Università degli Studi di Pisa
Facoltà di Scienze Matematiche Fisiche e Naturali
Corso di Laurea Magistrale in Fisica

Tesi di Laurea Magistrale

A Minimalist Model for Simulation of Structure and Dynamics of Helical Polypeptides

Candidato:
Giulia Lia Beatrice Spampinato

Relatore:
Dott. Valentina Tozzini

Anno Accademico 2012–2013

To my Zia Marisa and my Zio Ughino

CONTENTS

INTRODUCTION

Context

This Thesis reports a study in the framework of proteins dynamics computer simulations. In order to introduce the problems related to the *in silico* representation of proteins, the Chapter 1 of this Thesis is devoted to a description of their structural organization.

Proteins are molecular machines, building block and arms of a living cell. They are finely structured biomolecules highly specialized for functional roles. A protein is organized in hierarchical levels. The primary structure is a chain formed by amino acids (of 20 different types) linked together by peptide bonds in a specific sequence forming the polypeptide. The secondary structure describes local recurrent structural motifs shaping the polypeptide. The tertiary structure is the organization of secondary structures, through the interactions between residues often widely apart in the primary sequence. A quaternary structure describes how the tertiary structures of different polypeptide chains organize themselves.

The biological function of a protein depends on its overall 3D fold. The chain folds through a stepwise process, generally mirroring the hierarchical structural organization. To understand the final shape of a protein, the deep comprehension of the secondary structures and of their sequential determinants is thus mandatory, as it is the first step of this hierarchy.

Given the rigid geometry of the peptide bond, two internal variables (for each amino-acid) are sufficient to describe the conformation of the polypeptide's backbone. These are the dihedral angles Φ,Ψ describing the rotation around the two single bonds connecting the central amino-acid Carbon (C_α) with its neighboring amino- and carbossilic- groups along the chain. The distribution of the (Φ,Ψ) couples represented in a plane is called the Ramachandran plot (RP). The two main classes of secondary structures, namely helices and sheets, occupy well distinct areas of the RP. More difficult is the separation of the different sub-classes of helices (α-helix, 3_{10}-helix and π-helix), located in near and partially superposing areas of the RP. While the sheets are stabilized by hydrogen bonds connecting amino-acids belonging to different strands, often sequentially far apart, helices are stabilized by periodic and local intra-strand patterns of hydrogen bonds.

Any modeling and simulation approach must account for these structural features. Particularly important is the ability of discriminating and describe accurately all the different secondary structures, their dynamics, and possibly the free energy differences and transitions among them.

The most popular approaches used to address these issues are the atomistic molecular dynamics simulations based on empirical force fields. Within this framework, all the atomic degrees of freedom are treated explicitly and the atoms interacts among each other by means of empirical interactions mimicking chemical bonds, Van der Waals and electrostatic interactions. The parameters included in the model have been optimized in the last decades based on

higher accuracy calculations and on experimental data. Though the atomistic-empirical approaches have indeed given an unprecedented insight into the biomolecular processes and proven valuable tools to interpret experimental data, recently they start to reveal their limitations.

One obvious limitation is the computational cost: an hydrated proteins contains approximately $10^4 - 10^5$ atoms. This size can be conveniently addressed on single or few processors workstations, producing 10-100 ns long runs. Up to 2-3 orders of magnitude with respect to these scales can be gained on heavily parallel systems, allowing reaching the μsec scale for supra-bio molecular aggregates. However, to routinely address the biologically interesting scales (sub cellular, and msecs) one has to reduce the computational cost with more efficient algorithm or to wait for more powerful processors or innovative parallel architectures (although intrinsic limitation to both start to appear as well).

In addition, on the macroscopic time scales, in the few cases in which those have been reached, the empirical atomistic force field have started to reveal inaccuracies, specifically in reproducing the free energy differences between different secondary structures. A great effort is currently in the course to correct these inaccuracies, but it is not likely that this could happen without increasing the complexity of the force field.

Goal of the work

A possible way out to solve both problems, is to reconsider the model itself, and face it from a different point of view. Instead of making the system's Hamiltonian more complex, one could simplify it, keeping only the minimum necessary information. This would obviously solve the problem of computational cost. Less intuitively, it could help also solving the problem of inaccuracy on the statistical scale: a fewer number of parameters with minimal redundancy could be optimized in order to reproduce a given number of properties, in principle even with a larger accuracy than those of the atomistic force fields, in which, in order to maintain coherence, many parameters have to be modified in a correlated way.

This considered, several attempts exist in the literature to build for proteins at less than atomic resolution (i.e. Coarse Grained, CG, models). The road followed in this work is the one of the one-bead-per-amino-acid models (OB models), i.e. models in which a single interacting center for each amino-acid is used. In addition, the specific focus is on C_α based OB models. CG and the OB models are reviewed in Chapter 2 of this Thesis.

This work is included in a wider framework, aimed at optimizing a "minimalist model for proteins", i.e. model capable of reproducing accurately the structure of proteins with the minimal possible number of internal degrees of freedom (i.e. the coarser possible representation) and with the possibility of back-mapping to the atomistic representation. The idea of the CG models to represent efficiently macro-bio-molecular systems in computer simulations trace back to the seventies of the past century. From time to time these models have been reconsidered, especially recently, due to the need of bridging, through simulation, the macroscopic experimental data with the biochemistry. However, these models have never reach a standard (as done by the atomistic

models instead). This is also because their theoretical foundations have always received less attention than the applications.

This Thesis work, conversely, aims at shedding some light into the fundamental aspects and properties of the CG models, specifically the minimalist ones. The C_α OB models are good candidates for this role, and are those considered in this Thesis. Specifically this work focuses in analyzing the following three properties of these models: (i) possibility of back-mapping to the full atomistic representation, (ii) capability of describing all the different kinds of secondary structures and their dynamics/thermodynamics and (iii) in general, capability of predicting accurately structure and dynamics of the global fold of the protein. In addition, this work provides an optimized parameterization for this model, capable of satisfying (ii) and (iii) (within certain limits).

Strategy and tools

The strategy used to verify the above mentioned properties and to optimize the parameterization is the comparison of the results from simulations with experimental data. In Chapter 2 the bases of the molecular dynamics simulations and of the tools used to analyze their results are also reported.

However, an important part of the parameterization strategy is related to the choice of experimental data to which compare. In this work it was chosen to base parameterization on the structural information. The available structural experimental data from different sources are organized in a world-wide freely accessible database (the RCSB Protein Data Bank, or PDB). In this work, however, it was necessary to select data on the basis of primary and secondary structures, to reduce the amount of data to the minimal necessary information (namely to "coarse grain" the structures) and then to statistically analyze them, namely to build the distribution of the internal coordinates related to the OB-CG model.

No software tool was available for these tasks, thus a first original contribution of this Thesis work was to build this tool. A software package, SecStAnT, was built capable of downloading from the PDB data sets with user-defined properties (e.g., maximum-minimum size of a protein, prevalence (or not) of a given secondary structure, given sequential/structural diversity among proteins, and many others). At will, the SecStAnT can then "coarse grain" the structures at different levels included of course the one of the minimalist model, and analyze the distributions of internal variables, and their 2D and 3D correlations. The software is made freely available to the scientific community, under the BSD Open Source license. This software is a tool with its own utility even out of the context of the CG and minimalist models. In fact, to our knowledge, it is one of the most flexible structure selection and statistical analysis tools, specifically regarding the evaluation of correlations between internal variables. SecStAnT together with related results on the analysis of the secondary structure dependent internal variables analysis is described in Chapter 3.

Results

It is to be remarked that SecStAnT and other analyses reported in Chapter 3 are already original contributions, which are, in fact, submitted to the journal *Bioinformatics*. In this Chapter it was also shown that the internal variables of the C_α OB models are "good" variables that satisfy condition (i) (back-mapping) at least for the backbone. This was shown from the comparison of the RP built with the atomistic representation with its counterpart for the minimalist model, involving the two conformational internal variables of the C_α chain (i.e. the pseudo-bond angle θ and the pseudo-dihedral ϕ). Even in the minimalist representation, the secondary structures occupy separated areas in the (θ,ϕ) plane, indicating that the minimalist model can represent the secondary structures, and that back-mapping to the atomistic RP is possible.

However the main set of results is reported in Chapter 4, and are about the parameterization of the model in such a way that it is capable of reproducing the secondary structures with a high level of accuracy. In this part SecStAnT has been used to produce distributions of internal variables θ,ϕ and other involved in the description of the secondary structures (e.g. the distances between the third, fourth and fifth neighboring C_αs along the chain, related to the hydrogen bonds stabilizing the helices) and their correlations. These data have then been used as targets, and the parameterization has been optimized to reproduce them in the simulations of their minimalist model, run with DL_POLY. The parameters optimization is then carried out by means of a physically driven trial-and-error procedure. Simulations on the different kind of helices are then produced, also on the macroscopic time scales. These are shown to reproduce accurately all the known structural and dynamical features of these secondary structures.

The final goal is to produce a general model capable of describing all the secondary structures, and to combine them in tertiary structures. The force field of the model here optimized already contains a set of conformational terms directly related to the internal variables θ,ϕ, aimed at describing the general conformational flexibility of the backbone even in the case of weakly structured or de-structured proteins. Terms mimicking the hydrogen bonds stabilize the different secondary structures. In this work, those terms for the helical structures were optimized. In subsequent works the optimization of similar terms for the sheets structures could be addressed and finally combined with the helical ones in a sequence dependent fashion. In this phase, experimental information on the relative free energy of the different secondary structures can be included. In addition, the relative weight of the different secondary structures terms can be made sequence dependent, in order to give predictive power to the model also concerning the primary to secondary structures passage.

The Hamiltonian of the proposed model is composed by a minimal number of terms, whose meaning can be directly understood in terms of physical interactions (e.g. hydrogen bonds). This, together with the high accuracy, can be considered the main innovation of this model: a physically based parameterization allow to straightforwardly extend the model to include other secondary structures, giving to it generality and predictive power.

In Chapter 5, conclusions are reported, including a preliminary definition of the next steps of the research and possible applications.

ABBREVIATIONS

A number of abbreviations has been adopted in order to maintain consistency throughout. They are summarized here:

NMR Nuclear Magnetic Resonance

PDB Protein Data Bank

RCSB Research Collaboratory for Structural Bioinformatics

AA Amino Acid

RP Ramachandran Plot

DOF Degree Of Freedom

FF Force Field

CG Coarse Grain

OB One Bead

EN Elastic Network

BI Boltzmann Inversion

SP Statistical Potential

PMF Potential of Mean Force

MD Molecular Dynamics

RMSD Root Mean Square Displacement

RMSF Root Mean Square Fluctuations

NOTE

All the cartoon representations of proteins have been realizied with the VMD [45] software.

1 THE STRUCTURE OF PROTEINS

In this Chapter the main features of proteins are described. Protein systems have a typical hierarchical organization, structured in primary, secondary, tertiary and quaternary structure. The focus of this Thesis work is on modeling the secondary structure. Different secondary structures are analysed together with the forces maintaining them and their possible empirical descriptions. This Chapter also reports the state of the art about the definition of the secondary structure motifs and their prediction, starting from the primary structure. The Chapter is closed by a brief overview of the other levels of organization (tertiary and quaternary) and of the forces holding together the different parts of proteins.

1.1 INTRODUCTION TO PROTEINS

Proteins are finely structured biomolecules highly specialized for functional roles. They are molecular machines, buildings blocks and arms of living cells.
The synthesis of a protein, proceeding from the genetic information, occurs through a ribosome [56].
In an active protein the polypeptide is folded in a specific 3D structure. From the amino acid sequence, the polypeptide chain folds in different levels of 3D organization, showing a hierarchical structure. A chain is active when completely folded in a highly specific structure.
The main role played by proteins is the enzymatic catalysis of chemical conversions in and around the cell: they help chemical reactions of a living system to occur. Furthermore, they perform most of the cellular function: regulatory proteins control gene expressions; receptor proteins recognize in the lipidic membrane the right (ormonal) signals to communicate between cells; structural proteins maintain the structure of the cell and form the tissue that protect it; transfert protein transport other molecules. Many additional different specific proteins serve different functional roles. The huge variety of protein functions derives from the high specificity of the roles they play, while interacting with other molecules. Each specific relationship demands a fairly rigid spatial structure of the protein. For this reason, their biological functions are closely related to their three-dimensional (3D) structures.
According to the different enviromental conditions and the general structural features, proteins can be roughly divided into three classes [59]: globular proteins, usually found in acqueous environments; fibrous proteins and membrane proteins, lying in water-deficient environments.
Globular proteins are the most common and their structures are easier to be experimentally solved. They can be found with a variety of structures.
Fibrous proteins have also different structures and perform different roles in cells than globular proteins. They are constituents of fibers found in living

organism. They have the common role to confer strength and rigidity to the structure. The three main groups of fibrous proteins are: collagens, silk fibroin and keratins and all occupy pivotal roles within cells [15].
The last big class of proteins are the membrane proteins. They reside in a water-deficient membrane environment, although they usually partially project into water. They are firmly embedded within the hydrophobic bilayer, highly regular and highly hydrogen-bonded but restricted in size by the membrane thickness. Removal from this environment frequently results in a loss of structure and function. So, running experiments on them in their natural environment is a really hard task and the number of membrane proteins structures solved is relatively small, compared to globular proteins [15].
In conclusion, the 3D structure of a protein depends on its aminoacidic sequence and strictly determines its biological function.

1.2 EXPERIMENTAL DETERMINATION OF PROTEINS STRUCTURE

There are some different experimental tecniques to determine the 3D structure of a protein, the two main methods are the Xray crystallography and the Nuclear Magnetic Resonance (NMR).

1.2.1 Xray Crystallography

Xray crystallography extracts the atomic structure of a protein from the diffraction pattern generated by an Xray beam scattered from a crystal.
In a crystal of proteins, there is a protein in each lattice site. The X-ray radiation impacts on the crystal and it is elastically diffused in every direction. In some specific direction the diffused rays interfere constructively. Their intensity is recorded and the measure provides a form factor F_K, depending on the specific diffusion vector $\vec{K}$. In an ideal one dimensional crystal of particle with electronic density very localized around each lattice site, the form factor is proportional to the Fourier transform of the spatial density of electronic charge. The electronic charge density can then be obtained with an inverse Fourier transform of the measured form factor.
In a crystal of proteins, the electronic charge density is the sum of electronic densities of the single atoms. The structural form factor is the sum of atomic form factor weighted with a phase factor, which depends on the specific position of the atoms in the lattice cell. Since the structural form factor is measured and the atomic factors are known quantities, the atomic positions can be derived, by means of a fitting procedure. The atomic positions are in fact determined and refined through an iterative procedure, which uses atomic models. The output is a model solved to atomic level producing a theoretical electronic density, which best fit the experimental one. The resolution depends on the purity of the crystal and of the temperature (usually 100 K).
The problems related with this technique are:

- the position of the hydrogen (H) atoms is not solved. Their electronic density is not sufficient to reveal them;
- the molecules have to be crystallized. Their structure will be locked in a single configuration depending on the crystallization conditions. Furthermore, some proteins, e.g. membrane proteins, cannot be crystallized;
- the so called phase problem. From the diffraction pattern it is possible to obtain only the module of the (complex) structural form factor with a loss of information. However, with an iterative procedure starting from an approximate structure, the loss is compensated by the a priori knowledge of the system.

It is important to observe that this method is not able to reproduce the flexibility of a protein in a natural environment.

1.2.2 Nuclear Magnetic Resonance

The nuclear magnetic resonance (NMR) spectroscopy is able to analyse proteins in solution and then not constrained in a single structure.
This technique is based on the interaction between the magnetic moment of a nucleus and an applied external magnetic field. The molecule must have nuclei with non-zero spin, i.e. odd number of protons and neutrons. Some of the most used nuclei are for example H and ^{15}N with spin 1/2.
If a magnetic field ($\vec{B}_0$) is applied (along z) to a nucleus, its magnetic moment (μ) precedes at the Larmor frequency (ω_L):

$$\omega_L = B_0\gamma \tag{1}$$

$$\gamma = \frac{gZe}{2M} \tag{2}$$

where γ is the giromagnetic ratio, in which g is the g-factor (normally 1 for classic particles), Ze is the nuclear charge and M the mass of the nucleus.
A new magnetic field is then added, in the xy plane (orthogonal to $\vec{B}_0$), rotating at a frequency ω near the ω_L. For example, this field may be generated by an oscillating radio frequency (RF) circuit. As a consequence, the mean magnetization rotates in the direction orthogonal to z. The largest effect is reached in conditions of resonance ($\omega \sim \omega_L$). In this case, also the absorption of energy of the system from the circuit is at its maximum values.
If the external field $\vec{B}_0$ is fixed (e.g. 10 T), the hydrogen atoms should have all the same ω_L ($\sim$ 420 MHz). However, in each molecule, a single proton feels a local magnetic field $\vec{B}$, which is influenced by the chemical environment around the nucleus. Accordingly, the ω_L varies by a quantity defined as the chemical shift σ, such that:

$$\omega_L = (1-\sigma)B_0\gamma \tag{3}$$

σ depends on the local magnetic field sensed by the nucleus. It is a measure of the chemical environment surrounding the nucleus.
In a protein, the chemical shifts of different carbon atoms belonging to an amino-acid (see next section) are separately measured. Nearby nuclei influence each other causing a split of the principal peak of resonance. By analyzing

the split, bond distances, bond angles and dihedrals can be derived. Moreover, the two-dimensional map of correlation (of chemical shifts for different atoms) gives indications on the relative distance also between non bonded atoms.
In conclusion, analysing NMR spectra and maps, constraints are obtained on bond distances and angles together with non bonded atoms. Solve an entire structure is possible if there is a sufficient number of such constraints with respect to the structure's atoms. These are then included in a theoretical model, which is iteratively optimized until the constraints themselves are satisfied. The result of this procedure is not a single structure, but a set of possible models which satisfies the same constraints. The structural diversities among models reflect the actual conformational freedom of a molecule in solution, together with the thermal fluctuations of the structure. The quality of the structural determination is measured in terms of differences between models: the more the models are similar, the more the structures are accurate.
In conclusion, compared to the X-ray crystallography, the NMR solves structures with lower resolution, but it gives a more realistic image of the conformations assumed by the molecule in its natural environment.

1.2.3 The Protein Data Bank

Structures both from Xray and from NMR (as those from other experimental sources) can be deposited in the Protein Data Bank archive (PDB) [61][66]. This is a worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. The structures in the archive range from tiny proteins and DNA or RNA fragments to complex molecular machines, like the ribosomes.
The PDB archive, established in 1971 at Brookhaven National Laboratory, is freely available to the research community and it is weekly updated. After about thirty years, the Research Collaboratory for Structural Bioinformatics (RCSB) became responsible for its management. In 2003, the wwPDB was formed to maintain a single PDB archive of macromolecular structural data, that is freely and publicly available to the global community. It consists of organizations that act as storage, data processing and distribution centers for PDB data. In addition, the RCSB PDB supports a website [66], where visitors can perform complex queries on the data, run statistical analysis and chart the results.
Submitted structures undergo a revision before inclusion in the database. After acceptance each structure is identified by a PDB 4-letter entry name, that will uniquely identify the structure forever. Structures are deposit in a standard format (.pdb), readable from every visualization software of biomolecules [45]. This format contains bibliographic references, details about the first, secondary and tertiary structure of the protein, details about the specific experimental technique used to solve it and all the atomic coordinates. Appendix A gives a detailed description of the pdb file format.

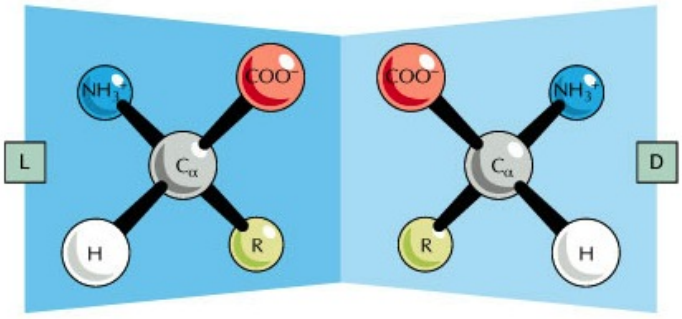


Figure 1: Schematic representation of the L- and D- isomers.

1.3 PRIMARY STRUCTURE: AMINO ACIDS AND PEPTIDE BOND

Amino acids (AA) are the fundamental elements, the building blocks, of proteins. They are organic compounds essentially formed by an amino (-NH_2) and a carboxyl (-COOH) group, bonded to a central carbon, named C_α. The same carbon atom is also linked to an hydrogen atom (H) and to a side-chain (R), specific for every different amino acid. The carbons forming the side chains are conventionally named with subsequent greek letter (β, γ, δ, ϵ,...).
The C_α is generally bonded to four different constituents, so forming a chiral center. In fact two isomers exist for each AA: L- and D- isomers (see figure 1). The two different isomers (called enantiomers) cannot be superimposed, the molecules are mirror images to each other. The amino acids found in natural proteins are all L-isomers. The Glycine is the only exception, because its side chain has only an hydrogen atom, implying that not all the four substituent are different, thus it is not chiral.

Another important feature of amino acids is their amphiprotic property, i.e. they can react both as acids and as bases, depending on the environment in which they are. With specific values of the (solution) pH, the carboxilic group ($-CO_2H$) can be deprotonated, becoming a negative carboxilates ($-CO_2^-$), and at the same time, the amino group ($-NH_2$) can be protonated, becoming an ammonium group ($^+NH_3-$). At neutral pH, the net charge of this molecular state is zero and the amino acid is in its zwitterionic state.
The amino acids are classified in five different groups, depending on specific properties of their side chains. In figure 2 all the amino acids are reported and the different groups are underlined [56]. One important characteristics is that amino acids with polar side chain are hydrophilic, while apolar side chains make amino acids hydrophobic. In table 2 four of the most important hydrophobicity scales are compared. AAs are named either with a three letter code or single letter code as in table 1.

Amino Acids bind to one another through the peptide bond. This is a chemical covalent bond, where the amino group of one AA reacts with the carboxyl group of the subsequent AA, releasing a water molecule (H_2O). This dehydration (or condensation reaction) leads to the formation of a bond between the carbon atom (C) of the first and the nitrogen atom (N) of the second amino acid (see figure 3, panel A). The lone pair of electrons on the N atom can delocalize,

Table 1: Amino acids codes

AA name	Abbreviation	Symbol	AA name	Abbreviation	Symbol
Glycine	Gly	G	Alanine	Ala	A
Proline	Pro	P	Valine	Val	V
Leucine	Leu	L	Isoleucine	Ile	I
Methionine	Met	M	Phenylalanine	Phe	F
Tyrosine	Tyr	Y	Tryptophan	Trp	W
Serine	Ser	S	Threonine	Thr	T
Cysteine	Cys	C	Aspargine	Asn	N
Glutamine	Gln	Q	Lysine	Lys	K
Histidine	His	H	Arginine	Arg	R
Aspartate	Asp	D	Glutamate	Glu	E

Table 2: Comparison of hydrophobic scales. On the top is reported the most hydrophobic amino acid. Scales in the second and in the fourth columns define the hydrophobic character as the tendency for a residue to be found inside of a protein, rather than on its surface. The other two scales are derived from the physicochemical properties of amino acid side chains.

Kyte and Doolittle [62]	Rose et al. [53]	Wolfenden et al.[10]	Janin [38]
Ile	Cys	Gly, Leu, Ile	Cys
Val		Val, Ala	Ile
	Phe,Ile		Val
Leu	Val	Phe	Leu, Phe
	Leu, Met, Trp	Cys	Met
Phe		Met	Ala, Gly, Trp
Cys			
Met, Ala	His	Thr, Ser	
	Tyr	Trp, Tyr	His, Ser
Gly	Ala		Thr
Thr, Ser	Gly	Pro	
Trp, Tyr	Thr		Tyr
Pro			Asn
		Asp, Lys, Gln	Asp
His	Ser	Glu, His	Gln, Glu
Asn, Gln	Pro, Arg	Asp	
Asp, Glu	Asn		
Lys	Gln, Asp, Glu		
			Arg
Arg	Lys	Arg	Lys

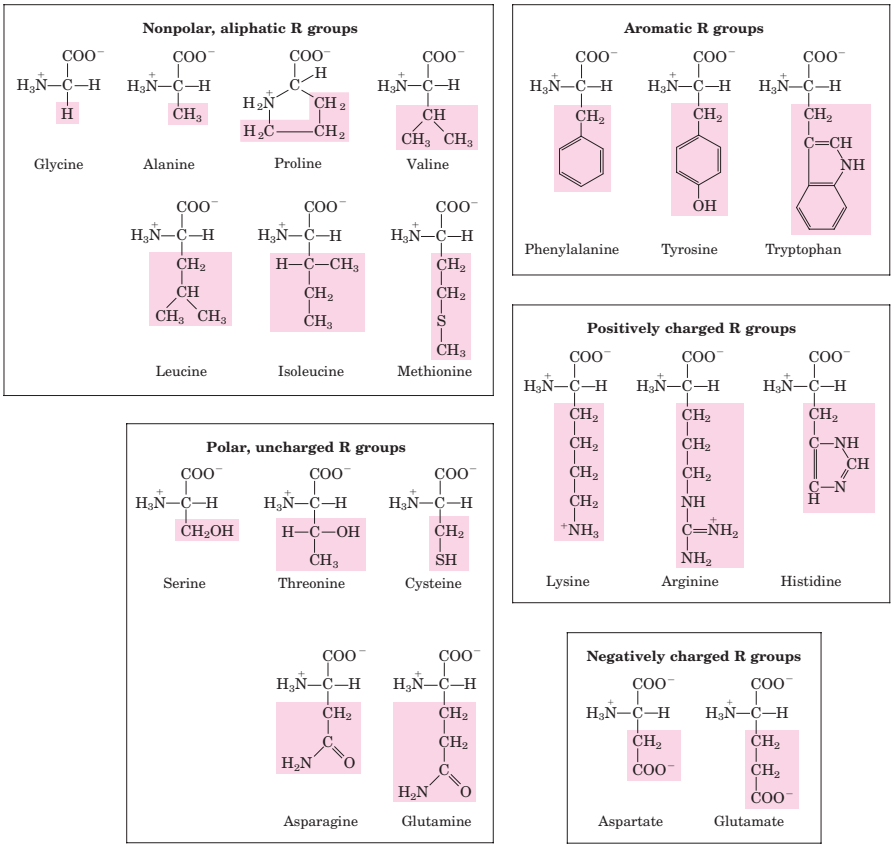


Figure 2: Natural Amino Acids structure. Structural formula of common amino acids, in the state of ionization that would predominate at pH 7.0. They are grouped into five main classes on the basis of the side chain (R-group, shaded portion). Non polar, aliphatic R groups are non-polar and hydrophobic, while aromatic side chains are relatively non-polar. Side chains of polar, uncharged amino acids are more soluble in water, or more hydrophilic, than those of the non-polar amino acids, because they contain functional groups that form hydrogen bonds with water. R-groups with positive or negative net charge are the most hydrophilic. [56]

giving to the group a partial character of double bond. It has in fact an intermediate length (1,32 Å) between the ordinary single C − N bond (1,45 Å) and the double C = N bond (1,25 Å). The main consequence of the partial character of double bond is that the peptide bond results rigid and planar. Rotation around it is not allowed, admitting in this way only two possible conformations of the atoms, related by a 180° angle.

In figure 3 (Panel B) the dihedral angle ω around the link is defined, it can assume only two values: 0° in cis conformation and 180° in trans conformation. This last is the most favourite, since in this arrangement the repulsion between atoms non bonded connected to the central C_α are minimized. The Proline is an exception because it has a ring side chain, so it is found with more probability than the other amino acids also in cis conformation.

The AA chain is called a polypeptide, which is characterized by the sequence in which AAs are connected, namely the primary structure. The sequences are conventionally reported and red from the N-terminus to the C-terminus, following the order in which they are synthesize by the ribosome. The poly[e[tide is then composed by two parts: the main chain and the side chains. The main chain is the backbone of the protein and maintains always the same composi-

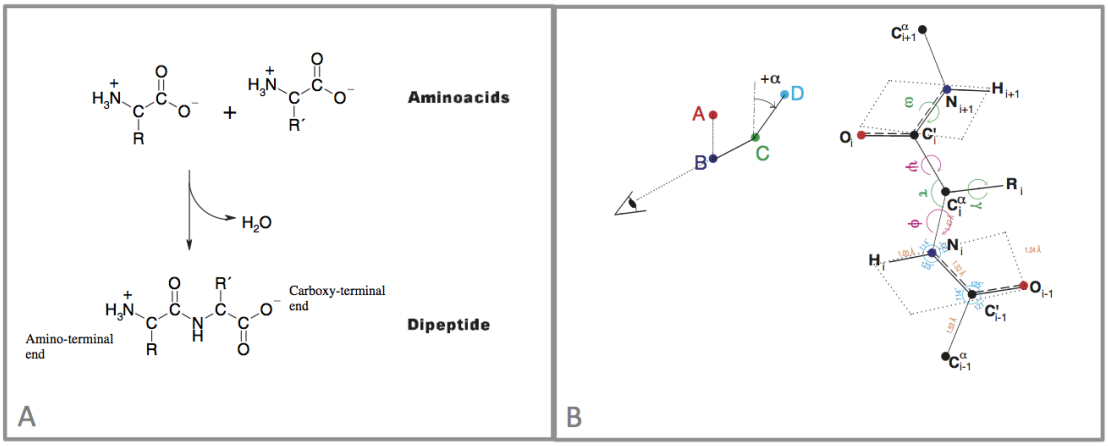


Figure 3: Panel A: dehydration reaction leading to the peptide bond. Panel B: Main angles and length defining a polypeptide chain. On the left: standard way to determine the dihedral angle of a bond (in this case between atoms B and C). The bond is oriented in the paper plane, so that neighbouring A and D atoms point upwards. Then measure the angle formed, clockwise is positive, anticlockwise negative. On the right: typical angles and length for a polypeptide chain. The dihedrals angle of the bonds between C′ ad N is fixed to 180° and N, H, C and O lie in the same plane. The bond angle of the C_α (τ) is $109,5°$. Also the dihedrals angles Φ and Ψ are showed, which determines the secondary structure. [20]

tion ($NH-C_\alpha-C'O$). The side chains are the parts depending on the sequence and specific to each single protein.

1.4 SECONDARY STRUCTURE

1.4.1 Ramachandran Plot and Hydrogen bond: How to Describe a Secondary Structure

Due to the rigidity of the peptide bond, the flexibility of the peptide chain can be imputed to the dihedral angles around the single bonds aside the central C_α. Namely, $C-N-C_\alpha-C$ defines the torsion angle Φ, while $N-C_\alpha-C-N$ defines the torsion angle Ψ, as it is shown in figure 3. The set of couples (Φ_i,Ψ_i) with i labeling the AAs along the chain uniquely determine the entire backbone conformation.
A convenient way to represent this information was first introduced by Ramachandran (1968, [81]), who reported these values in a 2D plot Φ vs Ψ (the Ramachandran Plot, RP). In figure 4 three Ramachandran plot are reported. It is clear that there are sterically forbidden (light yellow) and permitted areas (dark yellow up to black). The different occupied regions naturally conduct to the definition of different conformations of the peptide backbone. i.e. secondary structures. So each specific secondary structure corresponds to a values of (Φ,Ψ), occupying a restricted region of the (Φ,Ψ) plane.
The forces that stabilize the secondary structure are: steric hindrance between side chains of different amino acids; in some cases, electrostatic interactions between R, but, more than others, the intra backbone hydrogen bonds, i.e. those

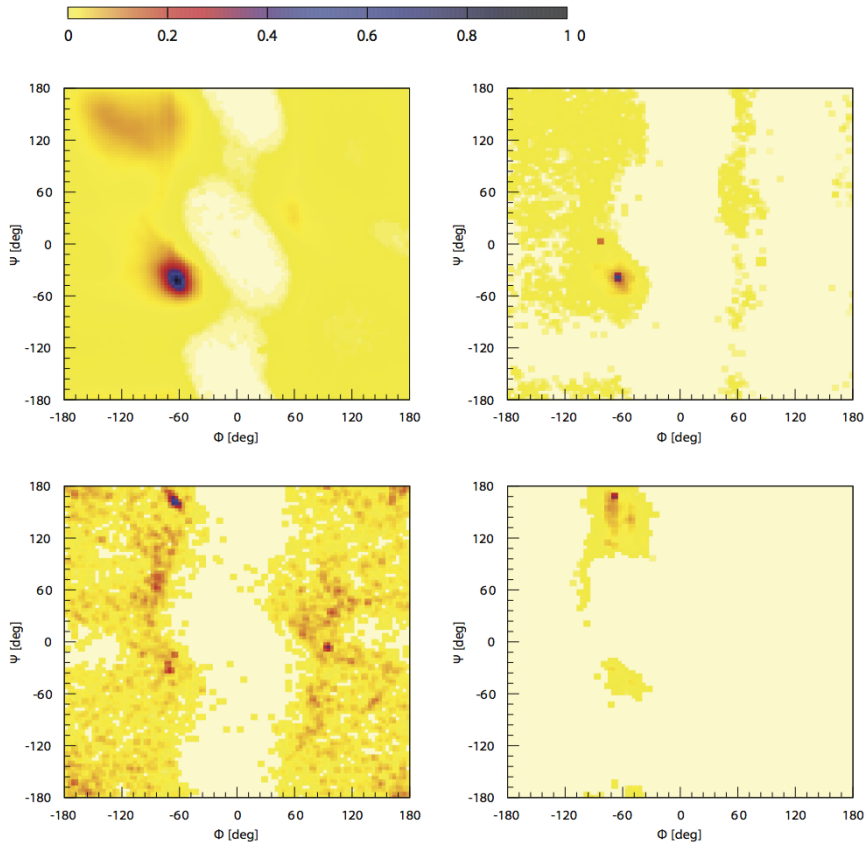


Figure 4: Ramachandran plots. The generic map (top left), the poly-Alanine map (top right), the poly-Glycine map (bottom left) and the poly-Proline map (bottom right). Maps are generated with SecStAnT (see Chapter 3), for a set composed by proteins solved with NMR. In the first map are clearly distinguished (red and blue) the two area corresponding to helical and extended conformations. Areas in light yellow are not permitted, while area in red and in blu are the most populated. The other three RPs are specific for single amino acids. The Glycine plot is symmetric and there are many conformations allowed. The poly-Proline instead, because of its rigidity, can assumes a little range of extended conformations.

having the NH backbone group as proton donor and C = O backbone group as H acceptor. So, the topology of the bond is a specific feature of different secondary structures and it is strictly related to the corresponding (Φ,Ψ) values.
In general, the hydrogen bond occur when one hydrogen (H) atom approaches some electronegative (electron attracting) atom, while it is chemically bonded (covalent bond) to another strong electronegative atom, like oxygen (O), nitrogen (N) or Fluorine (F) [59]. This interaction has an electromagnetic nature. In the protein backbone H is covalently bonded to the strongly electronegative N, which distorts the hydrogen electron cloud, attracting it. H acquires partial positive charge, while N a partially negative. Considering, at the same time, the double bond between C=O, here O is more electronegative than the carbon C, so there is a negative charge on it. The hydrogen bonding is manifested in the attractive interaction between the two partially charged atoms, H defined as donor and O defined as acceptor, $N-H\cdots O=C$.
The H-bond has strong directionality. Usually, the valence bond of the donor is directed at the acceptor atom to be involved in the hydrogen bond, while orientation of the acceptor group is less important. The H-bond energy is about

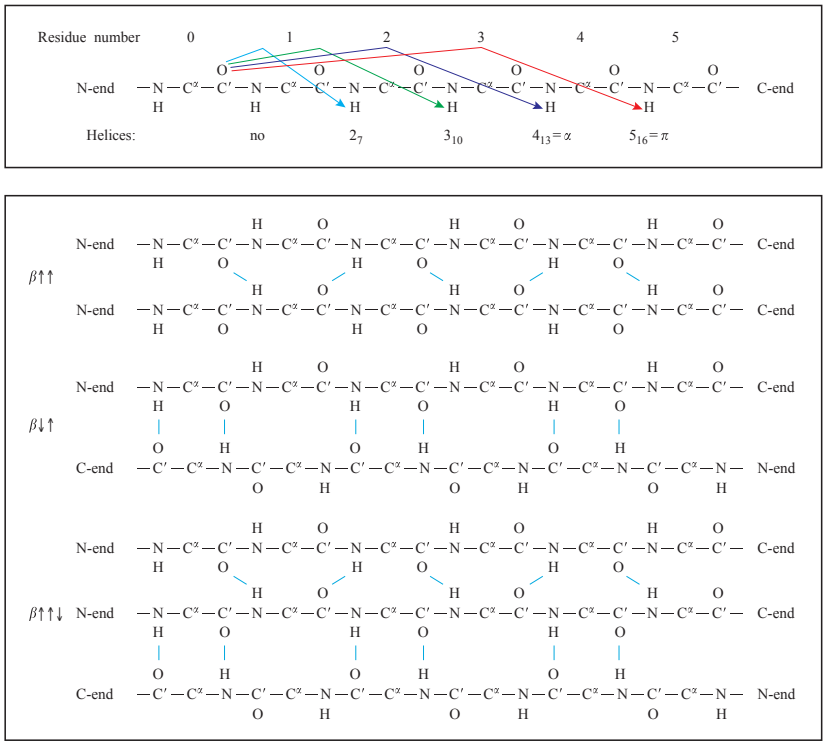


Figure 5: Hydrogen bonding patterns for secondary structures. Top box: helices are distinguished coloring the H-bonds in them. Bottom box: the H-bond pattern is given for three different class of β-sheets: parallel (β↑↑), antiparallel (β↑↓) and mixed (β↑↓) [59]

5 $kcal\ mol^{-1}$, although its variability range is large. In proteins, the intra-backbone H bonds along the same lead to its first fold in secondary structure.

1.4.2 Types of Secondary Structures

The secondary structure describes the local conformation of the amino acids in the protein chain. Different secondary structures are distinguished by regular arrangements of the main chain. There are three main classes of secondary structures: Helix, Sheets and Turns.
Helix can be right-handed (R) or left-handed (L), considering the positive orientation of their axis from N to C terminus. In the helices, the intra-backbone hydrogen bonding pattern is local and periodic, while in the sheets H-bonds can occur quite far in term of sequence. In turns are local but less regular. In figure 5, all the H-bonding patterns are schematically drawn, with different secondary structures put into evidence. The nomenclature of that bonds is usually described as n_m, where n is the number of amino acid residues per helical turn and m is the number of atoms involved in the cycle generated by the intramolecular H-bond.
The secondary structures are described in deeper detail in the following.

$3_{10}-Helix$

This structure was first proposed by Taylor in 1941 [6], ten years before the α-helix. It is characterized by three amino acids per turn and ten atoms in the

Table 3: Conformational parameters from the 3_{10}-helix. Φ,Ψ are the backbone dihedral angles. n is the number of residues per helical turn. d is the axial translation per residue. p is the pitch or axial translation for helical turn. The row (i,i+..) identifies how many residues divide the two amino acids hydrogen bonded.

	3_{10}-helix			
	Perutz [55]	Pauling et al. [34]	Crisma et al. [6]	Whitford [15]
Φ(°)	-49	-74	-57	-49
Ψ(°)	-26	-4	-30	-26
n	3.0	3.0	3.24	3.0
d (Å)	1.93	2.0	1.94	2.0
p (Å)	5.8	6.0	6.29	6.0
(i,i+...)	3	3	3	3

pseudo-ring formed by the intramolecular $C = 0 \cdots H - N$ hydrogen bond. In table 3, there are two different data for canonical helices, [55][34], that identify two different conformations with three residues for turn and they manifest themselves also in the same helix. These conformations are the consequence of the optimization of the van der Waals contacts and electrostatic interactions, under the constraint of keeping consecutive (i, i + 3) H-bonds [58]. However, this "ideal" conformation was never identified. Conversely, it is common to find irregular helices identified as 3_{10}-helices, which are increasingly irregular as the helix length increases.
Comparing in tables 3 and 4 the ideal parameters of 3.0_{10}-Helix and α-Helix, shows Φ,Ψ quite similar, falling in superposing regions of the Ramachadran map (figure 4). However, their intramolecular $C = 0 \cdots H - N$ H-bonding schemes are remarkably distinct: (i,i+3) for 3.0_{10}-Helix and (i,i+4) for α-Helix. The ternary 3.0_{10}-helix is more tightly bound and more elongated than α-Helix., as shown in figure 6
The experimental number of residue per turn (3.24) is intermediate between those of the theoretical 3.0_{10}-Helix and the (3.6_{13}) α-helix. In a perfect 3.0_{10}-helix the side chains on successive turns are exactly eclipsed, but the experimentally observed number of residues per turn does not superimpose side chains, inducing a slightly staggered disposition.
Finally the 3.0_{10}-helix is often found in proteins when short sequences fold into helical conformation or near a turn region. It is moreover found when a regular α-helix is distorted by the presence of unfavourable residues or to begin or end a regular α-helix. However, there is no forbidden region of the conformational space completely separating these two secondary structures. Thus, the α-helix may be gradually transformed into a 3.0_{10}-Helix (and vice versa) and, further, the main chain in an unfavorable conformation may slip into the other helix type. Thus, a role of the 3.0_{10}-helix as an intermediate in the mechanism of folding of α-helical proteins may be envisaged[6].

α-Helix

The 3.6_{13}-helix, aslo called α-helix, was first predicted, from theoretical studies, by Linus Pauling [34]. It is the most abundant secondary structure in proteins. Its Φ,Ψ values allow the backbone atoms to pack close together with few unfavourable contacts. With its large fractional number of residues per turn, it

Table 4: Conformational parameters from the α-helix. Definitions of the parameters as in table 3

	α-helix		
	Pauling et al. [34]	Crisma et al. [6]	Whitford [15]
Φ(°)	-58	-63	-57
Ψ(°)	-47	-42	-47
n	3.6	3.63	3.6
d (Å)	1.5	1.56	1.5
p (Å)	5.4	5.67	5.4
(i,i+...)	4	4	4

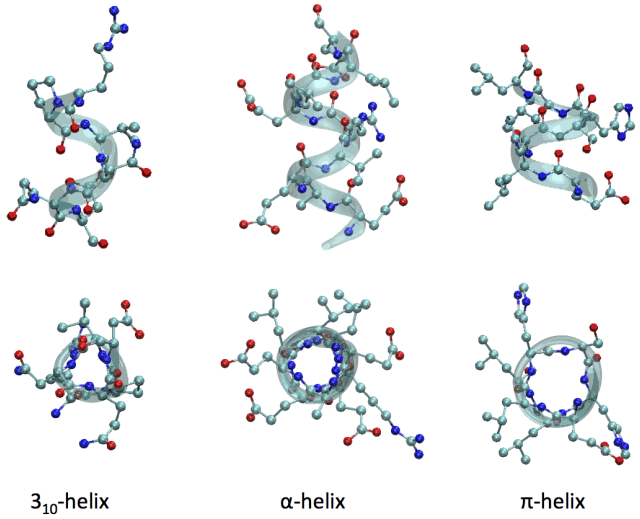


Figure 6: Different types of helices. Top: side view of the three different types of helices: α, 3_{10}, π. Bottom: top view of the same helices. Color codes: C atoms are cyan, N blue and O red. Hydrogen atoms are not solved. The backbone is highlighted with a ribbon.

requires two turns to position two side chains exactly one on top of the other on the same helical face. Moreover, they allow hydrogen bonding between the backbone carbonyl oxygen (acceptor) of one residue and the amide hydrogen of a residue four ahead in the polypeptide chain. The hydrogen bonds are 0,286 nm long [15] from oxygen to nitrogen atoms, linear and lie parallel to the helical axis.

The polarity of H-bonds and of peptide bonds and their alignment and periodicity give rise to a pronounced dipole moment in the α-helix. This dipole moment is also present in the other types of helices, but the α-helix are the longer and more stable, then this effect is particularly evident in this case.

Finally, it is worth mentioning that the first four NH groups and last four CO groups will normally lack backbone hydrogen bonds. For this reason, short helices often have distorted conformations and form alternative hydrogen bond patterns. In table 4 the values of Φ,Ψ reported in literature and other parameters are listed. In figure 6 it is shown the side view and the top view of this helix.

Table 5: Conformational parameters from the π-helix. Definitions of the parameters as in table 3

	π-helix		
	Low et al. [64]	Fodjie et al. [71]	Whitford [15]
Φ(°)	-57	-76	-57
Ψ(°)	-70	-41	-70
n	4.4	4.4	4.4
d (Å)	1.14	1.2	1.15
p (Å)	5.02	5.28	5.06
(i,i+...)	5	5	5

π-Helix

Initially hypothesized by Low and Baybutt in 1952 [64], this type of helix opens a thorny issue. It was first identified as conformationally unfavorable [81], then as rare, but with specific important functional roles [78] and finally as difficult to identify with standard algorithms [8][60], but more frequent than it is reported in literature [71].
The π-helix has hydrogen bonds formed between the CO and the NH groups separated by five residues (i,i+5). In a single turn it contains 16 atoms and 4.4 residues. They are mostly 5 residue in length. In figure 6 it is shown that different values for (Φ,Ψ) are identified as ideal, emphasizing the controversial issue of this secondary structure.
The conformation has been postulated to be less probable for three reasons [78]: first, it has unfavorable dihedral angles (Φ,Ψ); second, it has a 1 Å hole at the center of the helix creating a loss of van der Waals interactions; third, it needs to correctly align four residues to allow the collinear (i,i+5) hydrogen bond. This conformation has mostly be observed in the middle of α-helices. This fact has lead to a recent supposition [2] that naturally occurring π-helices are evolutionarily related to α-helices and that they derived from a single insertion of one amino acid in an α-helix. This insertion would lead to a conformational rearrangement to accommodate the residue in most, resulting in the formation of π-type H-bonding patterns from two until five, in some cases. The fact that π-helices are sandwiched in two α-helices is the principal reason of their difficult identification, since it is complex to identify defined boundaries between them. Finally it was observed [2],[78] that this specific secondary structure is conformationally ideal to absolve specific functional roles, as forming specialized binding sites within proteins.
Figure 6 reports the conformational parameters for different references. The three types of helices are shown in figure 6.

β-Sheet

The β-sheet was first described by Pauling and Corey [65]. This structure is composed by extended polypeptide chains, β-strands. This corresponds, in extended chains, to a rotation of 180° of subsequent planar peptide links, around the central chain axis. Therefore, the strand could be considered an helical arrangement, an extremely elongated form with two residues per turn. The side chains alternately project above and below the main plane of peptide bonds,

Table 6: Conformational parameters for β-sheets. Parallel sheets are distinguished from antiparallel ones. Definition of the parameters as in table 3

	Parallel β-strand		Antiparallel β-strand	
	Salemme [18]	Whitford [15]	Salemme [19]	Whitford
Φ(°)	-116	-119	-147	-139
Ψ(°)	112	+113	145	135
n	2	2	2	2
d (Å)	3.25	3.2	3.48	3.4

defining a second plane normal to the first. The backbone groups NH and CO are oriented approximatively orthogonal to the direction of the chain, available for interchain hydrogen bonding [48].
One individual β-strand is not very stable, because of the absence of stabilizing (intra-backbone) hydrogen bonds. Two or more of these chains can be brought together with 2.8Å hydrogen bonds between peptide groups of adjacent chains (see figure5). This results in a structure where the peptide groups are contiguously connected to create an array approximately planar of H-bonds and whose surfaces are occupied by side chain, projected orthogonal to the mean sheet plane. The ideal flat sheet can be seen as a regular, two-dimensional lattice, stabilized by covalent bonds in the direction of polypeptide chains and by H-bonds between or across the chains. The minimum energy configuration of this lattice originates from the simultaneous optimization of the conformational energies of the individual polypeptide chains and of the H-bonded interactions between them.
Depending on the relative sense amino (N) to carboxyl (C) of adjacent chains, two different sheet arrangements are distinguishable: parallel and antiparallel. Parallel sheets form when all the constituent chains have the same direction, while antiparallel when chains are oriented in opposite N to C direction, as it can be seen in figure 5. As it is clear from table 6, the parallel and antiparallel structure correspond to specific sets of (Φ,Ψ) backbone conformations, which are consistent with slightly different geometric requirements for interchain backbone H-bonding [23]. Furthermore, in parallel sheets, adjacent chains form an interconnected set of identical hydrogen-bonded rings, roughly trapezoidal in plan. Antiparallel sheets, on the other side, show a structure organized as a set of interconnected "small" and "large" H-bonded rings, so that each pair of rings translationally repeats along the chain axis direction.
Another possible configuration is composed by a mix of parallel and antiparallel sheet [19]. They in fact can be incorporated in most extended multiple strand sheets to form a mixed parallel/antiparallel sheet. In figure 7 β-sheets parallel antiparallel and mixed are shown.

The sheets of globular proteins are often found to twist in a right-handed directions, when viewed along the polypeptide chain axes. This global twist come from the assumption of the character locally left-handed helical of the single constituent chains. Because these twists, β-sheets typically conform to extended surfaces of complex curvature. The geometries observed reflect the equilibrium between two main compensated factors: the tendency of the polypeptide chain to twist to minimize the conformational energy and the necessity of preserving the interchain hydrogen-bonds, limiting the chain twist.

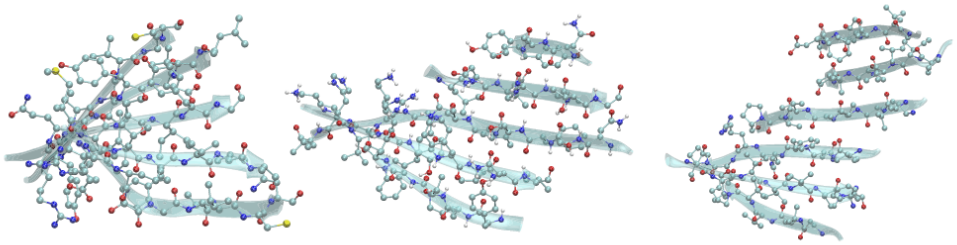


Figure 7: Sheet configurations. Three dispositions of strands in sheets: parallel, antiparallel and mixed. Color code as in figure 6

The conformational constraints imposed by the H-bonds network depend on the number and the disposition of the strands in the β-structure, but also on the parallel or antiparallel organization.
Parallel β-sheets are usually observed as multistrand arrangements and are found essentially in two geometric configurations, showed in B and C in figure 8. The observed structure is, in general, very regular.

The antiparallel β-sheets show instead a greater structural diversity, resulting from the fundamental differences between the two geometries. Antiparallel sheets are more conformationally flexible then parallel, admitting a large variety of twisted states, but preserving good interchain hydrogen bonds.
It is important to underline some relevant characteristics of the possible antiparallel configurations, all showed in figure 8. First, the polypeptide chain of a two-strand antiparallel sheet can twists preserving completely the integrity of interchain H-bonds. This is the reason for which the two strand structures are usually seen as composed by antiparallel strands. If a β-turn (see sec. 1.4.2) links the two strands, a β-hairpin is defined. Moreover, the antiparallel sheets have a pattern of H-bonds, which allows for large of flexibility to constituent strand. They in fact are able to incorporate coiled polypeptide chain, as in F in figure 8. Furthermore, particularly strong curvature radius changes are incompatible with regular H-bonding in a continuous sheet, but they can be accommodated by the introduction of a buldge residue (G in figure 8) in the β-structure.

Turns

Turns are weakly structured regions, which enable the backbone to change direction and eventually reverse back on itself. They are structural motif where the C_α of two residues separated by others (from one to five residues) are near, while other residues are not forming a regular element of secondary structure. They can be classified based on the number of residues in them, on the position of the H-bond and on the Φ,Ψ values of the corner residues. Each of them can be converted in its mirror image, changing the sign to all its dihedral angles. β-turns are the most common found type of these elements. They are four AA turns with two corner residues and a hydrogen bond between the first and the fourth residue. They were first identified by Venkatachala in 1968 [11], who defines them stereochemically. For maintaining the constraint of the formation

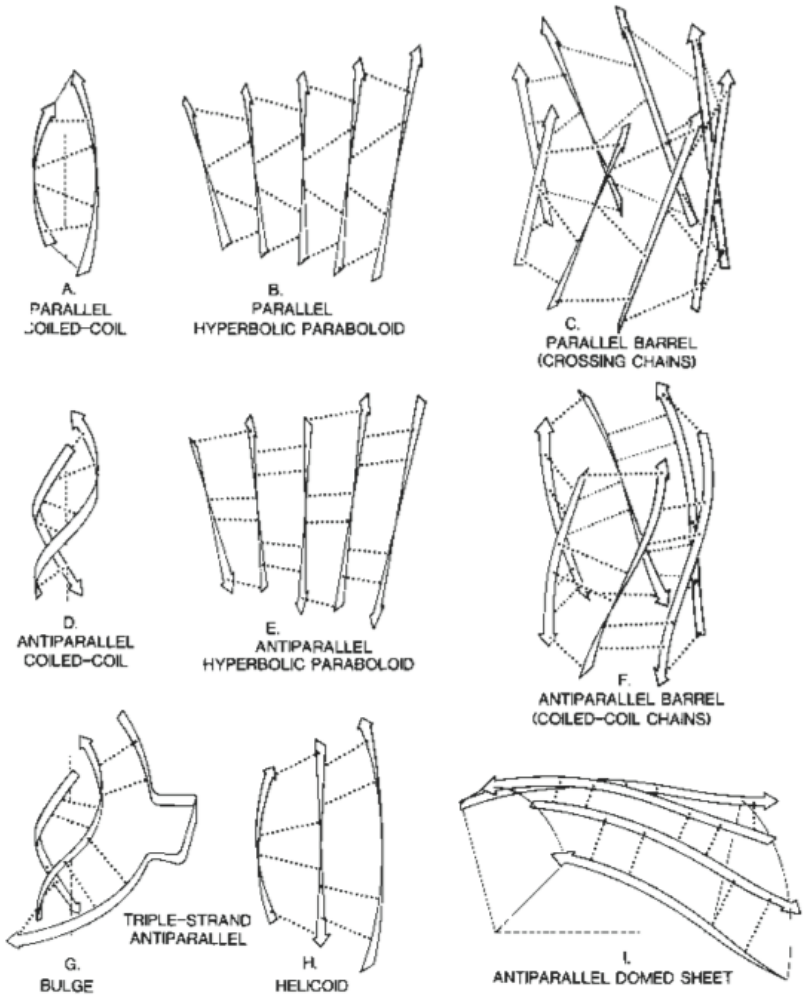


Figure 8: Geometric configurations of β-sheets [23]. Arrows indicate strands, while dotted lines H-bonds. Strands in sheets are usually linked by other elements of secondary structures, creating a superseconday element (see Sec. 1.5.3).

of the H-bond, the number of possible conformations is in fact limited to three main conformations [48]. Type I and Type II satisfy these stereochimical criteria, with all peptide bonds trans. These turns differ in the orientation of the peptide bond between the corner residues and then in the preferred side chain disposition there. In type I both the i+1 and the i+2 positions are occupied by L-residues, but preferentially Proline occupies the second place. They are the most prevalent in naturally occurring proteins. Type II has a Proline in i+1 and the i+2 position favors a Glycine, small polar L-residue, or a D-residue for steric reasons. Type III are single turns of 3_{10}-helix.
Turns are often sites of interactions among proteins, both because are topologically biased to occur on the surfaces of proteins and because their structure exposes side chains of corner residues optimally for molecular recognition.

Coils

Coils are associated to every amino-acidic sequence not folded in a specific secondary structure element, a chain without any distinct structure and any long-range order in the chain. They have extremely low density and large

volume. They are important as well as all other secondary structures because give to the chain the flexibility needed to form subsequent levels of folding, i.e. tertiary and quaternary structures.

1.4.3 Assigning the Secondary Structure

In the proteins structures included in the PDB, the identification of the secondary structures is normally performed by the author and included in the PDB file together with other information. However, several algorithms are available for the identification of protein secondary structure elements, given experimentally observed atomic coordinates of the protein. The two main algorithms are STRIDE [60] and DSSP [8].
DSSP uses a process of pattern recognition. The adopted method is the presence/absence of H-bonding between residues. Once identified the pattern of H-bonds, DSSP uses them to recognize the secondary structure. A detailed description of this algorithm is given in Appendix A.
STRIDE is a knowledge-based algorithm, which makes combined use of hydrogen bond energy and statistically derived backbone torsional angle information. The essential difference between these two methods is that the second uses mainly information about the geometry of the backbone, while the first uses only pattern recognition of hydrogen bonding network.

1.4.4 From Primary to Secondary Structure

Predicting the secondary structure from the sequence is a central issue for the comprehension of the protein global folding. One of the most popular algorithms for this aim has been proposed by Chou and Fasman on 1974 [28]. It is based on an analysis of amino acids secondary structures propensities. From the analysis of a set of 15 proteins [27], composed expecially by helical fragments, they obtained a set of parameters for α-helix (P_α) and for β-sheet (P_β) for the twenty naturally occurring amino acids, basically observing the probability of each AA of forming helices or sheets, and are related to their frequency of occurrrence of a given AA in a given secondary structure. The original parameters are reported in figure 9.
After Chou and Fasman many different studies on the amino acids propensities appeared. The more recent one is that of Fujiwara et al. in 2012 [50]. They analysed a bigger set of proteins and investigated if the amino acids propensities are affected by the type of AA neighboring along the sequence and the specific protein fold. The propensities were also calculated for exposed and buried sites, respectively. In figure 9 their parameters are reported. They found that the α-helix mean propensities have similar trends: buried or exposed sites do not influence so much the tendencies. On the other hand, mean propensities for exposed residues and buried residues for β-strand differs significantly and they depends strongly on the specific protein fold. So the β propensities are less transferable than the α ones.
From the propensities in figure 9, Chou and Fasman developed their algorithm to predict the secondary structure of a protein from the knowledge of its primary structure. They mainly consider these propensities, but the prediction is

Amino acid	α-helix			β-strand		
	Exposed residues	Buried residues	Total residues	Exposed residues	Buried residues	Total residues
V	0.83	0.89	0.91	2.31	1.57	2.00
I	0.96	1.01	1.04	2.02	1.39	1.79
L	1.16	1.27	1.28	1.18	0.93	1.15
M	1.03	1.29	1.26	1.01	0.84	1.01
P	0.48	0.41	0.44	0.49	0.42	0.40
A	1.43	1.37	1.41	0.48	0.72	0.75
C	0.63	0.85	0.85	1.24	1.07	1.36
F	0.88	0.99	1.00	1.50	1.10	1.4
Y	0.91	0.98	0.98	1.71	1.12	1.37
W	0.87	1.09	1.07	1.90	0.91	1.23
Q	1.34	1.21	1.26	0.96	0.82	0.72
S	0.74	0.80	0.76	0.86	0.85	0.81
T	0.72	0.84	0.78	1.58	1.08	1.21
N	0.74	0.77	0.73	0.71	0.76	0.63
H	0.90	0.85	0.87	1.15	0.98	0.99
D	0.91	0.73	0.82	0.61	0.76	0.55
K	1.25	1.13	1.17	1.14	0.98	0.76
E	1.51	1.25	1.39	0.89	0.86	0.65
R	1.31	1.13	1.21	1.27	0.82	0.85
G	0.28	0.59	0.44	0.41	0.81	0.67

Helical Residues[b]	P_α		β-Sheet Residues[c]	P_β	
$Glu^{(-)}$	1.53		Met	1.67	
Ala	1.45	H_α	Val	1.65	H_β
Leu	1.34		Ile	1.60	
$His^{(+)}$	1.24		Cys	1.30	
Met	1.20		Tyr	1.29	
Gln	1.17	h_α	Phe	1.28	
Trp	1.14		Gln	1.23	h_β
Val	1.14		Leu	1.22	
Phe	1.12		Thr	1.20	
$Lys^{(+)}$	1.07	I_α	Trp	1.19	
Ile	1.00		Ala	0.97	I_β
$Asp^{(-)}$	0.98		$Arg^{(+)}$	0.90	
Thr	0.82		Gly	0.81	i_β
Ser	0.79	i_α	$Asp^{(-)}$	0.80	
$Arg^{(+)}$	0.79		$Lys^{(+)}$	0.74	
Cys	0.77		Ser	0.72	
Asn	0.73	b_α	$His^{(+)}$	0.71	b_β
Tyr	0.61		Asn	0.65	
Pro	0.59	B_α	Pro	0.62	
Gly	0.53		$Glu^{(-)}$	0.26	B_β

Figure 9: Tables of conformational parameters: Fujiwara et al. [50] (on the left) Chou and Fasman [27] (on the right). Left: when $P_\alpha = 1$ the corresponding amino acids is contained equally in both the α-helical region and the protein. If $P_\alpha > 1$ the amino acid in question is more frequent in helical region than in the protein. The same is valid for β-strands. Right: helix former are identified by H_α strong, h_α normal, I_α weak. i_α defined indifferent helix former. B_α is strong and b_α normal helix breaker. The same code of capitals and small letters is valid for the sheet formers.

based also on non local properties, i.e. there are considered the propensities of groups of atoms not only consecutive in the primary sequence. The main problems of this algorithm are that the protein set analysed was really small and that the component of β-structures in it was particularly small.

From the Chou-Fasman algorithm a series of other more complex methods were developed and optimized. The GOR algorithm [3] predicts the secondary structure considering probability parameters derived from empiric studies of solved protein structure. More recent improvements are PHD [9] and PSIPRED [17], two of the most accurate secondary structure predictors, which use a neural network approach. Another possible way is the comparison between the foldings of the studied sequence and the sequence homologues, like in PROMETEUS [16].

One of the most recent and accurate algorithm is the BLAST-RT-RICO (Relaxed Threshold Rule Induction from Coverings) approach [67]. This method embeds a modified association rule learning approach and uses a multiple sequence alignment information, using the sequence as input for BLAST [5]. BLAST returns a list of proteins with significant sequence assignement. From this dataset proteins with known secondary structures are extracted and used to formulate a series of specific criteria, which enables to predict the possible secondary structure from the sequence analysed.

The Chou-Fasman algorithm is accurate to about 65%, while the BLAST-RT-RICO reaches the 88% of accuracy.

1.5 OTHER LEVELS OF ORGANIZATION: TERTIAY AND QUATERNARY STRUCTURE

1.5.1 Interactions stabilizing the global structure and the folding problem

A protein in its native state has to fold in the final three-dimensional conformation to be active. The fold arises from linking together secondary structures, forming a compact globular molecule. The stable form of a protein has clearly to establish more attractive than repulsive interactions. These forces are:

DISULFIDE BRIDGES They form strong covalent bonds between Cysteine side chains, often large separated in the primary sequence. They are broken only by high temperatures or acid pH.

THE HYDROPHOBIC EFFECT The hydrophobic residues of a protein in their linear state are in contact with water molecules of their solution. They are rejected by the solvent, resulting in an enhancement of interactions between non-polar molecules and the formation of hydrophobic clusters in water. Because the side chains of many residues are hydrophobic, this effect may contribute significantly to the intramolecular interactions.

CHARGE–CHARGE (ELECTROSTATIC) INTERACTIONS They occur between side chains of oppositely charged residues (Figure 2) as well as between NH_3^+ and COO^- at the ends of polypeptide chains. That charged residues are often found on the surfaces of proteins, where the interactions with water or solvent molecules really weaken these forces. They are described through the standard Coulomb potential.

HYROGEN BONDING Donor and acceptor groups of the backbone are generally completely occupied in forming intra-helical and intra-sheets H-bonds. However, side chains of several AA have donor or acceptor group, which can form H-bonds among each other or with the residual free groups of the backbone and stabilize the tertiary fold. Particularly important to the formation of this bond are side chains with hydroxil group of Thyrosine, Threonine, Serine and side chains with amide group of Glutamine and Aspargine. Frequently, atoms of side chains are hydrogen bonded to the water molecule trapped inside proteins. Other times, they appear shared between two donor and acceptor groups, in this case these bonds are defined bifurcated H-bonds.

VAN DER WAALS INTERACTIONS These forces including an attractive (dispersive) tail and a repulsive core, appear between adjacent, uncharged and non-bonded atoms and arise from the dipole induction due to fluctuations in the charge density of atoms. Although van der Waals forces are extremely weak, their large number arranged close together in proteins make these interactions significant to the maintenance of the folded state.

The folding process is extremely fast, usually completed in less than one nanosecond. This is surprising because the final conformations that a linear chain could reach are an astronomical number, Levinthal supposed 10^{300}, for the big variety of degree of freedom of a polypeptide chain. If a protein would fold trying

different conformations, it would be necessary a time longer of the estimated age of the universe, also folding a conformation for nanosecond. This paradox, "the Levinthal paradox", is partially solved observing that the fold is a stepwise process, in which different parts of the protein reach at the same time their configuration of energy minimum.

1.5.2 Tertiary Structure

The tertiary structure is the overall topology formed by the single polypeptide chain. It is defined as the organization of secondary structures in a more complex folds through the interactions between residues often largely distant in the primary sequence.
For little globular proteins of ~ 150 residues or less, the fold is a compact spheric molecule, composed by structural motif of secondary structure with little irregular structure. For proteins with more than ~ 150 residues, the tertiary structure may be organized around more than one structurate units, called domains. They can have different folds and are frequently linked together by extended regions, relatively destructurated, of the polypeptide.
A rigorous definition of protein domain does not exist. An acceptable one is the presence of an autonomously folding unit within a protein or a region of a protein showing structurally homology to others proteins. The recognition that proteins shows similar tertiary structures lead to the concept of structural homology and proteins can be grouped in related families. There are three main classes of domains: consistent mainly of α-helices, β-strands and domains that are mixed containing both.
In figure 10 all the hierachical levels of a folded proteins are showed and, in the third image, two different domains are underlined. It is important to note that domains arise from folding of a single polypeptide chain and this is the main difference between them and an element of quaternary structure.

1.5.3 Super-Secondary Structure

Discriminating secondary from super-secondary or super-secondary and tertiary structure is not always easy. Super-secondary structure could approximately be defined as an intermediate level of organization, which reflects groups of secondary structural elements but does not encompass all of the structural domain or tertiary fold.
In section 1.4.2 there is a general description of possible sheet conformation. The definition of these conformations as secondary or super-secondary structures is difficult. It may be said that a sheet composed by strands and other defined elements of seconday structures is a super-secondary motif.
Some super-secondary elements recognized are: β-α-β motif, four helix bundles, Greek-key motif and its variants and the β-meander. The β-meander motif is a series of antiparallel strands linked by a number of turns. The Greek motif is a variation of the β-meander, where four (or more) antiparallel strands are bonded in a way that the third and the fourth strands are externally placed. Elements of super-secondary structure are frequently used to allow protein domains to be classified by their structure.

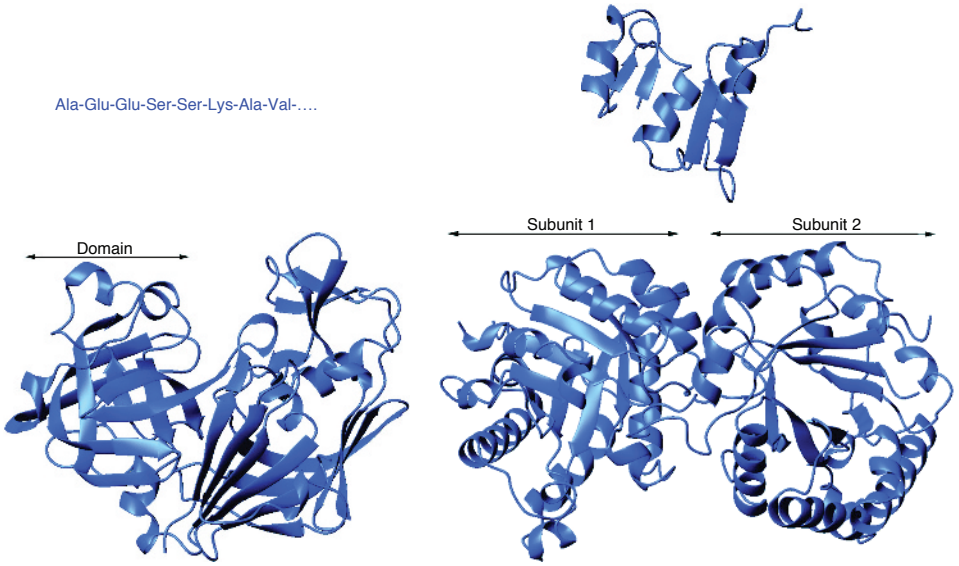


Figure 10: Hierarchical organization of a folded protein: primary sequence (top left); secondary structure, composed by helical, sheets and destructurated parts (top right); tertiary structure, divided into domains (bottom left); quaternary structure, with explicitly defined the constituent subunit (bottom right).[15]

1.5.4 Quaternary Structure

The quaternary structure is defined by the interaction between different polypeptide chains, making up the protein. The interactions are that responsible for the tertiary structure, like disulfide bonds and hydrogen bonds, with the exception that in this case they stabilize two or more chains.
Haemoglobin is the classic example of a protein with quaternary structure. It is a tetramer containing two different subunits. The correct functional activity requires a specific association of subunits. These subunits are kept together by non covalent weak interactions. Although individually weak, these interactions are large in number and lead to subunit assembly as well as gains in stability.
The quaternary structure characterizes a large number proteins and allows the formation of catalytic or binding sites at the interface between subunits, which are usually impossible for monomeric proteins. Furthermore, oligomeric proteins are generally more flexible, since they allow substrate induced conformational changes in the assembly and, as a consequence, opportunities for regulating the biological activity.
In figure 10 all the four different levels of organization of proteins are shown. This is an example of the complexity of the living system organization.

2 MODELING OF PROTEINS: THE COARSE GRAINED APPROACHES

After the description of protein system in the previous Chapter, the simulation tool is here introduced. Particular attention is given to the description of the Coarse Grained (CG) models and a brief review of the state-of-art on the main models in this class is reported. The fundamental concepts of CG simulation set up and running are then illustrated.

2.1 INTRODUCTION TO SIMULATIONS

As previously discussed, proteins are structurally very complex and processes involving them span a range of about ten orders of magnitude in the space domain and fifteen in the time domain. Simulation is a powerful tool for dealing with this kind of complexity: depending on the level of detail represented into the model, different scales can be addressed, as it will be clear in the following.

To set up a simulation, there are three main point to be addressed: definition of the degrees of freedom (DOF), definition of the potential energy of the system, exploration of the phase space, i.e. sampling of the allowed conformations. To choose the degree of freedom describing the system requires to set the resolution level at which to analyse it [84]. The highest level of resolution is the explicit treatment of the DOFs both atomic and electronic to describe chemical reactions. The Born–Oppenheimer approximation is usually assumed [39]: at any time, the Schrödinger equation for the electrons is solved in the external field generated by the atomic nuclei considered as frozen. Then, the effective electronic-structure-dependent potential energy functions determine the dynamics of the nuclei. Different quantum mechanic approaches (QM) model the electron-electron interactions in different ways. Some examples are Hartree-Fock theory and the Density functional theory [84]. With these approaches, it is possible to model small molecules (up to a few tens of atoms, $0.1 - 1$ nm) for maximum run lengths of 100 ps to 1 ns.

To extend the size of the addressable systems, the molecular mechanic (MM) scheme has to be adopted. In MM the electronic DOFs are implicitly treated and an empirical description of the inter-atomic forces (Force Field, FF) is adopted. The parameters of the FF are fitted on the QM potential energy functions evaluated in small molecules, which represent parts of the system, and experimental structural and thermodynamical data are also included. These methods, even with the computation power currently available, are not able to reach biologically interesting scales for a sufficient time. Big macromolecular assemblies are now reachable up to the $n$sec timescale: the μsec timescale is the current limit for simulations of single proteins.

The Coarse Grained (CG) [83] approach, i.e. reducing the number of internal variables used to describe the system, is an option to move beyond the above

limits. The level of CG can be modulated: the coarser the representation, the larger the saving in computational cost [82]. However, the elimination of internal degree of freedom implies that their effect must be implicitly considered in the effective forces acting between the explicit DOFs. This task becomes harder as the level of coarse graining is made stronger. Different recipes for the parameterization of the FFs were given, making the CG models landscape very complex. A critical issue is the combination of accuracy and predictive power in the FF. Accuracy is the ability to reproduce the native structure of the protein analysed, while the transferability/predictive power is the ability to describe the general dynamics of systems with different compositions and different configurations.
To further lower the level of resolution, the mesoscale models represent entire proteins or domains with single interacting centers. The purpose is the reproduction of only the slow dynamics of the system.
The highest level of reduction of degree of freedom is the continuum representation. The system is represented as a medium, while dynamics is described by a single functional variable depending on the spatial coordinates [84].
The different levels of system description can then be combined in multiscale approaches.
As described in the previous Chapter 1, proteins are system of various size. The model developed in this work is aimed at the reproduction of protein secondary structures with the less possible computational cost. This Chapter focuses on the description of the coarse grained models that implement the highest level of CG, still allowing an explicit description of the DOFs responsible for secondary structures. These are also called "minimalist models".

2.2 DEFINING THE DEGREES OF FREEDOM

Coarse graining means reducing the internal degrees of freedom of the system, describing a set of atoms with a single interacting center: a bead. This can be done in several different ways [26]. Formally, the CG procedure is described by:

$$Q_I = Q_I(\{r_i\} \in B_I) = \sum_{r_i \in B_i} M_{Ii} r_i \tag{4}$$

where $\{Q_i\}$ is the set of internal coordinates within the CG representation, $\{r_i\}$ is the set of the cartesian coordinates within the atomistic representation, B_I represents the set of atoms belonging to a given bead and defines the level of coarse graining. M_{Ii} is a rectangular matrix $n \times m$, with n total number of atoms and m total number of beads. The second equality implies the linear dependence of the CG coordinate on the AA coordinate. Linearity does not generally hold, but it generally it does when considering the cartesian CG coordinates [86].
Different definitions of B_I may imply different levels of CG resolutions. The 4-6 beads models (see figure 11) describe a single amino acid with four to six interacting center [54],[74],[75] . These are normally considered the least coarse representations. The positions of the beads depend on the specific model. It is possible to use more than one bead to describe the backbone and more than one bead to describe the side chain. The Ramachandran plots are always traceable.

Side chains are explicitly represented. This allows an easier representation of the side chain effects and simplifies the functional forms of the correspondent FF terms, although obviously their number increases.
The two bead models [13],[4],[63] use one bead to describe the backbone and one bead to describe the side chain of each amino acid of the polypeptide. It is no more possible to explicitly represent bonds, angles and dihedral angles of the backbone. The polypeptide conformation is described by pseudo-bonds, pseudo-angles and pseudo-dihedrals to describe the polypeptide conformation (see figure 11, angles θ,ϕ).
A further level of coarse graining (that considered in this Thesis) are the One Bead (OB) models. Here, a single interacting center for the whole amino acid is used. These models have some important advantages. First, this is the coarser level at which the different secondary structures can be explicitly defined. The demonstration of this statement for the specific class of the C_α based models is one of the issues addressed in this Thesis. This is an important point because many biological processes, occurring in and between proteins, involve a transition in the secondary structure as triggering step. Second, with respect to the all atom models or less levels of CG, they result in a significative saving in computational cost and an easy implementation of the model (which becomes a linear chain), still maintaining the possibility to include amino acid dependent terms to increase predictivity and transferability. Third, their resolution level matches exactly with the structural data at low resolution, allowing for a direct data exchange with those experiments. Finally, they are the most natural representation of a protein chain, because the amino acid is the building block of every protein.
On the other hand, with the OB CG approach is rather hard to reach a satisfactory accuracy in describing the local interactions. It is very difficult to include highly specific and strongly directional interactions (e.g H-bonds and side chains conformational flexibility effects), in the few parameters available, in a both predictive and transferable way.
Within the OB CG class it is possible to choose where place the center of the bead, e.g. on the center of mass of the residue, on its geometric center or on a specific atomic position. Choosing the C_α as center of the bead brings the additional advantage of a direct description of the secondary structure with pseudo-bonds and pseudo-dihedrals conformational internal variables. Moreover, this opens the opportunity of reverse mapping to the all-atom description, as it will be shown in the following.
The number of internal degree of freedom in this resolution level is extremely reduced. As showed in figure 11, only θ_i and ϕ_i describe the conformation, because the C_α-C_α pseudo-bond distance has a fixed value of 3.8 Å (for the trans peptide bond), independently from the specific secondary structure. The general form for the FF in these OB models can be cast in the general form:

$$\begin{aligned} U &= U^{bond}(\{r_{i,i+1}\}) + U^{back}(\{\theta_i, \phi_i\}) + U^{nb}(\{r_{i,j}\}) \\ &= \sum_i u^{bond}(r_{i,i+1}) + \sum_i u^{back}(\theta_i, \phi_i) + \sum_{i<j} u^{nb}(r_{i,j}) \end{aligned} \tag{5}$$

Here, $r_{i,j}$ is the distance between beads i and j, U^{bond} is the term describing the pseudo-peptide bond energy, often replaced with a sum of constraints; U^{back} describes the conformational energy and U^{nb} the non bonded interactions, i.e.

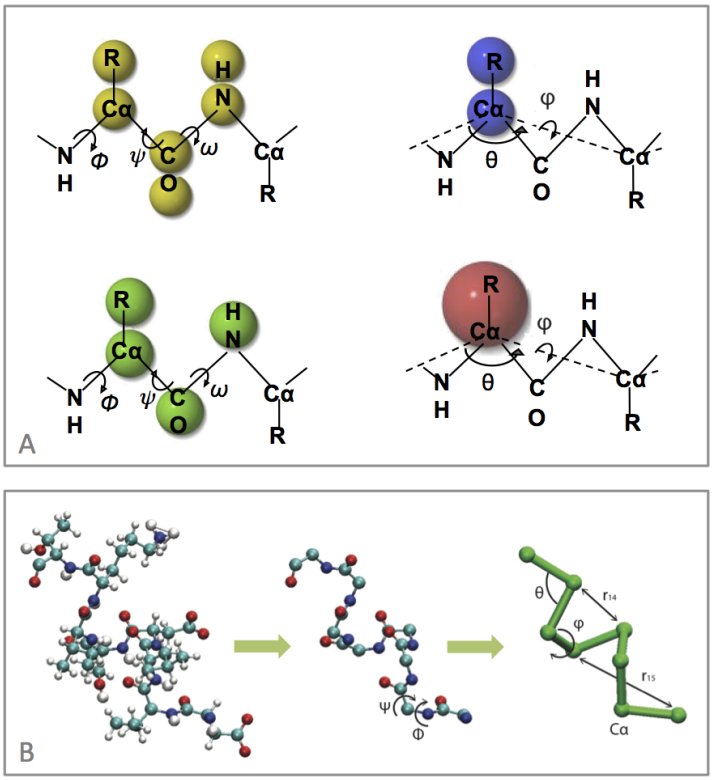


Figure 11: Levels and procedure of coarse graining. Panel A: different levels of coarse graining (C_α). Top left: six beads. Top right: two beads. Bottom left: four beads. Bottom right: one bead. The internal angles and dihedrals describing the conformation of the chain for each level are shown. For the two beads and the one bead the pseudo bonds and dihedrals are reported. Panel B: graphical explanation of the procedure of coarse graining. Left: all atom fragment of a protein. Center: backbone of the same fragment, with the dihedral conformational angles Φ,Ψ marked. Right: OB C_α-based model of the same fragment, with explicit θ and ϕ (conformational variables in this representation). Example of distances between beads not consecutive along the chain are also given, as r_{14} or r_{15}. The distance between subsequent C_α is constrained.

the interactions between beads not subsequent in the amino-acidic chain. This last term has to include many different effects: H-bonding, excluded volume and hydrophobicity interaction, electrostatics. For this reason it may be very complex and separated in subterms describing every effect. The different definitions of this term are specific for different models.

2.3 TOPOLOGICAL CONNECTIVITY

As previously introduced, in the one bead C_α-based models two subsequent beads are bonded with a term mimicking the peptide bond. This term could be harmonic, but also a constraint.

A good choice for the U^{nb} term is an hard task. It in fact has to account for many different interactions. Its exact definition is not clear and its implementation is model dependent.

A possible way could be to separate it on physical basis. It would then become

a sum of terms each specific for a different interaction, i.e. hydrogen bonds, electrostatics and so on. Unfortunately, it is not easy to a priori distinguish these interactions. Furthermore, it is very difficult to represent interactions as the hydrogen bond, which is local and directional, in an easy way.
An alternative strategy often adopted (in the so called "partially biased models"[42]) is to separate U^{nb} in two contributions, one for the local interactions and one for the non-local ones. The local part of the potential is then treated as topologically connected, also if it would not be, since in this way it is easier to insert a structural bias towards a reference structure and model those complex interactions.
The distinction between local and non-local part depends on the specific model. The most intuitive way is on the basis of a cutoff radius. This is the strategy adopted by the network models [57] and by the Gō models [31], for example. Here, all the local interactions, including the conformational terms, are topological connections, represented with harmonic bonds, whose parameterization is based on a single reference structure. The non-local interactions are generally represented as weakly attractive, repulsive or null [83].
Another possible way is to define the local interaction only between specific sites of the amino acidic chain. For example, the hydrogen bond could be placed between every bead i and the one four after along the chain, $(i, i+4)$, as in [41].
The ideal model should have only the conformational terms topologically biased and all the non covalent interactions totally unbiased. Models with a lot of topological bias are highly accurate, but specific for the specific structure, so they are poorly predictive and transferable. The less is the topological bias, the more the model is predictive, but the structural accuracy decreases. The goal of every model is to impose the topology to the system that realize the better compromise between predictive power and structural accuracy.

2.4 FORCE FIELD AND PARAMETERIZATION STRATEGIES

Once the number and the location of the beads are chosen, the next step is the choice of the functional forms for the terms of the Force Field (FF). The generic One Bead C_α-based FF can be finally written as :

$$U = U^{bond}(\{r_{i,i+1}\}) + U^{back}(\{\theta_i, \phi_i\}) + U^{nb,loc}(\{r_{i,j}\}) + U^{nb,non-loc}(\{r_{i,j}\}) \tag{6}$$

How many terms, and with which functional forms, are present in the FF is model dependent [83]. In table 12 a survey of the FF terms of the most popular minimalist models is reported. Once decided which terms have to be explicitly present, it is necessary to choose an analytic representation for each of them. Therefore, building the FF requires the optimization of a set of parameters in order to represent in the best way the real interactions between amino acids. Chosen a specific property of the system, an objective function describing it is individuated. For example the energy of the real system or specific geometric properties of the system. Then, the optimization phase follows to search for the parameters set for the FF terms which minimizes the difference between the value of this objective function (obtained by the simulations) and a reference (target) value. The parameterization optimization is thus defined by the chosen

Models	u^b	u^θ	u^ϕ	u^{Hb}	u^{Hyd}	u^{el}	Remarks
Elastic Networks (ANM) [35]			Harmonic: $\frac{1}{2}k\left(r_{i,j}-r_{i,j}^0\right)^2$				$r_{ij} < r_{cut}$ r_{cut} = 8-15 Å k ≈ 10-0.9 kcal/mol Å^2 $r^0_{i,j}$ and r_{cut} from reference structure
Gō-Models [31]	Harmonic or constraint	Harmonic Angle: $\frac{1}{2}k_\theta(\theta-\theta_0)^2$	Cosine Sum: $\sum_{n=1,3}K_n[1-\cos n(\phi-\phi_0)]$	L-J (12-10): $\varepsilon\left[\left(\frac{r^0_{i,j}}{r_{i,j}}\right)^{12}-\left(\frac{r^0_{i,j}}{r_{i,j}}\right)^{10}\right]$		Repulsive only: $\varepsilon\left(\frac{r^0_{i,j}}{r_{i,j}}\right)^{12}$	u^{hyd} is partially local (LJ) and partially non local (rep only) r_{cut} = 8 Å; ε = energy unit k = 100ε ; k_θ = 20ε ; K_n ≈ε structural parameters from reference structure
DMC [7]		As in Gō-models		$u^{nb}=\sum_{j-i>3}\varepsilon(c_i,c_j)\left[5\left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{12}-\delta(c_i,c_j)6\left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{10}\right]$			c_i,c_j = amino acid type $\sigma_{i,j}$ = extracted with BI, for 3 sets: |j-i| = 4; |j-i| = 5; |j-i| > 5 $\varepsilon(c_i,c_j)$= positive,energy unit $\delta(c_i,c_j)$=0(repulsive),1(attractive).
OB generic [41]	Constraint	Double-well $k_\theta\frac{1}{2}(\theta-\theta_0)^2+k'_\theta\frac{1}{3}(\theta-\theta_0)^3+k''_\theta\frac{1}{4}(\theta-\theta_0)^4$	Cosine sum: $A[1+\cos\phi]+B[1+\cos(3\phi)]+C\left[1+\cos\left(\phi+\frac{\pi}{4}\right)\right]+D[1+\cos(2\phi)]$	Morse potential: $u^{nb}=\varepsilon\left[\left(1-e^{-\alpha(r-\sigma)}\right)^2-1\right]$			$\theta_0 \approx 90°$; other u^θ param. determine position of the 2nd min. and the relative high of the two minima. Param in u^ϕ depend on specific sec struct. σ = 6.1 Å =bead diameter;α =0.7 Å^{-1}; ε = energy unit
Partially Biased [40,42]	Constraint	Double-well	Harmonic angle: $\frac{1}{2}k_\phi(\phi-\phi_0)$	$u^{nb,loc}$ + $u^{nb,non-loc}$ = Morse: $u^{nb}=\varepsilon\left[\left(1-e^{-\alpha(r-r_0)}\right)^2-1\right]$			ϕ_0 = 60°-180°;k_ϕ= 5 kcal/mol rad^2 $u^{nb,loc}$:ε = 6 kcal/mol; r_0 = 2.8 Å;α = 0.7 Å^{-1} $u^{nb,non-loc}$: ε = 0.202 kcal/mol; r_0 = 9.5 Å;α = 0.7 A^{-1} $r_{cut}(u^{nb,loc}$ vs $u^{nb,non-loc})$= 8 Å
YFSH [79]	Constraint	Harmonic angle	Cosine sum as Rocchia	$\sum_{hb}-\varepsilon_{hb}\exp(-(r_{i,j}-r_{hb})^2/\sigma^2_{hb})\exp[(\lvert r_{i,j}\cdot t_i\rvert-1)/\sigma^2_{hb}]\exp[(\lvert r_{i,j}\cdot t_j\rvert-1/\sigma^2_{hb})]$	$u^{nb}=\sum_{j-i>3}4\varepsilon S_1\left[\left(\frac{\sigma}{r_{i,j}}\right)^{12}-S_2\left(\frac{\sigma}{r_{i,j}}\right)^6\right]$		ε=enery unit ; k_θ=20ε/rad^2 ; θ_0=95°(helix),105°(otherwise) u^Φ, param spefic for sec str (phys-chem based) ε_{hb}=0.7ε ; r_{hb} =1.35 (helix),1.25(sheet) ; σ_{hb} = 0.5; t_i and t_j are ortoghonal vectors to (i-1,i,i+1) and (j-1,j,j+1) planes σ=1.16 ; S_1,S_2 = dependent on the type of interactions (bead type)
Alemani et al [21]	Harmonic	Double-well as Tozzini(2005-2007)	Cosine: $[1+\cos(n\phi+\delta)]$	Dipole-Dipole $\frac{\vec{\mu}_i\cdot\vec{\mu}_j}{\vec{r}^3}-\frac{3}{\vec{r}^5}(\vec{\mu}_i\cdot\vec{r})(\vec{\mu}_j\cdot\vec{r})$	Van der Waals interactions as in all atom FF		There is also a $V_{corr}(\phi_i,\phi_{i+1})=k_{corr}(\phi_i-\phi_{i+1})^2$ with k_{corr} > 0.4(helix) and = 0.6(sheet). μ_i and μ_j =dipole formed by triplets (i-1,i,i+1) and (j-1,j,j+1). r is the vector connecting them. This interaction accounts also for some of the electrostatics.

Figure 12: Summary of the minimalist models Force Fields. Model specific functional forms for each FF term are showed and the specific parameters used are given in the remarks column. Functional forms shared by more than one model are explicitly given only once together with correspondent parameters.

objective function, the source of the reference values and the algorithm used for the minimization.
In the following subsections the most important parameterization methodologies used for the minimalist models are reported. All these methods are often synergetically used to parameterize different terms of the same force field.

2.4.1 Single Structure Based

The most intuitive way to parameterize a FF term is to completely bias its structural parameters towards a reference structure, usually experimental. This strategy is used for instance in Elastic Network models (EN) [57] and Gō-models [31]. Here, the separation in local and non local of U^{nb} is based on a cutoff radius r_{cut}. If $r_{i,j} < r_{cut}$ in the reference structure, then the beads i and j will interact locally, while if $r_{i,j} > r_{cut}$, then the interaction between i and j will be considered non-local. The force field terms derived are treated in different manner, i.e. with different functional forms and parameters.
In the EN every interaction is described with a spring, whose elastic constant k in the original model [57] was fixed at the same value for every i and j. Two beads are interacting only if $r_{i,j} < r_{cut}$ and equilibrium distance is assigned at the values taken from the equilibrium structure. The non local interaction is null. The force field is composed only by a term $U^{nb,loc}$, which in part substitute conformational terms (θ,ϕ).
These extremely simple models are able to describe with good accuracy the slow fundamental dynamics of the analysed system [35]. These motions are those related to the biologic function, in fact EN models were used to analyse the equilibrium dynamic of huge systems like entire viruses [36].
This basic type of network admitted the presence of a single equilibrium configuration, reached with the simpler possible interactions (single well harmonic potential), though somehow unphysical. The subsequent models of this type tried to overcome these and other weaknesses [51],[24].
The Gō-models [31] have a parameterization procedure similar to EN, although with more complex functional forms and consequently a larger number of parameters. The conformational terms are here explicitly present (see table 12). The r_{cut} has the meaning of dividing the couples of amino acids that are in native contacts (in the folded structure), from that there are not in contact. The Gō-model was initially defined the minimally frustrated model for folding, because the local free energy minima different from the native structure were almost absent. The evolutions of this model try to include frustration and so the reproduction of intermediate states important for the complete folding. [52].
The single structure based results in high structural accuracy by definition and produce FFs able to describe the quasi-equilibrium dynamics or the global folding kinetics. The main problem of these models is that they are not predictive: they are inadequate to describe more general dynamical properties or structures different from the reference one.

2.4.2 Statistical Potentials and Statistics Based Parameterization

A possible parameterization strategy, which makes the FF terms more transferable and predictive, is to extract the parameters from a statistical set of structures. The parameters would be no more evaluated from a single reference structure. Conversely they would reflect the common features of all proteins in the dataset. So, the larger is the set, the more the parameters are transferable. The composition and source of data may influence the features of the force field.
How can one extract the statistical information included in a dataset? A procedure commonly adopted is the Boltzmann Inversion (BI), with the subsequent generation of Statistical Potentials (SP), also called knowledge based potentials. In the following, the justification historically used for these methodologies will be analysed.
Theoretically, the essential hypothesis assumed as true at the base of the BI are completeness of the FF, i.e. the sum of the terms describes exhaustively the interactions of the system, and orthogonality, i.e. no correlations between FF terms. The total energy of the system is so decomposable as sum of single internal variable terms, uncorrelated each other:

$$U(\{Q\}) = \sum_{l} U^{l}(\{Q_l\}) \tag{7}$$

where U is the total energy of the system, $\{Q\}$ is the entire set of coordinates, Q_l the single CG internal variable and U^l its corresponding energy (force field term).
The probability distribution of a single internal variable is then:

$$P(Q_l) \propto \int dQ_1, ..., dQ_{l-1}, dQ_{l+1}, ..dQ_n \exp\left(-\frac{U(\{Q\})}{kT}\right) = \exp\left(-\frac{U(\{Q_l\})}{kT}\right) \tag{8}$$

with n number of DOFs of the system, k the Boltzmann constant and T temperature of the entire system. The last equality is valid only if terms are completely uncorrelated. The equation (8) can be equivalently written as:

$$U(Q_l) = -kT \ln(P(Q_l)) + const \tag{9}$$

Equation 9 defines the potential of mean force (PMF), which is, in first approximation, the effective term that should be included in the FF. The next assumption is that $P(Q_l)$ can be evaluated from a statistical set of structures. In principle, this is true only if the set is populated following an equilibrium distribution. This is a controversial issue: there is no guarantee that this is true for a given set of structures downloaded from the PDB. The best one can do is to maintain the largest possible diversity in the set, in order not to introduce an artificial bias. In other words, it is assumed that the statistical potential obtained by BI evaluated on experimental set is a good approximation of the PMF. For every defined internal variable Q_l, the statistical potential (SP) is defined as:

$$U(Q_l) = -kT \ln\left(\frac{P(Q_l)}{P_0(Q_l)}\right) \tag{10}$$

where $P_0(Q_l)$ is the probability distribution of Q_l in a reference state of non interacting particles [37]. The idea is that the interaction energy between beads

deriving from a specific internal variable is the difference of energy between the interacting system and the ideal system [83], in which theoretically there are no interactions. The non interacting particles are randomly distributed in the 3D space. $P_0(Q_l)$ is usually Q independent and SP and BI differ only by an irrelevant constant. For some specific variables this is not true, it will be pointed out in the following.

Equation (10) gives an operative way to directly build the FF terms. In principle, in fact, the SPs may be directly used to fit the parameters for the corresponding force field analytic terms. This is true only in first approximation. First, the probability distribution are in fact amino acid dependent, so an accurate FF would requires the construction of about 20^3 different PMFs only for the construction of U^{back}. Second, the probability distribution functions depend on the specific statistical set chosen and from the origin experimental or computational of the data (i.e. generated with more accurate models). The biggest problem is that the force field terms are not uncorrelated.

There are different possible strategies to address these problems especially the one of correlation between force field terms. It is possible to first assign to all the FF terms the corresponding SPs and subsequently correct them on the basis of the results of simulations iteratively, until specific quantities (e.g. the probability distributions themselves or other observables) are reproduced. The correction is:

$$\Delta U = -kT \ln \left(\frac{P(Q_I)}{P^{n-1}(Q_I)} \right) \tag{11}$$

where $P(Q_I)$ is the target distribution (experimental) and $P^{n-1}(Q_I)$ is the distribution obtained by the simulation using the potential itself. The procedure is called Iterative Boltzmann Inversion (IBI) [22].

However, the physical meaning of these SPs was widely disputed, since their introduction [47],[37]. The main issues are: the interpretation of these potentials as true, physically valid potentials of mean force and the nature of the reference state and its optimal formulation. This discussion is out of the scopes of this work and is better addressed in [37], but to avoid ambiguity the term Statistical Potentials for the force field terms derived from statistical information extracted from datasets is used here.

Many recent models use SP as FF terms, with a simplified version of the iterative correction procedure. The so called 'partially biased' models were introduced to simulate the large dynamics of the HIV-1 protease [40],[42]. In [40], U^{back} is divided into two contributions, i.e. $U^{\theta}(\{\theta_i\})$ and $U^{\Phi}(\{\phi_i\})$. This is a division often used in the minimalist models (see table 12). For the specific case of the partially biased model in [40], u^{θ} is represented as a double well quartic potential (see table 12) where the equilibrium angle is the location of the first well ($\sim 90°$). The other parameters are amino acid type dependent and determine the position of the second well and its relative stability. BI procedure was used together with direct information extraction from experimental structures to parameterize u^{θ}. The parameters extracted are then optimized using the probability distributions as targets.

In the locally biased models, u^{nb}(see table 12) both local and non-local, are represented with a morse potential, but a partial bias in the structural parameters of $u^{nb,loc}$ is maintained to better reproduce the short range interactions: the equilibrium distances r_{ij}^0 differ for each ij couple and are taken from a

single reference structure. $U^{nb,loc}$ and $U^{nb,non-loc}$ are separated by a cutoff radius. The first contain mainly the hydrogen bond interactions, while the second mainly the hydrophobic and electrostatic interactions. Thus the local bias allows to represent the most complex terms of the FF in a very simple way. It also maintains the structure stable and gives a high level of structural accuracy. At the sime time, the unbiased term gives enough flexibility to the system: even the out of equilibrium dynamics can be simulate.
Other examples of models, which use statistical potentials are VAMM FF [87] and DMC models [7].
In conclusion, one can say that the BI-based methods realize the thermodynamic consistency of the CG model with a given statistical set of data. In fact, they use the target quantity to reproduce in the CG simulation the probability distribution which are related to the internal variable dependent free energies by the Boltzmann inversion. In particular if the data set is produced by atomistic simulations, this is the strategy to generate a CG model thermodynamically consistent with a given atomistic model, i.e. to thermodynamically match two different resolutions.

2.4.3 The Force Matching method

The Force Matching (FM) is an alternative approach where the targets are the forces acting on the CG sites [43], instead of the thermodynamic properties of the system. In the FM, the input data are typically obtained from trajectories of atomistic simulations. The quality of the parameterization depends on the quality of the atomistic Force Field used and on the extension of the phase space sampling by the atomistic simulation. This method was rigorously formulated and optimized [33],[25] and finally named multiscale CG method (MS-CG).
It consist in the minimization of the functional:

$$\chi^2(\{\vec{F}\}) = \frac{1}{3N}\left\langle \sum_{i=1}^{N} |\vec{F}_I(\{Q(\{q\})\}) - \vec{f}_I(\{q\})| \right\rangle \tag{12}$$

where $\vec{F}_I$ are the CG forces on CG sites I, while $\vec{f}_I$ are the forces on the CG sites evaluated from the AA simulations. The average is on the atomistic simulation trajectory. This functional must be minimized with respect to the parameters of $\vec{F}$. This equation realizes the mechanical consistency between the atomistic and the CG representation.

2.4.4 Physics-Chemistry-Based Parameterization

To accurately assess the input data and the distribution probability functions is not an easy task, because it implies that the data included in the statistical set are distributed according to the thermal equilibrium. This is really difficult to achieve, especially if the set come from a simulation. Using data from simulations includes possible additional systematic errors due to the atomistic force field, which is a model itself.
A possible way to improve accuracy is to include in the parameterization elements based on the known chemical and physical properties of the amino acids

and/or thermodynamics data from experiment. Starting with a reasonable approximation of the FF, the parameters can be optimized until the convergence between simulated and target quantities is reasonably reached. This strategy can be also implemented with multi-variable fit procedures. However, as the number of parameters or the number of target observables increase, the procedure becomes more numerically unstable and computationally expensive.
A model of this class is the one by Sorenson and Head-Gordon [77],[79]. In its last version (named YFSH model from the names of the authors) [79], U^{back} is separated in its two components (see table 12). u_θ is harmonic with secondary structure based parameters and θ_0 is different from helical or extended structures. u_ϕ is a complex cosine sum, with A,B,C,D (see table 12) secondary structure dependent (including the turns). The dihedral term is considered as the one that mainly determines the secondary structure. Its parameters are chosen to stabilize one or the other secondary structure, and assigned based on the secondary structure propensity of the amino acids. The stability of the folding is also determined by the non-bonded term (see table 12). Here, the parameters are assigned based on the 'flavor' of the amino acids: hydrophobic, hydrophilic and neutral. As the dihedral term, also the non bonded one can be assigned solely based on the sequence.
The YFSH model can be put in the class of those highly predictive and transferable (with low structural bias), but its high transferability is paid with a low structural accuracy. The functional forms of the single FF terms are rather crude with respect to the BI or FM based FFs. Other example of force fields based on this methodology is the one of Alemani et al [21]. The force field is in table (12). The secondary structure is maintained also by and additional term that correlates subsequent dihedrals.

2.4.5 The Hydrogen Bond Term

Considered the focus of this work, it is important to underline some of the possible ways to implement the terms corresponding to the hydrogen bonds. These are fundamental to the maintenance of the secondary structures. This concept comes naturally from the consideration that different patterns of H-bonding define different secondary structures. So, if the goal is a good representation of the secondary structures, it is necessary to add terms that explicitly maintain the pattern of H-bonds.
In some models the H-bond is implicit in the U^{nb} or in the local part of the non bonded term [40]. For instance, in the latest version of the YFSH models [79], the explicit hydrogen bond term depends on the distance r_{ij}. This potential implements the interactions between two planes: the one defined by the three subsequent (C_α-centered) beads $(i-1, i, i+1)$ and the plane defined by other three subsequent beads $(j-1, j, j+1)$. This potential stabilizes helices and sheets, through the relative orientation of those two planes. The parameters are assigned based of a statistical sets of secondary structures taken from the PDB. The hydrogen bond potential is evaluated for all $i-j$ bead-pairs, whose amino acids are capable of forming hydrogen bonds: at each bead the hydrogen bond forming capability is assigned (amino acid based) from three possible types: helical (designated A), sheet (designated B), or none (designated C). After the assignement of the flavour, there are some rules with which the different flavours

interact, e.g. for a bead assigned B, the hydrogen bond potential is evaluated between itself and all B-beads situated within a cutoff distance of 3.0 length units.
Also in the model of Alemani et al. [21] there is an accurate description of the hydrogen bond term, represented by a dipolar interaction between peptide dipoles $\vec{\mu}_{ij}$ (see table 12). The $\vec{\mu}_{ij}$ are located approximately midway between beads i and i+1 and their orientation depend on the orientation of the plane formed by the three beads $(i-1, i, i+1)$.
The two hydrogen bond terms described are similar for the use of the mutual distance between triplets of beads, but they differ in some aspects. First, in Alemani et al. the term is entirely physics based, imputing the hydrogen bond to a simple dipolar interaction. Second, the hydrogen bond of this model does not explicitly depend on the secondary structure, which enters however implicitly in the definition of the orientation of $\vec{\mu}$. Here, the H-bond helps in the stabilization, but the α versus β propensity included mainly by the conformational terms. This model is capable of reproducing stable secondary structures and the transitions among them, but a systematic amino-acid-dependent parameterization was not established. Thus, it still lacks the predictive power of the FYHG model.
The last type of hydrogen bonding analyzed is the one in the FF of Seno et al. [1]. This is not really an OB minimalist model. In fact, the beads are placed on the C_αs, but they are connected by a tube with a non zero radius. In this case, a hydrogen bond is recognized if some geometric constraint are satisfied. The H-bonds are separated into local, between beads two residues apart along the chain, and non-local, between beads more than two amino acids apart. The local H-bonds are more stabilizing the conformation than the non-local one. A new element is the cooperativity effect: there is an energetic price if consecutive H-bonds are formed. A hydrogen bond is assigned on the basis of geometrical constraints. Considering the planes formed by three non collinear C_α, a H-bond is established between the two central bead if the vectors normal to the two respective planes are parallel each other and parallel to r_{ij}, distance between the two amino acids considered.

2.5 EXPLORING THE CONFORMATIONAL SPACE: CLASSICAL MOLECULAR DYNAMICS

Once the potential energy is defined, it can be used to sample the conformational space. There are many methods to do this. Within the molecular dynamics (MD) framework, this is done through the numerical resolution of the equation of motion. In the following the basis of the method and the computational procedure are elucidated.

2.5.1 Molecular Dynamics Fundamentals

In the CG OB C_α-based models, a simple classical mechanical approach is used to describe the motion of the beads. The Newton's equation of motion, to be numerically solved, is:

$$\vec{F}_i = m_i \vec{a}_i = m_i \ddot{\vec{r}}_i = -\nabla_i U(\vec{r}) \tag{13}$$

where $\vec{F}_i$ is the force acting on particle i, m_i is the mass of particle i, $\vec{a}_i$ is its acceleration and $\ddot{\vec{r}}_i$ is the second derivative of the particle position $\vec{r}$ with respect to time. $U(\vec{r})$ is the potential energy function and the force is determined by its gradient.

Computational Algorithms for Molecular Dynamics

A standard method to numerically integrate the differential equation (13) is the so called finite-difference approach: the molecular coordinates and velocities at a time $t + \Delta t$ are obtained from the molecular coordinates and velocities at an earlier time t. The equations are solved on a step-by-step basis.
The most common integration algorithm is due to Verlet [46]. The basis of this integrator is the sum between two Taylor expansions, around $(t + \Delta t)$ and around $t - \Delta t$, that gives:

$$\vec{r}_{n+1} = 2\vec{r}_n - \vec{r}_{n-1} + \frac{\vec{F}_n}{m}\Delta t^2 + O(\Delta t^4) \tag{14}$$

where $\vec{r}_n$ indicates the position at step n (at time t), $\vec{r}_{n+1}$ indicates the position at the next step, $n + 1$ (at time $t + \Delta t$), and $O(\Delta t^4)$ is the term of order Δt^4. So, the current force $\vec{F}_n$ is calculated from the current position $\vec{r}_n$. Then, $\vec{r}_n$ and $\vec{r}_{n-1}$ are used together with the just obtained $\vec{F}_n$ to calculate the position in the next step, $\vec{r}_{n+1}$, according to equation (14). This procedure is repeated for each timestep for each bead in the system. Subtracting the two Taylor expansions $\vec{r}_{n+1}$ and $\vec{r}_{n-1}$ yields a complementary algorithm for propagating the velocities.
In the Verlet algorithm the position integration is quite accurate and independent of the velocity propagation. On the other hand, the velocity propagation is subject to relatively large errors and $\vec{v}_n$ can be computed only if $\vec{r}_{n+1}$ is already known. Furthermore, the Verlet algorithm is not "self-starting": a lower order Taylor expansion is often used to initiate the propagation. Finally, it must be modified to incorporate velocity-dependent forces or temperature scaling.
To overcome these disadvantages, particularly to improve the velocity evaluation, the leap-frog algorithm was proposed. It is so called for its half-step scheme: velocities are evaluated at the mid-point of the position evaluation and vice versa [69] [14] . The algorithm can be written as:

$$\vec{r}_{n+1} = \vec{r}_n + \vec{v}_{n+1/2}\Delta t \tag{15}$$

$$\vec{v}_{n+1/2} = \vec{v}_{n-1/2} + \frac{\vec{F}_n}{m}\Delta t \tag{16}$$

where $\vec{v}_{n\pm 1/2}$ stands for the velocity at the mid-step time $(t \pm 1/2\Delta t)$. This algorithm involves three steps: first, $\vec{F}_n$ is calculated from $\vec{r}_n$; second $\vec{F}_n$ together with $\vec{v}_{n-1/2}$ are used to obtain the next mid-step velocity $\vec{v}_{n+1/2}$; third,

$\vec{r}_n$ and $\vec{v}_{n+1/2}$ are used to compute the position in the next step, $\vec{r}_{n+1}$. The current velocity $\vec{v}_n$, which is necessary for calculating the kinetic energy, can be calculated as:

$$\vec{v}_n = \frac{\vec{v}_{n+1/2} + \vec{v}_{n-1/2}}{2} \tag{17}$$

The advantage of the leap frog algorithm is that it improves the velocities evaluation, giving a useful handle for controlling the simulation temperature, via velocity scaling. The disadvantage is that it still does not handle the velocities in a completely satisfactory manner, because the velocities at time t are only approximated by equation (17).
The size of the time step is an important parameter that determines the magnitude of the error associated with each of the foregoing integration algorithms. A small time step means better integration quality, but more integration steps are required for the same length of simulation. In general, one would like to choose the largest possible time step that still ensure an accurate simulation.

Constrained Dynamics

Constrained dynamics enables individual internal coordinates of a system to be fixed, during the simulation, without affecting the other internal degrees of freedom.
Usually in MD, constraints are used to fix the bonds to their equilibrium value. This allows the use of longer simulation time steps Δt, if the constrained bonds correspond to the highest frequency modes. The most commonly used method for applying constraints is the SHAKE procedure [44]. Here, after each timestep, the atoms positions are modified in order to satisfy the constraints. Another commonly used algorithm for constrained dynamics is RATTLE [32].
Additional details are given in Appendix B.

2.5.2 Molecular Dynamics in Canonical Ensemble

In the simplest version, Molecular Dynamics simulations are performed at constant number of particle, volume and energy, i.e. the microcanonical ensemble (also called the NVE ensemble). This is not the condition in which real systems are found. So, the necessity of simulating in more realistic conditions, such as in the canonical (NVT) ensemble, arises. The instantaneous temperature is related to the kinetic energy, according to the equipartition theorem, by:

$$T(t) = \frac{1}{k_B N_{DOF}} \sum_{i=1}^{N_{DOF}} m_i |\vec{v}_i|^2 \tag{18}$$

where $N_{DOF} = 3N - n$ is the number of unconstrained degrees of freedom in the system, with N number of beads and n number of constraints.
The simplest way to keep constant the temperature of the system is velocity scaling [49]. If $T(t)$ is the system temperature at time t and the velocities are

multiplied by a factor λ, then the associated temperature change can be calculated as:

$$\Delta T = \frac{1}{k_B N_{DOF}} \sum_{i=1}^{N_{DOF}} m_i(\lambda v_i)^2 - \frac{1}{k_B N_{DOF}} \sum_{i=1}^{N_{DOF}} m_i(v_i)^2 \tag{19}$$

$$\Delta T = (\lambda^2 - 1)T(t) \tag{20}$$

$$\lambda = \sqrt{\frac{T_0}{T(t)}} \tag{21}$$

with $T(t)$ current temperature, as calculated with the kinetic energy, and T_0 the target temperature. The temperature is then controlled by rescaling the velocities at each timestep by a factor λ.
A further refinement of the velocity rescaling approach was proposed by Berendsen et al. [30]. To maintain the temperature, the system is coupled to an external heat bath with fixed temperature T_0. The velocities are scaled at each step with a more specific scaling factor, with an additional parameter that controls the relaxation towards T_0 and make the method more numerically robust.
The biggest problem or these two method is that they do not sample the canonical ensemble. To overcome this limit, the Nose thermostat is often adopted [72]. The molecular system is placed in contact with a thermal reservoir in an unique Lagrangian formalism. Energy is allowed to flow dynamically from the reservoir to the system and back. Additional details are reported in Appendix B.

2.6 ANALYSIS OF THE MOLECULAR DYNAMICS TRAJECTORIES

The final output of a molecular dynamics simulation is a trajectory of the atomic positions as function of time. Some properties obtainable from these data are the Root Mean Square Displacement (RMSD) from a reference configuration and the Root Mean Square Fluctuations (RMSF).
For a set of N atoms at time t and with respect to a reference conformation, the RMSD is defined as:

$$RMSD(t) = \sqrt{\frac{\sum_{i=0}^{N} |\vec{r}_i(t) - \vec{r}_i^0|}{N}} \tag{22}$$

Here $|\vec{r}_i(t) - \vec{r}_i^0|$ is the displacement from the reference position $\vec{r}_i^0$. The RMSF calculates the displacement of the atoms along the trajectory. Usually $\vec{r}_i^0$ is the first configuration of the trajectory. If so defined, $RMSD^2$ increases linearly in time in case of diffusional behaviour. However, in order to maintain the internal fluctuations, the net translations and rotations are removed (if present) and in this case the RMSD fluctuates around a constant, measuring the overall average fluctuations and increasing with temperature or due to possible structural transitions.
On the other hand, the RMSF are the fluctuations of an atom (or bead), defined as:

$$RMSF_i = \sqrt{< (\vec{r}_i - \vec{r}_i^0)^2 >} \tag{23}$$

here $\vec{r}_i$ is the vector position of the ith atom and the brackets are for a time average. The RMSF indicates which atoms (beads) of the structure are more fluctuating.

3 THE MINIMALIST MODEL I: THEORETICAL BASIS AND PARAMETERIZATION PROCEDURE

In this Chapter, the first original part of this work is presented. The strategy and the tools adopted to build the model are introduced and the results of the statistical analysis are discussed.

3.1 FEATURES OF THE MODEL

The goal of this Thesis is to build a minimalist CG C_α-based model to reproduce primarily the secondary structures (these must be properly combined in tertiary structures to obtain the protein 3D fold, but this is a matter of subsequent work). For this reason, relatively small peptides (7 to 20 amino acids) are considered, which are the typical sequence length of secondary structures.
To achieve the best compromise between accuracy and predictive power, the force field is parameterized as sum of statistical potentials derived from the statistical distributions of the internal coordinates of the model. Such distributions are used as direct input or as target quantity for the parameters optimization. The accuracy of the model is thus based on the quality of the statistical distributions, determined by the statistical relevance of the dataset (i.e. number and diversity of included structures) and its composition in terms of sequence or secondary structures.
In order to build these datasets, SecStAnT is here developed [80],[73]. It can in fact efficiently create from the PDB [66] (i.e. the most comprehensive database for biomolecular structures) datasets of protein secondary structures, selected according to structural criterion or amino acid sequence information (see next section), at different levels of resolution (atomistic or CG). Moreover, it can evaluate a large number of internal variables distributions, together with two and three body correlation functions. This last feature is exploited to systematically address the problem of FF terms interdependency.
It is important to underline that the datasets, which SecStAnT produces on-demand, are always updated by construction to the latest results of the experimental research on proteins as represented in PDB. Furthermore, any dataset is always reproducible in a short time, because SecStAnT automatizes the search procedure of search by parsing and fragmentation of all the PDB entries corresponding to a specific query.
In the following, SecStAnT is described in detail and the distributions and the correlations obtained for each specific secondary structure are theoretically and experimentally discussed.

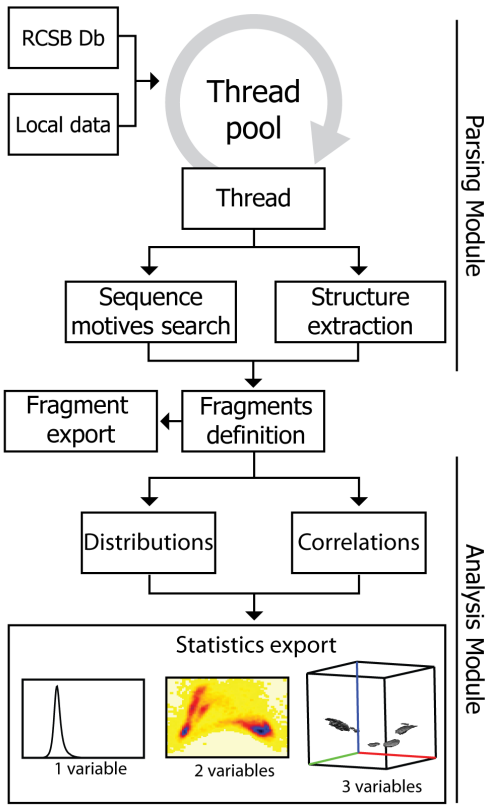


Figure 13: Schematic illustration of SecStAnT workflow. The process is separated in two modules. In the parsing module, input data are downloaded and processed by a thread pool. Each entry is fragmented by user-defined primary and secondary structure criterions. In the analysis module, each fragment is then saved separately and a series of statistics is calculated

3.2 SECSTANT: SECONDARY STRUCTURE ANALYSIS TOOL

SecStAnT is written in Java, it is available at [73] and the source code is distributed under the BSD2 Open Source licence.
The program is roughly composed of two modules, as described in figure 13. The first one, the parsing module, performs the data set building, extracting structures from PDB and fragmenting them in elements with defined secondary structures. The input selection is performed through a graphical interface by combining secondary structure information with any other selection criterion available on the RCSB advanced search interface as, for instance, the experimental method for the structure determination, the release year, etc.
The downloading process is performed through RCSB FTP interface, according to the server guidelines; a cache mechanism is implemented in order to avoid multiple downloads of a single entry.
As anticipated, the queries consist of secondary structures composition and sequence motifs. Secondary structures can be selected either based on the information included into the PDB file itself (provided by the PDB file author) [61], or on the DSSP file (provided by RSCB and based on the DSSP algorithm) [8]. Primary structure is defined by standard regular expression search. During the extraction process, information on primary, secondary and super-secondary structures (when available) is mined and stored. Either the whole proteins

or only the structure fragments with the selected secondary structure can be stored in hierarchical organized folders for future consultation. A detailed description of the output dataset organization is given in Appendix C.
In the analysis module, the fragments data set is used to build different kinds of distributions of internal variables and their correlations. SecStAnT is a tool designed to perform the statistical analysis for one bead C_α-based models, so distribution and correlations of the internal variables are implemented for chains containing only C_α. The output format (described in detail in Appendix C) is given in numerical form, conveniently readable by a large number of commonly used graphics software packages.
In the next sections the potentiality of SecStAnT in this second module are described.

3.3 BACKBONE CONFORMATION DESCRIPTION: ALL ATOM AND COARSE GRAIN

SecStAnT is able not only to separate secondary structures, but also to coarse grain the structure of different levels, namely all-atom, backbone only and minimalist (C_α only). Distributions can be evaluated for a number of pre-defined internal variables. Correlations are calculated for the C_α-based fragments, but it is also possible to generate the Ramachandran (Φ,Ψ)-plot for the all-atom and backbone only level.
Exploiting the possibility to extract primary sequence based fragments, the Ramachandran plots for the 20 different amino acids can be generated. The maps were calculated for all the fragment of repeated specific amino acids, e.g. polyAla, polyGli or PolyPro, contained in all the proteins deposited in the RCSB Protein Data Bank. The Ramachandran plots for the 20 amino acids are reported in appendix A. By definition, the Ramachandran plot displays densely populated areas corresponding to the main secondary structures, in which the (Φ,Ψ) pairs assume typical values reported in table 7.
In the C_α based representation of the protein backbone, the dihedral angles Φ,Ψ are no more explicitly represented. The internal variables describing the conformations are now θ and ϕ, respectively the angle between three subsequent C_α and the dihedral between four subsequent C_α, as showed in figure 11. The (θ,ϕ) map can be considered the OB-CG analogous of the RP plot.
SecStAnT is then able to generate the 2D maps that define the conformations of a polypeptide chain in the all atom configuration, as well as in the OB CG ones. The tool thus provides a direct comparison among the conformational variables in the two spaces (Φ,Ψ) and (θ,ϕ).
The relation between these two maps, has been previously analyzed in [41]. An analytical correspondence of the all-atom to the CG internal backbone coordinates that allows for an explicit mapping of the Ramachandran plot onto the

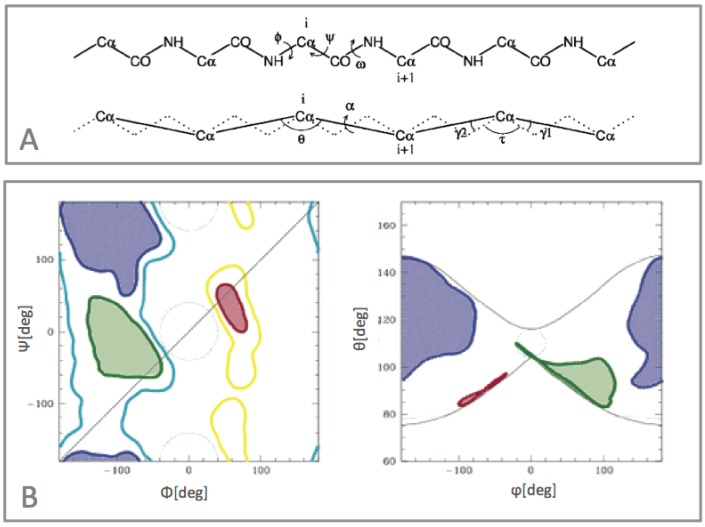


Figure 14: Backbone representation. Panel A: all atom conformational degrees of freedom and the correspondent ones in the CG representation. In the upper part, the all atom dihedral conformational angles (Φ,Ψ) are reported. In the lower part, the θ,ϕ variables for a OB CG C_α-based models and the angles defining the CG procedure (τ,γ_1,γ_2) are reported. Typical values are $\tau = 111°$,$\gamma_1 = 20.7°$ and $\gamma_2 = 14.7°$. Here, the symmetric approximation $\gamma_1 = \gamma_2 = 16° = \gamma$ is used. Panel B: Mapping of the RP into the (θ,ϕ)-map. The allowed conformations are in green (right-handed helices), red (left-handed helices) and blue (extended strands).

new (θ,ϕ) conformational density plot was derived. Under some simplifying conditions the following relationships can be derived:

$$\phi = \Phi + \Psi + \pi + \gamma(\sin\Phi + \sin\Psi) - \gamma(\tau - \frac{\pi}{2})(\sin\Phi + \sin\Psi) + \\ + \frac{1}{4}\gamma^2(\sin 2\Phi + \sin 2\Psi + 8\sin(\Phi + \Psi)) \tag{24}$$

$$\cos\theta = \cos\tau(\cos^2\gamma - \sin^\gamma\cos\Phi\cos\Psi) + \sin\tau(\cos\gamma\sin\gamma(\cos\Phi\cos\Psi)) + \\ - (\sin^2\gamma\sin\Phi\sin\Psi) \tag{25}$$

(see figure 14 (Panel A) for the definition of parameters and variables). Uniform secondary structures are assumed and the directionality of the polypeptide chain is neglected. This implies that the two angles θ− and θ+ near a given ϕ dihedral have the same values. The two maps (θ−,ϕ) and (θ+,ϕ) coincide in this approximation.
Using equations 24, a uniform density in the all atom plane is mapped onto a non-uniform butterfly-shaped image in the CG plane [41], as shown in figure 14 (Panel B). Because of the mapping, the allowed areas corresponding to secondary structures are re-shaped and re-sized (figure 14, Panel B). However, due to the specific relative location of the forbidden and allowed secondary structure areas in the (Φ,Ψ) plane, these remain separated even in the (θ,ϕ) plane. Consequently, the backbone conformation is uniquely determined for each (θ,ϕ) couple and the all atom backbone conformation can be uniquely reconstructed from the CG one.
SecStAnT automatically generates the results that numerically confirm this theoretical result. A sample of these results is shown in figure 15, reporting the RP (first line), (θ−, ϕ) and (θ+,ϕ) plots (second and third line) for generic (first

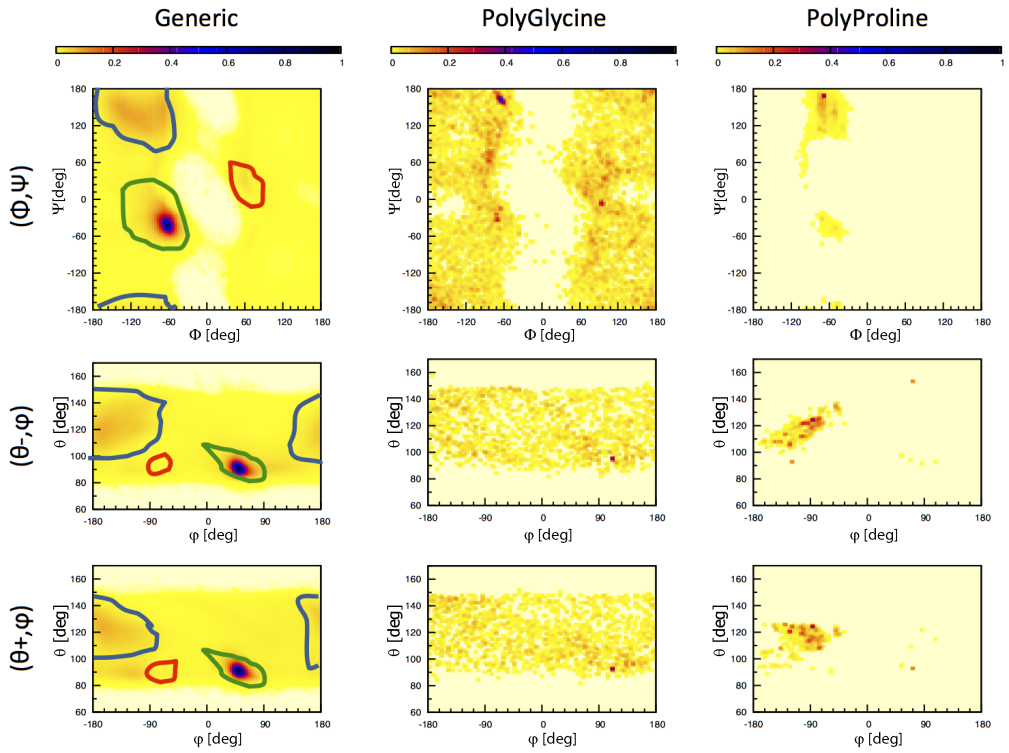


Figure 15: Conformational plots. Here are reported for generic amino acid, polyGlycine and PolyProline (three different columns) in the first line the Ramachandran plot, in the second line the $(\theta-,\phi)$-plot and in the third line the $(\theta+,\phi)$-plot. The color bar is also reported: light yellow correspond to forbidden areas, while dark yellow to area with few occurrence.

column), polyGli (second column) and polyPro (third column) datasets. Observing the generic maps in figure 15 (first column), some considerations can be done. First, the ϕ axis is completely spanned, while the θ one is confined and the θ range limitations are coherent with the theoretical prediction of [41]. Second, the different secondary structures are clearly distinguishable: in figure 15 the helical region and the extended one are marked in different colors; the left and right handed helix are also separate. For a generic amino acid, the three different types of helices (3_{10},α and π) are all in the same helical region, because the values for the conformational angles are similar. In the following Chapters, the different correlations for specific secondary structures based datasets will be analysed.

The third important observation is that the two maps $(\theta-,\phi)$ and $(\theta+,\phi)$ are not superimposable, underlining the directionality of the chain. The difference come from the intrinsic asymmetry of the backbone. Locally, the CO and NH groups occupy different sides of the C_α. This local asymmetry is shown by the polyGlycine map. In its $(\theta-,\phi)$ and $(\theta+,\phi)$ maps in figure 15, the butterfly shape appears following the red points, but there are also red points out of the "wings". The intrinsic directionality of the backbone leads to a non symmetric representation in the θ,ϕ space.

Furthermore, the CG maps are not symmetric with respect to $\phi = 0$, because of the chirality of the amino acids. The peptide bond is mostly in the trans configuration, so the left-handed helix are less populated than the right-handed. The same happens for the extended structure. The strands are mainly in the

Table 7: Summary of protein secondary structures. Conformational variables for the most common secondary structures. The θ and ϕ values are from the structures built with Avogadro [29]

Structure	Φ[°]	Ψ[°]	θ[°]	ϕ[°]
Anti-parallel sheet [15]	-139	135	131	179
β-strand [83]	-120	120	121	178
Parallel sheet [15]	-119	113	119	177
$3_{10} - helix$ [6]	-57	-30	85	69
$3_{10} - helix$ [29]	-74	-4	87	83
α-helix [34]	-58	-47	92	52
α-helix [6]	-63	-42	92	51
π-helix [29]	-57	-70	96	26
π-helix [83]	-30	-90	100	34
PolyProline II [83]	-75	-145	119	109
PolyProline I [83] cis	-71	160	100	96
Left-handed α-helix [83]	57	47	92	-52

left part of the (θ,ϕ)-plot and this is the confirmation that these are structures slightly left handed.
Finally, with respect to the Ramachandran map, the (θ,ϕ) plot has a more direct interpretation: the ϕ dihedral directly represent the helicity, thus $\phi = \pm 180°$ corresponds to flat strands, $\phi = 0°$ to rings, while positive and negative ϕs correspond to right or left handed structures with different degrees of helicity.
In conclusion, it is important to underline that the presence of a map with the same level of information about the secondary structure in the all-atom and in the coarse grained representation provides a useful tool to the validation of the results of simulation. The correct sampling of the conformational space will be a meaningful comparison.

3.4 THEORETICAL CORRELATIONS BETWEEN DEGREES OF FREEDOM

The (θ,ϕ) maps for the generic dataset in figure 15 shows that for the one bead C_α based CG models the two conformational internal variables are strongly correlated. This correlation is more evident in the presence of defined secondary structures. Consequently, one could infer that it is generated by real interactions, specifically those stabilizing the secondary structures, normally hydrogen bonds. It is then useful to analyse, the geometric relations among θ, ϕ and others internal variables related to those interactions. These relations are analytically derived, when possible, otherwise numerically evaluated.
The first considered is the relationship between θ and the distance (r_{1-3}) between the first and the third bead forming the angle (see figure 16, panel A). If all the peptide-bonds are in trans conformation (left figure in panel A, figure

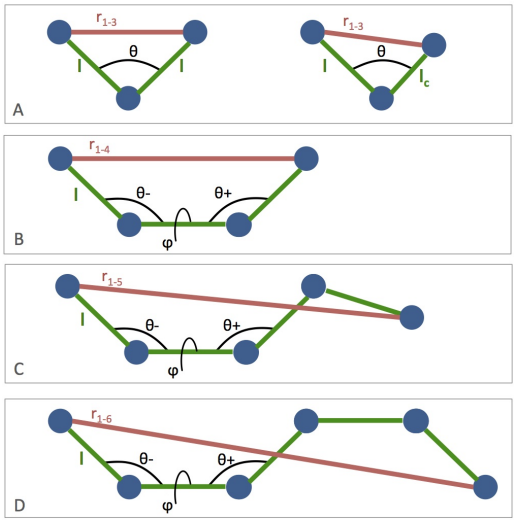


Figure 16: Internal degrees of freedom. Simple representation of the internal variables included in the different correlations. In panel A there are the two possibilities of trans on the left and cis on the right conformations for the peptide pseudobond between second and third bead.

16) then the distance C_α-C_α is $l = 3.8$Å and the three beads form an isosceles triangle, so:

$$r_{1-3} = 2l \sin\left(\frac{\theta}{2}\right) \tag{26}$$

Considering one pseudo-peptide bond in cis conformation (right figure, panel A, figure 16), this is a scalene triangle. Then, the Carnot formula leads to:

$$r_{1-3} = \sqrt{l_c^2 + l^2 - 2ll_c \cos\theta} \tag{27}$$

where $l_c = 2.9$Å is the length of the peptide bond in the cis conformation. The variables r_{1-3} and θ are in one-to-one correspondence, consequently they represent the same degree of freedom. It is sometimes more useful to use one or the other, or even both, as it will be clear in the next Chapter.
Another analytically derivable relation is between the dihedral angle ϕ, the bond angle θ and the distance r_{1-4} between the first and the fourth bead, as in panel B of figure 16. This relation has been previously calculated [83]. The analytical derivation assumed that the secondary structure is regular, so $\theta+$ and $\theta-$ are considered equal. The bond length is always considered in the trans conformation and then fixed at $l = 3.8$Å. Under these approximations:

$$r_{1-4} = l\sqrt{(1-2\cos\theta)^2 + 4\sin^2\theta \sin^2\frac{\phi}{2}} \tag{28}$$

this equation can be explicitly solved with respect to ϕ or θ. It is useful to plot $\theta(\phi)$ lines of constant values of r_{1-4} (see figure 17A). This was done both using 28 and numerically.

For the numerical calculation an unphysical toy model of a peptide has been built, with 20 beads C_α based. This structure is forced to assume all the different conformation corresponding to all the different values of (θ, ϕ), imposed to the entire chain. The beads are set with no interactions except their linkage in the chain. For each different (θ, ϕ) imposed, the distances between the four

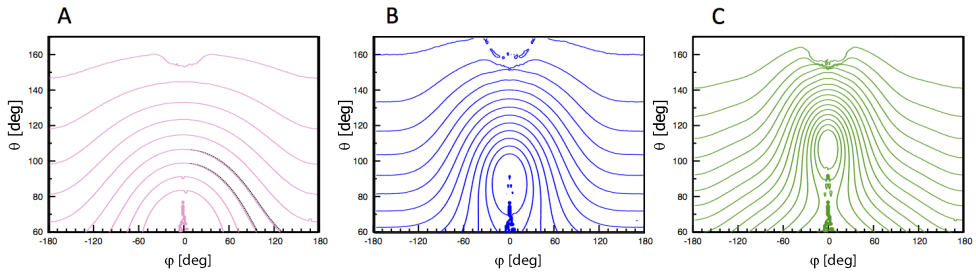


Figure 17: Theoretical correlation functions. Representation of the three correlation functions. These are the lines in the θ,ϕ space corresponding to constant values for the different distances. Plot A: (θ, ϕ)-correlation with constant r_{1-4}. With pink lines the computational correlations are drawn for constant r_{1-4}, increasing by one starting from 3 Å up to 11 Å, while with black dashed lines there are the two analytical lines for r_{1-4} = 5, 6Å. Plot B: (θ, ϕ)-correlation with constant r_{1-5}. Blue lines draw computationally calculated constant r_{1-5} correlations, increasing by one starting from 3Å up to 15Å. Plot C: (θ, ϕ)-correlation with constant r_{1-6}. Green lines are the correlation lines at constant r_{1-6}, increasing by one from 3Å up to 18Å.

subsequent beads in the center of the structure have been measured. At the same time, the distances are calculated between the first and the fifth beads and between the first and the sixth beads in the middle of the structure. The 4 to 6 central beads are chosen because it is supposed that the center of the structure is more regular than their edges. This procedure gives $\theta(\phi)$ (or $\phi(\theta)$) relations for fixed values of given distances. In figure 17 the lines connecting all the point in the conformational space with the same values of the three distances r_{1-4}, r_{1-5}, and r_{1-6} are shown, one for each graph (A,B,C).

3.5 STATISTICAL ANALYSIS

SecStAnT was used to generate secondary structure specific distributions, that will be subsequent used to aid the model parameterization. Specific datasets for each secondary structure has been built. These datasets differ for the experimental method used to solve the protein (X-ray or NMR) and for the algorithm used to identify the specific secondary structure in it (DSSP or directly from the PDB secondary structure entry). The selection of experimental method is done directly on the RCSB server, other selections are made by SecStAnT. Search results were filtered by the RCSB server imposing a structure similarity less than 30%. An example of RCSB query is reported in appendix C.
Once the dataset is built, the second module of SecStAnT can perform different distribution and correlations for different internal DOFs. The distributions for the conformational variables θ and ϕ have been calculated, together with every repeated distance between beads, i.e. $r_{i,i+n}$ with n=1,..,6. Finally, the distribution of all the distance $r_{i,j}$ with $j > i$ and the distribution for $r_{i,j}$ with $j > i+2$ have been computed. Algorithmic details are reported in appendix C.
SecStAnT is able to perform two or three variable correlations and provides some different ways to visualize them. It is possible to analyse volumetric data (see appendix C), but also to analyse the correlation between θ and ϕ for different values of r_{1-4} or r_{1-5}.

Table 8: Distribution parameters for 3_{10}-helix. These values are obtained using Octave [70]. The interquartile range (iqr) was computed as difference between the upper and lower quartile.

			$\theta[^\circ]$	$\phi[^\circ]$	r_{1-3}[Å]	r_{1-4}[Å]	r_{1-5}[Å]	r_{1-6}[Å]
X-ray	DSSP	max	89.7	68	5.33	5.75	8.2	10.2
		iqr	2.2	18	0.27	0.72	0.8	0.6
	PDB	max	89.7	63	5.42	5.6	7.76	10.2
		iqr	6.6	19.8	0.36	0.63	1.4	1.2
NMR	DSSP	max	88.11	64	5.33	5.61	8.12	9.92
		iqr	6.12	24	0.27	0.86	1.44	1
	PDB	max	88.1	60	5.33	5.33	8	9.92
		iqr	8.55	32	0.45	0.99	2	1.8

The relations between single variable distributions, e.g. r_{1-4} and the two variables correlations, e.g. θ,ϕ, are :

$$P(\theta, \phi) \propto \int P(\theta, \phi, r_{1-4}) dr_{1-4} \tag{29}$$

$$P(r_{1-4}) \propto \int P(\theta, \phi, r_{1-4}) d\theta d\phi \tag{30}$$

In the following the distributions and the correlations obtained for each secondary structure are discussed. Only relevant data are reported for chosen datasets.

3.5.1 3_{10}-Helix

Four datasets for this structure were built, fetching 15716 proteins for the X-ray dataset and 1881 for the NMR dataset. Using SecStAnT for each experimental method, a dataset of 3_{10}-helices identified using DSSP algorithm was created together with a dataset built directly using the information contained in the PDB files.
As discussed in Chapter 1, this type of helix is not completely separable from the α-helix. The ideal values for the bond angle and the dihedral angle are similar and often happens that a 3_{10}-helix starts or ends an α-helix. The exact edge between these two structures is not well defined.
Another problem is that they are very short. The average length is 7 amino acids, so the distributions of $r_{i,i+4}$ and $r_{i,i+6}$ have poor statistics and are noisy. In this case, the secondary peak is imputable on the impurities of α-helices wrongly identified as 3_{10}, and this is confirmed by the location of the secondary peak in other distributions, especially in the distribution of r_{1-6}.
From the distributions in figure 18 emerges the difficult definition of this helix. The peaks are not always at the same value, as confirmed by the table 8. Other distributions for these datasets are in appendix C. The most realistic data are those coming from NMR, because at variance with X-ray data, they are solved in more physiologic conditions, and from the DSSP algorithm. These lead to the most peaked curves and the most coherent maxima with the ideal data.

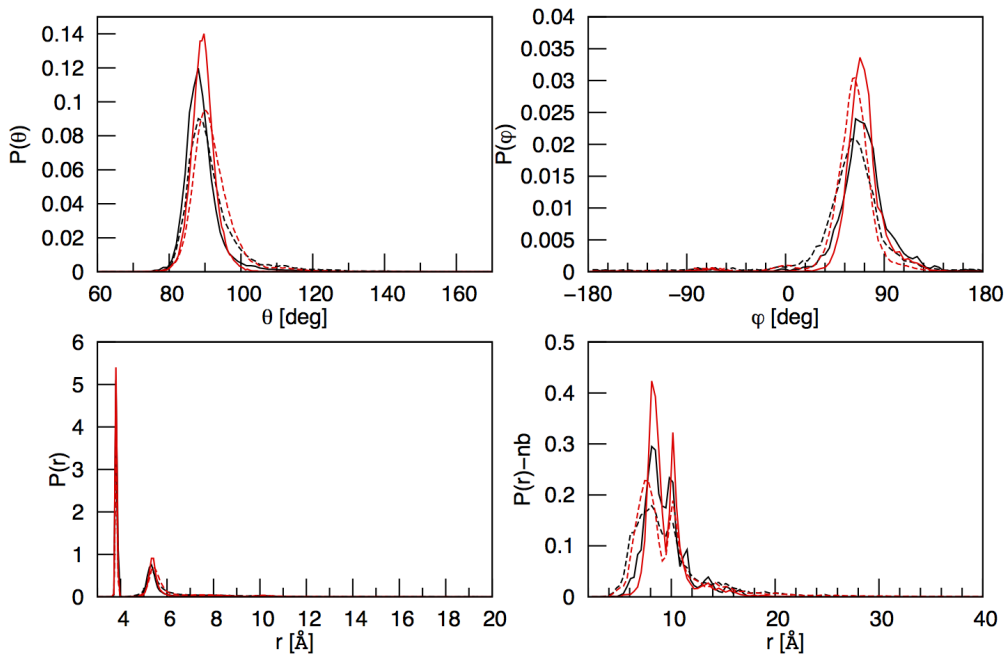


Figure 18: Internal DOFs distributions for 3_{10}-helices. Distributions of θ (top-left), ϕ (top-right), $P(r)$ for r distance between every two C_α (bottom left) and $P(r) - nb$ for r distance between every i and j with $j > i + 2$ (bottom right). Different lines are for different datasets: red for Xray and black for NMR, solid for DSSP and dashed for PDB direct information.

In figure 19 (Panel B) the Ramachandran plot and the CG conformational maps are reported. The occupied areas in the RP for the 3_{10}-helices are located at similar values of θ and slightly larger values for ϕ than those of α-helices (see next section). The two correlations are not completely distinguishable.
The clearer difference among correlation areas of the two helices is their slope in the (θ,ϕ). In the 3_{10}-helix the H-bonds would be between the CO group of an amino acid and the NH group of the amino acid three bead after along the chain. As it will be better shown in the next Chapters, such a H-bond should be represented in the C_α-based models by bonds between second and third neighbors along the chain. This implies that constant r_{1-3} and r_{1-4} correlation lines could represent the correlations in these helices. These lines are represented in black and pink in figure 19 (Panel A). As it can be seen, the slope of the correlation plot is between the two lines for constant r_{1-3} and r_{1-4}. Considerations about the correlation maps are more understandable in the following case of the α-helix.
In appendix C there are the additional distributions and correlations for this kind of helix.

3.5.2 α-Helix

The α-helix is the most common and abundant type of helix. 17051 proteins for the Xray and 3614 proteins for the NMR can be found in the PDB.
In figure 20 there are the distributions for every dataset (color codes in caption). These data are quite similar for all the datasets. It is to be noted that

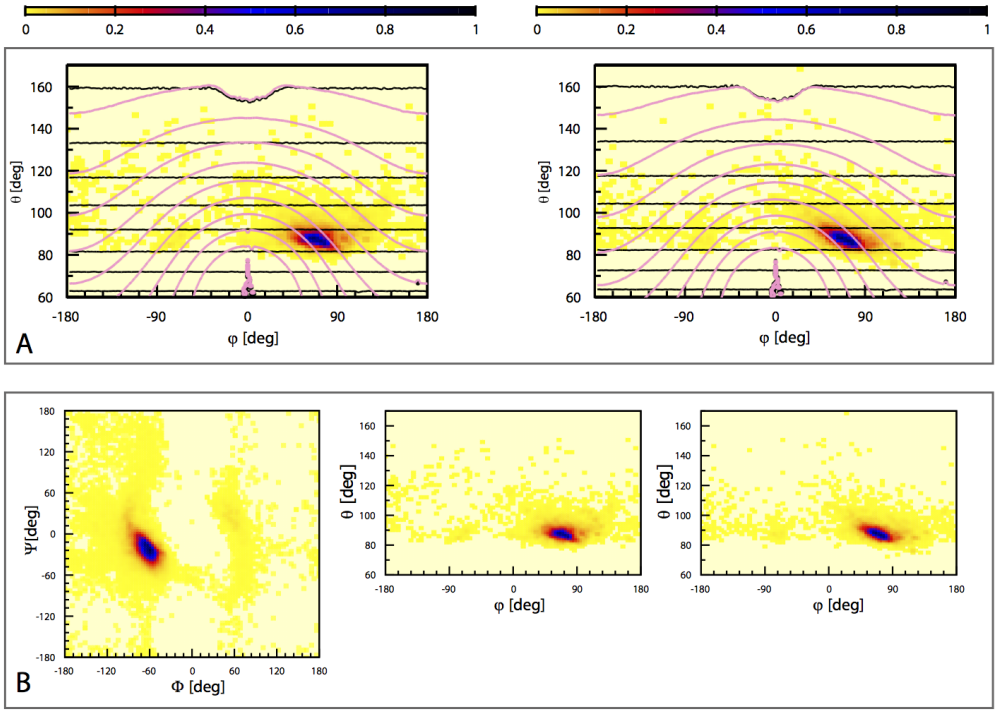


Figure 19: Conformational plots for 3_{10}-helix. Panel A: Correlation plots for ($\theta-$,ϕ) on the left and ($\theta+$,ϕ) on the right for NMR structures. In pink the constant r_{1-4} lines are superimposed, while the same is done for the r_{1-3} constant line in black. Both the two datasets are composed extracting structures directly from the PDB secondary structure information. Color bar is on the top. Panel B: Conformational plots for a dataset of NMR 3_{10}-helices. The distinction of the structure is made with the direct information of the PDB file. Left (Φ,Ψ) map, center ($\theta-$,ϕ) map, right ($\theta+$,ϕ) map. The data are normalized to the maximum count.

in the distance dependent distribution ($P(r)$) the first peak, corresponding to the distance between two consecutive C_α, is at 3.8 Å. There are not peptide bonds in cis conformation forming a regular α-helix. This is also confirmed by the correlation line in the (r_{1-3}, θ)-map in figure 21, where for the α-helix there is only one correlation line (pink correlation plot), while for the dataset of unstructured proteins there is a net component also of cis peptide bond (yellow correlation plot).

Each distribution is characterized in table 9. Other distributions are reported in Appendix C

In figure 22 (Panel B) the Ramachandran plot and the ($\theta-$, ϕ) and ($\theta+$, ϕ) maps for the dataset of NMR are reported. As it can be seen, the directionality of the peptide and the chirality of the C_α are not so relevant in this case.

As previously explained (see Chapter 1), the α-helix has a pattern of H-bonds that binds the CO group of one amino acid and the HN group of the amino acid four after along the chain, i.e. (i,i+4). As in the case of 3_{10}-helix, at least two distances are related to this bond in the C_α representation, namely r_{1-4} and r_{1-5}. In figure 22 the correlation lines computationally calculated are superimposed to the (θ,ϕ) maps of X-ray to outline that the slope of the correlation plot is midway between the correlation lines at constant r_{1-4} and r_{1-5}.

In figure 23 (panel A) the different sampling of this space available for X-ray

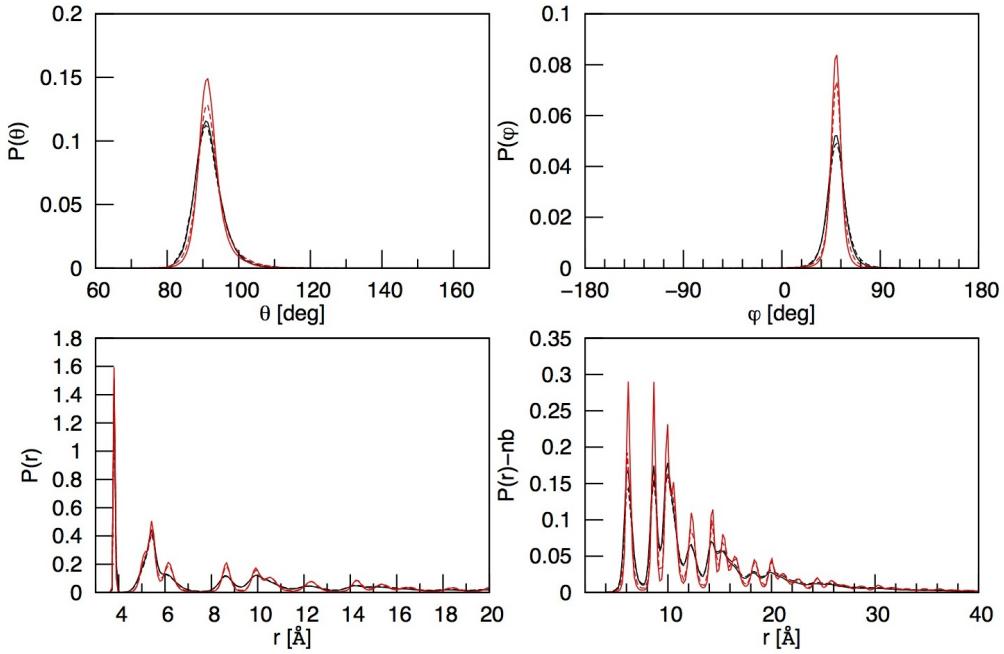


Figure 20: Internal DOFs distributions for α-helices. These four graphs show the distributions of θ (top-left), ϕ (top-right), $P(r)$ for r distance between every two C_α (bottom left) and $P(r) - nb$ for r distance between every i and j with $j > i + 2$ (bottom right). Different lines are for different datasets: red for Xray and black for NMR, solid for DSSP and dashed for PDB direct information.

and NMR is shown. The maps on the top line in fact are more extended in the space than the ones in the bottom line, where the distribution is extremely concentrated. This is the effect of the crystallization in a static configuration of the structures.
All these data confirm that the α-helix is a well defined structure, because the distributions are well peaked and the correlation plots are localized in the (θ,ϕ)-plane.

3.5.3 π-Helix

The exact definition of an ideal π-helix is a disputed issue. This is confirmed by the composition of the datasets. First of all, there were only few (less than ten) helices directly recognized as π in the PDB secondary structure entries. Using the DSSP algorithm, 550 π-helices were found in NMR data and 179 in Xray data. Another problem is the length of the recognized π-helices. The longest helix in the NMR dataset has 5 C_α, while in the Xray dataset 7. This is particularly problematic because the characterizing H-bonding pattern is $(i, i+5)$. Then, there are no data for the r_{1-6} distribution in the NMR dataset and for the Xray one they are not reliable. Looking at the distributions in figure 24, it becomes even clearer how the identification of the structure is tough task. The distribution are large and for different experimental methods different maxima are found (see table 10). Other distributions are reported in appendix C
The correlation plots (figure 25) have similar problems. The center of the populated region is really near that of the α-helices and the distinction between the two areas is not possible. However, these data are not completely reliable,

Table 9: Distribution parameters for α-helix. These values are obtained using Octave [70]. The interquartile range (iqr) was computed as difference between the upper and lower quartile.

			θ[°]	ϕ[°]	r_{1-3}[Å]	r_{1-4}[Å]	r_{1-5}[Å]	r_{1-6}[Å]
X-ray	DSSP	max	91	50.4	5.42	5.15	6.14	8.66
		iqr	2.2	7.2	0.27	0.36	0.36	0.27
	PDB	max	91.35	50.4	5.42	5.15	6.14	8.66
		iqr	5.95	9	0.36	0.36	0.45	0.45
NMR	DSSP	max	90.8	50.4	5.42	5.15	6.14	8.57
		iqr	5.5	12.6	0.36	0.45	0.63	0.54
	PDB	max	90.8	50.4	5.42	5.15	6.05	8.66
		iqr	5.5	12,6	0.36	0.54	0.63	0.63

Table 10: Distribution parameters for π-helix. These values are obtained using Octave [70]. The interquartile range (iqr) was computed as difference between the upper and lower quartile.

			θ[°]	ϕ[°]	r_{1-3}[Å]	r_{1-4}[Å]	r_{1-5}[Å]	r_{1-6}[Å]
X-ray	DSSP	max	100	28.8	5.78	5.96	4.88	6.32
		iqr	18	28.8	0.9	1.08	0.72	0.72
NMR	DSSP	max	100.33	36	5.78	5.6	4.52	
		iqr	18.34	36	0.72	1.44	1.44	

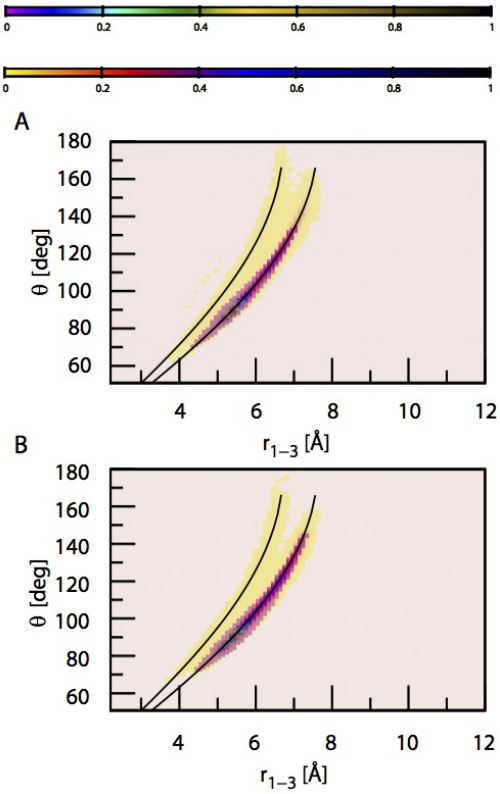


Figure 21: (r_{1-3}, θ)-maps for unstructured proteins and α-helices. In panel A, there is the overlapped correlation of X-ray α-helix and unstructured dataset. In panel B the same correlations are showed for the corresponding NMR datasets. The analytical correlation lines at constant r_{1-3} are superimposed. The correlation area for the α-helix is pink (higher color bar on the top), while the correlation area for the unstructured dataset is yellow (lower color bar on the top). The data are normalized to the maximum count.

because of the uncertain definition of the π-helices.

3.5.4 Unstructured Chains

The dataset of the unstructured proteins were constructed including lower possible percentage of helices and sheets (i.e. less than 20%). 1155 proteins for Xray and 837 for NMR were found. The intention here is to ideally analyse the situation in which no secondary structure is formed and then no H-bonds are established in the chain. In figure 26 only the NMR data are reported. These are in fact the most unstructured, because these proteins are less easily crystallizable. However, there are still two more evident peaks (especially for the dashed line) at the values of θ specific for helices and sheets, indicating that even unstructured dataset still conserves a bias towards the main secondary structures.
The distribution of the dihedral angle has the same meaning. It spans all the ϕ points, but there are more counts near the helical and the extended structures. Looking at the solid line in the distance dependent "non-bonded" distributions (bottom left in figure 26), there is always a repeated pattern for the solid line (directly PDB identification of the structure) probably attributable to a strong

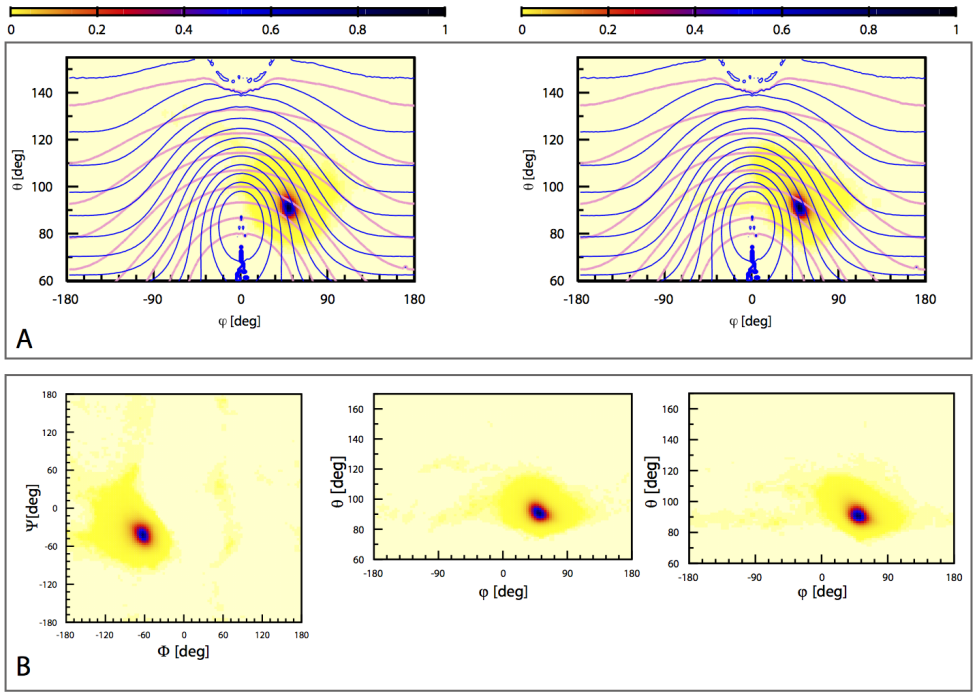


Figure 22: Conformational plots for α-helix. Panel A: Correlation plots for (θ−,ϕ) on the left and (θ+,ϕ) on the right for X-ray structures. In pink the constant r_{1-4} lines are superimposed, while the same is done for the r_{1-5} constant line in blue. Both the datasets are composed extracting structures directly from the PDB secondary structure information. Panel B: Conformational plots for a dataset of NMR α-helices. The distinction of the structure is made with the direct information of the PDB file. Left (Φ,Ψ) map, center (θ−,ϕ) map, right (θ+,ϕ) map. Color bars on the top. The data are normalized to the maximum count.

presence of extended structures (β-strand more than helices). On the other hand the dashed (DSSP) distribution shows only two main peaks.
In the (r_{1-3}, θ) correlation plot in figure 21 two correlation lines are visible. This is due to the presence of peptide bonds in the cis conformation for the dataset of the unstructured proteins. It is shown that the correlation formula for the trans and for the cis conformations are perfectly superimposed to the corresponding correlated areas.
In figure 27 there are the Ramachandran plot and the CG conformational plots for the unstructured proteins. Here the observations made for the distributions are confirmed. There are clearly distinguishable region weakly helical and region of extended structures, while the conformational space is however widely sampled. Other distributions and correlations are reported in Appendix C.
More detailed information for the unstructured proteins data set are given in the 3D $(r_{1-4}, \theta+, \phi)$ map. In the 3D map the highly populated regions distributed in the volume can be visualized with iso-values surfaces (in grey, in the upper part of figure 28), making the separation between secondary structures even more immediate than in the 2D map. This is a consequence of choosing immediately physically interpretable variables for the 3D map building. In fact, the r_{1-4} is a particularly important variable especially for helices, being associated to the formation of local hydrogen bonds stabilizing them. For this reason,

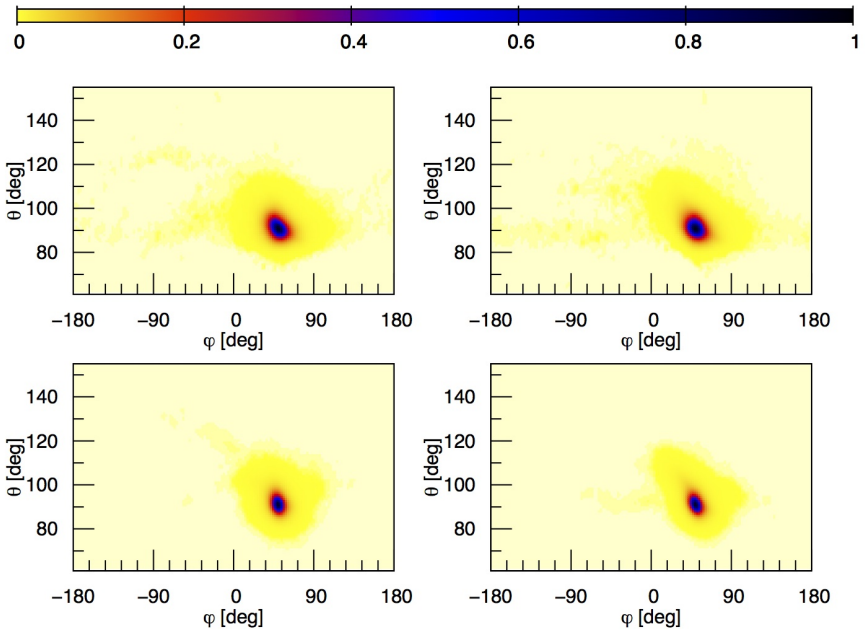


Figure 23: (θ, ϕ) maps for α-helix. $(\theta-,\phi)$ on the left and $(\theta+,\phi)$ on the right. In the top line there are data from the NMR dataset, while in the bottom line the X-ray ones. All the two datasets are composed extracting structures directly from the PDB secondary structure information. Color bar is on the top. The data are normalized to the maximum count. Here are reported the same data of figure 22 to underline the difference among the two experimental datasets.

an alternative visualization of the 3D map by means of the iso-variable sections, e.g. the iso-r_{1-4} (lower part of figure 28) is also particularly interesting. Figure 28 reports the sections corresponding to three relevant values of the single variable r_{1-4} distribution (red dots A, B, C in the top right plot). 2D maps of these slices are also reported in the three bottom plots (corresponding letters) each with its colors bar. By definition, the single variable r_{1-4} distribution (right upper plot) is the integral over θ and ϕ of the 3D map. The 3D representation is generated with VMD [45] from the CUBE file. The surface corresponding to the helical region (residually populated also in the "unstructured" data set) is a roughly ellipsoid shape located at $r_{1-4} = 5.75$Å. This is also confirmed by the $r_{1-4} = 5.75$Å section (plot A figure 28), in which a high concentration in the helical area is observed. In this plot it is also possible to observe that a upside-down parabolic shape is populated (red-blue shades). As previously seen, this kind of correlation is induced among the variables θ and ϕ keeping constant r_{1-4}. At higher levels of r_{1-4} other structures appear, first a transition region (plot B) and then the extended structures region (plot C).

3.5.5 Strand and β-Sheets

The four datasets corresponding to different combination of NMR or Xray and secondary structure algorithm (PDB or DSSP) were built. There were found

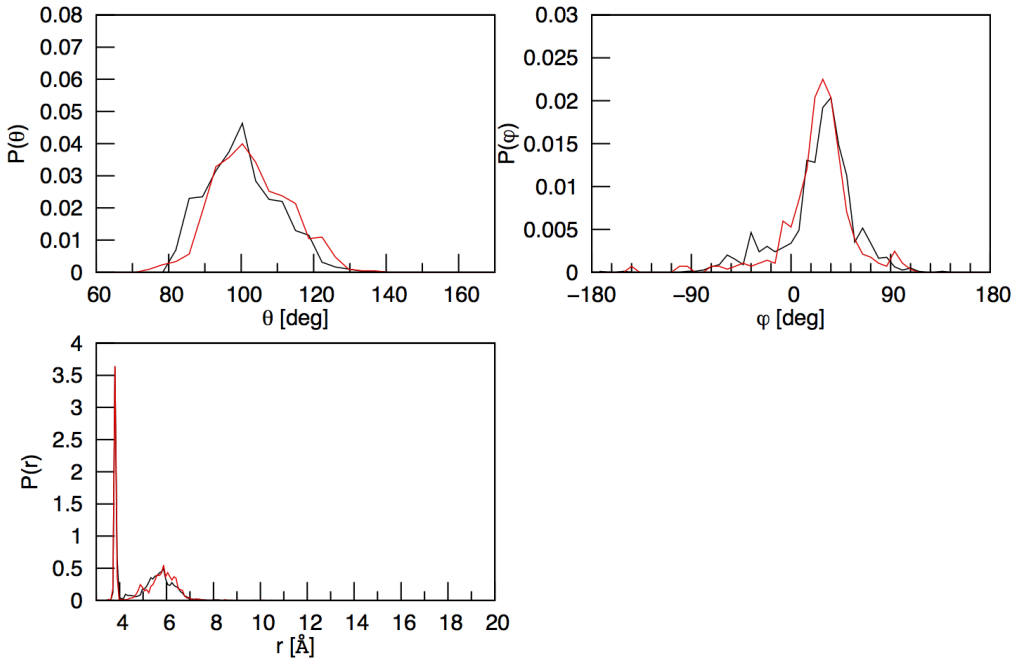


Figure 24: Internal DOFs distributions for π-helices. Distributions of θ (top-left), ϕ (top-right), $P(r)$ for r distance between every two C_α (bottom left). Different lines are for different datasets: red for Xray and black for NMR, solid for DSSP

2270 proteins for the NMR dataset and 15270 for the Xray one.
Using the DSSP algorithm, it is not possible to distinguish if a strand is parallel or antiparallel to the previous one in a sheet. Furthermore, in the PDB format there is also specified the best H-bond solved, i.e. which atom is bonded with which atom in the previous strand. These features lead to the chioce of analysing better the datasets built using the information directly included in the PDB. However, in appendix C there are the graphics for the four datasets.

In figure 29 the bond angle and dihedral angle distributions (top line) both for parallel and antiparallel strands are shown. These distributions are quite similar, as it is confirmed also by the table 11. In figure 29 the dihedral distribution shows a little peak in $\phi = 0$. This happens for two reasons: first, the presence of buldges in strands (see cap 1), second the presence of pieces of turns binding two different strands, wrongly recognized as strands. The same structures are responsible for the second little peak at ~ 7Å in the distribution of r_{1-4} (bottom left).

The pattern of H-bonds for both parallel and antiparallel sheets is inter-strand instead than intra-strand, as it was for the helical structures. Then, to reproduce these bonds, it would be necessary to analyse these interstrand distances near the H-bonds. Two different repeated structure for parallel and antiparallel sheets were identified, as in figure 30. Then, using the information contained in the PDB, the positions of all the H-bonds between two subsequent strands were labeled. Because the H-bonds are between the CO and HN groups of amino acids on two different strands (see figure 30), all the distances around the two beads supposed H-bonded were calculated.

In figure 31 there are the distributions for some of the interstrand distances defined in figure 30. Other distributions are reported in Appendix C. In the graphics for parallel sheets (panel B) there are tails, not real marked. These are

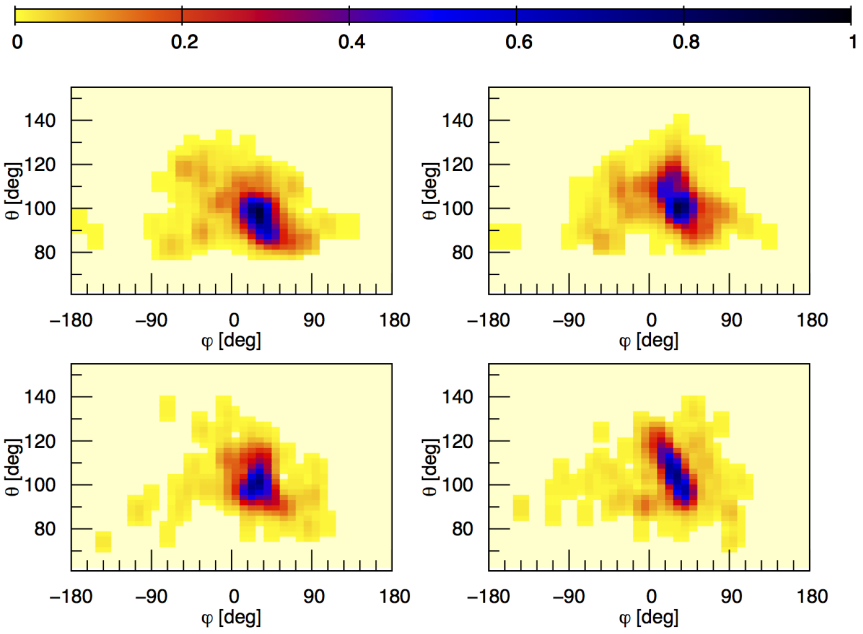


Figure 25: (θ, ϕ) maps for π-helix. Correlation plots for $(\theta-,\phi)$ on the left and $(\theta+,\phi)$ on the right. In the top line there are data from the NMR dataset, while in the bottom line the X-ray ones. All the two datasets are composed extracting structures with the DSSP algorithm. Color bar is on the top. The data are normalized to the maximum count.

due to the twisting of the sheets. This effect is more evident for the antiparallel sheets (panel A), which can assume many different conformations (see Chapter 1). In the distribution for d_4 in panel A of figure 31, there are clearly discernible two secondary peaks other than the principal one. These are the distances assumed by the curving atoms in particularly twisted sheets.

In figure 32 there are the Ramachandran plot and the conformational CG plots (θ, ϕ) for a dataset of NMR BDB β-strands. The extended regions are clearly discernible. As previously noted, the directionality of the chain is here strongest than in the case of helical structure. The two CG maps in fact are not superimposable. The correlation relations in this case are not valid, confirming that the H-bonds, which causes the correlations, are not intra but interstrand.

3.6 SUMMARY

In this Chapter, the statistical analysis constituting the theoretical foundation of the model was discussed. The next step is to build a minimalist model with a force field composed by statistical potentials. These potentials will be initially derived based on BI from these distributions, then optimized targeting them and the correlation plots.

The dataset that better describes each secondary structure has to be identified. For the α-helices all the datasets leads to the same distributions and correla-

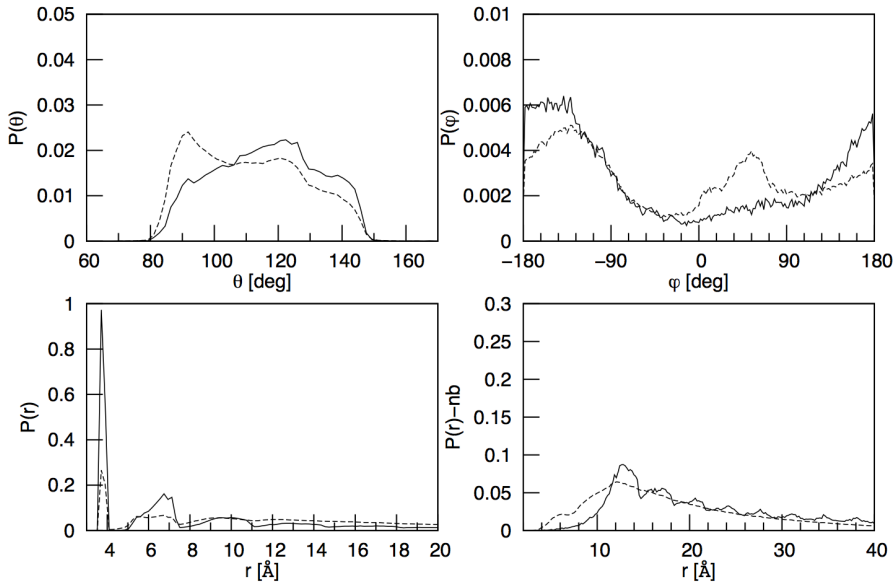


Figure 26: Internal DOFs distributions for unstructured fragments. Distributions of θ (top-left), ϕ (top-right), $P(r)$ for r distance between every two C_α (bottom left) and $P(r) - nb$ for r distance between every i and j with $j > i + 2$ (bottom right). Different lines are for different datasets: black for NMR, solid for DSSP and dashed for PDB direct information.

tions, any choice is thus valid. In the case of the 3_{10}-helices, the data solved with NMR in solution are the most realistic and it was here demonstrated that the DSSP dataset has less impurities of α-helices than the PDB one. For the π-helices the X-ray data has to be chosen, because in the NMR ones there are structures up to 5 amino acids. However, also the X-ray DSSP dataset is not reliable, because the logest structure has 7 amino acids.
From the distributions obtained in the unstructured datasets, it could be argued that the β-sheets and the helices are the two possible conformations that an unstructured chain can assume. The parameterization of a field able to reproduce all the secondary structures, can be divided in two parts. The first part is the accurate reproduction of the helical area of the conformational space. Here, the three helices have quite similar θ,ϕ parameters, but they are distinguished on the basis of their intra-strand H-bonding pattern. The second part is the optimization of the area of the (θ,ϕ)-space allowed to the extended conformations. Here, the structures are stabilized by inter-strand H-bonds, then the correlations dependent on specific intra-strand distances cannot be used.
In the following Chapter, the fields for the accurate reproduction of the different types of helices will be optimized.

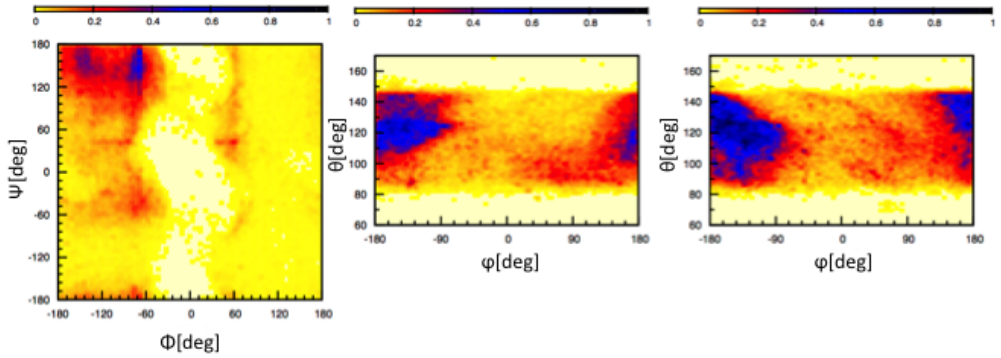


Figure 27: Conformational plots for unstructured fragments solved with NMR. The distinction of the structure is made with DSSP algorithm. Left (Φ,Ψ) map, center (θ−,ϕ) map, right (θ+,ϕ) map. The data are normalized to the maximum count. Color bar on the top.

Table 11: Distribution parameters for parallel and antiparallel strands. These values are obtained using Octave. [70]. The interquartile range (iqr) was computed as difference between the upper and lower quartile.

				θ[°]	ϕ[°]	r_{1-4}[Å]	r_{1-5}[Å]
Antiparallel strands	X-ray	PDB	max	121.5	-167.4	10.1	13.43
			iqr	12.1	115.2	0.72	1.08
	NMR	PDB	max	123.8	-163.8	10.19	13.25
			iqr	16.06	127.8	0.72	1.08
Parallel strands	X-ray	PDB	max	119.95	-163.8	10.01	13.25
			iqr	15.95	41.4	0.63	0.9
	NMR	PDB	max	121.05	-163.8	10.01	13.16
			iqr	15.4	41.4	0.72	0.9

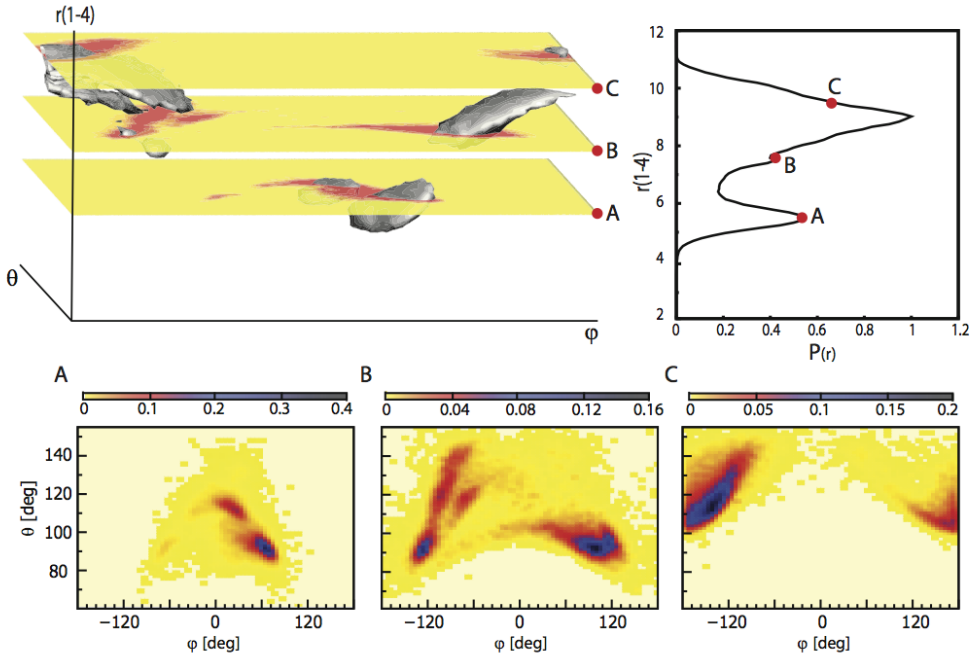


Figure 28: $(r_{1-4}, \theta+, \phi)$ map for the X-ray PDB unstructured proteins. An iso-surface (level=120) is represented in grey and three $r_{1-4} = const$ sections are in color. The three r_{1-4} values are chosen corresponding to three relevant values of the single variable r_{1-4} distribution (red dots A, B, C in the top right plot). 2D maps of these slices are also reported in the three bottom plots (corresponding letters) each with its colors bar. By definition, the single variable r_{1-4} distribution (right upper plot) is the (renormalized) integral over θ and ϕ of the 3D map. The 3D representation is generated with VMD from the CUBE file.

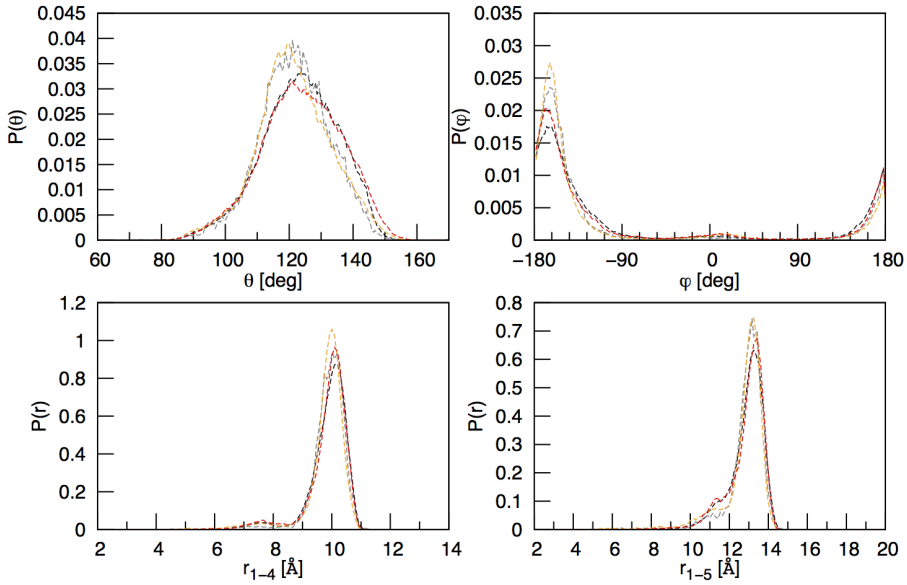


Figure 29: Internal DOFs distributions for parallel and antiparallel strands. Distributions for bond angle θ (top left), dihedral angle ϕ (top right), r_{1-4} (bottom left) and r_{1-5} (bottom right). Here are reported only structures directly identified from the PDB entries (dashed lines). Red: Xray antiparallel strands. Black: NMR antiparallel strands. Orange: Xray parallel strands. Grey: NMR parallel strands.

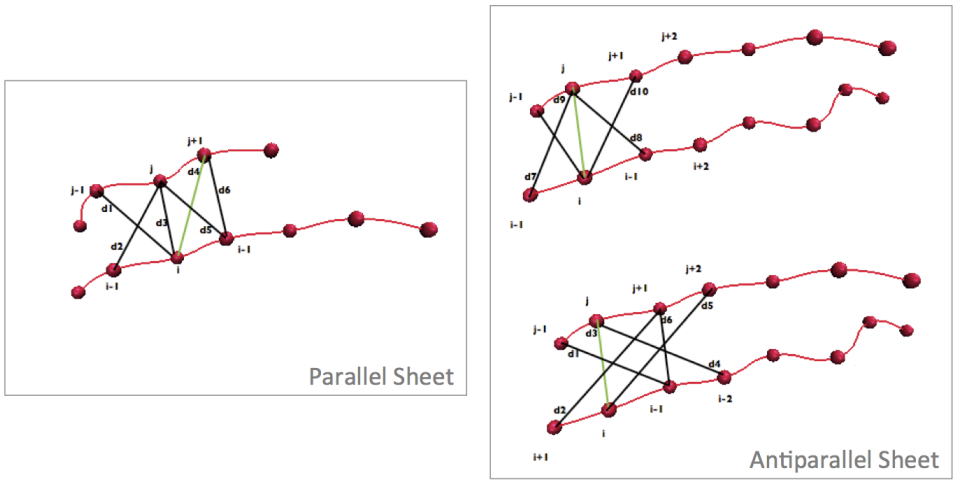


Figure 30: Schematic representation of two strands. Here are showed the chains for a parallel (left) and for an antiparallel (right) two strands sheet, with only the C_α explicit. The green lines connecting the two strands are the supposed H-bonds. Black lines near the hydrogen bonds are the distances measured to build the distributions of figure 31.

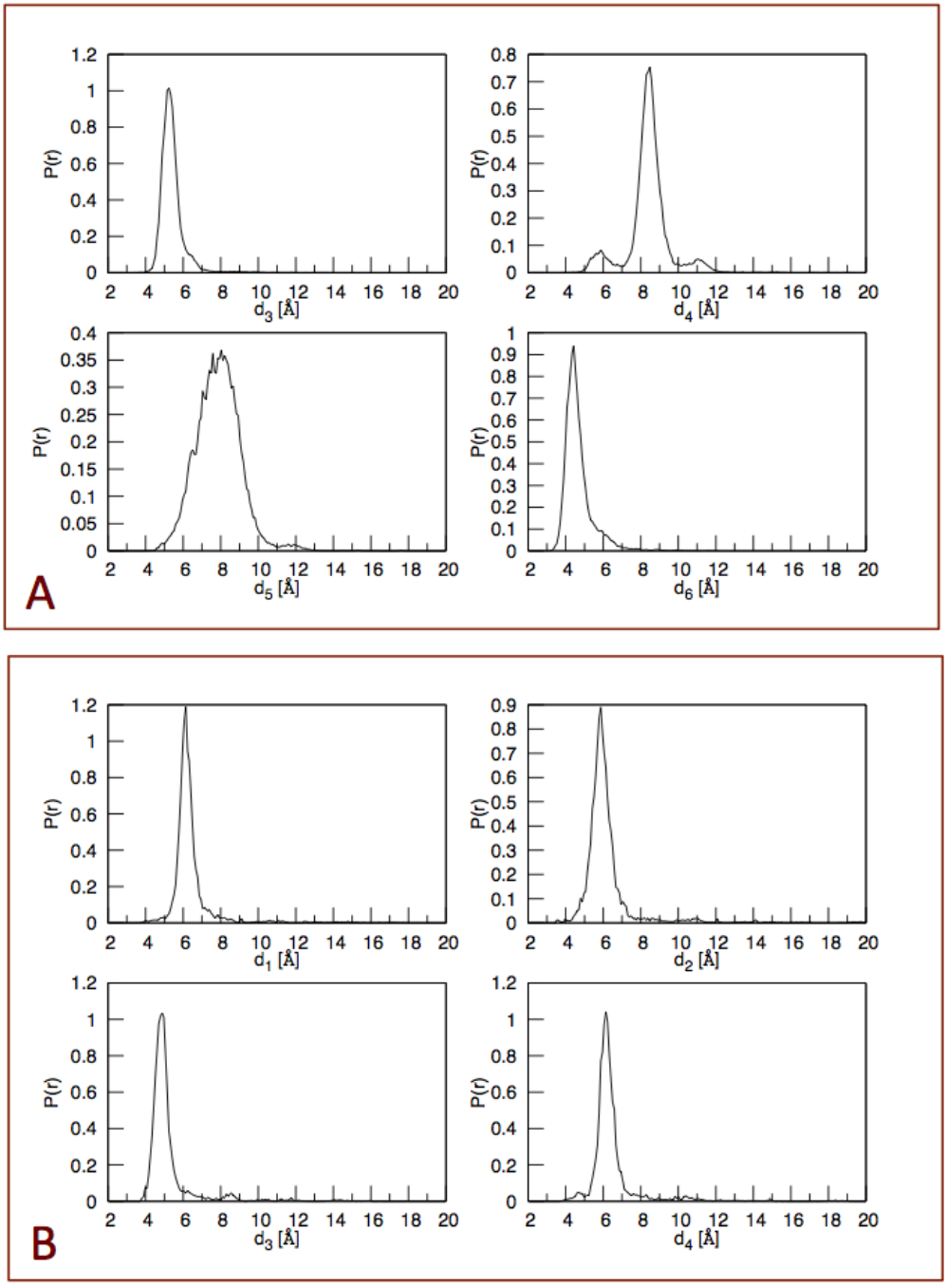


Figure 31: Inter-strand distances distributions. Panel A: Distributions dependent on distances between two antiparallel strands. Panel B Distributions dependent on distances between two parallel strands. Different distances are identified with the same name as in figure 30. Structures are solved with NMR and identified directly from the PDB.

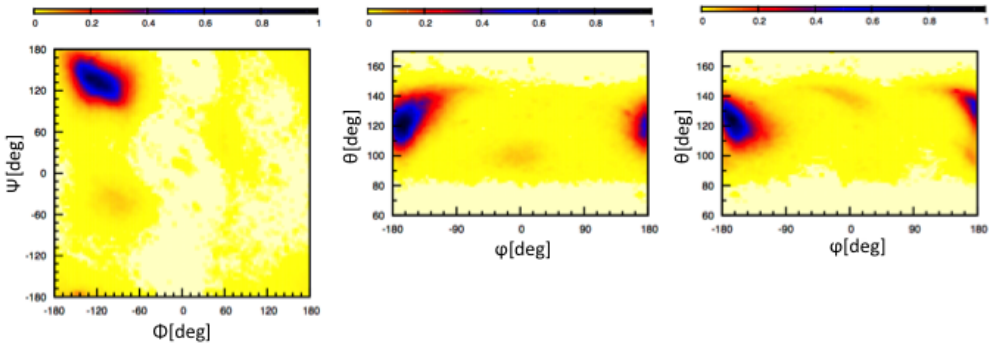


Figure 32: Conformational plots for extended strands solved with NMR. The distinction of the structure is made using the PDB entries. Left (Φ,Ψ) map, center (θ−,ϕ) map, right (θ+,ϕ) map. The data are normalized to the maximum count.

4 THE MINIMALIST MODEL II: PARAMETERS OPTIMIZATION FOR HELICES AND SIMULATIONS RESULTS

In this Chapter a minimalist model capable of reproducing the structure and dynamics of the different types of helices is proposed. After an introduction to the strategy used to optimize the force fields, the simulation results are reported and commented.

4.1 INTRODUCTION

The model here reported is aimed at accurately reproducing the secondary structures of a peptide. The statistical analysis reported in the previous Chapter has revealed that two different areas are discernible in the conformational space (θ, ϕ): one for the helical structures and one for the extended structures. From a physical-chemical point of view, these conformations are characterized by different patterns of hydrogen bonds, specifically occurring intra-strand (for helices) and inter-strand (for extended structures forming sheets). This suggests to add into the FF specific terms to describe these patterns to these FF terms describing the behavior of a polypeptide in absence of specific H-bonding patterns.
This work focuses on the reproduction of the helical part of the conformational map as a first step. The three helices (3_{10}, α, π) occupy a continuous area in the (θ,ϕ)-plane. This suggests a model in which the U^{θ} and U^{ϕ} terms are the same for all the helices and the distinction among them depends on the presence of distance dependent force field terms mimicking the specific intra-strand hydrogen bonds (different for the helices).
The goal is here achieved in two steps: first, fields with only the conformational terms and parameters are obtained directly from the Boltzmann inversion of the correspondent distributions of the three helices; second, a force field with equal U^{θ} and U^{ϕ} for the three types of helices is optimized and other new terms, describing the hydrogen bond, are included and tuned. The field that optimally reproduces the distributions and the correlations specific for each kind of helix is obtained. The field is then tested for different length of the peptide from 11 to 20 amino acids and the resulting data are shown.

4.2 STARTING STRUCTURES

As described in the previous Chapter, datasets of every secondary structure present in the RCSB Protein Data Bank were collected and analysed, using SecStAnT. It further allows to select larger and more regular helices. From the

NMR datasets, fragments of typical length for each type of helix are then extracted. For the α helix a typical length is between 10 and 20 amino acids. For the 3_{10}-helix this is 7 amino acids. Two ideal structures were chosen, one of 7 C_α, to test the fields in a realistic system, and one of 17 C_α, to assess stability on larger peptides. Finally, for the π helix in the NMR dataset there are fragments up to 5 amino acids, while in the Xray one up to 7. The simulation actually tested on a helix of 12 amino acids, built with the software Avogadro [29], a free cross-platform molecule editor.
These structures, considered representative for each helix type, are used as starting and reference structure. Upon coarse graining to the C_α-based representation, and since at this level no sequence information is included, to each bead an average AA mass is assigned (115 a.u.).

4.3 FORCE FIELD AND TOPOLOGICAL CONNECTIVITY

In the model here proposed, the distance between two subsequent beads is fixed at the length of the trans peptide bonds, i.e. 3.8 Å. They are treated as constraints. As anticipated in the previous section, the FF consists of an intrinsic conformational part depending on the pseudo-bond angles θ_i between three subsequent C_αs and on the pseudo-dihedrals between four subsequent C_αs. In addition, terms describing the H-bond are introduced and treated as topologically connected (with the known helical topology) with detachable functional forms, to allow H-bond breaking. The general form of the force field is then:

$$U = U^\theta(\theta_i) + U^\phi(\phi_i) + U^{hb}(r_{i,i+n}) + U^{nb}(r_{i,j}) \tag{31}$$

where $r_{i,j}$ is the distance between beads i and j, θ_i is the bond angle between beads $(i-1, i, i+1)$ and ϕ_i is the dihedral angle between beads $(i-1, i, i+1, i+2)$. The non bonded term (U^{nb}) is taken from the optimized force field in [85] and will not be further refined in this work. It was however checked that its contribution is the less relevant, being the weights of the other terms much higher in the global energy. The following functional forms have been used, which reproduce the Boltzmann inverted statistical data:

$$U^\theta = \sum_\theta u^\theta = \sum_\theta k_\theta \frac{1}{2}(\cos\theta - \cos\theta_0)^2 \tag{32}$$

$$U^\phi = \sum_\phi u^\phi = \sum_\phi A_\phi[1 - \cos(\phi - \phi_0)] \tag{33}$$

$$U^{hb} = \sum_{hb} u^{hb} = \sum_{i,j \in S_{hb}} \epsilon\{[1 - exp(-\alpha(r_{i,j} - r_0))]^2 - 1\} \tag{34}$$

Depending on the type of helices, different couples i,j are included in the set S_{hb} of the hydrogen bonding terms. For instance, for α-helices the H-bond is between the NH group of the amino acid i and the CO group of AA $j = i+4$. This implies that in the C_α-based model both the (i,i+4) and (i,i+5) couples of C_αs are involved, at least, in a H-bond. Also the number of couples included is optimized here. r_0, ϵ, and α are helix type dependent and, for each helix, they depend also on which interaction (i,i+n) they represent.

Finally, the u^{nb} term has a double-well form, created by combining single well forms (f_1, f_2) according to:

$$u^{nb} = \frac{1}{2}\left[f_1(r_{i,j}) + (f_2(r_{i,j}) - \Delta)\right] - \frac{1}{2}\sqrt{\left[f_1(r_{i,j}) - (f_2(r_{i,j}) - \Delta)\right]^2 + \lambda^2} \quad (35)$$

where Δ is the energy difference between f_1 and f_2 at $r \to \infty$ and λ is kept very small and is needed to smoothen the discontinuity at the barrier. u^{nb} is applies to all the beads of the chain. f_1 and f_2 in this case are two morse potentials:

$$f(r) = \epsilon\{[1 - exp(-\alpha(r - r_0))]^2 - 1\} \quad (36)$$

The parameters used for f_1 are: $\epsilon_1 = 25 kcal/mol$, $\alpha_1 = 0.1 Å^{-1}$ and $r_0 = 6.1 Å$; while for f_2 are: $\epsilon_2 = 0.1 kcal/mol$, $\alpha_1 = 0.7 Å^{-1}$ and $r_0 = 9.64 Å$. Moreover, $\Delta = 25.3$ and $\lambda = 0.1$. As previously said, these parameters were optimized in [85] for non bonded inter-chain interactions of a "generic" amino acid. In addition, as it will be clearer in the following, this term has a minor role in determining the intra-helical structure, due to the presence of other force field terms.

4.4 PARAMETERIZATION STRATEGY

The initial guess for the parameters of each force field terms was directly extracted from the distributions of the correspondent internal variables. As seen in Chapter 2, using equation 10, the Statistical Potentials could be derived from the distributions. In equation 10 $P_0(Q_l)$ is the probability distribution of Q_l, internal variable, in a reference state with non interacting particles. SecStAnT is able to generate the Boltzmann inversion of each distribution, but also to generate the potential of mean force for different ideal reference states. The infinite system of non interacting particles has been chosen as the ideal reference state. This means:

$$P_0(\theta) = \frac{\sin\theta}{2} \quad (37)$$

$$P_0(\phi) = 2\pi \quad (38)$$

$$P_0(r) = 4\pi r^2 \quad (39)$$

For every internal variable directly used in the force field, its inverted distribution was fit with the most appropriate analytical function that will be the same directly used in the simulation. In figure 33 the fit for the α-helix bond angle and dihedral angle is showed.

4.5 SIMULATIONS PROTOCOL

The simulations were performed using the DL_POLY software package [76] (see appendix D). In-house developed software was used to build the input and to analyse the output files. DL_POLY input files are CONFIG, FIELD and CONTROL, while among the output files there are the trajectory (HISTORY) and energy terms (STATIS). The coordinates identified as ideals in the CONFIG

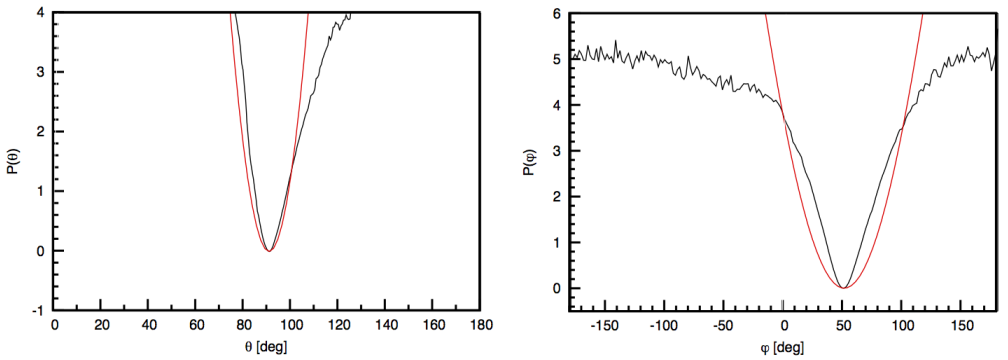


Figure 33: Examples of fit. Fit procedure of the PMFs for the distributions of bond angle (left) and dihedral angle (right) of the α-helix NMR dataset. The helical structures are obtained with the PDB direct information. In black there are the U(Q) obtained directly with equation 10, in red there are the functions of fit. The parameters for the fit are the same used for the initial states.

Table 12: Summary of the protocol used for the simulations. The different force field tested are distinguished.

FF	Phase	T(K)	Integrator	Δ t (ps)	Duration (ns)
FF1	Minimization		Leap Frog	0.001	1
	Equilibration	300	Leap Frog	0.001	5
	Production Run	300	Leap Frog	0.001	15
FF2	Minimization		Leap Frog	0.01	1
	Equilibration	300	Leap Frog	0.01	5
	Production Run	300	Leap Frog	0.01	50

format file are converted using only the C_α positions. Starting from the initial structure, a minimization was performed for 1 ns to relax the system in its equilibrium minimum. An equilibration is then performed for 5 ns, in which the system is gradually heated up to 300 K. The required temperature is maintained by coupling the system to a Nose-Hoover thermostat (see Appendix B). For all the runs the leap frog Verlet integrator is adopted. In table 12 there is a summary of the simulation protocols used for the different force field tested.

4.6 FIRST FORCE FIELDS SET

The first fields generated are composed only of the conformational terms (θ,ϕ), specific for each type of helix, with the parameters directly obtained from the Boltzmann inversion, with no explicit hydrogen bonding term. The first force field used (FF1) has the form:

$$U = U^{\theta} + U^{\phi} + U^{nb} \tag{40}$$

Different sets of parameters for different helices are in table 13.

Table 13: Parameters for the initial force field tested. Force constant are given in $kcal/(molÅ^2)$, values of equilibrium angles and dihedral in deg (°).

FF1	3_{10}-helix	α-helix	π-helix
u^θ	$k_\theta = 100$	$k_\theta = 200$	$k_\theta = 20$
	$\theta_0 = 88$	$\theta_0 = 90.08$	$\theta_0 = 100$
u^Φ	$A_\Phi = 10$	$A_\Phi = 35$	$A_\Phi = 10$
	$\phi_0 = 64$	$\phi_0 = 50.4$	$\phi_0 = 36$

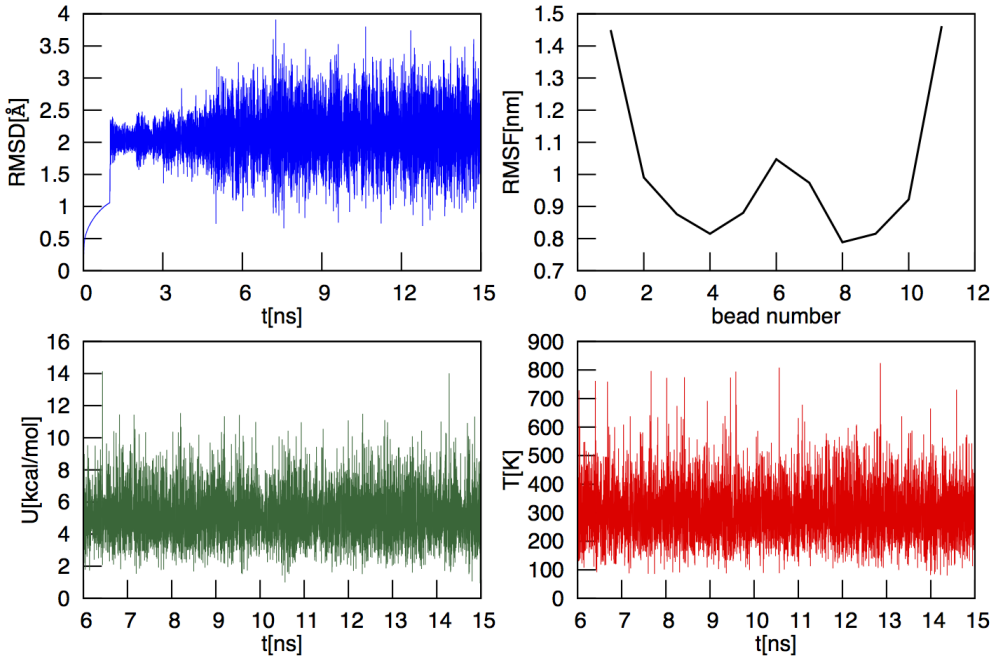


Figure 34: Simulation results for the initial field of the 3_{10}-helix. The system used is a peptide of 11 amino acids. Top left: RMSD of the entire simulation. The equilibration process is also reported. Top right: RMSF of all the peptide beads. Bottom left: potential energy U of the system for only the production run. Bottom right: temperature (K) of the system for the production run.

4.6.1 3_{10}-Helix

The NMR structures, recognized with DSSP algorithm were used to extract the parameters for the FF1 in table 13 for the 3_{10}-helix. The high values of the constant forces chosen force the use of a timestep of 0.001 ps. The simulation protocol is summarized in table 12. The output values for energy and temperature are shown in figure 34, where it is also possible to see the stability of the fluctuations (RMSD) of the structure in the production phase. In the top right plot of figure 34 there is the RMSF. Besides the obviously mobile extremes, also the central part of the helix fluctuates more than the two intermediate segments between center and extremes. This means that there are two some sort of stable pins and the central part can flex itself more freely than those pins. This relationship is also observed in real structures, which bend in the same way.
These parameters were tested for two different length of 3_{10}-helix: 11 amino acids and 17 amino acids. Here, all the results for the 11 amino acid long he-

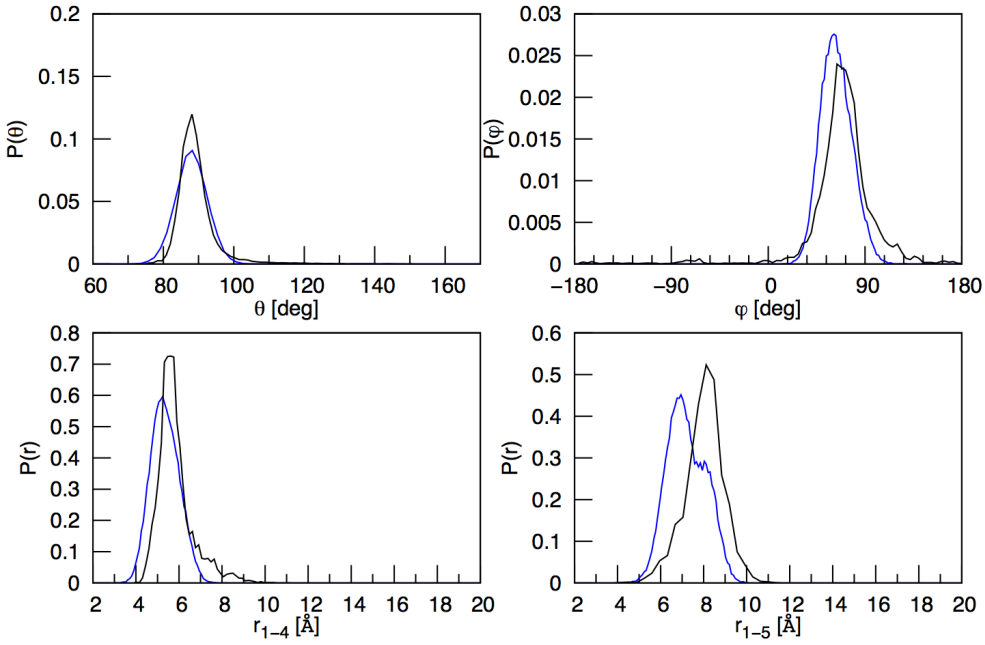


Figure 35: Comparison of experimental and simulated distributions for the 3_{10}-helix. Black: experimental distributions obtained from the dataset of 3_{10}-helix solved with NMR and recognized with DSSP. Blue: distributions obtained with simulation of the 11 beads long 3_{10}-helix with FF1.

lix are reported. In appendix D the results for the other length are included. The distributions in figure 35 are calculated with SecStAnT, even those from simulations. In this case, the frames are considered as individual structure. The agreement between experiment and simulation is rather rough, which was some how expected, since the hydrogen bond terms are not included in this preliminary model.

The (θ,ϕ) plots (said also conformational or correlation plots) in figure 36 reports simulation data (lower row) compared to experimental data (upper row). As it can be seen, the allowed areas in the simulations are too large and differently shaped rather elliptic with axes orthogonal to θ and ϕ axis, while the experimental one is elongated along an oblique line.

In conclusion, this field is not accurate enough to reproduce the 3_10-helix fine structure. However, the responsibility of this lack of accuracy cannot be attributed to the values of the conformational parameters: the distributions of θ and ϕ are quite well centered and the force constant are already in a range that is physically too large. The stabilization of the helix has thus to be attributed to some other interaction.

4.6.2 α-Helix

The parameterization used for the FF1 of α-helices is reported in table 13. The constant forces are even higher than in the 3_{10} helix case. The simulation protocol used is reported in table 12. The simulated polypeptide is 20 beads long.

In figure 37 the outputs of the simulation are reported. In the RMSF plot is

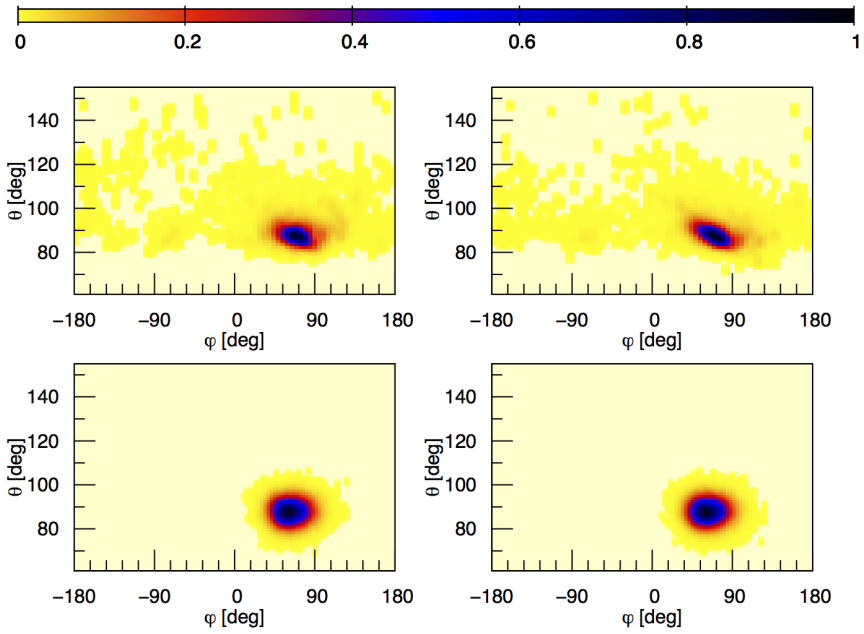


Figure 36: Comparison of experimental and simulated (θ,ϕ) plots. Top line: (θ−,ϕ) map on the left and (θ+,ϕ) plot on the right for the experimental NMR DSSP dataset. Bottom line: correlation plots corresponding to the top line for the simulation results. The simulation was performed with the 11 beads long 3_{10}-helix and with FF1. Color bar on the top.

always visible the effect of the high flexibility of the central part of the chain.

In figure 38 there is the comparison between the experimental data (black) and the data obtained from the simulation (blue). The distributions (together with the subsequent correlations) are always calculated with SecStAnT. In this case, the coherence with the distributions is better than in the 3_{10} case. This is also probably due to the fact that the available statistics for the α-helices is of higher quality than the one for the 3_{10}-helices and that α-helices are on average longer and more regular.

It is important to note that these simulations confirm the adequacy of the non bonded term, taken from [85] and not re-optimized here. This is apparent in the distribution of the P(r) with r distance between every beads, reported in figure 39. Here, the peaks correspondent to repeated pattern of helices are well reproduced.

Finally, in figure 40 the correlation maps are shown. Also in this case, the allowed area is roughly correctly delimited, but the slope of the correlation is not reproduced. The allowed area appears circular instead of elongated along on oblique line.

In conclusion, even for α-helices the FF1 is not accurate and shows the lack of terms that help the reproduction of the accurate correlations between the conformational DOFs.

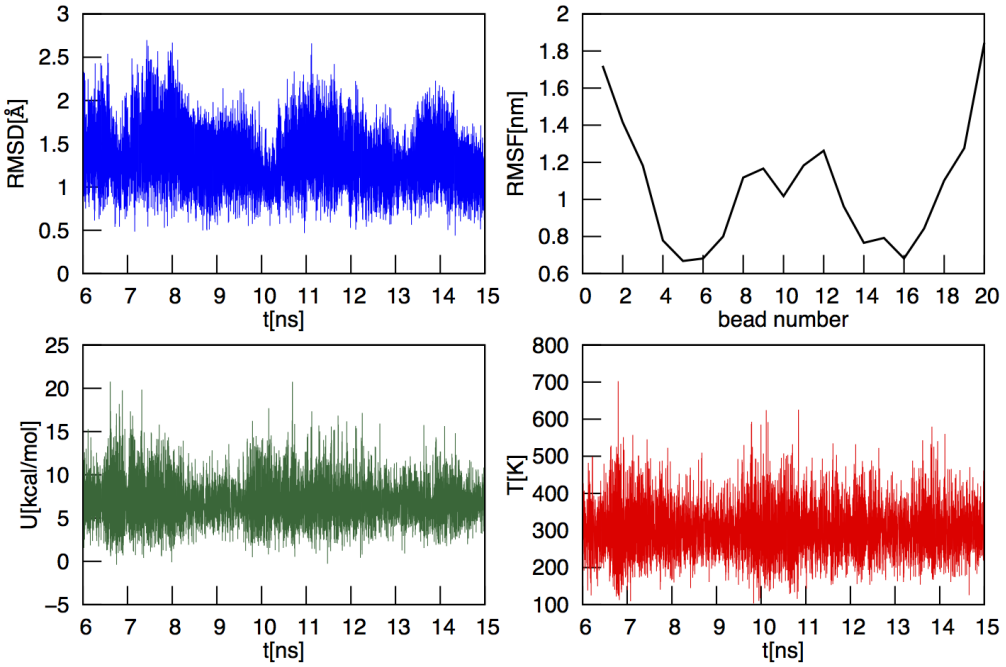


Figure 37: Simulation results for the initial field of the α-helix. The system used is a peptide of 20 amino acids. Top left: RMSD of the production run. Top right: RMSF of all the peptide beads. Bottom left: potential energy U of the system. Bottom right: temperature (K) of the system.

4.6.3 π-Helix

The starting structure used in simulations is a 12 long helix generated with Avogadro [29], since no "regular" π-helices were found in the PDB. The parameters have been directly obtained from the Xray dataset of π-helices recognized with DSSP algorithm. A starting field is proposed also for this type of helix. The parameterization is reported in table 13. Since the experimental distributions have low statistics, the presented data are only an indicative result obtained from the direct Boltzmann inversion of the distributions for internal DOFs. The simulation protocol used is reported in table 12. The simulation outputs are reported in appendix D.

In figure 41 there is the comparison of experimental and simulated distributions (Panel A) and correlations (Panel B). Given the short length of the solved π-helices the distributions reliable are the ones involving up to four beads. However, also in this case (θ, ϕ, r_{1-4}) the FF1 leads to distributions with different peak with respect to experimental ones. Furthermore, although the experimental data are not reliable, the correlation plots (Panel B), obtained with the simulation, show a too vertical allowed area.

4.7 OPTIMIZED FORCE FIELDS

Experimental analysis and simulations with the preliminary FFs obtained from direct BI lead to two important considerations:

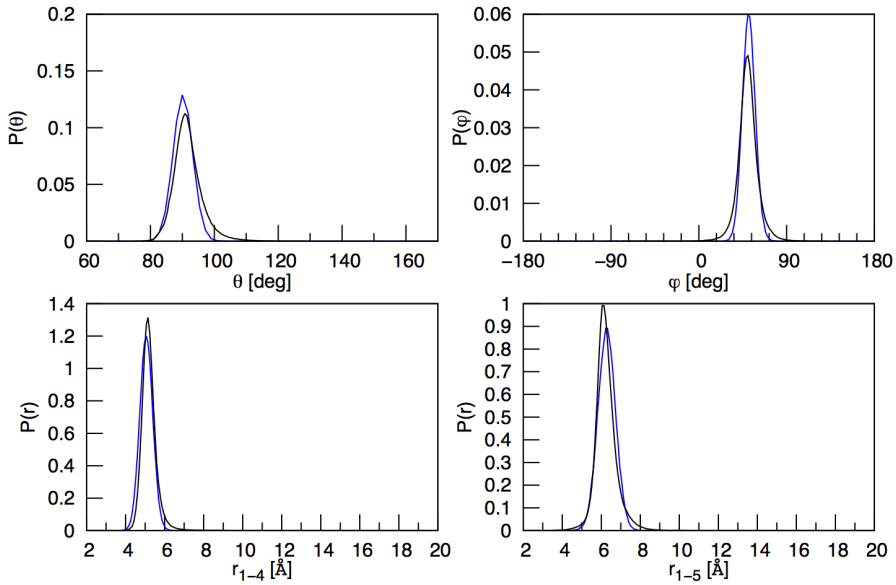


Figure 38: Comparison of experimental and simulated distributions for the α-helix. Black: experimental distributions obtained from the dataset of α-helix solved with NMR and recognized directly from PDB. Blue: distributions obtained with simulation of the 20 beads long α-helix with FF1.

- three different types of helices occupy quite the same area of the (θ,ϕ) plot;
- the conformational force field terms alone cannot accurately reproduce structures, as the correlation plot is in particular imprecise.

The above two points suggest a force field with U^{θ} and U^{ϕ} equal for all the different helices, with the only scope of maintain the structure in the helical area of the correlation map. The stabilization of the specific helix is due to the formation of the hydrogen bonding pattern. In the previous Chapter, the analysis of the correlation maps showed that there are at least two recurrent distances involved in the H-bonds for each helix. These two distances are usually the ones on the sides of the all atom hydrogen bond, as shown in figure 42. Then, it is straightforward to include these distances as explicit terms into U^{hb}.
In this new version of the FF, the θ_0 and ϕ_0 of u^{θ} and u^{ϕ} were taken at the average value of the three helices to account for the structures of all of them. The parameters are in table 15. The constant forces were fit to the three Boltzmann inverted distributions of the three helices with a single function. A range of possible parameters were derived, to be subsequently optimized in the presence of the U^{hb} term. The distributions for the π-helix were marginally considered due to their low statistics and reliability.
The parameterization of u^{hb} was obtained from the Boltzmann inversion of the correspondent distributions and then optimized. The specific $u_{i,i+n}$ terms included depend on the kind of helix, as shown in table 14: going from 3_{10} to π

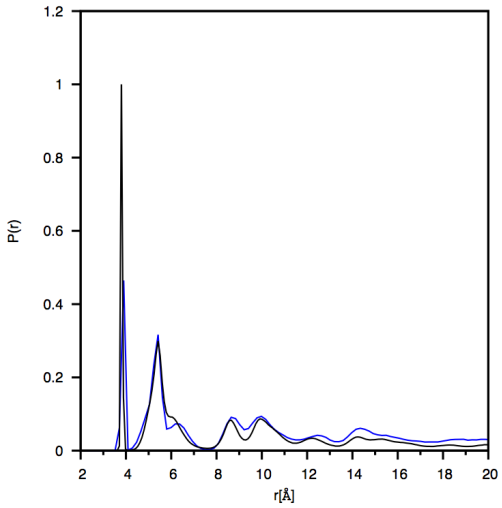


Figure 39: Comparison of experimental and simulated P(r) distribution. Black: experimental distribution obtained from the dataset of α-helix solved with NMR and recognized directly from PDB. Blue: distribution obtained with simulation of the 20 beads long α-helix with FF1.

Table 14: Distances reproducing the hydrogen bond for each type of helix. The distances mimicking the hydrogen bond are marked in black, while the distances added to stabilize the simulated structure are in green.

Helix type	r_{1-3}	r_{1-4}	r_{1-5}	r_{1-6}
$3_{10} - helix$	√	√		
α-helix	√	√	√	
π-helix		√	√	√

helix the average n increases, due to the fact that the H-bonds move along the chain. The final parameterization is thus the result of the overall optimization of the new parameters together with the other force field terms, where the sum of the energy of the relative H-bond terms is limited to $\sim 5-7$ $kcal/mol$. A summary of the distances constrained for each helix is reported in table 14 and the parameters for u^{hb} in table 15

Simulations showed that for the 3_{10}-helix, two constrained distances are enough, while to maintain the helical structure for the α-helix, it was necessary to introduce another distance dependent term, as reported in table 14.

Finally, for the π-helix the experimental data are very few and elusive, thus the performed simulations can be considered a prediction in view of comparison with forthcoming experimental data, whose reliability was tested on the other two kinds of helices. The parameterization for the optimized force field (FF2) is in table 15. The results obtained are reported in the following sections.

4.7.1 3_{10}-Helix

In the case of the 3_{10}-helix, the H-bonding pattern involves the r_{1-3} and r_{1-4} distances, i.e. bonds between C_α separated by one and two beads, respectively. In the experimental (θ,φ) plot for these helices the slope of the allowed area is mid way between the two correlation lines at constant r_{1-3} and at constant

Table 15: Parameters for the optimized force fields. Force constant are given in $\mathrm{kcal}/(\mathrm{mol}\text{Å}^2)$, values of equilibrium angles and dihedral in deg (°). Equilibrium distances are given in Å, while α in Å^{-1} and ϵ in $\mathrm{kcal}/\mathrm{mol}$

FF2	3_{10}-helix	α-helix	π-helix
u^{θ}	$k_\theta = 10$ $\theta_0 = 92$	$k_\theta = 10$ $\theta_0 = 92$	$k_\theta = 10$ $\theta_0 = 92$
u^{ϕ}	$A_\phi = 1$ $\phi_0 = 50$	$A_\phi = 1$ $\phi_0 = 50$	$A_\phi = 1$ $\phi_0 = 50$
$u^{r_{1-3}}$	$\epsilon = 6.5$ $\alpha = 1.4$ $r_0 = 5.33$	$\epsilon = 6.5$ $\alpha = 1.3$ $r_0 = 5.42$	
$u^{r_{1-4}}$	$\epsilon = 3.5$ $\alpha = 0.75$ $r_0 = 5.61$	$\epsilon = 3.5$ $\alpha = 0.8$ $r_0 = 5.15$	$\epsilon = 3.5$ $\alpha = 0.8$ $r_0 = 5.96$
$u^{r_{1-5}}$		$\epsilon = 2.6$ $\alpha = 0.66$ $r_0 = 6.05$	$\epsilon = 2.6$ $\alpha = 0.66$ $r_0 = 4.88$
$u^{r_{1-6}}$			$\epsilon = 2$ $\alpha = 0.6$ $r_0 = 6.32$

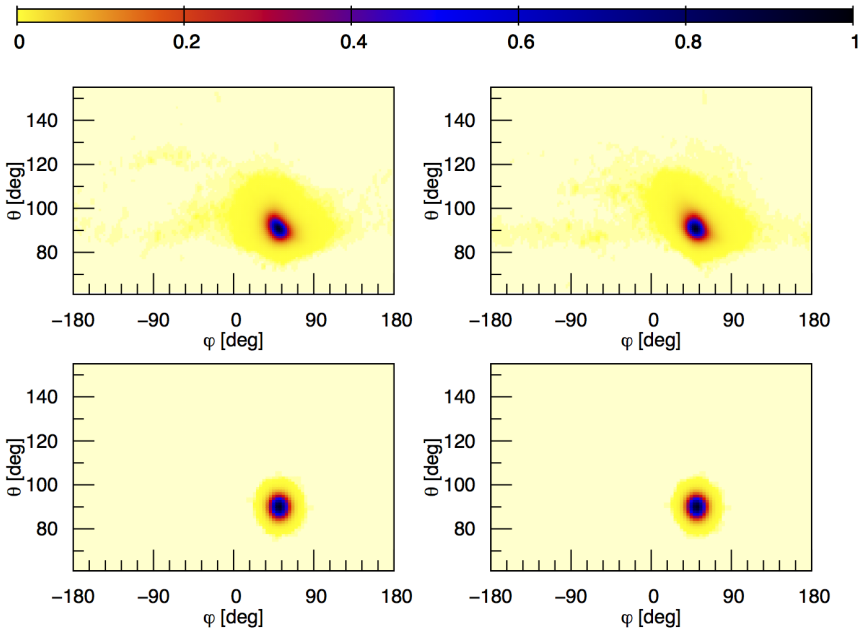


Figure 40: Comparison of experimental and simulated (θ,ϕ)-plots for α-helices. Top line: (θ−,ϕ) map on the left and (θ+,ϕ) map on the right for the experimental NMR PDB dataset. Bottom line: correlation plots corrisponding to the top line for the simulation results. The simulation was performed with the 20 beads long α-helix and with FF1. Color bar on the top.

r_{1-4} (see figure 19 in previous Chapter). Thus the FF terms corresponding to these two terms where included. The optimized parameters are showed in table 15. The simulations were run using the protocol of table 12. The outputs of the simulation are reported in appendix D.

In figure 43 (Panel A) the distributions obtained with the optimized field FF2 are reported. It must be observed that the datasets for the 3_{10} are composed also of a part of α-helices wrong solved as 3_{10}-helices, thus in this case the maximum values for the peak more than the width of the distributions was targeted. In the distribution dependent on the distance r_{1-5}, the effect of the presence of the α-helices can be seen in the left side of the experimental curve. The distribution obtained from the simulation also reproduces a little secondary peak, indicating that even in the model, metastable α-like configuration are possible and from time to time explored in the simulations.

In figure 43 (Panel B) the comparison of the experimental (upper) and theoretical (lower) (θ,ϕ)-plot is reported. As it can be seen, the slope and the extension of the correlation plot is well represented.

The same filed was tested also with the helix of 17 amino acids with the same simulation protocol. The results were similar and are reported in appendix D.

It must be reminded that the good reproduction of the correlation plots is specifically due to the inclusion of the correct topology of H-bonding and it is an element of novelty of this with respect to other similar models.

4.7.2 α-Helix

For the α-helix the hydrogen bond is between the NH group of one amino acid and the CO group of the one four beads later along the chain (i,i+4). This implies that it involves distances r_{1-4} and r_{1-5} in the C_α representation (see figure 42). Then, two terms depending on the two distances r_{1-4} and r_{1-5} was added to the force field. After some preliminary simulations, it has been clear that the stabilization of the α-helix requires the inclusion of another distance dependent term, namely u^{1-3}. The final force field with the correspondent parameterization is reported in table 15. The simulation was performed using the protocol in table 12. The simulation outputs are reported in appendix D.
In figure 44 (Panel A) the distributions obtained for the α-helix are shown. In this case, the distributions are well reproduced both in the maximum peak and in the width. Looking at figure 44 (Panel B), also the conformational correlation plots seem well reproduced, both in the slope and in the extension of the allowed area. Within this model, the α-helix is the one reproduced with best accuracy. Also in this case, it is noticeable the good reproduction of the elongation and the slope of the correlation plot.

4.8 APPLICATION TO THE π-HELIX

Since the π-helix data are not statistically relevant, in this case the parameters set, optimized in previous sections, is used "as is" without changes except for the topology of the H-bonds. This helix has the hydrogen bond between an amino acid and the one five later in the chain (i,i+5). The C_α-C_α distances involved in this bond are then the r_{1-5} and the r_{1-6} (see figure 42). The parameterizations of the same distance in the α and 3_{10} helices needed for them are quite similar. The same parameters are thus used also for this field. As in the case of α-helices, it appear that an additional distance needed to be included, namely r_{1-4} here. The simulation protocol is the same used for the other helices, summarized in table 12. The parameters are in table 15.
In figure 45 (Panel A) the resulting distributions are compared to the experimental Xray DSSP one. As previously said, the experimental data contained helices of at most 7 amino acids. Only the bond angle and dihedral angle distributions are then considered quite reliable. The FF reproduces quite well these distributions, in spite of the high noise level in the experimental data. Analogously, the slope in the correlation plot (see figure 45, Panel B) is quite well reproduced, although the experimental plot is more noisy.

4.9 SUMMARY

In this Chapter a simplified C_α-based models that reproduce accurately the helical secondary structures (with their probability distributions) are proposed. The corresponding force fields are composed of statistically derived potentials. The problem of the correlation among terms is here addressed by including new terms mimicking the physical forces, like the backbone H-bond. It is shown that these naturally account for the correlations among conformational

terms, correctly reproducing the (θ,ϕ)-plots, which are the equivalent of the Ramachandran plots in the atomistic representation.
The improvement of the model obtained including those terms is apparent from the comparison of figure 43 and figure 44 : both for 3_{10} and α-helices the calculated single variables distributions, as well as the correlations well reproduce the experimental ones, when the appropriate pseudo-hydrogen bonds are included. It is to be remarked that the correct reproduction of these correlations generally receive little attention in the parameterization of CG models, mainly because, at variance with the Ramachandran plot, the experimental (θ,ϕ) map is not often considered a validation criterion.
Concerning the π-helix, as said, the comparison with experiment is less good, but this can be attributed to the poor statistical relevance and accuracy of its dataset. In this sense, the present model could be considered a prediction to be compared with new data as soon as they are available.
The results shown in this Chapter are obtained by using additional terms depending on a single scalar variable, allowing for the use of simple functional forms in place of multi-variable ones or complex anisotropic potentials. This is a big advantage not only because the analytical forms are simpler and more intuitive, but also because single variable functions are computationally more simple and efficient and they can be directly included into any MD code, such as the DL_POLY package. In spite of the simplicity of the FF, high accuracy is obtained and even many-body effects, such as the correlations, can be reproduced. This is the effect of the accurate physically driven choice of the terms representing the hydrogen bonding.

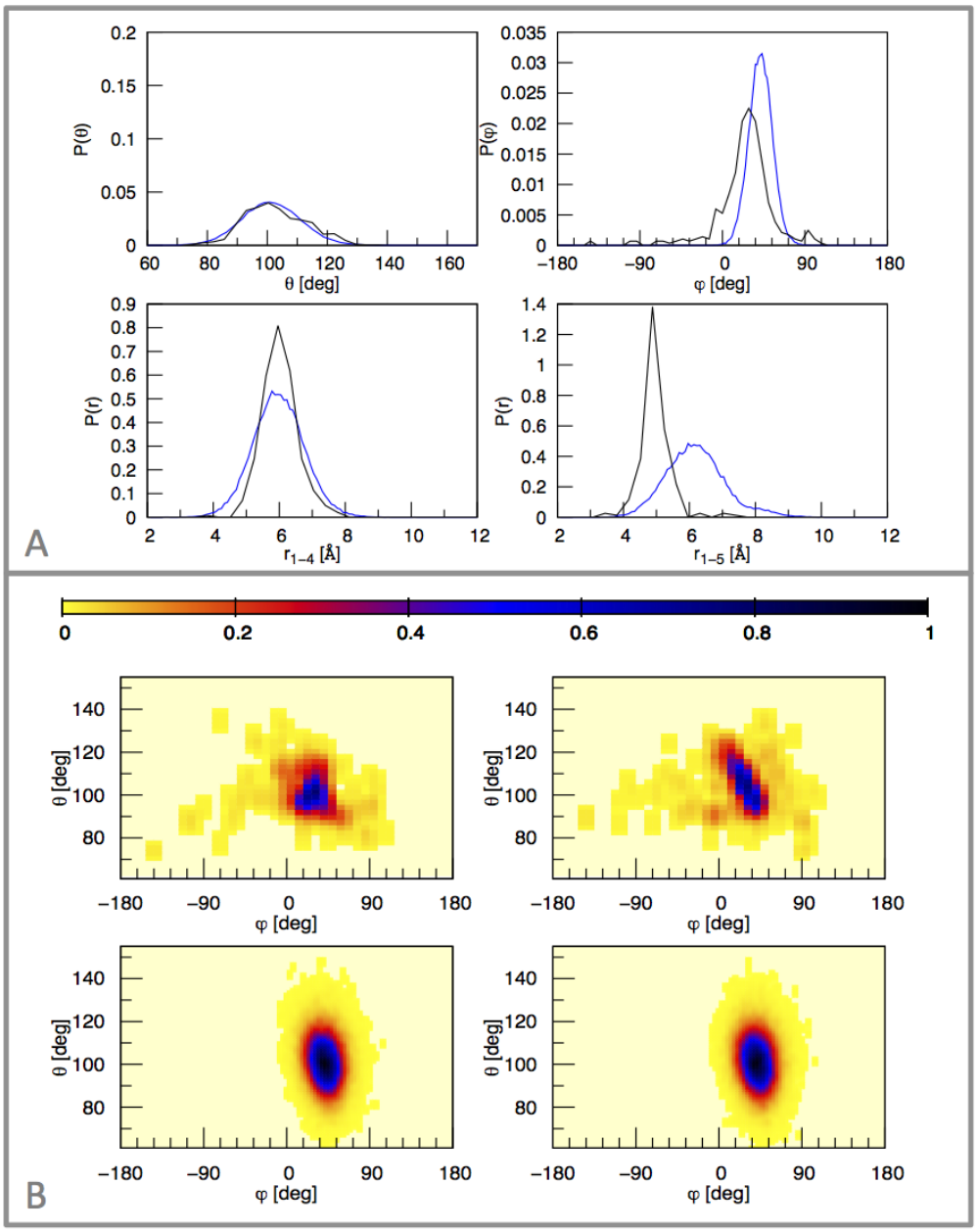


Figure 41: Comparison of experimental and simulated distributions distributions and correlations for the π-helix. (Panel A) Black: experimental distributions obtained from the dataset of pi-helices solved with X-ray and recognized with DSSP. Blue: distributions obtained from the simulation. (Panel B) Top line: (θ−,ϕ) map on the left and (θ+,ϕ) map on the right for the experimental dataset. Bottom line: correlation plots, corresponding to the top line, for the simulation results. Color bar on the top. The simulation was performed with the 12 beads long π-helix and with FF1.

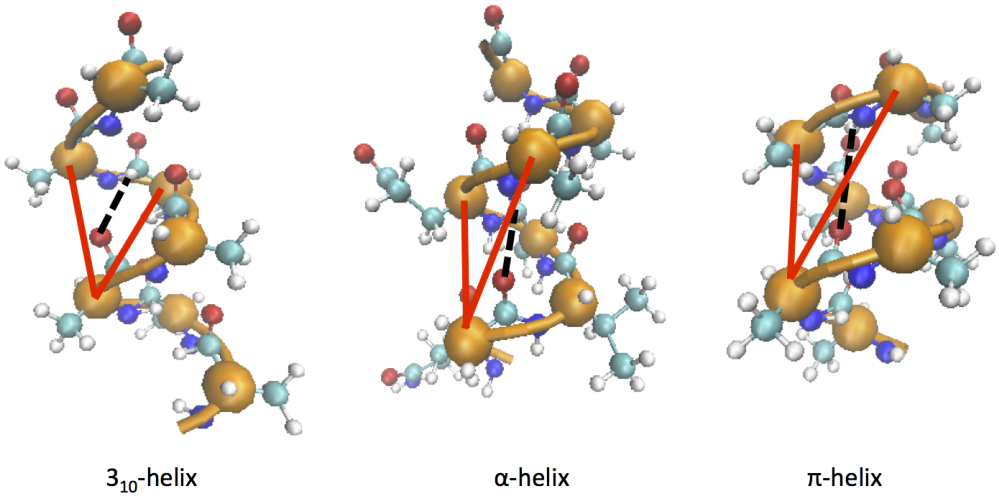


Figure 42: Helical pattern of H-bonds. For each type of helix, the entire backbone is showed together with the CG representation in orange. Color code for backbone: red O, blue N, cyan C, white H. The hydrogen bonds in the all atom representation are outlined with dashed black line, while the distances representing it in the CG representation are outlined in red.

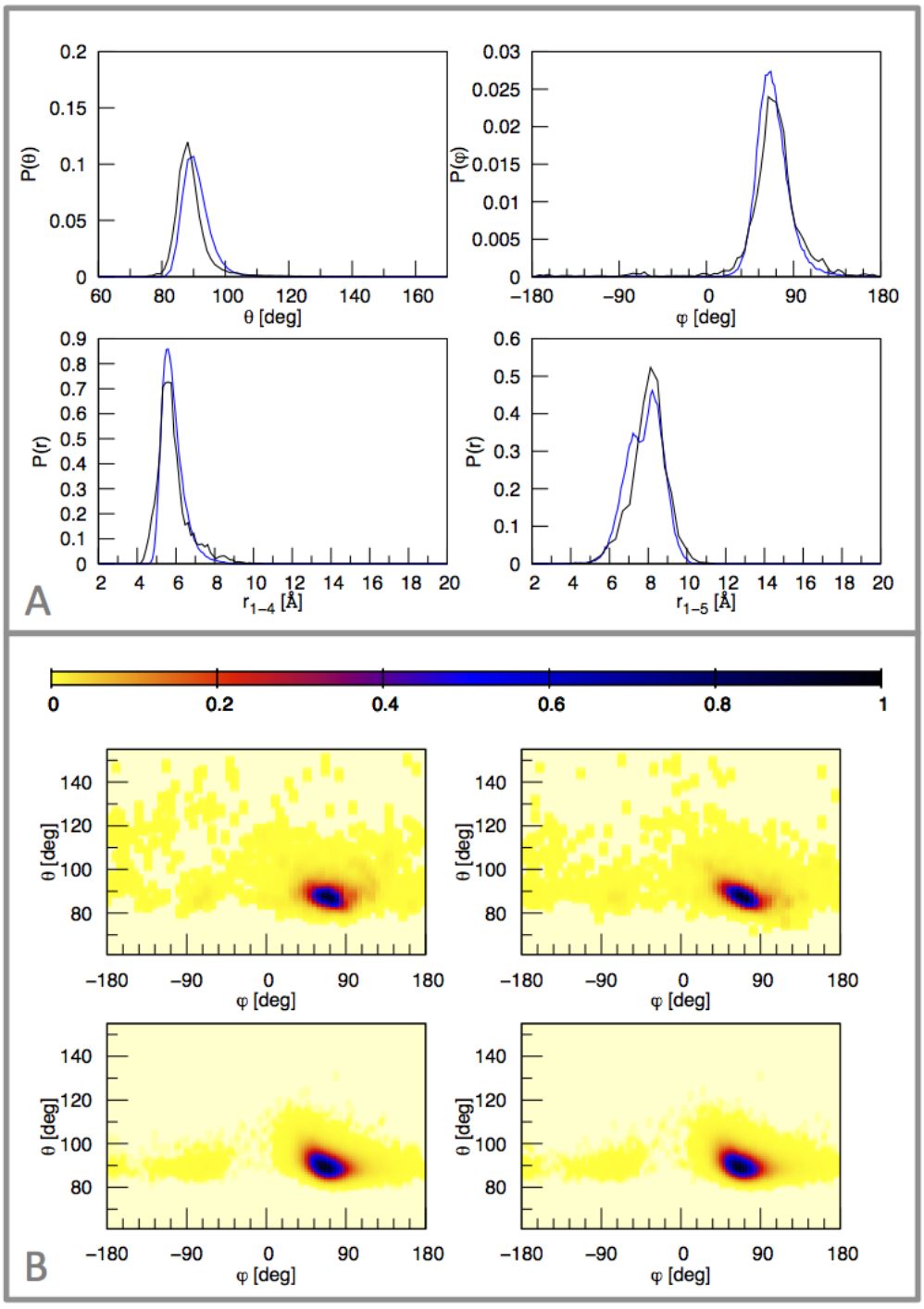


Figure 43: Comparison of experimental and simulated distributions distributions and correlations for the 3_{10}-helix. (Panel A) Black: experimental distributions obtained from the dataset of 3_{10}-helices solved with NMR and recognized with DSSP. Blue: distributions obtained from the simulation. (Panel B) Top line: ($\theta-$,ϕ) map on the left and ($\theta+$,ϕ) map on the right for the experimental dataset. Bottom line: correlation plots, corresponding to the top line, for the simulation results. Color bar on the top. The simulation was performed with the 11 beads long 3_{10}-helix and with FF2.

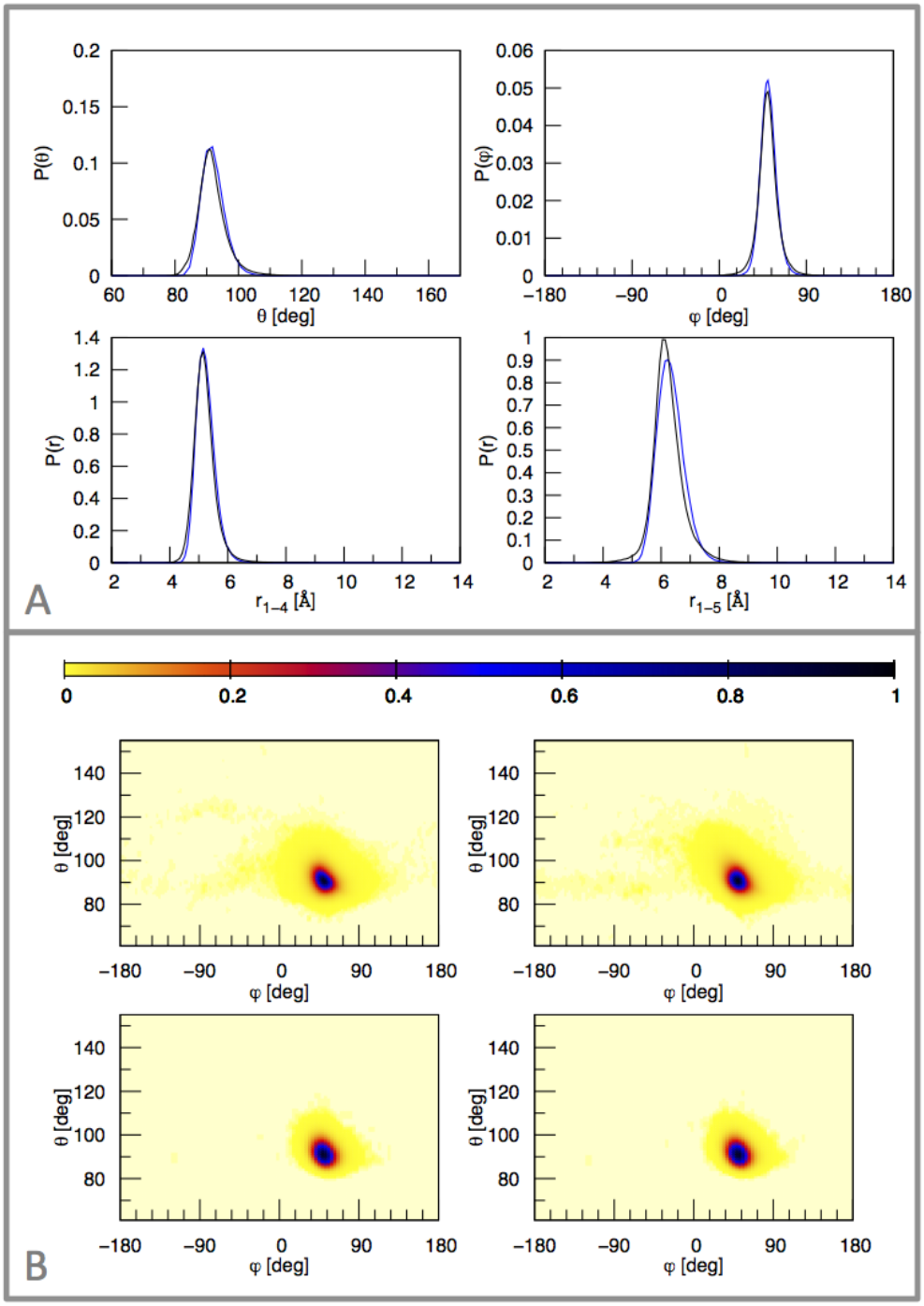


Figure 44: Comparison of experimental and simulated distributions distributions and correlations for the α-helix. (Panel A) Black: experimental distributions obtained from the dataset of α-helices solved with NMR and recognized directly from the PDB. Blue: distributions obtained from the simulation. (Panel B) Top line: (θ−,ϕ) map on the left and (θ+,ϕ) map on the right for the experimental dataset. Bottom line: correlation plots, corresponding to the top line, for the simulation results. Color bar on the top. The simulation was performed with the 20 beads long 3_{10}-helix and with FF2.

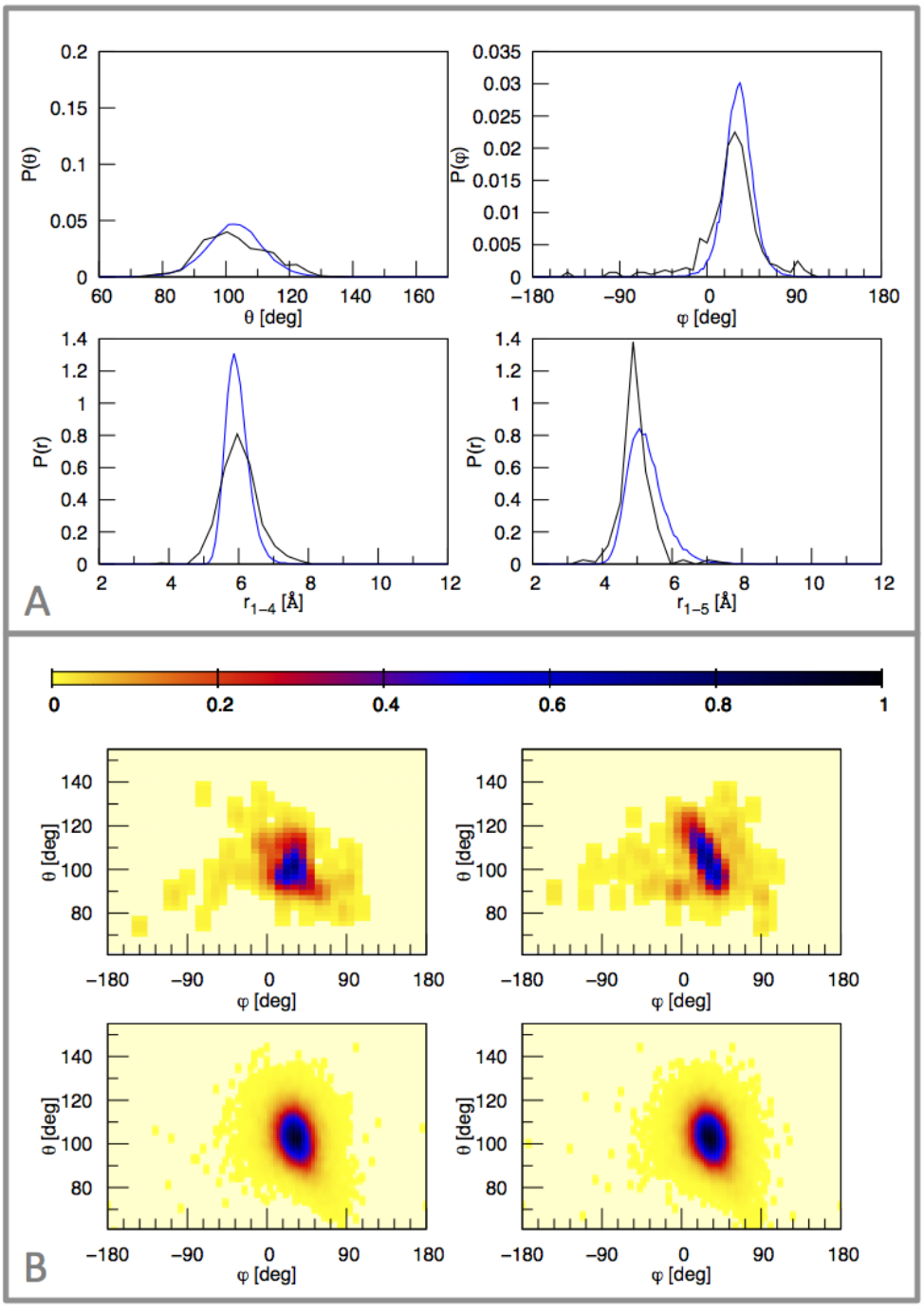


Figure 45: Comparison of experimental and simulated distributions distributions and correlations for the π-helix. (Panel A) Black: experimental distributions obtained from the dataset of π-helices solved with NMR and recognized directly from the PDB. Blue: distributions obtained from the simulation. (Panel B) Top line: (θ−,ϕ) map on the left and (θ+,ϕ) map on the right for the experimental dataset. Bottom line: correlation plots, corresponding to the top line, for the simulation results. Color bar on the top. The simulation was performed with the 12 beads long pi-helix and with FF2.

5 CONCLUSIONS AND PERSPECTIVES

As previously stated, this Thesis work is included in a more general framework aimed at producing a general model capable of describing all the secondary structures, and to combine them in tertiary structures. This is clearly a long time scale project, a part of which has been completed with this work, precisely the one regarding the helices. A model has been produced capable of describing with a high degree of accuracy the three main different types of helices. Specifically, the structure, dynamics and distributions and correlations of internal variables of α-helices and 3_{10}-helices, from simulations compare well with available experimental data, indicating the high accuracy of this parameterization. For the π-type helix the experimental data are very few and elusive, thus our simulations can be considered a prediction whose reliability is tested on the other two kinds of helices, in view of comparison with forthcoming experimental data.

The model here presented includes some elements of innovation with respect to similar previous one. The above mentioned results are achieved with a minimal number of terms in the Hamiltonian, whose meaning can be directly understood in terms of physical interactions. Starting from the choice of the force field terms and of their functional forms, its building is, in fact, strongly "physics based": at variance with previous models, a relatively small weight is given to the conformational terms of the force fields (related to the internal variables θ,ϕ) aimed at reproducing only the generic tendency of the backbone to form coils in absence of specific hydrogen bonding interactions, while the formation of ordered secondary structures is mainly imputed to force field terms representing the hydrogen bonds. This follows the real biochemistry of the proteins and their hierarchical structural organization.

Given the strategy used to build the model, it is already naturally equipped to describe the general conformational flexibility of the backbone even in the case of weakly structured or de-structured proteins, beyond the helices. The extension to sheet structures must proceed through the inclusion of terms describing the hydrogen bonding network of these structures. These could be finally combined with the helical hydrogen bonding terms. The relative weight of the two terms will determine the preferred conformation (helical or sheet-like). These weight could be added in a sequence dependent way, exploiting the available (and accurate) algorithm of prediction of secondary structure from the sequence. This would give to the model the capability of predicting the folding from the sequence, at least at the secondary structure level. In this stage, also experimental information about the relative stability of the different secondary structures could be included, in order to predict also the correct thermodynamics of the system. Finally, in order to accurately combine the secondary structures in the tertiary fold, additional sequence dependent information on the long range interactions (hydrophobicity, electrostatics) will be added to the model.

Though being part of a larger project, these results are complete, since the current model is capable of simulating all kind of helices. In addition, several side-results have been produced. Some of them are related to the fundamentals of the minimalist models. It was shown that (backbone) back-mapping is possible within this approach, giving the possibility of going back to the atomistic representation at any time. This means that the model can be safely and coherently included in a multi-scale approach.

A general strategy to build and parameterize the model is also suggested, which can be applied to different cases (e.g. nucleic acids). This focuses on the inclusion of experimental data from different sources (structural, thermodynamics etc) in a physics based fashion. This have the advantage of identifying more easily the terms and parameters responsible for a given behavior of the model, of reflecting the natural organization of these complex molecules and to be, in a word, more adherent to reality. It is a common belief among the CG modelers that combining accuracy and predictive power/transferability in these low resolution model is a formidable task. Probably it is, indeed, but this approach gives an alternative strategy to try to obtain the best from these simple models.

Another side result (submitted for publication) was the creation of the SecStAnT software, originally aimed at helping the minimalist model parameterization, but soon revealing its intrinsic potentialities. In fact, organization of the huge amount of structural data currently available and mining useful data from it is useful *per se*, not only for minimalist model parameterization. For this reason, SecStAnT is already currently provided with a number of tasks going beyond those used for this Thesis work, and will be further enriched in the future. SecStAnT is capable of treating different resolution from the atomistic to the minimalist, to analyze up to 3D correlation of a number of internal variables (not only related to the minimalist model), and is able to create Ramachandran Plots and their CG equivalent, often used to validate models (even the atomistic ones). In addition, it is provided of a user friendly interface, also designed to connect it to the molecular dynamics simulations code. In fact, an immediate future development is the automatization of the communication between SecStAnT and DL_POLY. This could be done by a new software module, mining structural data with SecStAnT, preparing the input for DL_POLY from them, passing the output of simulations (i.e. structures from simulation trajectories) to SecStAnT which would perform again the analysis and correct the input for DL_POLY and repeat the cycle (as in the Iterative Boltzmann Inversion). This would also allow automating the parameters optimization, by including recursive force field corrections from the comparison between simulation and experimental data, always preserving the possibility for the user to interact with the procedure. This work set the ground not only for a general minimalist model protein structures and dynamics, but also provides a procedure to optimize it and possibly to improve its parameterization when new data are available.

A APPENDIX

A.1 PDB FILE FORMAT

Listing 1: Example of PDB file

```
HEADER    IMMUNE SYSTEM                           25-FEB-13   2M5H
TITLE NMR STRUCTURE NOTE:SOLUTION STRUCTURE OF MONOMERIC HUMAN FAM96A
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: MIP18 FAMILY PROTEIN FAM96A;
COMPND   3 CHAIN: A;
...
...
REMARK 210 EXPERIMENTAL DETAILS
REMARK 210  EXPERIMENT TYPE                : NMR
REMARK 210  TEMPERATURE           (KELVIN) : 298
REMARK 210  PH                             : 7.0
...
...
HELIX    1   1 MET A   33  THR A   46  1                   14
HELIX    2   2 SER A   63  SER A   65  5                    3
SHEET    1   A 3 VAL A  67  ASN A  73  0
SHEET    2   A 3 GLU A  76  PHE A  83 -1  O  ILE A  80   N  GLU A  68
SHEET    3   A 3 LYS A 111  ILE A 118  1  O  LYS A 113   N  VAL A  79
...
...
MODEL        1
ATOM    1  N   MET A  27      2.022   1.021   8.639  1.00  0.00   N
ATOM    2  CA  MET A  27      1.496   1.839   7.505  1.00  0.00   C
ATOM    3  C   MET A  27      0.867   0.950   6.421  1.00  0.00   C
...
...
TER    2277      HIS A 165
ENDMDL
MODEL        2
ATOM    1  N   MET A  27     -2.816  -0.216   5.224  1.00  0.00   N
...
...
TER    2277      HIS A 165
ENDMDL
MASTER      203    0    0    6    3    0    0    6 1135    1    0   11
END
```

Listing 1 shows some important lines of a typical pdb file for a NMR structure. It begins with an header, containing some general informations about the structure, like title, date of deposition, type and number of molecules and experimental method used to solve it. The REMARK lines add general information and the different number of remarks distinguishes the type of information given. For example, REMARK 210 gives some details of the NMR technique

used.
In lines starting with HELIX or SHEETS the authors report the different secondary structures. HELIX records show the number of the helix, the name and the number of the starting and the terminal residue of the helix and the corresponding chain. Finally, an integer number on the tenth column identifies the specific type of helix. In this way, it is possible to distinguish between α (1),3_{10} (5) and π(3) helices.
Lines starting with SHEET identify strands in different β-sheets, for these reasons they are grouped with different letters in the third column, that divides the sheet. Also in these entries there are number and name of the amino acids starting and ending the sheet as well as the chain name. In the eleventh column, 1 defines a strand parallel to the previous, while -1 defines a antiparallel strand. 0 is assigned to the first strand of the sheet. The last seventh columns report the better hydrogen bond solved for the current strand bonded with the previous. There are listed in order: the atom H-bonded in the current strand, name, chain and number of the amino acids to which it belongs and the same information for the atom H-bonded in the previous strand.
For X-ray crystallography there is just one model for every structure, while for NMR there are more different models. Each model starts with the MODEL line and a growing integer number and finishes with the ENDMDL entry. Between these two lines there are all information about atoms solved. The ATOM entry in fact reports: a growing number identifying the atom; the atom name; name, chain and number of the amino acid to which it belongs; three coordinates to define its specific position and other information. The end of each chain is identified by the TER entry.
The MASTER line is a sort of summary of the structure and the END entry closes the file. A more detailed description of the file format is available at [89].

A.2 DSSP

The DSSP method finds the simplest possible pattern of hydrogen bonds able to discriminate between different types of secondary structures. It defines a specific structure through the presence/absence of H-bonding between residues. The H-bond is defined with only one parameter: a cutoff in the bond energy. To define a H bond, the DSSP algorithm uses an electrostatic model and the process of definition can be simplified as follows:

1. Considering two H-bonding groups, partial charges are placed on the C, O ($+q_1, -q_1$) and N, H ($-q_2, +q_2$) atoms, with $q_1 = 0,42e$ and $q_2 = 0,20e$, e being the unit electron charge;

2. the electrostatic interaction energy is calculated using 41:

$$E = q_1 q_2 \cdot [\frac{1}{r(ON)} + \frac{1}{r(CH)} - \frac{1}{r(OH)} - \frac{1}{r(CN)}] \cdot f \tag{41}$$

   where r(AB) is the distance between the atoms A and B (Å), f (=332) is the dimensional factor and E is measured in $kcal/mol$.

3. An H bond is assigned between the group CO of residue i and the group NH of residue j, if the calculated E is lower than a cutoff. It uses $-0,5 kcal/mol$

as cutoff because it's known that a good H bond has $E = -3kcal/mol$. A large enough cutoff is chosen to admit the possibility of bifurcated H bonds and errors in coordinates.

4. Finally a relationship is defined between this one-parameter definition and a more complicated two-parameter description: one distance (d = N$\cdots$O) and one angle (the angle (θ) between the direction N-H and N$\cdots$O). So an ideal H bond with $E = -3kcal/mol$ has $d = 2,9$Å and $\theta = 0°$.

Each feature of the same structure defined by patterns of H bonds are included in a hierarchy. So, once the position in the sequence of the H bonded residues is defined, two elementary H bonds patterns are identified:

n-turn It is a single H bond between the C=O group of residue i and the group NH of residue i+n with n = 3,4,5. If there is, an n-Turn is assigned to i.

bridge It is composed by two non overlapping stretches of three residues each: (i-1,i,i+1) and (j-1,j,j+1). This pattern takes into account the H bonds between residues not consecutive in the chain. There are two types of bridges (parallel and antiparallel) and a bridge is assigned between residues i and j if there are two H bonds characteristic of β-structures. So a parallel bridge is identified if there are H bonds between [i-1,j and j,i+1] or between [j-1,i and i,j+1] and an antiparallel bridge if they are between [i,j and j,i] or between [i-1,j+1 and j-1,i+1].

Once mapped the amino acid sequence in the chain of patterns, some cooperative H bond patterns are recognized:

helices A minimal helix is defined by two consecutive n-turn. For example, to assign an α-helix there has to be a 4-turn between residues 1 and 5 and the same turn between residues 2 and 6. Conventionally, two consecutive 3-turn define a minimal 3_{10} helix, two consecutive 4-turn define a α helix and two consecutive 5-turn define a π helix. Overlaps of minimal helices are then considered longer helices. The DSSP nomenclature is reported in tabs, where every turn is defined with its name in its line and in another line (SUMMARY) the residues belonging to different helices are marked with appropriate letters: H(α), G(3_{10}) and I (π).

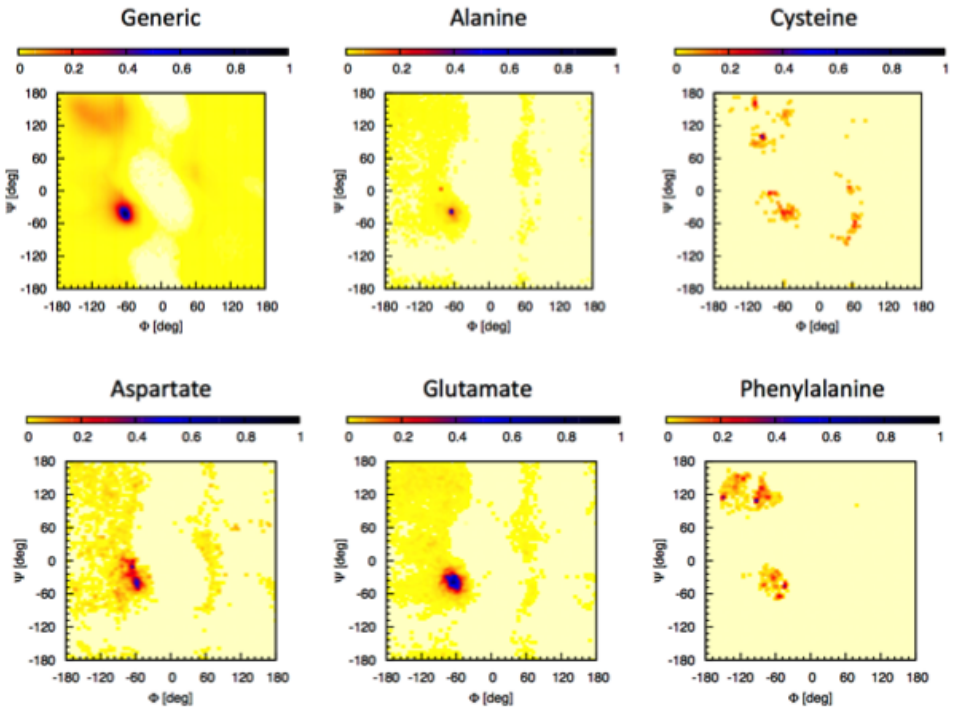
β-ladders and β-sheets The DSSP defines the term *Ladder* as set of one or more consecutive bridges of identical type. A *Sheet* then is a set of one or more ladders connected by shared residues. So in the DSSP's tabs each residue is identified with the name of the sheet to which it belongs (line SHEET) and also with the name of the ladders (at most two) to which it belongs (line BRIDGE). Finally, in the line SUMMARY residues in single bridges (ladder one residue long) are differently marked then residues in ladders (named extended and marked with E). In that way continuous stretches of E are β-strands.

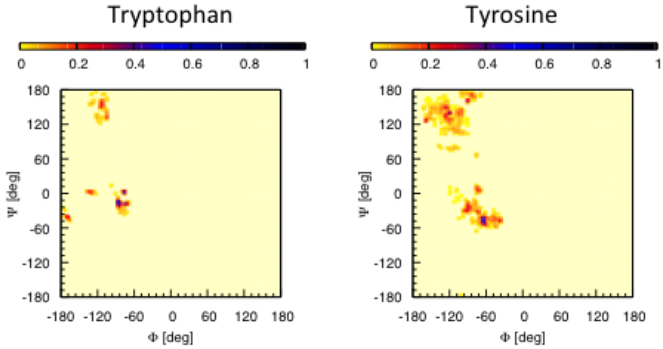
Furthermore, the DSSP algorithm is also able to account for some irregularities of secondary structures like kink in helices and β-bulges in β-structures.

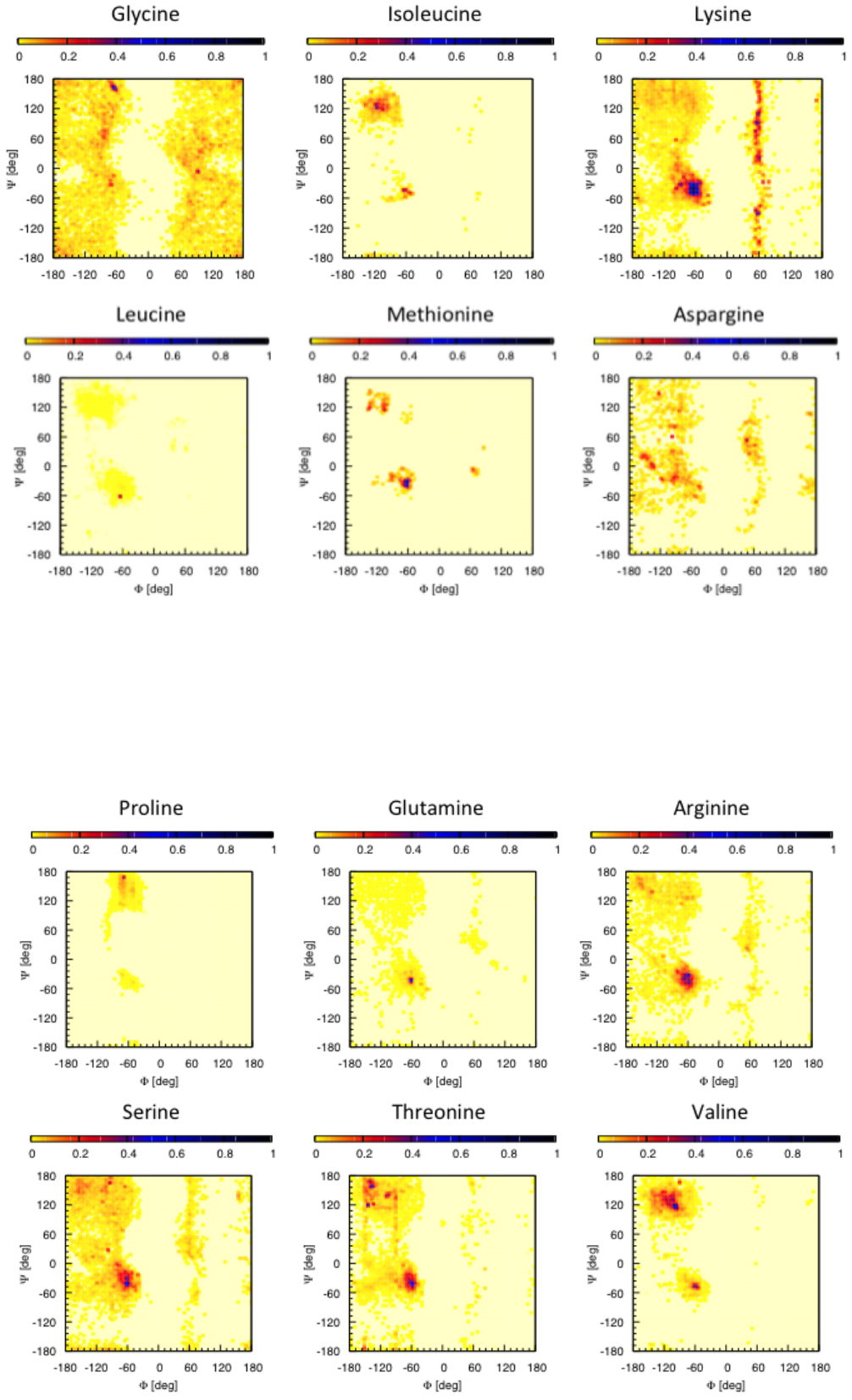
It's important to underline that every structural feature is defined independently from the others and the overlaps are solved by establishing a summary (SUMMARY) of secondary structure. It assign a single state to every residue by giving priority following α,ladder,sheet,3_{10},π, turn. Pieces of 3- or 5- helix reduced to less than minimal size due to overlaps are labeled turns.

A.3 RAMACHANDRAN PLOTS

In this section are reported the Ramachandran plots obtained with SecStAnT for the natural amino acids. The RP for Histidin was not created, because there were not enough data.






Glycine
Isoleucine
Lysine
Leucine
Methionine
Aspargine
Proline
Glutamine
Arginine
Serine
Threonine
Valine
Ψ [deg]
Φ [deg]
0 0.2 0.4 0.6 0.8 1
180 120 60 0 -60 -120 -180
-180 -120 -60 0 60 120 180

B APPENDIX

B.1 CONSTRAINED DYNAMICS

In classical molecular dynamics simulations, it is common practice to not represent the intra-molecular bonds with terms of the force field, because these bonds have very high vibration frequencies, which would be treated in a quantum mechanical way. An alternative, often used, is the constraining of the bond length to fixed distance.
The most commonly used method for applying constraints is the SHAKE procedure [44]. It is a step-by-step algorithm. First, the beads in the system are moved with the integration algorithm chosen (Verlet is usually used with SHAKE), assuming an absence of constraint forces. Second, the deviation of each obtained bondlength from the constrained ones are calculated. This deviation is then used for evaluate a corresponding constraint force that (restrospectively) correct the bond lengths. Third, after the correction has been applied to all bonds, every bond lengths is checked. If the largest deviation found exceeds the desired tolerance, the correction calculation is repeated. Finally, the second and third steps are repeated until convergence for all bond lengths is reached.
This algorithm may have convergence problems if applied to large planar groups and its implementation could hinder the efficiency of computing.
Another algorithm to implement constraint dynamics is RATTLE [32]. It has two parts: the first constrains the bondlength and the second adds an additional constraint to the velocities of the atoms in the constrained bond.

B.2 NOSE–HOOVER THERMOSTAT

In the following there is the detailed description of thermostat most used in this Thesis work, the Nose-Hoover thermostat [88], [88]. The system is coupled to a heat bath. This is considered an integral part of the system by addition of an artificial variable s associated with a mass parameter Q. The magnitude of Q defines the strength of the coupling to the heat bath: it influences the thermal fluctuations. the variable s may be considered a time-scaling parameter.
For the extended system, Nose introduced the Hamiltonian:

$$H_{(Nose)} = H_0\left(q, \frac{p}{s}\right) + gkT\ln s + \frac{p_s^2}{2Q} \tag{42}$$

Here p_s is the momentum associated to s, g is a parameter related to the degrees of freedom in the system, while H_0 is the Hamiltonian for a classical many body system, considering also the artificial variable, i.e.:

$$H_0\left(q, \frac{p}{s}\right) = \sum_i \frac{(p_i/s)^2}{2m_i} - U(q) \tag{43}$$

Then the equations of motion are:

$$\begin{aligned}
\dot{r}_i &= \frac{\partial H_{Nose}}{\partial p} = \frac{p_i}{ms^2} \\
\dot{p}_i &= \frac{\partial H_{Nose}}{\partial r} = F_i(r) \\
\dot{s} &= \frac{\partial H_{Nose}}{\partial p_s} = \frac{p_s}{Q} \\
\dot{p}_s &= \frac{\partial H_{Nose}}{\partial s} = \sum \frac{p_i^2}{ms^3} - \frac{gkT}{s}
\end{aligned} \tag{44}$$

These equations can be rewritten in a simpler form considering s as time-scaling parameter, i.e. $dt_{old} = sdt_{new}$. Then, the equations given before can be expressed as:

$$\begin{aligned}
\dot{r}_i &= \frac{\partial H_{Nose}}{\partial p} = \frac{p_i}{ms} \\
\dot{p}_i &= \frac{\partial H_{Nose}}{\partial r} = F_i(r)s \\
\dot{s} &= \frac{\partial H_{Nose}}{\partial p_s} = \frac{p_s s}{Q} \\
\dot{p}_s &= \frac{\partial H_{Nose}}{\partial s} = \sum \frac{p_i^2}{ms^2} - gkT
\end{aligned} \tag{45}$$

It is possible to write also:

$$\ddot{r}_i = \frac{d\dot{r}_i}{dt} = \frac{\dot{p}_i}{ms} - \frac{p_i \dot{s}}{ms^2} = \frac{F_i}{m_i} - \xi \dot{r}_i \tag{46}$$

Here $\xi = p_s/Q$ can be considered as a friction coefficient. It accounts for the action of the thermostat and evolves in time $\propto (T(t) - T_0)/Q$.

All the thermostats share some problems. The most important one arise when simulating large complex having different components (so different degrees of freedom) with different dynamics. The exchange of kinetic energy between them could be too slow, leading to different temperature for different components. One example is the phenomenon of the "hot-solvent,cold-solute", when simulating large assemblies in solvent. The temperature of the solute is lower than that of the solvent, even though the overall temperature of the system is at the desired value. A possible solution is to apply temperature coupling separately to the solute and to the solvent, but the problem of unequal distribution of energy between the various components of the system may still remain. This type of problems, however, is typical of big systems or systems in which the solvent is explicit.

C APPENDIX

C.1 SECSTANT: DEFINITION OF THE OUTPUT DATA FORMAT

At the first startup, a cache folder is generated in the same directory of the executable file. In this folder, every downloaded file will be saved in order to avoid multiple downloads of a single entry. In the results folder, secondary structure's fragments are organized in a hierarchical order. For each structure, a folder named as secondary structure's short name is created. In this folder, a log describing in details the extraction process is generated.
Each fragment is named on the base of the original structure, the Model, the chain and a counter. Model defines the NMR model number; for non-NMR data it is equal to 1 and the counter is an integer number necessary for the separations of every fragment. A "Statistics" folder is created in the results folder where for each secondary and super-secondary structure a file is generated for each distribution or correlation chosen.
For each file format a brief description is given:

PDB FRAGMENTS OUTPUT FORMAT In the header section, only the line corresponding to the definition of the secondary structure (if any) is saved. After the header, ATOM lines are saves as described by the standard PDB file format.

DISTRIBUTIONS Distributions are organized in a plain text file. Different columns correspond to different normalization strategies or to the Boltzmann inversion data, when required.

TWO VARIABLES CORRELATIONS Two different file formats are available for this type of correlations: a gnuplot-compatible format and a comma separated value (CSV) format. The gnuplot-compatible format is composed by a sequence of 4-tuples. The first two elements are the coordinates in the 2D space (e.g. θ,ϕ) and the last two are the correlations events and their normalization. The 2D space is represented as a sequence of couples ordered by (ϕ,θ), where θ varies faster. After each variable cycle there is an empty row. In the CSV format, data is represented by a matrix N x M where N is the number of ϕ bins and M is the number of θ bins. Only raw data are reported.

THREE VARIABLES CORRELATIONS The file is a sequence of 5-tuples. The first three elements are the coordinates in the 3D space (e.g. r_{1-4},ϕ,θ) and the last two are the correlations events and their normalization. The 3D space is represented as a sequence of triples ordered by (r_{1-4},ϕ,θ), where θ varies faster. After each variable cycle there is an empty row (i.e. after each ϕ cycle there are two blank lines).

CUBE FORMAT SecStAnT is able to use the CUBE format to display three variables correlations. The CUBE format is originally thought to represent volumetric data of atoms sets, like electrostatic potentials and orbitals. In SecStAnT the information about electrostatic potentials are replaced by the counting of occurrence of correlation events between variables. Therefore, in order to make the format compatible, one dummy atom is inserted at the origin of axes. In this way there are no atoms in the cube space of correlations data. The file format is fully described at [12].

C.2 RCSB QUERY

In listing 2 there is reported am example of query for a general dataset of X-ray proteins. This is the result of an advances query in the RCSB server and it is the input necessary for SecStAnT to start the download process.

Listing 2: Example XML query.

```
<orgPdbCompositeQuery version="1.0"> <resultCount>82735</resultCount> <
    queryId>A5428B4</queryId>
<queryRefinement>
...
...
<containsProtein>Y</containsProtein>
<containsDna>N</containsDna>
<containsRna>N</containsRna>
...
...
<queryType>org.pdb.query.simple.HomologueReductionQuery</queryType>
<description>Homologue Removal - 30 Identity Cutoff of Chain Type: there is
     a Protein chain but not any DNA or RNA or Hybrid
and
Revised between 2001-01-01 and 2013-04-05
and
Experimental Method is X-RAY </description>
<queryId>null</queryId>
<resultCount>11672</resultCount>
<runtimeStart>2013-04-05T08:16:56Z</runtimeStart> <runtimeMilliseconds
    >1913</runtimeMilliseconds>
<identityCutoff>30</identityCutoff>
</orgPdbQuery>
</queryRefinement>
</orgPdbCompositeQuery>!
```

C.3 ALGORITHMIC DETAILS FOR THE CALCULATION OF DISTRIBUTIONS AND CORRELATIONS

ßThe angle θ_i between three consecutive beads is calculated by:

$$\theta_i = \arccos\left(\frac{\vec{r}_{(i-1)-i} \cdot \vec{r}_{i-(i+1)}}{r_{(i-1)-i} r_{i-(i+1)}}\right) \tag{47}$$

where $\vec{r}_{(i-1)-i}$ is the vector joining bead i and $i-1$, while $r_{(i-1)-i}$ is its magnitude.
Finally the dihedral angle ϕ_i between four consecutive C_α is defined as the angle between two planes: the first one is described by the vector joining the first and the second residues $\vec{r}_1$ and the second with the third $\vec{r}_2$, while the second one is described by the vector joining the second and the third residue $\vec{r}_2$ and the one joining the last two residues $\vec{r}_3$. By definition, the dihedral angle is calculated by intersecting the dihedral with a plan normal to $\vec{r}_2$; in other words the angle between $\vec{r}_1 \times \vec{r}_2 = \vec{n}_1$ and $\vec{r}_2 \times \vec{r}_3 = \vec{n}_2$. At this point the dihedral angle is obtained from:

$$\phi_i = \frac{s}{|s|} \arccos\left(\frac{\vec{n}_1 \cdot \vec{n}_2}{n_1 n_2}\right) \tag{48}$$

where s is the projection of the vector $\vec{n}_1 \times \vec{n}_2$ on $\vec{r}_2$ and $|s|$ is the absolute value of s. Here it is used to calculate the sign of ϕ_i.

C.4 DISTRIBUTIONS AND CORRELATIONS

In this section, distributions and correlations plot made with SecStAnT are reported. The data here shown complete the results commented in Chapter 3, including different datasets, distributions/correlations of different internal variables, secondary structures evaluated with different algorithms, as explained in the captions. At which secondary structure the graphs correspond is also reported in caption.
The data shown for sheets were not used in this work, since the model for β-sheets was not optimized. However, they could be useful for the future development of this work.

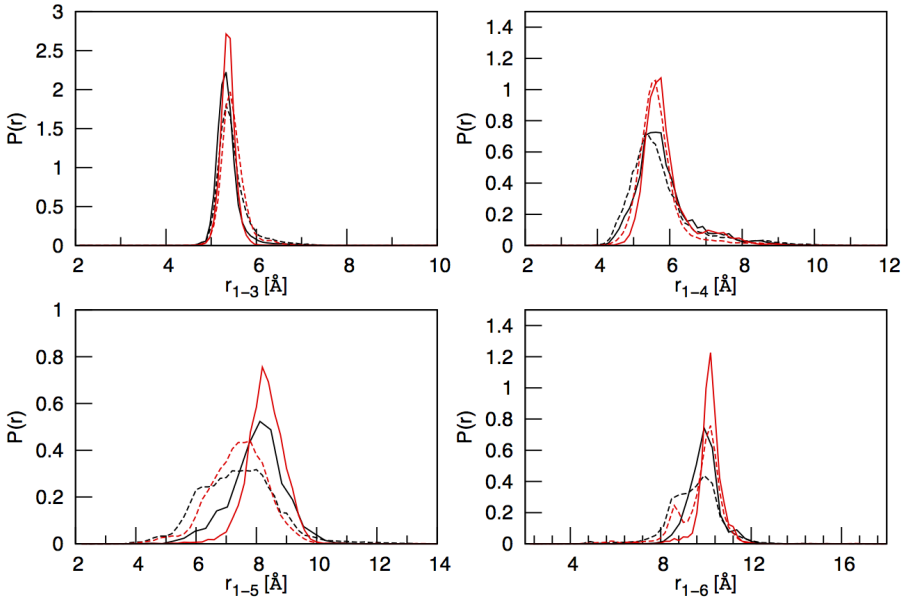


Figure 46: Distance distributions for 3_{10}-helix: repeated distance between beads i and i+2 beads after in the chain (r_{1-3}, top left), i and i+3 (r_{1-4}, top right), i and i+4 (r_{1-5}, bottom left), i and i+5 (r_{1-6}, bottom right). Color codes are: red for Xray or black for NMR and solid line for DSSP and dashed line for PDB.

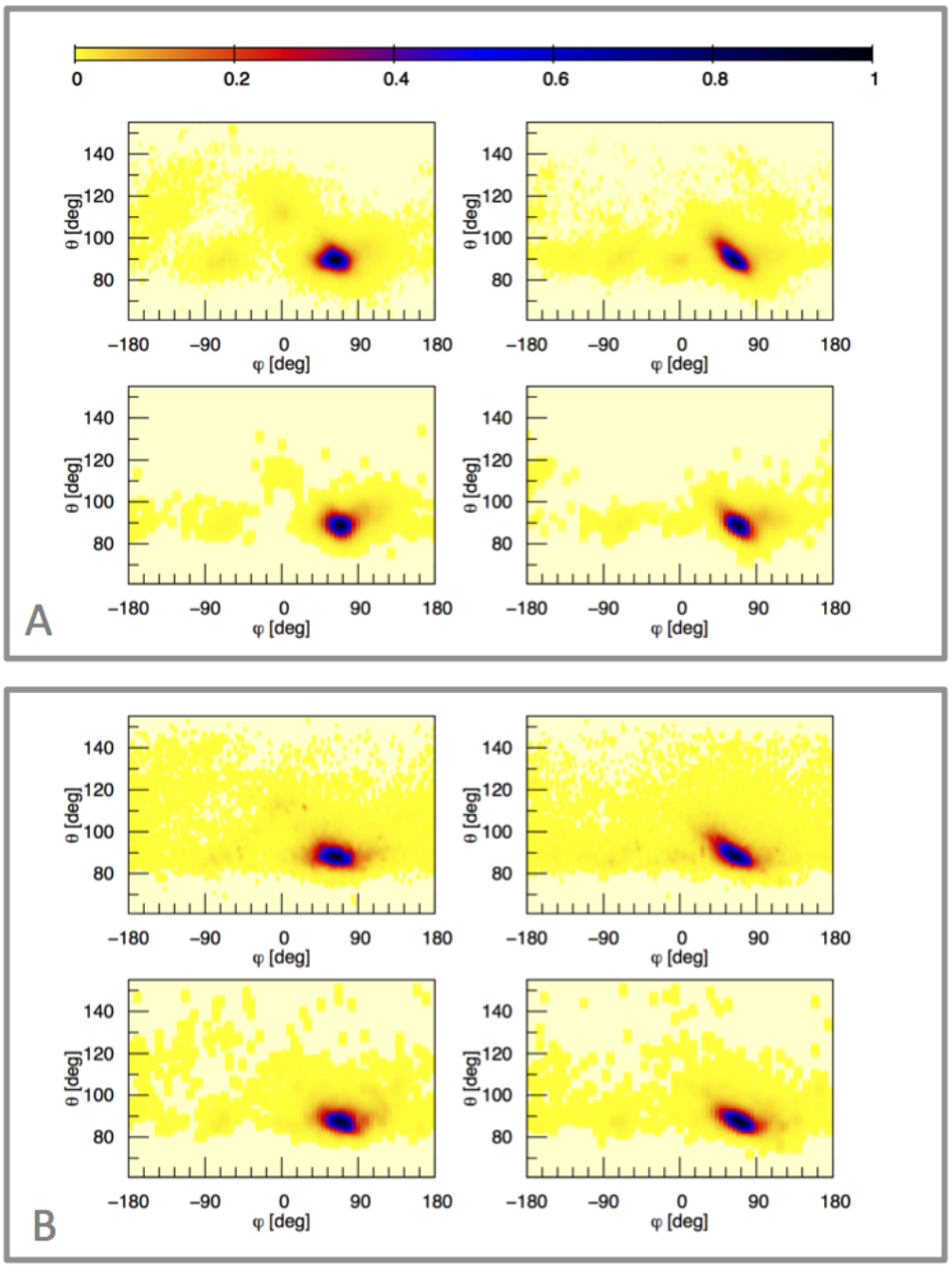


Figure 47: Experimental (θ,ϕ) maps for 310-helix. Panel A: Xray. Panel B:NMR. For both panel: Top line PDB data. Bottom line DSSP data. Correlation plots for (θ−,ϕ) on the left column and (θ+,ϕ) on the right column. Color bar is on the top.

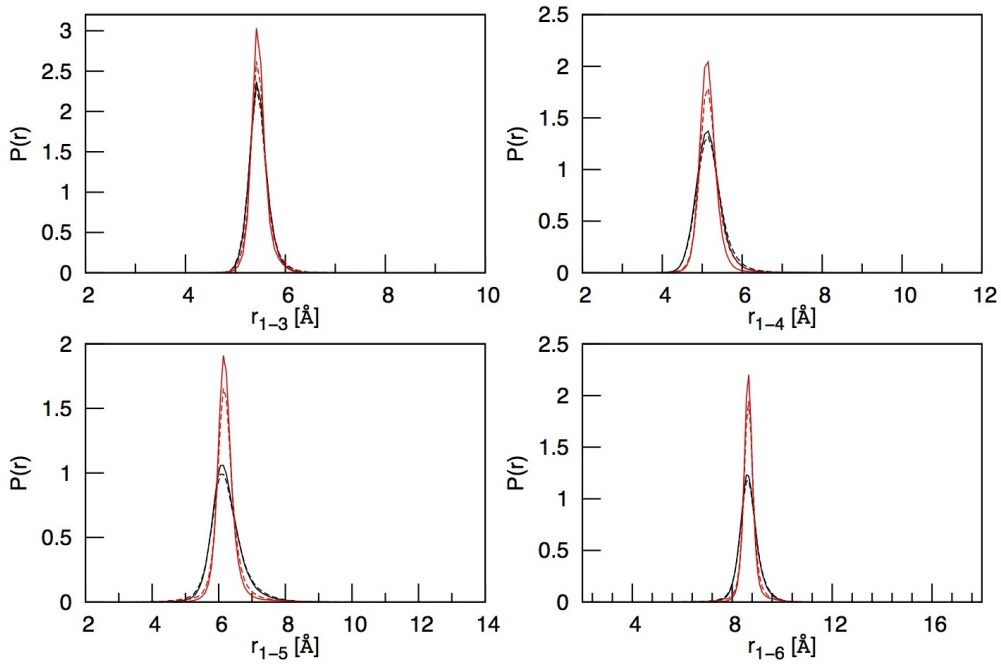


Figure 48: Distance distributions for α-helix: repeated distance between beads i and i+2 beads after in the chain (r_{1-3}, top left), i and i+3 (r_{1-4}, top right), i and i+4 (r_{1-5}, bottom left), i and i+5 (r_{1-6}, bottom right). Color codes are: red for Xray or black for NMR and solid line for DSSP and dashed line for PDB.

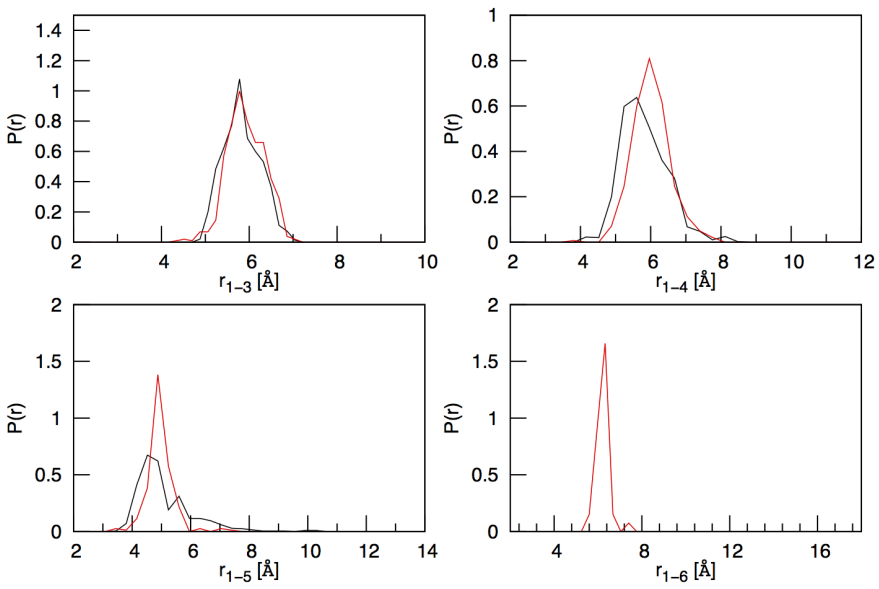


Figure 49: Distance distributions for π-helix: repeated distance between beads i and i+2 beads after in the chain (r_{1-3}, top left), i and i+3 (r_{1-4}, top right), i and i+4 (r_{1-5}, bottom left), i and i+5 (r_{1-6}, bottom right). Color codes are: red for Xray or black for NMR and solid line for DSSP.

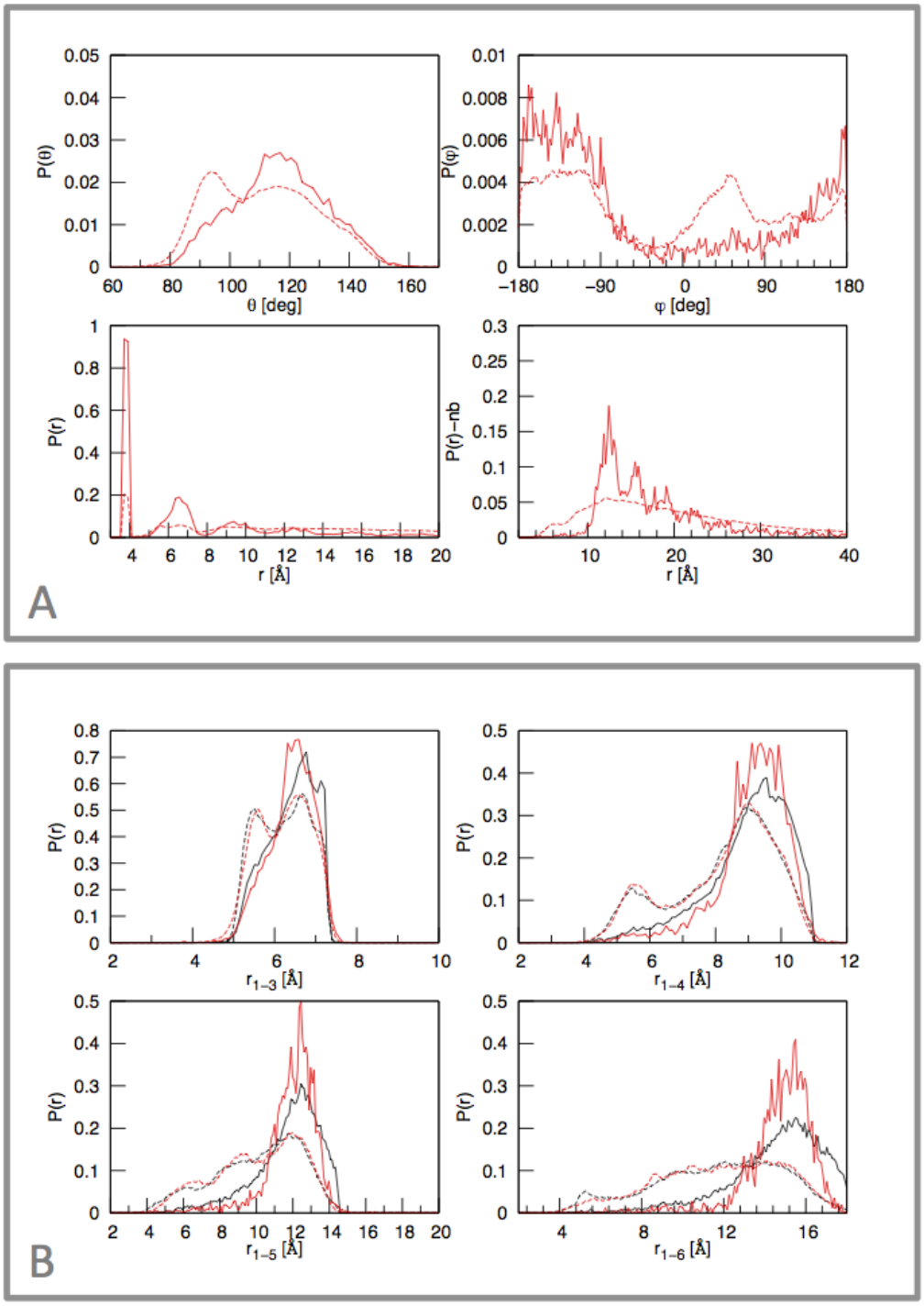


Figure 50: Internal DOFs distributions for unstructured proteins. Panel A: distributions of θ (top-left), ϕ (top-right), $P(r)$ for r distance between every two C_α (bottom left) and $P(r) - nb$ for r distance between every i and j with $j > i + 2$ (bottom right). Panel B: distributions for the repeated distance between beads i and $i+2$ beads after in the chain (r_{1-3}, top left), i and $i+3$ (r_{1-4}, top right), i and $i+4$ (r_{1-5}, bottom left), i and $i+5$ (r_{1-6}, bottom right). Color codes are: red for Xray or black for NMR and solid line for DSSP and dashed line for PDB.

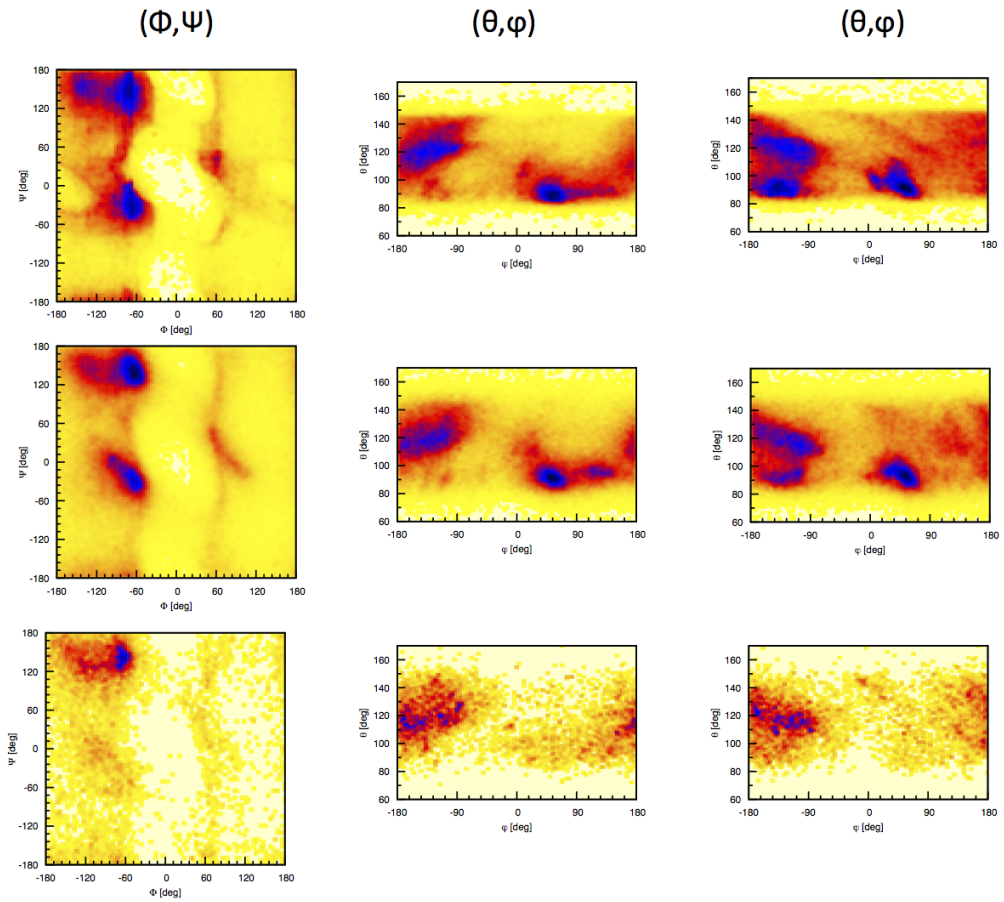


Figure 51: Conformational plots for unstructured proteins. Top line: data from solved with NMR and structures identified with PDB direct information. Center line: data from X-ray and structures distinguished with DSSP algorithm. Bottom line: data from X-ray and structures identified by PDB direct information. Columns are explained on the top. Color bar is the same as in figure 47.

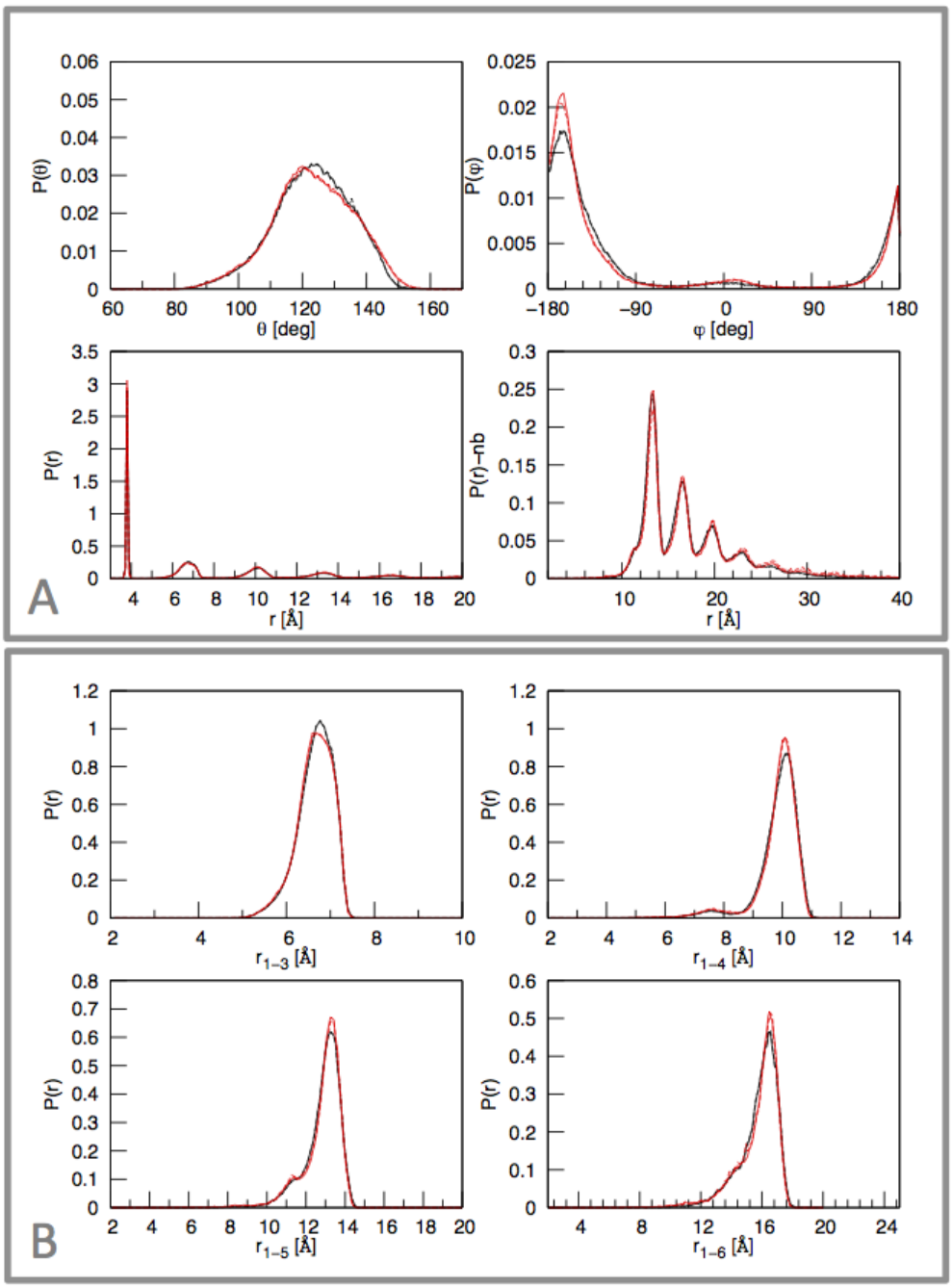


Figure 52: Internal DOFs distributions for strand. Panel A: distributions of θ (top-left), ϕ (top-right), $P(r)$ for r distance between every two C_α (bottom left) and $P(r) - nb$ for r distance between every i and j with $j > i + 2$ (bottom right). Panel B: distributions for the repeated distance between beads i and i+2 beads after in the chain (r_{1-3}, top left), i and i+3 (r_{1-4}, top right), i and i+4 (r_{1-5}, bottom left), i and i+5 (r_{1-6}, bottom right). Color codes are: red for Xray or black for NMR and solid line for DSSP and dashed line for PDB. There is no distinction between parallel or antiparallel strands

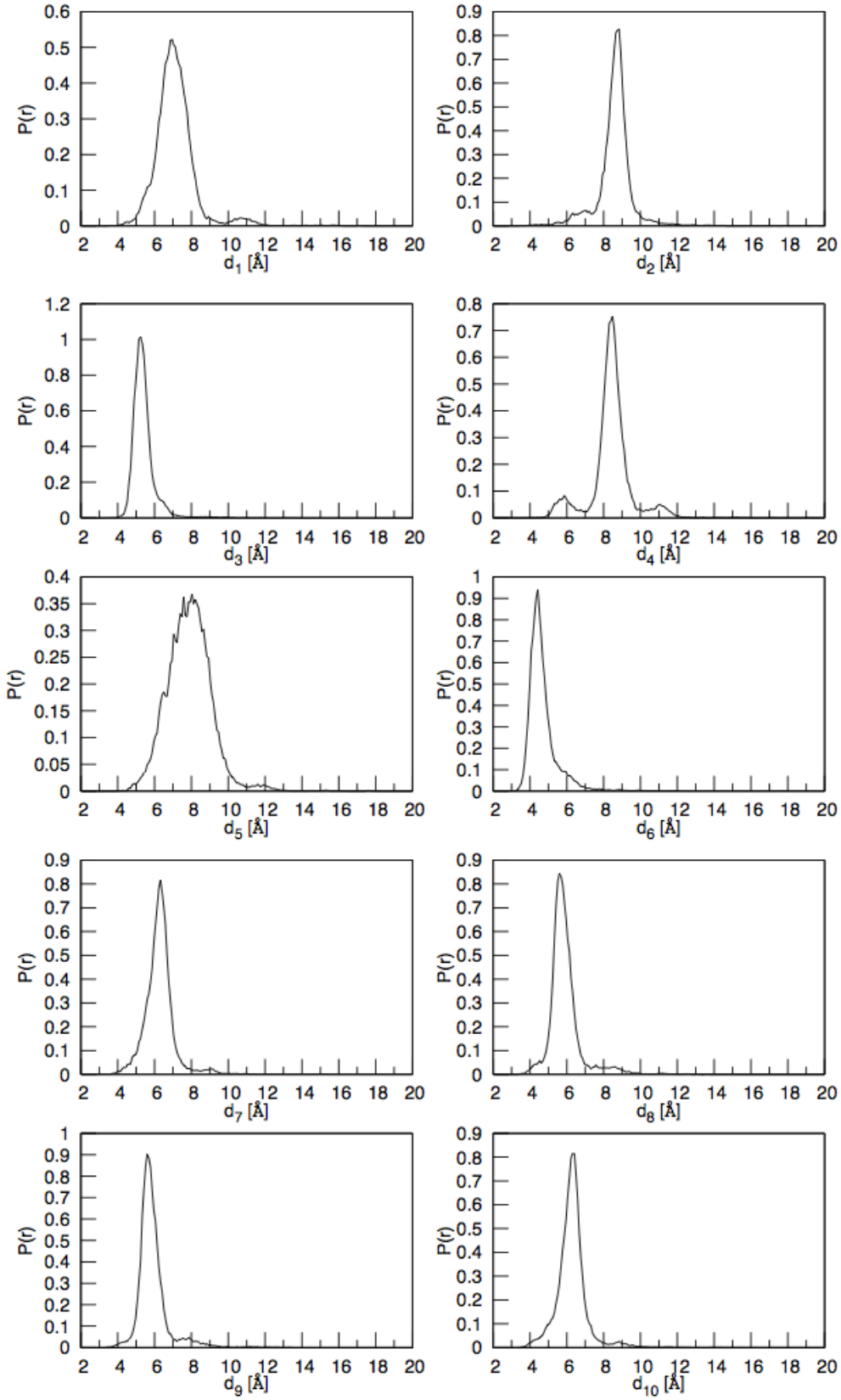


Figure 53: Inter-strand distances distributions for antiparallel sheets. Different distances are identified with the same name as in figure 30. Structures are solved with NMR.

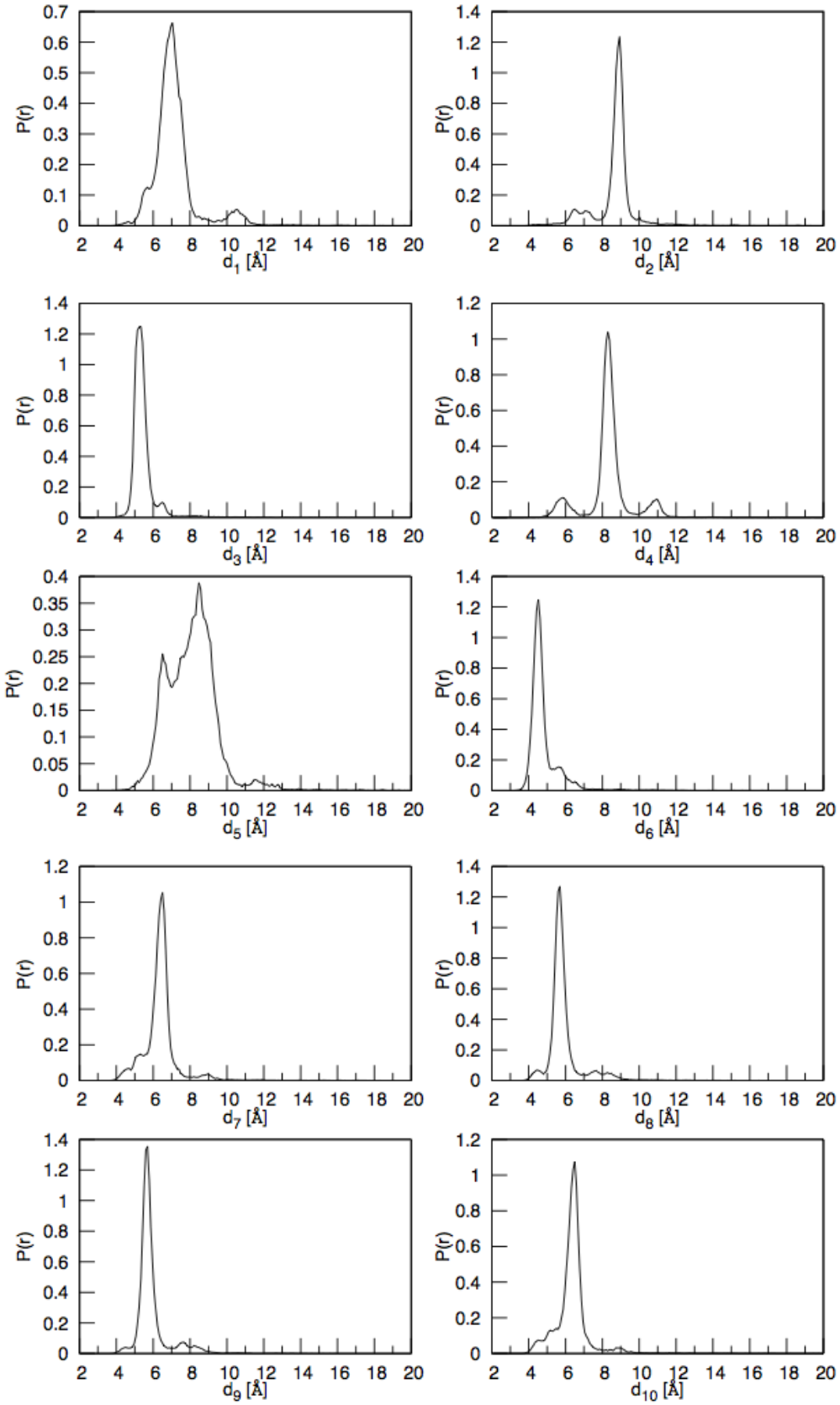


Figure 54: Inter-strand distances distributions for antiparallel sheets. Different distances are identified with the same name as in figure 30. Structures are solved with Xray.

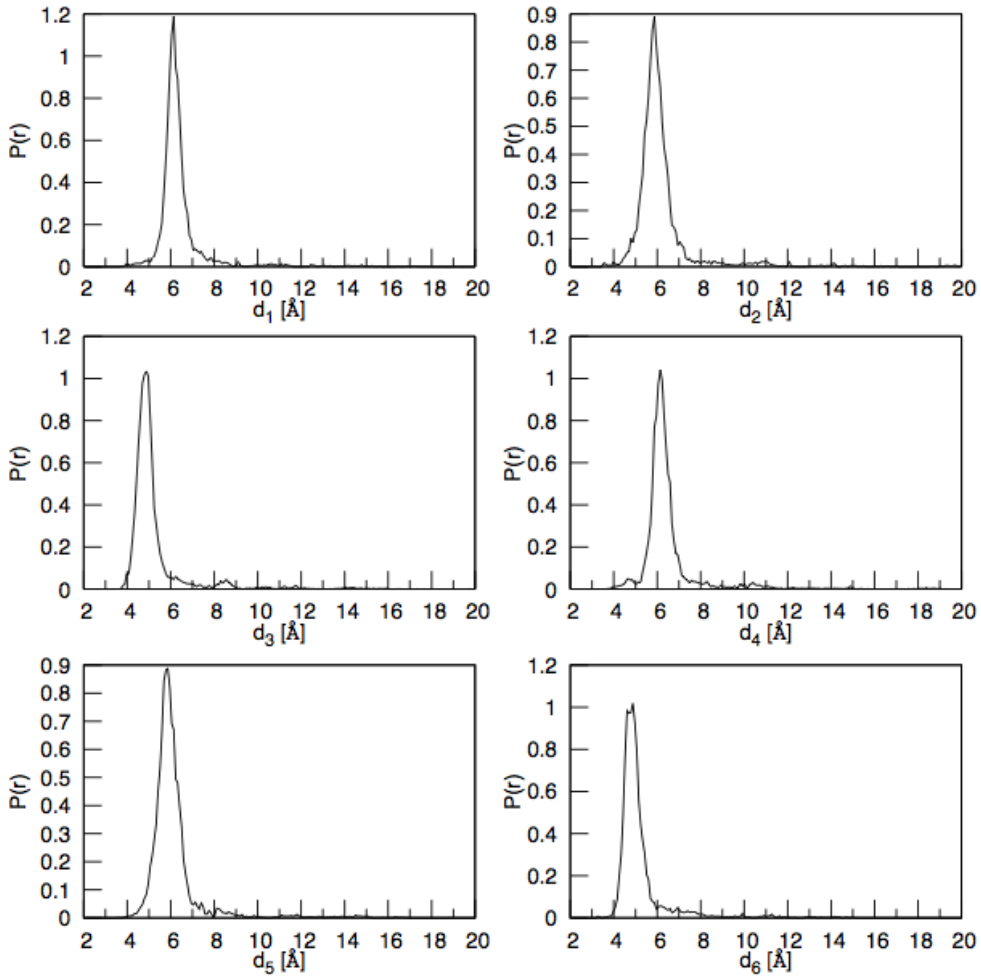


Figure 55: Inter-strand distances distributions for parallel sheets. Different distances are identified with the same name as in figure 30. Structures are solved with NMR.

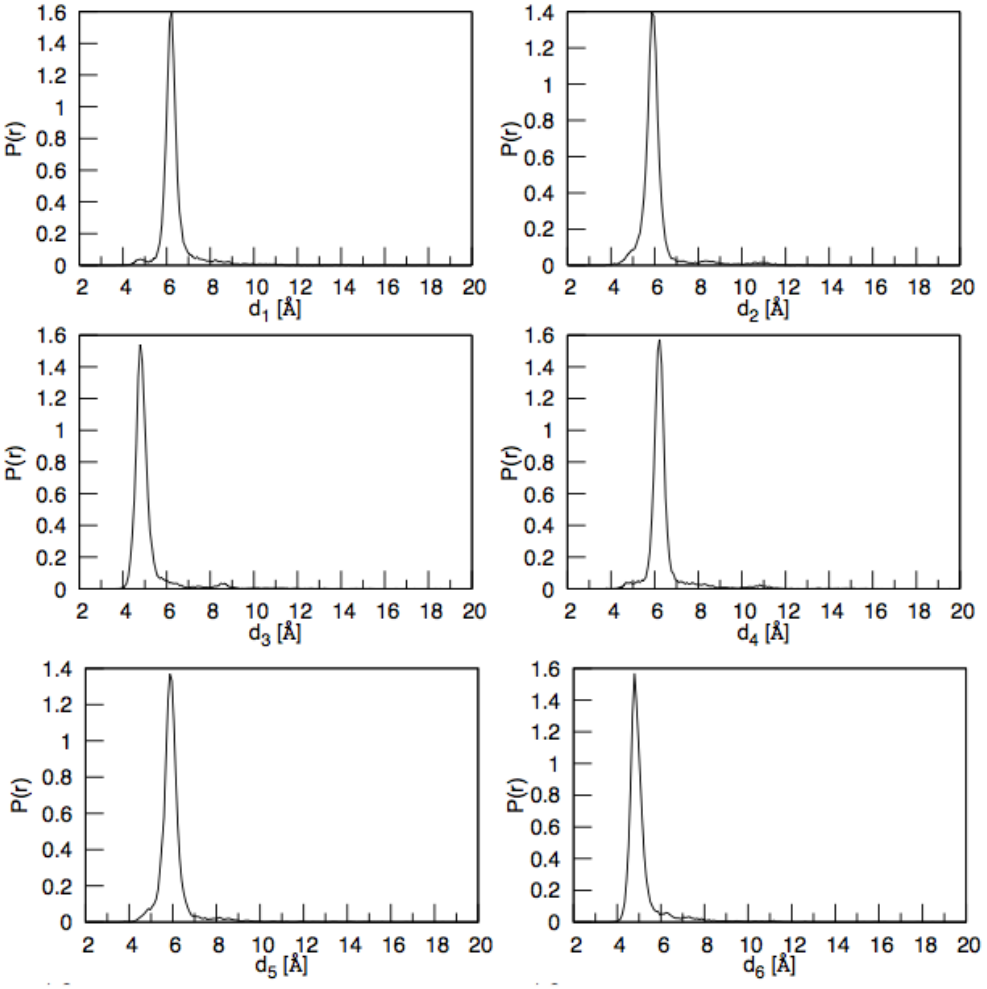


Figure 56: Inter-strand distances distributions for antiparallel sheets. Different distances are identified with the same name as in figure 30. Structures are solved with Xray.

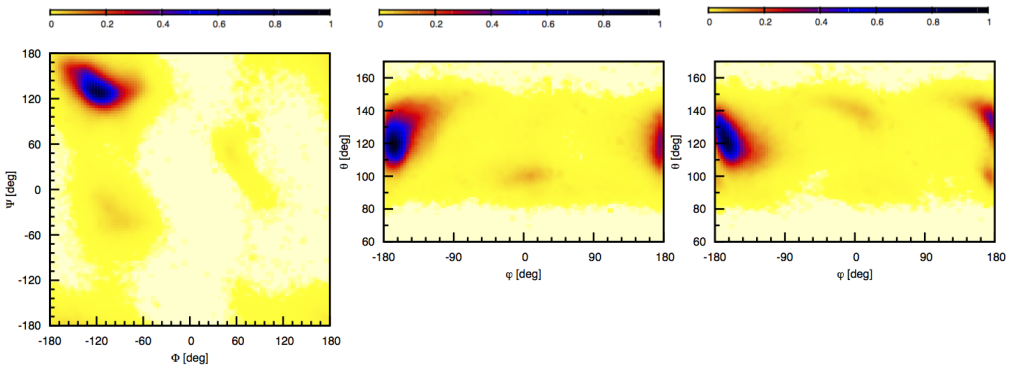


Figure 57: Conformational plots for extended strands. Conformational plots for a dataset of Xray β-strands. The distinction of the structure is made using the PDB entries. Left (Φ,Ψ) map, center (θ−,ϕ) map, right (θ+,ϕ) map. The data are normalized to the maximum count.

D APPENDIX

D.1 DL_POLY

DL_POLY is a general purpose parallel molecular dynamics simulation package developed at Daresbury Laboratory by W. Smith and I.T. Todorov. DL_POLY Classic is freely distributed under BSD2 licence.
To execute it there are required three input files named CONFIG, FIELD, CONTROL.
In the CONTROL file there are all the directives to run the simulation as the temperature, the timestep and the print range on output files. An example of CONTROL file is reported in listing 3.

Listing 3: Example of CONTROL file

```
steps          5000000
timestep    0.010
restart
...
...
ensemble nvt hoover 0.5
temperature      300
...
print  400
traj 1 400 1
...
finish
```

The CONFIG file contains the information about the initial configuration of the system, types of amino acids contained and their positions. An example of CONFIG file is reported in listing 4.

Listing 4: Example of CONFIG file

```
 Polypeptide model
         2         0        20  -89201.2616280
CAX              1
    -4.807728850         10.88150629        -9.394864917
 -0.348834394195      0.404144586899E-01  0.949274808032
   130.381996169       1776.51406099       237.920388075
```

In the FIELD file all the interactions between amino acids are reported together with their functional forms and parameters. The FIELD file contains all the Force Field information. It is divided in two main parts: first the topologically connected interactions are listed, i.e. the potentials depending on the specific bead positions like U^{θ}, then the non bonded interactions, i.e. U^{nb}.

Listing 5: Example of FIELD file

```
 1 Bead Model
 units  kcal
 molecular types 1
 protein
 nummols 1
atoms   20
CAX         115.0000      0.0000    1    0
CAX         115.0000      0.0000    1    0
...
...
 constraints          19
    1    2    3.8100
    2    3    3.8100
...
...
bonds   51
mors    1    3       6.500     5.42000     1.30000
mors    1    4       3.500     5.15000     0.80000
...
...
angles  18
hcos    1    2    3      10.000    92.00000
hcos    2    3    4      10.000    92.00000
...
...
 finish
   vdw   1
     CAX     CAX dblw   25.00000   0.10000  0.61D+01   0.10000   0.70000
         0.96D+01
 close
```

DL_POLY returns some different output files. The more important are the following. The HISTORY files contains coordinates and velocities for each step of simulation. The OUTPUT file is divided in seven sections. It contains after an header and a summary of the three input files, the status of the simulation and a summary of the statistical data. Finally, in the STATIS file there are all the informations about the various contribution to the energy of the system.

D.2 ADDITIONAL SIMULATION RESULTS FOR THE FIRST AND OPTIMIZED FORCE FIELD SETS

In this section additional simulations results are reported to complete those commented in Chapter 4. A detailed description is in the captions.

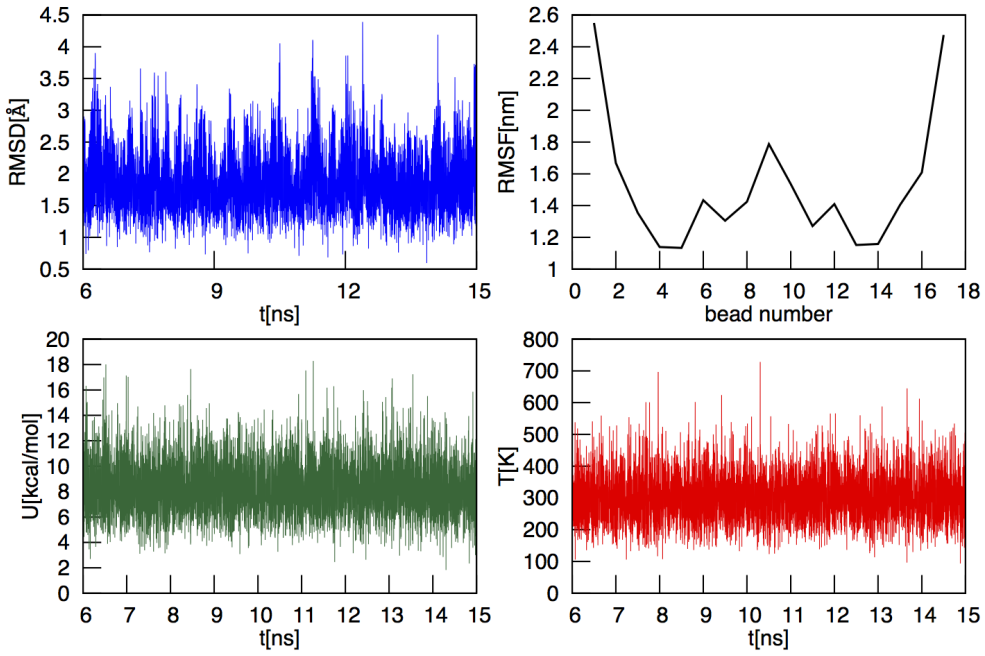


Figure 58: Simulation results for the initial field of the 3_{10}-helix. The system used is a peptide of 17 amino acids. Top left: RMSD of the production run. Top right: RMSF of all the peptide beads. Bottom left: potential energy U of the system for only the production run. Bottom right: temperature (K) of the system for the production run.

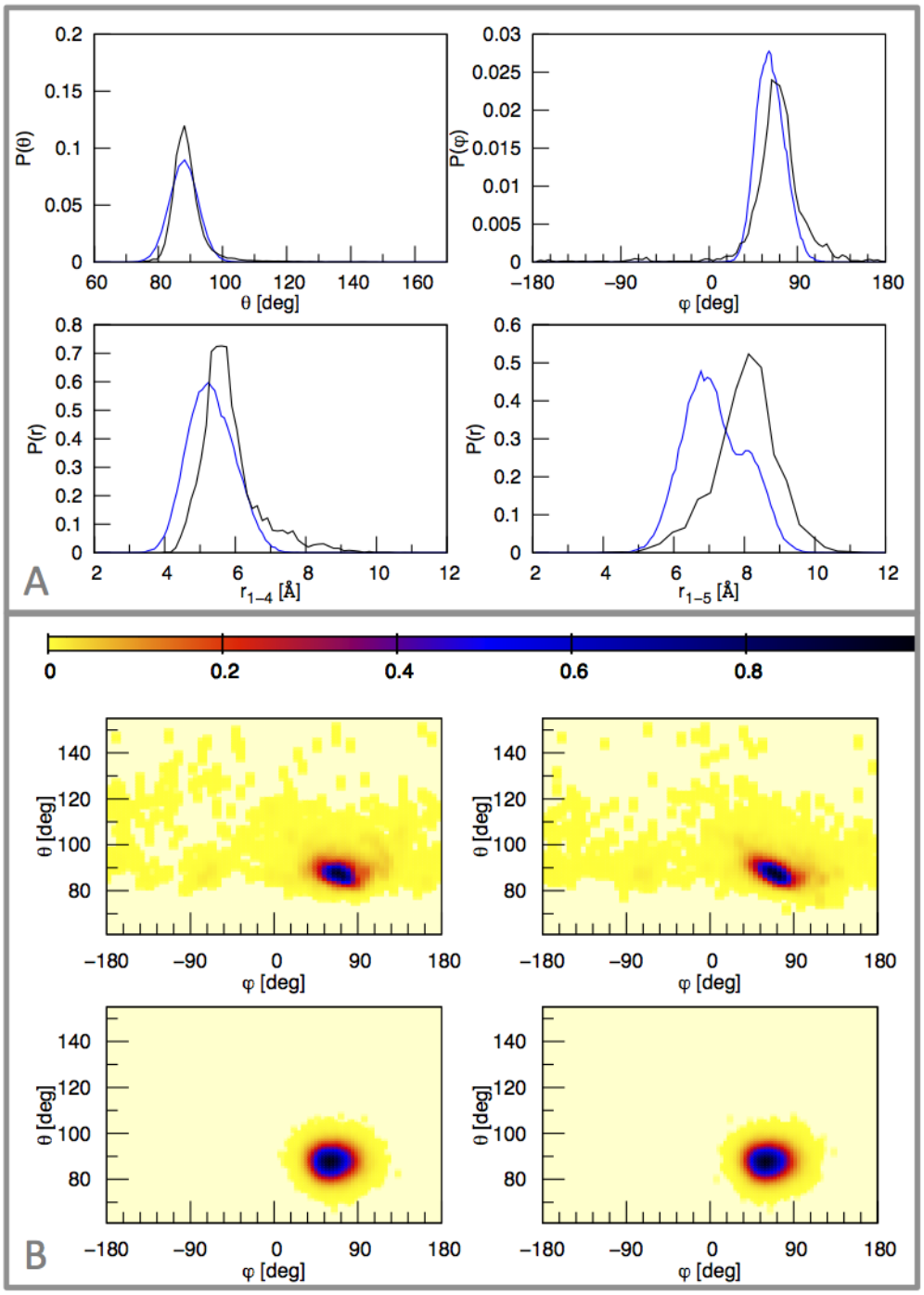


Figure 59: Comparison of experimental and simulated distributions and correlations for the 3_{10}-helix. (Panel A) Black: experimental distributions obtained from the dataset of 3_{10}-helix solved with NMR and recognized with DSSP. Blue: distributions obtained with simulation of the 17 beads long 3_{10}-helix with FF1. (Panel B) Top line: ($\theta-$,ϕ) map on the left and ($\theta-$,ϕ) plot on the right for the experimental NMR DSSP dataset. Bottom line: correlation plots corresponding to the top line for the simulation results. The simulation was performed with the 17 beads long 3_{10}-helix and with FF1. Color bar on the top.

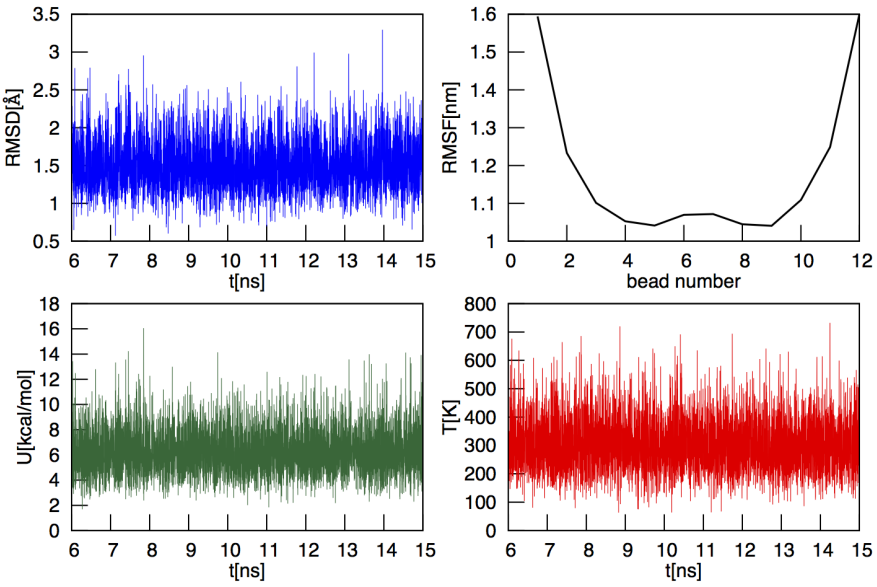


Figure 60: Simulation results for the initial field of the π-helix. The system used is a peptide of 12 amino acids. Top left: RMSD of the production run. Top right: RMSF of all the peptide beads. Bottom left: total energy of the system for the production run. Bottom right: temperature of the system for the production run.

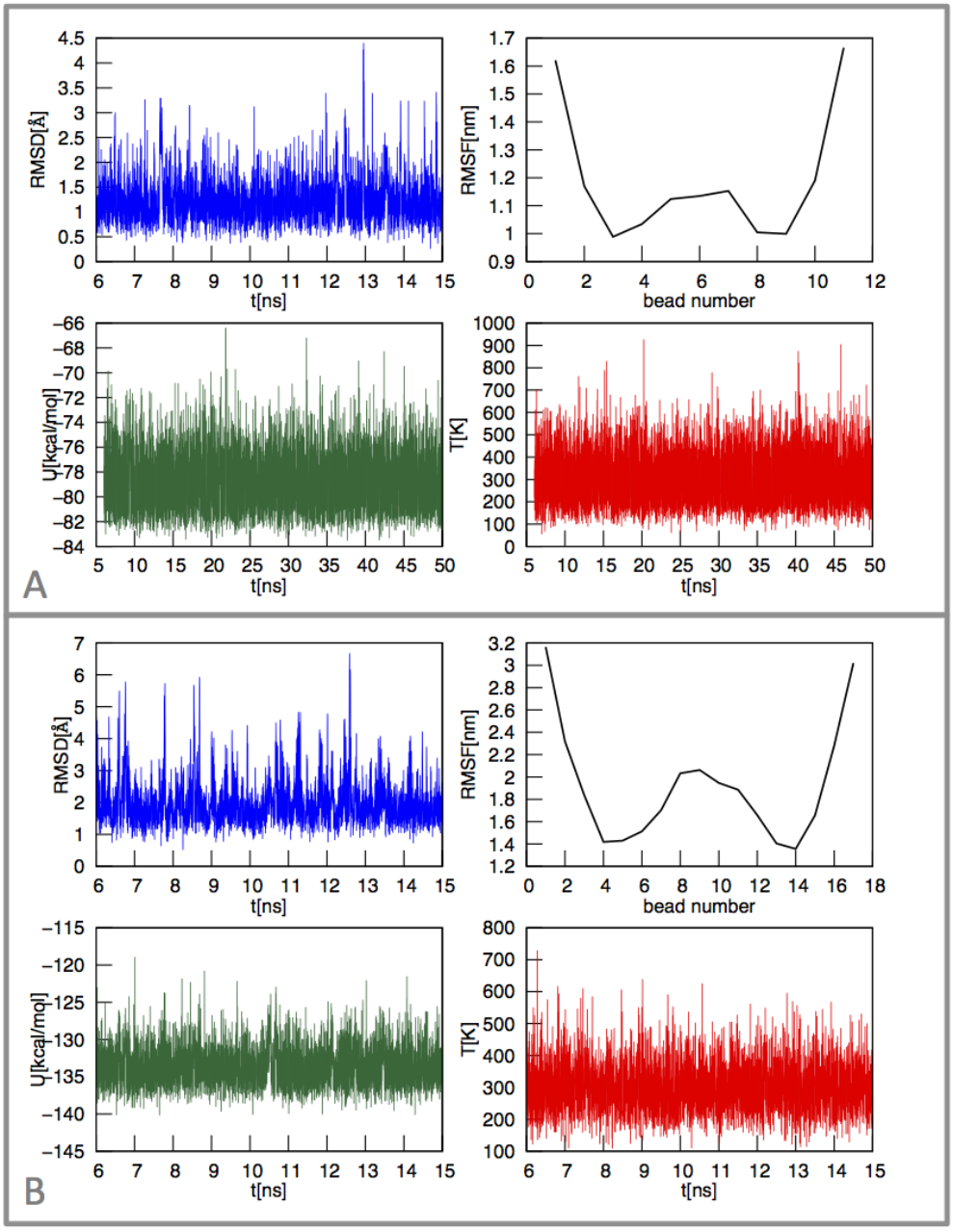


Figure 61: Simulation results for the optimized field of the 3_{10}-helix. Panel A: Simulation results for the 11 amino acids long helix. Panel B: Simulation results for the 17 amino acids long helix.

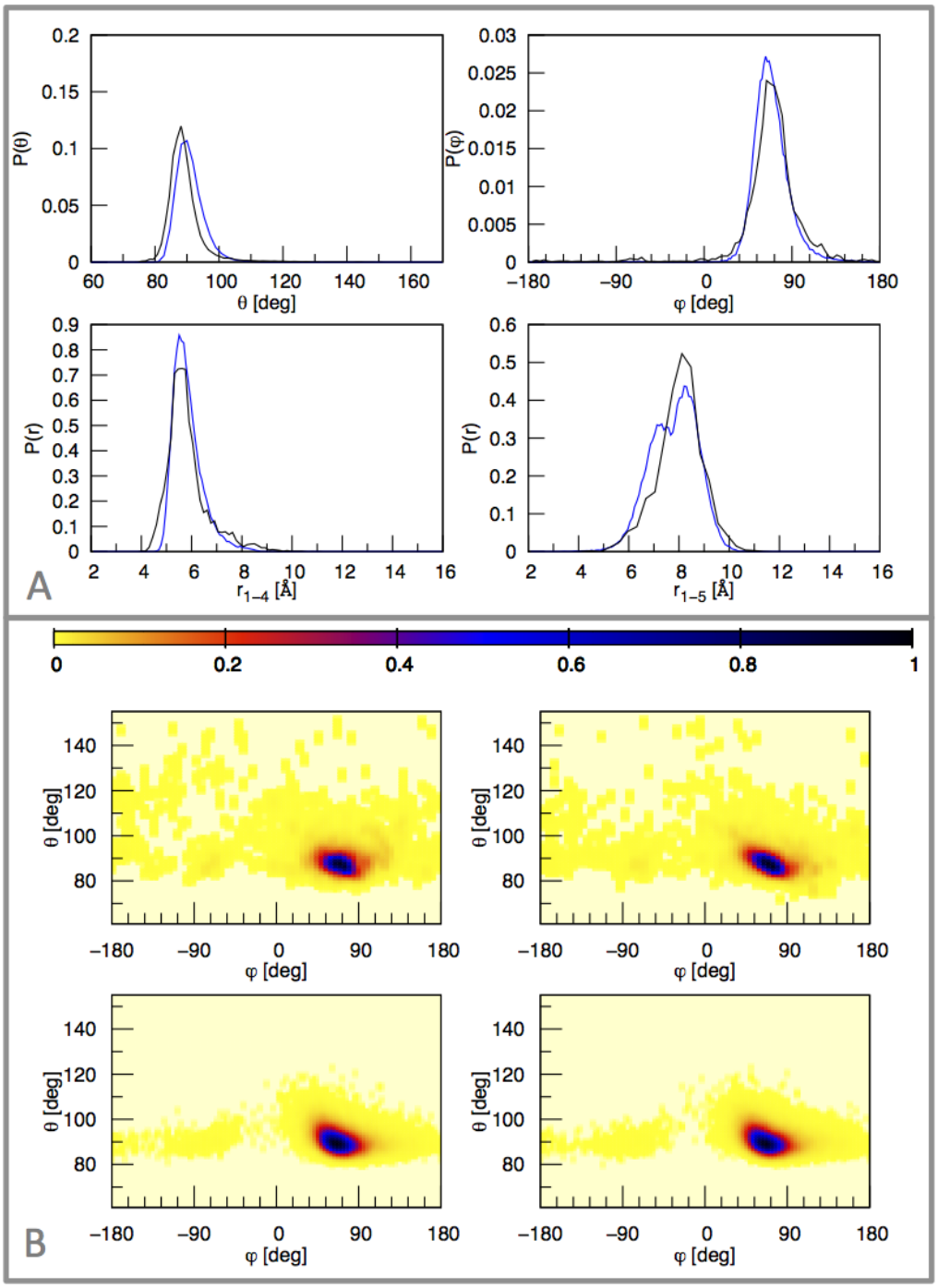


Figure 62: Comparison of experimental and simulated distributions and correlations for the 3_{10}-helix. (Panel A) Black: experimental distributions obtained from the dataset of 3_{10}-helix solved with NMR and recognized with DSSP. Blue: distributions obtained with simulation of the 17 beads long 3_{10}-helix with FF2. (Panel B) Top line: ($\theta-$,ϕ) map on the left and ($\theta-$,ϕ) plot on the right for the experimental NMR DSSP dataset. Bottom line: correlation plots corresponding to the top line for the simulation results. The simulation was performed with the 17 beads long 3_{10}-helix and with FF2. Color bar on the top.

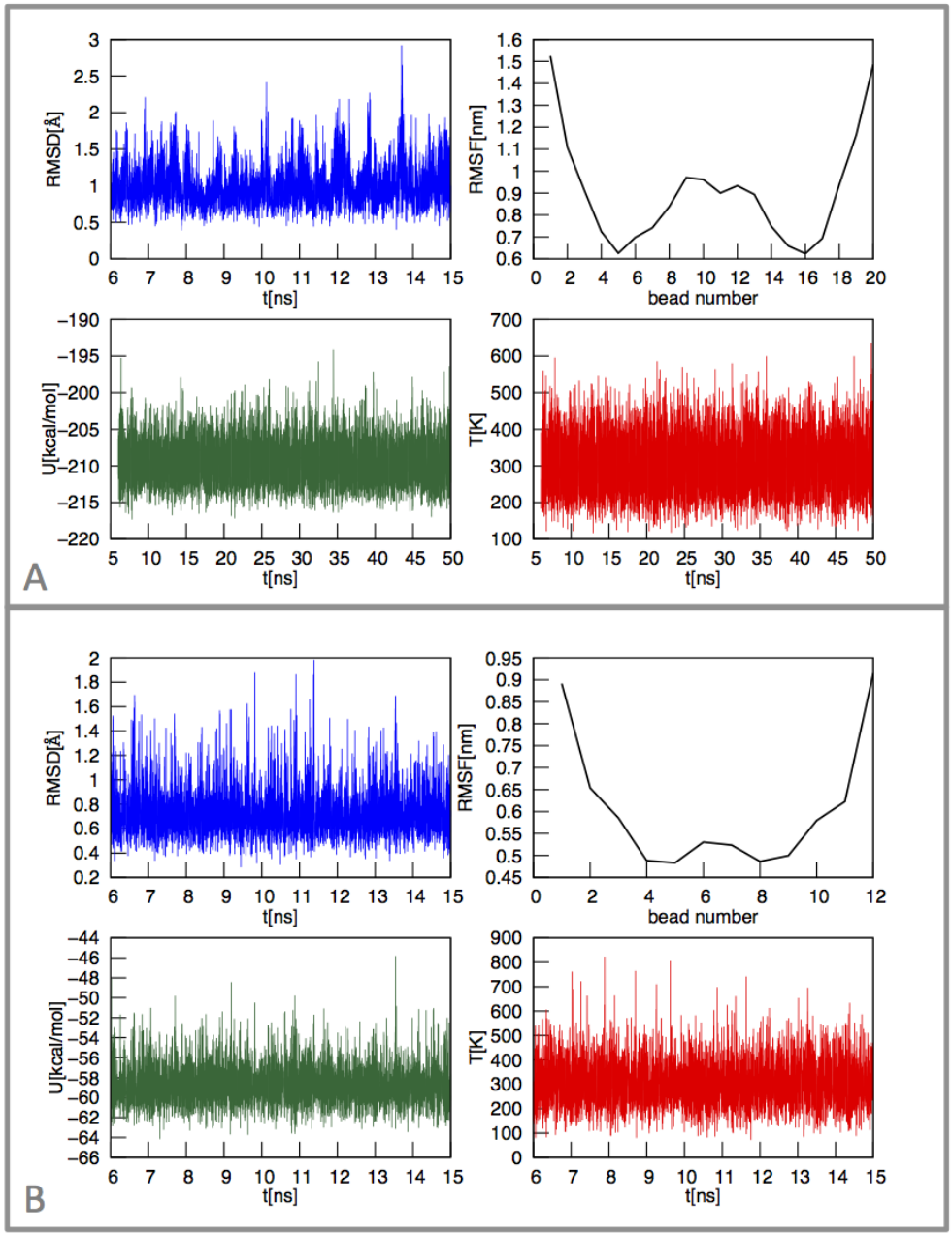


Figure 63: Simulation results for the optimized field of the α-helix (Panel A) and of the π-helix (Panel B)

BIBLIOGRAPHY

[1] Hoang TX; Trovato A; Seno F; Banavar JR; Maritan A. "Geometry and symmetry presculpt the free-energy landscape of proteins". In: *PNAS* 101 (2004), 7960–7964 (cit. on p. 34).

[2] Cooley RB; Arp D; Karplus AP. "Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices enhancing protein functionality". In: *J. Mol. Biol.* 404 (2010), pp. 232–246 (cit. on p. 13).

[3] Garnier J; Osguthorpe DJ; Robson B. "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins". In: *J. Mol. Biol.* 120 (1978), pp. 97–120 (cit. on p. 18).

[4] Mukherjee A; Bagchi B. "Correlation between rate of folding, energy landscape and topology in the folding of a model protein HP-36". In: *J. Chem. Phys.* 118 (2002), pp. 4733–4747 (cit. on p. 25).

[5] *BLAST: Basic Local Alignment Search Tool*. 2013. URL: http://blast.ncbi.nlm.nih.gov/Blast.cgi (cit. on p. 18).

[6] Crisma M; Formaggio F; Moretto A; Toniolo C. "Peptide helices based on α-amino acids". In: *Biopolymers* 84 (2006), pp. 3–12 (cit. on pp. 10–12, 44).

[7] Das P; Matysiak S; Clementi C. "Balancing energy and entropy: a minimalist model for the characterization of protein folding landscapes". In: *Proc. Natl. Acad. Sci. U.S.A.* 102 (2005), 10141–10146 (cit. on p. 32).

[8] Kabsch W; Sander C. "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". In: *Biopolymers* 22 (1983), pp. 2577–2637 (cit. on pp. 13, 17, 40).

[9] Rost B; Sander C. "Prediction of protein secondary structure at better than 70% accuracy". In: *J. Mol. Biol.* 232 (1993), pp. 584–599 (cit. on p. 18).

[10] Wolfenden R; Andersson L; Cullis P; Southgate C. "Affinities of amino acid side chains for solvent water". In: *Biochemistry* 20 (1981), pp. 849–855 (cit. on p. 6).

[11] Venkatachala CM. "Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units". In: *Biopolymers* 6 (1968), pp. 1425–1436 (cit. on p. 15).

[12] *Cube Plugin*. 2012. URL: http://www.ks.uiuc.edu/Research/vmd/plugins/molfile/cubeplugin.html (cit. on p. 94).

[13] Klimov DK; Thirumalai D. "Mechanisms and kinetics of β-hairpin formation". In: *Proc. Natl. Acad. Sci. U.S.A.* 97 (2000), 2544–2549 (cit. on p. 25).

[14] Potter D. *Computational Physics*. WILEY, 1972 (cit. on p. 35).

[15] Whitford D. *Proteins. Structure and functions*. WILEY, 2005 (cit. on pp. 2, 11–14, 21, 44).

[16] Montgomerie S; Sundararaj S; Gallin WJ; Wishart DS. "Improving the accuracy of protein secondary structure prediction using structural alignment". In: *BMC Bioinformatics* 7 (2006), pp. 301–314 (cit. on p. 18).

[17] Jones DT. "Protein secondary structure prediction based on position-specific scoring matrices". In: *J. Mol. Biol.* 292 (1999), pp. 195–202 (cit. on p. 18).

[18] Salemme FR; Weatherford DW. "Conformational and geometrical properties of β-sheets in proteins. I. Parallel β-sheets". In: *J. Mol. Biol.* 146 (1981), pp. 101–117 (cit. on p. 14).

[19] Salemme FR; Weatherford DW. "Conformational and geometrical properties of β-sheets in proteins. II. Antiparallel and mixed β-sheets". In: *J. Mol. Biol.* 146 (1981), pp. 119–141 (cit. on p. 14).

[20] Buxbaum E. *Fundamentals of protein structure and function*. Springer, 2007 (cit. on p. 8).

[21] Alemani D; Collu F; and Cascella M; Dal Peraro M. "A nonradial coarse-grained potential for proteins produces naturally stable secondary structure elements". In: *J. Chem. Theor. Comput.* 6 (2010), pp. 315–324 (cit. on pp. 33, 34).

[22] Reith D; Pütz M; Müller-Plathe F. "Deriving effective mesoscale potentials from atomistic simulations". In: *J. Comput. Chem.* 24 (2003), pp. 1624–1636 (cit. on p. 31).

[23] Salemme FR. "Structural properties of protein β-sheets". In: *Prog. Biophys. Mol. Biol.* 42 (1983), pp. 95–133 (cit. on pp. 14, 16).

[24] Lyman E; Pfaendtner J; Voth GA. "Systematic multiscale parameterization of Heterogeneous Elastic Network models of proteins". In: *Biophys. J.* 95 (2008), 4183–4192 (cit. on p. 29).

[25] Noid WG; Liu P; Wang Y; Chu JW; Ayton GS; Izvekov S; Andersen HC; Voth GA. "The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models". In: *J. Chem. Phys.* 128 (2008), p. 244115 (cit. on p. 32).

[26] Voth GA. *Coarse-Graining of condensed phase and biomolecular systems*. CRC Press, 2008 (cit. on p. 24).

[27] Chou PY; Fasman GD. "Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins". In: *Biochemistry* 13(2) (1974), pp. 211–222 (cit. on pp. 17, 18).

[28] Chou PY; Fasman GD. "Prediction of protein conformation". In: *Biochemistry* 13(2) (1974), pp. 222–245 (cit. on p. 17).

[29] Hanwell MD; Curtis DE; Lonie DC; Vandermeersch T; Zurek E; Hutchison GR. "Avogadro: an advanced semantic chemical editor, visualization, and analysis platform". In: *J. Cheminform.* 4:17 (2012) (cit. on pp. 44, 64, 70).

[30] Berendsen HJC; Postma JPM; van Gunsteren WF; Hermans J. "Molecular-dynamics with coupling to an external bath". In: *J. Chem. Phys.* 81 (1984), pp. 3684–3690 (cit. on p. 37).

[31] Gō N; Scheraga HA. "On the use of classical statistical mechanics in the treatment of polymer chain conformation". In: *Macromolecules* 9(4) (1976), pp. 535–542 (cit. on pp. 27, 29).

[32] Andersen HC. "Rattle: A "Velocity" version of the Shake algorithm for molecular dynamics calculations". In: *J. Comput. Phys.* 52 (2001), pp. 24–34 (cit. on pp. 36, 91).

[33] Noid WG; Chu JW; Ayton GS; Krishna V; Izvekov S; Voth GA; Das A; Andersen HC. "The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models". In: *J Chem. Phys.* 128 (2008), p. 244114 (cit. on p. 32).

[34] Pauling L; Corey RB; Branson HR. "The structure of proteins: two hydrogen-bonded helical congurations of the polypeptide chain". In: *Proc Natl Acad Sci USA* 37 (1951), pp. 205–211 (cit. on pp. 11, 12, 44).

[35] Atilgan AR; Durell SR; Jernigan RL; Demirel MC; Keskin O; Bahar I. "Anisotropy of fluctuation dynamics of proteins with an Elastic Network model". In: *Biophys. J.* 80 (2001), pp. 505–515 (cit. on p. 29).

[36] Chennubhotla C; Rader AJ; Lee-Wei Yang LW; Bahar I. "Elastic Network models for understanding biomolecular machinery : from enzymes to supramolecular assemblies". In: *Phys. Biol.* 2 (2005), S173–S180 (cit. on p. 29).

[37] Hamelryck T; Borg M; Paluszewski M; Paulsen J; Frellsen J; Andreetta C; Boomsma W; Bottaro S; Ferkinghoff-Borg J. "Potential of Mean Force for protein structure prediction vindicated, formalized and generalized". In: *Plos One* 5(11) (2010), e13714 (cit. on pp. 30, 31).

[38] Janin J. "Surface and inside volumes in globular proteins". In: *Nature* 227 (1979), pp. 491–492 (cit. on p. 6).

[39] Marx D; Hutter J. *Ab Initio molecular dynamics: theory and implementation*. NIC Series, 2000 (cit. on p. 23).

[40] Tozzini V; McCammon JA. "A coarse grained model for the dynamics of flap opening in HIV-1 protease". In: *Chem. Phys. Lett.* 413 (2005), pp. 123–128 (cit. on pp. 31, 33).

[41] Tozzini V; Rocchia W; McCammon JA. "Mapping All-Atom models onto One-Bead Coarse-Grained models: general properties and applications to a minimal polypeptide model". In: *J. Chem. Theory Comput.* 2 (2006), pp. 667–673 (cit. on pp. 27, 41–43).

[42] Tozzini V; Trylska J; Chang CE; McCammon JA. "Flap opening dynamics in HIV-1 protease explored with a Coarse-Grained model". In: *J. Struct. Biol.* 157 (2007), pp. 606–615 (cit. on pp. 27, 31).

[43] Ercolessi F; Adams JB. "Interatomic potentials from first-principles calculations: the force-matching method". In: *Europhys. Lett.* 26 (1994), pp. 583–588 (cit. on p. 32).

[44] Ryckaert JP; Ciccotti G; Berendsen JC. "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes". In: *J. Comput. Phys.* 23 (1977), pp. 327–341 (cit. on pp. 36, 91).

[45] Humphrey W; Dalke A; Schulten K. "VMD - Visual Molecular Dynamics". In: *J. Mol. Graph.* 14 (1996), pp. 33–38 (cit. on pp. xi, 4, 54).

[46] Verlet L. "Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules". In: *Phys. Rev.* 159 (1967), pp. 98–103 (cit. on p. 35).

[47] Sippl LJ. "Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins". In: *J. Mol. Biol.* 213 (1990), pp. 859–883 (cit. on p. 31).

[48] Rotondi KS; Gierasch LM. "Natural polypeptide scaffolds: β-sheets, β-turns, and β-hairpins". In: *Biopolymers* 84 (2006), pp. 13–22 (cit. on pp. 14, 16).

[49] Becker OM; MacKerell AD; Roux B; Watanabe M. *Computational Biochemistry and Biophysics*. Marcel Dekker, Inc., 2001 (cit. on p. 36).

[50] Fujiwara K; Toda H; Ikeguchi M. "Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type". In: *BMC Struct. Biol.* 12 (2012) (cit. on pp. 17, 18).

[51] Maragakis P; Karplus M. "Large amplitude conformational change in proteins explored with a Plastic Network model: adenylate kinase". In: *J. Mol.Biol.* 352 (2005), pp. 807–822 (cit. on p. 29).

[52] Nakagawa N; Peyrard M. "Modeling protein thermodynamics and fluctuations at the mesoscale". In: *Phys. Rev. E.* 74 (2006), p. 04191 (cit. on p. 29).

[53] Rose G; Geselowitz A; Lesser G; Lee R; Zehfus M. "Hydrophobicity of amino acid residues in globular proteins". In: *Science* 229 (1985), pp. 834–838 (cit. on p. 6).

[54] Zacharias M. "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction". In: *Prot. Sci.* 11 (2003), 2714–2726 (cit. on p. 24).

[55] Perutz MF. "New X-ray evidence on the configuration of polypeptide chains." In: *Nature* 167 (1951), 1053–1054 (cit. on p. 11).

[56] Nelson DL; Cox MM. *Lehninger. Principles of biochemistry*. 5th edition. W. H. Freeman & Company, 2008 (cit. on pp. 1, 5, 7).

[57] Tirion MM. "Large amplitude elastic motions in proteins from a single-parameter, atomic analysis". In: *Phys. Rev. Lett.* 77 (1996), pp. 1905–1908 (cit. on pp. 27, 29).

[58] Enkhbayar P; Hikichi K; Osaki M; Kretsinger RH; Matsushima N. "3_{10}-helices in proteins are parahelices". In: *Proteins* 64 (2006), 691–699 (cit. on p. 11).

[59] Finkelstein VA; Ptitsyn O. *Protein physics. A course of lectures*. ACADEMIC PRESS, 2002 (cit. on pp. 1, 9, 10).

[60] Frishman D; Argos P. "Knowledge-Based Protein Secondary Structure Assignment". In: *Proteins* 23 (1995), pp. 566–579 (cit. on pp. 13, 17).

[61] Berman HM; Westbrook J; Feng Z; Gilliland G; Bhat TN; Weissig H; Shindyalov IN; Bourne PE. "The Protein Data Bank". In: *Nucl. Acids Res.* 28 (2000), pp. 235–242 (cit. on pp. 4, 40).

[62] Kyte J; Doolite R. "A simple method for displaying the hydropathic character of a protein". In: *J. Mol. Biol.* 157 (1982), pp. 105–132 (cit. on p. 6).

[63] Majek P; Elber R. "A coarse-grained potential for fold recognition and molecular dynamics simulations of proteins". In: *Proteins* 76 (2009), pp. 822–836 (cit. on p. 25).

[64] Low BW; Baybutt RB. "The π-helix. A hydrogen bonded configuration of the polypeptide chain". In: *J. Am. Chem. Soc.* 74 (1952), pp. 5806–5807 (cit. on p. 13).

[65] Pauling L; Corey RB. "Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets". In: *Proc. natl. Acad. Sci. U.S.A.* 37 (1951), pp. 729–740 (cit. on p. 13).

[66] *RCSB PDB-Protein Data Bank*. 2013. URL: http://www.rcsb.org/pdb/home/home.do (cit. on pp. 4, 39).

[67] Lee L; Leopold JL; Frank RL. "Protein Secondary Structure Prediction Using BLAST and Relaxed Threshold Rule Induction from Coverings". In: *CIBCB* (2011) (cit. on p. 18).

[68] Cotterill RMJ. *Biophysics. An introduction*. WILEY, 2002.

[69] Hockney RW. "Potential calculation and some applications". In: *Methods Comput. Phys.* 9 (1970), p. 136 (cit. on p. 35).

[70] Eaton JW; Bateman D; Hauberg S. *GNU Octave manual version 3*. Network Theory Limited, 2008 (cit. on pp. 47, 51, 58).

[71] Fodje MN; Al-Karadaghi S. "Occurrence, conformational features and amino acid propensities for the π-helix". In: *Prot. Eng.* 15 (2002), pp. 353–358 (cit. on p. 13).

[72] Nosé S. "A unified formulation of the constant temperature molecular dynamics methods". In: *J. Chem. Phys.* 81 (1984), pp. 511–520 (cit. on p. 37).

[73] *SecStAnT: Secondary Structure Analysis Tool*. 2013. URL: http://secstant.sourceforge.net (cit. on pp. 39, 40).

[74] Monticelli L; Kandasamy SK; Periole X; Larson RG; Tieleman DP; Marrink SJ. "The MARTINI coarse-grained force field: extension to proteins". In: *J. Chem. Theory. Comput.* 4 (2008), pp. 819–834 (cit. on p. 24).

[75] Ha-Duong T. "Protein backbone dynamics simulations using coarse-grained bonded potentials and simplified hydrogen bonds". In: *J. Chem. Theory. Comput.* 6 (2010), pp. 761–773 (cit. on p. 24).

[76] Smith W; Forester T. "DL_POLY 2.0: A general purpose parallel molecular dynamics simulation package". In: *J. Mol. Graph.* 14 (1996), pp. 136–141 (cit. on p. 65).

[77] Sorenson JM; Head-Gordon T. "Protein engineering study of protein L by simulation". In: *J. Comput. Biol.* 9 (2002), pp. 35–54 (cit. on p. 33).

[78] Weaver T. "The π-helix translates structure into function". In: *Prot. Sci.* 9 (2000), pp. 201–206 (cit. on p. 13).

[79] Yap EH; Fawzi NL; Head-Gordon T. "A coarse-grained α-Carbon protein model with anisotropic hydrogen-bonding". In: *Proteins* 70 (2008), pp. 626–638 (cit. on p. 33).

[80] Maccari G; Spampinato GLB; Tozzini V. "SecStAnT: Secondary Structure Analysis Tool for data selection, statistics and models building". In: *Bioinformatics* (). Submitted (cit. on p. 39).

[81] Ramachandran GN; Sasisekharan V. "Conformation of polypeptides and proteins". In: *Adv. Protein Chem.* 23 (1968), 283–437 (cit. on pp. 8, 13).

[82] Tozzini V. "Coarse-Grained models for proteins". In: *Curr. Opin. Struct. Biol.* 15 (2005), pp. 144–150 (cit. on p. 24).

[83] Tozzini V. "Minimalist models for proteins: a comparative analysis". In: *Q. Rev. Biophys.* 43 (2010), pp. 333–371 (cit. on pp. 23, 27, 31, 44, 45).

[84] Tozzini V. "Multiscale modeling of proteins". In: *Acc. Chem. Res.* 43 (2009), pp. 220–230 (cit. on pp. 23, 24).

[85] Trovato F; Nifosi R; Di Fenza A; Tozzini V. "A minimalist model of proteins diffusion and interactions: the Green Fluorescent Protein within the cytoplasm". In: *Macromolecules* (). Submitted (cit. on pp. 64, 65, 69).

[86] Trovato F; Tozzini V. "Minimalist models for biopolymers: open problems, latest advances and perspectives". In: *AIP Conf. Proc.* 1456 (2012), pp. 187–200 (cit. on p. 24).

[87] Korkuta A; Hendrickson WA. "A force field for virtual atom molecular mechanics of proteins". In: *Proc. Natl. Acad. Sci. U.S.A.* 106 (2009), 15667–15672 (cit. on p. 32).

[88] Hoover WJ. "Canonical dynamics: equilibrium phase-space distributions". In: *Phys. Rev. A.* 31(3) (1985), 1695–1697 (cit. on p. 91).

[89] *WWPDB: World Wide Protein Data Bank*. 2013. URL: http://www.wwpdb.org/ (cit. on p. 86).