PDB structures is reported in Figure 2.27. As it can be noticed, two main peaks with the typical chirality of α helix and a coil-sheets conformations are present both in X-ray and NMR structures of the protein databank, in accord with the percentages reported in Table 2.2.

Table 2.3: Percentage of the different turns for the X-ray and NMR protein structures, according to STRIDE.

Turn	**Xray**	**NMR**
I	33.39	19.58
I'	4.12	2.10
II	11.37	5.81
II'	2.02	1.42
IV	35.36	59.54
VIa	0.02	0.07
VIb	0.00	0.01
VIII	9.94	6.41
γ	0.56	4.32
γ'	0.05	0.72

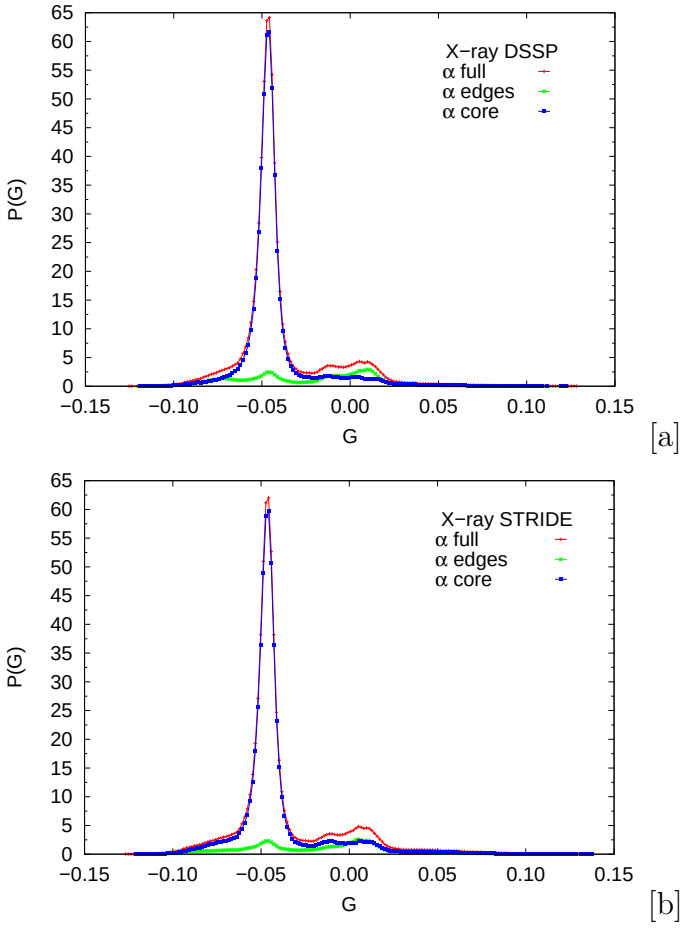


Figure 2.17: Chirality index distributions for the α helices classified using DSSP [a] and STRIDE [b], sampled from the PDB X-ray structures; it is worth to note a more defined distribution when the edges between one structure and the following one, which correspond also to the edges of the distribution, are removed.

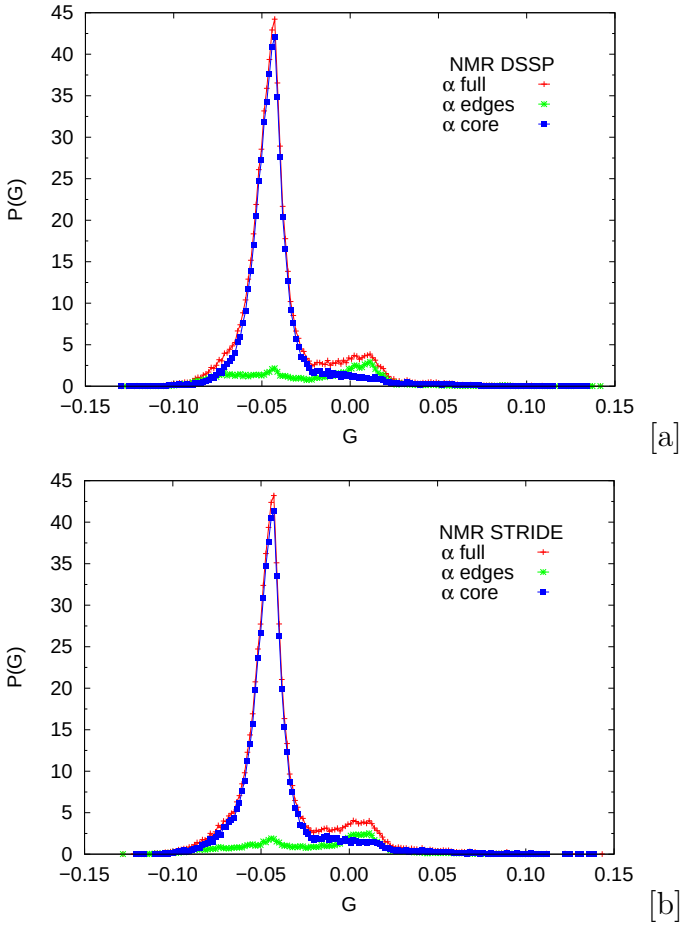
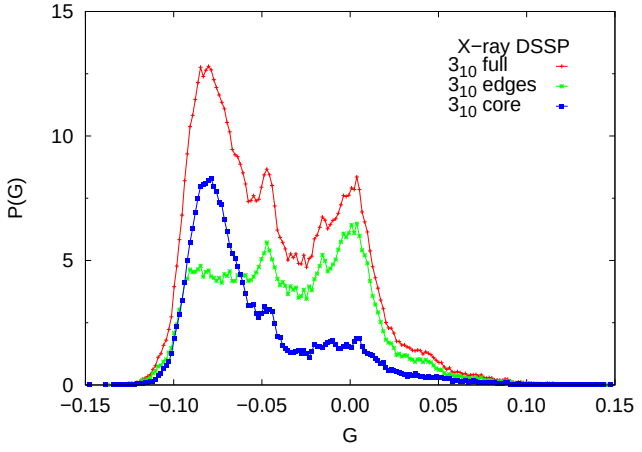
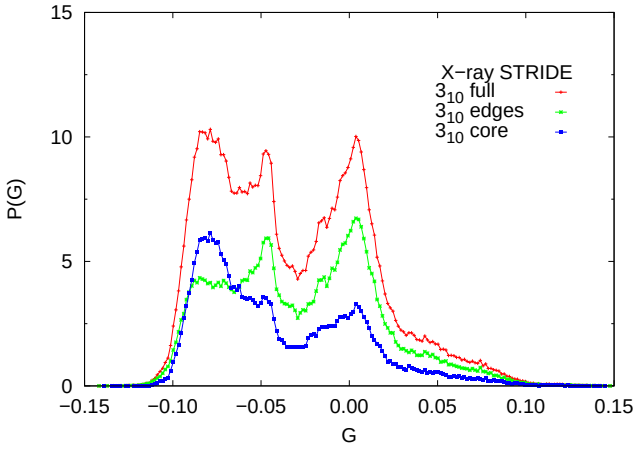


Figure 2.18: Chirality index distributions for the α helices classified using DSSP [a] and STRIDE [b], sampled from the PDB NMR structures; it is worth to note a more defined distribution when the edges between one structure and the following one, which correspond also to the edges of the distribution, are removed.



[a]



[b]

Figure 2.19: Chirality index distributions for the 3_{10} helices classified using DSSP [a] and STRIDE [b], sampled from the PDB X-ray structures; it is worth to note a more defined distribution for the 3_{10} helices classified using DSSP algorithm, moreover both distributions reveal a not proper chirality of a 3_{10} helix, as shown from the peak centered at values of the G chirality index approaching zero, typical of coil structures. Such a peak lowers when the edges are removed, but still persists in the STRIDE distribution.

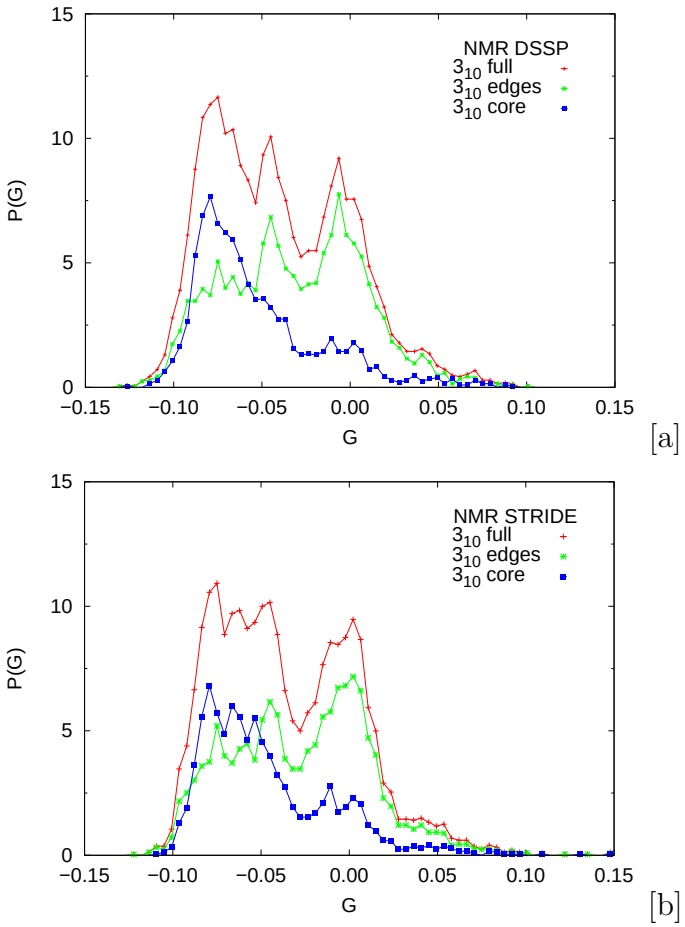


Figure 2.20: Chirality index distributions for the 3_{10} helices classified using DSSP [a] and STRIDE [b], sampled from the PDB NMR structures; both distributions are less defined if compared to the X-ray ones of Figure 2.19.

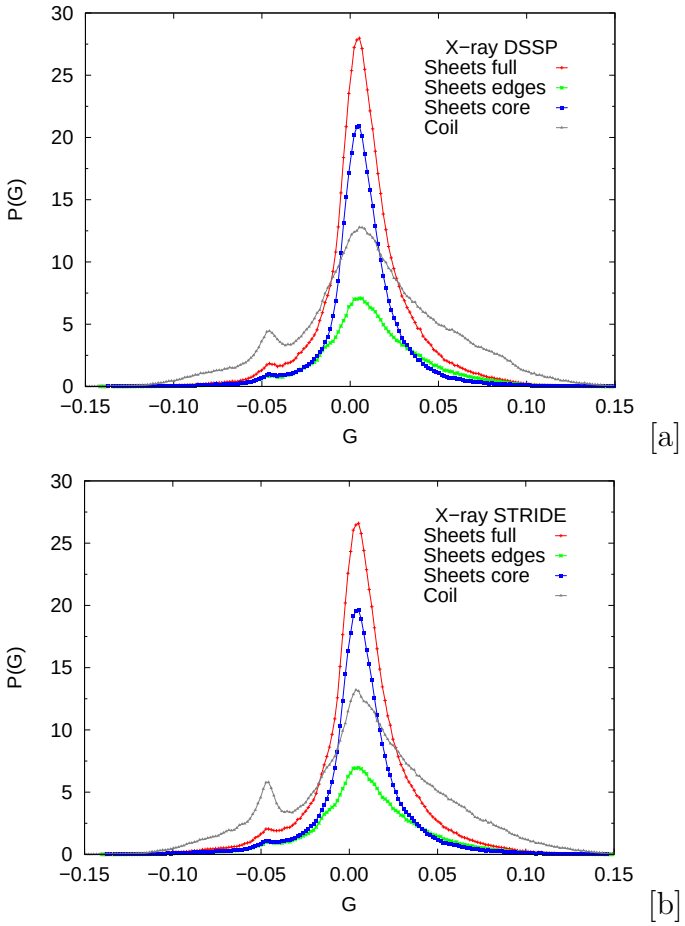


Figure 2.21: Chirality index distributions for the β sheets classified using DSSP [a] and STRIDE [b], sampled from the PDB X-ray structures; it is worth to note a good description of the β sheets structures using both DSSP [a] and STRIDE [b].

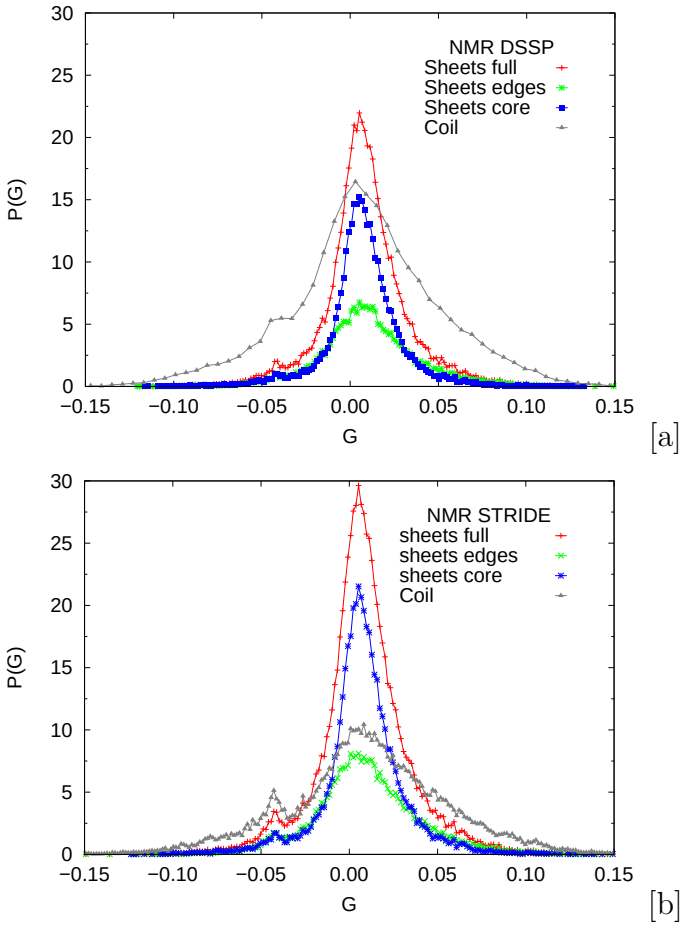


Figure 2.22: Chirality index distributions for the β sheets classified using DSSP [a] and STRIDE [b], sampled from the PDB NMR structures; it is worth to note a good description of the β sheets structures using both DSSP [a] and STRIDE [b].

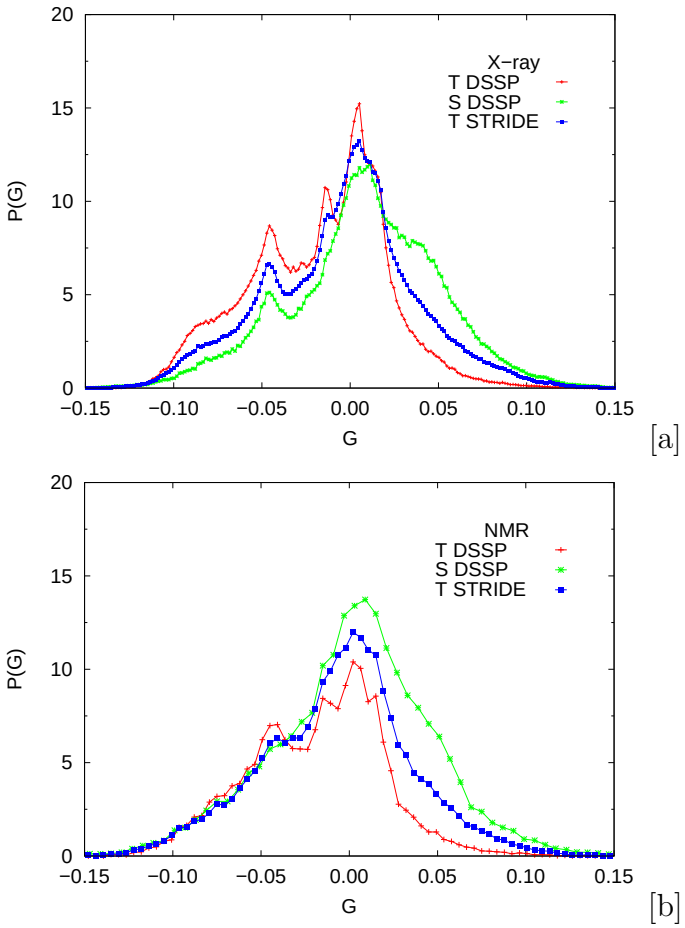


Figure 2.23: Chirality index distributions for the turn conformation classified using DSSP and STRIDE, sampled from the PDB of X-ray [a] and NMR [b] structures. The turn conformations, according to the percentage reported in Table 2.2, are more favored using the STRIDE algorithm with respect to the DSSP one. Bend distribution, classified with DSSP, is shown to make a comparison with the turn distribution according to STRIDE.

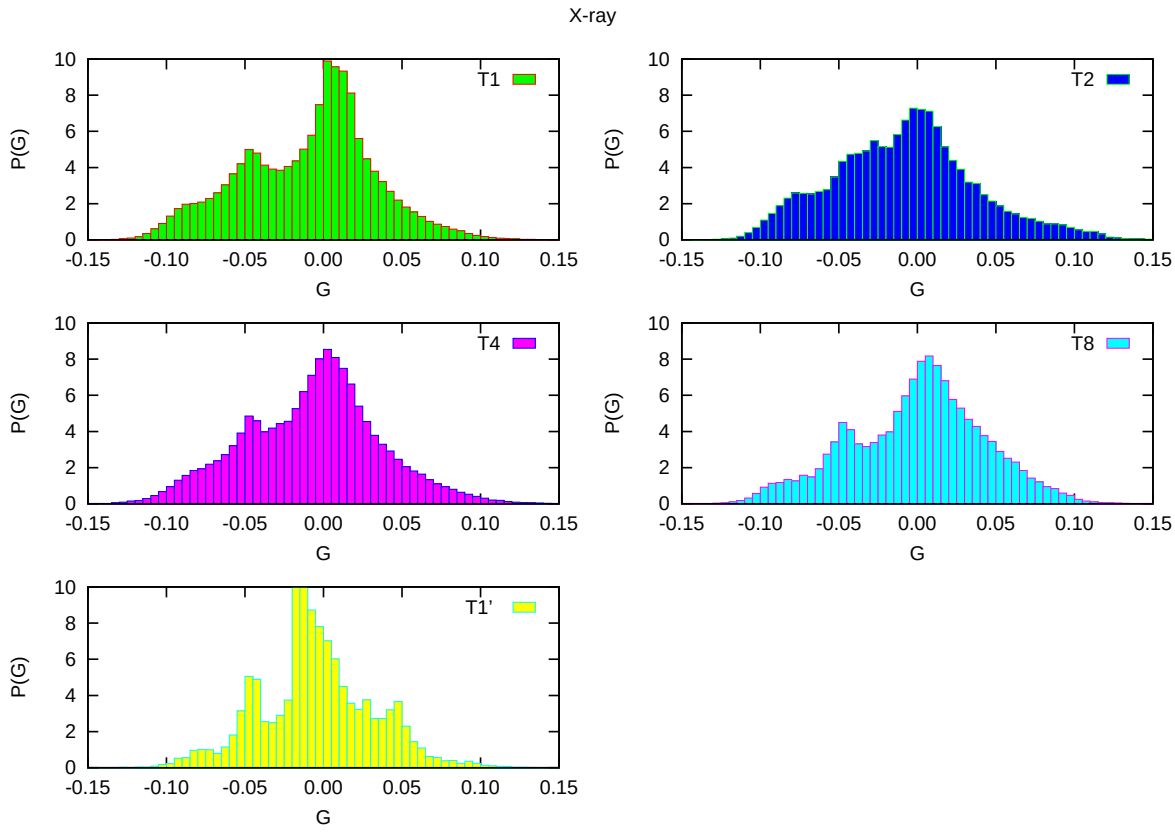


Figure 2.24: Chirality index distributions for the different turns, classified according to STRIDE, concerning the X-ray dataset. It is worth noting the same shape of the chirality indexes for the different turns, probably denoting the wide average range of ϕ and ψ angles.

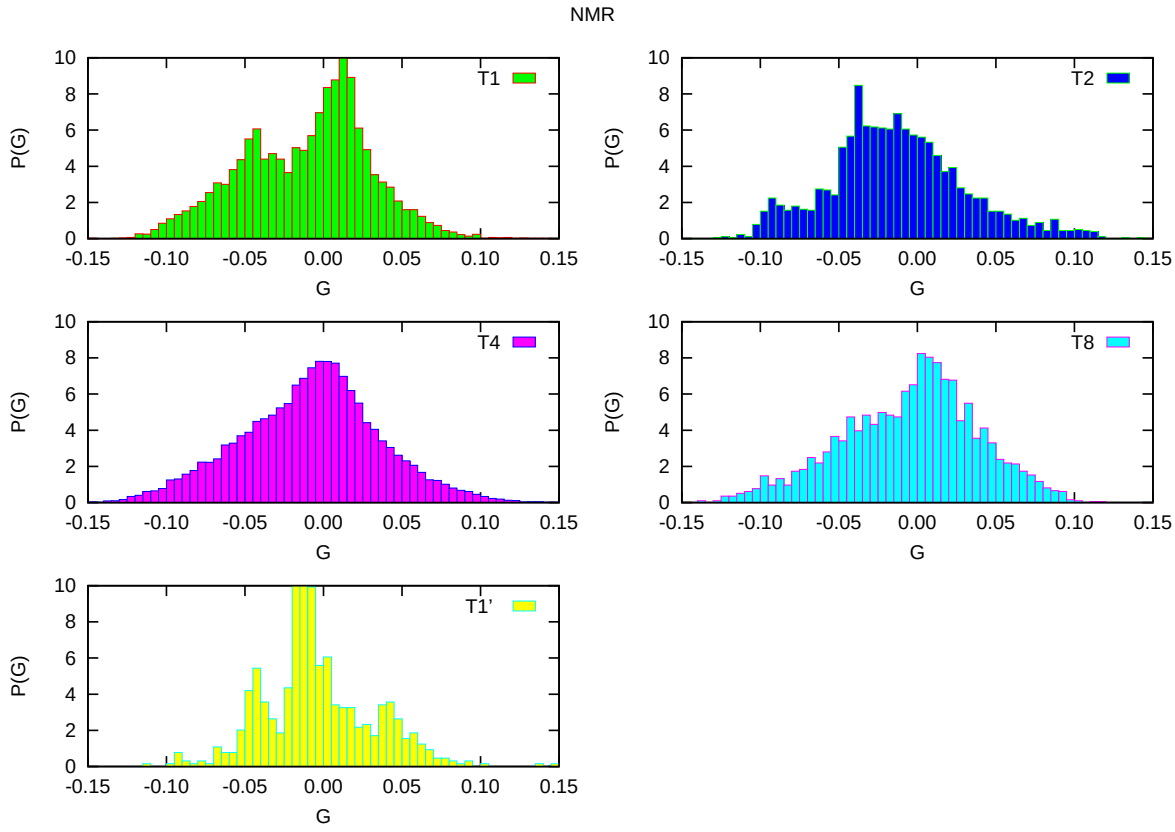
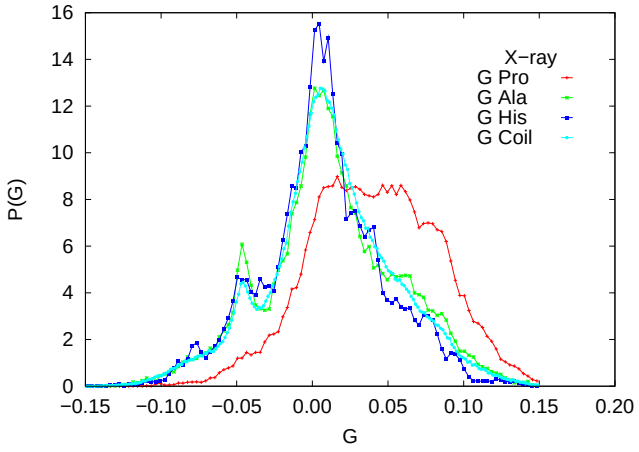
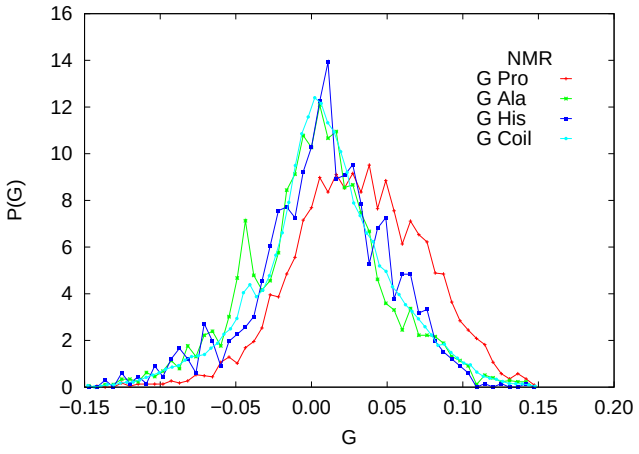


Figure 2.25: Chirality index distributions for the different turns, classified according to STRIDE, concerning the NMR dataset. The peak with negative chirality indexes belonging to type I β turn distributions results to be more populated if compared to that one of the X-ray one (see Figure 2.24). Moreover, Type IV β Turn results to be lacking of the second peak with negative chirality indexes values, present instead in the X-ray dataset.



[a]



[b]

Figure 2.26: Chirality index distributions for Proline, Alanine Histidine, all classified as coils, and Coils, calculated for the PDB X-ray [a] and NMR [b] dataset. The chirality of proline is very different from that one of alanine and histidine. These two latter ones are more similar to the coil distribution according to DSSP.

Structure	$\langle G \rangle$	σ_G	N_R
α helix	-0.04	0.01	>3
3_{10} helix	-0.08	0.02	> 3
β Sheets	0.00	0.02	≥ 2

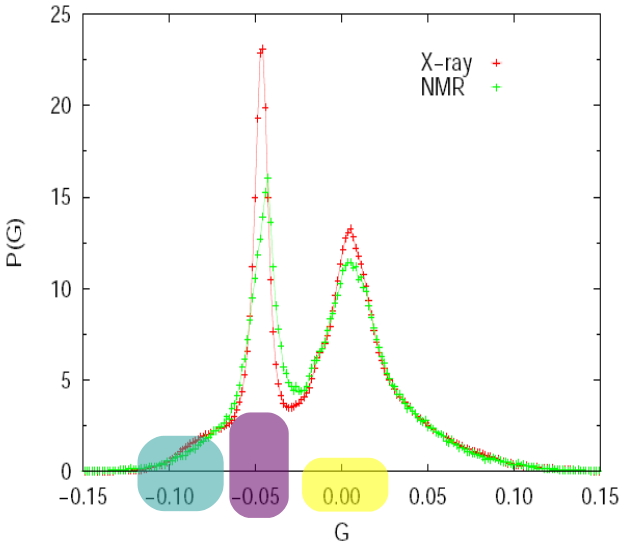


Figure 2.27: Average G values and the relative standard deviations of G for the most adopted PDB secondary structures, obtained by a gaussian fit of the G index distributions. The chirality index distributions of the PDB Xray and NMR dataset is also reported. It is worth to note two main regions centered approximately at chirality index values of -0.05 and 0.00, namely α helix and sheets-coil secondary structures, according to the percentage reported in Table 2.2.

2.3.3 A scoring function for tridimensional protein structure based on conditional probability of G_i, G_{i+1}

In this section it is considered how the chirality of a particular residue can influence the chirality and thus the secondary structure of the following one, suggesting how persistent is a structure. Consequently, for each of the twenty amino acid types AA, the conditional probability $P^{NMR}(G_{i+1}|G_{i,AA})$, i.e. the normalized occurence in the dataset of having, for an amino acid $i+1$, a given value of G_{i+1} once fixed the type of the preceeding amino acid i to AA and its chirality index to G_i, was then evaluated. These probability maps can be employed in the definition of scoring functions that allows to measure how is compliant a given protein structure to the dataset that originated the maps.

In Figures 2.28 [a-d] and 2.29 [a-d] the map of the conditional probability $\mathrm{P}(\mathrm{G}_{i+1}|\mathrm{G}_i)$ is reported for the DSSP secondary structures of the X-ray protein dataset. As shown in the previous section, α helix (Figure 2.28 [a]) shows the narrowest range of G_i, G_{i+1} values, around (-0.05, -0.05), typical of α helix (see Figures 2.17, 2.18 [a]-[d]), which here denoted the peculiar persistence of this motif. On the other hand, 3_{10} helix (Figure 2.28 [b]), shows a broad map, which becomes narrower for turn structure (Figure 2.28 [c]). β sheets (Figure 2.28 [d]) show a spread map around zero, which is consistent with the absence of chirality for these structures. π helix (Figure 2.29 [a]), probably because of its rare occurrence, shows a narrow map with values of G_i, G_{i+1} approaching zero, as expected from its typical chirality (see Table 2.4). Finally the coil, bend and β bridge maps (Figure 2.29 [b-d]) show spread maps, considering the wide range of dihedral angles which these conformations cover.

The conditional probability was also calculated separately for each of the twenty amino acids and reported in Figures 2.30-2.34. In Figure 2.30 [a] the *natural* conditional probability is reported for all the L-amino acids inside the protein databank. Here, the main narrow range is that one of α helix, previously shown in Figure 2.28 [a], which is found mainly in Ala, Leu, Ile residues (Figure 2.30 [b, d]), known to have a high α helix propensity [47], in Met, Cys and Val (Figure 2.31 [a-c]) and in less extent in the other amino acids, with the exception of glycine and proline (see Figures 2.32 [d], 2.33 [d]). On the contrary, proline residue (Figure 2.33 [d]) presents in the map a very narrow range of G_i, G_{i+1} values of (0.15, 0.13), typical of polyproline II structures. This range is also present mainly in the Tyr and His maps (Figures 2.33 [a], 2.33 [e]), although with a

slight shift with respect to the G_i, G_{i+1} values of the Pro map. After obtaining all these maps, it can be introduced a function which weighs each consecutive couple of values of amino acids in a given protein structure, according to the conditional probability described above and takes the average of all the weights:

$$G_{score} = \frac{1}{N-5} \sum_{i=3}^{N-3} P^(G_{i+1}|G_{i,AA}) \tag{2.6}$$

As the chirality index is a local function of the coordinates, this score gives an information on the likeliness of the secondary structure, and ranges from 0 (very unlikely) to 1 (perfect match with the dataset). It is worth noting that even for the proteins of the dataset itself, the compliance is not perfect, but has a broad non-gaussian distribution ranging approximately from 0.05 to 0.3 with a maximum at around $G_{score} = 0.11$. To understand the origin of the shape of the G_{score} distribution, it was decomposed in separate contributions according to the SCOP classification of protein folds [48] (see Table 2.5), which classifies the structure in 12 classes, being the dataset mainly constitued of α, β, α+β and α / β classes.

In Figures 2.35 and 2.36 the G_{score} distribution concerning X-ray and NMR structures is reported, suggesting that α proteins adopt higher G_{score} than that of β ones, consistent with the rigidity of α helix and the plasticity of β sheets. In addition, native proteins adopt a value around 0.16 concerning X-ray and around 0.10 for NMR ones. As expected, the more crystalline X-ray structures show higher values of G_{score}, with respect to NMR structures, in which thermal motions could be taken into account and therefore showing slower values. The value of the G_{score} may give precious information about the correct conformation adopted from a protein during molecular dynamics simulation, by checking the evolution of this value during the time, an example of this application will be given in chapter 4.

Table 2.4: Average G values and relative standard deviations of G for ideal secondary structures, involving at least N_R residues. Each structure was built by sampling ϕ and ψ angles from a gaussian distribution, centered on the ideal ϕ and ψ values with sigma=15 degree (see reference [46]).

Structure	$\langle G \rangle$	σ_G	N_R
α helix	-.04	0.02	>3
3_{10} helix	-.07	0.01	> 3
β Turn I	-.07	0.01	2,3
β Sheets	+.00	0.01	≥ 2
PPII	+.10	0.03	>3
π helix	-.01	0.02	>3

Table 2.5: SCOP classifications of proteins of the X-ray and NMR datasets.

Class	**Xray**	**NMR**
All β	25.6	14.0
α / β	22.3	3.3
α + β	16.8	15.2
Not Classified	16.1	16.6
All α	13.9	13.6
Multi-domain α and β	2.1	-
Small Proteins	1.4	19.7
Membrane and cell surface proteins and peptides	0.8	1.1
Coiled coil proteins	0.6	0.7
Designed Proteins	0.1	2.1
Peptides	0.1	13.4
Low resolution protein structures	0.1	0.3

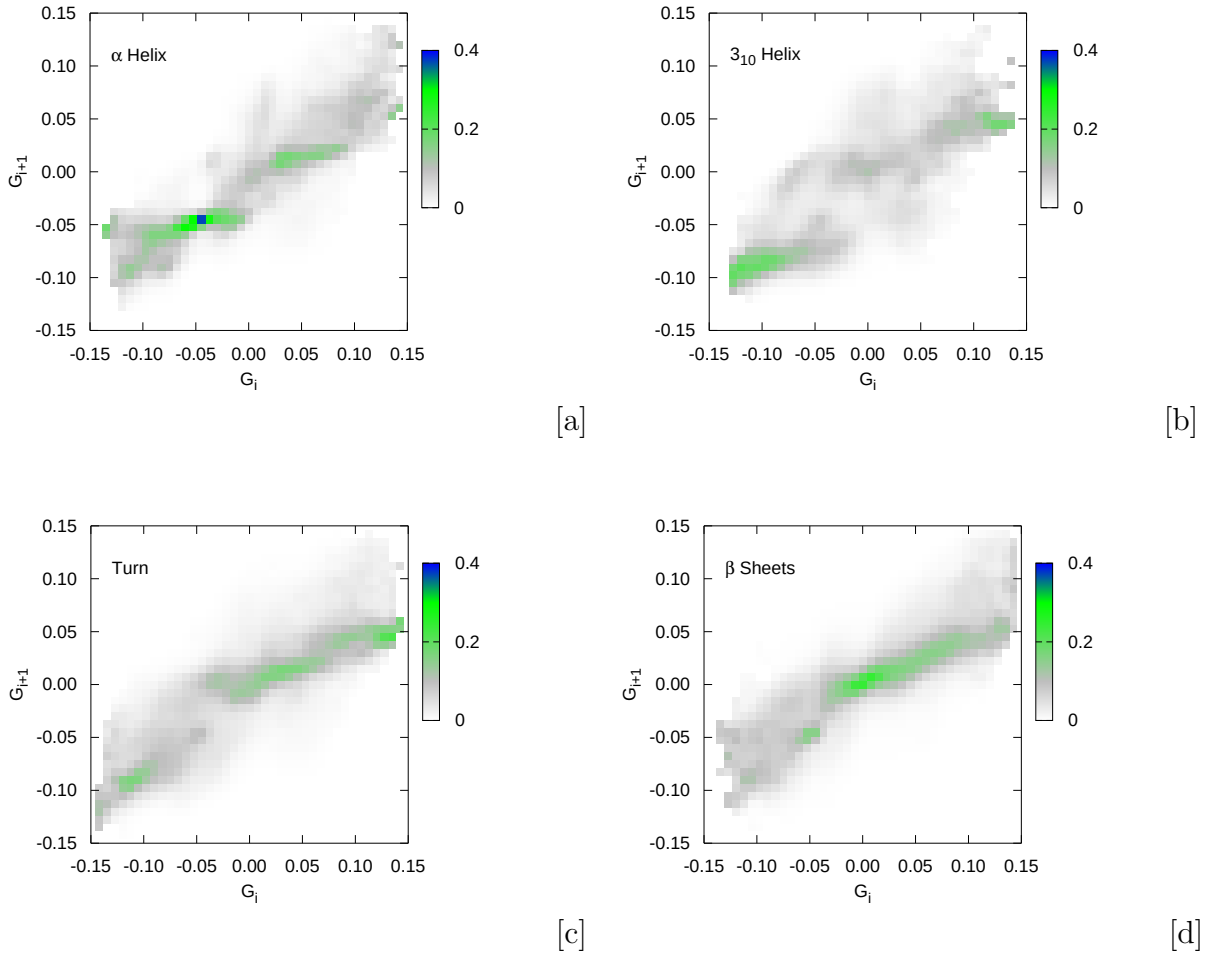
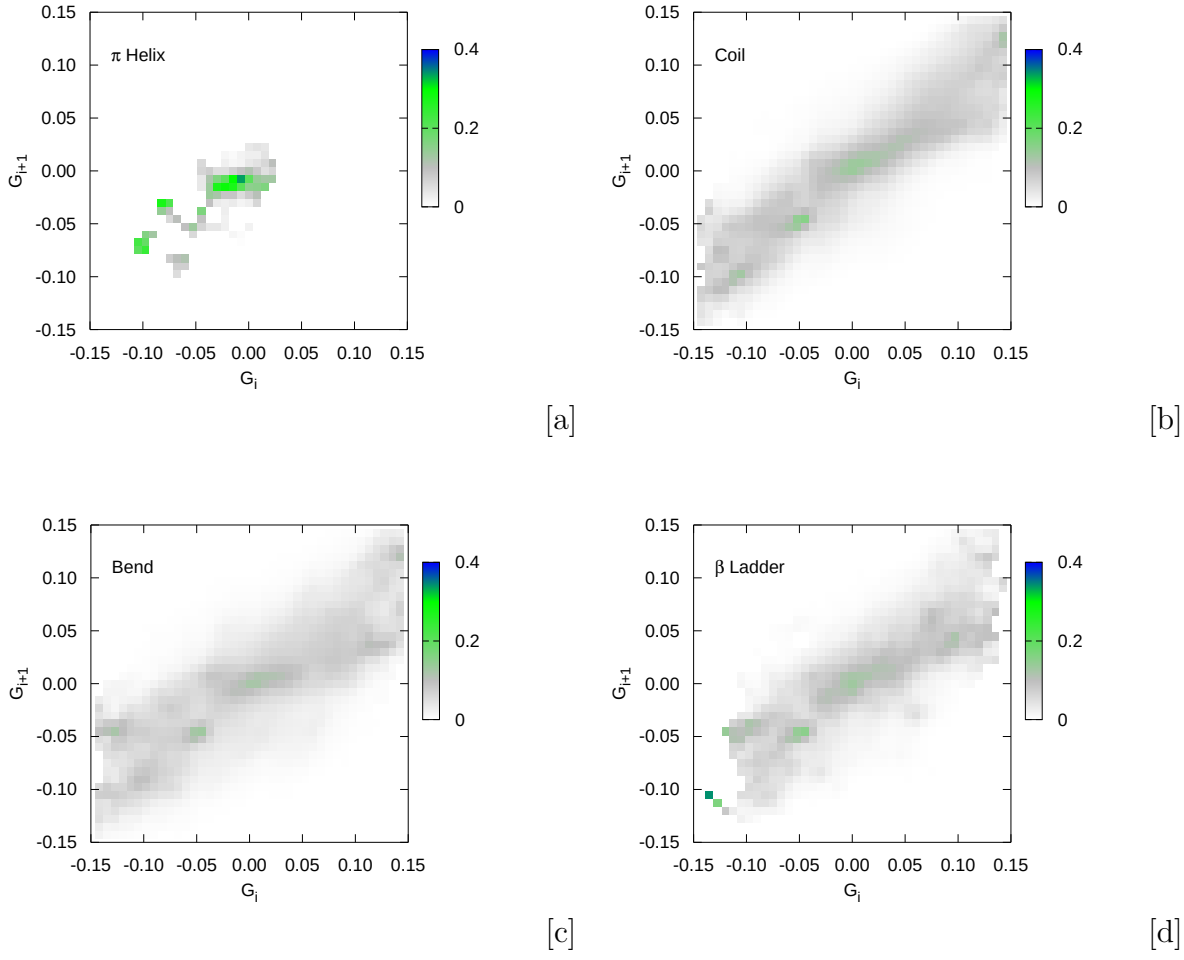


Figure 2.28: Conditional probability P($G_{i+1}|G_i$) calculated among the PDB database.[a] α helices structures show a narrow range around (-0.05, -0.05); [b] a wide range around (-0.10, -0.10) is present for 3_{10} helices structures; [c] a better defined range with respect to 3_{10} helix is present for Turn regions at (0.15, 0.05) and (-0.12, -0.10); [d] β sheets present a range with zero chirality. The G_i, G_{i+1} occurrence is shown with a color code ranging from green to red.



[a] [b]

[c] [d]

Figure 2.29: Conditional probability $P(G_{i+1}|G_i)$ calculated among the PDB database. [a] π helices shows more ranges adopted, being the main one with zero chirality. Excluding π helix, which rarely occurs, coil [b], bend [c] and Bridge regions [d] (apart one region at (-0.13, -0.10) show wide ranges of G_i, G_{i+1} values. The G_i, G_{i+1} occurrence is shown with a color code ranging from green to red.

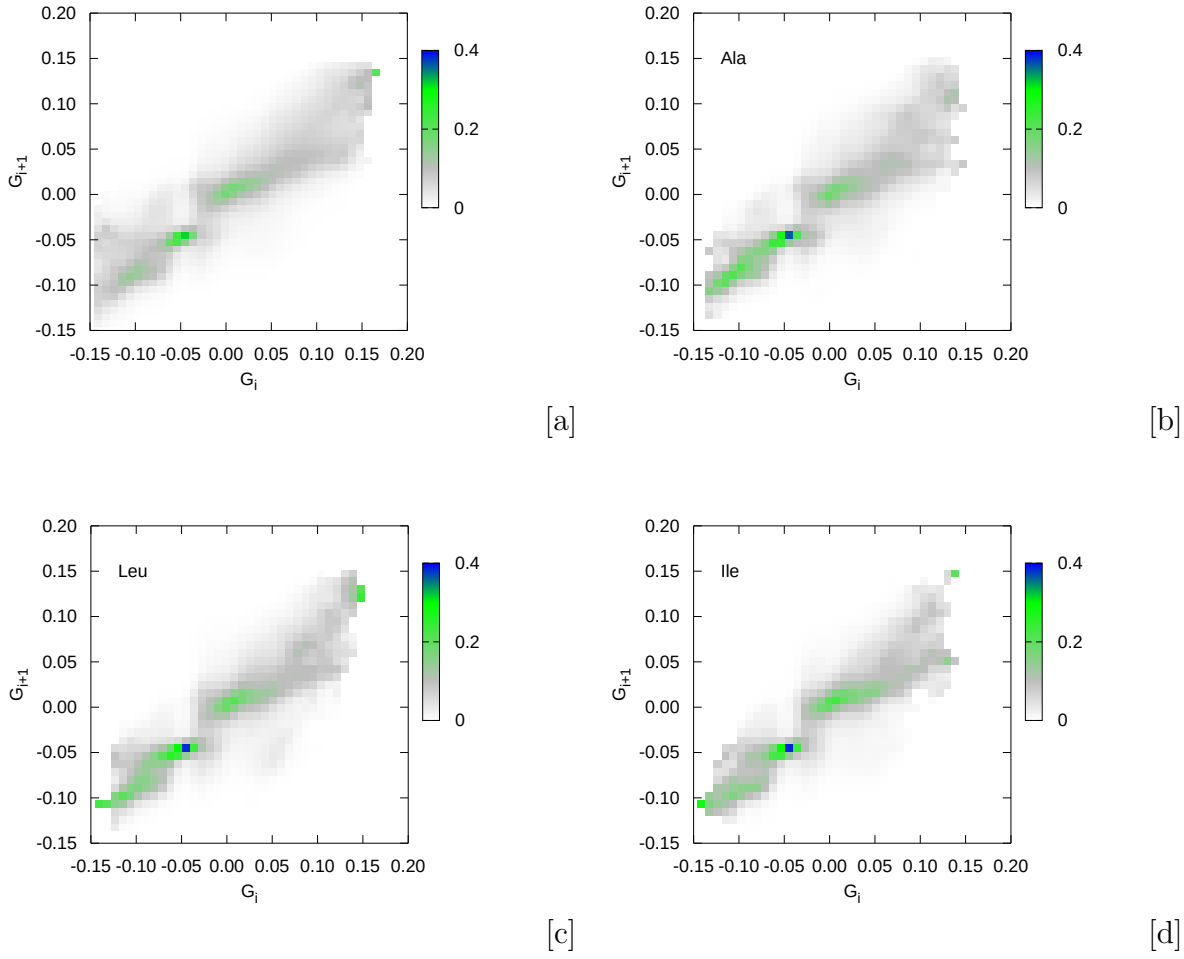


Figure 2.30: Conditional probability $P(G_{i+1}|G_i)$ calculated among the amino acid PDB database. [a] Conditional probability as union of those one of the twenty amino acids. The α helix range, here shown at G_i, G_{i+1} values (-0.05, -0.05), is present for Ala [b], Ile [c] and Leu residues [d]. These latter ones ([c], [d]) show also negative ranges of G_i, G_{i+1} values, typical of turn and 3_{10} regions. The G_i, G_{i+1} occurrence is shown with a color code ranging from green to red.

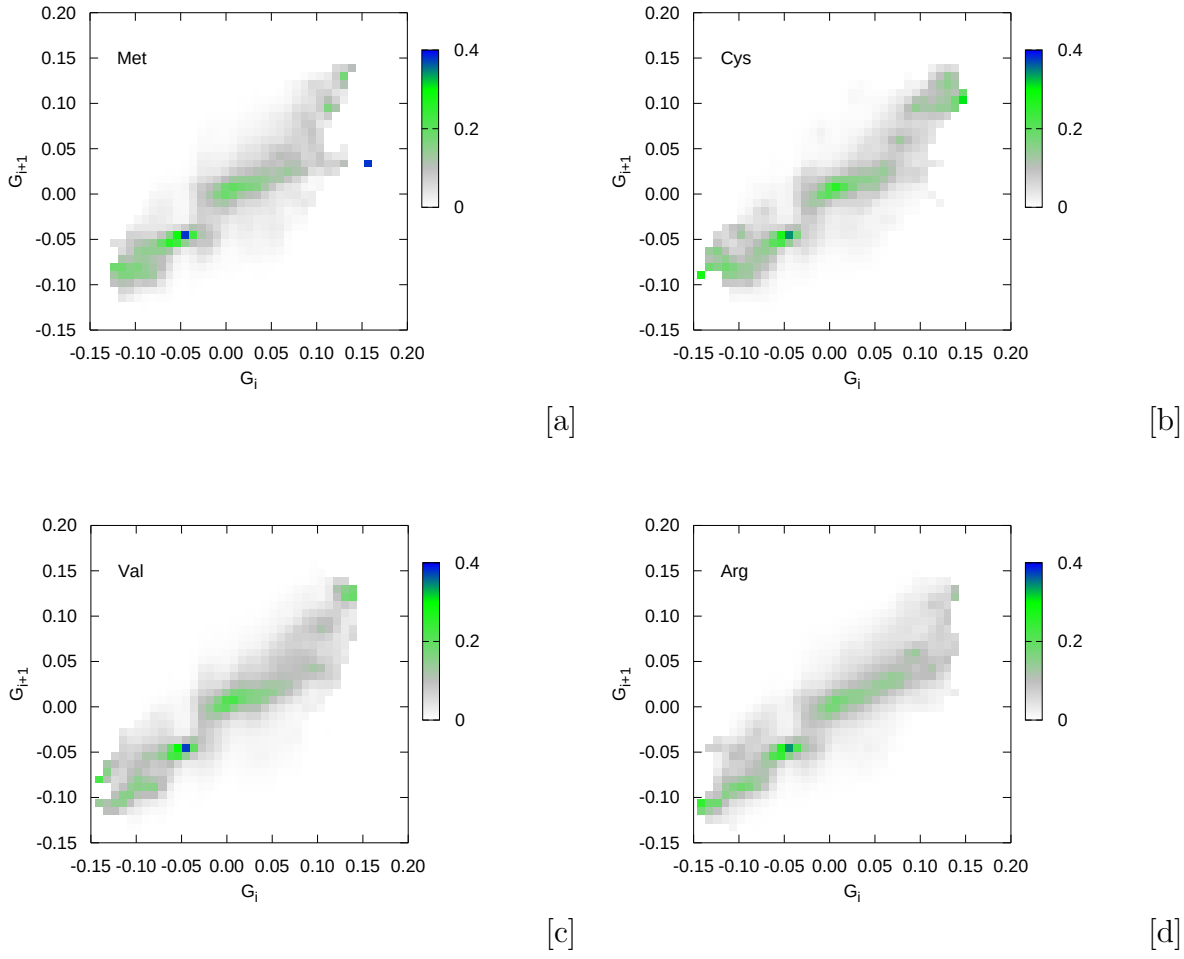


Figure 2.31: Conditional probability $P(G_{i+1}|G_i)$ calculated among the amino acid PDB database. The α helix range, here shown at G_i, G_{i+1} values (-0.05, -0.05), is present for Met [a], Cys [b], Val [c] and Arg residues [d]. In addition, in the Met map [a], a range at (0.15, 0.02) underlines the boundary with coil state. Cys [b], Arg [d] and in less extent Val [c] show also negative ranges of G_i, G_{i+1} values, typical of turn and 3_{10} regions. The G_i, G_{i+1} occurrence is shown with a color code ranging from green to red.

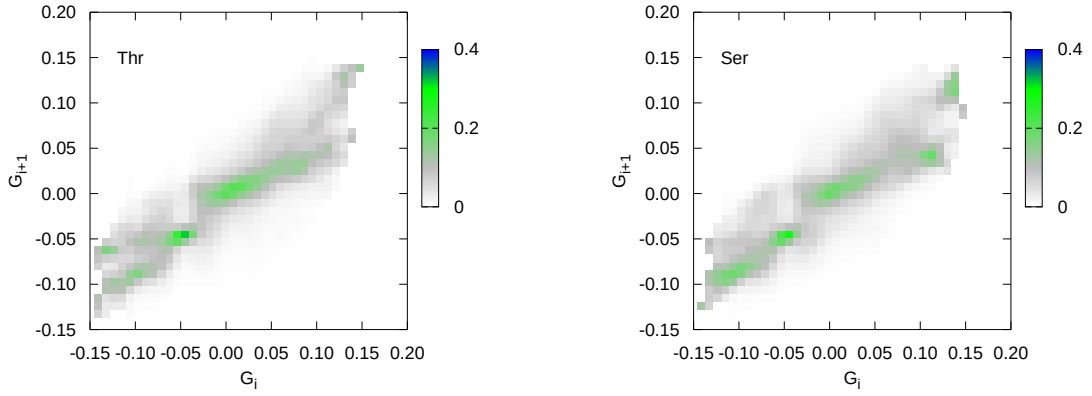


Figure 2.32: Conditional probability $P(G_{i+1}|G_i)$ calculated among the amino acid PDB database. The α helix range, here shown at G_i, G_{i+1} values (-0.05, -0.05), is present for Thr [a], and in less extent for Ser [b] and Lys [c]. Glycine shows the broadest map, as expected because of the high flexibility of this residue. The G_i, G_{i+1} occurrence is shown with a color code ranging from green to red.

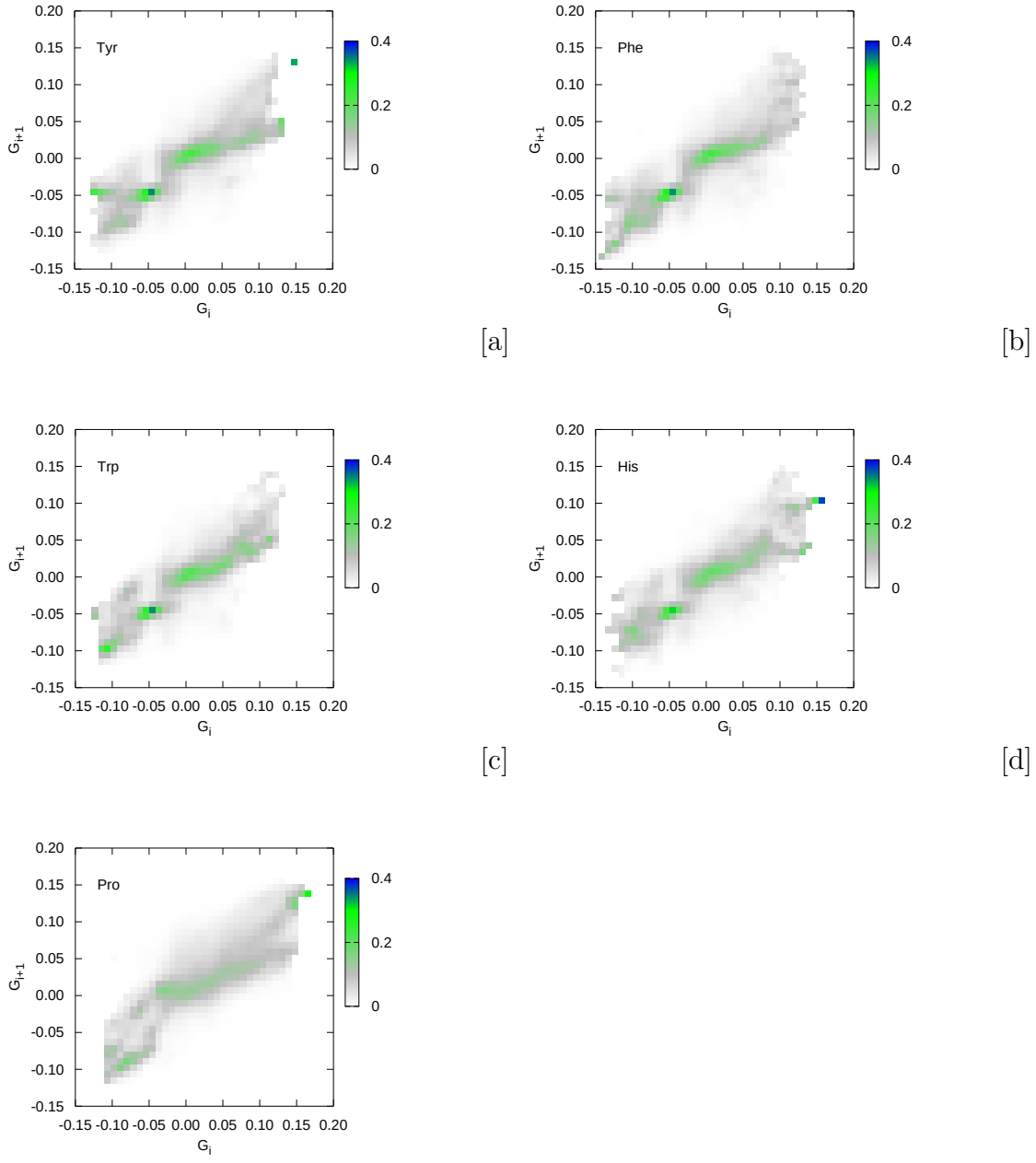


Figure 2.33: Conditional probability $P(G_{i+1}|G_i)$ calculated among the amino acid PDB database. The α helix range, here shown at G_i, G_{i+1} values (-0.05, -0.05), is present for Tyr [a], Phe [b], Trp [c] and in less extent for His [d]. Pro residue map [e] does not show, as correct, the range of α helix, but rather exhibits the PPII range (0.15, 0.13), also found in the Tyr [a] and His [b] maps. The G_i, G_{i+1} occurrence is shown with a color code ranging from green to red.

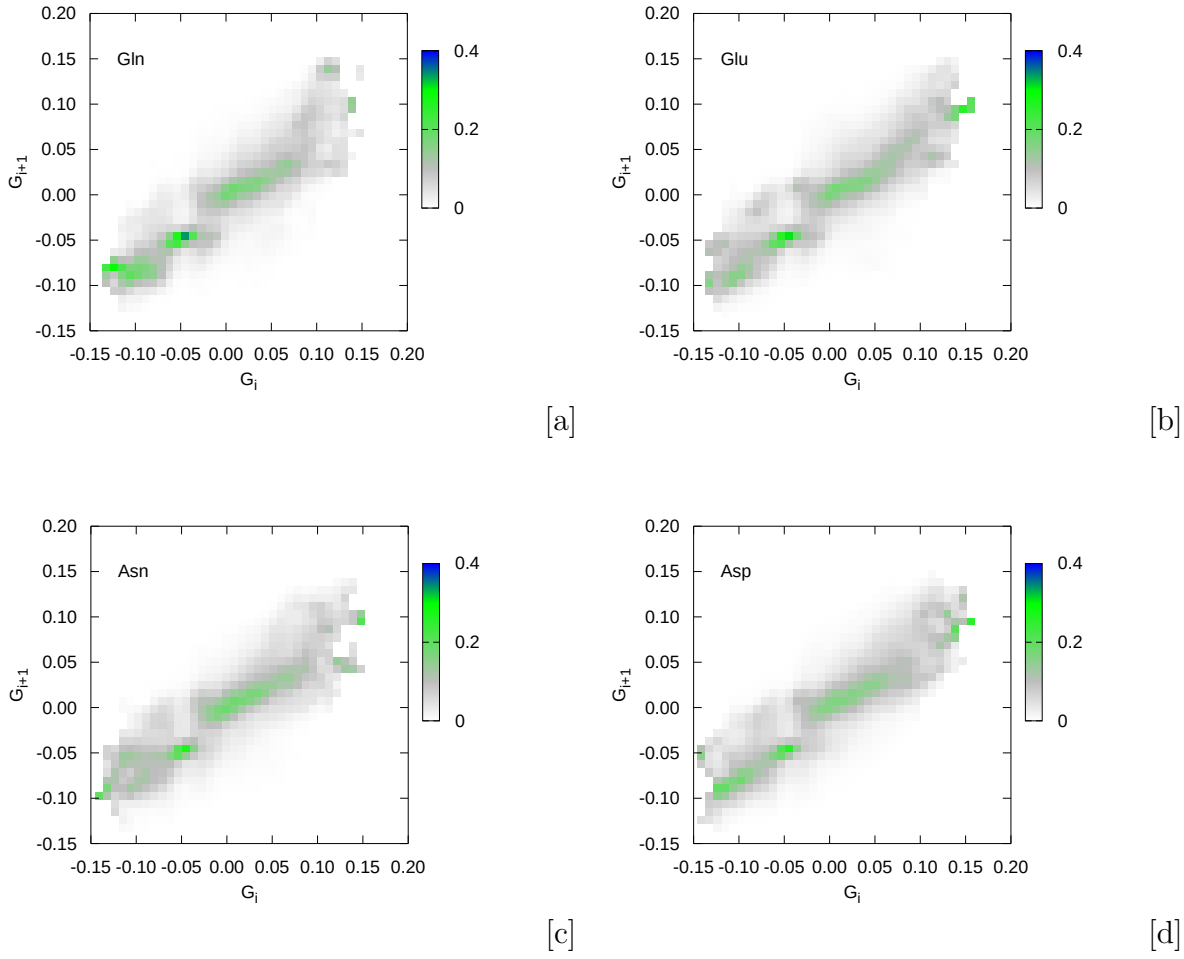


Figure 2.34: Conditional probability $P(G_{i+1}|G_i)$ calculated among the amino acid PDB database. The α helix range, here shown at G_i, G_{i+1} values (-0.05, -0.05), is present mainly in the Gln [a] and Glu [b] maps and in less extent in those one of Asp [c] and Asn [d].

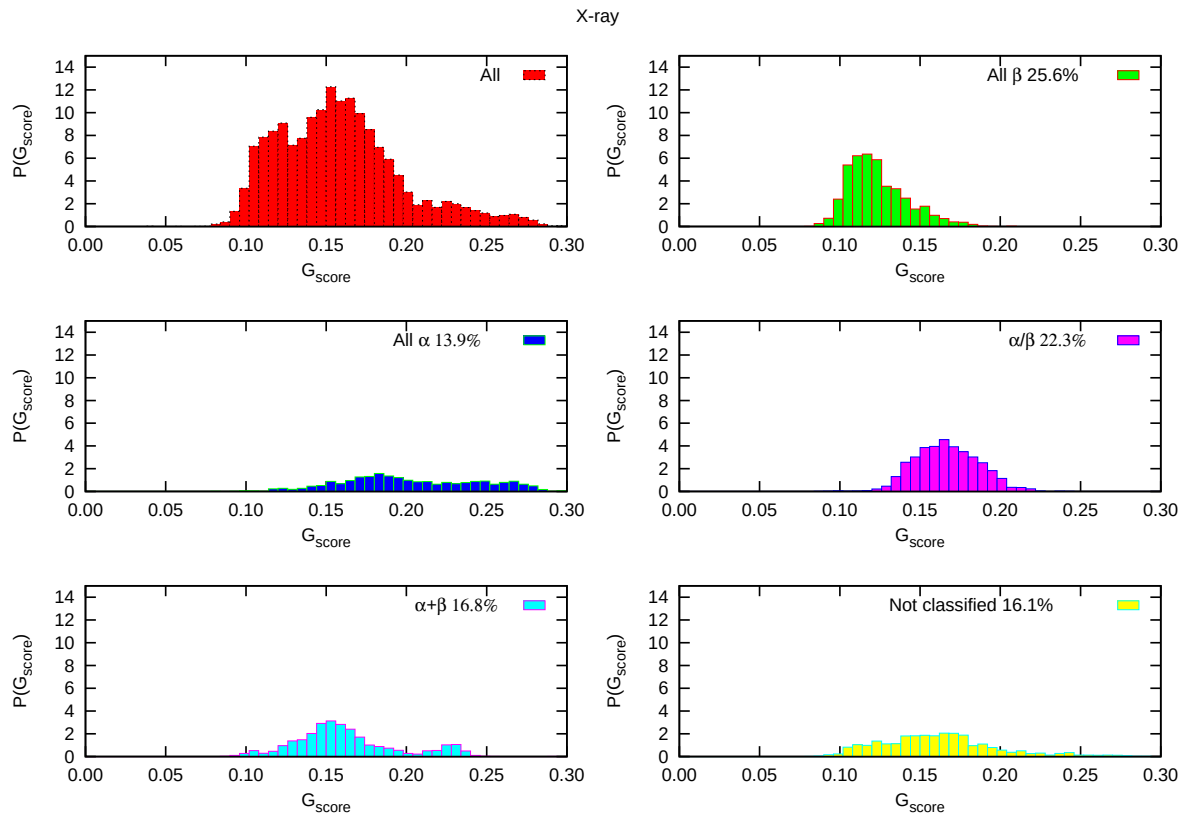


Figure 2.35: G_{score} distribution for the most populated classes of X-ray dataset, according to SCOP classification. It is worth noting higher values of G_{score} adopted from α helix proteins with respect the β ones.

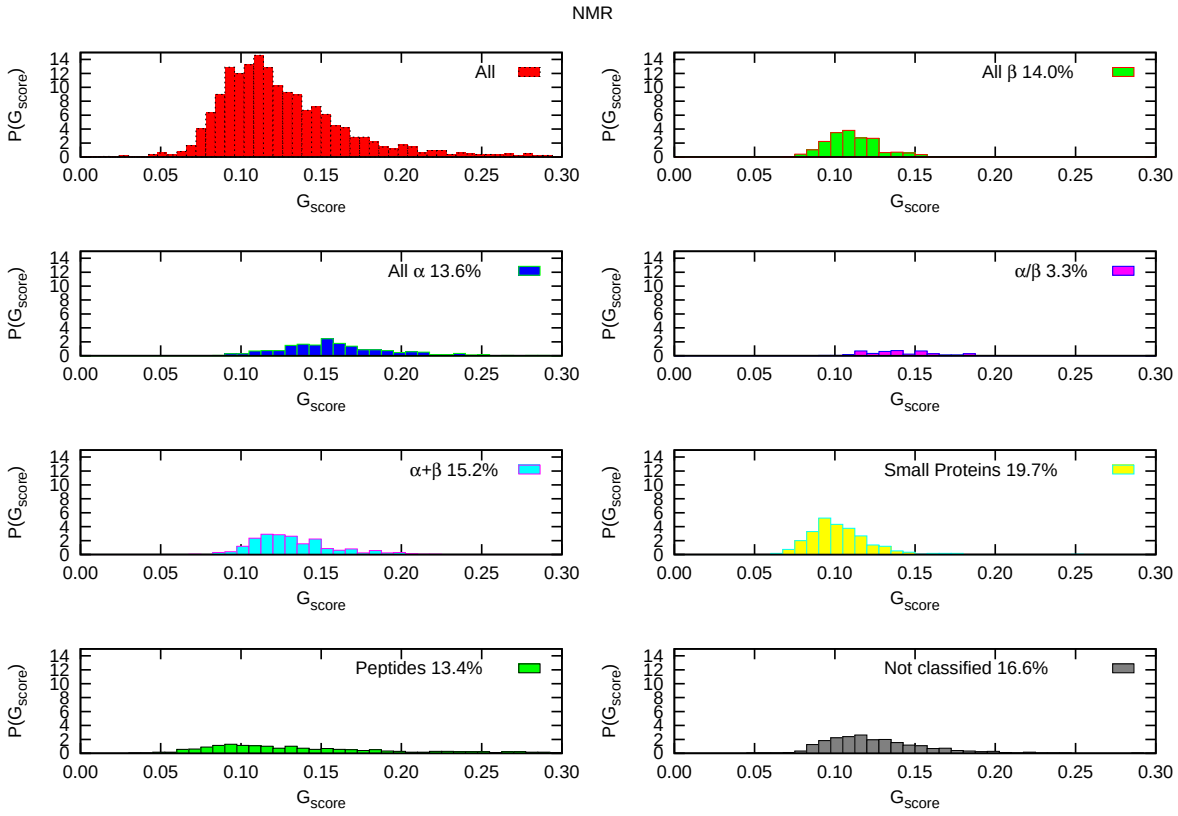


Figure 2.36: G_{score} distribution for the most populated classes of NMR dataset, according to SCOP classification. It is worth noting higher values of G_{score} adopted from α helix proteins with respect the β ones. The NMR α proteins adopt smaller G_{score} values than that one found for X-ray α proteins (Figure 2.35, as expected from the a higher mobility.

2.3.4 The persistence of the secondary structure and its correlation with the amino acid types

As a final investigation about the secondary structure adopted by the native proteins, it can be interesting to quantify the correlation between a particular amino acid and the persistence of its adopted secondary structure, here reflected in the values of chirality index adopted. A concise way of expressing such a correlation is to calculate the *Shannon* or *Information Entropy* [49], concerning the combined probability P(G_i, G_{i+1}):

$$H_2(G_i, G_{i+1}) = -\sum P(G_i, G_{i+1}) \log P(G_i, G_{i+1}) \tag{2.7}$$

and the single residue probability P(G_i):

$$H_1(G_i) = -\sum P(G_i) \log P(G_i) \tag{2.8}$$

where $P(G_i)$ and $P(G_i, G_{i+1})$ are normalized.
As it can be noticed from Table 2.6 Proline and Glycine show the higher H_2 entropy, strengthening the knowledge that, in many proteins, Proline and Glycine act as point breakers of the secondary structure. The low values of entropy of Alanine, Methionine, Isoleucine, instead, reflect the α helix propensity scale of these amino acids [50]. Of course, the entropy scale is not properly an α helix propensity scale. From the tendence of α helix in preserving the structure and from the abundance of α helices in the dataset, it derives that amino acids adopting α helix possess low values of entropy. This is clearly shown from the Shannon Entropy calculated for both the secondary structure distribution of the G_i, G_{i+1} chirality index and of G_i index (Tables 2.7, 2.8), which stress the idea that α helix is the most rigid and conserved conformation, especially with DSSP algorithm rather than the STRIDE one, which shows some differences from DSSP in the NMR set (see Table 2.7, 2.8). On the contrary, 3_{10} seems to be more flexible with respect to the other two helices, α and π.

Looking at the H_1 entropy for the edges reported in Table 2.8, in which the full, edges and the core (full-edges) distributions are reported, β sheets show smaller values of entropy, as the chirality index distribution suggested for such structures (see i.e. Figures 2.17, 2.21). The H_1 values of α helix edges, instead, are higher that those ones of β sheets. This indicates that the initial and final residues of α helix, classified

with DSSP and STRIDE, do not properly belong to α helix structure, but rather to a boundary with other conformational states, like left handed helices, such as PPII, or coil states.

Table 2.6: Shannon Relative entropy values for the twenty L-amino acids.
[a] Xray dataset, minimum entropy value 0.65;
[b] NMR dataset, minimum entropy value 0.67;
[c] Xray dataset, minimum entropy value 0.80;
[d] NMR dataset, minimum entropy value 0.83.

Amino Acids	$\mathbf{H}_2^a$	$\mathbf{H}_2^b$	$\mathbf{H}_1^c$	$\mathbf{H}_1^d$
PRO	0.110	0.110	0.083	0.059
GLY	0.089	0.085	0.075	0.055
SER	0.077	0.088	0.058	0.055
ASP	0.073	0.071	0.052	0.034
ASN	0.064	0.059	0.046	0.033
HIS	0.051	0.048	0.043	0.040
THR	0.051	0.046	0.040	0.030
LYS	0.043	0.046	0.037	0.030
GLU	0.035	0.037	0.021	0.011
CYS	0.035	0.074	0.034	0.046
TYR	0.034	0.031	0.030	0.017
ARG	0.033	0.034	0.028	0.024
PHE	0.033	0.016	0.029	0.008
GLN	0.031	0.023	0.025	0.011
TRP	0.027	0.025	0.027	0.009
ALA	0.012	0.023	0.011	0.008
LEU	0.008	0.010	0.017	0.009
VAL	0.007	0.006	0.002	0.000
MET	0.001	0.004	0.009	0.005
ILE	0.000	0.000	0.000	0.000

Table 2.7: Shannon Entropy (H_2) values for the secondary structures, according to DSSP and STRIDE classifications.
[a] DSSP, Xray dataset, minimum entropy value 0.54;
[b] DSSP, NMR dataset, minimum entropy value 0.47;
[c] STRIDE, Xray dataset, minimum entropy value 0.52;
[d] STRIDE, NMR dataset, minimum entropy value 0.21.

Structure	$\mathbf{H}_2^a$	$\mathbf{H}_2^b$	$\mathbf{H}_2^c$	$\mathbf{H}_2^d$
Bend	0.30	0.27	-	-
Coil	0.27	0.23	0.24	0.23
Bridge	0.26	0.19	0.24	0.18
Turn	0.25	0.21	0.24	0.22
3_{10}	0.25	0.21	0.23	0.18
Sheets	0.12	0.19	0.12	0.09
π helix	0.02	-	0.01	-
α Helix	0.00	0.00	0.00	0.00

Table 2.8: Shannon Entropy (H_1) values for the full-edge-core structures according to DSSP and STRIDE calssifications.
[a] DSSP classification of Xray dataset, minimum entropy values for full, edges and core 0.44, 0.54, 0.38 for full, edges and core respectively;
[b] STRIDE classification of X-ray dataset, minimum entropy values 0.46, 0.53, 0.41 for full, edges and core respectively;
[c] DSSP classification of Xray dataset, minimum entropy values for full, edges and core 0.55, 0.60, 0.45 for full, edges and core respectively;
[d] STRIDE classification of X-ray dataset, minimum entropy values 0.38, 0.37, 0.29 for full, edges and core respectively.

Xray						
SS	**full**[a]	**edges**[a]	**core**[a]	**full**[b]	**edges**[b]	**core**[b]
Bend	0.20	0.11	0.26	-	-	-
Coil	0.20	0.10	0.27	0.24	0.15	0.32
Bridge	0.23	0.14	0.30	0.17	0.11	0.23
Turn	0.23	0.14	0.30	0.18	0.11	0.23
3_{10}	0.17	0.08	0.19	0.22	0.14	0.26
Sheets	0.07	0.03	0.10	0.09	0.06	0.12
π helix	0.05	0.00	0.06	0.03	0.00	0.03
α Helix	0.00	0.05	0.00	0.00	0.11	0.00
NMR						
SS	**full**[c]	**edges**[c]	**core**[c]	**full**[d]	**edges**[d]	**core**[d]
Bend	0.25	0.16	0.30	-	-	-
Coil	0.23	0.15	0.30	0.23	0.10	0.30
Bridge	0.18	0.10	0.25	0.19	0.07	0.20
Turn	0.20	0.11	0.27	0.22	0.09	0.27
3_{10}	0.19	0.12	0.20	0.17	0.06	0.18
Sheets	0.06	0.03	0.09	0.07	0.00	0.08
π helix	0.07	0.00	0.05	-	-	-
α Helix	0.00	0.08	0.00	0.00	0.04	0.00

2.3.5 Conclusions

The chirality index previously proposed [46], was used to analyze a set of not obsolete protein structures contained in the protein databank. In particular it was shown the capability in correlating the chirality of a particular amino acid, with its preferred secondary structure and the persistence of a given structure as a function of the amino acids involved. To analyze this phenomenon, the conditional probability of G_i, G_{i+1}, $\mathrm{P}(G_{i+1}|G_i)$, was introduced. From this, it is possible to identify how the chirality of amino acid i influences the chirality and thus the secondary structure of the following one. To assess how compliant is a structure, the G_{score} quantity is introduced as a sum of all the conditional probability for the amino acids belonging to a protein structure. This quantity may help in studying the conformation adopted from a protein during molecular dynamics simulation and thus the equilibration process in the first steps of the run.
Moreover, all the amino acids were classified as a function of their capability in preserving the secondary structure, by using the *Shannon Entropy* of (G_i, G_{i+1}). From this analysis, Proline and Glycine are the most likely secondary structure breaker, as usual happens in many protein structures. Concerning the *Shannon Entropy* of the secondary structure, α helix is the most rigid among the protein conformations. This finding could clarify the role of helices in the misfolding pathway: the more flexible 3_{10} helix, instead of α, could in fact be involved in the early stage of these extremely important processes, explaining why a rigid conformation like α helix is able to unravel and thus misfolding in a pathogenic non *native* structure.

Notes
[*1]www.rcsb.org
[*2] ftp://ftp.cmbi.ru.nl/pub/molbio/data/dssp

Bibliography

[1] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, *J. Mol. Biol.*, **1963**, *7*, 95–99.

[2] T. Schlick, *Molecular Modeling and Simulation;* New York: Springer; 2002. 656 p.

[3] B. Zagrovic, J. Lipfert, E. J. Sorin, I. S. Millett, W. F. van Gunsteren, S. Doniach, and V. S. Pande, *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 11698–11703.

[4] A. Pietropaolo, L. Raiola, L. Muccioli, G. Tiberio, C. Zannoni, R. Fattorusso C. Isernia, D. La Mendola, G. Pappalardo, and E. Rizzarelli, *Chem. Phys. Lett.*, **2007**, *442*, 110–118.

[5] W. Kabsch and C. Sander, *Biopolymers*, **1983**, *22*, 2577–2637.

[6] K, Gō and N. Mizuguchi, *Protein Eng.*, **1995**, *8*, 353–362.

[7] C. A. F. Anderson, A. G. Palmer, S. Brunakand, and B. Rost, *Structure*, **2002**, *10*, 175–184.

[8] D. Frishman and P. Argos, *Proteins*, **1995**, *23*, 566–579.

[9] J. Martin, G. Letellier, A. Marin, J. F. Taly, A. G. de Brevern, and J. F. Gibrat, *BMC Struct. Biol.*, **2005**, *5*, 17.

[10] F. Dupuis, J. F. Sadoc, and J. P. Mornon, *Proteins*, **2004**, *55*, 519–528.

[11] F. M. Richards and C. M. Kundrot, *Proteins*, **1988**, *3*, 71–84.

[12] I. Majumdar, S. Sri Krishna, and N. V. Grishin, *BMC Bioinformatics*, **2005**, *6*, 202.

[13] B. J. Stapley, and T. R. Creamer, *Protein Sci.*, *1999*, *8*, 587–595.

[14] M. V. Cubellis, F. Caillez, T. L. Blundell, and S. C. Lovell, *Proteins*, **2005**,*58*, 880–892.

[15] J. I. Kwiecińska, and M. Cieplak, *J. Phys. Condens. Matter.*, **2005**, *17*, 1565–1580.

[16] E. Ruch, *Angew. Chem. Int. Ed. Engl.*, **1977**, *16*, 65–72.

[17] A. B. Buda and K. A. Mislow, *J. Am. Chem. Soc.*, *1992*, *114*, 6006–6012.

[18] M. A. Osipov, B. T. Pickup, and D. A. Dunmur, *Mol. Phys.*, **1995**, *84*, 1193–1206.

[19] A. Ferrarini and P. L. Nordio, *J. Chem. Soc. Perkin. Trans.*, **1998**, *2*, 455–460.

[20] A. B. Harris, R. D. Kamien, and T. C. Lubensky, *Rev. Mod. Phys.*, **1999**, *71*, 1745-1757.

[21] M. Solymosi, R. J. Low, M. Grayson, and M. P. Neal, *J. Chem. Phys.*, **2002**, *116*, 9875–9881.

[22] S. M. Todd, A. Ferrarini, and G. J. Moro, *Phys. Chem. Chem. Phys.*, **2001**, *3*, 5535–5541.

[23] M. P. Neal, M. Solymosi, M. R. Wilson, and D. J. Earl, *J. Chem. Phys.*, **2003**, *119*, 3567–3573.

[24] D. J. Earl and M. R. Wilson, *J. Chem. Phys.*, **2003**, *119*, 10280–10288.

[25] R. Berardi, G. Cainelli, P. Galletti, D. Giacomini, A. Gualandi, L. Muccioli and C. Zannoni, *J. Am. Chem. Soc.*, **2005**, *127*, 10699–10706.

[26] A. R. Kinjo, K. Horimoto, and K. Nishikawa, *Proteins*, **2005**, *58*, 158–165.

[27] L. Pauling, R. B. Corey, and H. R. Branson, *Proc. Natl. Acad. Sci. USA*, **1951**, *37*, 205–211.

[28] D. N. Marti, J. Schaller, and M. Llinás, *Biochemistry*, **1999**, *38*, 15741–15755.

[29] K. C. Chou, *Anal. Biochem.*, **2000**, *286*, 1–16.

[30] A. A. Adzhubei and M. J. E. Stemnberg, *J. Mol. Biol.*, **1993**, *229*, 472–493.

[31] C. Branden and J. Tooze, *Introduction to Protein Structure;* New York, NY: Garland Publishing; 1999. 410 p.

[32] L. Calzolai, D. A. Lysek, D. R. Perez, P. Güntert, and K. Wüthrich, *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 651–655.

[33] J. F. Bazan, R. J. Fletterick, M. P. McKinley, and S. B. Prusiner, *Protein Eng.*, **1987**, *1*, 125–135.

[34] S. A. Adcock and J. A. McCammon, *Chem. Rev.*, **2006**, *106*, 1589–1615.

[35] P. Procacci, E. Paci, T. Darden, and M. Marchi, *J. Comput. Chem.*, **1997**, *18*, 1848–1862.

[36] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *J. Am. Chem. Soc.*, **1995**, *117*, 5179–5197.

[37] M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids;* Oxford University Press, Oxford; 1989.

[38] S. A. Nosé, *J. Chem. Phys.*, 1984, *81*, 511–519.

[39] W. G. Hoover, *Phys. Rev. A*, **1985**, *31*, 1695–1697.

[40] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, *Intermolecular Forces;* Reidel, Dordrecht: B. Pullman, editor; 1981.

[41] P. Procacci and M. Marchi, *J. Chem. Phys.*, **1996**, *104*, 3003–3012.

[42] W. Humphrey, A. Dalke, and K. Schulten,*J. Molec. Graphics*, **1996**,*14*, 33–38.

[43] www2.fci.unibo.it/%7Eadriana

[44] V. Daggett, *Chem. Rev.*, **2006**, *106*, 1898–1916.

[45] D. J. Wales, *Energy Landscapes. With Applications to Clusters, Biomolecules and Glasses;* Cambridge: Cambridge U. P; 2003. 692 p.

[46] A.Pietropaolo, L. Muccioli, R. Berardi, and C. Zannoni, *Proteins*, **2008**, *70*, 667-677.

[47] J. K. Myers, C. N. Pace, and J. M. Scholtz, *Proc. Natl. Acad. Sci. USA*, **1997**, *94*, 2833–2837.

[48] C. Orengo, T. P. Flores, W. R. Taylor, and J. M. Thornton. *Protein Eng.*, 6:485–500, 1993.

[49] E. T. Jaynes. *Phys. Rev.*, 106:620–630, 1957.

[50] C. N. Pace and J. M. Scholtz. *Biophys. J.*, 75:422–427, 1998.

Chapter 3

Structure determination using NMR: the role of Residual Dipolar Couplings in the protein refinement

3.1 Introduction

In chapters 1 and 2, the study of protein structure with a computational approach was described. In this chapter, instead, the structure determination of proteins by means of NMR investigations is introduced. A number of new refinement strategies, aimed at both facilitating NMR structure determination and increasing the accuracy of the resulting structures, are actually feasible. These include direct refinement against three–bond coupling constants and ^{13}C and ^{1}H shifts. More recently, methods have been developed to obtain structural restraints that characterize long range order; these methods include the residual dipolar contributions to one–bond hetero–nuclear couplings arising from small degrees of alignment of molecules in a magnetic field. In the following, it is shown how Residual Dipolar Couplings, RDCs, may help in improving the structure resolution.

NMR structural biologists are always seeking ways to increase the size limit of biological macromolecules that are amenable to study and to expand the range of biological questions that can be addressed. Recent methods such as TROSY [1] and protein labeling strategies [2] as well as the availability of higher magnetic fields have dramatically increased the size of macromolecules that can be studied by NMR. However, the ability to study larger macromolecules, in of itself, still does not allow one to answer many

relevant questions, particularly those pertaining to the global structure and domain interactions. This is due to the fact that until recently the principal NMR data for structure determination were the NOE and scalar J couplings, which are entirely local in nature.

Any structure determination by NMR tries to find the global minimum region of a target function E_{tot} given by: $E_{tot}=E_{cov} + E_{vdw} + E_{NMR}$, where E_{cov}, E_{vdw}, and E_{NMR} are terms representing the covalent geometry (bonds, angles, planarity, and chirality), the nonbonded contacts, and the experimental NMR restraints, respectively [3]. Algorithms currently used include simulated annealing in both Cartesian [4, 5] and torsion angle space [6], metric matrix distance geometry [7], and minimization with a variable target function in torsion angle space [8]. The main source of geometric information contained in the experimental NMR restraints is provided by the nuclear Overhauser effect (NOE). The NOE (at short mixing times) is proportional to the inverse sixth power of the distance between the protons, thus its intensity falls off very rapidly with increasing distance between proton pairs. Consequently, NOEs usually are observed only for proton pairs separated by maximum 5 or 6 Å. Despite the short range nature of the observed interactions, approximate interproton distance restraints, derived from NOE measurements, can be highly conformationally restrictive, particularly when they involve residues that are far apart in the sequence but close together in space [3, 9]. Systematic bias arising from the different algorithms used to calculate the structures may be introduced via the first two terms, E_{cov} and E_{vdw}, in the equation above. The values of bond lengths, bond angles, planes, and chirality are known to very high accuracy, so it is clear that the deviations from idealized geometry, as represented by the term E_{cov}, should be kept very small. The second term, E_{vdw}, representing the nonbonded contacts, is associated with considerably more uncertainty than the covalent geometry [10, 11]. Given the numerous ways to represent E_{vdw} (for example, a simple van der Waals repulsion term or a complete empirical energy function including a van der Waals Lennard-Jones 6-12 potential), it is evident that variability is introduced via E_{vdw}. It is therefore essential to ensure that the calculated structures display good nonbonded contacts. The uncertainties associated with the covalent geometry and van der Waals terms can introduce errors of 0.3 Å in the coordinates [11]. The major determinant of accuracy, however, resides in the number and quality of the experimental NMR restraints that enter into the third term, E_{NMR}. Although a high resolution, carefully refined an x-ray structure of a given protein may not be identical to the *true* solution structure, it is likely

to be reasonably close in many instances, as evidenced, for example, by the excellent agreement (1 Hz rms deviation) between the experimentally determined values of $^3J_{HN\alpha}$ three bond coupling constants in solution and their corresponding calculated values from crystal structures [12–14]. Moreover, it is generally the case that three-bond coupling constants, ^{13}C secondary shifts, and ^1H shifts calculated from high resolution crystal structures agree better with the experimentally measured values than those calculated from the corresponding NMR structures (refined in the absence of coupling constant and chemical shift restraints) [10], [12–15]. It is therefore instructive to examine the dependence of the backbone rms difference between NMR and x-ray structures on the precision of the NMR structures [10]. The accuracy of NMR structures will be affected by errors in the interproton distance restraints. These errors can arise from two sources: (i) misassignments and (i) errors in distance estimates. Errors due to misassignments may be quite common in low resolution NMR structures. Fortunately, in many cases, these errors are of relatively minor consequence and do not result in the generation of an incorrect fold. Systematic errors in distance estimates may be introduced in attempts to obtain precise distance restraints. For example, interactive relaxation matrix analysis of the NOE intensities [17] and direct refinement against the NOE intensities [18,19], while accounting for spin diffusion, can result in systematic errors from several sources such as the presence of internal motions (not only on the picosecond time scale but also on the nanosecond to millisecond time scales), insufficient time for complete relaxation back to equilibrium to occur between successive scans, and differential efficiency of magnetization transfer between protons and their attached heteronucleus in multidimensional heteronuclear NOE experiments [11]. In the case of experimental structures calculated with an incomplete set of NOE restraints (i.e., comprising 90% of the structurally useful NOEs), there is no doubt that errors, arising both from misassignments as well as from the incorrect classification of NOEs into the various loose approximate distance ranges, will occur, resulting in less accurate structures. This loss in accuracy is due to the fact that, until a significant degree of redundancy is present in the NOE restraints, such errors often can be accommodated readily without unduly comprising the agreement with either the experimental NMR restraints or the restraints for covalent geometry and non-bonded contacts.

For large molecules, having short correlation time, the NOE cross peak and the exchange cross peak is the same. It is therefore impossible to distinguish NOE from chemical exchange. In this case, the ROESY (NOESY in the rotating frame) pulse se-

quence should be used. Differently from NOE that can be positive (for small molecules), negative (for large molecules) or null (if the correlation time happens to cancel the NOE), the ROE (NOE in the rotating frame) is always positive. Scalar J couplings are related to torsion angles by the Karplus [20] equation, $^3J(\lambda)=A\cos^2(\lambda)+B\cos(\lambda)+C$, where ^3J is the three bond coupling constant, λ is the torsion angle corresponding to the bond coupling, and A, B, C are constants obtained by nonlinear optimization to yield the best fit between experimental ^3J values and values calculated from a series of very high resolution x-ray structures. The coupling constants can be converted directly into loose torsion angle restraints [3]. Alternatively, direct refinement against coupling constants can be achieved by adding the potential $E_J=k_J(J_{obs} - J_{calc})^2$, where k_J is a force constant and J_{obs} and J_{calc} are the observed and calculated values of the coupling constants. From the standpoint of refinement, the most useful coupling constant, in so far that it can be measured accurately and easily by quantitative J correlation spectroscopy and that its Karplus relationship has been parametrized reliably, is the $^3J_{HN\alpha}$ coupling, which is related directly to the backbone torsion angle [21].

Nowadays, it is common practice to refine structures using Residual Dipolar Couplings, RDCs. RDCs have dramatically altered the types of applications to which NMR methods can be applied. RDCs are complementary to NOEs; they provide orientational information, both short range and long range. Similar to NOEs, RDCs are utilized as restraints in molecular dynamics calculations. In contrast to an NOE, which provides a distance restraint between two atoms, an RDC contains distance information as well as angles formed by a vector connecting the two atoms within a tensor axis system. However, within the past few years there has been an explosion in the number of systems and problems that have been studied using RDCs and many of these applications address unresolved structural discrepancies among previous NMR structures, crystal structures, and other biophysical data.

3.1.1 Theoretical Framework

In NMR, the r^{-6} dependence of NOEs means that NOEs can usually be detected only between protons within 5 Å. This information is both short range and local; the presence of two pairs of NOEs does not provide any information on how they are related to each other. In contrast, because the dipolar coupling is defined in terms of a molecular coordinate frame, the measurement of two dipolar couplings provides orientational in-

formation on how each dipole is related to the molecular coordinate frame and in turn, to each other. The dipolar coupling is a useful phenomenon by means of characterizing a structure because it depends on distance, orientation, and dynamics. Dipolar couplings have long been a mainstay in solid-state NMR, but recent developments have made them routine in solution NMR. The dipolar coupling is a through-space interaction that arises between any two magnetically active nuclei. As a result of the effects of Brownian motion, dipolar couplings average to zero under isotropic conditions and are only observed under anisotropic conditions. For two dipole-coupled nuclei, A and B, the observable dipolar coupling in solution, D_{AB}, can be expressed as:

$$D_{AB}(\theta,\phi) = \frac{1}{2} D_{AB}^{max} \left[A_a^{AB} \left\{ (3cos^2\theta - 1) + \frac{3}{2} R(sin^2\theta cos2\phi) \right\} \right] \tag{3.1}$$

$\mathrm{A}_a{}^{AB}$ and R are the axial and rhombic components, respectively, of the molecular alignment tensor, **A**, in the principal coordinate frame. According to typical convention, the magnitudes of the principal components are $|\mathrm{A}_{zz}| \geq |\mathrm{A}_{yy}| \geq |\mathrm{A}_{xx}|$. $\mathrm{A}_a{}^{AB}$ is equal to $1/3[\mathrm{A}_{zz}{}^{AB}$- $(\mathrm{A}_{xx}{}^{AB} + \mathrm{A}_{yy}{}^{AB})/2]$ and $\mathrm{A}_r{}^{AB}$ is equal to $1/3[\mathrm{A}_{xx}{}^{AB}$ - $\mathrm{A}_{yy}{}^{AB}]$. $\mathrm{A}_a{}^{AB}$ is in units of hertz and R, which is equal to $\mathrm{A}_r{}^{AB}/\mathrm{A}_a{}^{AB}$, is unitless and always positive. θ is the angle between the internuclear bond vector and the z axis of the alignment tensor, ϕ is the angle between the projection of the internuclear bond vector onto the x-y plane and the x axis. D_{AB} is equal to a

$$D_{AB}^{max} = -\left(\frac{\mu_0 h}{16\pi^3}\right) \gamma_A \gamma_B \left\langle S(cos\theta_{AB}) r_{AB}^{-3} \right\rangle \tag{3.2}$$

where μ_0 is the permeability in a vacuum, h is Planck's constant, S is the generalized order parameter, γ_A, γ_B are the gyromagnetic ratios of the two nuclei, β_{AB} is the angle between the internuclear vector, AB, and the director axis, $|r^{-3}|$ is the inverse cube of the internuclear distance. In the applications presented here, $\langle r^{-3} \rangle$ between directly bonded nuclei is known and S is generally assumed to be constant, thus $\langle r^{-3} \rangle = \langle r \rangle^{-3}$ and the θ and ϕ angles are the only variables that contribute to the values of the RDC. To extract dipolar coupling data, the molecule must behave anisotropically. Otherwise, there is no preferred orientation and the average value of D_{AB} is 0.

In solution NMR, solutes behave according to Brownian motion and the dipolar interaction averages to zero. To obtain RDCs in solution, a cosolute is needed that causes a partial alignment and a net nonzero average value without causing severe

coupling interactions and distorted spectra. This allows RDCs to be observed while retaining the overall simplicity of solution NMR spectra [22]. Dipolar couplings on fully aligned samples such as solids are typically tens of kHz, whereas dipolar couplings from partially aligned solution samples are usually under 100 Hz. RDCs can be observed in molecules that have a sufficiently large magnetic susceptibility anisotropy such as metalloproteins with paramagnetic centers or diamagnetic systems such as DNA where the small anisotropy in each base is additive over the entire molecule. The magnetic susceptibility causes a field-dependent alignment of molecules [23].

3.1.2 Measurements of RDCs

Liquid crystals for the purposes of alignment in NMR were first introduced in 1963 by Saupe [24, 25] to study small molecules, but the concentrations used led to multiple dipolar couplings for individual nuclei, thus making the spectra more difficult to resolve. Bicelles were introduced in the early 1990s and have since been used extensively to achieve a sufficient degree of alignment [26, 27]. Bicelles are disk- shaped particles that are made from the detergents DMPC and DHPC*, typically in a ratio of 3:1. The concentrations in NMR samples are usually 5% (w/v), but the degree of protein alignment can be tuned by adjusting the bicelle concentration. The alignment of bicelles is temperature dependent. At room temperature the bicelles behave isotropically, but at higher temperatures (37 C) they take on a liquid crystal behavior, aligning with their normal, perpendicular to the direction of the magnetic field [28]. The mechanism by which the neutral bicelles exert their orienting properties is thought to be primarily due to steric hindrance [29]. The degree of alignment can be determined by measuring the 2H quadrupolar splitting in the HDO resonance. The splitting arises from exchange between isotropic bulk H_2O and aligned H_2O molecules associated with the aligned bicelle. In addition to bicelles, many other types of alignment media and protocols have been developed. These include bicelles using different detergents, phage particles, purple membrane fragments, strain-induced gels, and CPCl/hexanol*.

In NMR spectra, RDCs appear as an additional contribution to the scalar J coupling splitting. The magnitude of D_{AB} can be positive or negative and must be determined by taking the difference of the splitting under anisotropic conditions (J + D) and under isotropic conditions (J). Methods for measuring RDCs have been described previously [30] - [32]. It is possible to alter some of the alignment cosolutes, thus producing a

different alignment tensor. Bicelles, for example, can be doped with small charged amphiphiles to alter their charge. CTAB* confers a positive charge on bicelles, whereas SDS* confers a negative charge. In addition, salt and pH can also change the alignment tensor. The choice of which aligning medium to use is protein dependent and usually determined empirically.

* see List of abbreviations

3.1.3 Determination of A_a and R

To use RDCs in any type of structure refinement, good estimates for A_a and R in Equation 3.1 must be available. There are several methods for determining A_a and R, and the choice of which one to use depends in part on whether a reasonably accurate structure is available prior to the refinement. In the histogram method demonstrated by Clore et al. [33], the RDCs are measured, normalized to account for the properties of different nuclei, and plotted in a histogram. This histogram closely resembles a chemical shift anisotropy (CSA) powder pattern spectrum characteristic of solid-state NMR spectra, where the values of the chemical shift tensor can be estimated from the pattern. Values for A_{zz}, A_{yy}, and A_{xx}, are taken from the three extrema of the histogram (see Figure 3.1). These values can be used with Equations 3.3, 3.4, and 3.5 to solve for A_a and R.

$$A - B \text{ lies along } D_{zz} : \theta = 0 \; A_{zz} = 2A \tag{3.3}$$

$$A - B \text{ lies along } D_{yy} : \theta = 90°, \phi = 90° \; A_{yy} = -A_a \left\{ 1 + \frac{3}{2} R \right\} \tag{3.4}$$

$$A - B \text{ lies along } D_{zz} : \theta = 90°, \phi = 0° \; A_{zz} = -A_a \left\{ 1 - \frac{3}{2} R \right\} \tag{3.5}$$

This method can be used in cases where no previous structural information is available. The key to using this method successfully is that the ensemble of RDCs must sample a wide range of θ and ϕ. For many types of biomolecules this is unlikely to be the case. Another approach put forth by Clore et al. [34] makes use of a grid search to determine A_a and R. First, A_a is estimated by taking the average of the low RDC values:

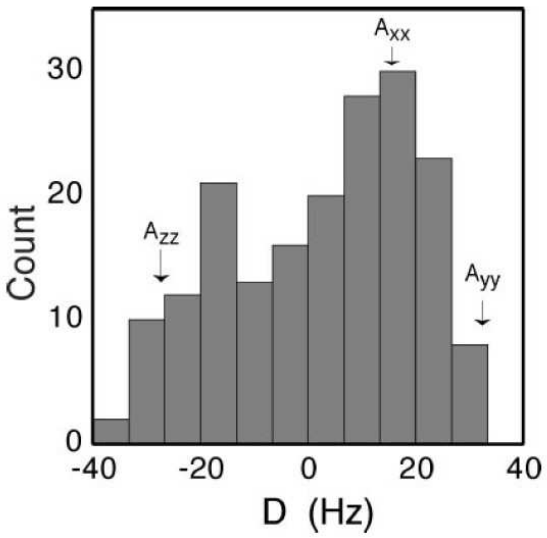


Figure 3.1: A histogram of normalized RDCs, ${}^1\mathrm{D}_{NH}$, ${}^1D_{H\alpha C\alpha}$ ${}^1D_{CN}$, ${}^1D_{C'C\alpha}$, for the KH3 domain from ribonucleoprotein [35].

$$A_a = \frac{-A_{min}}{1 + 1.5R} \tag{3.6}$$

Then a series of short simulated annealing calculations are performed with the value of A_a from Equation 3.6 while varying the value of R. The premise of this method is that a structure refined in a simulated annealing protocol with the correct value of A_a and R will have the lowest overall energy.

In cases where a fairly accurate structure is available, single value decomposition can be used to fit RDCs to a series of linear equations to determine the direction cosines that are then used to determine the three principal components of the Saupe order matrix, $\mathbf{A}_{ij}$ [36]. The $\mathbf{A}_{ij}$ order matrix using Cartesian coordinates is a 3 $\times$ 3 matrix in which the subscripts i and j refer to the x, y, or z axes of the alignment tensor. Since the order matrix is both symmetric and traceless only five RDCs are required to define it. The three principal components of the alignment tensor are determined by diagonalizing the matrix. Since the matrix contains five unknowns, measurement of five individual RDCs within a rigid structure is sufficient to determine the alignment

tensor. ${D_{AB}}^{max}$ in Equation 3.2 contains the generalized order parameter S, which is assumed to be constant. Thus, if the structural region used for the five measurements is not sufficiently rigid, then the calculated alignment tensor is actually an average alignment tensor. If the input structure is not good enough, it can be difficult to obtain a unique solution for A_a, R, and the Euler angles. Some of the above methods may be difficult to apply to RNA and DNA. These molecules have a lower proton density and are accompanied by limited spectral dispersion, and because of the high helical content, the RDCs do not always adequately sample enough orientations to obtain unbiased solutions. Warren and Moore [37] have presented a method for determining A_a and R specifically for oligonucleotides. This protocol is based on the maximum-likelihood method [38], a strategy also used to map efficiently the conformational space of flexible molecule in solution [39]. An initial family of structures is calculated using only NOE and dihedral angle restraints [40]. This family is used to generate a range of R values. A_a is determined for each value of R by the histogram method. Another family of structures is calculated with RDC restraints using this range of R and A_a values. The refined structures are used as input and this protocol is repeated until the range of values for R and A_a converge.

3.1.4 Data Refinement

For the purpose of using RDCs to calculate NMR structures, RDCs are usually used not in initial structure calculations but rather in a refinement stage of structure calculations. The reasons are that the potential energy surface is very rough and including RDCs initially may trap the structure into a false minimum, leading to convergence problems [41]. A module for incorporating RDCs into structure calculations has been developed for use in XPLOR-NIH [42], but many softwares like CYANA [43] or variants are commonly used. This protocol includes a target function in the form of a quadratic harmonic potential,

$$E_{dip} = K_{dip}(\mathbf{D^{calc}} - \mathbf{D^{meas}})^2 \tag{3.7}$$

where D^{calc} and D^{meas} are the calculated and measured RDCs, respectively, k_{dip} is the force constant, and E_{dip} is the dipolar energy. The force constant should be chosen so that the dipolar RMS is equal to the error in the measured RDCs, approximately 0.2–2.0 Hz.

The alignment tensor is specified by a four-atom pseudo molecule, OXYZ. O represents the center of the molecule and the atoms X, Y, and Z represent the three orthogonal axes of the tensor. The orientation of the alignment tensor with respect to the molecular coordinate frame is determined during the simulated annealing [44]. The simulated annealing protocol attempts to shift the bond vector orientations to minimize the difference between measured and calculated RDCs in a manner that is consistent with other experimental data without distorting the covalent structure.

Inclusion of RDC restraints generally improves the precision of families of structures (RMSD from x-ray structure). A common measure of improvement is an increase in the number of residues that falls in the most favored region on a Ramachandran plot. The Q factor and R_{dip} (equations 3.8 and 3.9, respectively) are qualitative measures of the agreement between RDCs that are not used in the structure calculation with the other structural restraints used during the simulated annealing calculations. The calculated RDCs can be determined from a refined structure or another structure that is being compared, such as a homologous crystal structure. The Q factor is:

$$Q = \left\{ \frac{\sum_{i=1,N}(D_i^{meas} - D_i^{calc})^2}{\sum_{i=1,N}(D_i^{meas})^2} \right\}^{1/2} \tag{3.8}$$

R_{dip} is the same as the Q factor but in a form similar to the crystallographic free R factor [45].

$$R_{dip} = \left\{ \frac{5(D_i^{meas} - D_i^{calc})^2}{[2(D_a^{AB})^2(4 + 3R^2)]} \right\}^{1/2} \tag{3.9}$$

Both the Q factor and R_{dip} vary from 0 to 1; a lower value indicates better agreement. An NMR structure refined with dipolar couplings should have a Q factor as low as 0.16 [46].

3.1.5 Determination of protein folds from RDCs

Identifying NOEs is an extremely time-consuming endeavor. Although, a protein structure is difficult to calculate with only RDCs as the experimental restraints, they can be used to expedite this process by determining a protein fold. Determination of the protein fold is in of itself quite valuable, especially in the era of proteomics, in which

determination of the fold of an unknown protein often yields the first clues regarding its function. Some of the concepts behind protein fold determination are useful for other problems such as identifying different ligand binding conformations.

Homologous RDCs can be used to identify the ϕ and ψ backbone dihedral angles that are associated with a particular conformation that can then be used to construct a model of the protein [47]. The protein is broken into overlapping seven residue fragments and the RDCs from each fragment are compared to calculated RDCs from seven-residue fragments in high-resolution crystal structures in the PDB. Twenty matches are selected for each fragment on the basis of a 2 value that is calculated from the agreement between the measured and calculated RDCs and, to a lesser extent, from chemical shift agreement. This process is repeated for every possible seven-residue fragment in the protein, yielding 100 pairs of ϕ, ψ angles for each residue. A protein model is then calculated on the basis of these derived ϕ and ψ angles. Cases where the spread in ϕ, ψ angles from the PDB matches is not narrow indicate that alternative conformations might satisfy the same RDCs, and those ϕ, ψ angles should be used with caution.

In cases of proteins that have multiple binding partners, the binding mode can be established with $^1D_{NH}$ RDCs instead of binding assays or full-structure determinations. Sequence alone predicts the mode of binding reasonably well, but RDCs provide a more robust analysis in cases where there are more than two potential hydrophobic anchoring residues.

3.2 The effect of RDCs on the lysozyme structure resolution

To study the influence of RDCs on protein structure refinement and to quantify the entity of this effect, different sets of structure calculations including or not RDCs were carried out. As a model, the lysozyme protein was chosen. This, because of its relative large size (129 amino acids) and because lysozyme possesses defined secondary structure elements, with both region of helix (α and 3_{10}) and β sheets. Moreover, it is a protein extensively studied by means of both NMR and high resolution X-ray, implying that the structure is known with high resolution and thus being a good model for testing the role of residual dypolar coupling in the structure determination.

3.2.1 Structure Calculations

The N-H^N RDCs and the NOEs sets

Preliminarily, the role of RDCs was studied chosing a restricted sets of NOEs restraints, involving H^N-H^N, H^N-H^α, H^N-H^B(ALA), H^N-H^G(VAL), H^N-H^D(LEU), H^N-H^G(ILE), H^N-H^D(ILE). This, because such NOEs are easy to identify with only a ^{15}N labelled protein and 3D spectra recorded. The sets chosen differ from the addition of RDCs; in particular set a includes 3 set of N-H^N RDCs at pH 3.8 and 308 K, in the media, 6% polyacrilamide gel, 5% c12E6/hexanol and [D13OPC]: [DHOPC]: [CTAB]=30:10:1, while set b includes 3 set of N-H^N RDCs at pH 6.5 and 308 K, in the media ester bicelles, 7.5 % DHPC/DMPC ester bicelles doped with CTAB and 10mg/ml PF1 bacteriophage in 500 mm NaCl. In set c any NOEs contacts were added; set d includes both the two sets of RDCs of a and b.

The inclusion of C^α-H^α, C^α-C', C'-N RDCs and the other NOEs sets

The inclusion of carbons data was also analyzed. In particular, set e includes the 680 NOEs, previously introduced and 1 set of N-H^N, C^α-H^α, C^α-C', C'-N of [D13OPC]: [DHOPC]: [CTAB]=30:10:1, and 2 set of N-H^N RDCs at pH 3.8 and 308 K namely 6% polyacrilamide gel, 5% c12E6/hexanol at pH 3.8 and 308 K together with the RDCs of set b; set f does not include RDCs; set g includes the 680 NOEs restraints and the RDCs of set d, starting from the 100 structures of set f.

The inclusion of the full set of NOEs restraints together with RDCs

In order to investigate the entity of the improvement with larger data available, the NOEs restraints previously reported by Schwalbe et al. [48] were also added, In particular, set h includes these latter NOEs [48] and the sets of N-H^N RDCs included in set d; Set i includes such NOEs and all the RDCs of set e, thus including carbon data.

Calculation details

All the structures were calculated using XPLOR-NIH [49] and the PARALLDG5.1 [50] force field. A harmonic potential was used for the dipolar coupling restraints. The estimations of D_a and R were obtained for the 6 sets of dipolar couplings using the grid search module in XPLOR-NIH [49] (D_a= 7.36, R= 0.08 for the [D13OPC]: [DHOPC]:

[CTAB]=30:10:1 data; D_a= 4.37, R= 0.41 for 6% polyacrilamide gel; D_a= 9.75, R= 0.41 for 5% c12E6/hexanol; D_a= -8.70, R= 0.11 for the 10mg/ml PF1 bacteriophage in 500 mm NaCl; D_a= 13.39, R= 0.15 for 7.5 % DHPC/DMPC ester bicelles doped with CTAB and D_a= 15.96, R= 0.32 for ester bicelles.) A simulated annealing protocol starting from random coordinates was used; 25000 steps of cooling from 1500 K to 100 K, followed by a final 5000 steps of energy minimization were carried out; a geometry distance protocol was also used for set f. From the 100 structures obtained, 50 of them with the lowest energy were analyzed. The stereochemical quality of the structures was analyzed with the program PROCHECK [51] and the secondary structure regions were identified with the chirality analysis [52] (see chapter 2). To assess the quality of the structure, the Q factor was calculated as shown in section 3.1.4, equation 3.8, for the following media: medium I :[D13OPC]: [DHOPC]: [CTAB]=30:10:1 at pH 3.8; medium II: 10mg/ml PF1 bacteriophage in 500 mm NaCl at pH 6.5.

3.2.2 Structure analysis

The effect of adding N-H^N RDCs

The analysis of the structures from the different sets of calculations reveal strong differences which are a function of the constraints used. In particular, the sets with only NOEs constraints show a low accuracy (RMSD from the mean structure), and interestingly the RMSD from the X-ray structure is very high in magnitude (see Table 3.1). These data underline that the minimum of energy sampled from the calculations with only the limited set of NOEs restraints, is not adopted by the crystal state of lysozyme. Adding the N-H^N RDCs constraints the resolution of the structure improves, mainly for set a, as it can be noticed from Table 3.1.

To further analyze the conformational space of lysozyme using only the limited set of NOEs restraints, structure calculations with a distance geometry protocol was carried out for set f. The ensembles sampled are in two different minima of energy and only one of them is similar to that one adopted by the X-ray structure (see Table 3.2). Furthermore, the percentage of residues in the most favoured regions of Ramachandran map is lower for set f if compared with those one of the sets including RDCs restraints.

The effect of adding C^{α}-H^{α}, C^{α}-C' and C'-N RDCs

In order to assess the quality of the structures calculated using RDCs restraints, further calculations were carried out. It was reported that the structure of lysozyme does not show any variations from acidic to neutral pH [53], having only one histidine which could influence the structure at these values of pH. Consequently, all the RDCs sets were included to study the improvement of the quality of the structures. In particular, set d includes only N-H^{N} RDCs, while set e include also the C^{α}-H^{α}, C^{α}-C' and C'-N ones. The improvement of the quality of structures is striking, as it is possible to notice from Table 3.2 and Figures 3.2-3.5.

The influence of the full set of NOEs and RDCs, included in set h, was also investigated finding that it gives a significant improvement in the resolution, and in the percentage of residues in the most favoured regions of the Ramachandran plots, see Table 3.3. In addition, the role of the carbon RDCs inside the full set of NOEs (set i) was tested, finding slightly better statistical parameters (see Table 3.1 and Table 3.3) and a better accuracy and precision among the backbone of the structures with respect to the Schwalbe ones [48].

Table 3.1: Structural statistics for the calculated ensembles of 50 low energies structures of set a-c. The Schwalbe and the previous structure values are shown as comparison. Set a: 680 NOEs (99 long, 425 short and 156 intra range contacts) and 3 set of N-H^N RDCs at pH 3.8 and 308 K including 6% polyacrilamide gel, 5% c12E6/hexanol and [D13OPC]: [DHOPC]: [CTAB]=30:10:1; Set b: 680 NOEs (99 long, 425 short and 156 intra range contacts) and 3 set of of N-H^N RDCs at pH 6.5 and 308 K including ester bicelles, 7.5 % DHPC/DMPC ester bicelles doped with CTAB and 10mg/ml PF1 bacteriophage in 500 mm NaCl; Set c: 680 NOEs (99 long, 425 short and 156 intra range contacts). [1] Calculated for Backbone; [2] Calculated including side chains; [3] Calculated using the program PROCHECK [51]. medium I :[D13OPC]: [DHOPC]: [CTAB]=30:10:1 at pH 3.8; medium II: 10mg/ml PF1 bacteriophage in 500 mm NaCl at pH 6.5.

	Schwalbe	**NOE+RDC**^{a}	**NOE+RDC**^{b}	**NOE**^{c}
RMS deviation from ideal covalent geometry				
Bonds (Å)	.00333± .00007	.00279 ± .00023	.00302 ± .00031	.0024 ± .00015
Angles (deg)	.492 ± .0069	.44 ± .02	.48 ± .03	.36 ± .01
Impropers (deg)	.384 ± .0079	.37 ± .03	.41 ± .05	.24 ± .02
RMS deviation from experimental restraints				
NOE(Å)	.0439± .0009	.046 ± .009	.049 ± .007	.0386 ± .003
Dihedrals (deg)	.66 ± .045	.5 ± .3	.4 ± .3	.5 ± .1
Dipolar (Hz)	1.16 ± .042	1.3 ± .1	1.3 ± .09	-
Structural quality				
Accuracy[1]	.5 ± .1	1.5 ± .4	1.7 ± .4	2.2 ± .7
Accuracy[2]	.7 ± .2	2.0 ± .5	2.1 ± .6	2.7 ± .8
Precision[1]	1.5 ± .1	2.4 ± .6	2.7 ± .6	11 ± 1
Precision[2]	1.8 ± .2	3.0 ± .7	3.3 ± .7	12 ± 1
Percentage of residues in the most favoured regions of the Ramachandran plots[3]				
	74.2	78.8	73.5	73.5
Q factor				
medium I	.37 ± .04	.19 ± .03	.35 ± .07	.96 ± .01
medium II	.35 ± .03	.33 ± .05	.18 ± .02	.85 ± .04

Table 3.2: Structural statistics for the calculated ensembles of 50 low energies structures of set d-h. Set d: 680 NOEs (99 long, 425 short and 156 intra range contacts), 3 set of N-H^N RDCs at pH 3.8 and 308 K including 6% polyacrilamide gel, 5% c12E6/hexanol and [D13OPC]: [DHOPC]: [CTAB]=30:10:1; and 3 set of of N-H^N RDCs at pH 6.5 and 308 K including ester bicelles, 7.5 % DHPC/DMPC ester bicelle doped with CTAB and 10mg/ml PF1 bacteriophage 500 mm NaCl; Set e: 680 NOEs (99 long, 425 short and 156 intra range contacts), 2 set of N-H^N RDCs at pH 3.8 and 308 K including 6% polyacrilamide gel, 5% c12E6/hexanol; 1 set of N-H^N, C^α-H^α, C^α-C',C'-N at pH 3.8 and 308 K of [D13OPC]: [DHOPC]: [CTAB]=30:10:1; and 3 set of of N-H^N RDCs at pH 6.5 and 308 K including ester bicelles, 7.5 % DHPC/DMPC ester bicelles doped with CTAB and 10mg/ml PF1 bacteriophage 500 mm NaCl; Set f: 680 NOE (99 long, 425 short and 156 intra range contacts) calculated using a more robust Xplor protocol with respect of set c; Set g: 680 NOEs (99 long, 425 short and 156 intra range contacts) and the RDCs of set d, starting from the 100 structures of set f. [1] Calculated for Backbone; [2] Calculated including side chains; [3] Calculated using the program PROCHECK [51]. medium I: [D13OPC]: [DHOPC]: [CTAB]=30:10:1 at pH 3.8; medium II: 10mg/ml PF1 bacteriophage in 500 mm NaCl at pH 6.5.

	NOE+RDC^d	**NOE+RDC**^{e}	**NOE**^{f}	**NOE+RDC**^{g}
RMS deviation from ideal covalent geometry				
Bonds (Å)	.0032 ± .0003	.0031 ± .00022	.0041 ± .00081	.0039 ± .00051
Angles (deg)	.53 ± .03	.58 ± .02	.54 ± .07	.59 ± .05
Impropers (deg)	.44 ± .03	.52 ± .03	.5 ± .1	.5 ± .1
RMS deviation from experimental restraints				
NOE(Å)	.05 ± .008	.048 ± .005	.07 ± .01	.07 ± .01
Dihedrals (deg)	.5 ± .2	.3 ± .2	2.5 ± .8	.6 ± .2
Dipolar (Hz)	1.67 ± .05	1.74 ± .04	-	1.8 ± .1
Structural quality				
Accuracy[1]	1.4 ± .4	1.1 ± .3	5 ± 2	2.2 ± .5
Accuracy[2]	1.9 ± .5	1.4 ± .4	6 ± 2	2.7 ± .5
Precision[1]	2.3 ± .5	1.9 ± .4	6 ± 3	3.2 ± .6
Precision[2]	2.9 ± .7	2.4 ± .5	7 ± 3	3.9 ± .6
Percentage of residues in the most favoured regions of the Ramachandran plots[3]				
	74.3	78.8	55.8	76.1
Q factor				
medium I	.23 ± .02	.19 ± .01	.85 ± .09	.26 ± .03
medium II	.16 ± .01	.16 ± .01	.85 ± .07	.18 ± .02

Table 3.3: Structural statistics for the calculated ensembles of 50 low energies structures of set h-i. Set h: NOEs of Schwalbe et al. [48] and the sets of RDCs included in set d. Set i: NOEs of Schwalbe et al. [48] and 2 set of N-H^N RDCs at pH 3.8 and 308 K including 6% polyacrilamide gel, 5% c12E6/hexanol; 1 set of N-H^N, C^α-H^α, C^α-C',C'-N at pH 3.8 and 308 K of [D13OPC]: [DHOPC]: [CTAB]=30:10:1. [1] Calculated for Backbone; [2] Calculated including side chains; [3] Calculated using the program PROCHECK [51]. medium I: [D13OPC]: [DHOPC]: [CTAB]=30:10:1 at pH 3.8; medium II: 10mg/ml PF1 bacteriophage in 500 mm NaCl at pH 6.5.

	NOE+RDC^h	**NOE+RDC^i**	**1993**
RMS deviation from ideal covalent geometry			
Bonds (Å)	.0047 ± .00024	.0041 ± .00011	.009 ± .004
Angles (deg)	.69 ± .03	.58 ± .02	2.76 ± .07
Impropers (deg)	.64 ± .06	.51 ± .02	.23 ± .02
RMS deviation from experimental restraints			
NOE(Å)	.058 ± .003	.050 ± .001	.079 ± .012
Dihedrals (deg)	1.5 ± .1	1.55 ± .06	-
Dipolar (Hz)	1.71 ± .05	1.62 ± .04	-
Structural quality			
Accuracy[1]	.7 ± .2	.5 ± .1	1.8 ± .2
Accuracy[2]	1.1 ± .3	.8 ± .2	1.9 ± .3
Precision[1]	1.2 ± .3	1.0 ± .2	2.4 ± .3
Precision[2]	1.8 ± .3	1.5 ± .3	2.7 ± .4
Percentage of residues in the most favoured regions of the Ramachandran plots[3]			
	85.0	82.3	53.1
Q factor			
medium I	.23 ± .02	.22 ± .02	.86 ± .04
medium II	.17 ± .01	.31 ± .03	.78 ± .04

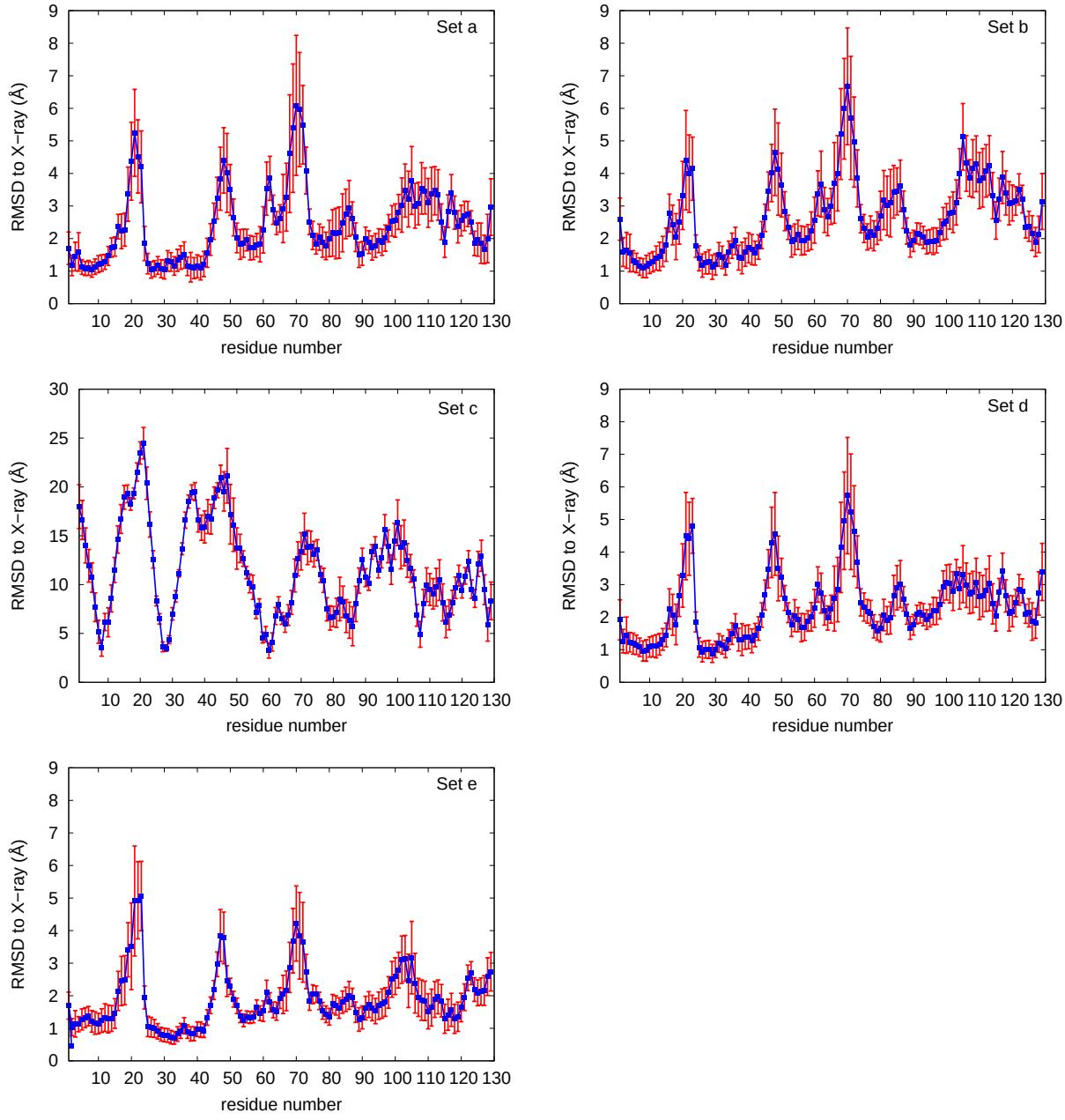


Figure 3.2: RMSD values from the X-ray (pdb code 193L) for the structures of set a-e.

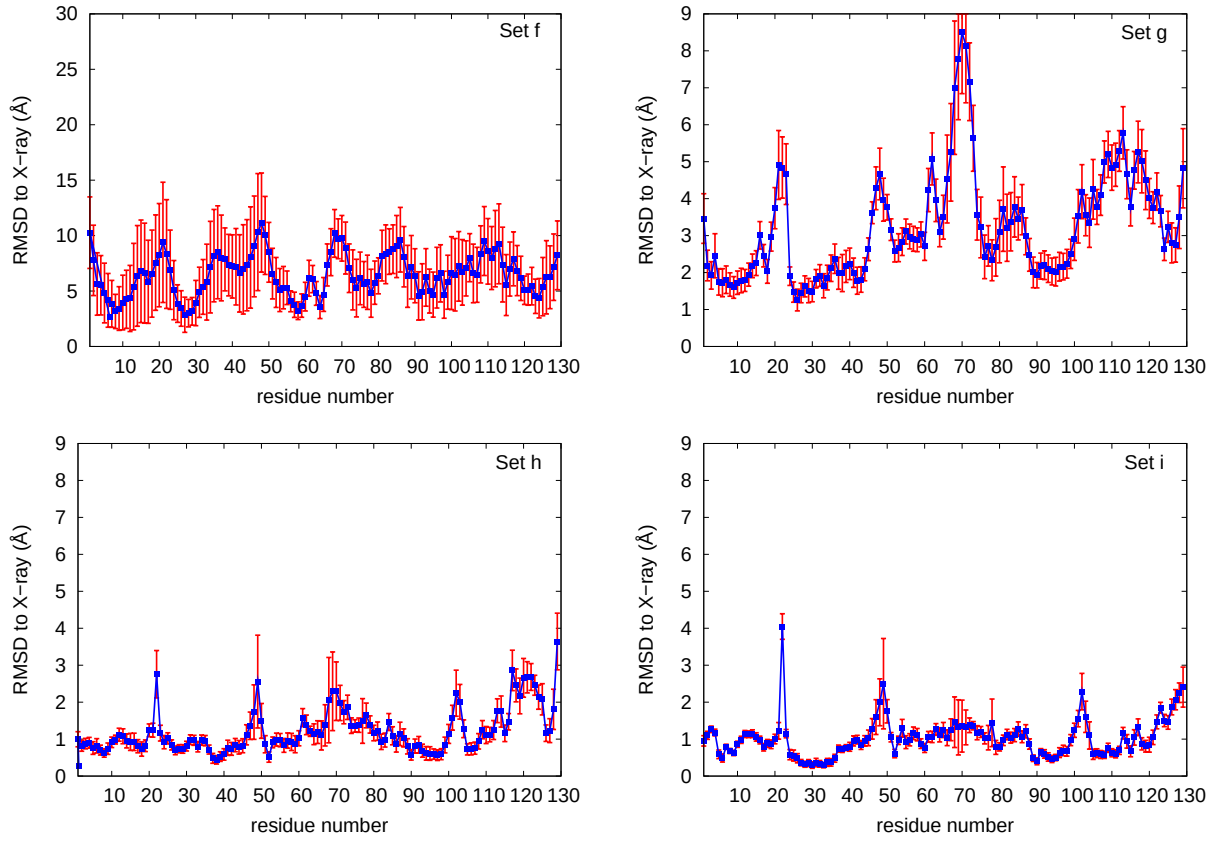


Figure 3.3: RMSD values from the X-ray (pdb code 193L) for the structures of set f-i.

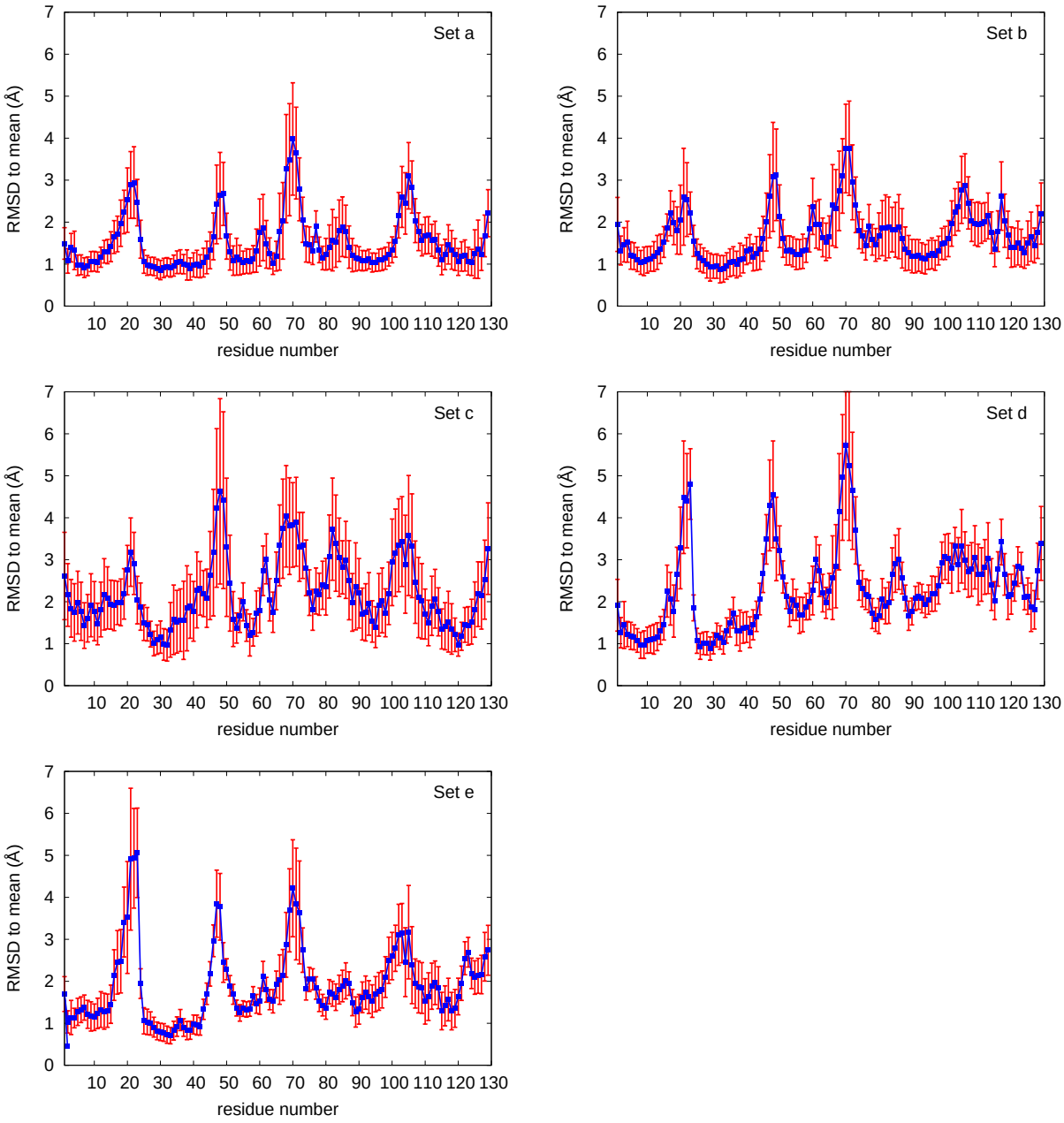


Figure 3.4: RMSD values from the mean for the structures of set a-e.

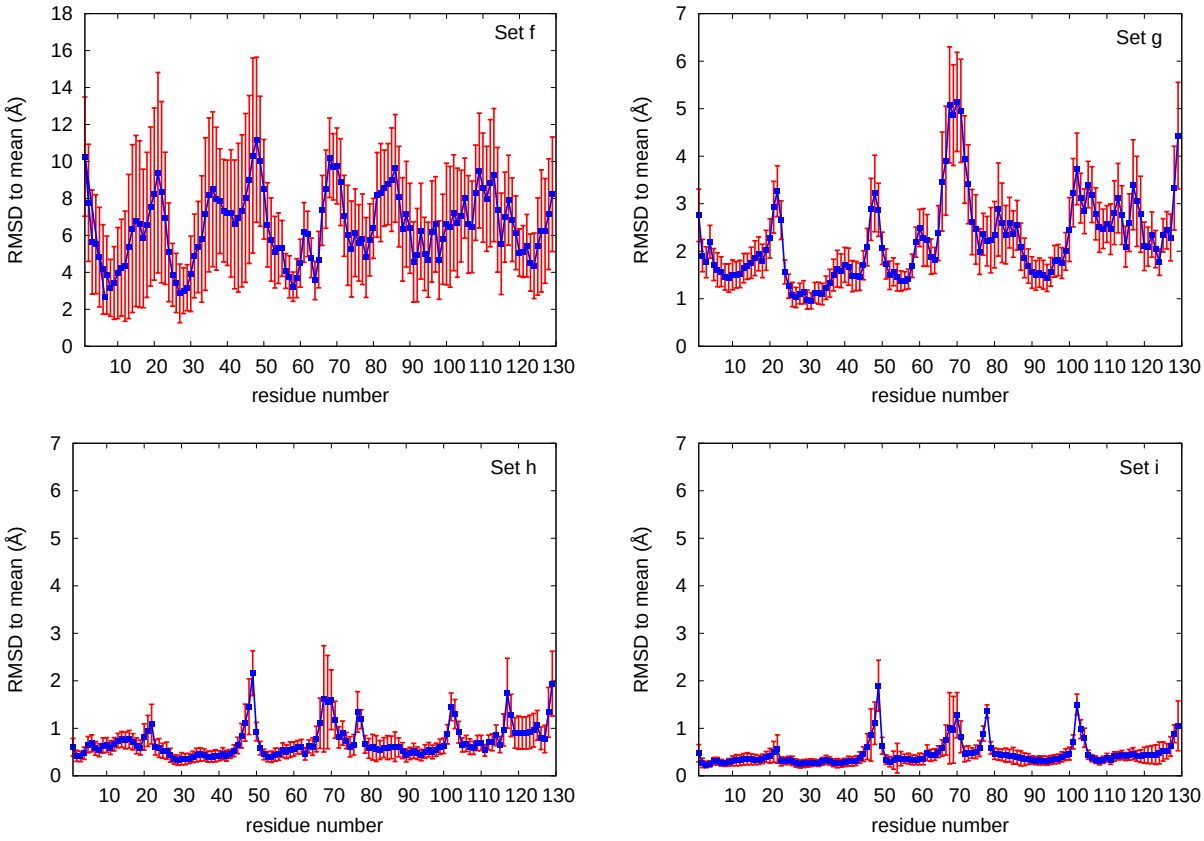


Figure 3.5: RMSD values from the mean for the structures of set f-i.

3.2.3 The set validation

In order to validate all the structures obtained, the Δ NOEs ($|NOE_{exp} - NOE_{calc}|$) distributions were taking into account. From here, the inclusion of carbon RDCs further improves the quality of the structures. This can be seen clearly from histograms of Δ NOEs, reported in Figures 3.6 and 3.7. Set e shows the sharpest one. This is when all the RDCs are included together with the carbon data. Furthermore, the Δ NOEs histograms distributions of structures calculated without RDCs are broader than those one including RDCs restraints. These data fits well with the Δ RDCs ($|RDC_{exp} - RDC_{calc}|$) along the lysozyme backbone, reported in Figures 3.8-3.11. In particular medium II seems to have a significant role in the improvement of the structures, as noticed from Figures 3.10, and 3.11. Furthermore, looking at the Δ RDCs of set a and b of medium I and II respectively, larger values are found in the C-terminal region of structures of set b rather than set a (3.8[a], 3.10[b]). This is due to the absence of RDCs data in medium II in the 100-105 region, which is instead present in medium I. This explains the better statistical quality of set a, which includes the medium I, with respect to set b, reported in Tables 3.1. Both of the representative structures are shown in Figure 3.12 and interestingly different alignments of lysozyme are present by using different alignment media.

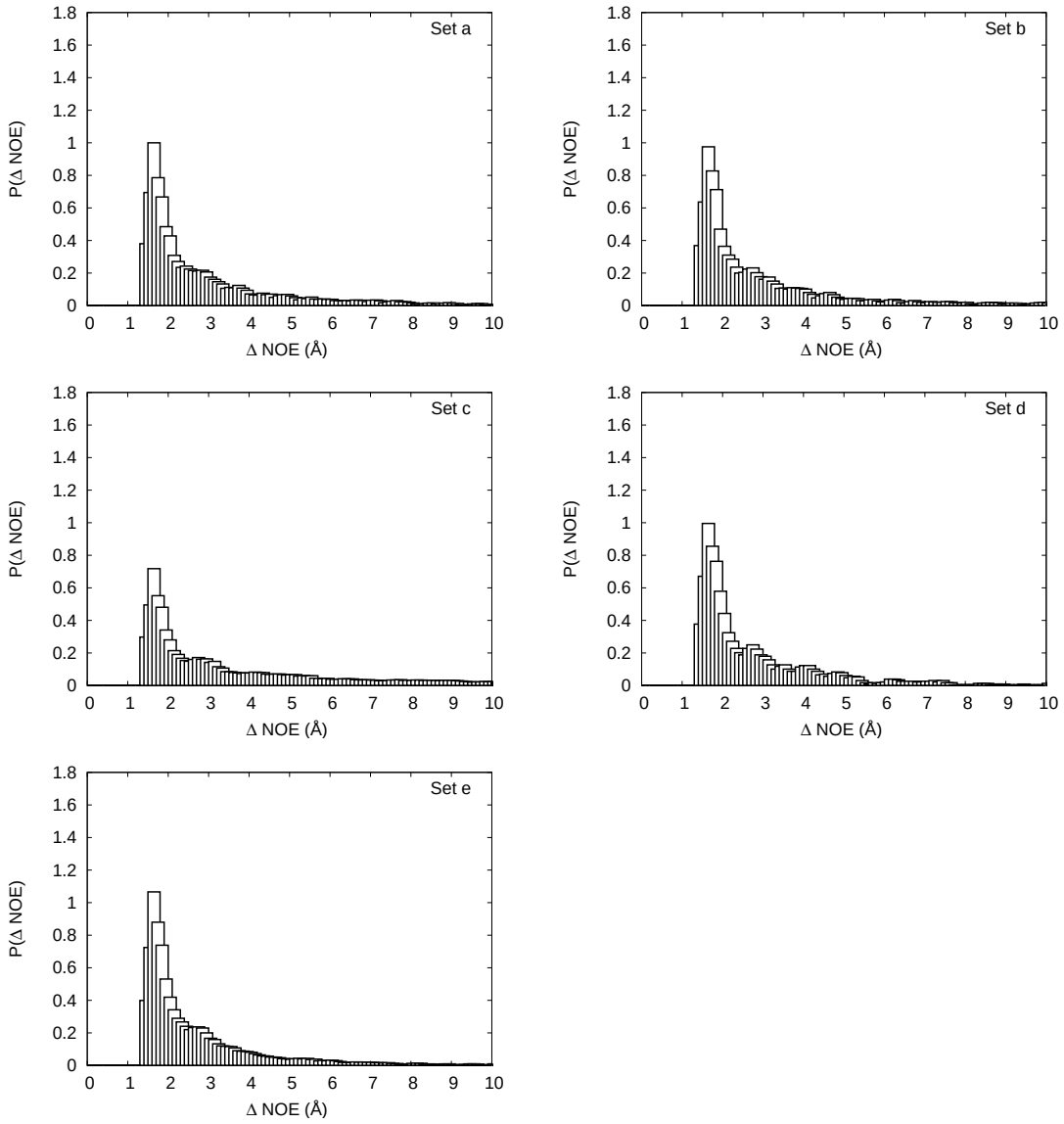


Figure 3.6: Δ NOEs ($|NOE_{exp} - NOE_{calc}|$) distributions for set a-e. It is possible to notice a broad distribution for the structures obtained with only the NOE constraints (set c).

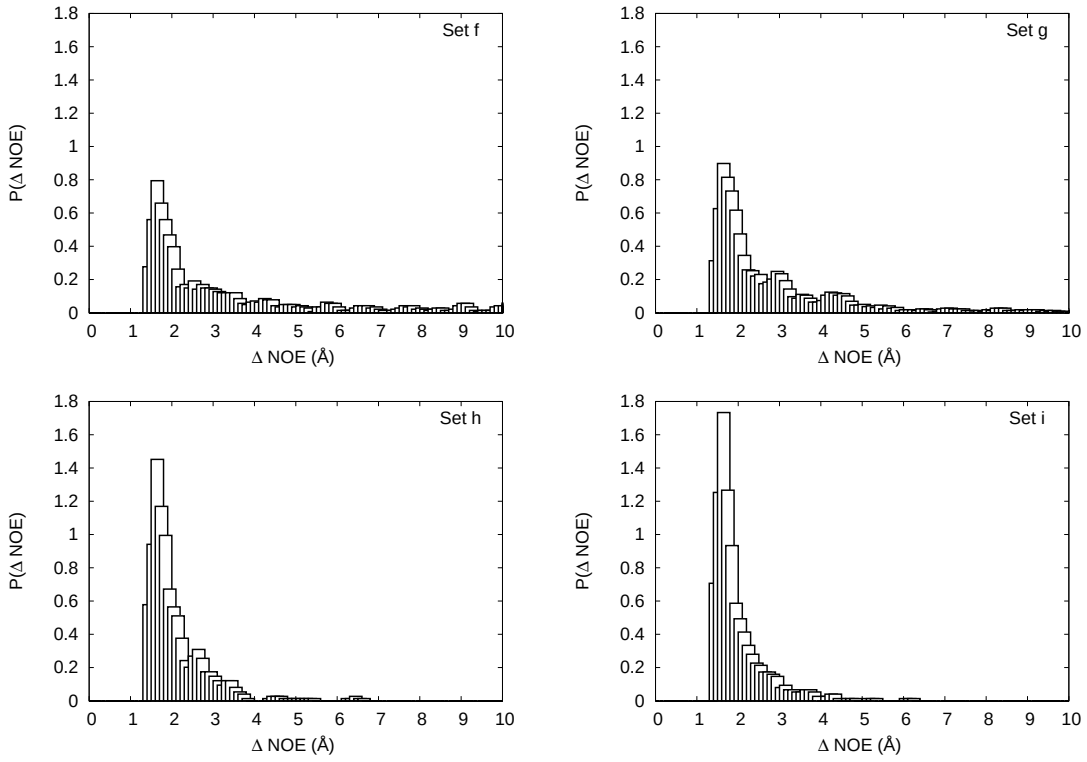


Figure 3.7: Δ NOEs ($|NOE_{exp} - NOE_{calc}|$) distributions for set f-i. It is possible to notice the sharp distribution for set h and i (including RDCs), while a broad distributions for the structures obtained with only the NOE constraints (set f). Set h and i have been calculated with all the NOEs constraints and set g includes the structures of set f perturbed with RDCs, being the distribution of set g sharper if compared with that one of set f.

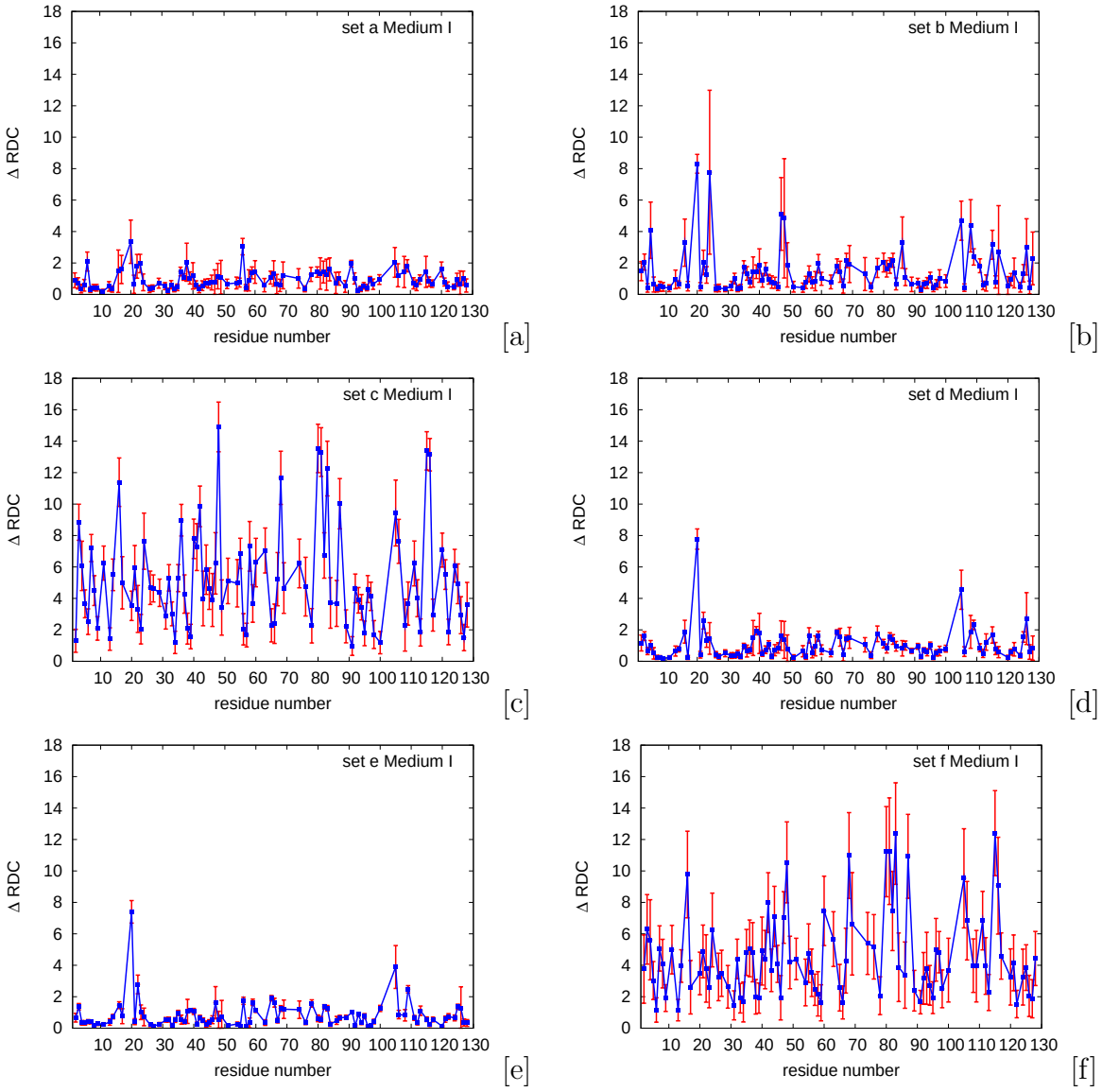


Figure 3.8: Δ RDCs ($|RDC_{exp} - RDC_{calc}|$) along the backbone of lysozyme for set a-f with respect to medium I. It is possible to notice the high variations for the structures obtained with only the NOE constraints (set c, f).

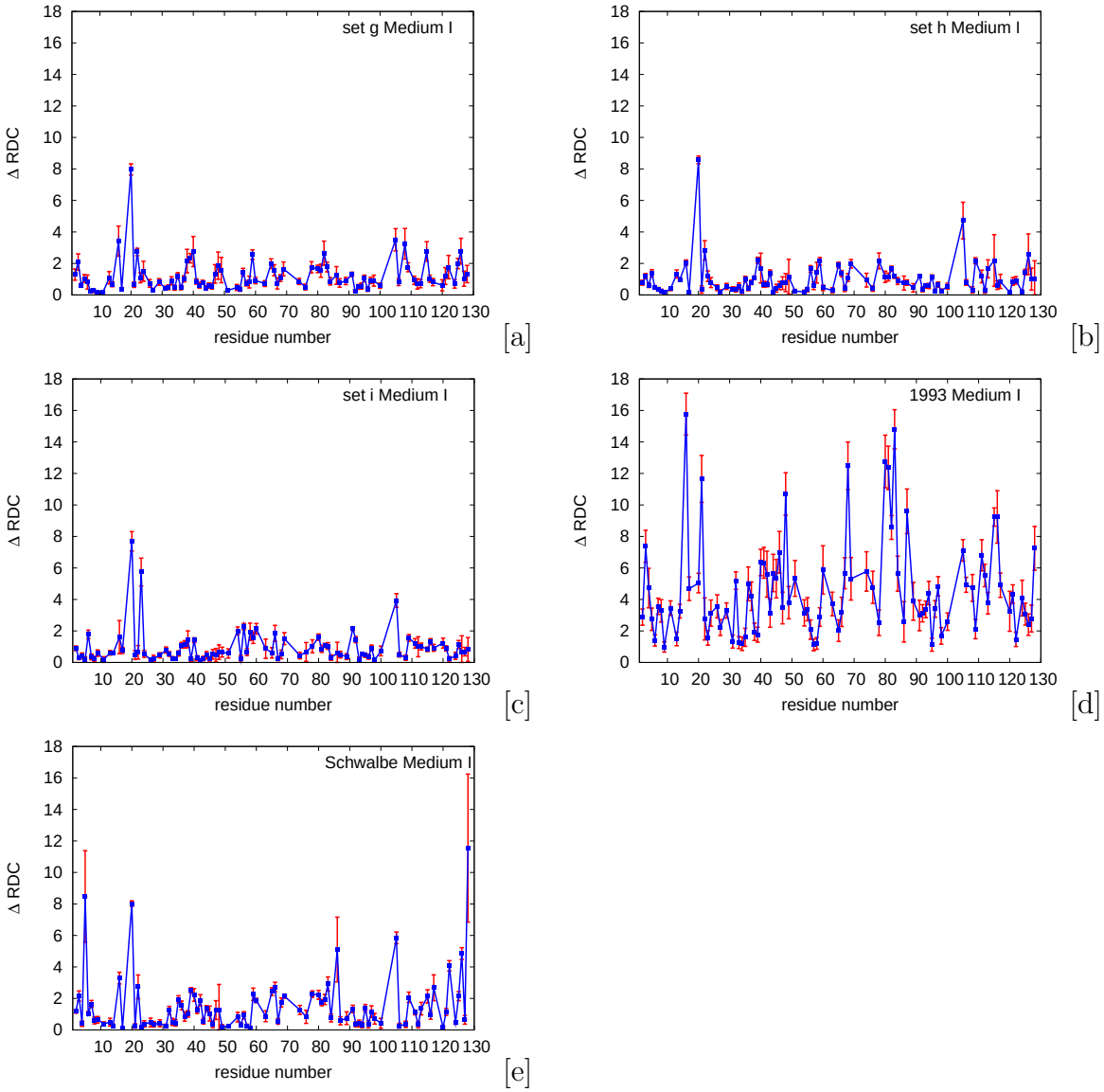


Figure 3.9: Δ RDCs ($|RDC_{exp} - RDC_{calc}|$) along the backbone of lysozyme for set g-i with respect to medium I. 1993 and Schwalbe Δ RDCs (l,m) are shown as comparison. It is possible to notice the high variations for the 1993 structures [l] obtained with only the NOE constraints.

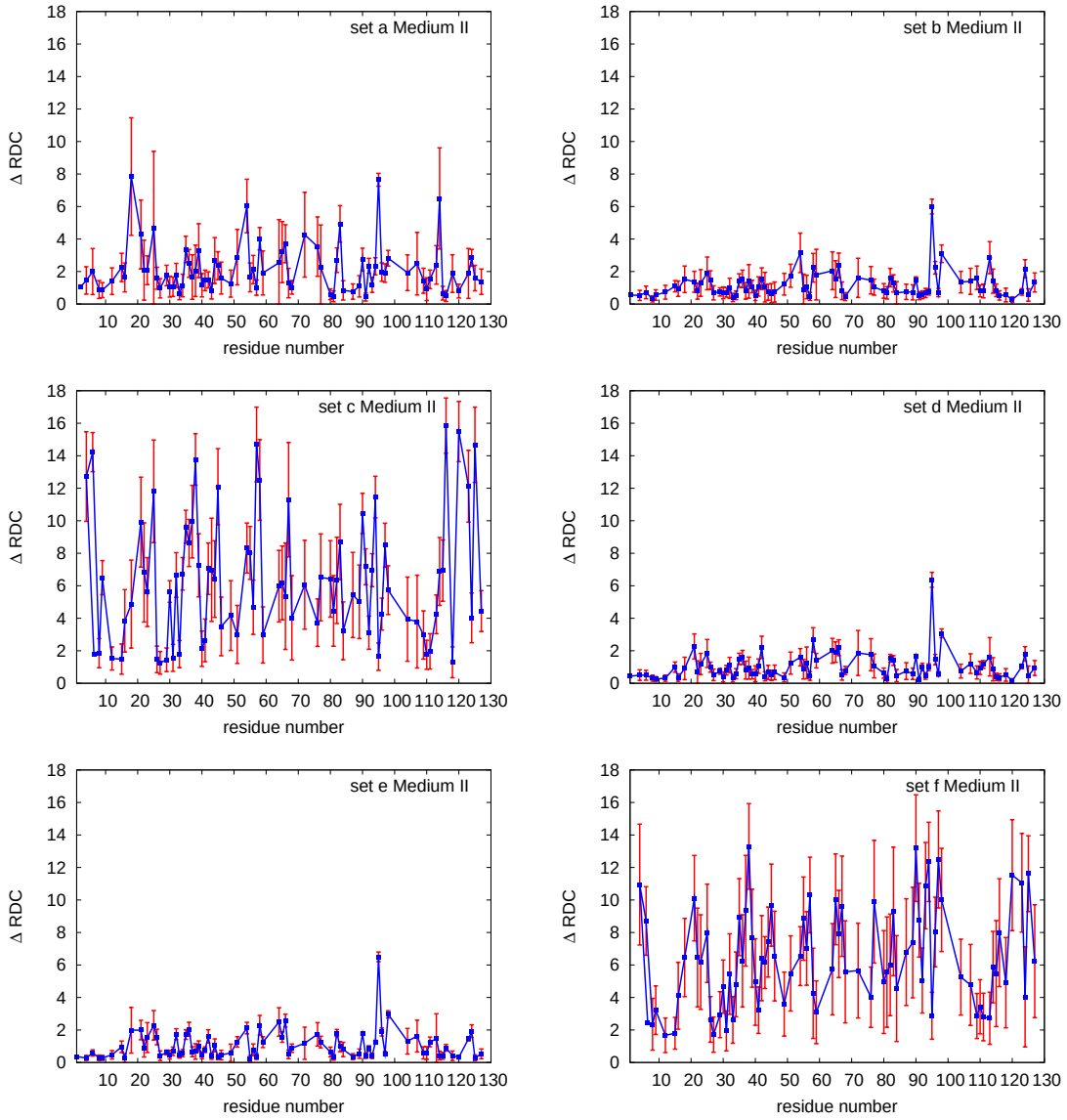


Figure 3.10: Δ RDCs ($|RDC_{exp} - RDC_{calc}|$) along the lysozyme backbone for set a-e with respect to medium II. 1993 and Schwalbe Δ RDCs (i,l) are shown as comparison.

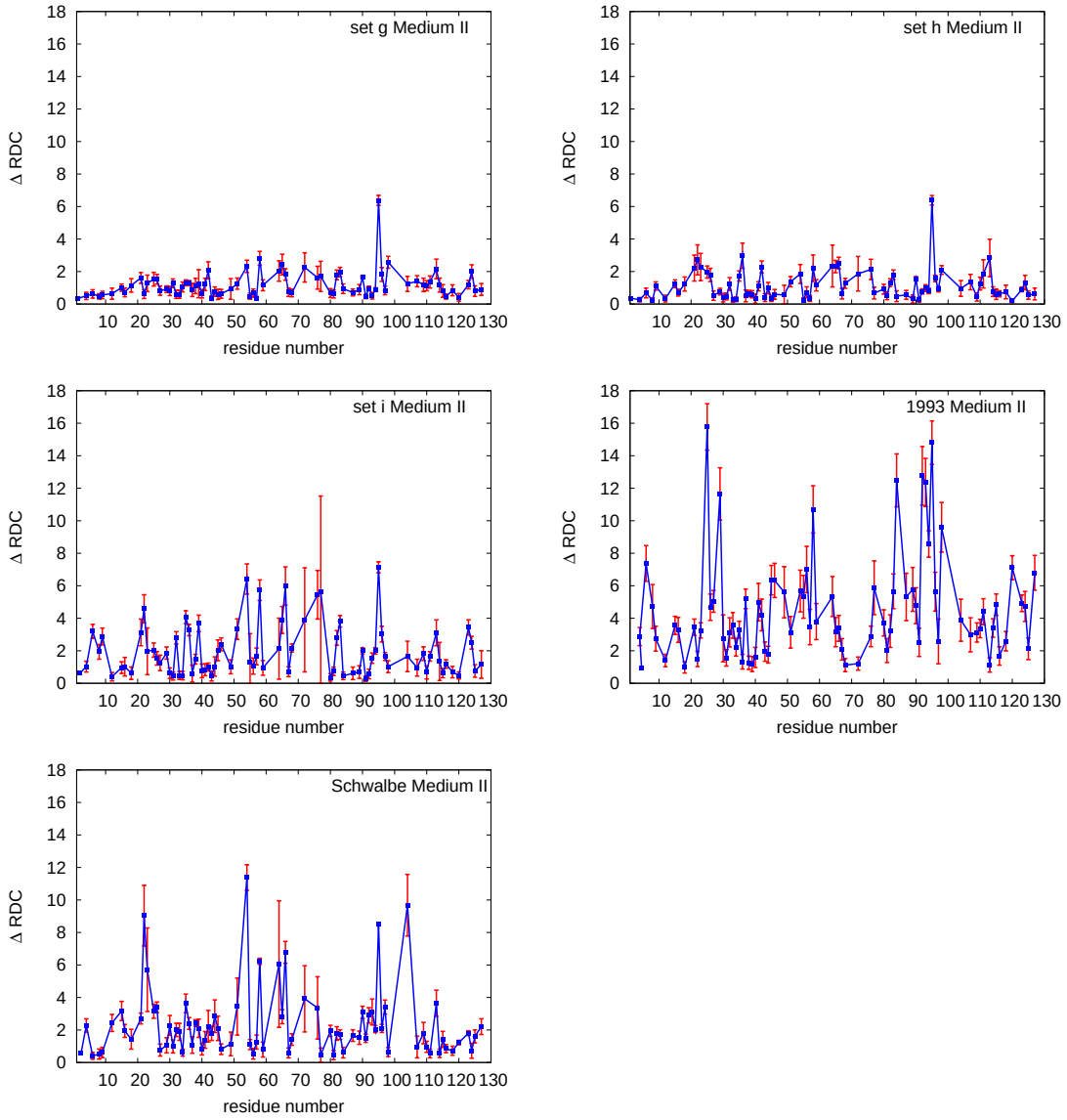


Figure 3.11: Δ RDCs ($|RDC_{exp} - RDC_{calc}|$) along the lysozyme backbone for set f-i with respect to medium II. 1993 and Schwalbe Δ RDCs (l,m) are shown as comparison.

Figure 3.12: Different alignment of the representative lysozyme structures of set a (up) and b (bottom), in medium I and II, respectively.

Secondary structure analysis using chirality index

As a final investigation to assess the quality of the secondary structure, all the structures of the different sets were compared with respect to the X-ray structure (pdb code 193L) using the chirality analysis [52]. In particular the chirality index along the lysozyme backbone for the different sets confirms the statistical results of the RMSD variations along the backbone reported in Figures 3.2-3.5 and gives further indication about the different secondary structure regions of the lysozyme. In particular, the region of 3_{10} helix involving residues 80-84 is only presents in the sets including RDCs restraints (sets a, b, d, e Figures 3.13, 3.14). This region adopts an α helix conformation in set c and it is both 3_{10} and α helix in set f. The other α helix regions are present in all the sets, as shown by the negative oscillations at around -0.05, as well as the three sheet-turn-sheet-turn-sheet regions at around 40-60 are preserved. The other flexible regions are detected at around residue 20 and 107, as it is possible to notice from the high standard deviations in Figures 3.14, of set d-e. Set f shows instead, as said before, two sets of structures, noticed from the oscillations of the standard deviations of the chirality index. In order to test of the quality of the structures with only NOEs constraints, a perturbation of the structures of set f, obtained using only NOEs constraints, was added using the RDCs constraints of set d. The improvement of the structures is noteworthy, as it is shown in Table 3.2, Figures 3.3[g], 3.5[g], 3.7[g], moreover the RDCs improve the 80-84 region adopting a 3_{10}, with respect to set f, as also shown in Figure 3.14 for set gII, in which the chirality index of the most similar structures of set g to the X-ray one (namely the structures with low precision value), is shown.

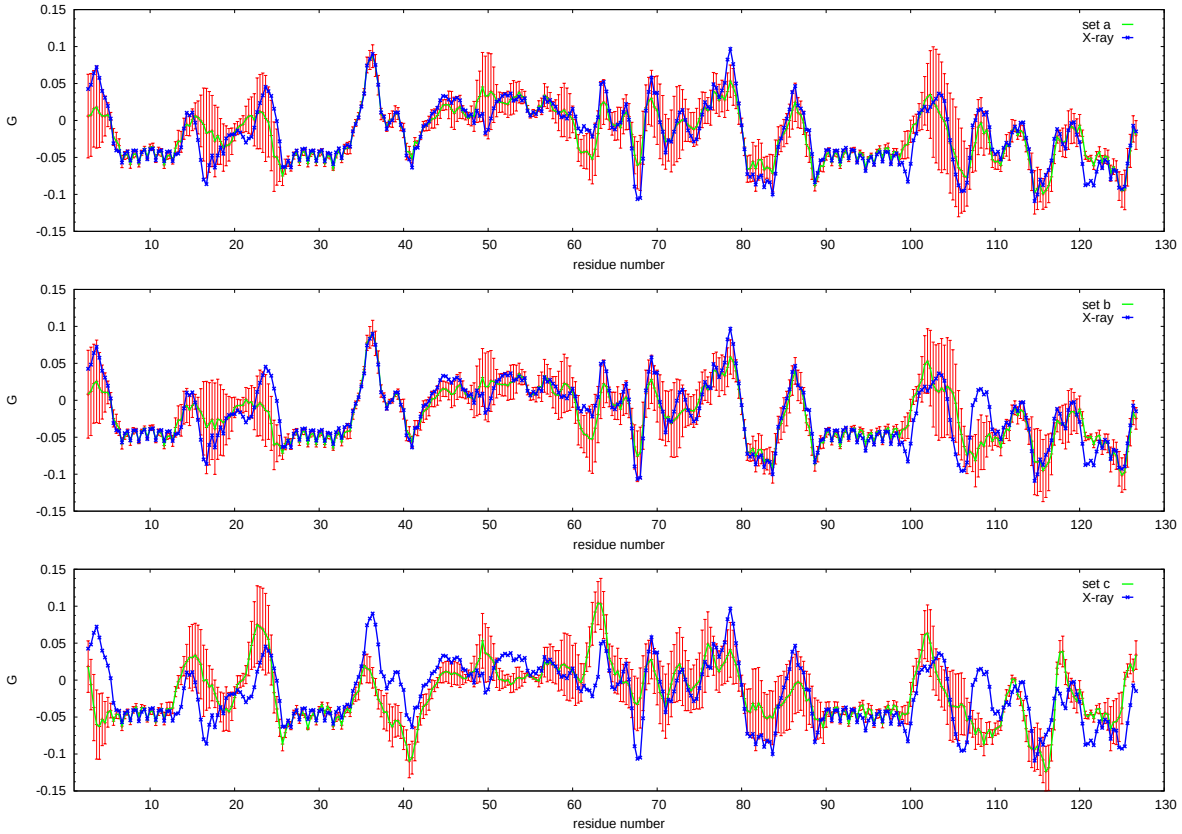


Figure 3.13: From up to bottom: Chirality index averaged among the structures of set a-c; standard deviations of the chirality index among the backbone are also shown with the errorbars to underline the mobility of the residues. It can be noticed the 3_{10} helix in the 80-84 region underlined by negative oscillations around -0.08, only in the structures of set a and b which include RDCs. In set c instead an α helix in the same region is present, identified by the negative oscillations around -0.05.

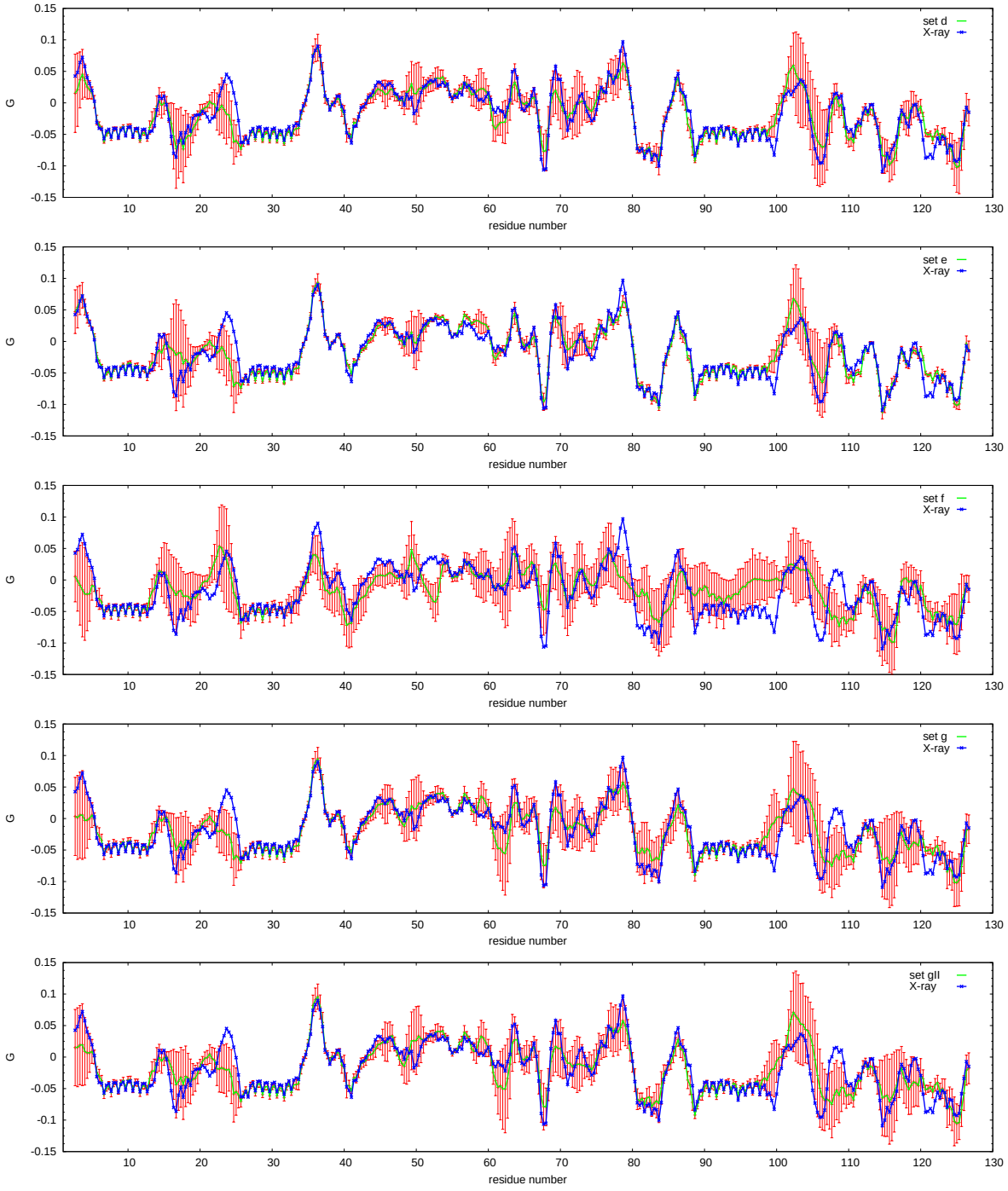


Figure 3.14: From up to bottom: Chirality index averaged among the structures of set d-gII; standard deviations of the chirality index among the backbone are also shown with the errorbars to underline the mobility of the residues. Set f shows two set of different minimized structures, underlined from the high sigma deviations in the chirality index.3_{10} helix in set g for residues 80-84 is more defined with respect to set f, this is also shown from the chirality index of the most similar structures to the X-ray one in set gII.

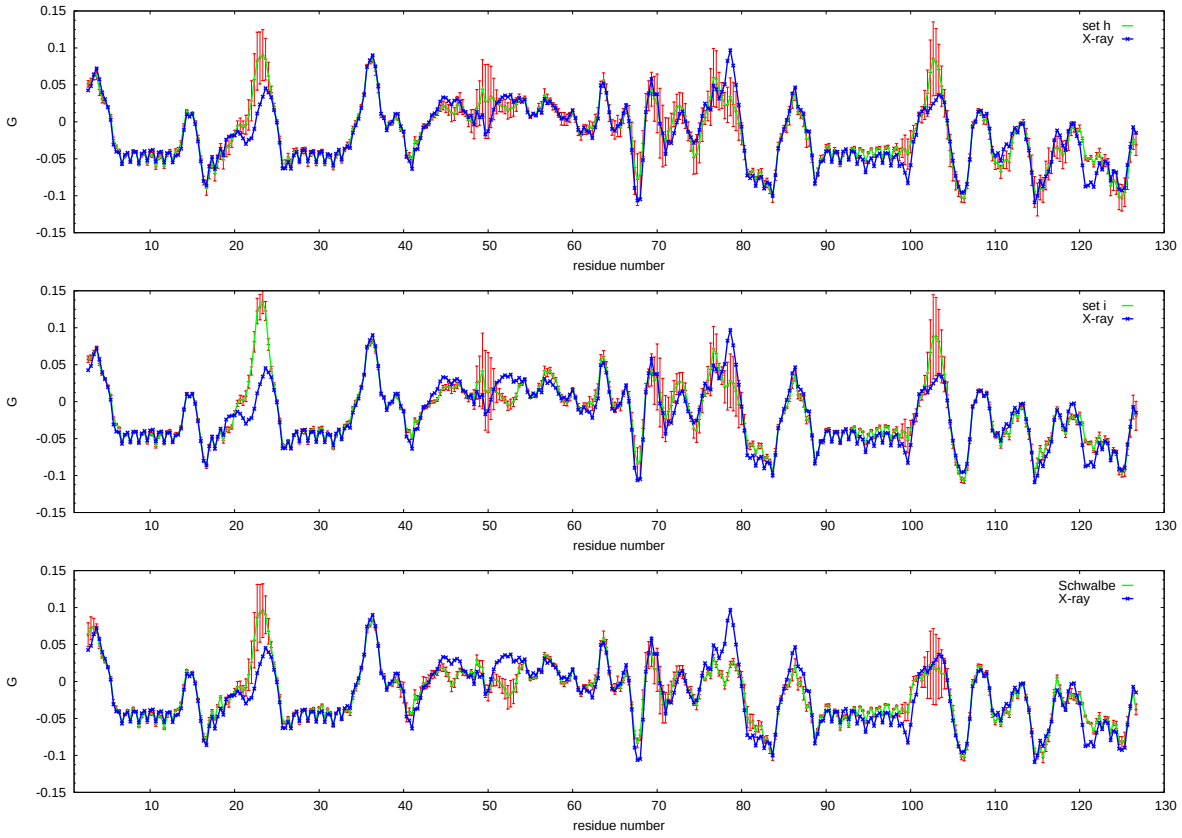


Figure 3.15: From up to bottom: Chirality index averaged among the structures of set h , set i and Schwalbe structures; standard deviations of the chirality index among the backbone are also shown with the errorbars to underline the mobility of the residues.

3.3 Conclusions

The role of Residual Dypolar Couplings (RDCs) in NMR structure refinement is discussed. From the data shown here, RDCs give a deep improvement in the structure resolution. This effect further increases by adding carbon backbone RDCs, which contain the ϕ and ψ dihedral information. This is inferred from both the Δ NOEs ($|NOE_{exp} - NOE_{calc}|$) and the Δ RDCs ($|RDC_{exp} - RDC_{calc}|$) histograms, being very sharp when carbon RDCs are included. Moreover, RDCs help in reaching the correct conformational minimum of energy. In particular, the region of 3_{10} helix involving residues 80-84 is only presents in the sets including RDCs. This is shown from the chirality analysis which reveals the typical values of such structure.

All these results show that RDCs help the annealing in reaching the correct energy minimum of the NMR ensemble and that including a small set of NOEs, families of uncorrect structure could be obtained.

* List of abbreviations:

RDCs : residual dipolar couplings;

NMR : nuclear magnetic resonance;

TROSY : transverse relaxation-optimized spectroscopy;

NOE : nuclear Overhauser effect;

D13OPC: 1,2-O-ditetradecyl-sn-glycero-3-phosphocholine;

DHOPC : 1,2-dihexyl-sn-glycero-3-phosphatidyl-choline;

CTAB : cetyl trimethyl ammonium bromide;

DMPC : dimyristoyl-phosphatidyl choline;

DHPC : dihexanoyl-phosphatidyl choline;

C12E6 : mono n-dodecylether;

SDS : sodium dodecyl sulfate;

Bibliography

[1] K. Pervushin, R. Riek, G. Wider, and K. Wüthrich, *Proc. Natl. Acad. Sci. USA*, **1997**, *94*, 5822–2833.

[2] K. H. Gardner and L. E. Kay, *Annu. Rev. Biophys. Biomol. Struct.*, **1998**, *27*, 357–406.

[3] G. M. Clore and A. M. Gronenborn, *CRC Crit. Rev. Biochem. Mol. Biol.*, **1989**, *24*, 479-564.

[4] G. M. Clore, A. T. Brünger, M. Karplus and A. M. Gronenborn, *J. Mol. Biol.*, **1986**, *191*, 523-551.

[5] M. Nilges, G. M. Clore, and A. M. Gronenborn, *FEBS Lett.*, **1988**, *229*, 317-324.

[6] E. G. Stein, L. M. Rice, and A. T. Brünger, *J. Magn. Reson.*,**1997**, *124*, 154-164.

[7] T. F. Havel and K. Wüthrich, *J. Mol. Biol.*, **1985**, *182*, 381-394

[8] W. Braun, *Q. Rev. Biophys.*, **1987**, *19*, 115-157.

[9] G. M. Clore and A. M. Gronenborn,*Science*, **1991**, *252*, 1390-1399.

[10] A. M. Gronenborn and G. M. Clore, *CRC Crit. Rev. Biochem. Mol. Biol.*,**1995**, *30*, 351-385.

[11] G. M. Clore, M. A. Robien, and A. M. Gronenborn, *J. Mol. Biol.*, **1993**, *231*, 81-102.

[12] D. S. Garrett, J. Kuszewski, T. J. Hancock, P. J. Lodi, G. W. Vuister, A. M. Gronenborn, and G. M. Clore, *J. Magn. Reson.*, **1994**, *104*, 99-103.

[13] K. Bartik, C. M. Dobson, and C. Redfield, *Eur. J. Biochem.*, **1993**, *215*, 255-266.

[14] A. C. Wang and A. Bax, *J. Am. Chem. Soc.*, **1996**, *118*, 2483-2494.

[15] J. Kuszewski, A. M. Gronenborn, and G. M. Clore, *J. Magn. Reson.*, **1996**, *112*, 79-81.

[16] V. Luzzati, *Acta Crystallogr.*, **1952**, *5*, 802-810.

[17] B. A. Borgias, M. Gochin, D. J. Kerwood, and T. L. James, *Progr. NMR Spectrosc.*, **1990**, 22, 83-100.

[18] P. Yip and D. A. Case, *J. Magn. Reson.*, **1991**, *83*, 643-648.

[19] M. Nilges, P. Habbazettl, A. T. Brünger, and T. A. Holak, *J. Mol. Biol.*, **1991**, *219*, 499-510.

[20] M. Karplus, *J. Am. Chem. Soc.*, **1963**, *85*, 2870.

[21] A. Bax, G. W. Vuister, S. Grzesiek, F. Delaglio, A. C. Wang, R. Tschudin, and G. Zhu, *Methods Enzymol.*, **1994**, *239*, 79-106.

[22] N. Tjandra and A. Bax, *Science*, **1997**, *278*, 1111–1114.

[23] C. Gayathri, A.A. Bothnerby, P. C. M. Vanzijl, and C. Maclean, *Chem. Phys. Lett.*, **1982**, *87*, 192–196.

[24] A. Saupe and G. Englert, *Phys. Rev. Lett.*, **1963**, *11*, 462–464.

[25] A. Saupe, *Angew. Chem. Int. Ed.* , **1968**, *7*, 97–111.

[26] C. R. Sanders, B. J. Hare, K. P. Howard, and J. H. Prestegard, *Prog. Nucl. Magn. Reson. Spectrosc.*, **1994**, *26*, 421–444.

[27] C. R. Sanders and J. P. Schwonek, *Biochemistry*, **1992**, *31*, 8898–8905.

[28] M. Ottiger and A. Bax, *J. Biomol. NMR*, **1998**, *12*, 361–372.

[29] M. Zweckstetter and A. Bax, *J. Am. Chem. Soc.*, **2000**, *122*, 3791–3792.

[30] A. Bax, G. Kontaxis, and N. Tjandra, *Methods Enzymol.*, **2001**, *339*, 127–174.

[31] E. de Alba and N. Tjandra, *Prog. Nucl. Magn. Reson. Spectrosc.*, **2002**, *40*, 175-197.

[32] J. H. Prestegard and H. M. Al-Hashimi, *Q. Rev. Biophys.*,**2000**, *33*, 371-424.

[33] G. M. Clore, A. M. Gronenborn, and A. Bax, *J. Magn. Reson.*, **1998**, *133*, 216-221.

[34] G. M. Clore, A. M. Gronenborn, and N. Tjandra, *J. Magn. Reson.*, **1998**, *131*, 10571–10572.

[35] L. Banci, I. Bertini, G. G. Savellini, A. Romagnoli, P. Turano, et al. *Proteins*, **1997**, *29*, 68–76.

[36] J. A. Losonczi, M. Andrec, M. W. F. Fischer, and J. H. Prestegard, **1999**, *J. Magn. Reson.*, *138*, 334–342.

[37] J. J. Warren and P. B. Moore, *J. Biomol. NMR*, **2001**, *20*, 311–332.

[38] J. J. Warren and P. B. Moore, *J. Magn. Reson.*, **2001**, *149*, 271–275.

[39] R. Berardi, F. Spinozzi, and C. Zannoni, **1998**, *J. Phys. Chem*, *109*, 3742–3759.

[40] N. Tjandra, J. G. Omichinski, A. M. Gronenborn, G. M. Clore, and A. Bax, *Nat. Strut. Biol.*, **1997**, *4*, 732–738.

[41] J. Meiler, N. Blomberg, M. Nilges, and C. Griesinger, *J. Biomol. NMR*, **2000**, *16*, 245–252.

[42] C. D. Schwieters, J. J. Kuszewski, N. Tjandra, and G. M. Clore, *J. Magn. Reson.*, **2003**, *160*, 65–73.

[43] T. Herman, P. Güntert, and K. Wüthrich, *J. Mol. Biol.*, **2002**, *319*, 209–227.

[44] W. Y. Choy, M. Tollinger, G. A. Mueller, and L. E. Kay, *J. Biomol. NMR*, **2001**, *21*, 131–140.

[45] G. M. Clore and D. S. Garrett, *J. Am. Chem. Soc.*, **1999**, *121*, 9008–9012.

[46] G. Cornilescu, J. L. Marquardt, M. Ottiger, and A. Bax, *J. Am. Chem. Soc.*, **1998**, *120*, 10571–10572.

[47] F. Delaglio, G. Kontaxis, and A. Bax, *J. Am. Chem. Soc.*, **2000**, *122*, 2142–2143.

[48] H. Schwalbe, S. B. Grimshaw, A. Spencer, M. Buck, J. Boyd, C. M. Dobson, C. Redfield, and L. J. Smith, *Proteins Sci.*, **2001**, *10*, 677–688.

[49] C. D. Schwieters, J. J. Kuszewski, N. Tjandra, and G.M. Clore, *J. Magn. Res.*, **2003**, *160*, 66–74.

[50] J. P. Linge and M. Nilges, *Journal Biomol. NMR*, **1999**, *13*, 51–59.

[51] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, *J. Appl. Crystallogr.*, **1993**, *26*, 283–291.

[52] A. Pietropaolo, L. Muccioli, R. Berardi, and C. Zannoni, *Proteins*, **2008**, *70*, 667-677.

[53] P. Polverino De Laureto, E. Frare, R. Gottardo, H. Van Dael, and A. Fontana, *Proteins Sci.*, **2002**, *11*, 2932–2946.

Chapter 4

The fold of prion protein

4.1 Prion and protein misfolding

As seen before in the Introduction, the protein folding is a very important mechanism for biological systems. Errors in this process give rise to misfolded structures, which can be lethal. The understanding of how a conformational change occurs, producing a pathogenic protein, is one of the main purposes of proteomics. A correct folding involves a particular energy minimum, corresponding to a given conformation of the protein. A misfolding, which means an incorrect folding, is very dangerous for the cell. However, there is a control and a defense system, represented by *Molecular Chaperons*, namely some special proteins which assist the correct folding of native structures. Besides, there is also a further control system, ubiquitin depending, that degrades the misfolded proteins in the *Proteosome*, so called because these proteins could not carry out the function dictated by the genetic code and thus degraded. Sometimes, for uncleared reasons, this process does not happen. Therefore, misfolded proteins tend to aggregate and the insoluble aggregates may form the *amyloid plaques*. Many proteins, unfortunately, tend to misfold and aggregate, giving rise to diseases having a great social impact (Table 4.1).

The prion disease, the only one to be infectious, includes the *Creutzfeldt-Jacob disease* in humans, the *Mad cow disease* in cattles and *Scrapie* in sheeps, but interestingly it seems to be spared to non mammals [1–3]. Such disorder involves the misfolding of PrP^{C} (Prion protein cellular) in the infectious scrapie isoform, PrP^{Sc} [4], causing the disease. Both isoforms are codified by the same chromosomial gene. Consequently, it was proposed that the conversion from PrP^{C} to PrP^{Sc} occurs through a conformational

Table 4.1: Misfolding related diseases.

disease	**involved protein**
cystic fibrosis	CFTR
Marfan syndrome	fibrillin
Amyotrophic lateral sclerosis	Superoxide Dismutase
Scurvy	Collagen
Maple sirup urine disease	alpha ketoacid dehydrogenase (BCKAD)
Cancer	p53
Osteogenesis imperfecta	α procollagen type I
Scrapie/Creutzfeldt-Jacob/familiar insomnia	prion protein
Alzheimer's disease	β Amyloid
Familiar Amyloidosis	Transthyretin/Lysozime
Cataract	α Crystalline
Type II diabetes	Amylin
Parkinson Disease	α synuclein
Familiar Hypercholesterolemia	LDL receptor
α_1-Antitrypsin deficiency	α_1-Antitrypsin
Tay-Sachs disease	Hexosaminidase
Retinitis Pigmentosa	Rodopsine
Donhoue disease	insulin receptor

mechanism, since both isoforms have the same amino acidic sequence [5]. Prion disease has been associated with the conversion of the α helix rich prion protein (PrP^C) into a β sheet rich insoluble conformer (PrP^{Sc}), thought to be infectious. In this regard, a possible interaction with another protein, called *Protein X*, so far not yet defined, was hypothesized. Such a protein is thought to convert the PrP^C to the Scrapie isoform, probably explaining why the disorder occurs since, up to now, there is no evidence of the presence of a virus [4].

Despite the uncertainties about the transmission of the disease, there is a strong evidence that the prion expression and the Scrapie isoform conversion are essential for the neurodegeneration. It was observed, in fact, that mice whose normal prion expression gets knocked out cannot be infected [6]. Furthermore, neurons PrP^C gene knockout, are resistant to PrP^{Sc} neurotoxicity, but are more exposed to oxidative stress [7]. The subcellular location of PrP^{Sc} has not been determined yet. However, it seems that PrP^{Sc} is present in the Golgi apparatus from which it goes anchoring on the cell membrane

Table 4.2: Main differences between PrP^C and PrP^{Sc}.

$\mathbf{PrP}^C$	$\mathbf{PrP}^{Sc}$
poorly aggregable	highly self-aggregable and polymerizable
completely digested from proteases	partial digestion with proteases and phospholipases
soluble and disrupted by detergents	insoluble in detergents
location only in cell membrane	cellular location intracytoplasmatic
cellular turnover very quick	slow and stable synthesis
α helices rich conformation	β sheets rich conformation
easy disruption	strong resistance to physico-chemical agents

surface, like PrP^C, by a signal C-terminal sequence. The two prionic isoforms show some intriguing features (Table 4.2). It was observed, in fact, that while PrP^C is entirely degraded by proteases, PrP^{Sc} is degraded into a stable fragment [8]. Therefore, the protein can be incorporated into an aggregate, growing up continuously forming plaques like the amyloids in the infected brain (Figure 4.1).

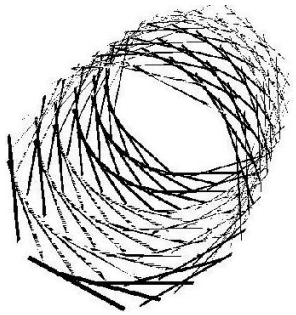


Figure 4.1: Structure of the amyloid plaques aggregate.

The word *Amyloid* is used to show some extra and intra cellular fibrils related to a disorder. These deposits are stored in different kinds of tissues and are distinguished by their morphological properties [9]. The normally soluble proteins conversion to amyloid plaques is a common feature for many disorders, amyloidosis related, like Alzheimer's disease, the type II diabetes and Bovine Spongiform Encephalopathy (BSE) [8]. Recent works improved our knowledge of how proteins can afford amyloid plaques. It was shown experimentally that both globular and unstructured proteins, start the fibrillation adopting a partially structured conformation. Although the amyloidogenic proteins are very different each other with respect to the amino acid sequence and native conformation,

fibrils are structurally similar [8]. Electronic microscopic techniques show that fibrils from a tissue or formed by the mature protein have a diameter of about 5-13 nm and furthermore they are stiff and not very branched (Figure 4.2) [8]. X-ray diffraction studies suggest a cross β structure, where perpendicular β strands and hydrogen bonds, parallel to the main axis of a given fibril, are present [10]. Recently, a possible fibrillation mech-

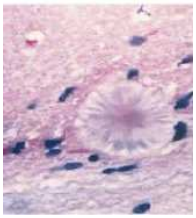


Figure 4.2: Amyloid plaques structure obtained from electronic microscopy.

anism implying an intermediate with an ordered secondary structure and a non native tertiary one has been suggested [11]. Probably, a partially ordered conformation gives rise to highly specific intermolecular interactions, like hydrogen bond and hydrophobic interactions necessary to yield fibrillation and oligomerization [6]; moreover, the intermediate needs to have a very high propensity to assume a β sheet rich conformation, favouring the self assembly [8]. The solutions of these fibrils studied by CD spectroscopy show a minimum around 215 nm, typical of a β strand [12]. Furthermore, from X-ray diffraction studies it was observed that myoglobine fibrils contain β strands, which are oriented perpendicular to the main axis of the fibril [12]. These results show how a protein like myoglobine can adopt a very different structure, highly organized, typical of amyloid plaques. This underlines that the property of forming fibrils is common in many proteins and is present in partial denaturation conditions. Furthermore, it was also observed that, in order to produce the generation of fibril agglomerates, the presence of glutamine [13] and of interchain hydrogen bonds through the glutamines is necessary. It is worth noting that both allow to interact with bivalent metals, such as copper(II). Many sistems like mammal prion [14], β amyloid [15], implied in the Alzheimer formation, α synuclein [16], involved in the Parkinson's disease, the immunoglobuline light chain [17] and the β2 microglobulin [18], related to the *Hemodialysis-associated amyloidosis*, possess these necessary conditions for the formation of amyloid plaques.

4.2 The role of prion on the cellular metabolism

Despite a very high number of papers, actually only few information are known about the prion protein physiological function. Although the physiological role of prion protein is still an open question, it has been suggested that PrP^C is involved in oxidative stress protection [19], in apoptosis [20], in cellular signaling [21], in membrane excitability and synaptic transmission [22], in transport and copper metabolism [23], but it is still unclear how all these functions can be carried out by the same protein. The PrP^C was found on the cellular surface and it is believed that the endocytosis may influence the physiological function [24]. The prion protein, as well as the membrane glycoproteins, resides in some sites of plasmatic membrane, rich in cholesterol and sphingolipids [25]. The mechanism of prion endocytosis is not, so far, clear. The absence of intracellular amino acid sequences shows that the protein is hampered in directly interacting with other adapter proteins for the endocytosis mediated by chlatrins [24]. Pauly and Harris were the first ones to show that micromolar copper(II) concentrations stimulate the chicken PrP endocytosis [26]. The binding of copper(II) ion to the PrP^C induces a change in conformation, but how this could be a signal for the internalization is unclear.

In 1998 D.R. Brown and coworkers [27] studied the antioxidant activity of recombinant chicken and mouse prion protein observing that, if copper(II) was present during the folding process, the two analogues strongly inhibited the superoxide ion activity. Furthermore, the kinetic constant value referring to the reaction of superoxide inhibition by prion protein is equal to 4×10^8 M^{-1} s^{-1}, lower than that of cytosolic Cu-Zn SOD (Superoxide Dismutase) of about one unit. The zinc ion, which in the Cu-Zn SOD has a structural role, does not seem to be involved in the prion SOD-like function because the folding in presence of both copper and zinc does not show any increase of SOD activity, although a recent paper indicates that zinc ion regulates the copper coordination in the octa-repeat region [28]. The prion SOD-like function is supported by the fact that in the synapsis, where the concentration of PrP is very high, the presence of Cu-Zn SOD was not observed. Therefore, prion may be the main or the only SOD protein. The conformational change caused by the binding of copper(II) to the PrP^C could be a switch for the expression of SOD activity. Consequently, an intriguing hypothesis could be that the increased levels of PrP^{Sc} may modifies the ROS (reactive oxygen species) metabolism, causing the disease (Figure 4.3) [29].

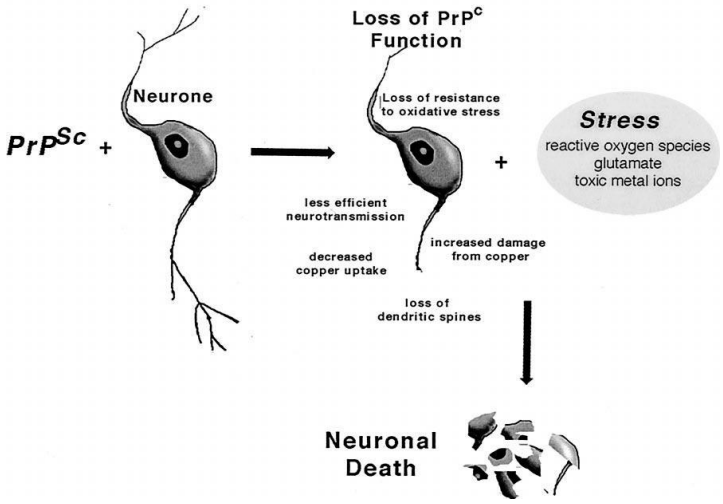


Figure 4.3: Possible mechanism of neural necrosis.

4.3 The structure of mammal prion protein

The prion protein is codified by a gene called *PRNP*, localized on the short arm of chromosome 20, featured by two exons separated by a single intron [30]. Such a protein (PrP^C) is a cell surface glycoprotein expressed in brain, spinal cord and several peripheral tissues [31, 32]. A signal peptide bound to the N-terminal region drives the protein entrance inside the endoplasmic reticulum. The anchoring to the cell membrane is regulated by a second signal peptide located in the C-terminal region. Both the two signal sequences are broken before the arrival of the protein to the cell membrane. Furthermore the glycosylation, which takes place in the endoplasmic reticulum, begins after the cleavage of the C-terminal signal peptide [23]. The glycosylation site is generally a Serine, more exactly Serine 231 in human prion and Serine 249 in chicken prion [33]. Moreover, two glycosylation sites, namely Asparagine 181 and 197, are present in the mature protein. A site of cleavage is also present in a hydrophobic region of the N-terminal tail, where the protein is hydrolyzed during the cellular methabolism. Two cysteine residues form a disulphide bridge, which are fundamental for the protein folding [34]. They bind together, in fact, two regions of the protein which are folded with an elicoidal structure. Similarly, following the behavior of many membrane proteins, PrP^C is synthesized in the rough endoplasmic reticulum and it reaches the cell surface, passing through the Golgi apparatus.

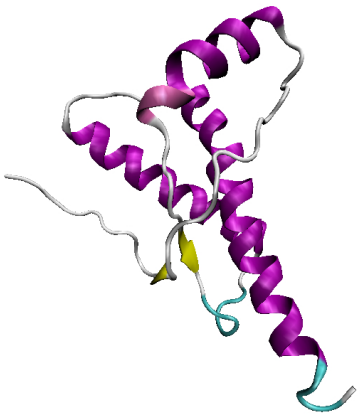


Figure 4.4: Tridimesional structure of the human prion protein (119-230). α helices are shown in violet, β strands in yellow, unordered regions in white and turns in green.

The prion conformation was studied by a recombinant analogue [35,36]. NMR studies show a structure prevalently in the α helix form [37]. In particular, the globular domain contains three α helices, being helices 2 and 3 bound by a single disulphide bridge and two regions with antiparallel β sheets near the helix 1 (Figure 4.4). Fast Fourier Transform InfraRed (FT-IR) and Circular Dichroism (CD) spectroscopy revealed that α helix amount in the PrP^C is about 42%, whereas the β sheets presence is about 3% [38], being these values very different in the Scrapie form, where α helix amount is lower than 30%, whereas that one of β sheets reaches 43% [38]. The tertiary structure of the human prion protein (Figure 4.4) shows also a long and flexible N-terminal tail. Such N-terminal half of the apoprotein is unstructured with a very high mobility of the backbone. Residues 60-91 are called *Octarepeat*, because of eight repetitions of the amino acid motif *PHGGGWGQ*. This region is one of the mostly conserved part in mammalian prion and it binds copper(II) in the entire protein and in a similar way in synthetic octarepeat fragments, involving four copper(II) ions cooperatively, each of them with an equal coordination geometry [39], although the PrP106-126 sequence shows a higher metal-binding affinity than the octarepeat fragments [40].

4.4 The Avian prion

The analysis of 27 mammalian and 9 avian PrPs revealed high conservation of the flexible regions of prion proteins, encompassing the N-terminal part characterized by the presence of peptide repeats [41]. The chicken prion protein (ChPrP) was the first non mammalian prion whose gene was isolated and it is expressed in the spinal cord and brain cells [42]. Such a prion has a high sequence homology with human prion in the central part (106-207) but it is very different in the N-terminal region. In addition, an increment in the number of amino acids is present in the C-terminal region (Figure 4.5). Although avian species also express prion protein, it has not been reported

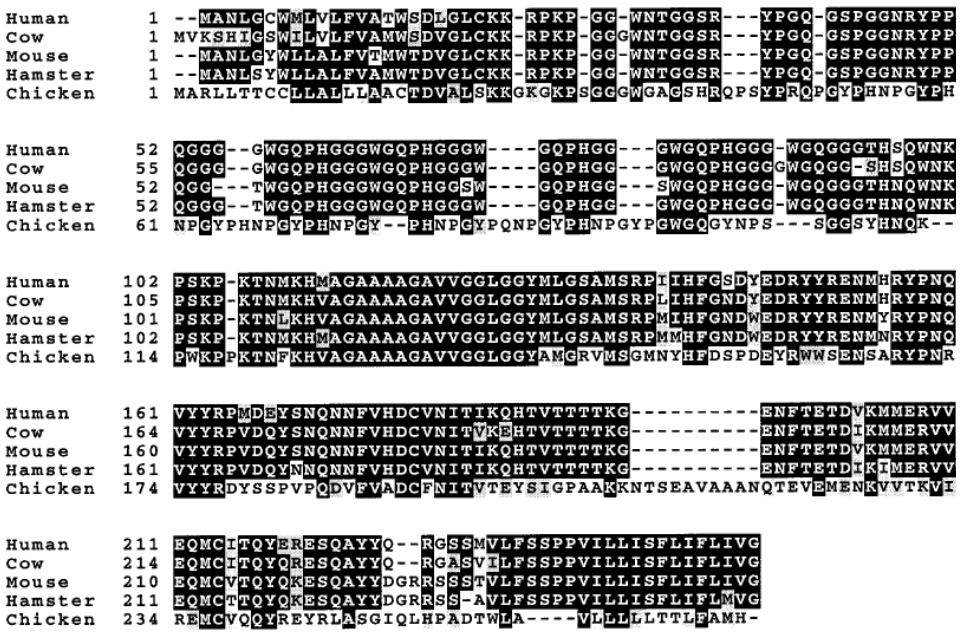


Figure 4.5: Amino Acid sequences of human, cow, mouse, hamster and chicken prion.

any evidence of neurodegenerative disorders among them [1–3]. Moreover, despite the low sequence identity to mammalian PrP (around 33% between chicken and human prion protein), the molecular architecture of the globular core is preserved among the two species (Figure 4.7), with a long flexibly disordered tail attached to the N-terminal end of the globular domain [43], suggesting the importance of these properties on prion function.

The mammalian prion essential features are conserved (Figure 4.6) [44]. In particular both proteins possess: i) multiple N-glycosylated sites; ii) an amino-terminal signal

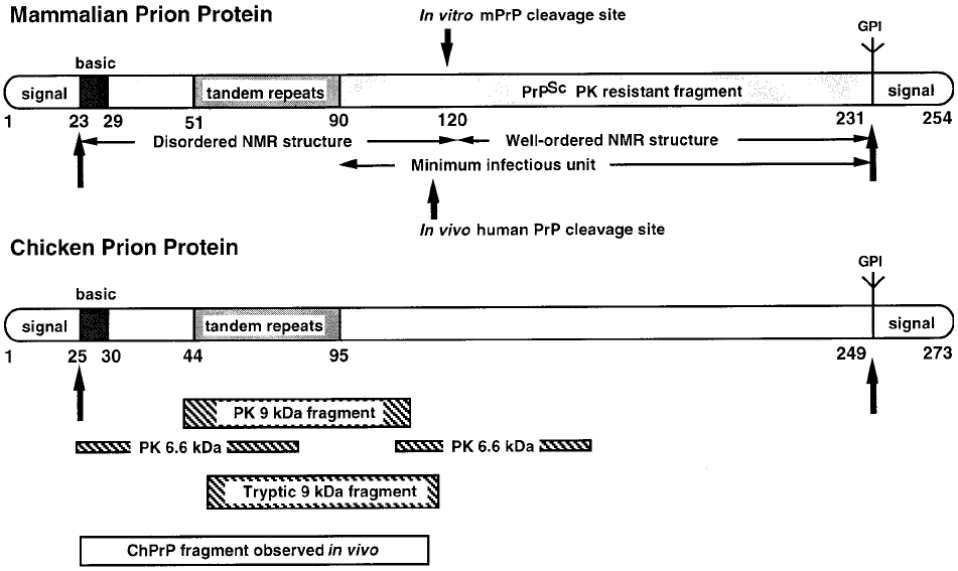


Figure 4.6: Main features of mammalian and chicken prion.

sequence that is removed in the mature protein; iii) a carboxy-terminal signal that is eliminated when the mature protein is linked to *Glycosyl Phosphatidyl inositol* (GPI), and iv) an N-terminal domain featured by tandem amino acid repeats, octameric in mammalian prion, as seen in section 4.3, and hexameric in the avian one, respectively $(PHGGGWGQ)_4$ and $(PHNPGY)_7$.

Such N-terminal region, made up of polypeptide repeats, has been analyzed by proteolysis [33]. Differently than mammalian homologues, the digestion of chicken prion protein with trypsine or proteinase K, produces peptide fragments stable to further proteolysis [33], suggesting that they adopt a different structure than those of mammalian prion tandem repeats. One of these fragments comprises the 49-129 sequence, consisting of a large part of the N-terminal domain [33]. This resistance to proteolysis may suggest a compact domain of the proline/glycine rich N-terminus, although the hexarepeat amino acid sequence seems to show no tendency towards a particular structured conformation [33]. It is worth underlying also that the mammalian octarepeat peptides contain 50% of glycine and 12% of proline residues while the chicken hexarepeats encompass 16% of glycine and 33% of proline. Therefore, the higher flexibility conferred by glycine residues might explain why the mammalian prion protein N-terminal tandem amino acid repeats do not form a stable protease-resistant domain [33]. The differences of sequence between chicken and mammalian prion could help to understand some similarities, as

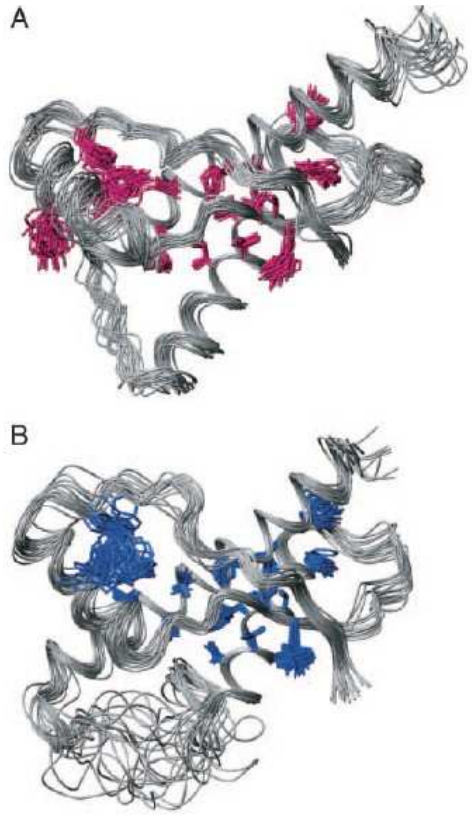


Figure 4.7: NMR structures of the globular domains of HuPrP(A) and ChPrP(B), side chains shown in pink and blue, respectively.

in particular the not well defined N-terminal secondary structure [42]. Experimental studies indicate that the N-terminal region of PrP^C plays a regulatory role in the PrP^C-PrP^{Sc} conversion [45, 46]. Furtherly, the N-terminal half of chicken prion protein is essential for the anterograde axonal transport [47] and has initially been described to induce acetylcholine-receptor activity [2], driving also the clathrin-coated pits endocytosis [48], supposingly because of the abundance of glycines, prolines and the possibility of forming β turn motifs. Besides the wide conformational freedom of the N-terminal backbone, the significant content of proline residues may in principle bring about further complications in the structure determination of this part of protein, due to the presence of an equilibrium between cis and trans form in each Xxx-Pro peptide bond, even if the stabilization of a preferred isomer may be assisted by other residue side chains, particularly histidine and tyrosine. In native and folded proteins the Xxx-Pro bond exists

essentially in either cis or trans form because the interactions with neighbouring groups favours one of them, but an equilibrium exists in unfolded proteins. In small peptides containing proline residues, the two forms are almost isoenergetic and are both present, with an observed increase of trans isomer going from polar to non-polar solvents [49–51]. In solution, polyproline peptides have been shown to exist predominantly in the Poly-L-proline II (PPII) conformation, a left-handed helix in which all the residues are in the trans form. Also short peptides not containing proline are supposed to partially adopt this conformation, but a PPII helix seems to be stabilized in protein domains having an intra-chain PXXP motif (see chapter 2). One of these motifs has been identified in mammalian as well as avian prion proteins and it corresponds to the region 101-104 in mouse PrP^C and 107-111 in chicken PrP^C [52]. CD data concerning the entire protein and different peptide fragments containing hexarepeat sequences clearly suggest the presence of more than one conformation and not simply of a random coil one [33, 53, 54]. In particular, as it will be also shown in the following, this has been pointed out by previous studies on bis-hexarepeats, different mono hexarepeats and analogues with single residue mutation (i.e. a tyrosine replaced by phenylalanine), where the CD spectral shapes and thus the conformational equilibria have been found to be strongly dependent on pH. It has been also suggested that histidine and tyrosine residues play a role in stabilizing one of the conformers and that the location of the PXXP motif along the peptide backbone may also determine the conformational features. Considering all these points, a more structured conformation adopted by chicken hexarepeat region with respect to the mammal analogue could explain the different behavior towards proteolysis. To prove this speculation, molecular dynamics studies was carried out to further study the chicken prion protein structure [55–57]. A computational study of the conformation of the prion protein N-terminal part is actually of relevant interest, especially considering that the available NMR structure of chicken prion protein is only restricted to the globular core sequence (128-242), being the assignment of the N- and C-terminal tails hampered until now by the absence of rigid secondary structure elements. Therefore, in the following sections the study of the avian prion protein structure is shown. In particular, in section 4.5 the NMR and pH dependent molecular dynamics investigations of the mono-hexarepeat, Ac-PHNPGY-NH_2, (4.5) are shown for all the isomers due the Xxx-Pro peptide bond, either cis or trans. In section 4.6 the dynamics and circular dichroism spectra of the longer tetra-hexarepeat fragment, Ac-$(PHNPGY)_4$-NH_2 as a function of pH is discussed and finally in section 4.7 the dynamics of the overall chicken

prion structure, ChPrP1-267 is investigated.

4.5 Conformational features of the N-terminal domain: PHNPGY

4.5.1 NMR measurements

The peptide under study was synthesized as previously reported [58]. All the experiments were carried out at 500 MHz on a VARIAN INOVA UNITY PLUS located at the Department of Chemical Sciences, Catania (Italy), and on a VARIAN UNITY 500 spectrometer, located at the Department of Environmental Sciences, Caserta (Italy). Spectra were processed using the VARIAN VnmrJ and XEASY [59] software. Sample solutions (5mM) were prepared in TFE-d3/H2O (80/20 v/v) and H2O/D2O 90/10 (v/v). NMR spectra for the three-dimensional structure determination were collected at 300 K and referenced to external TMS ($\delta = 0$ ppm); the pH of the aqueous solution was adjusted at 4.2. The dependence of the amide chemical shifts on the temperature was observed in the range 300-311 K. Furthermore, for all the solvent systems the possible occurrence of aggregation was verified by analyzing the cross peak patterns in ROESY spectra, recorded at a peptide concentration of 0.8 mM. Deuterated D2O (99.9 % relative isotopic abundance) and TFE-d3 (99%) were purchased from Cambridge Isotope Laboratories. Mono (1D) and two dimensional (2D) spectra were accumulated with a spectral width of 6000 Hz in H2O/D2O and 4800 Hz in TFE/H2O (80/20 v/v). 2D experiments DQFCOSY, TOCSY, ROESY and NOESY [60] were recorded in the phase sensitive mode using the States-Haberkorn method. Water suppression was achieved by DPFGSE sequence [61]. TOCSY, NOESY and ROESY spectra were acquired with mixing time of 70, 250 and 150 ms, respectively. Typically, 64 transients of 4K data points were collected for each of the 256 increments; the data were zero filled to 1K in ω_1. Squared shifted sine-bell functions were applied in both dimensions prior to Fourier transformation and baseline correction.

4.5.2 Structure calculations

Due to unfavourable correlation time, NOESY experiments were scarcely informative and not suitable for structure calculations. Therefore distance restraints for structure

calculations were derived from the cross-peak intensities in ROESY spectra, recorded in H2O/D2O and TFE/H2O (80/20 v/v). The ROESY cross peaks were manually integrated using the XEASY software [59] and converted to upper distance constraints according to an inverse sixth power peak volume-to-distance relationship for the backbone and to an inverse fourth power function for side chains, by using the CALIBA module of the CYANA program [62]. Distance constraints together with the obtained scalar coupling constants were then used by the GRIDSEARCH module, implemented in CYANA, to generate a set of allowable dihedral angles. Structure calculations, obtained by using the torsion angle dynamics protocol of CYANA, were then started from 100 randomized conformers. The 20 conformers with the lowest CYANA target function were further refined by means of unrestrained energy minimization in vacuo, using the GROMOS 96 [115] force field (FF) within the program SPDB viewer [63]. Several cycles of steepest descent were repeated until the energy difference between two successive steps was less than 10^{-3} kJ mol^{-1}. The structure analysis has been performed with the program MOLMOL [64]. Consistently with suggestions derived from a recent work [37] the conformers were minimized in a water shell and in a box with TFE/H2O 80/20 with a conjugate gradient method and tolerance of 10^{-3} kJ mol^{-1}, using ORAC 4.0 program [65] and the Amber94 FF [66].

4.5.3 CD measurements

CD spectra were recorded on a JASCO 810 spectropolarimeter at a scan rate of 50 nm/minute and 0.1 nm resolution. The pathlengths were 1 or 0.1 cm, in the 190-800 nm range. The spectra were recorded as an average of 10 or 20 scans. The CD instrument was calibrated with ammonium (+)-camphor-10-sulfonate. Peptide solutions were prepared in water in a concentration range of 10^{-5}- 10^{-6} mol dm^{-3} and varying the pH by addition of a diluted solution of potassium hydroxide or hydrochloric acid. Tempetarure was ranged from 277 K to 348 K, with an incubation time of 15 minutes at each temperature prior to data collection.

4.5.4 Molecular Dynamics

All the simulations were run in water using ORAC 4.0 [65] and the Amber94 FF [66]; a cubic sample containing one 1-HexaPY chain and 1184 water molecules with periodic boundary conditions (PBC) was studied in the isothermal-isobaric ensemble (NPT, P=1

atm, T=300 K); temperature was controlled with a Nosé-Hoover thermostat [67] and the SPC model [68] was used for water. The ESP charges of deprotonated tyrosine, not available in the FF, were calculated at HF/6-31G* level for the N-acetyl-(L)-tyrosinate amide after geometry optimization, following the Amber charge fitting philosophy. An r-RESPA multiple time-step algorithm with a potential subdivision specifically tuned for proteins [65] was used for integrating the equations of motion, using a time step of 10 fs. Due to its partial double bond character, the torsional barrier around the peptide bond Xxx-Pro is not thermally overcome [69] in the 1-100 nanosecond time scale, typical of MD simulations. Consequently, the different isomers, originated from the Asn-Pro and the Ac-Pro ω torsion angles (N and C-termini were capped with acetyl and amide groups, respectively), were simulated separately. For each isomer, three different protonation states were studied to mimic the different pH conditions. Taking into consideration the pKa of histidine and tyrosine (6.34 and 9.77 respectively [58]), it is possible to assume in the simulations that at acidic pH, the histidine is protonated and tyrosine is in its neutral form (labelled H^+Y), while at neutral pH both histidine, in the δ form, and tyrosine are in their neutral states (labelled HY) and finally that at basic pH histidine is in the δ neutral form and tyrosine is deprotonated (labelled HY^-). One chloride ion and one sodium ion were added at acidic and basic pH conditions respectively, to ensure charge neutrality. The total simulation time was about one hundred nanoseconds for each of the twelve simulations of the four isomers in three different protonation states. Each run was equilibrated for at least 30 ns to avoid a starting configuration bias and, after equilibration, both volume and total energy were checked fluctuated around their average value, without systematic drifts. The trajectory analysis was performed on 73 ns-long production runs, with configurations stored every 5 ps. In order to assess if the force field was able to predict PPII conformations, 30 ns simulation of 9-mer poly-L proline was carried out in the zwitterion form (NPT, P=1 atm, T=300 K) using a cubic box with 2123 explicit water molecules and PBC. The PPII structure (ϕ= -75, ψ= 145) [70] was retained for the whole simulation time, with the exception of the C-terminal proline, whose average ψ angles resulted in equilibrium between -30 and 150 values. During the 1-Hexapy simulations, the presence of hydrogen bonds between carbonyl oxygens and amide hydrogens belonging to different amino acids was estimated following the DSSP criteria [71]. To investigate their persistence we also calculated the variation of a given O-H distance and the angle between the C-O and N-H vectors (τ) as a function of time, on one hand to understand the orientation of the two atoms involved

in the hydrogen bond and on the other to calculate the τ angle/O-H distance/energy surface so as to decide if a hydrogen bond was present. The hydrogen bond energy used in this work is purely electrostatic and it is expressed as follows [71]:

$$E_{HB} = uq_1q_2\left(-\frac{1}{r_{CN}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} + \frac{1}{r_{NO}}\right) \tag{4.1}$$

where q1 and q2 are respectively 0.42e and 0.20e, with e the unit electron charge, u=332 the conversion factor from $e^2/Å$ to kcal/mol and r_{CN}, r_{CH}, r_{OH} and r_{NO} the inter-residue distances in Å; a hydrogen bond is considered to be present when $E_{HB} \leq$ -0.5 kcal/mol, $\tau \leq 60°$ and $r_{OH} \leq 3$ Å.

4.5.5 The structure adopted by the mono-hexarepeat fragment

NMR Analysis and Structure Determination

The conformational behavior in solution of the 1-HexaPY hexapeptide was studied by proton NMR spectroscopy in two different solvents: water (H2O/D2O 90/10), because of its biological relevance and trifluoroethanol (TFE)/H2O (80/20 v/v). The TFE/H2O 80/20 mixture, often used as a membrane-mimicking solvent (see, e.g. [72]) and as a hydrogen bond promoting solvent, here can on one hand facilitate the analysis of NMR data in water and on the other contrast it by modifying the H-bond propensity and favouring intramolecular interactions. Identification of the complete spin systems in both solvents was readily accomplished by the homonuclear J-correlated 2D techniques TOCSY and DQF-COSY. Since these were not sufficient to determine their position in the sequence, sequential NOE connectivities between backbone protons were employed to unambiguously assign all the amino acids and their position in the peptide chain. Complete chemical shift assignments in H2O/D2O and TFE/H2O (80/20 v/v) are reported in Table 4.3 and 4.4 respectively.
$^3J_{NH\alpha CH}$ and $^3J_{\alpha CH\beta CH}$ coupling constants and temperature coefficients are reported in Table 4.5.

A key feature of 1-HexaPY is the presence of two proline residues in the sequence; the first one is located at the beginning of the chain, the other at a central position in the Ac-Pro1-His2-Asn3-Pro4-Gly5-Tyr6-NH_2 sequence: consequently Pro4 plays a major role in determining the distribution of conformers in solution. Starting with the TFE/H2O solvent, the hexapeptide shows a well resolved 1D 1H spectrum, characterized by a good

Table 4.3: Proton chemical shifts (ppm) for the major conformer family (up) and the minor conformer family (bottom) of 1-HexaPY at 300 K in H2O/D2O (90/10, v/v). The values for the Acetyl and CONH2 groups are 2.10 ppm and 7.55/7.05 ppm up), 2.10 ppm and 7.55/7.05 ppm bottom).

Major conformer					
AA	**NH**	**αCH**	**βCH**	**γCH**	**Others**
Pro[1]		4.31	2.20/1.91	1.79	δCH_2 3.61
His[2]	8.58	4.71	3.24/3.13	/	H(2) 8.59 H(4) 7.27
Asn[3]	8.47	4.95	2.82/2.66	/	γNH_2 7.60/6.93
Pro[4]		4.41	2.25/2.02	1.95	δCH_2 3.85/3.73
Gly[5]	8.41	3.91/3.84	/	/	/
Tyr[6]	7.91	4.52	3.05/2.95	/	H(2,6) 7.13 H(3,5) 6.84
Minor conformer					
AA	NH	αCH	βCH	γCH	Others
Pro[1]		4.47	2.34/1.92	1.72	δCH_2 3.45
His[2]	8.76	4.70	3.20/3.14	/	H(2) 8.61 H(4) 7.29
Asn[3]	8.63	4.77	2.82/2.65	/	γNH_2 7.60/6.95
Pro[4]		4.20	2.15/1.85	1.65	δCH_2 3.55
Gly[5]	8.49	3.90	/	/	/
Tyr[6]	8.00	4.47	3.08/2.95	/	H(2,6) 7.13 H(3,5) 6.80

dispersion of the proton resonances and a predominant set of sharp peaks for amide protons. Its features indicate the existence of two main families of conformers: a dominant one representing 95% of the population, and a minor one covering less than 5%. Two observable sets of resonances, presumably due to the cis-trans isomers about the Xxx-Pro bond and with relative populations 83% and 17%, are clearly present also in the proton spectrum of 1-HexaPY in H2O/D2O. The hexapeptide assumes a trans conformation for the Ac-Pro and Asn-Pro peptide bonds in TFE/H2O, as indicated by the strong NOEs between the H^{δ} of Pro1 and Ac, and between the H^{δ} of Pro4 and H^{α} of Asn3; also in H2O/D2O, both the Xxx-Pro bonds of the major conformer family assume a trans conformation as confirmed by the presence of two NOE contacts between the H^{δ} of Pro1 and Ac, and between the H^{δ} of Pro4 and H^{α} of Asn3 (see Tables 4.6 and 4.7). The presence of a less populated conformer group is presumably due to a cis arrangement of the

Table 4.4: Proton chemical shifts (ppm) for 1-HexaPY at 300 K in TFE/H2O (80/20, v/v). The values for the Acetyl and CONH2 groups are 2.07 ppm and 7.14/6.54 ppm, respectively.

AA	**NH**	**αCH**	**βCH**	**γCH**	**Others**
Pro[1]	-	4.30	2.18/1.88	1.96	δCH_2 3.60/3.52
His[2]	7.93	4.68	3.29/3.13	/	H(2) 8.43 H(4) 7.21
Asn[3]	8.16	4.92	2.81/2.68	/	γNH_2 7.30/6.48
Pro[4]	-	4.41	2.25/1.95	2.00	δCH_2 3.79/3.73
Gly[5]	8.19	3.82/3.79	-	-	-
Tyr[6]	7.57	4.54	3.08/2.95	/	H(2,6) 7.09 H(3,5) 6.81

Asn3-Pro4 peptide bond since an Ac-Pro1 NOE contact is not observed in the ROESY spectrum and, moreover, the H^{α} resonance of Asn3 of the minor conformer is obscured by the water signal. All amide protons present a linear and negative dependence of their chemical shifts on temperature; the backbone amide protons of residues His2 and Tyr6 in TFE/H2O and Tyr6 in H2O/D2O show low temperature gradients in both solutions (Table 4.5). In particular, the value obtained for the Tyr6 was the lowest and identical in both solvent systems indicating that this amide proton has a good tendency to be solvent shielded likely by its aromatic side chain, as the up-field shifted chemical shift of Tyr6 amide proton also suggests. The scalar coupling constants, $^3J_{NH\alpha CH}$, for residues His2, Asn3 and Tyr6 show values in the range 6-8 Hz (Table 4.5), not evidencing the formation of a canonical folded conformation. On the basis of $^3J_{\alpha CH\beta CH}$ coupling constants and αCH-βCH, NH-βCH NOEs, the preferred conformations of the side chains were identified [73] as a trans arrangement for all the amino acids except for Asn3 which, only in TFE/H2O, appears to prefer a gauche conformation. These data were confirmed by an analysis of rotamer populations [74] reported in Table 4.10. After a careful analysis of the ROESY spectra, 50 proton-proton NOE cross peaks were assigned and integrated for 1-HexaPY in TFE/H2O (80/20 v/v), and 37 in H2O/D2O, for the principal conformer family. A list of the relevant distances derived from the integration of the ROESY peaks is reported in Tables 4.6 and 4.7; it should be noted that in both solvents a correlation between the NH of Gly5 and the NH of Tyr6, corresponding to a calculated distance of about 3 Å, is clearly observed. Furthermore, only in TFE/H2O a medium range NOE value could be assigned as a weak correlation between the Pro4 αCH and the Tyr6

Table 4.5: Temperature coefficients (ppb/K) and ^3J coupling constants (Hz) for 1-HexaPY in TFE/H2O (80/20, v/v) and in H2O/D2O (90/10, v/v). In H2O/D2O data are reported only for the major conformer family.

	TFE/H2O 80/20				**H2O/D2O 90/10**			
AA	$\Delta\delta/\Delta$**t**	$^3J_{NH\alpha}$	$^3J_{\alpha\beta}$	$^3J_{\alpha\beta'}$	$\Delta\delta/\Delta$**t**	$^3J_{NH\alpha}$	$^3J_{\alpha\beta}$	$^3J_{\alpha\beta'}$
Pro	-	-	5.3	8.5	-	-	5.1	8.5
His	-4.9	7.8	5.5	7.5	-5.6	7.8	6.1	8.6
Asn	-6.7	6.8	7.4	6.3	-6.5	7.0	6.3	8.1
Asn γ	-7.0/-6.2	-	-	-	-4.6/-4.6	-	-	-
Pro	-	-	5.1	8.5	-	-	5.5	8.5
Gly	-5.7	5.8/5.8	-	-	-6.3	6.0	-	-
Tyr	-4.5	7.5	6.3	8.0	-4.5	7.3	7.2	8.4
$CONH_2$	-/-6.2	-	-	-	-4.6/-3.7	-	-	-

amide proton. Four $^3J_{NH\alpha H}$ and ten $^3J_{\alpha H\beta H}$ coupling constants were extracted from the 1H monodimensional spectra of 1-HexaPY either in TFE/H2O (80/20 v/v) and H2O/D2O (Table 4.5). Stereospecific assignments for His2 (βCH2 protons) and Asn3 (βCH2 and δCH2 protons) of 1-HexaPY in TFE/H2O (80/20 v/v) were derived from the input data, using the GRIDSEARCH module of CYANA software. The final input files for the structure calculation contained 26 meaningful distance constraints (16 intraresidue and 10 short-range) and 13 angle constraints for the 1-HexaPY in H2O/D2O, and concerning the study in TFE/H2O (80/20 v/v), 26 meaningful distance constraints (15 intraresidue, 10 short- and 1 medium-range) and 23 angle constraints. These constraints were then used to generate a total of 100 structures and among them the 20 structures with the lowest target function values were selected and energy minimized. The best 20 CYANA conformers of 1-HexaPY in H2O/D2O (target function 0.77) and TFE/H2O (80/20 v/v) (target function 1.05) were then minimized in vacuo and in a solvent shell; In TFE/H2O (80/20 v/v), the NMR structure of 1-HexaPY is well defined (RMSD 0.16 in the 1-4 region; RMSD= 0.38 in the region 1-5) and the average dihedral angles of the 20 NMR structures minimized in vacuum and in the proper solvent are reported in Tables 4.8 and 4.9. In this solvent system Pro4 assumes ϕ and ψ angles close to the PPII conformation, while Pro1, His2, Asn3 conformational angles only reflect a propensity towards this elongated conformation. The NMR structure in H2O/D2O is

reasonably defined (RMSD=0.43 in the 1-4 region; RMSD= 0.59 in the region 1-5) and elements of PPII conformation are present (Figure 4.8, a and b), which include the Asn3 and Pro4 residues; the presence of Gly5 causes the break of this secondary structure and allows Tyr6 to experience an higher conformational freedom.

Table 4.6: Relevant ROESY (150 ms)-derived distances intra-residues (up) and inter-residues (bottom) for 1-HexaPY in H2O/D2O (90/10, v/v).

PRO1	HB2	PRO1	HA	3.02
PRO1	HB3	PRO1	HA	2.93
HIS	HN	HIS	HA	3.24
HIS	HN	HIS	HB3	2.77
HIS	HN	HIS	HB2	3.11
HIS	HA	HIS	HB3	2.49
HIS	HB3	HIS	HD2	3.27
HIS	HB2	HIS	HD2	3.48
ASN	HN	ASN	HA	2.83
ASN	HN	ASN	HB2	3.05
ASN	HN	ASN	HB3	2.62
ASN	HA	ASN	HB2	2.43
ASN	HB2	ASN	HD21	3.76
ASN	HB3	ASN	HD21	2.99
PRO4	HB2	PRO4	HA	2.40
PRO4	HB3	PRO4	HA	3.86
PRO4	HD3	PRO4	HB2	5.50
GLY	HN	GLY	HA2	2.46
GLY	HN	GLY	HA1	2.65
TYR	HN	TYR	HA	2.99
TYR	HN	TYR	HB2	2.93
TYR	HN	TYR	HB3	2.52
TYR	QD	TYR	HB3	4.63
TYR	QD	TYR	HB2	4.66
TYR	HA	TYR	QD	4.76

PRO1	HA	HIS	HN	2.49
HIS	HA	ASN	HN	2.40
ASN	HA	PRO4	HD2	2.40
ASN	HA	PRO4	HD3	2.40
ASN	HB2	PRO4	HD3	5.50
ASN	HB3	PRO4	HD3	5.50
PRO4	HA	GLY	HN	2.46
GLY	HN	TYR	HN	3.02
GLY	HA2	TYR	HN	2.68
GLY	HA1	TYR	HN	2.68

Table 4.7: Relevant ROESY (150 ms)-derived distances intra-residues (up) and inter-residues (b) (Å) for 1-HexaPY in TFE/H2O (80/20 v/v).

PRO1	HB2	PRO1	HA	3.24
HIS	HN	HIS	HA	2.86
HIS	HN	HIS	HB3	3.21
HIS	HN	HIS	HB2	3.70
HIS	HA	HIS	HB2	2.40
HIS	HA	HIS	HB3	3.52
HIS	HB3	HIS	HD2	4.01
HIS	HB2	HIS	HD2	4.07
HIS	HA	HIS	HD2	4.51
ASN	HN	ASN	HA	2.86
ASN	HN	ASN	HB2	2.90
ASN	HN	ASN	HB3	2.90
ASN	HA	ASN	HB3	2.96
ASN	HA	ASN	HB2	2.62
ASN	HB2	ASN	HD21	3.83
ASN	HB3	ASN	HD21	3.27
PRO4	HB2	PRO4	HA	2.65
PRO4	QG	PRO4	HA	5.11
GLY	HN	GLY	HA2	3.27
GLY	HN	GLY	HA1	2.80
TYR	HN	TYR	HA	2.93
TYR	HN	TYR	HB2	3.11
TYR	HN	TYR	HB3	2.90
TYR	HA	TYR	HB3	2.90
TYR	HA	TYR	HB2	3.45
TYR	HA	TYR	QD	4.94
TYR	HB2	TYR	QD	5.22
TYR	HB3	TYR	QD	5.25

PRO1	HA	HIS	HN	2.49
HIS	HA	ASN	HN	2.40
HIS	HB3	ASN	HN	2.40
HIS	HB2	ASN	HN	2.40
ASN	HA	PRO4	QD	5.50
PRO4	HA	TYR	HN	5.50
PRO4	HA	GLY	HN	2.46
GLY	HN	TYR	HN	3.02
GLY	HA2	TYR	HN	2.68
GLY	HA1	TYR	HN	2.68

Table 4.8: Average dihedral angles obtained from the 20 NMR structures minimized in vacuo for 1-HexaPY in H2O/D2O (90/10, v/v) up) and in TFE/H2O (80/20, v/v) (bottom).

AA	ϕ			ψ		
Pro^1	-75	±	1	80	±	43
His^2	-92	±	78	98	±	17
Asn^3	-73	±	7	142	±	4
Pro^4	-73	±	5	124	±	30
Gly^5	7	±	88	0	±	78
Tyr^6	-93	±	26	94	±	50

AA	ϕ			ψ		
Pro^1	-75	±	1	177	±	1
His^2	-54	±	1	173	±	1
Asn^3	-52	±	1	125	±	1
Pro^4	-75	±	1	126	±	30
Gly^5	-79	±	87	-1.5	±	15
Tyr^6	100	±	28	65	±	67

Table 4.9: Average dihedral angles obtained from the 20 NMR structures minimized in solvent for 1-HexaPY in H2O/D2O (90/10, v/v) up) and in TFE/H2O (80/20, v/v) (bottom).

AA	ϕ			ψ		
Pro^1	-71	±	10	103	±	60
His^2	-90	±	45	89	±	28
Asn^3	-65	±	21	142	±	9
Pro^4	-68	±	7	138	±	45
Gly^5	-173	±	70	1	±	56
Tyr^6	-111	±	40	67	±	64

AA	ϕ			ψ		
Pro^1	-69	±	7	155	±	42
His^2	-87	±	28	148	±	33
Asn^3	-39	±	51	117	±	18
Pro^4	-72	±	8	135	±	40
Gly^5	-62	±	52	20	±	30
Tyr^6	124	±	33	110	±	66

Table 4.10: Distribution of χ_1 side chain rotamers for 1-HexaPY in H2O/D2O (90/10, v/v) and TFE/H2O (80/20 v/v.), being **t** and **g** trans and gauche, respectively.

AA	$\mathbf{P}_I$ **(g-)**	$\mathbf{P}_{II}$ **(t)**	$\mathbf{P}_{III}$ **(g+)**
Pro	0.23	0.54	0.23
His	0.32	0.55	0.13
Asn	0.34	0.50	0.16
Pro	0.26	0.54	0.20
Tyr	0.42	0.53	0.05

AA	$\mathbf{P}_I$ **(g-)**	$\mathbf{P}_{II}$ **(t)**	$\mathbf{P}_{III}$ **(g+)**
Pro	0.25	0.53	0.22
His	0.26	0.45	0.29
Asn	0.44	0.34	0.22
Pro	0.23	0.53	0.24
Tyr	0.34	0.49	0.17

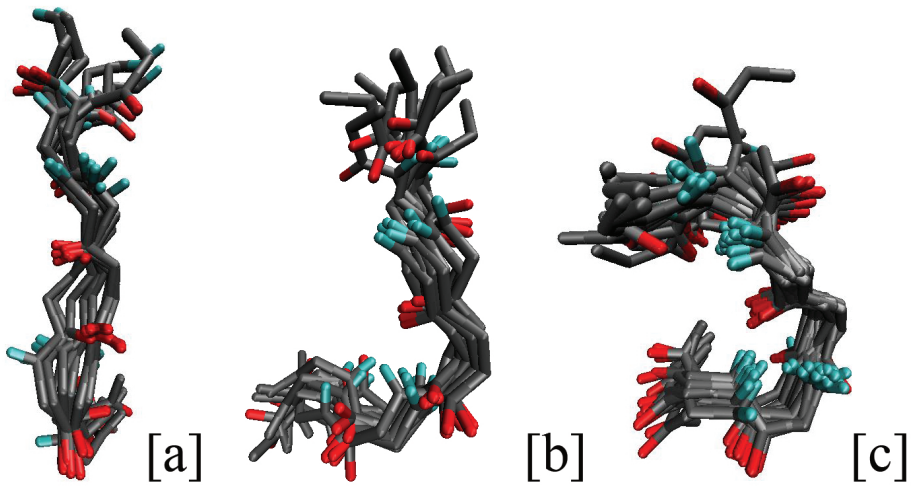


Figure 4.8: From left to right: 1-HexaPY trans-trans conformers obtained from CYANA structures minimized in a water shell at acidic pH: disordered [a] and turn-like [b]. 1-HexaPY trans-trans [c] isomer obtained from MD simulations in the H^+Y form sampled every 2.5 ns. 1-HexaPY trans-trans MD isomer [c] shows the turn conformer, underlined also in the NMR structures [b].

Circular Dichroism Investigations

CD spectra of 1-HexaPY (*Ac-PHNPGY-NH$_2$*), previously reported [58], show a very strong pH dependence. In particular, proceeding from acidic to basic pH values, it is possible to observe an intensity increase of the band at around 210 nm and a decrease of $\Delta\epsilon$ at 230 nm. Such a band, approximately at around 230 nm, is very common in structures containing PPII helices [75], [76], although it is still unclear the correlation between this band and the presence of PPII conformation. Furthermore, the CD spectra of 1-HexaPY in aqueous solution does not show the presence of *classical* secondary structures as those ones of α helix or β sheets. Therefore, in order to investigate the dependence of conformational equilibria on pH and to evaluate the relative ratio of different conformers as a function of pH, a set of CD spectra of 1-HexaPY was analyzed at different temperature and pH conditions using the convex constraint analysis, (CCA) [77], [78]. CCA operates on a set of CD spectra, finding the common components among them, constituting the basis set. In this context 42 spectra at various temperature and pH values was used. The basis set found is made up by three components (Figure 4.9), which better minimize the least square error of CD data fit. After the basis set was obtained, the relative molar fractions of the resulting three CCA components over the 42 CD spectra was estimated using the Lincomb program [77].

Component 1 is stabilized on increasing the temperature, from acidic to neutral pH values, decreasing at basic pH values. Furthermore, two minima approximately at 208 and 216 nm are detected, being negative in the far UV-range. This behavior fits well with type I β turn CD spectra of model peptide systems [76], [79], [80], despite the less intense positive band at around 195 nm

Component 2 is the main one at basic pH values and moderate temperatures, since it decreases on temperature increasing. Such a component has a deep minimum at around 200 nm and a shoulder approximately at 220 nm, typical of unordered conformations [76]. A weak positive maximum is present at around 246 nm, probably due to tyrosine side chain aromatic contributions.

Component 3 shows, as component 2, an unordered like spectrum, despite a less intensity, with a positive band centered at around 228 nm and a negative minimum centered at around 205 nm [81]. Such a component is the most important at acidic pH values. Furthermore, it decreases on increasing the temperature as previously reported [82], [83]. β turn conformation involves intra-residue interactions between i and i+3 residues. ΔH°

and ΔS° values (Table 4.11) for component 3-component 1 equilibrium extrapolated from Van't Hoff equation, using the populations reported in Table 4.12, are both positive, as expected because of the solvent interaction stabilization of component 3 with respect to component 1. Therefore the main contribution to lower the free energy is given by entropy and since the enthalpy has a positive sign, equilibria are easily shifted on temperature increasing.

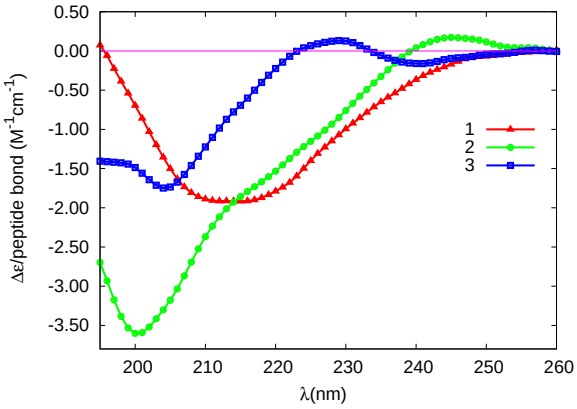


Figure 4.9: The basis set obtained from CCA analysis. Component one is shown in red, component 2 is shown in green and component 3 is shown in blue.

Table 4.11: ΔH° and ΔS° values for Component 3 $\rightleftharpoons$ Component 1 equilibrium.

pH	ΔH° **(KJ** mol^{-1}**)**	ΔS° **(J** $mol^{-1}K^{-1}$**)**
5	39	116
6	40	129
7	37	125
8	60	203
9	74	249
10	81	266

Table 4.12: Coefficients obtained from CCA analysis of 1-HexaPY at different temperature and pH values.

Components	**277 K**	**298 K**	**308 K**	**318 K**	**328 K**	**338 K**	**348 K**
			pH5				
1	0.02	0.12	0.20	0.28	0.35	0.40	0.46
2	0.16	0.24	0.22	0.22	0.22	0.22	0.22
3	0.82	0.64	0.58	0.50	0.43	0.38	0.32
			pH6				
1	0.09	0.22	0.27	0.33	0.41	0.46	0.58
2	0.26	0.26	0.25	0.30	0.25	0.28	0.28
3	0.65	0.52	0.48	0.37	0.34	0.26	0.14
			pH7				
1	0.19	0.32	0.38	0.45	0.49	0.51	0.58
2	0.30	0.31	0.31	0.30	0.30	0.31	0.28
3	0.51	0.37	0.31	0.25	0.21	0.18	0.14
			pH8				
1	0.19	0.32	0.38	0.45	0.50	0.52	0.59
2	0.33	0.46	0.42	0.39	0.39	0.38	0.33
3	0.48	0.22	0.20	0.16	0.11	0.10	0.08
			pH9				
1	0.16	0.26	0.36	0.42	0.47	0.54	0.55
2	0.45	0.61	0.53	0.48	0.48	0.45	0.42
3	0.39	0.13	0.11	0.10	0.05	0.01	0.03
			pH10				
1	0.01	0.22	0.25	0.30	0.43	0.54	0.56
2	0.62	0.16	0.67	0.65	0.54	0.44	0.42
3	0.37	0.12	0.08	0.05	0.03	0.02	0.02

Molecular Dynamics simulations

The NMR and CD study discussed above give a fairly clear picture on the averaged conformation adopted by the 1-HexaPY peptide chain. However, for technical difficulties the NMR study was limited to acidic pH values, thus allowing only a partial interpretation of other available data, such as the CD spectra that cover a wide range of pH values. In order to get a more complete understanding at a molecular level and, in particular, to further investigate the pH dependence of 1-HexaPY conformational states in aqueous solution, a fully atomistic Molecular Dynamics (MD) study was then carried out in water. It is known that the trans conformation of proline tends to stabilize type I β turn (ϕ_{i+1}= -60, ψ_{i+1}= -30, ϕ_{i+2} = -90, ψ_{i+2} = 0) [84] and PPII structures (ϕ= -75, ψ= 145) [70]; conversely the cis conformation of proline stabilizes a special type of turn, devoid of intrachain hydrogen bonds, called VI [85] (VIa1 ϕ_{i+1}= -60, ψ_{i+1}= 120, ϕ_{i+2}= -90, ψ_{i+2}= 0, VIa2 ϕ_{i+1}= -120, ψ_{i+1}= 120, ϕ_{i+2}= -60, ψ_{i+2}= 0, VIb ϕ_{i+1}= -135, ψ_{i+1}= 135, ϕ_{i+2}= -75, ψ_{i+2}= 160). To claim the presence of such conformers, the existence of consecutive amino acid residues with the typical angles is required, thus the percentage of type I β turn and PPII was calculated considering both 2, 3, 4 consecutive residues (see Table 4.13), using the standard tolerance interval of $\pm$ 30° for all the dihedrals. During the MD simulations, no more than two consecutive residues with type I β turn ϕ and ψ dihedrals were found, indicating a high conformational flexibility of the hexarepeat. The time evolution of the conformers (Figure 4.10), calculated taking into account only two residues, proves the fast flipping of the backbone dihedrals, showing also a prevalence of angles typical of β turn and unordered structures. The analysis of i-i+3 hydrogen (H) bonds and of intrapeptide contacts (evaluated as $C_{\alpha i}$-$C_{\alpha i+3}$ distances inferior to 7 Å [86]), revealed the pivotal role of asparagine and of the conformation of prolines on the peptide structure: in fact, H-bonds and contacts always involve asparagine, despite its tendency to adopt PPII dihedrals registered both by NMR and MD, and they exist only in regions containing trans prolines (cf H-bond percentages in Table 4.14). In particular a NPGY β turn was found fulfilling the DSSP criteria [71], and a second turn region was detected, albeit in lower extent, in the AcPHN sequence. The existence of turn structures is indicated also by $C_{\alpha i}$-$C_{\alpha i+3}$ distances, quite large for PHNP and HNPG segments, and short in the NPGY C-terminal sequence (Table 4.18). Summarising, the peptide shows a tendency to adopt β turn conformations, which on the other hand exist only in a fraction of the trajectory and quickly interconvert with

unordered ones.

Table 4.13: MD percentage of two consecutive amino acids with typical ϕ and ψ dihedral angles of PPII and type I β turn (β I). Type VI β turn (β VIb) is regarded only to trans-cis and cis-cis isomers. The remaining percentage is referred to unordered dihedrals.

H^+Y										
	trans-trans		**cis-trans**		**trans-cis**			**cis-cis**		
Residues	β I	PPII	β I	PPII	β I	PPII	β VIb	β I	PPII	β VIb
PRO-HIS	44.4	0.6	-	7.2	42.1	1.0	-	30.8	-	-
HIS-ASN	-	0.8	-	5.9	-	-	-	-	0.3	-
ASN-PRO	-	1.5	-	-	-	12.6	20.1	-	3.7	33.0
PRO-GLY	61.8	-	18.1	-	-	1.0	-	-	0.4	-
GLY-TYR	11.8	-	7.2	-	24.0	-	4.0	23.6	-	-
average	23.6	0.6	5.1	2.6	13.2	2.9	-	10.9	0.9	6.6
HY										
PRO-HIS	46.3	-	16.8	-	30.9	1.2	-	24.7	-	-
HIS-ASN	-	-	-	8.6	-	3.4	-	-	1.4	-
ASN-PRO	-	0.8	-	0.1	-	12.0	17.6	-	4.8	37.3
PRO-GLY	64.4	-	54.2	-	-	1.2	-	-	1.5	-
GLY-TYR	10.9	-	18.2	-	7.4	-	-	29.0	-	-
average	24.3	0.2	17.8	1.7	7.7	3.6	-	10.7	1.5	7.5
HY^-										
PRO-HIS	36.2	0.6	4.4	2.9	32.6	-	-	15.0	1.2	-
HIS-ASN	-	1.5	-	9.1	-	-	-	-	9.7	-
ASN-PRO	-	-	-	0.8	-	-	-	-	5.3	16.3
PRO-GLY	9.5	-	14.2	-	1.0	-	-	5.5	-	-
GLY-TYR	2.3	-	4.9	-	0.2	-	-	24.6	-	-
average	9.6	0.4	4.7	2.6	6.8	0.0	0.0	9.0	3.2	3.3

Table 4.14: pH dependent MD hydrogen bond percentages for 1-HexaPY isomers found between i and i+3 residues.

H^+Y				
H-bond$_{i-i+3}$	**trans-trans**	**cis-trans**	**trans-cis**	**cis-cis**
ACE$_{CO}$-ASN$_{NH}$	76%	0%	57%	0%
ASN$_{CO}$-TYR$_{NH}$	68%	21%	0%	0%
HY				
ACE$_{CO}$-ASN$_{NH}$	84%	0%	38%	0%
ASN$_{CO}$-TYR$_{NH}$	70%	64%	0%	0%
HY^-				
ACE$_{CO}$-ASN$_{NH}$	49%	0%	22%	0%
ASN$_{CO}$-TYR$_{NH}$	8%	14%	0%	0%

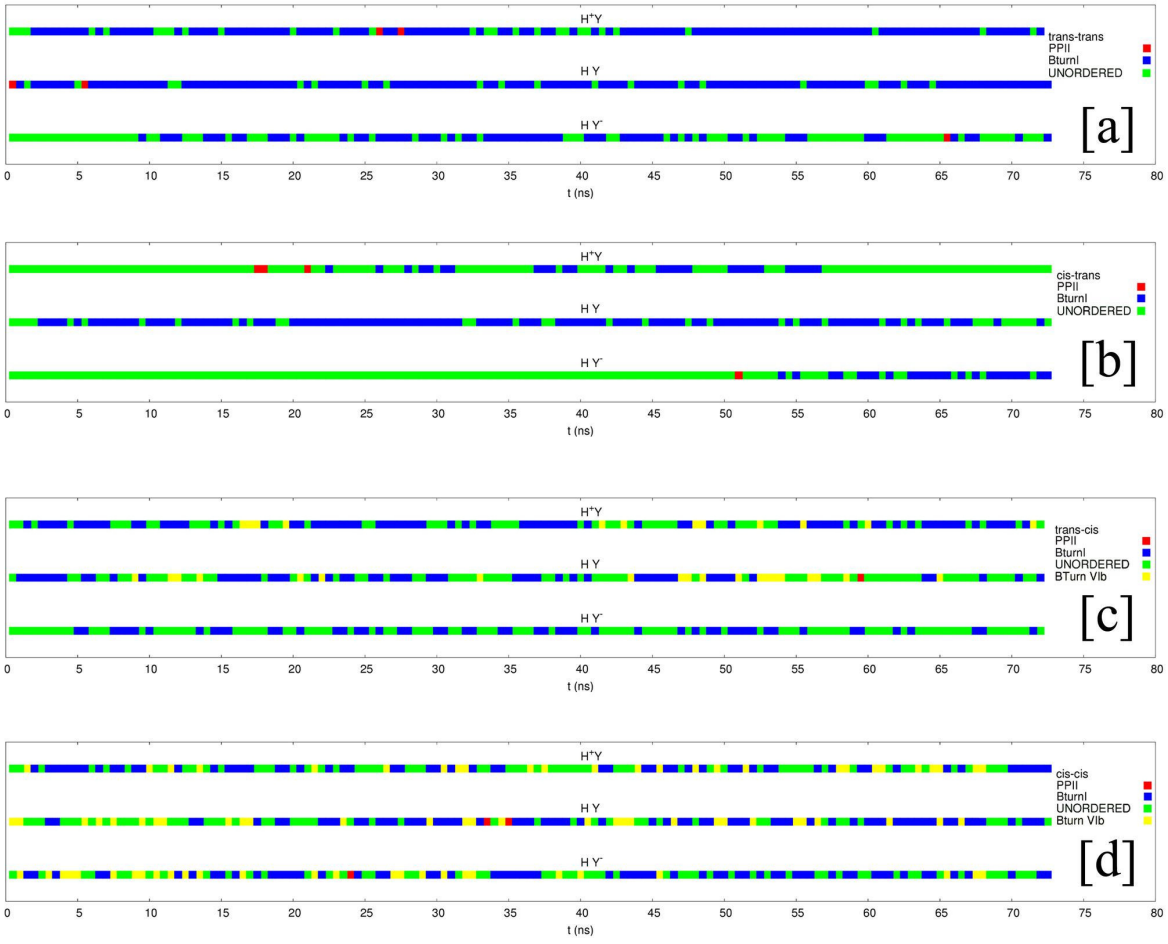


Figure 4.10: Time evolution of 1-HexaPY conformers for each of the four isomers (trans-trans [a], cis-trans [b], trans-cis [c], cis-cis [d]) in the three different protonation states obtained from MD simulations. The calculation is based on at least two consecutive residues with the characteristic dihedral angles concerning type I β turn and three consecutive residues for PPII.

pH-dependent features

The Ramachandran maps of all the isomeric forms, calculated from the simulations, present also strong variations as a function of pH, the most striking feature being the presence of distinct regions of local minima (see Tables 4.15-4.17). For an easier comparison with the NMR results in water, in Figure 4.11 the experimental dihedrals and the MD distribution obtained by the trans-trans simulation at acidic pH is reported. The two maps show a nice agreement, in particular in highlighting that the Asn3 and Pro4 residues assume polyproline II dihedrals. Besides unordered structures, type I β turn results to be the principal conformer at pH lower than tyrosine pKa (9,77 [58], see Table 4.13). The probability of turn conformation in the NPGY and Ac-PHN regions is high when tyrosine is not deprotonated and the involved proline is in trans conformation, mainly in the trans-trans isomer (Figures 4.8b, 4.8c, 4.12a, 4.12b) and in lower amount, in the cis-trans isomer at neutral and acidic pH conditions. The occurrence of such conformations decreases as pH increases, giving raise to unordered structures. The conformational change with pH is likely driven by the interaction between the phenolate oxygen of tyrosine (available only at basic pH) and the side chain amide hydrogen of asparagine. This interaction determines a tilt of the glycine residue, breaking the turn structure (Figure 4.12 a,b) in favour of the unordered ones (Figure 4.12c). The conformational change therefore can be also tracked following the variations of glycine (ϕ and ψ distribution angles minima from -97,1 in the H^+Y and in the HY form, to 137,-15 in the HY^- form (see Tables 4.15-4.17), and more evidently from the distribution of $C_{\alpha i}$-$C_{\alpha i+3}$ distances in the NPGY region (Figure 4.13). This result agrees with the CD spectra blue shift for the minimum around 205 nm at pH 10 and 298 K previously reported [58], and explains the driving force of the phenomenon. The presence of the NPGY turn at acidic pH, visible also in the NMR minimized structures (both in vacuum and in water, Figure 4.8b), is compatible with the ROESY correlation between the NH of Gly5 and the NH of Tyr6. As a final test of the agreement between the MD and the NMR structures at acidic pH the average backbone RMSD values of the 20 NMR structures divided in two groups (disordered and turn as in Figure 4.8 [a] and [b]) with respect to the MD configurations sampled each 2 ns is reported in Table 4.19. Group [b] shows the lowest RMSDs, confirming again the presence of the turn conformation in the NPGY sequence.

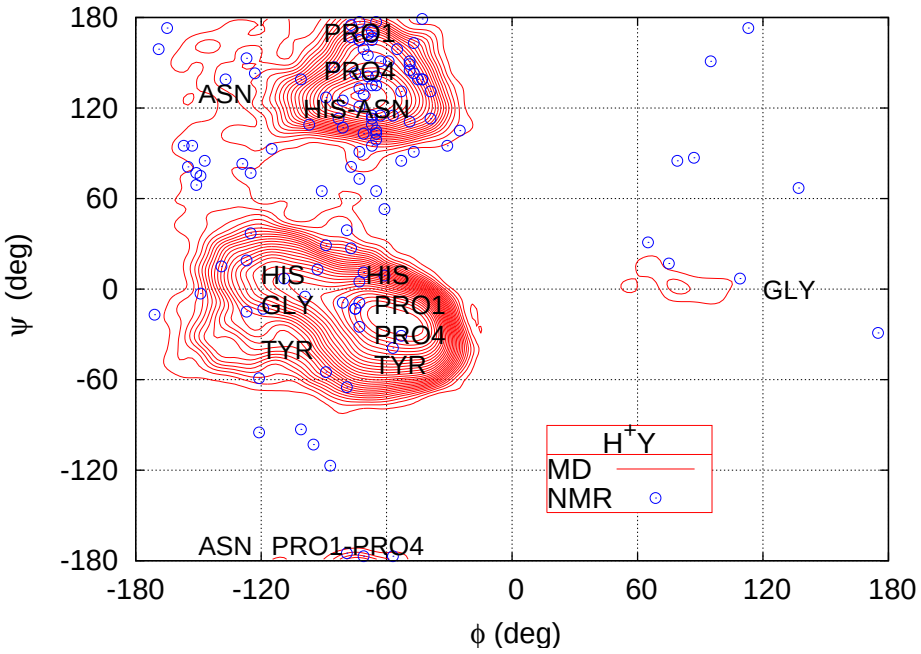


Figure 4.11: Superposition of the backbone dihedrals of the 20 NMR structures minimized in water and the MD (ϕ, ψ) distributions of the 1-HexaPY obtained by the trans-trans simulations at acidic pH.

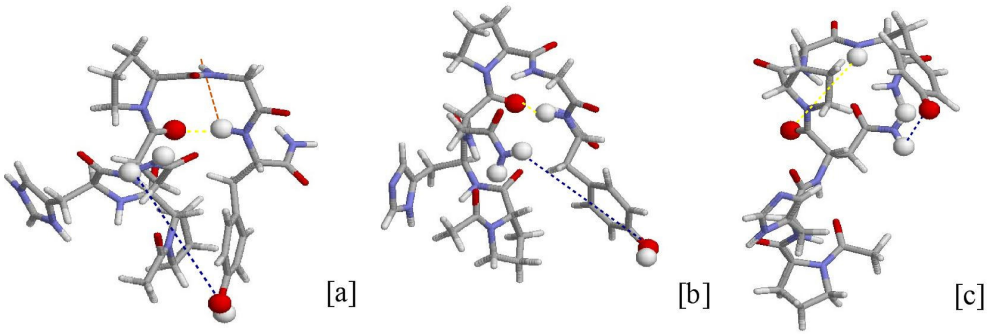


Figure 4.12: 1-HexaPY trans-trans typical conformations in the H^+Y [a], HY [b] and HY^- [c] form obtained from MD simulations. It is worth noting the short distance between the phenolate oxygen of tyrosine and the amide side chain hydrogen of asparagine residue (blue line) only in the HY^- form, where tyrosine is deprotonated. In the H^+Y and HY forms, instead, a turn structure is stabilized by an intra-chain $ASN_{C=O}$-TYR_{NH} H-bond (yellow line). The orange line shows the distance between the NH of Gly5 and the NH of Tyr6, corresponding to a calculated distance of about 3 Å.

Table 4.15: Principal minima of (ϕ, ψ) dihedral angle distribution (deg) and their relative energy values (E, kcal/mol) for 1-HexaPY isomers obtained from MD simulation.

H^+Y												
	trans-trans			**cis-trans**			**trans-cis**			**cis-cis**		
N	ϕ	ψ	E	ϕ	ψ	E	ϕ	ψ	E	ϕ	ψ	E
Pro I												
1	-43	-23	0.0	-71	167	0.0	-47	-23	0.0	-73	-13	0.0
2	-69	157	1.7	-51	-161	0.7	-71	163	1.5	-	-	-
His												
1	-67	-23	0.0	-63	-37	0.0	-119	5	0.0	-111	-21	0.0
2	-99	-5	0.7	-137	-11	1.2	-75	-11	0.9	-75	163	1.1
3	-69	153	2.0	-79	147	1.4	-71	161	2.0	-71	-41	1.6
Asn												
1	-73	129	0.0	-63	125	0.0	-137	95	0.0	-135	101	0.0
2	-107	113	1.5	-91	-179	1.7	-63	145	0.9	-71	143	2.0
3	-	-	-	-	-	-	33	84	1.3	-	-	-
Pro IV												
1	-53	-23	0.0	-67	-13	0.0	-73	167	0.0	-71	165	0.0
2	-65	-175	2.3	-83	-15	1.1	-73	-179	0.5	-	-	-
Gly												
1	-97	-1	0.0	135	-11	0.0	-55	-21	0.0	-51	-23	0.0
2	-68	-12	0.3	-68	-15	0.1	-81	12	1.1	-103	5	1.3
3	90	1	1.9	-114	-8	0.1	-113	-1	1.3	130	1	1.4
4	-	-	-	-	-	-	-72	161	1.5	-71	126	1.6
Tyr												
1	-67	-39	0.0	-75	-19	0.0	-77	-13	0.0	-77	-13	0.0
2	-92	-21	0.8	-94	-12	0.2	-119	3	1.6	-113	1	0.6
3	-61	179	1.9	-68	142	1.8	-	-	-	-	-	-

Table 4.16: Principal minima of (ϕ, ψ) dihedral angle distribution (deg) and their relative energy values (E, kcal/mol) for 1-HexaPY isomers obtained from MD simulation.

HY												
	trans-trans			**cis-trans**			**trans-cis**			**cis-cis**		
N	ϕ	ψ	E	ϕ	ψ	E	ϕ	ψ	E	ϕ	ψ	E
Pro I												
1	-47	-19	0.0	-73	-17	0.0	-49	-21	0.0	-71	-15	0.0
2	-71	163	2.0	-69	177	0.6	-73	167	0.7	-	-	-
His												
1	-69	-23	0.0	-69	-35	0.0	-119	-7	0.0	-115	-21	0.0
2	-91	-17	0.8	-112	-21	0.4	-79	-13	0.7	-70	164	0.9
3	-71	164	1.8	-66	157	0.7	-68	124	1.2	-65	-21	1.1
4	-	-	-	-	-	-	-	-	-	-134	173	1.4
Asn												
1	-75	125	0.0	-57	123	0.0	-137	93	0.0	-133	103	0.0
2	-101	133	1.0	-123	99	1.2	-64	143	0.5	-59	139	1.6
3	-	-	-	-	-	-	37	82	1.7	-	-	-
Pro IV												
1	-57	-19	0.0	-57	-21	0.0	-73	165	0.0	-73	161	0.0
2	-66	145	2.3	-57	131	2.0	-75	-11	1.0	-	-	-
Gly												
1	-97	1	0.0	-91	1	0.0	-49	-25	0.0	-53	-19	0.0
2	-79	-3	0.2	-69	-12	0.1	135	-6	0.1	144	-77	0.7
3	-	-	-	78	10	1.5	57	-98	0.4	141	-152	1.2
4	-	-	-	-	-	-	82	15	0.4	-70	167	1.5
5	-	-	-	-	-	-	140	-171	0.4	-	-	-
6	-	-	-	-	-	-	156	177	0.4	-	-	-
7	-	-	-	-	-	-	-105	8	0.6	-	-	-
8	-	-	-	-	-	-	-67	161	0.7	-	-	-
Tyr												
1	-63	-39	0.0	-73	27	0.0	-73	-19	0.0	-53	-19	0.0
2	-131	-31	1.3	-125	-24	0.7	-116	-1	0.3	-126	2	0.4
3	-68	154	1.9	-75	161	1.7	-148	159	1.5	-148	168	1.5
4	-	-	-	-	-	-	-76	159	1.7	-72	161	1.6

Table 4.17: Principal minima of (ϕ, ψ) dihedral angle distribution (deg) and their relative energy values (E, kcal/mol) for 1-HexaPY isomers obtained from MD simulation.

						HY$^-$						
	trans-trans			**cis-trans**			**trans-cis**			**cis-cis**		
N	ϕ	ψ	E	ϕ	ψ	E	ϕ	ψ	E	ϕ	ψ	E
						Pro I						
1	-49	-21	0.0	-71	157	0.0	-47	-29	0.0	-71	-13	0.0
2	-67	161	1.2	-77	-13	0.3	-69	163	1.6	-73	161	1.3
						His						
1	-67	-19	0.0	-67	-37	0.0	-117	-27	0.0	-119	-19	0.0
2	-117	3	0.5	-65	156	0.6	-67	-13	1.0	-74	157	0.6
3	-68	161	1.7	-112	-19	0.6	-	-	-	-67	-12	1.7
						Asn						
1	-67	133	0.0	-67	129	0.0	-141	109	0.0	-59	139	0.0
2	-137	150	1.5	-137	121	1.1	-71	139	1.0	-133	98	0.3
						Pro IV						
1	-49	-21	0.0	-51	-21	0.0	-71	-17	0.0	-73	165	0.0
2	-	-	0.0	-62	163	2.3	-	-	-	-79	-15	0.1
						Gly						
1	137	-15	0.0	137	-15	0.0	129	23	0.0	-53	-27	0.0
2	-95	10	1.1	-90	4	0.6	-157	23	1.3	146	-15	1.0
3	-66	-7	1.7	-62	-22	0.9	-95	5	1.7	-115	5	1.3
						Tyr						
1	-79	-15	0.0	-73	-15	0.0	-63	-9	0.0	-63	-19	0.0

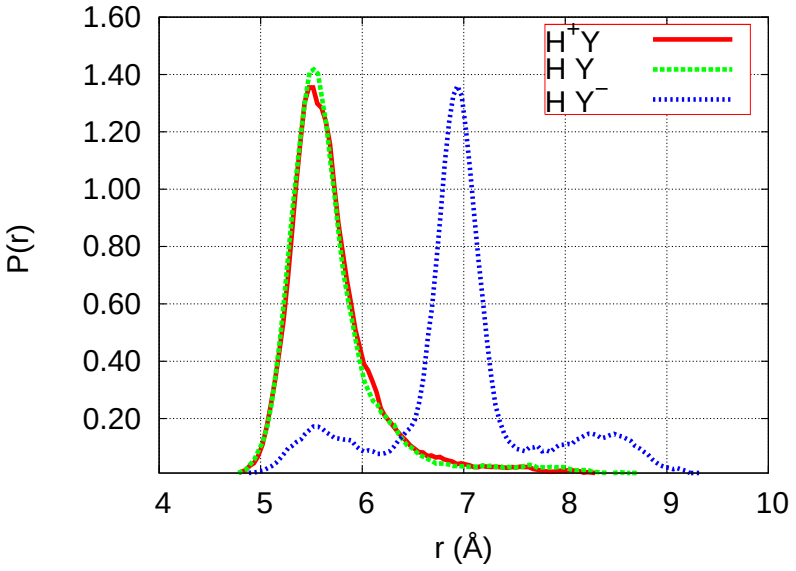


Figure 4.13: $C_{\alpha i}$-$C_{\alpha i+3}$ distance distribution concerning the NPGY region of 1-HexaPY trans-trans in the three different protonation states. It can be noticed the short distance in the H^+Y and HY forms and the larger distance distribution in the HY^- form.

Table 4.18: $C_{\alpha i}$-$C_{\alpha i+3}$ distances in 1-HexaPY isomers, ordered according to their probability. For acetyl group, methyl carbon is considered.

Residues												
	trans-trans			**cis-trans**			**trans-cis**			**cis-cis**		
						H^+Y						
AC-PRO-HIS-ASN	5.9	8.7	-	7.8	-	-	5.0	8.6	-	4.6	6.3	-
PRO-HIS-ASN-PRO	8.0	-	-	7.3	8.5	5.7	7.3	8.9	-	6.4	8.7	-
HIS-ASN-PRO-GLY	8.6	-	-	8.9	-	-	4.2	7.7	-	4.1	7.3	-
ASN-PRO-GLY-TYR	5.4	-	-	6.9	5.6	8.4	7.5	4.9	8.3	7.4	8.4	-
						HY						
AC-PRO-HIS-ASN	6.0	8.4	-	4.6	7.8	6.3	5.0	8.5	10.7	4.6	6.2	-
PRO-HIS-ASN-PRO	7.4	-	-	8.5	5.6	-	6.8	8.6	-	6.7	8.3	-
HIS-ASN-PRO-GLY	8.7	-	-	8.8	-	-	4.1	5.2	6.9	4.4	5.8	-
ASN-PRO-GLY-TYR	5.4	-	-	5.5	8.3	-	7.7	8.3	6.5	7.4	6.4	8.4
						HY^-						
AC-PRO-HIS-ASN	5.9	5.0	8.8	7.8	6.5	4.6	4.9	6.0	9.7	6.4	4.7	8.2
PRO-HIS-ASN-PRO	8.8	7.3	-	7.3	8.9	5.7	6.8	8.6	-	8.7	6.7	-
HIS-ASN-PRO-GLY	8.8	-	-	8.9	-	-	3.9	5.8	-	5.7	4.2	7.5
ASN-PRO-GLY-TYR	7.0	5.5	8.3	6.9	5.7	8.3	7.3	6.3	-	6.3	5.6	8.0

Table 4.19: Average backbone RMSD values (Å) of the two clusters among the 20 NMR structures, [a] and [b] as in Figure 4.8, from the MD configurations sampled every 2 ns. Cluster [b] shows the lowest RMSDs, singling out the C-terminal turn conformation. The accuracy of NMR and MD ensembles is 1.17 Å and 0.59 respectively. The average RMSD between NMR structures and the mean of MD configurations is 1.87 Å.

Structures [a]	**Structures [b]**
2.72	1.37
2.46	1.11
2.24	1.08
2.10	1.26
2.10	1.15
2.34	1.88
2.54	0.88
2.11	1.58
2.41	1.17
-	1.17
-	0.89

4.5.6 Brief summary on the mono-hexarepeat section

In this section, NMR and MD simulations were applied to the investigation of the conformational behavior of the avian prion hexarepeat Ac-PHNPGY-NH_2. NMR experiments indicated that the trans-trans isomer is the predominant species in H2O/D2O at pH 4.2 (83%), followed presumably by the trans-cis one (17%), while in the less polar TFE/H2O environment the relative ratio is further increased to 95 to 5. From the analysis of the best 20 NMR structures minimized in vacuum and water, an averaged conformation with characteristic polyproline II dihedrals emerged for the Asn3 and Pro4 residues. The presence of such dihedrals inside a small region of the peptide is not related to a more extended sequence, and the local conformation does not propagate along the backbone, which is found instead to have more frequently either an unordered or a turn-like structure in the NPGY region (Figure 4.8, a and b). MD simulations of the trans-trans isomer in water succeeded in explaining this behavior on the basis of the hydrogen bond distribution, hard to detect from NMR investigations of peptides. In particular, an i-i+3 turn structure is found in the NPGY region at acidic and physiological pH with Asn3 and Pro4 angles typical of polyproline II. Moreover, MD calculations were performed for all four isomers in water at acidic, neutral and basic pH conditions, showing in all the cases a fast interconversion among the accessible conformations. An interesting dependence of the peptide shape on the tyrosine residue deprotonation was also detected (Figure 4.12, 4.13), mainly for the trans-trans isomer. The deprotonation causes a shortening of the distance between the phenolate oxygen of tyrosine and the amide side chain hydrogens of asparagine thus tilting the glycine residue. This appears to be the driving force for the increase of unordered structures at basic pH, explaining the relative blue shift in the CD spectrum [58]. The presence of turns inside the hexarepeat fragment has also been suggested by secondary structure prediction algorithms and it is believed to play a crucial role in the endocytosis processes [87], as the same conformation is also adopted by the NPXY internalization signal of the LDL receptor [88], and tyrosine-containing motifs, essential for coated pit-mediated endocytosis, have been found to adopt an i-i+3 β turn conformation, especially when present in the last position of the hexarepeat [89]. All these experimental studies are not in contrast and seemingly are reinforced by the finding that, depending on the environmental conditions, the avian prion PHNPGY hexarepeat may adopt a turned conformation which includes an i-i+3 hydrogen bond. The present combined NMR and MD approach has provided a significant comprehen-

sion of the origins of the actual conformational distribution in solution, that represents a first step in understanding the structural features of the N-terminal region of the avian prion protein.

4.6 Unveiling the role of histidine and tyrosine residues on the conformation of the avian prion hexarepeat domain: a further look on the more extended tetra-hexarepeat fragment

4.6.1 Peptide synthesis and purification

The peptide Ac-$(PHNPGY)_4$-NH_2 (TetraHexaPY) was synthesized with N- and C-termini blocked on a Pioneer^{TM} Peptide Synthesizer. All residues were introduced according to the HATU/DIEA activation method starting from an Fmoc chemistry on Fmoc-PAL-PEG-PS resin (substitution 0.25 mmol/g, 0.1 mmol scale synthesis, 0.4 g of resin). The synthesis was carried out under a four-fold excess of amino acid at every cycle. Removal of Fmoc protection during synthesis was achieved by means of 20% piperidine solution in DMF. N-terminal acetylation was performed by treating the fully assembled and protected peptide resin (after removal of the N-terminal Fmoc group) with a solution containing acetic anhydride (6% v/v) and DIEA (5% v/v) in DMF. The peptide was cleaved off from the resin and deprotected by treatment with a mixture of water/triisopropylsilane/trifluoroacetic acid (95/2.5/2.5 v/v) for 1.5 hours at room temperature. The solution containing the free peptide was filtered off from the resin and concentrated in vacuo at 30° C. The peptide was precipitated with freshly distilled diethyl ether. The precipitate was then filtered, dried under vacuum, re-dissolved in water and lyophilised. The resulting crude peptide was purified by preparative reversed-phase high-performance liquid chromatography (Rp-HPLC). Rp-HPLC was carried out by means of a Varian PrepStar 200 model SD-1 chromatography system equipped with a Prostar photodiode array detector with detection at 222 nm. Purification was performed by eluting with solvent A (0.1% TFA in water) and B (0.1% TFA in acetonitrile) on a Vydac C18 250x22 mm column (300 Å pore size, 10-15 mm particle size), at flow rate of 10 mL/min. The peptide TetraHexaPY was eluted using a linear gradient (0-20%)

in solvent B. The elution profiles were monitored at 222 nm and 278 nm, and the peptide fractions were collected and lyophilised. Sample identity was confirmed by ESI-MS (Calculated mass TetraHexaPY C126H161N37O33 M=2720.21; found m/z $[M+2H]^{2+}$= 1361.10; $[M+3H]^{3+}$= 907.73; $[M+4H]^{4+}$=681.05).

4.6.2 Potentiometric measurements

Potentiometric titrations were performed with a computer-controlled Metrohm digital pH meter (Model 654) and a Hamilton digital dispenser (mod Microlabm). The titration cell (2.5 ml) was thermostated at 25.0 ± 0.2 C and all solutions were kept under an atmosphere of argon, which was bubbled through another solution under the same conditions of ionic strength and temperature. A KOH solution was added through a Hamilton burette equipped with 0.25 or 0.50 cm^3 syringes. The combined microelectrode (ORION 9103SC) was calibrated on the pH= -log $[H^+]$ scale by titrating HNO_3 with CO_2 free base. The ionic strength of all solutions was adjusted at 0.10 mol dm^{-3} (KNO_3). The analytical concentrations of TetraHexaPY ranged from 2.5 x 10^{-3} to 5.0 x 10^{-3} mol dm^{-3}. Stability constants for proton complexes were calculated from three or four titrations carried out over the pH range 2.5-10.6. Calculations of the electrode system, E°, E_j and K_W values as well as ligand purity were determined by the least square ACBA computer program [90]. The protonation constants were calculated by means of the HYPERQUAD program [91].

4.6.3 CD measurements

CD spectra were recorded on a JASCO 810 spectropolarimeter at a scan rate of 50 nm/minute and 0.1 nm resolution. The pathlengths were 1 or 0.1 cm, in the 190-800 nm range. The spectra were recorded as an average of 10 or 20 scans. The CD instrument was calibrated with ammonium (+)-camphor-10-sulfonate. Peptide solutions were prepared in water in a concentration range of 10^{-5}- 10^{-6} mol dm^{-3} and varying the pH by addition of a diluted solution of potassium hydroxide or hydrochloric acid.

4.6.4 Molecular Dynamics

An extensive Molecular Dynamics (MD) study of the fragment Ac-$(PHNPGY)_4$-NH_2 (TetraHexaPY) was carried out in water. Taking into consideration the pKa of histi-

dine and tyrosine residues (see Table 4.20), by assuming acidic pH in the simulation, histidines are protonated and tyrosines are in the neutral state (labelled $LH_8{}^{4+}$), at neutral pH both histidines, (protonated at the δ nitrogen) and tyrosines are in the neutral state (labelled LH_4) and finally at basic pH histidines are in the neutral state and tyrosines are deprotonated (labelled L^{4-}). Four chloride ions and four sodium ions were added at acidic and basic pH conditions respectively, to ensure charge neutrality in the simulation box. All the simulations were run in water using GROMACS 3.3 [92] and the Amber94 force field [66], using the SPC model [42] for water. The ESP charges of deprotonated tyrosine, not available in the FF, were calculated as previously suggested [55]. The starting configuration of TetraHexaPY (Figure 4.14) was built by linking four times the most representative NMR structure of MonoHexaPY, with all prolines in trans conformation, previously reported [55]; a cubic box containing one TetraHexaPY chain and 4338 water molecules with periodic boundary conditions (PBC) was used for simulations in the isothermal-isobaric ensemble (NPT, P=1 atm, T=300 K), with the temperature controlled using a Berendsen thermostat [28]. Long runs of about 150 nanoseconds for each of the three protonation states were performed. Preliminarily, to assess the effective equilibration of the peptide, the hydrogen bonds formation was analyzed following L. J. Smith et al. [93], and the evolution of the end-to-end distances with time, reported respectively in Figure 4.15 [a] and [b]. The number of hydrogen bonds increases, while the end-to end distances (Figure 4.15 [b]) show a fast decrease in the first 40 ns, with a reorganization time of 10 ns (40 ns considering also the 30 ns of the equilibration) for the LH_4 state, and then remain substantially stable for each case. Accordingly, the first 40 ns were considered as equilibration runs and totally excluded them from the production analysis, to avoid a starting configuration bias. Moreover, after equilibration, volume, total energy, the number of H-bonds and end-to end distances were checked fluctuated around their average value, without systematic drifts. The trajectory analysis was thus performed on 110 ns-long production runs, with configurations stored every 2.5 ps.

Table 4.20: Protonation constants (log β^*) and pK values of TetraHexaPY (T= 298.0 K, I=0.1 mol dm^{-3} KNO_3). $^*3\sigma x10^{-2}$ values are shown in parentheses.

Species	**logβ**	**pK**	**Site of protonation**
HL^{3-}	10.65 (3)	10.65	OH group of Tyr6,
HL^{2-}	20.64 (3)	9.99	Tyr12,
H_3L^-	30.40 (6)	9.76	Tyr18 and Tyr24
H_4L	39.45 (6)	9.05	
H_5L^+	46.3 (9)	6.81	Imidazole of His2,
H_6L^{2+}	52.6 (9)	6.32	His8,
H_7L^{3+}	58.6 (9)	6.01	His14 and His20
H_8L^{4+}	63.7 (9)	5.13	

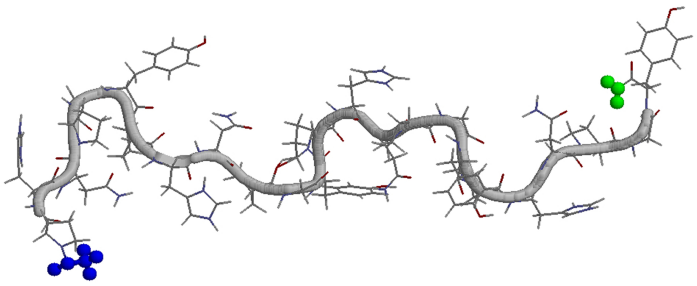


Figure 4.14: Starting configuration of TetraHexaPY, where the N-terminal Acetyl and the C-terminal amide are shown in blue and in green respectively.

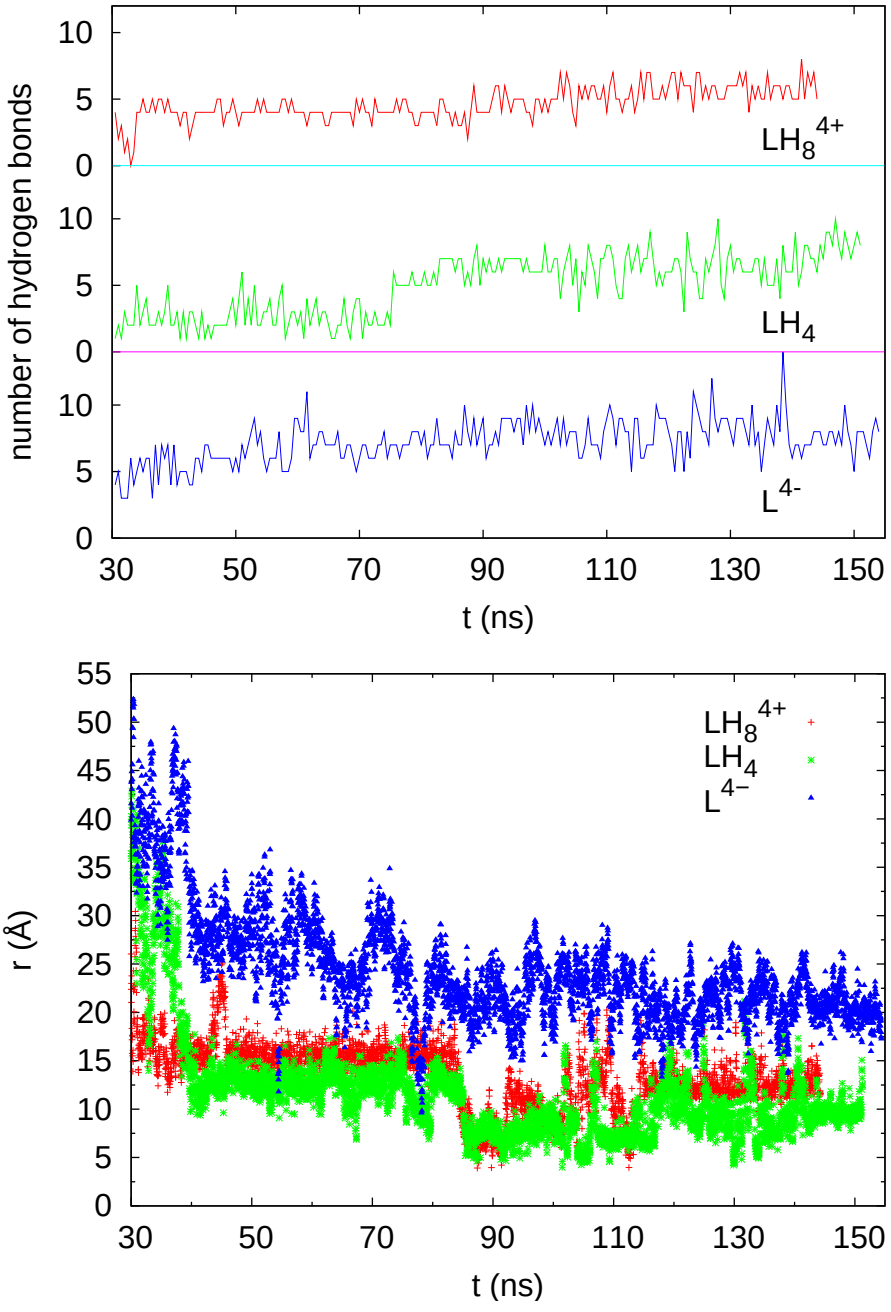


Figure 4.15: Time evolution of hydrogen bonds number [a] and of the end-to-end distance relative to the C_α carbons [b]. Both of them are useful to check the equilibration of the starting configuration of the peptide, initially with an elongated structure and thus having few hydrogen bonds. This starting configuration is clearly lost during the run.

4.6.5 The structure adopted by the tetra-hexarepeat fragment

Protonation constants and far UV-CD of TetraHexaPY

The protonation constants of the TetraHexaPY are reported in Table 4.20; it is clear that the peptide contains eight protonation sites, being their assignments shown in Table 4.20. The phenolic-OH side chain of tyrosine residues have the highest pK values and their deprotonation takes place in overlapping processes between pH 9 and 11. Protonation of the four imidazole-N atoms takes place in the pH range 5-7 and shows as well overlapping processes; the average pK value is 6.1, which is the range of the imidazole pK values reported in literature. Far UV-CD spectra of TetraHexaPY peptide, carried out at different pH values, are reported in Figure 4.16. At pH 4 the spectrum is broad with a minimum at 203 nm, a weak shoulder around 216 nm and a maximum at 230 nm. Generally, this shape indicates an equilibrium between different conformations, suggesting the presence of other secondary structure elements besides the random coil. At this value of pH, all histidyl residues are protonated and the spectrum shape is similar to that found for shorter peptide fragments (MonoHexaPY and BisHexaPY) as previously reported [58]. For those fragments, the presence of both random coil and β-turn structures was suggested [55], coherently with the primary sequence that encompasses a PXXP motif, generally supposed to favor β-turn and/or polyproline II structure [94]. However, while an intensity enhancement of the minimum at 200 nm and the disappearing of the maximum at 230 nm was observed on increasing the pH for mono and bis-hexarepeats, a different trend is observed here (Figure 4.16) for the four-tandem repeat. In fact, up to pH 7 any variation in the band centered at 230 nm was not observed, while there is a shift of the minimum towards 200 nm, together with a decrease of the band intensity. At basic pH values a general broadening of the spectra appear, with a significant decrease of the signal at 230 nm. In addition, at pH 10 a maximum at 250 nm is observed, that can be easily attributed to the deprotonation of tyrosine residues. The corresponding decrease and then disappearance of the maximum at 230 nm at basic pH allows to relate this latter band to a positive signal of the phenolic group of tyrosine residues, fading out with their progressive deprotonation. Actually it is well known that aromatic side-chains can give rise to a contribute to the far UV-CD spectra of peptides and that this is red-shifted for the phenolate ion with respect to the phenol [95, 96] (in this case, from 230 to 250 nm). Besides the evidences of deprotonation of specific residues, secondary structure variations can also be identified

as pH increases: the strong positive band at 190 nm (Figure 4.16) and the shoulder found at 216 nm are features typical of β turn like conformations [76, 79, 80], which appear to be predominant at neutral and basic pH. Concerning the mono and bis-hexarepeat, the folded states were destabilized at basic pH values, although, similarly, the structure adopted at physiological pH was mainly the type I β turn. On the whole, the trend observed by increasing the pH indicates that a β-turn conformation prevails at basic pH and that the conformational equilibria are strongly dependent on protonation steps of histidines and tyrosines.

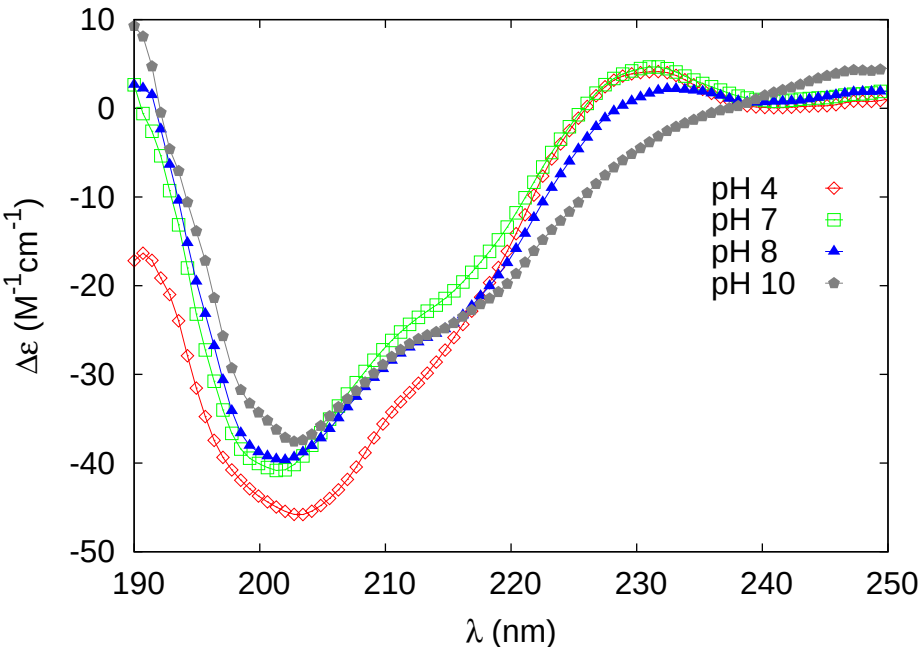


Figure 4.16: CD spectra of TetraHexaPY (4 x 10^{-6} mol dm^{-3}) as a function of pH. It is possible to observe the intense and broad negative peak at acidic pH, which instead becomes less intense increasing the pH. The band at around 190 nm, typical of type I β-turn structures, turns from negative (acidic pH) to zero (neutral pH) to highly positive (basic pH) values. Also the shoulder at around 216 nm, typical of type I β turn region becomes evident with the increasing of pH.

Molecular Dynamics simulations

As above mentioned, NMR studies carried out on the entire protein did not provide any particularly relevant data on the conformation of tandem hexarepeats. Therefore, in order to better understand the conformational equilibria involved in solution suggested from CD measurements, and to further rationalize the effects of pH, a fully atomistic molecular dynamics study was carried out in water. After verifying the effective equilibration of the peptide backbone as already described, the different conformational states of TetraHexaPY were analyzed on the production trajectories. In order to identify the contacts occurring inside the peptide structure, the average contact map concerning the Cα-C$_\alpha$ interatomic distances was calculated between two non consecutive amino acids, using a cutoff window of 8.5 Å [97]. Looking at the contact maps, reported in Figure 4.17, end-to end contributes were found at acidic and mainly at neutral pH, thus attesting the presence of a bent structure in the LH_4 form and moreover, long range contacts are present in the left region of the map concerning LH_4 and $LH_8{}^{4+}$ species. To better characterize the different conformations adopted by TetraHexaPY emerging from the contact map, and to probe the role of specific amino acids, it was subsequently analyzed the local secondary structure during the simulation using the DSSP algorithm [71]. The resulting time evolution of TetraHexaPY conformers is shown in Figure 4.18. i-i+3 hydrogen bonds were prevalently found, as in the mono repeat sequence [55], with few i-i+4 hydrogen bonds in the 16-20 and 11-15 regions (Figure 4.18) only for the LH_4 and L^{4-} state. In particular, at acidic pH two consecutive i-i+3 hydrogen bonds are present in the GYPHN sequence (5-8 and the 6-9, Figure 4.18) characteristic of a 3_{10} helix [71], which involves residues 6-8 (Figure 4.19). Two single i-i+3 hydrogen bonds were also found in the 9-12 (NPGY) and 21-24 (NPGY) regions (Figure 4.18), corresponding to a type I β turn structure for residues 10-11 and 22-23. Significantly, the number of residues involved in turn conformation increases with histidines deprotonation (see Tables 4.21, 4.22, 4.23). This is in good agreement with the formation of the shoulder at around 216 nm and the positive values approximately at 190 nm, observed in the CD spectra on increasing the values of pH. At neutral pH, the β-turn conformation is basically driven by the interaction between the deprotonated imidazole of histidine 8 and the phenol hydrogen of tyrosine 18, causing a strong tilting of the peptide backbone (Figure 4.19), disrupting the 3_{10} helix structure in the 6-8 YPH region, found at acidic pH. Such β-turn conformation leads to the formation of a family of conformers in which

a new 3_{10} helix is stabilized in the 17-20 GYPH region, carrying tyrosine 24 inside the backbone. Consequently, the phenolic hydroxyl groups of Tyr6, Tyr18 and Tyr24 get close, as shown in Figure 4.19. At basic pH, the deprotonation of tyrosine 24 determines another weak bond with the amide side chain hydrogens of asparagine 21 (Figure 4.19), previously found also in the MonoHexaPY at the same pH [55]. Consequently, glycine 17 and hystidine 20 stay much closer, thus making possible the formation of turns in the 17-20 GYPH regions, as underlined from Figure 4.18 for residues 18, 19 and 20 of L^{4-}. Here, the deprotonated Tyrosine 18 is stable in a turn conformation if compared to the LH_4 and $LH_8{}^{4+}$ states, in which the protonated Tyrosines 18 spend the initial time in a bend, interconverting later to turn and a 3_{10} helix structure.

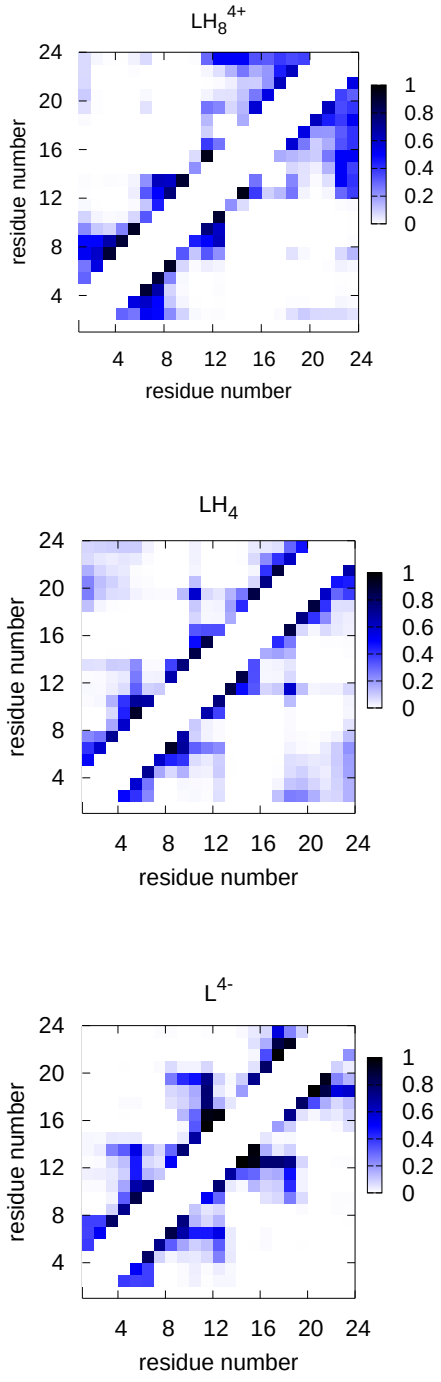


Figure 4.17: Contact maps of TetraHexaPY as a function of pH. It is worth to note a more compact structure at neutral pH (LH_4), underlined from the N- and C-terminal contacts and that the L^{4-} state shows the most expanded structure. The relative occurrence of the contacts is shown with a color code ranging from white to black.

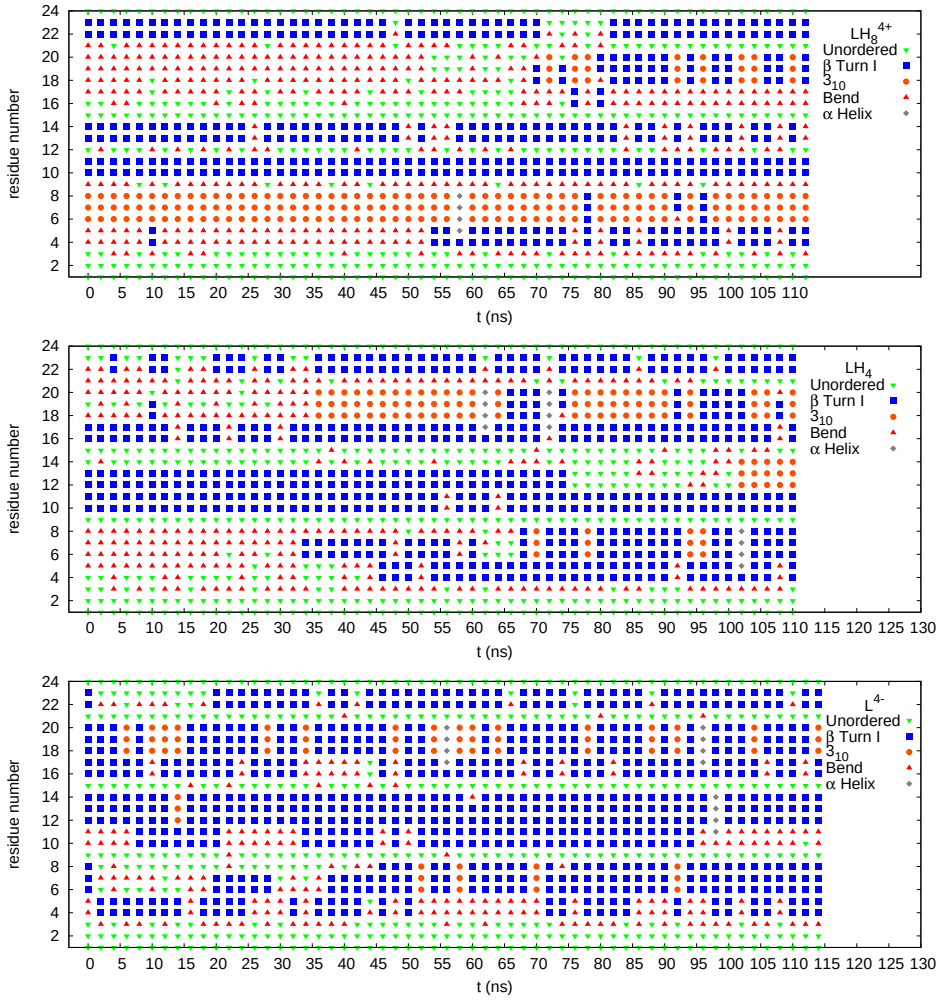


Figure 4.18: Time evolution of TetraHexaPY conformers in the three different protonation states obtained from MD simulations according to the DSSP criteria [71]. Turn regions increase with the histidine and tyrosine deprotonation, while the 3_{10} helix in the 6-8 region is present only in the $LH_8{}^{4+}$ protonation state.

Table 4.21: Percentage of the different conformations (Turn, 3_{10} helix, α Helix, Bend and Coil) adopted by the TetraHexaPY at acidic pH.

Residues	**Turn**	**3_{10} Helix**	**α Helix**	**Bend**	**Coil**
Pro1	-	-	-	-	100
His2	-	-	-	-	100
Asn3	-	-	-	47.4	52.6
Pro4	45.6	-	-	54.4	-
Gly5	43.9	-	1.7	54.4	-
Tyr6	3.5	93	1.7	1.8	-
Pro7	5.3	93	1.7	-	-
His8	5.3	93	1.7	-	-
Asn9	-	-	-	86.0	14.0
Pro10	100	-	-	-	-
Gly11	100	-	-	-	-
Tyr12	-	-	-	38.6	61.4
Pro13	82.5	-	-	17.5	-
His14	82.5	-	-	17.5	-
Asn15	-	-	-	-	100
Pro16	3.5	-	-	45.6	50.9
Gly17	3.5	-	-	79.0	17.5
Tyr18	22.8	14.0	-	56.2	7.0
Pro19	24.6	14.0	-	52.6	8.8
His20	21.0	14.0	-	52.6	12.4
Asn21	-	-	-	54.4	45.6
Pro22	89.5	-	-	7.0	3.5
Gly23	89.5	-	-	-	10.5
Tyr24	-	-	-	-	100
Total	30.2	13.4	-	27.7	28.7

Table 4.22: Percentage of the different conformations (Turn, 3_{10} helix, α Helix, Bend and Coil) adopted by the TetraHexaPY at neutral pH.

Residues	**Turn**	**3_{10} Helix**	**α Helix**	**Bend**	**Coil**
Pro1	-	-	-	-	100
His2	-	-	-	-	100
Asn3	-	-	-	73.2	26.8
Pro4	53.6	-	-	19.6	26.8
Gly5	51.8	-	1.8	39.3	7.1
Tyr6	51.8	7.1	1.8	30.4	8.9
Pro7	51.8	7.1	1.8	35.7	3.6
His8	30.4	7.1	1.8	57.1	3.6
Asn9	-	-	-	-	100
Pro10	96.4	-	-	3.6	-
Gly11	96.4	-	-	3.6	-
Tyr12	67.9	8.9	-	3.6	19.6
Pro13	67.9	8.9	-	7.1	16.1
His14	-	8.9	-	23.2	67.9
Asn15	-	-	-	12.5	87.5
Pro16	91.1	-	-	8.9	-
Gly17	89.3	-	3.6	7.1	-
Tyr18	17.9	46.4	3.6	32.1	-
Pro19	19.6	46.4	3.6	14.3	16.1
His20	16.1	46.4	3.6	30.3	3.6
Asn21	-	-	-	73.2	26.8
Pro22	75.0	-	-	23.2	1.8
Gly23	75.0	-	-	-	25
Tyr24	-	-	-	-	100
Total	39.7	7.8	0.9	20.7	30.9

Table 4.23: Percentage of the different conformations (Turn, 3_{10} helix, α Helix, Bend and Coil) adopted by the TetraHexaPY at basic pH.

Residues	**Turn**	**3_{10} Helix**	**α Helix**	**Bend**	**Coil**
Pro1	-	-	-	-	100
His2	-	-	-	-	100
Asn3	-	-	-	55.2	44.8
Pro4	55.2	-	-	44.8	-
Gly5	55.2	-	-	43.1	1.7
Tyr6	70.7	6.9	-	13.8	8.6
Pro7	70.7	6.9	-	17.2	5.2
His8	53.5	6.9	-	10.3	29.3
Asn9	-	-	-	3.4	96.6
Pro10	62.1	-	-	36.2	1.7
Gly11	62.1	-	1.7	36.2	-
Tyr12	96.6	1.7	1.7	-	-
Pro13	96.6	1.7	1.7	-	-
His14	94.9	1.7	1.7	1.7	-
Asn15	-	-	-	8.6	91.4
Pro16	75.9	-	-	22.4	1.7
Gly17	72.5	-	3.4	22.4	1.7
Tyr18	69.0	26.6	3.4	-	-
Pro19	69.0	26.6	3.4	-	-
His20	69.0	26.6	3.4	-	-
Asn21	-	-	-	5.2	94.8
Pro22	74.1	-	-	19.0	6.9
Gly23	74.1	-	-	-	25.9
Tyr24	-	-	-	-	100
Total	50.9	4.5	0.9	14.1	29.6

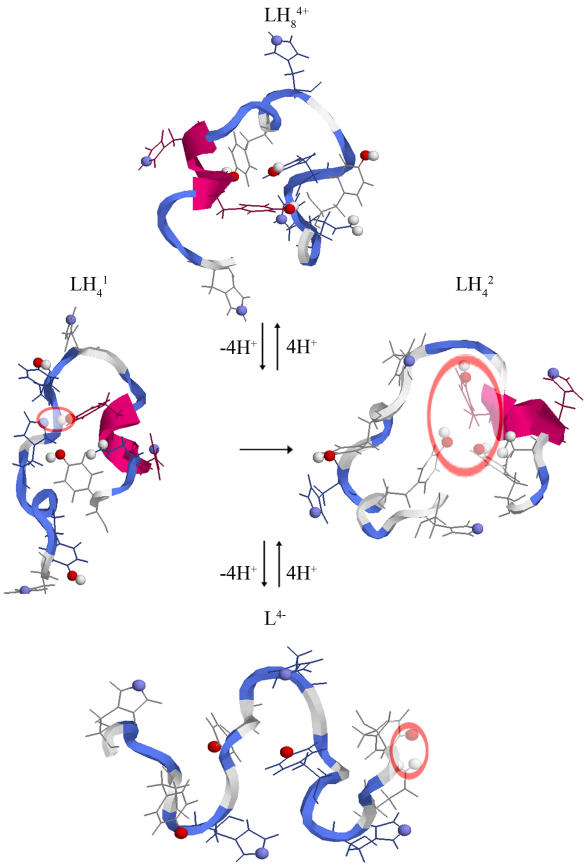


Figure 4.19: TetraHexaPY typical conformations in the $LH_8{}^{4+}$, in the LH_4 and in the L^{4-} form obtained from MD simulations. It is worth to note the 3_{10} helix, in the 6-8 YPH region, featured by two consecutive hydrogen bonds (5-8, 6-9, up $LH_8{}^{4+}$ state). Upon the deprotonation of the four histidines, the 6-9 hydrogen bond, inside the 3_{10} helix, is disrupted because of the interaction between the imidazole nitrogen of histidine 8 and the phenol hydrogen of tyrosine 18 ($LH_4{}^1$). Such a conformation leads to the formation of another one in which the 3_{10} helix, in the 17-20 region, causes the involvement of tyrosines 6, 18 and 24 in a hydrogen bond network, as shown in $LH_4{}^2$. At basic pH, when the four tyrosines are deprotonated (L^{4-}, bottom), a new interaction between the phenolate oxygen of tyrosine 24 and the side chain amide hydrogens of asparagine 21, causes a bending which stabilizes a turn structure in the 17-20 GYPH region and at the same time it provokes a tilting of the backbone. The different structures are shown according to a color code: blue for turn, violet for 3_{10} helix and gray for coil regions. The phenolate hydrogen and oxygen of tyrosines 6, 12, 18 and 24 are shown respectively in white and red, the imidazole nitrogens of histidines 2, 8, 14, 20 are shown in silver blue and the amide hydrogens of asparagine 21 are shown in white. The side chains hydrogen bonds are circled in red. N and C termini are shown respectively from left to right.

Chirality analysis

As a complementary test of the secondary structure, the conformations of TetraHexaPY was investigated in terms of their chirality, using a methodology recently proposed [98], largely described in chapter 2. The method consists in dividing up the entire backbone sequence in a number of fragments and assigning to each one a chirality index calculated from their geometry.

In chapter 2 it is shown that the chirality index allows assigning the motif type to the fragment, (see Table 4.24) complementing the DSSP classification. In Figure 4.20 the pattern of G along the TetraHexaPY (N, C_α, C) backbone atoms, with N_A equal to 15, corresponding to a window of five residues at once involved in the calculation. From the trend of the chirality index, averaged among the trajectories, the structures previous found using DSSP analysis can be recognized. For all the protonation states, the C-terminal region shows a pattern with chirality index approaching zero, typical of coil structures. At acidic pH the broad negative peak centered at residue 7 (Pro), involving the 6-8 YPH region, confirms the presence of a 3_{10} helix structure [98] while the turn region centered approximately at residue 11 (Gly), namely the 9-12 NPGY region, is characterized by a negative sharp peak. At neutral pH a turn region is found in the 4-6 PGY region, although in weak extent, as indicated by the wide standard deviations and the value of the negative peak, close to the higher values of the typical range for this motif [-0.1:-0.06]. Two other better characterized turn regions, signaled by negative peaks, are centered on residue 11 and 17 (Gly). At basic pH the negative peaks found at neutral pH show lower values of the standard deviations, underlining an enhancement of turn regions inside the peptide. Moreover, looking at the chirality pattern along the backbone, it is possible to notice that the regions which strongly differ in the chirality index values concerning the $LH_8{}^{4+}$ state, on one hand, and LH_4, L^{4-}, on the other one, are the 4-8 PGYP and the C-terminal region, in which His8 and Tyr24 are mainly involved, thus singling out the pivotal region for the conformational change. The strong pH dependence of the local chirality evolution as a function of the simulation time was also examined. As an example, in Figure 4.21 the time evolution of G for Tyrosine18 is reported. Here, the chirality index shows negative values, consistent with the presence of turn regions in the LH_4 and L^{4-} states and of a 3_{10} helix after 35 ns in LH_4. This is shown by lower values with respect to the L^{4-} form, in which the 3_{10} helix is less stable and thus the number of residues with the negative chirality,

typical of this secondary structure, are fewer (see Table 4.24) than the LH_4 neutral state. Looking at the $LH_8{}^{4+}$ protonated state, it is possible to observe instead that, the chirality index is approaching zero, as it would be expected for a coil region, till approximately 75 ns, thus demonstrating the strong pH dependence of the conformational states of this molecular system. In summary, the chirality analysis is consistent with the time evolution of the TetraHexaPY conformers reported in Figure 4.18, and highlights a significant contribution of turn structures progressively increasing with pH, expressed here by a shift towards more negative G values.

Table 4.24: Average G values and relative standard deviations of G for ideal secondary structures, involving at least N_R residues. Each structure was built by sampling ϕ and ψ angles from a gaussian distribution, centered on the ideal ϕ and ψ values with sigma=15 degree (see reference [46]).

Structure	$\langle G \rangle$	σ_G	N_R
α helix	-.04	0.02	>3
3_{10} helix	-.07	0.01	> 3
β Turn I	-.07	0.01	2,3
β Sheets	+.00	0.01	≥ 2
PPII	+.10	0.03	>3
π helix	-.01	0.02	>3

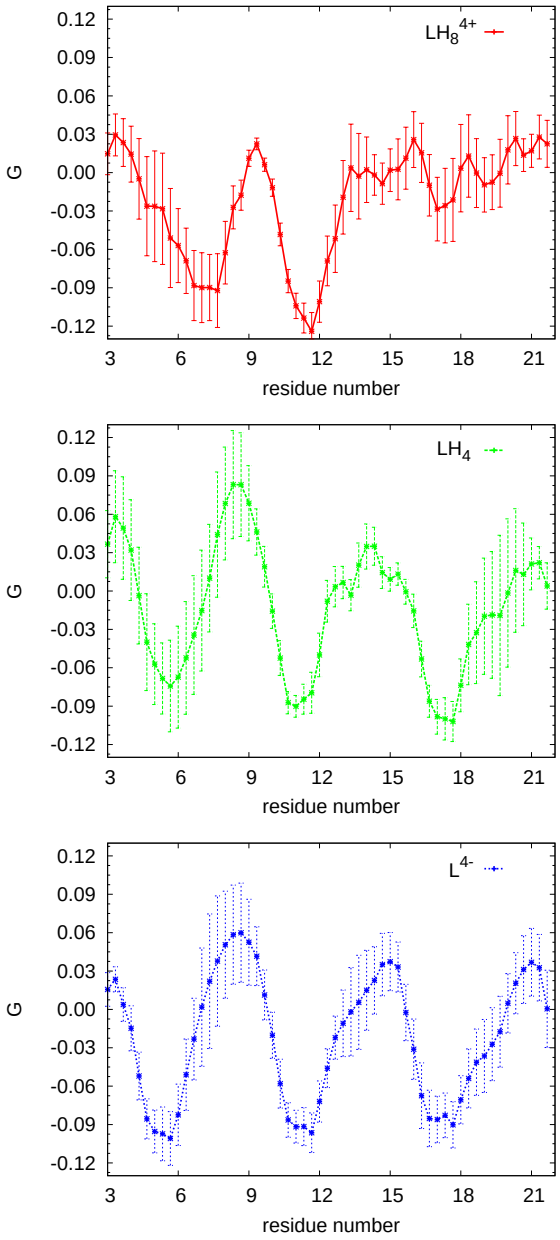


Figure 4.20: Chirality index, G, averaged over the trajectories of TetraHexaPY as a function of pH. The more negative peaks (-0.1< G <-0.06) underline the presence of turn regions, while the positive peak, centered in the asparagine 9 residue, underlines a small amount of polyproline structure in the 7-9 PHN region. The index shows a marked difference concerning the number of turn regions in the $LH_8{}^{4+}$ state with respect to the LH_4 and the L^{4-} ones. Error bars are reported for each of the G values as standard deviations on the ensemble of trajectories.

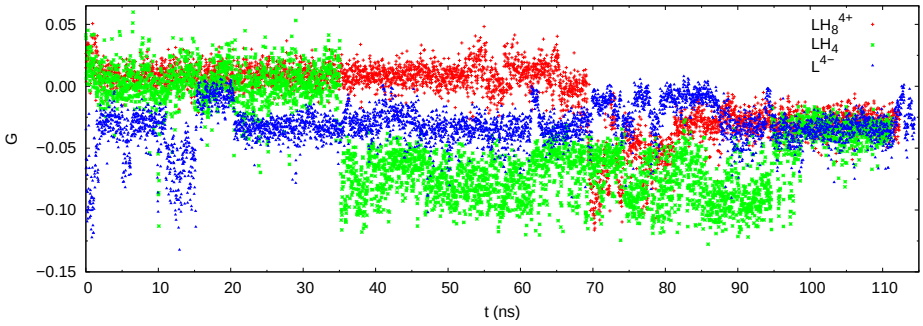


Figure 4.21: Time evolution of the chirality index G for Tyrosine 18. It should be noted the strong dependence on pH of the local chirality for tyrosine 18, pointed out by different patterns in the three different protonation states, $LH_8{}^{4+}$, LH_4, L^{4-}.

4.6.6 Brief summary on the tetra-hexarepeat section

The presence of turns inside the hexarepeat peptide fragment was suggested in the previous section concerning the mono hexarepeat Ac-PHNPGY-NH2 [55] and in other different studies [87, 89] as well. In this section, it is shown that the longer hexarepeat fragment $(PHNPGY)_4$ is essentially turn rich and that the turn formation is driven by the increase of pH. This trend is consistently disclosed by the analysis of the pH dependent CD spectra and by MD simulations at different pH, where C_α contacts, H-bond patterns and local chirality indices were monitored. The high number of prolines of course stabilizes a turn structure, but here it is shown that a crucial role is played by the histidine residues, particularly histidine 8 which, if deprotonated, is able to stabilize turn regions in the peptide backbone. The CD and MD results for TetraHexaPY at different pH values also stress the essential role played by tyrosine OH group in the peptide secondary structure. It is worth noting that tyrosine residues play a regulatory role in the endocytosis process [89] and are present in avian tandem hexarepeats (16% like glycine residues), but not in mammalian octarepeats. Moreover, the chirality pattern of the tetra hexarepeat region, in particular for the LH_4 and L^{4-} states, possesses a periodic-like shape, thus reflecting the periodicity in the primary structure; it is therefore likely that a similar pattern could be adopted also by the full repeat region. Within this hypothesis, it is possible to foresee a different biological behavior of the N-terminal domain of avian and mammal prion protein: the presence of tyrosines in the avian protein allows forming a compact hydrogen bond network, which could be probably responsible for its high resistance to proteases. This peculiar feature is largely explained in the following section.

4.7 A glimpse of the full avian prion protein structure: exploring the flexibility and rigidity inside the protein domains

4.7.1 Introduction to the reading of the section

The sections discussed till here concern the study of protein fragments, namely selected regions of the prion protein, in this particular case the hexa-repeat region, whose im-

portance was largely explained in section 4.4. Handling peptide fragments is easier for experimental measures, i.e. synthesis, potentiometric titrations, the study of coordination properties with metal ions and of course this approach permits to have a simple model which represents a specific site of the protein. However, these simplified systems could be somewhat different, depending on the presence of non-covalent interactions inside the protein, which may influence the region under study.

Although a number of Molecular Dynamics (MD) studies were carried out on PrP, [101–104], such investigations were concerned only with the study of the globular core structure, excluding the N- and C-terminal regions. In this regard, an investigation of the full structure of the avian prion protein could give a deeper comprehension of the phenomena surrounding the globular core. Bearing this in mind, here it is shown the effect of the N- and C-terminal missing sites linked to the globular ChPrP128-242 NMR structure on the globular conformation and the conformational preferences of such regions inside the full avian prion protein, ChPrP1-267, by using MD simulations at physiological pH. This, as far as we know, is the first computational study in which the full ChPrP1-267 sequence is investigated. To complement the existing simulation analysis tools, a recently proposed protein chirality index was used [98], (see chapter 2 for details), which allows for an easy visualization of secondary structure patterns. From this index, it is possible to investigate how persistent is a secondary structure inside the protein backbone.

4.7.2 Simulation Details and Chirality calculation

Molecular dynamics of the whole chicken prion protein, ChPrP1-267, was carried out in water at neutral pH. In these conditions, the ionizable amino acids such as histidines (protonated at the δ nytrogen) and tyrosines are considered in the neutral state; glutammic and aspartic acids are negatively charged and lysine and arginine residues are positively charged. In addition, eight chloride ions were added to ensure charge neutrality in the simulation box. The simulation was run in water using the GROMACS 3.3 package [92], the OPLS-AA force field [105] for the protein, the SPC model for water [42], and Particle Mesh-Ewald (PME) for long range electrostatic interactions [106]. The starting configuration of ChPrP1-267 was built by linking the best representative NMR model of the globular core, pdb code 1U3M [37], to the N- and C-terminal regions, 1-127 and 243-267 respectively, both of them energy minimized before the linking. A

long run of about 130 ns was then performed for a cubic box containing the protein and 26560 water molecules with periodic boundary conditions (PBC), with 2 fs of time step, using the isothermal isobaric ensemble (NPT, P=1 atm, T=300 K), being the temperature and pressure controlled by a Berendsen thermostat and barostat [107]. After the first 20 ns of equilibration, the trajectory analysis was performed on a 113 ns production run, with configurations stored every 5 ps.

An important part of the simulation will be to analyze the conformational features of ChPrP and their evolution in time. In particular this will be pointed out in changes of motif. It was shown that through the evaluation of a local chirality index it is possible to assign the motif type to a fragment, (see Table 4.24) complementing the DSSP classification. In summary, the method consists in dividing up the entire backbone in fragments and computing for each of them a chirality index calculated from the backbone atoms coordinates; the instantaneous value of the index can then be compared with the characteristic values for ideal secondary structures (here reported in table 4.24). The chirality index was then calculated as shown in the previous section and in chapter 2. It shall apply later on the index G (see chapter 2) to the prion protein trajectories. However, another application used is the G_{score} quantity (see chapter 2, section 2.3.3), quite important when, as in our case, a structure for the protein to be studied is not available either from X-ray or NMR to provide a starting configuration, as shown in the following section.

4.7.3 The protein equilibration

Preliminarily, in order to assess the equilibration of the two chains linked to the globular core, the time evolution of the end-to-end distance was checked and it is reported in Figure 4.22. The end-to-end distance, after showing a fast decrease in the first few nanoseconds, remains substantially stable for the whole run. Accordingly, the initial 20 ns were considered as equilibration run and excluded them from the production analysis, in order to avoid a starting configuration bias. Moreover, after equilibration, volume, total energy and end-to-end distance were checked fluctuated around their average value without systematic drifts. In addition, the G_{score} value was checked during the MD simulation and as it can be appreciated from Figures 4.23 and 4.24, the protein trajectories show G_{score} values consistent with those one extracted from the NMR protein database. Furthermore, the istograms of G_{score} (Figure 4.24) show a good agreement with the

G_{score} of the NMR dataset, thus attesting a not unduly biased ϕ and ψ dihedrals of the two chains bound to the globular core. It is possible to notice that this is not necessarily the case, e.g. chirality indexes sampled randomly from a gaussian distribution centered in zero chirality index value ($\sigma = 0.3$), are shown by comparison. The results indicate that random chirality indexes, assuming these as unfolded structures, fall out the range of 0.08:0.18, adopted by the NMR protein database.

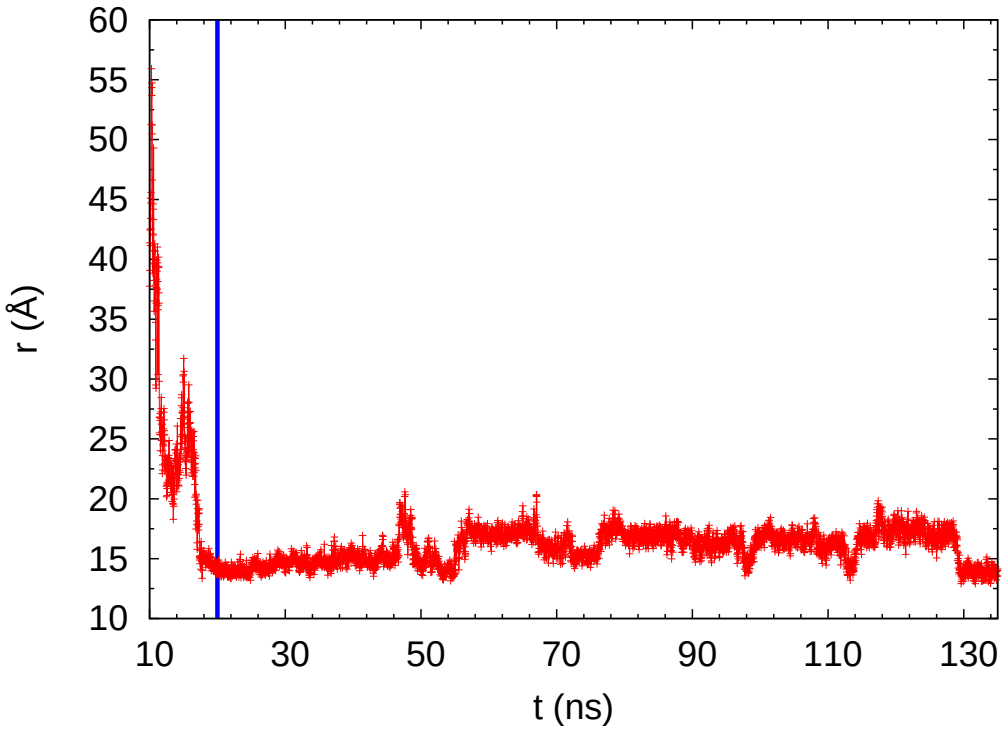


Figure 4.22: Time evolution of the end-to end distance for ChPrP1-267. The vertical blue line indicates the beginning of the production trajectory.

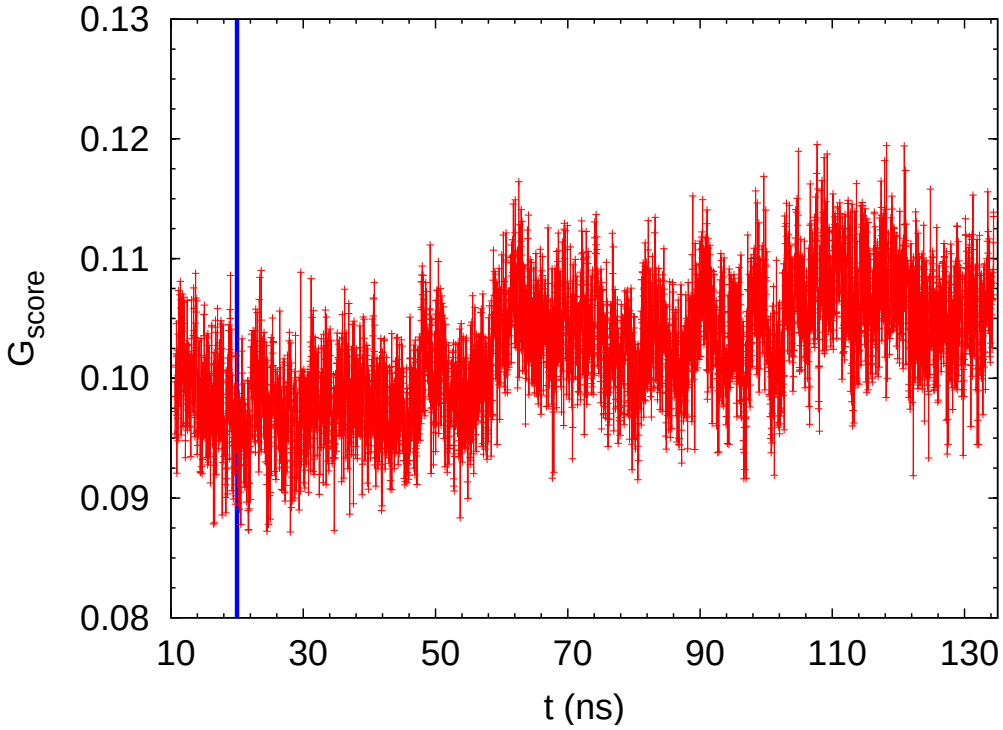


Figure 4.23: Time evolution of the G_{score} for ChPrP1-267. The G_{score} shows an average value of about 0.11, consistent with the values typical of the NMR dataset. The vertical blue line indicates the beginning of the production trajectory.

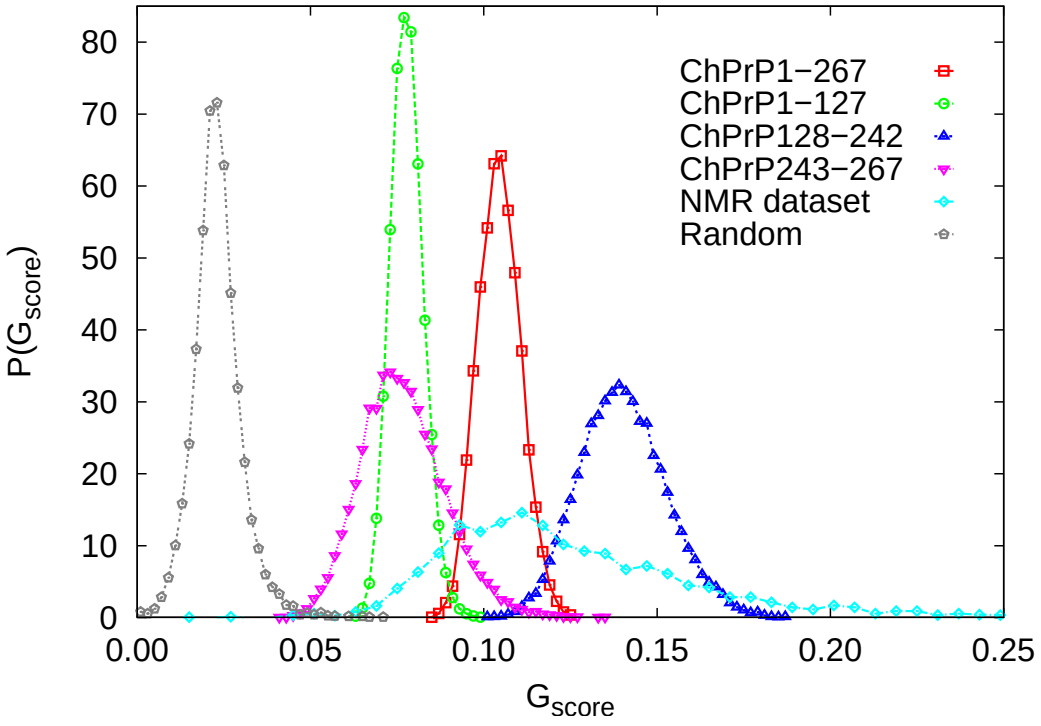
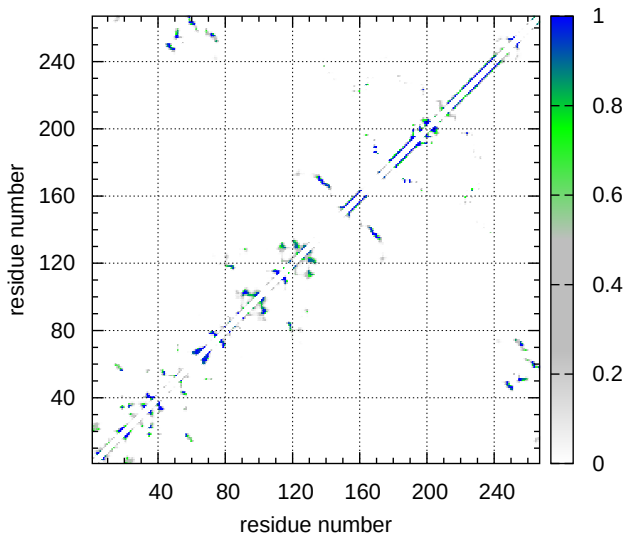


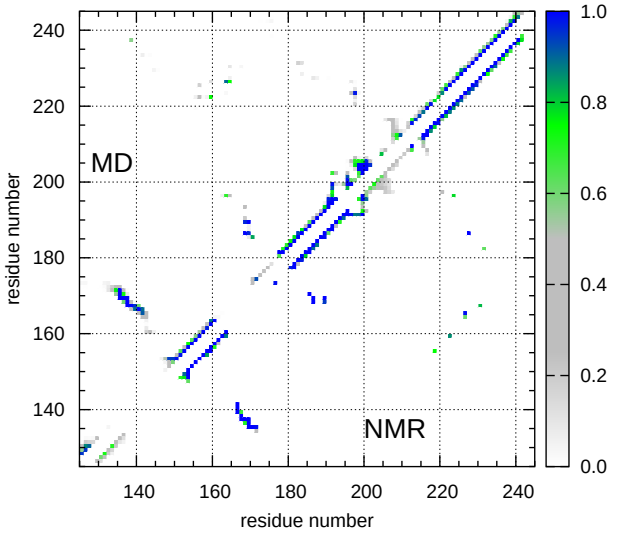
Figure 4.24: G_{score} calculated for ChPrP1-267, ChPrP1-127, ChPrP128-242, ChPrP243-267 and by sampling randomly chirality indexes from a gaussian distribution centered at zero chirality index value. It is worth noting a lower G_{score} value for the N- and C-terminal region and the higher G_{score} value for the globular core, starting from the NMR structure of reference [32] (pdb code 1U3M). The G_{score} of ChPrP1-267 (red curve) lays between those of the globular core and the N- and C-terminal regions, compatible with the cumulative distribution calculated from the NMR dataset (cyan curve).

4.7.4 The overall structure of the avian prion protein ChPrP1-267

After verifying the effective equilibration of the protein, its conformational states were analyzed along the production trajectories.
First, the average number of contacts between pairs of non consecutive amino acids were identified, evaluated through the C_α- C_α distances, using a cutoff of 8.5 Å [108]. In Figure 4.25 [a] the contact map are reported for the full protein; here all the secondary structures present in the globular core can be recognized: the spread in off-diagonal contacts between strands 132-144 and 166-173 underlines the presence of a short , β sheets, while the three α helices correspond to the contacts along the diagonal line. In addition, the map reveals diffuse contacts between N and C-terminal parts, in particular between residues 47-249, 51-255, 59-263 and 72-256, as suggested by the relatively short end-to-end distance in Figure 4.22. Such contacts appear after the first 20 ns of the simulation (equilibration) and last for the whole production time. Also in the N-terminal part some contacts appear, notably some 3_{10} helices and turns and a β bridge. Focusing on the globular core, in figure 4.25 [a] the comparison between its contact map as obtained by MD simulation and from the 20 NMR structures [32] is reported. Some difference appears: the 3_{10} helix inside the globular core, involving residues 197-199, is much more intense in MD with respect to the NMR structure, as well as the immediately following turn (around 203); on the contrary the small helix centered in 135 almost disappears. More interestingly, the contacts between residues involved in the β sheets region increase but also become more irregular if compared with the NMR structure, and a contact appears beween residues 139 and 237 (at one edge of the sheet and helix 3 respectively). Both the structural irregularity of the sheet and the presence of the latter contact (termed β bulge 1) have been deemed to be designed by nature for depressing edge-to-edge interprotein dimerization [109,110]; considering also that avian prion has a slightly shorter β sheet with respect to the mammal ones [32]. This suggests a relatively low tendency to amyloid aggregation.



[a]



[b]

Figure 4.25: [a] ChPrP1-267 contact map calculated from MD simulations. The contacts around residues 130-160 underline the presence of β sheets. Less spread contacts are found around residues 60-260, indicating two opposite non overlapping sequences. α helices are found along the diagonal line (approximately at 160, 200 and 240 residue number). The intense points near the diagonal line indicate the presence of turns and 3_{10} helices.
[b] Comparison from the contact map of the globular core (ChPrP128-242) as obtained from MD trajectories and from the NMR structures [32]. The contact map concerning the ChPrP128-242 globular core from MD simulations [b] reveals the stabilization of secondary structure elements with respect to the contact map from NMR structure of [b]. These involve the 3_{10} helix for 197-199 residues, between helix 2 and 3; an increment of the number of residues adopting the β sheet conformation is also shown from the wider contacts around 130-160 residues.

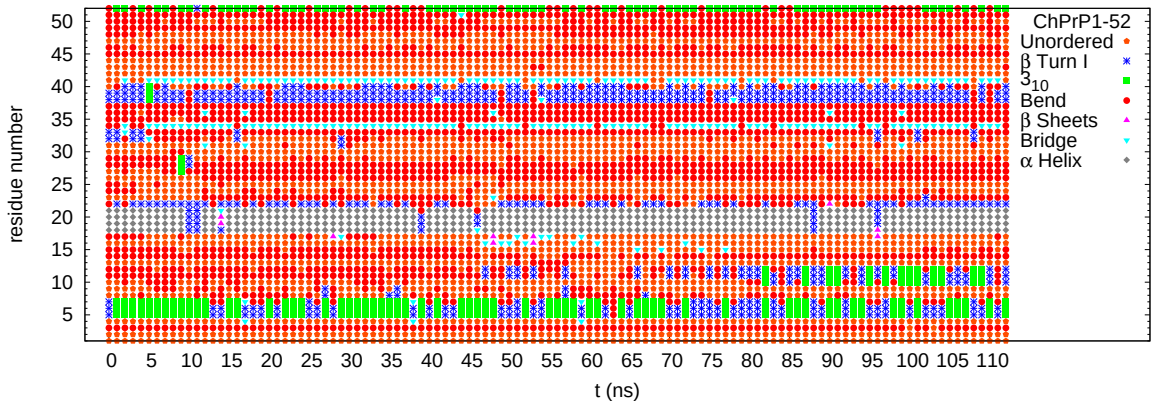


Figure 4.26: Time evolution of secondary structure in the 1-52 region according to DSSP criteria [5]. Unordered states, turns, 3_{10} helix (residues 5-7) and α helix (residues 18-21) are present. Isolated β bridges are found for residues 16-17,34,41.

In order to unambigously assign the secondary structure and to characterize its time dependence, it was monitored it by using the DSSP algorithm [5]. First we need to focuse our attention on the N-terminal hexarepeat region, ChPrP53-88, owing to its high flexibility and its likely biological function. As previously found by us for the single [55] and the tetra-hexarepeat [56], specific residues prefer to adopt the type I β turn and 3_{10} helix structures, in this case 53-54, 64-65 for turn and 67-69 for 3_{10} helix and, in lower extent, 83-84 and 86-88.

A short but persistent β sheet involving residues 136-137;169-170, was found. Interestingly, in the mouse prion protein the β sheet region was found, instead, to undergo disruption [111], revealing a role of the sheets in stabilizing the fold prion protein. In addition, time evolution was analyzed for the three α helices, the main feature of the prion protein. The second helix, namely ChPrP178-195, results to be the most rigid and preserved one. This is inferred from the comparison of the time evolution of the secondary structures of the helices, reported in Figures 4.28-4.30, where several residues of helix 1 and, more exstensively, the beginning of helix 3 experience frequent interconversion between α, β turn and 3_{10} helix, which is known to be very flexible. The C-terminal part, 243-267 is substantially unordered, except residues 251-254, adopting 3_{10} helix and turn structures, and 264-265 adopting turns and β bridge conformation.

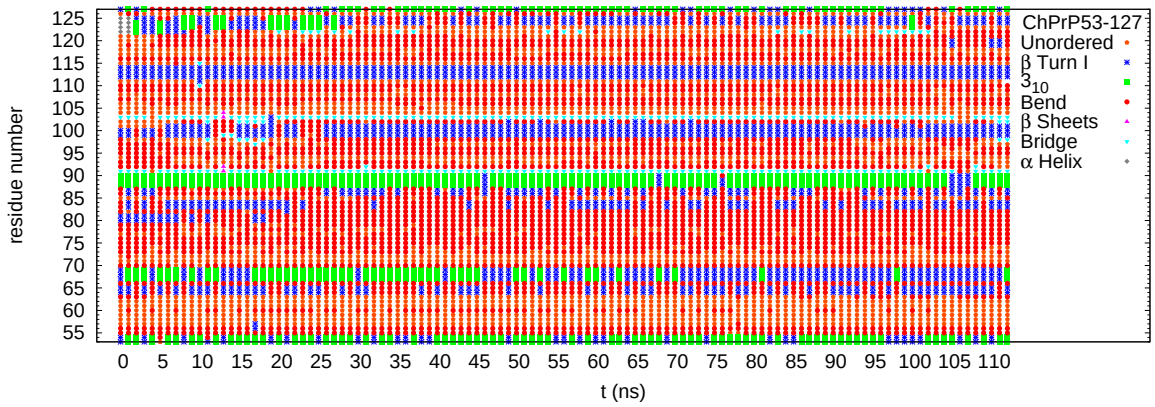


Figure 4.27: Time evolution of secondary structure in the 53-127 region, according to DSSP criteria [5]. In the hexarepeat region (ChPrP53-88) unordered states are present, together with turns and 3_{10} helices, which are abundant especially around residues 53-54 and 67-69, namely the second and the third repeat, and in less extent in the final one (residues 83-88).

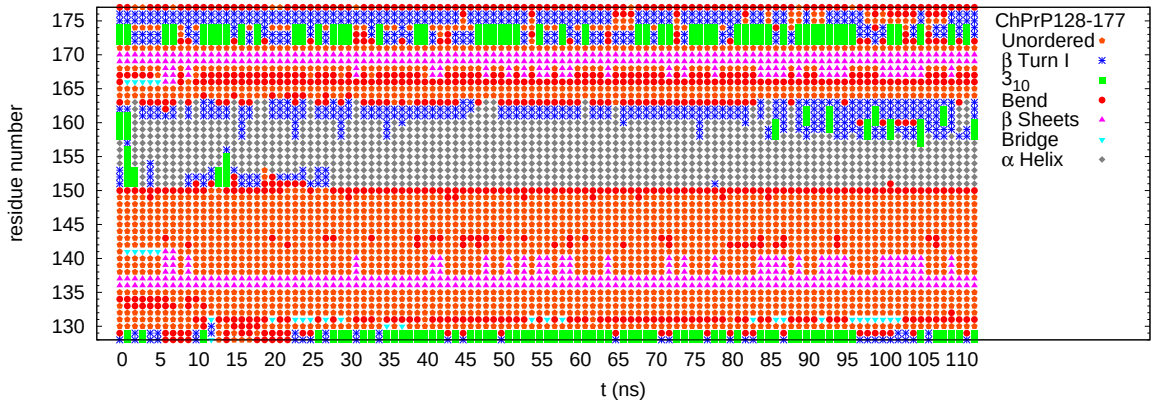


Figure 4.28: Time evolution of secondary structure in the 128-177 region according to DSSP criteria [5]. The helix 1 (ChPrP150-162) conformation is retained during MD simulations, except for the initial and the final portions of the helix. β sheets (136-137, 169-170), involve also residues 138-141 and 165-167. 3_{10} helix is found prevalently in residues 172-174.

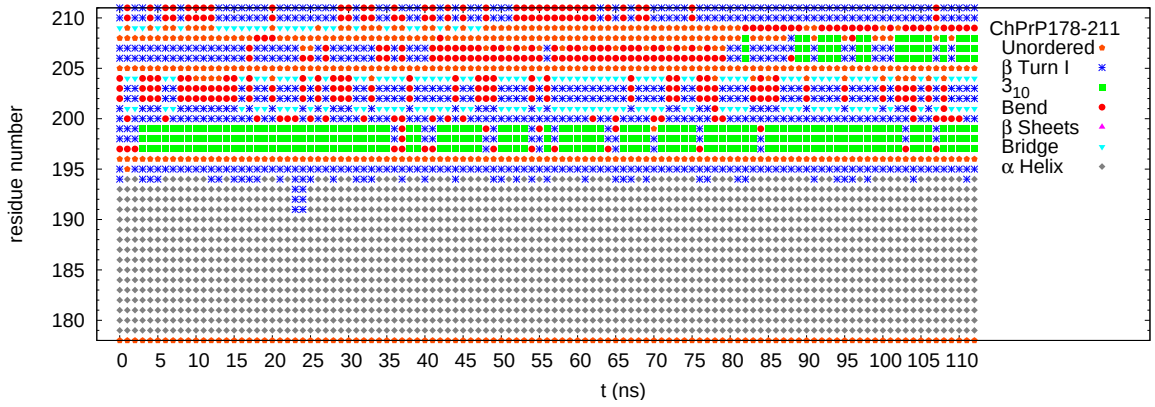


Figure 4.29: Time evolution of secondary structure in the 178-211 region, according to DSSP criteria [5]. It is worth noting the low flexibility of the helix 2 region (ChPrP178-195), that preserves its helical structure during the MD simulation. A 3_{10} helix is found for residues 197-199, while β bridges are detected for residues 201, 204 and in less extent for residue 209.

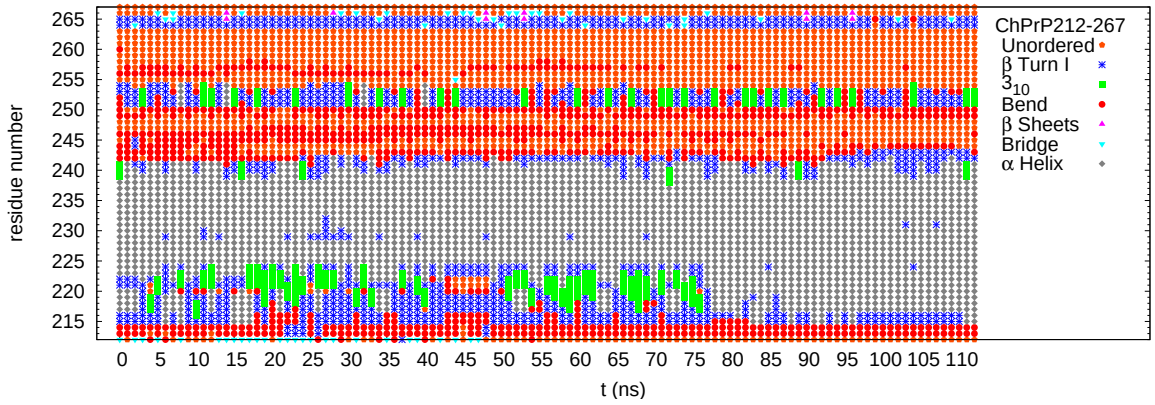


Figure 4.30: Time evolution of secondary structure in the 212-267 region according to DSSP criteria [5]. In the helix 3 region (ChPrP212-242), a high flexibility emerges. This helix spans also 3_{10} helices and turn regions inside its core.

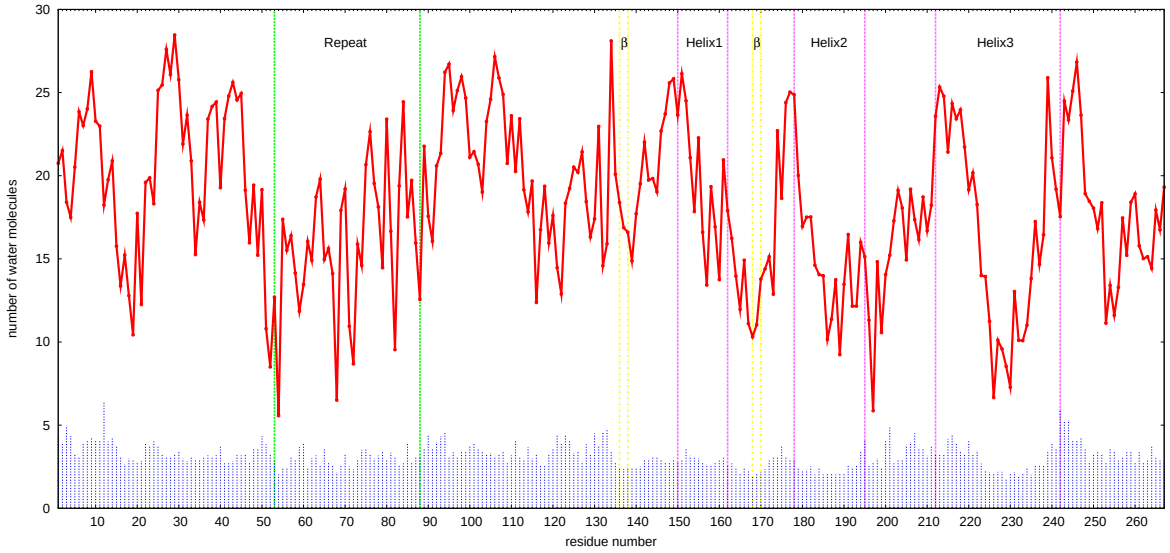


Figure 4.31: Number of water molecules along the backbone. The local minima are relative to regions with well defined secondary structures, i.e. the hexa-repeat (turn and 3_{10} helix), helix 1, helix 2 and helix 3 after residue 226, consistent with the chirality analysis. Istograms below each of the values are reported as standard deviations on the ensemble of trajectories.

The solvation of ChPrP was also investigated by counting the number of water molecules centers of mass inside a cylinder of radius of 12 Å, having as axis the nitrogen-carbonyl carbon distance vector of tne amino acid and as height its modulus. This very intuitive measure of solvation is affected by the size of the amino acid side chain and by the backbone dihedral angles which can reduce the height of the sampling cylinder. In figure 4.31, many regions of local minima are detected. These are centered in Valine 168, Valine 226, Arginine 228 and Glycine 197. Furtherly, Valine 168 was also found in a previous work [110] to be a tight water, as also here standard deviations reveal. In general, poorly hydrated regions with low standard deviations are found where the secondary structure is well defined, i.e. the repeat region (turn and 3_{10} helix), helix 1 and more extensively helix 2 and helix 3 after residue 226, consistent with that revealed from chirality analysis. Besides, all three helices present high solvation sites at their N-terminal, while the long connecting loop between helices 2 and 3 is scarcely accessible to water as indicated also in [110].

The average chirality index G was calculated for the three helices (Figures 4.32-4.34), revealing, in more detail, their flexibility, evidenced by the error bars, which indicate the secondary structure variation, higher for helix 3 (Figure 4.34). Furthermore, helix 1 presents typical α-helix chirality values (cf Table 4.24) after residue 152, coexisting with a 3_{10} (see also Table 4.24), helix 2 preserves its helix structure from 180 to 192 and finally helix 3 shows an α-helix structure from residue 226 till residue 238. Such interconversions are also apparent in Figure 4.35, in which it is possible to visualize the chirality index evolution for selected central residues (156, 186, 223). Here, the central residue of helix 3 shows an interconversion between 3_{10} helix and α helix after 60 ns, to a coil-bend structure (understandable from the lowering toward negative values of the chirality index) around 75-90 ns and finally turning again to an α helix.
In order to have an overview of the global conformation adopted by the avian prion protein in solution, the complete chirality pattern along the backbone, averaged on the trajectories, is also shown (Figure 4.36). As previously indicated from the DSSP assignment concerning the hexarepat region, the N-terminal part is rich in turns and 3_{10} helices; in addition positive peaks are present (Figure 4.36), pointing to a poly-L-proline II like structure (cf Table 4.24), involving typically no more than three residues, i.e. residues 113, 178. It is worth noting that these positive peaks often coincide with highly solvated amino acids in Figure 4.31.

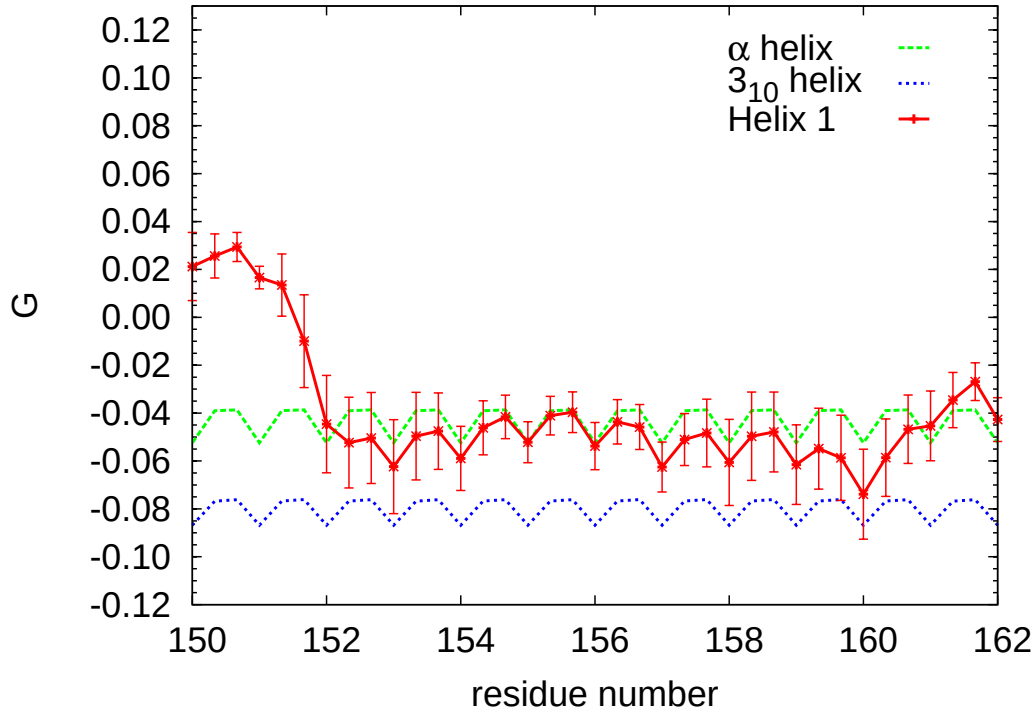


Figure 4.32: Average chirality index G of helix 1 (ChPrP150-162). The values indicate the coexistence with 3_{10} helix, understandable also from the standard deviations, spanning more negative G values than the α helix, see Table 4.24. Error bars are reported for each of the G values as standard deviations, referring to the ensemble of trajectories. Ideal α and 3_{10} helices patterns (dotted lines) are shown as comparison.

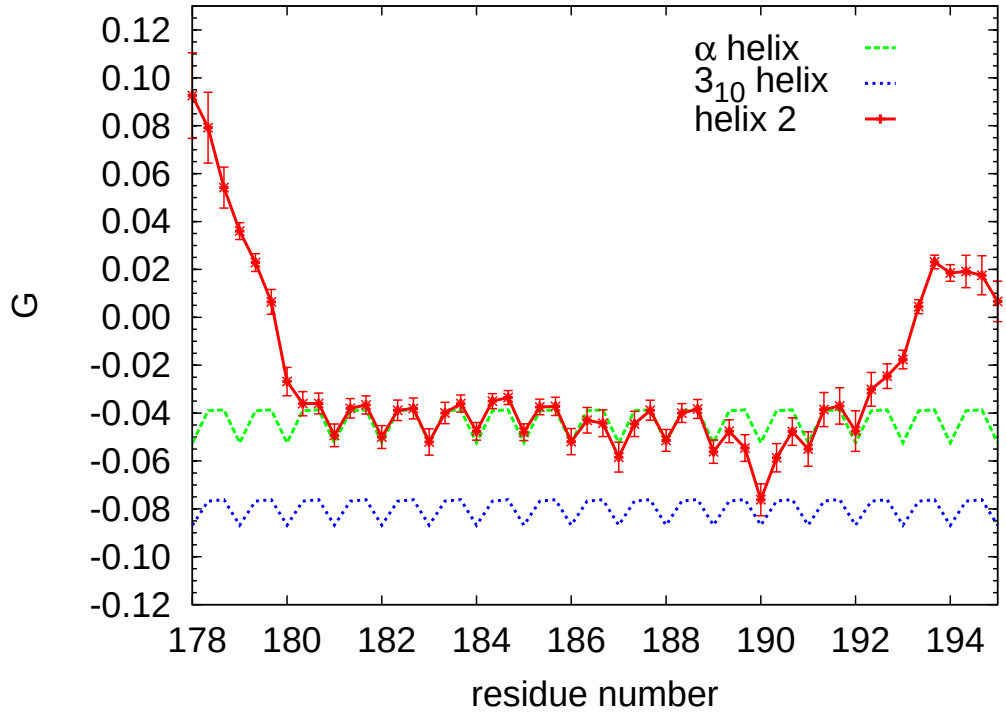


Figure 4.33: Average chirality index G of helix 2 (ChPrP178-195). As the standard deviations show, the structure is not really flexible, however the α helix structure stops at residue 192 and thus not involve residues 193-195. Error bars are reported for each of the G values, as standard deviations, on the ensemble of trajectories. Ideal α and 3_{10} helices patterns (dotted lines) are shown as comparison.

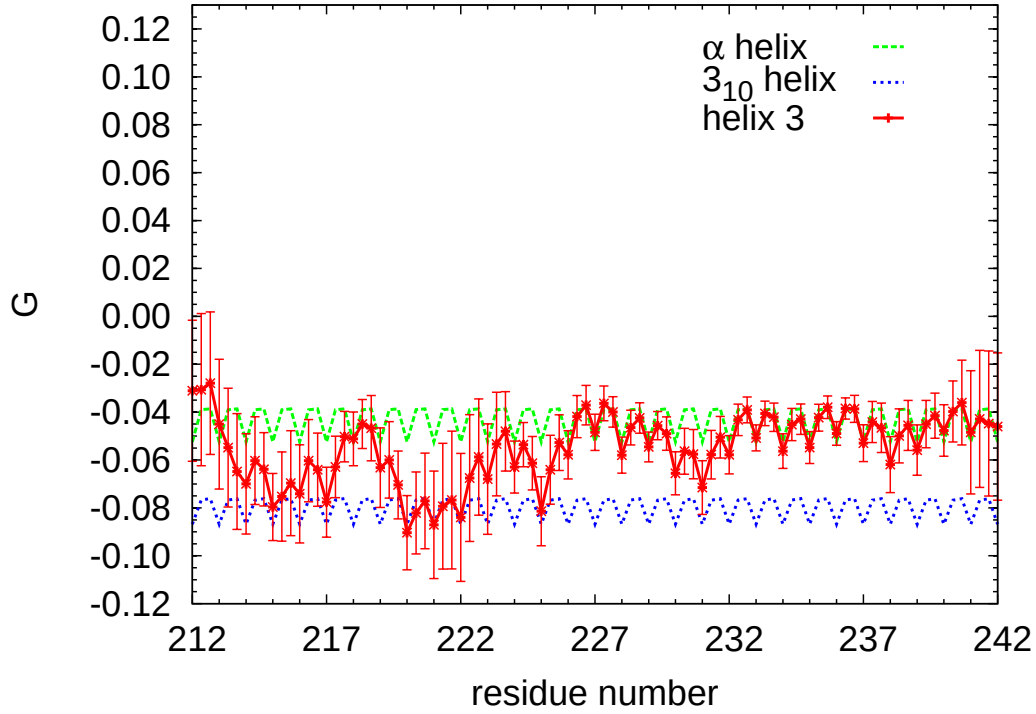


Figure 4.34: Average chirality index G of helix 3 (ChPrP212-242). Error bars are reported for each one of the G values, as standard deviations, on the ensemble of trajectories. Ideal α and 3_{10} helices patterns are shown as comparison. The large standard deviations show that this helix is very flexible, except the 227-240 region.

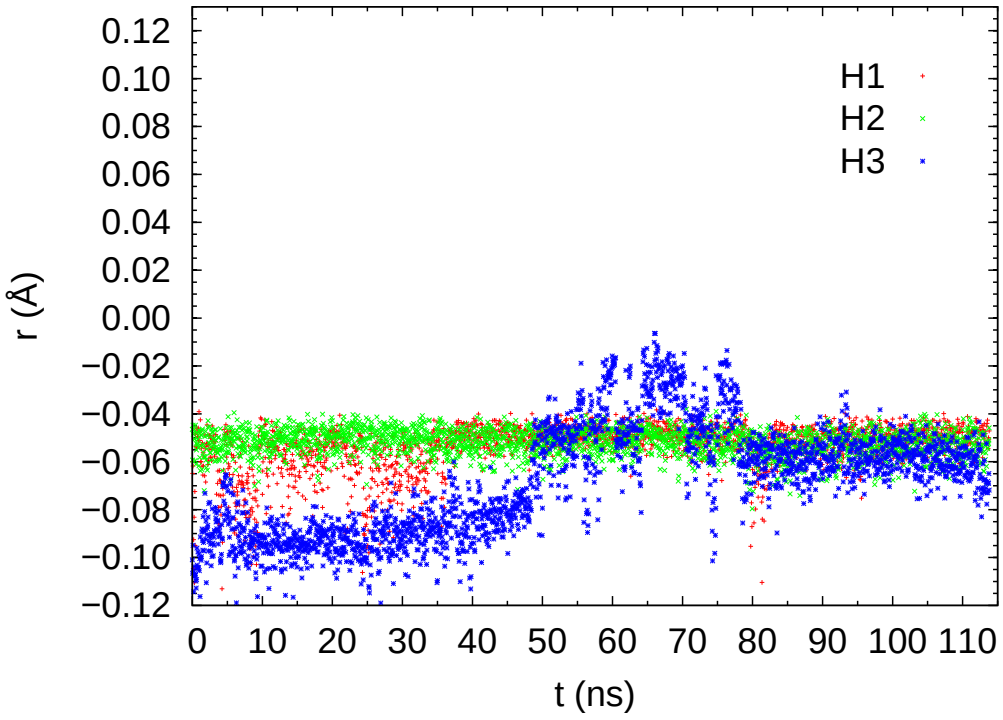


Figure 4.35: Time evolution of the chirality index, G, relative to the central residues of the three helices. Consistently with the average chirality index trend, the main residue of helix 3 is structured as a 3_{10} helix for the first 60 ns of the simulation, converting to an α helix between 60-75 ns, then to a bend-coil between 75-90 ns and finally turning again to an α helix. Helix 1 shows the presence of a 3_{10} helix in the first 40 ns, while helix 2 shows only α helix chirality index values.

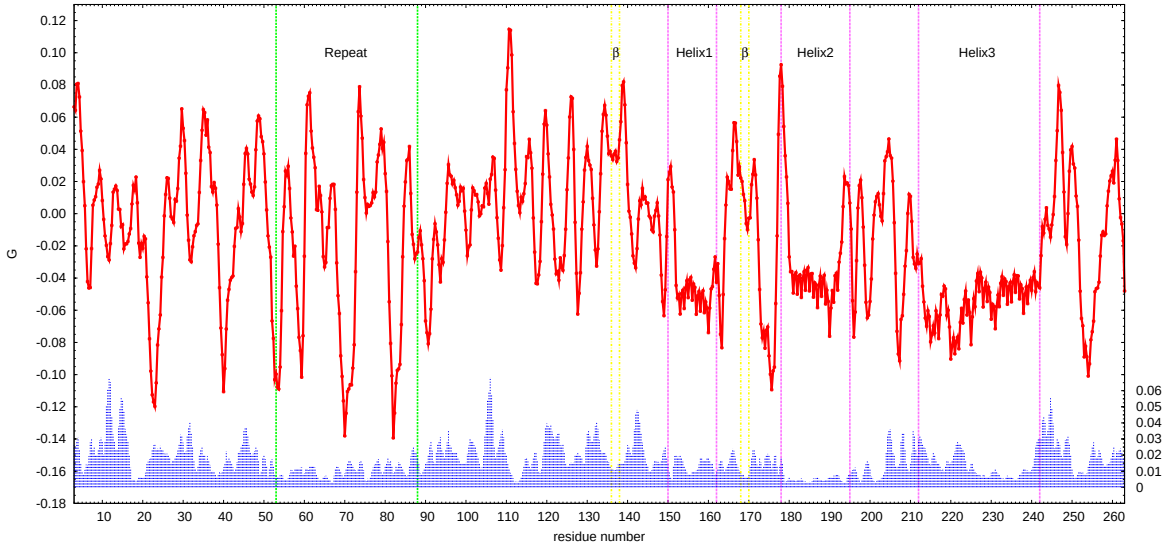


Figure 4.36: Chirality index, G, averaged among the trajectories of the ChPrP1-267. The more negative peaks (-0.1< G <-0.06) underline the presence of turn regions; the three negative oscillations centered at -0.05 G values underline the presence of α helix (main residues: 156, 186, 227). 3_{10} helices are recognized from the involvement of more than three residues having its typical chirality (See Table 4.24). The more evident central residues of such helices are: 23, 53, 70, 82, 175, 254. The high positive peaks, namely G values greater than 0.05, are typical of a polyproline II conformation, which is present, although it involves a little extent of residues, along the backbone. Istograms below each of the G values are reported as standard deviations on the ensemble of trajectories.

The standard deviations in the N- and C-terminal regions reach high values, owing to their flexibility, but it is interesting that in the hexa-repeat region, such deviations become smaller, comparable to those ones of helix 1 and helix 3. Helix 2, instead, is the most rigid of the three helices, as the standard deviations suggest. It is worth noting that ChPrP helix 2 possesses a proline residue at the beginning of the helix (residue 178), this is clearly shown by the average chirality index along the helix 2 residues of Figure 4.33, where the first value is strongly positive, typical of proline residue. Proline usually makes the structure stiff and it was reported to act as a fold protection preventing non-native interactions [112].

In HuPrP helix 2 the proline residue is replaced by an histidine residue in position 187, and this may support the high rigidity of ChPrP helix 2 with respect to the human prion helix 2, that it was reported to be rather flexible [113, 114]. Focusing again on the hexarepeat region, since in our simulation study of the tetrarepeat fragment a hydrogen bond was found between the imidazole nitrogen of the first histidine and the phenolic hydrogen of the third tyrosine, determining a loop conformation in this region [56], in Table 4.25 the calculated NOE contacts between the imidazole nitrogen and the phenolic hydrogen of selected histidine and tyrosine residues, are reported respectively. In the full avian prion protein, it is found again such an interaction, even if shifted of one repeat, as it involves the imidazole nitrogen of the fourth histidine and the phenolic hydrogen of the second tyrosine, as it can be appreciated from Figure 4.37 and from the calculated NOE distances, reported in Table 4.25. Finally, in Figure 4.38 the chirality index pattern averaged on the trajectories is reported for ChPrP53-88 and the averaged chirality index of the tetra-repeat fragment dynamics, previously reported [56]. The superimposition of the chirality pattern is surprisingly consistent with the periodicity of this system, that tends to adopt 3_{10} helix for residues 68-73 and 80-85 and a Type I β turn region around residue 59.

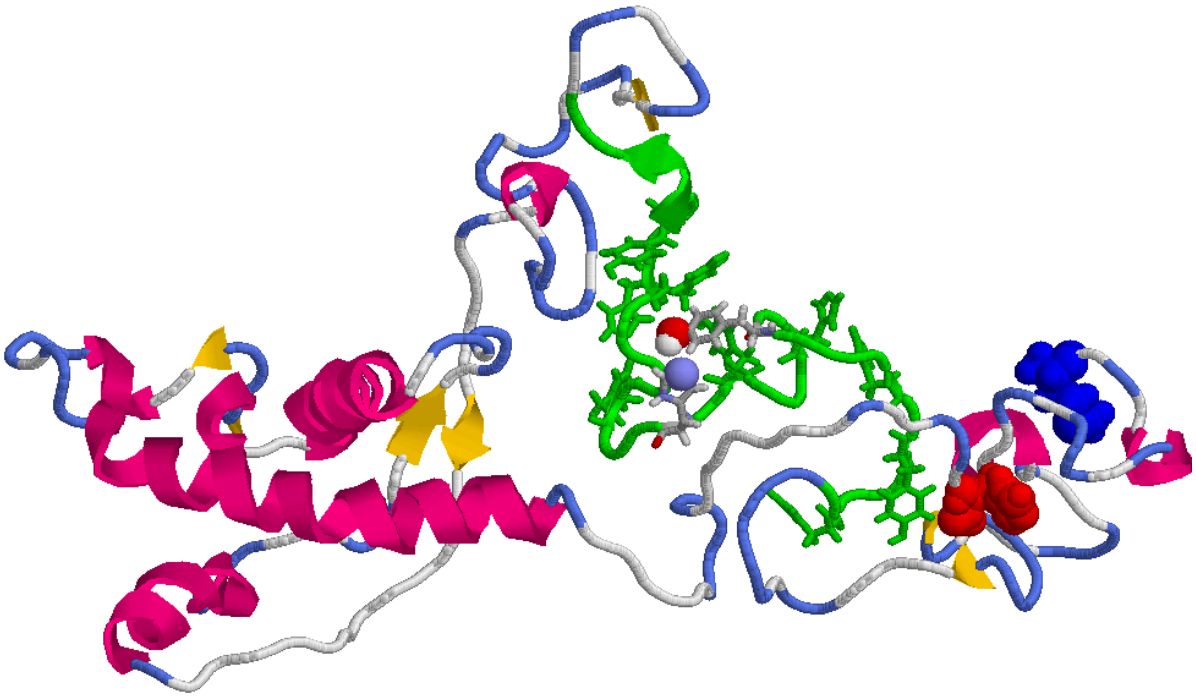


Figure 4.37: ChPrP1-267 typical conformation sampled from the MD simulations. The hexarepeat region, colored in green, shows the hydrogen bond between the imidazole nitrogen, N_ϵ, of the histidine 64 and the phenolic hydrogen of tyrosine 72. This interaction occurs also in the tetrarepeat fragment, previously reported [56].

Table 4.25: Tyr-His HH-NE2 and Tyr-Tyr HH-HH Noe contact distances, calculated as $\left\langle \frac{1}{r^6} \right\rangle^{-\frac{1}{6}}$, for tyrosine and histidine respectively. For sake of clarity the first, second and third order momenta, together with he skewness, of the $\frac{1}{r^6}$ distribution are reported.

Amino Acids	$\mathbf{r}_{NOE}$**(Å)***	$\mu_1 10^{-4}$	$\mu_2 10^{-5}$	μ_3	**Skewness**
Tyr 64-His 72	3.12	11	2.3	0.0	-0.69
Tyr 76-His 84	4.80	0.81	0.17	0.0	-0.19
Tyr 64-Tyr 82	6.37	0.15	0.0044	0.0	-0.21
Tyr 64-Tyr 88	6.52	0.13	0.0027	0.0	-0.23
Tyr 70-Tyr 82	7.01	0.08	0.0014	0.0	-0.23

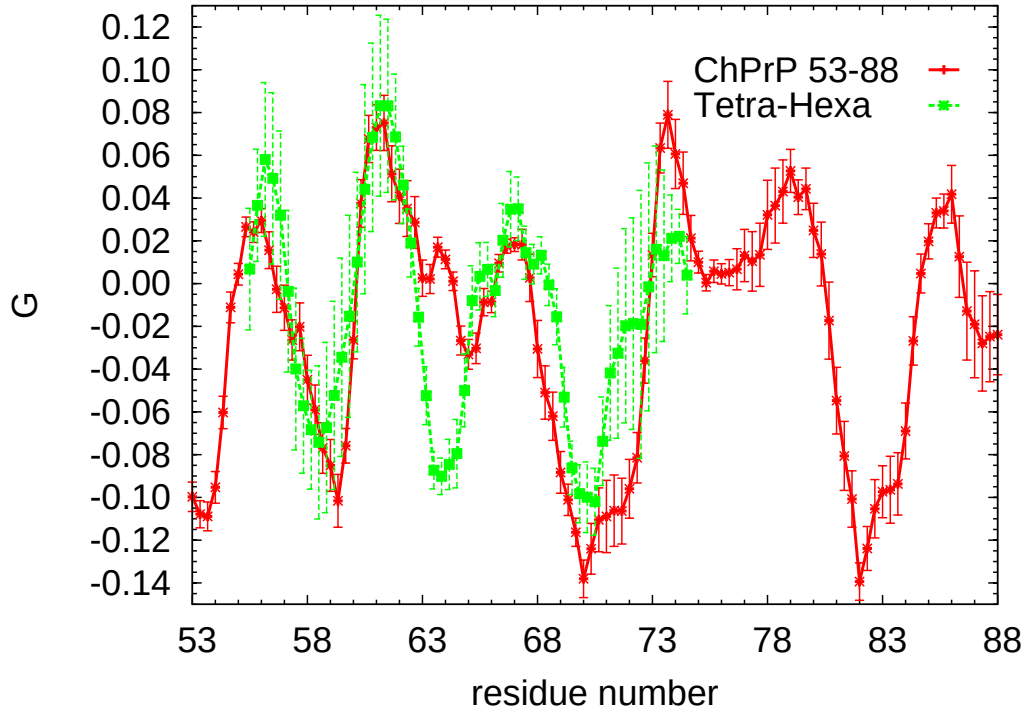


Figure 4.38: Chirality index, G, averaged among the trajectories of the ChPrP1-267, focused on the hexarepeat region, ChPrP53-88. The chirality index pattern of the tetrarepeat fragment, previously simulated [56], is also shown by comparison. First, the index shows a periodical pattern, as found in the tetrarepeat fragment and, consistently with this latter one, turn regions are frequently populated together with 3_{10} helices (68-73 and 80-85). Error bars are reported for each one of the G values as standard deviations on the ensemble of trajectories.

4.8 Conclusions

The conformation adopted at neutral pH by the full avian prion protein structure ChPrP1-267 in solution at neutral pH is reported. From our results, the three α helices show different flexibility, in particular helix 2 is very rigid, as shown both from chirality and DSSP analysis, while helix 1, which interestingly was found to be stable in the mouse prion [111] and, prevalently, helix 3, do not completely mantain the α helix structure, resulting in a coexistence with 3_{10} helix and highly flexible states. Differently than the human prion protein, in which the most rigid helix was found to be helix 3, avian helix 2 possesses a proline residue in the first position, that could be the cause of its high rigidity. On the contrary the avian β sheet is quite short and mobile, partly because of an interaction with helix 3 that prevents a perfect parallelism between the two forming strands.

Also the hexa-repeat region, in which a periodical conformation is adopted, presents unexpectedly high rigidity. This is pointed out especially by the low standard deviations of the chirality analysis, which also reveals the abundance of 3_{10} helices for residues 68-73 and 80-85 and a type I β turn structure, prevalently found in residue 59. Moreover, a hydrogen bond was detected mainly between the imidazole nitrogen of histidine 72 and the phenolic hydrogen of tyrosine 64. It is important to note that this bond was also found previously [56], in the trajectories of the tetra-repeat fragment and cannot be formed by the mammal sequence because of its lack of the required residues.

The solvation study is consistent with the chirality analysis, revealing almost the same pattern for the residues, found using backbone chirality. In particular, the regions with defined secondary structure, as the repeat region and the helix ones, present local minima and again helix 2 show low values of standard deviations, and the residues with polyproline II chirality values usually correspond to local solvation maxima.

These results suggest different conformational preferences of the avian prion protein with respect to the mammal one. The finding that a periodical structured conformation is adopted in the hexa-repeat region of the avian prion protein may be correlated to its high resistance to protease [33]. Such a structured conformation of the N-terminal tail, together with the lower flexibility of ChPrP helix 2 with respect to the mammal prion analogue, and the plasticity of the avian β sheet, could somehow hamper the interconversion leading to the pathogenic PrP^{Sc} isoform, explaining the rarity of prion diseases registered for avians.

The results reported in these chapter, which reveal a periodic conformation in the hexarepeat region, are encouranging every research that aims at understanding proteins by studying only limited segments of them, and goes to partial support of the idea that folding begins from the initial formation of local secondary structures that subsequently assemble in tertiary and quaternary structures.

Bibliography

[1] F. Wopfner, G. Weidenhofer, R. Schneider, A. von Brunn, S. Gilch, T. F. Schwarz, T. Werner, and H. M. Schatzl, *J. Mol. Biol.*, **1999**, *289*, 1163–1178.

[2] D. A. Harris, D. L. Falls, F. A. Johnson, and G. D. Fischbach, *Proc. Natl. Acad. Sci. USA*, **1991**, *88*, 7664–7668.

[3] J. M. Gabriel, B. Oesch, H. Kretzschamer, M. Scott, and S. B. Prusiner, *Proc. Natl. Acad. Sci. USA*, **1992**, *89*, 9097–90101.

[4] S. B. Prusiner, *Proc. Natl. Acad. Sci. USA*, **1998**, *95*, 13363–13383.

[5] N. Stahl, M. A. Baldwin, D. B. Teplow, L. Hood, B. W. Gibson, A. L. Burlingame, and S. B. Prusiner, *Biochemistry*, **1993**, *32*, 1991–2002.

[6] H. Buëler, M. Fisher, Y. Lang, H. Bluethmann, H. P. Lipp, S. J. DeArmond, S. B. Prusiner, M. Aguet, and C. Weissmann, *Nature*, **1992**, *356*, 577–582.

[7] S. Brandner, S. Isenmann, A. Raeber, M. Fischer, A. Sailer, Y. Kobayashi, S. Marino, C. Weissmann, and A. Aguzzi, *Nature*, **1996**, *379*, 339–343.

[8] J. C. Rochet and P. Jr Lansbury, *Curr. Opin. Struct. Biol.*, **2000**, *10*, 60–68.

[9] C. M. Dobson, *Trends Biochem. Sci.*, **1999**, *24*, 329–332.

[10] M. Sunde and C. Blake, *Adv. Prot. Chem.*, **1997**, *50*, 123–159.

[11] B. Schuler, R. Rachel, and R. Seckler, *J. Biol. Chem.*, **1999**, *274*, 18589–18596.

[12] M. Fändrich, M. A. Fletcher, and C. M. Dobson, *Nature*, **2001**, *410*, 165–166.

[13] E. Scherzinger, A. Sittler, K. Schweiger, V. Heiser, R. Lurz, R. Hasenbank, G. P. Bates, H. Lehrach, and E. Wanker, *Proc. Natl. Acad. Sci. USA*, **1999**, *96*, 4604–4609.

[14] M. F. Jobling, X. Huang, L. R. Stewart, K. J. Barnham, C. Curtain, I. Volitakis, M. Perugini, A. R. White, R. A. Cherny, C. L. Masters, C. Barrow, S. J. Collins, A. Bush, and R. Cappai, *Biochemistry*, **2001**, *40*, 8073–8084.

[15] T. Miura, K. Suzuki, N. Kohata, and H. Takeuchi, *Biochemistry*, **2000**, *39*, 7024–7031.

[16] N. V. Uversky, J. Li, and L. A. Fink, *J. Biol. Chem.*, **2001**, *276*, 44284–44296.

[17] P. D. Davis, G. Gallo, M. S. Vogen, J. L. Dul, K. L. Sciarretta, A. Kumar, R. Raffen F. J. Stevens, and Y. Argon, *J. Mol. Biol.*, **2001**, *313*, 1021–1034.

[18] C. J. Morgan, M. Gelfand, C. Atreya, and A. D. Miranker, *J. Mol. Biol.*, **2001**, *309*, 339–345.

[19] D.R. Brown, C. Clive, and S. J. Haswell, *J.Neurochem.*, **2001**, *76*, 69–76.

[20] R. Chiesa, B. Drisaldi, E. Quaglio, A. Migheli, P. Piccardo, B. Ghetti, and D. Harris, *Proc. Natl. Acad. Sci. USA*, **2000**, *97*, 5574–5579.

[21] S. R. Mouillet, M. Ermonval, C. Chebassier, J. L. Laplanche, S. Lehmann, J. M. Launay, and O. Kellermann, *Science*, **2000**, *289*, 1925–1928.

[22] J. Herms, T. Tings, S. Gall, A. Madlung, A. Giese, H. Siebert, P. Schurmann, O. Windl, N. Brose, and H. Kretzschmar, *J. Neurosci.*, **1999**, *19*, 8866–8875.

[23] D. R. Brown, K. Qin, J. W. Herms, A. Madlung, J. Manson, R. Strome, P. E. Fraser, T. Kruck, A. von Bohlen, W. Schultz Schaeffer, A. Giese, D. Westaway, and H. Kretzschmar, *Nature*, **1997**, *390*, 684–687.

[24] R. R. Brentani, R. Linden, A. C. Magalhaes, M. A. M. Prado, V. F. Prado, and A. Silva, *J. Mol. Biol.*, **2004**, *336*, 1175–1183.

[25] K. Simons and E. Ikonen, *Nature*, **1997**, *387*, 569–572.

[26] P. C. Pauly and D. A. Harris, *J. Biol. Chem.*, **1998**, *273*, 33107–33110.

[27] D. R. Brown, B. Wong, F. Hafiz, C. clive, S. J. Haswell, and I. M. Jonew, *Biochem. J.*, **1999**, *344*, 1–5.

[28] E. D. Walter, D. J. Stevens, M. P. Visconte, and G. L. Millhauser, *J. Am. Chem. Soc.*, **2007**, *129*, 15440-15441.

[29] D. R. Brown, *Brain Res. Bull.*, **2001**, *55*, 165–173.

[30] S. B. Prusiner and M. R. Scott, *Ann. Rev. Genet.*, **1997**, *31*, 139–175.

[31] H. A. Kretzshmar, S. B. Prusiner, L. E. Stowring, and S. J. DeArmond, *Am. J. Pathol.*, **1986**, *122*, 1–5.

[32] B. W. Caughey, A. Dong, K. S. Bhat, D. Ernst, S. F. Hayes, and W. S. Caughey, *Biochemistry*, **1991**, *30*, 7672–7680.

[33] E. M. Marcotte and D. Eisemberg, *Biochemistry*, **1999**, *38*, 667–676.

[34] N. Sthal and S. B. Prusiner, *Faseb J.*, **1991**, *5*, 2799–2807.

[35] F. Lopez Garcia, R. Zahn, R. Riek, and K. Wüthrich, *Proc. Natl. Acad. Sci. USA*, **1997**, *94*, 8334–8339.

[36] D. G. Donne, J. H. Viles, D. Groth, I. Mehlhorn, T. L. James, F. E. Cohen, S. B. Prusiner, P. E. Wright, and H. J. Dyson, *Proc. Natl. Acad. Sci. USA*, **1997**, *94*, 13452–13457.

[37] L. Calzolai, D. A. Lysek, D. R. Perez, P. Güntert, and K. Wüthrich, *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 651–655.

[38] K. M. Pan, M. Baldwin, J. Nguyen, M. Gasset, A. Serban, D. Groth, I. Melhorn, Z. Huang, R. J. Fletterick, F. E. Cohen, and S. B. Prusiner, *Proc. Natl. Acad. Sci. USA*, **1993**, *90*, 10962–10966.

[39] J. H. Viles, F. E. Cohen, S. B. Prusiner, D. B. Goodin, P. E. Wright, and H. J. Dyson, *Proc. Natl. Acad. Sci. USA*, **1999**, *96*, 2042–2047.

[40] G. Di Natale, G. Grasso, G. Impellizzeri, D. La Mendola, G. Micera, N. Mihala, Z. Nagy, K. Osz, G. Pappalardo, V. Rigó, E. Rizzarelli, D. Sanna, I. Sóvágó, *Inorg. Chem.* **2005**, *44*, 7214–7225.

[41] F. Wopfner, G. Weidenhofer, R. Schneider, A. von Brenn, S. Gilch, T.F. Schwarz, T. Werner, and H. M. Schatzl, *J. Mol. Biol.*, **1999**, *289*, 1163–1178.

[42] D. A. Harris, D. L. Falls, F. A. Johnson, and G. D. Fischbach, *Proc. Natl. Acad. Sci. USA*, **1991**, *88*, 7664–7668.

[43] L. Calzolai, D. A. Lysek, D. R. Perez, P. Güntert, and K. Wüthrich, *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 651–655.

[44] D. A. Harris, M. T. Huber, P. van Dijken, S. L. Shyng, B. T. Chait, and R. Wang, *Biochemistry*, **1993**, *32*, 1009–1016.

[45] M. Nunziante, S. Gilch, and H. M. Schätzl, *J. Biol. Chem.*, **2003**, *278*, 3726–3734.

[46] K. N. Frankenfield, E. T. Powers, and J. W. Kelly, *Protein Sci.*, **2005**, *14*, 2154–2166.

[47] R. Butowt, P. Davies P., and D. R. Brown, *J. Neurosci. Res.*, **2007**, *85*, 2567–2579.

[48] S. L Shyng, K. L. Moulder, A. Lesko, and D. A. Harris, *J. Biol. Chem.*, **1995**, *270*, 14793–14800.

[49] A. Radzicka, S. A. Acheson, and R. Wolfenden, *Bioorg. Chem.*, **1992**, *20*, 382–386.

[50] J. S. Jhon and Y. K. Kang, *J. Phys. Chem. A*, **1999**, *103*, 5436–5439.

[51] C. Benzi, R. Improta, G. Scalmani, and V. Barone, *J. Comput. Chem.*, **2002**, *23*, 341-350.

[52] D. A. Lysek and K. Wüthrich, *Biochemistry*, **2004**, *23*, 10393–10399.

[53] C. J. Smith, A. F. Drake, B. A. Banfield, G. B. Bloomberg, M. S. Palmer, A. R. Clarke, and J. Collinge, *FEBS Lett.*, **1997**, *405*, 378-384.

[54] M. P. Hornshaw, J. R. McDermott, J. M. Candy, and J. H. Lakey, *Biochem. Biophys. Res. Comm.*, **1995**, *214*, 993-999.

[55] A. Pietropaolo, L. Raiola, L. Muccioli, G. Tiberio, C. Zannoni, R. Fattorusso, C. Isernia, D. La Mendola, G. Pappalardo, and E. Rizzarelli, *Chem. Phys. Lett.*, **2007**, *442*, 110–118.

[56] A. Pietropaolo, L. Muccioli, C. Zannoni, D. La Mendola, G. Maccarrone, G. Pappalardo, and E. Rizzarelli, *J. Phys. Chem. B*, accepted **2008**.

[57] A. Pietropaolo, L. Muccioli, C. Zannoni, and E. Rizzarelli, *to be submitted*, **2008**.

[58] D. La Mendola, R. P. Bonomo, G. Impellizzeri, G. Maccarrone, G. Pappalardo, E. Rizzarelli, A. Pietropaolo, and V. Zito, *J. Biol. Inorg. Chem.*, **2005**, *10*, 463–475.

[59] C. Bartels, T. Xia, M. Billeter, and K. Wüthrich, *J. Biomol. NMR*, **1995**, *5*, 1–10.

[60] J. Cavanagh, W. J. Fairbrother, A. G. Palmer III, and N. J. Skelton, *Protein NMR Spectroscopy, Principles and Practice,;* Academic Press, San diego, 1996.

[61] T. L. Hwang and A. J. Shaka, *J. Magnet. Reson. A*, **1995**, *112*, 275–279.

[62] T. Herman, P. Güntert, and K. Wüthrich, *J. Mol. Biol.*, **2002**, *319*, 209–227.

[63] N. Guex and M. C. Peitsch, *Electrophoresis*, **1997**, *18*, 2714–2723.

[64] R. Koradi, M. Billeter, and K. Wüthrich, *J. Mol. Graph.*, **1996**, *14*, 51–55.

[65] P. Procacci, E. Paci, T. A. Darden, and M. Marchi, *J. Comput. Chem.*, **1997**, *18*, 1848–1862.

[66] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *J. Am. Chem. Soc.*, **1995**, *117*, 5179–5197.

[67] D. Frenkel and B. Smit, *Understanding Molecular Simulations: From Algorithms to Applications;* New York: Academic, 2000.

[68] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, *Intermolecular Forces;* Reidel, Dordrecht: B. Pullman, editor; 1981.

[69] C. Dugave and L. Demange, *Chem. Rev.*, **2003**, *103*, 2475–2532.

[70] A. A Adzhubei and M. J. E. Stemnberg, *J. Mol. Biol.*, **1993**, *229*, 472–493.

[71] W. Kabsch and C. Sander, *Biopolymers*, **1983**, *22*, 2577–2637

[72] S. Doniach, *Chem. Rev.*, **2001**, *101*, 11698–11703.

[73] G. Wagner, W. Braun, F. T. Havel, T. Schaumann, N. Go, and K. Wüthrich, *J. Mol. Biol.*, **1987**, *196*, 611–639.

[74] O. Jardetzky and G. C. K. Roberts, *NMR in Molecular Biology;* Academic Press, New York, 1981.

[75] N. Berova, K. Nakanishi, and R. Woody, *Cicular dichroism: Principles and Applications;* Wiley-Vch, New York, 1986.

[76] R. W. Woody, *In Circular Dichroism and the conformational analysis of biomolecules (Fasman, G.D., ed);* pp 25-67, Plenum Press, New York, 1996.

[77] A. Perczel, K.Park, and G.D. Fasman, *Anal. Biochem.*, **1992**, *203*, 83–93.

[78] A. Perczel, M. Hollosi, G. Tusnady, and G.D. Fasman, *Protein Eng.*, **1991**, *4*, 669–679.

[79] S. Brahms and J. Brahms, *J. Mol. Biol.*, **1980**, *138*, 149–178.

[80] J. Bandekar, D. J. Evans, S. Krimm, S. J. Leach, S. Lee, J. R. McQuie, E. Minasian, G. Nemethy, M. S. Pottle, H. A. Sheraga, E. R. Stimson, and R. W. Woody, *Int J. Pept. Protein Res.*, **1982**, *19*, 187–205.

[81] R. W. Woody, *Adv. Biophys. Chem.*, **1992**, *2*, 37–39.

[82] E. A. Bienkiewicz, A.-Y. M. Woody, and R. W. Woody, *J. Mol. Biol.*, **2000**, *297*, 119–133.

[83] K. MA and K. Wang, *Biochem. J.*, **2003**, *374*, 687–695.

[84] G. Hutchinson and J. M. Thornton, *Protein Sci.*, **1994**, *3*, 2207–2216.

[85] Y. Che and G. R. Marshall, *Biopolymers*, **2006**, *81*, 392–406.

[86] G. D. Rose, L. M. Gierasch, and J. A. Smith, *Adv. Protein Chem.*, **1985**, *37*, 1–109.

[87] J. F. Bazan, R. J. Fletterick, M. P. McKinley, and S. B. Prusiner, *Protein Eng.*, **1987**, *1*, 125–135.

[88] A. Bansal and L. M. Gierasch, *Cell*, **1991**, *67*, 1195–1201.

[89] J. P. Paccaud, W. Reith, B. Johansson, K. E. Magnusson, B. Mach, and J. L. Carpenter, *J. Biol. Chem.*, **1993**, *268*, 23191–23196.

[90] G. Arena, E. Rizzarelli, S. Sammartano, and C. Riganò, *Talanta*, **1979**, *26*, 1.

[91] P. Gans, A. Sabatini, and A. Vacca, *Talanta*, **1996**, *43*, 1739-1753.

[92] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and J. C. Berendsen, *J. Comput. Chem.*, **1995**, *26*, 1701–1718.

[93] L. J. Smith, X. Daura, and W. F. van Gusteren, *Proteins*, **2002**, *48*, 487-496.

[94] S.M. Feller, R. Ren, H. Hanafusa, and D. Baltimore, *Trends Biochem. Sci.*, **1994**, *19*, 453–458.

[95] C. Krittanai and W. C. Johnson, *Anal. Biochem.*, **1997**, *253*, 57-64.

[96] A. Chakrabartty, T. Kortemme, S. Padmanabhan, and R. L. Baldwin, *Biochemistry*, **1993**, *32*, 5560-5565.

[97] M. Vendruscolo, R. Najmanovich, and E. Domany, *Phys. Rev. E*, **1999**, *82*, 656-659.

[98] A. Pietropaolo, L. Muccioli, R. Berardi, and C. Zannoni, *Proteins*, **2008** *70*, 667–677.

[99] M. Solymosi, R. J. Low, M. Grayson, and M. P. Neal, *J. Chem. Phys.*, **2002**, *116*, 9875–9881.

[100] M. A. Osipov, B. T. Pickup, and D. A. Dunmur, *Mol. Phys.*, **1995**, *84*, 1193–1206.

[101] E. Langella, R. Improta R, and V. Barone, *Biochem. J.*, **2004**, *87*, 3623–3632.

[102] A. De Simone, G. G. Dodson, C. S. Verma, A. Zagari, and F. Fraternali, *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 7535–7540.

[103] M. Pappalardo, D. Milardi, D. Grasso, and C. La Rosa, *New J. Chem.*, **2007**, *31*, 901–905.

[104] H. F. Ji and H. Y. Zhang, *Biochem. Biophys. Res. Comm.*, **2007**, *359*, 790–794.

[105] G. A. Kaminski GA, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, *J. Phys. Chem.*, **2001**, *105*, 6474–6487.

[106] T. Darden, D. York, and L. Pedersen, *J. Chem. Phys.*, **1993**, *98*, 10089–10092.

[107] H. J. C. Berendsen, J. P. M. Postma, A. Di Nola, and J. R. Haak, *J. Chem. Phys.*, **1984**, *81*, 3684–3690.

[108] M. Vendruscolo, R. Najmanovich, and E. Domany, *Phys. Rev. E*, **1999**, *82*, 656–659.

[109] J. S. Richardson and D. C. Richardson, *Proc. Natl. Acad. Sci. USA*, **2002**, *99*, 2754–2759.

[110] A. De Simone, G. G. Dodson, A. Zagari, and F. Fraternali, *FEBS Lett.*, **2006**, *580*, 2488–2494.

[111] A. Barducci, R. Chelli, P. Procacci, V. Schettino, F. L. Gervasio, and M. Parrinello, *J. Amer. Chem. Soc.*, **2006**, *128*, 2705–2710.

[112] R. A.George and J. Heringa, *Protein Eng.*, **2003**, *15*, 871–879.

[113] R. I. Dima and D. Thirumalai, *Biochem. J.*, **2002**, *83*, 1268–1280.

[114] M. Pappalardo, D. Milardi, C. La Rosa, C. Zannoni, and D. Grasso E. Rizzarelli, *Chem. Phys. Lett.*, **2004**, *390*, 511–516.

[115] W. F. van Gunsteren and H. J. C. Berendsen; *Groningen Molecular Simulation (GROMOS) Library Manual;* Biomos, Groningen, **1987**.

This thesis was supported by
FIRB; Grant number: RBNE03PX83; PRIN, Grant number: 2005035119; PRIN2006, Grant number: 033492; Marco Polo project 2006 (visit at Chemistry Research Laboratory, University of Oxford)

I would like to remember all the people that I met in these three years during my PhD in Bologna.
Many thanks to Prof. Claudio Zannoni for having taught me physical chemistry and to Prof. Enrico Rizzarelli for having taught me bioinorganic chemistry, passing me the curiosity in studying proteins.
Many thanks to Dr. Luca Muccioli, for having taught me the computational tricks, the ethical guidelines for a good paper and to drink coffee, I have never drank coffee before working with computer and sorry if sometimes, because of coffee, I was a bit noisy!

Thanks to Dr. Roberto Berardi, aka Bebo, for having helped me in math and with linux, everytime I needed.
Thanks to Dr. Silvia Orlandi for her cheerful smile.
Thanks so much to all the people in my lab for the nice company.

Thanks to all the people that worked with me in the Italian PhD Association, ADI.

Cisco, thank you for having been there, when I was there.

I wish also to thank the nice people known in the Chemistry Research Laboratory (CRL) of Oxford.
Many thanks to Dr. Lorna J. Smith for having followed me so kindly in Oxford.
Thanks also to Dr. Christina Redfield for having shown me how to set an NMR experiment.

Thanks also to Sandra for the fun in the lab and for our friendship which I hope will last in the years.
To Ramani for her really nice lunch.
To Giuseppe, Lion, the S. Hilda's flatmates and all the chinese people that I knew. Thanks for not having left me alone during the months passed in UK!
Finally, many thanks to my dad for passing so much time with me talking about chemistry and also to my mum and my brother, for kindly talking with me everytime I feel alone.

Ad Maiora Semper.