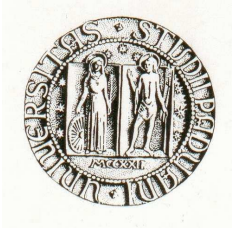
UNIVERSITÀ DEGLI STUDI DI PADOVA
Dipartimento di Fisica G. Galilei



XXI cycle – Ph. D. thesis

Solvent Induced Interactions in Biopolymers: Origin of Secondary Motifs

School director
Ch.mo Prof. Attilio Stella

Supervisors
Ch.mo Prof. Amos Maritan
Ch.mo Prof. Achille Giacometti
Dr. Antonio Trovato

Candidate
Chiara Poletto

Contents

Introduction

As the radius of knowledge expands,
the circumference of ignorance increases
(japanese saying)

The intricate interplay of the biochemical processes supporting life in cells represents a reservoir of fascinating puzzles for the modern physical research. Recently, biological phenomena of increasingly complexity are becoming subject of theoretical investigations, we can cite, for example, the operation of biomolecular machines [1] or the networks of molecule interactions responsible for gene regulation or signal transduction [2, 3]. Despite these advancements, however, a classical problem of this field, namely protein folding, represents so far a controversial matter.

Proteins are linear polymeric chains that adopt a unique three-dimensional structure, the native state, which allows them to carry out specific biological functions [4]. Given the suitable conditions, the folding in the native state is a spontaneous and reversible process for most of the small globular proteins. This result was obtained in the Anfinsen's experiments in the 1970s [5], and it provides the basis for a physical approach to the problem: protein folding, in the case of small globular proteins, is a thermodynamic process driven by the minimisation of the free energy of the system composed by the protein and by the aqueous solution in which the protein is immersed.

Somehow the information to fold is contained in the quantum-chemical Hamiltonian of the protein atoms interacting among themselves and with the surrounding water molecules. Given the huge number of degrees of freedom of the system, however, many efforts must be devoted to simplify the problem introducing intra-protein and protein-solvent effective interactions in order to describe the system at a mesoscopic scale [6]. The folding transition is a cooperative phenomenon, involving many weak interactions, which suggests that no single degree of freedom on its own might have any special importance for folding. This collective character indicates that a coarse grained approach, oriented to an identification of the average properties of the protein behaviour, may be appropriate to reach a more comprehensive view of the process.

At the mesoscopic scales described by semi-empirical classical interactions, however, the protein molecule still appears a complicated system. Many conflicting constraints, both geometric and energetic, have to be satisfied when the protein folds into its compact conformation: the different chemical characters of the side chains, and their different affinities with water solution, the steric hindrances of atoms with different Van der Waals radii and the propensity to form peptide hydrogen bonds. In this intricate scenario, it is quite astonishing that, instead of presenting the frustrate behaviour typical of disordered systems [7], the protein molecule reaches its unique equilibrium state on the time scale of the order of $10^{-1}s$ or even $10^{-3}s$ for the small globular proteins [8].

The spontaneity of the process, which follows from the thermodynamic hypothesis, suggests that the fundamental physics underlying it may be simpler than the molecule complexity would lead us to expect. The emerging simplicity of the protein behaviour is reflected also in the regularity of the native state conformations. The protein three-dimensional structures consist of a compact arrangement of α-helices and β-sheets, the so called secondary strucutres characterised by highly ordered geometries. These protein building-blocks transcend the chemical details of the sequence, being the product of factors common to all the proteins: the peptide hydrogen [9] bond and the steric hindrance [10]. Even the overall conformation of native states is not so strictly bound to the amino acid sequence specificity: the correspondence between sequences and folds is indeed many-to-one [11]. For small single-domain proteins, a

limited menu of native state topologies exists [12]; their conformations were preserved by evolution, which seems to prove that they satisfy an optimality principle [13].

In an attempt to explain these issues the principle of minimal frustrations was invoked [14]. It was hypothesised that the amino acid sequences were selected by evolution, among all the possible combination of the twenty types of amino acids, in order to allow a fast and reproducible folding. Each native fold is the product of its own sequence, well fitting all the sequence physicochemical features. This idea is able to justify the net difference between proteins and random heteropolymers which are characterised by a glassy behaviour. It stresses, nevertheless, the central role of the amino acid sequence in determining the protein conformation, and the peculiarity of each protein with respect to others. Therefore this view is unable to explain the origin of the general structural and thermodynamic features characterising the whole class of globular proteins.

More recently a different perspective has emerged [15], many experimental facts and theoretical considerations have shifted the focus from the amino acid sequence to the topology of the protein conformations. Not only large mutations in the sequence can leave the native state topology unchanged; but also many aspects of the folding kinetics, such as the folding rate and the distribution of structures in the transition states ensemble, result to be strongly correlated with the native state topology and, at the same time, quite insensitive to large-scale sequence changes, provided that the latter do not prevent the structure stability.

In this context the tube-like model was proposed few years ago [16, 17]. Despite its extreme simplicity it has been able to give important insights regarding the connection between the complex heterogeneity of the chemical details and the universal character of many protein aspects. The concept of thickness of a biopolymer has revealed itself to be the correct ingredient in order to capture the intrinsic anisotropy of the protein chain enhanced by the steric hindrance of the backbone atoms. The early analysis on the model showed that a thick homopolymer, self-interacting with a mere square-wall attraction is capable to reproduce the basic motifs of the protein conformations, the α-helices

and the β-sheets [18, 19]. A refined version of the model presents metastable state conformations, consisting in protein-like arrangements of secondary motifs; this results were reached with the only addition of an interaction mimicking the formation of peptide hydrogen bonds, without involving, also in this case, any sequence specificity.

The main analysis and results on the tube-like model are reviewed in this thesis. The ideas which have stimulated this research line and the subsequent developments will be extensively discussed in this work. An innovative scenario is suggested by these results: protein native conformations may be selected by their ability of optimising the energetic and geometric constraints shared by all proteins. Common requirements, such as steric hindrance of the backbone atoms, optimisation of peptide hydrogen bond formation, predetermine a collection of few conformations, among which each sequence choose its proper native state [20]. This could explain the relative small number of native state topologies (compared with the number of compact featureless conformations), the fact that the latter are formed by few recurrent motifs and the compatibility of each topology with many different sequences.

The tube-like model has demonstrated the ability to capture the general geometric and energetic features of proteins, and it has encouraged further developments. A factor strongly affecting protein behaviour was missing in the early versions of the model: the explicit interaction between the protein molecule and the solvent. The most recent works were thus devoted to fill this gap [21, 22], and they represent the main subject of this thesis.

The importance of considering explicit solvent for the folding of globular proteins is crucial. As a consequence of the dipolar nature of the water molecule the intra-molecule interactions result greatly diminished with respect to the case in which the solvent is not present. Moreover a crucial push towards a compact conformation arises just because of the interaction between the aqueous solution and the protein residues: many of them try to hide from the water owing to their non-polar character. This results in a global phenomenon, namely the hydrophobic effect, that was recognised since many decades as one of the most important driving forces of folding [23, 24]. In order to address the average behaviour determined by this complex process,

an effective interaction was introduced in the tube-polymer [21] that is simply proportional to the area of the tube regions hidden from the solvent particles.

Beside this effect, another physical process has been recently individuated as having a central role in the folding process [25, 26]. The tendency of solvent molecules to maximise their translational entropy gives rise to the so called depletion forces, which are responsible in colloidal systems of entropically driven phase transitions [27, 28], as revealed by experiments conducted early in the last century. The mechanism is very simple: in a situation in which macromolecules are immersed in a solvent of smaller particles, the former are induced to diminish their volume occupation in order to increase the free volume available to the latter, and hence their entropy; this drive the macromolecules to aggregate determining a phenomenon called flocculation. An elegant explanation of the process was given by the fundamental contribution by Asakura and Oosawa [29, 30] that still represents the paradigm for the theoretical investigations on this phenomenon. The depletion forces are omnipresent in living systems, and recently their relevance for biological processes was recognised in many works. In particular it was suggested that this forces could lead the protein to collapse in a compact configuration, and even that secondary structure motifs could be selected since they have the ability of minimising the volume occupation of the polymer. In order to explore this possibility we have adapted the Asakura-Oosawa paradigm to the case of a tube-like polymer in a solution of hard spheres [22]. The results of this theoretical and computational analysis will be exhaustively illustrated.

Plan of the thesis:

The plan of this thesis is as follows: in chapter 1 we introduce the main properties of globular proteins. Particular regard is then given to the most crucial advances in the understanding of protein folding, in the past half century. A review of the tube-like polymer model is the subject of chapter 2: the main studies of the model conducted in the last years are illustrated, from the starting ideas to the more sophisticate developments. In particular, the last part of

the chapter is devoted to the treatment of the effective interaction introduced for the tube-polymer in order to model the hydrophobic effect. Chapter 3 is entirely addressed to the depletion interaction. After an introduction of the Asakura-Oosawa paradigm, we discuss in detail the theoretical analysis developed in order to adapt it to the tube-polymer case. We conclude the chapter discussing the numerical simulation done in order to find the conformations minimising the effective potential so introduced.

Finally in the conclusions of this thesis we discuss the future perspectives of these initial theoretical works on the protein-solvent interaction in the context of the tube-like polymer. Investigation in this directions are conducted by other research groups by means of a morphometric approach [31, 32, 33, 34]. The method has demonstrated a high efficacy in extracting quantitative informations from the shape of complex patterns, and therefore is suitable to analyse the connection between the solvation free energy of a protein immersed in solution and its three-dimensional conformation. A possible merging between this method and the tube-like polymer is considered.

Chapter 1

A physical approach to protein folding

The ability of cellular subunits to accomplish highly specific functions determine the astonishing efficiency of the biological machinery. The cellular environment is crowded of macromolecules completely different in shape, size and chemical character; such a diversity enable them to interact with each other in a very selected manner, despite the extremely high concentration.

Most of the macromolecules operating inside the organisms are proteins: a common chemical nature underlies their heterogeneity [4, 35]. All of them consist of long polymers composed by twenty different types of monomer units called *amino acids*, but they differ each other by the three-dimensional conformation in which they fold after the biosynthesis and in which they are biologically active. According to the sequence of amino acids proper of each protein, this conformation varies and hence the functions the protein is able to perform.

The *protein folding* represents a process of a fundamental importance: if the protein fails to reach its proper conformation it becomes inevitably unable to carry out its biological functions; many serious diseases indeed are caused by proteins misfolding. The remarkable rapidity and reproducibility of this process still present many elements which are not yet completely understood. It is commonly recognised that the folding of the protein chain is a *thermodynamically driven* process [5], and that the biologically active protein conformation represents the energetic balance of various kinds of interactions between pro-

tein groups, and between these groups and the surrounding medium. If on the one hand this justifies a physical approach on the other one the high complexity of the system seems to prevent any possibility of solution to the problem, from a microscopical point of view.

Despite the high complexity of the protein molecule and the importance of the chemical details, proteins share common aspects [9, 10], that result very important in determining their three-dimensional structure. Regular motifs, helices and sheets, form the building-blocks of all the protein three-dimensional structures: such recurrent motifs, called secondary structures, are the product of factors independent of amino acid specificity. It suggests that a deep comprehension of the matter can be reached only distinguishing the property peculiar of each protein and responsible of its functional specificity from the features proper of the overall class of protein molecules. A coarse grained view proper of statistical mechanics can be very useful for this purposes: very simplified model can be able to enlighten the main mechanism driving the protein behaviour.

In this chapter we want to give a rapid introduction to proteins, briefly describing their biological role, their chemical composition and their conformational properties. We are mainly interested on the general characteristics that make the proteins a roughly homogeneous class of molecules. In the last section we will face the protein folding problem, dealing with both the known facts and the open issues related to it. Finally we will give a rapid introduction to the theoretical approaches to the matter, analysing the role that a simplified model can have in this framework.

1.1 Introduction to proteins

Proteins are long heteropolymers able to fold into three-dimensional conformations, called *native state*, properly designed by evolution to perform highly specific biological functions [4].

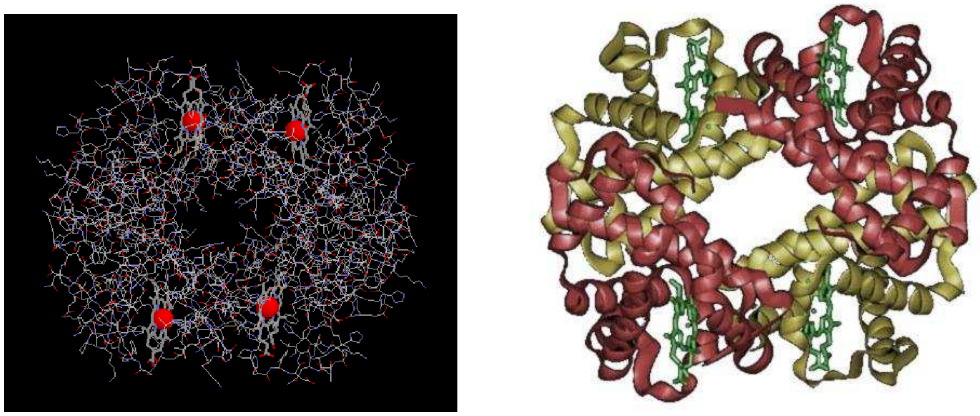
During the protein biosynthesis the amino acids are assembled together in a specific sequence according to the rule of the genetic code. The final product

of this process, the polypeptide chain, immersed in the cellular environment, undergoes a folding process reaching the protein proper native state. The functions each protein is able to carry out are determined by the three-dimensional conformation of the native state. For this reason protein biosynthesis and protein folding represent the processes by which the information encoded in DNA reach its full expression and hence their deep understanding represents an important challenge of the modern research.

In order to face the complexity of the protein macromolecule several levels of organisation are distinguished [4, 35]. The *primary structure* is the chemical sequence of amino acids along the polypeptide chain. The *secondary structures* represent the local ordered motifs occurring in most of proteins, the compact packing of which determines the unique three-dimensional native conformation, the *tertiary structure*. Finally, large proteins are composed by *domains*, small globular regions separated by a few amino acids. The arrangement of one respect to another is called the *quaternary structure*. In the following sections we will illustrate each of these levels in detail.

Even though proteins are a relatively homogeneous class of molecules, their native state can be completely different as regards shape and size and are able to carry out an enormous variety of biological functions. Most natural proteins are soluble in the physiological, almost aqueous, solution; they assume roughly spherical shapes, and thus are usually referred to as **globular proteins**. Example of globular proteins are enzymes, proteins for transmitting signal and regulating biological processes (i.e. hormones, insulin), immunoglobulin and hemoglobin (in Fig. 1.1). **Fibrous proteins** otherwise are insoluble and form aggregates with regular and extended conformations. They play structural roles, such as collagen (in Fig. 1.6), the best example, that is the main constituent of higher animal frameworks: bones, tendons, skins, cartilage, and membrane supporting tissues. A third class is represented by **membrane proteins**, associated to the impermeable lipidic bilayers forming the cellular membranes: the hydrophobic environment of membranes surrounding them, determine their strong difference respect to other classes of proteins. Finally a peculiar class of proteins was recently discovered [36], the **intrinsically unstructured proteins**, that are characterised by their lack of stable tertiary

structure as isolated subunits. The flexibility, that derives by their disordered character, enhances the ability to bind to their molecular targets. Many of them undergo transitions to more ordered states upon binding to their targets.



(a) (b)

Figure 1.1: Structure of hemoglobin, the protein responsible of the oxygen transport in the red blood cells, as resolved by X-ray diffraction data. In Fig. a the all atom representation is shown. In Fig. b we report a cartoon representation pointing out the regular secondary motifs compounding the native conformation. Though more schematic, the second view give us more informations about the protein structure. Hemoglobin is a globular protein with a diameter of approximately 6 nm, it is composed by 4 chains (two chains with 141 amino acids, and two with 146) folded in distinct domains and assembled together in a determined quaternary structure. Each domain is associated with an *heme* group (visible in both the pictures) responsible of the oxygen binding.

The three-dimensional structure of globular proteins can be resolved experimentally using NMR spectroscopy, taking advantage of their solubility, or X-ray diffraction. The other two classes of proteins are otherwise very hard to be characterised experimentally; many aspects related to their three-dimensional structures and their folding process are hence rather obscure. For this reason globular proteins represent the paradigm for the biophysical studies about protein conformation and folding kinetics, and they represented the point of reference for our studies.

The folding of globular proteins occurs in the cytoplasm, essentially an aqueous environment. It is a fundamental result of Anson and Anfinsen [5, 37, 38] the *reversibility* of protein unfolding: they proved in vitro that, after

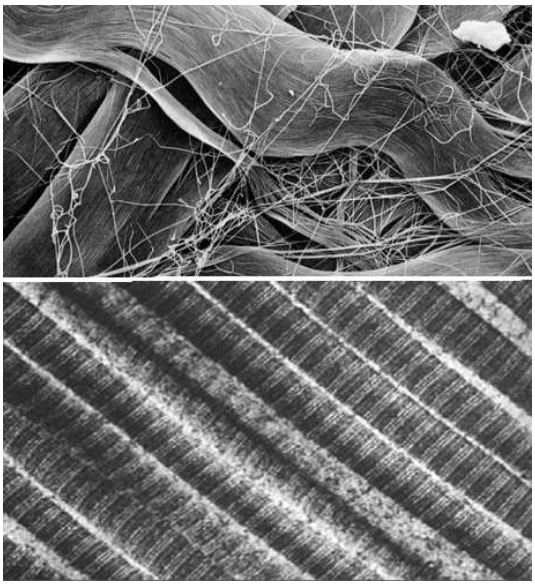


Figure 1.2: At top, collagen fibers in scanning electron microscope image are shown. At bottom a much higher-magnification image from transmission electronic microscope shows the characteristic banding pattern of individual collagen fibril that make up the larger anatomic fiber. A collagen molecule, the fibril subunit, is approximately 300 nm long and 1.5 nm in diameter and it is made up of three polypeptide strands, each possessing the conformation of a left-handed helix.

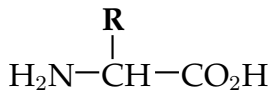
inducing an unfolding by altering the physiological conditions (pH, temperature, pressure), the protein is able to re-fold in its correct native state once physiological conditions are restored. This fact shows that the protein folding is a *thermodynamically driven* process with the unique native structure of each protein representing the thermodynamic *ground state* of the system composed by the protein molecule and solvent molecules, taking into account the various types of interactions affecting them. This behaviour is typical of small proteins. The large multidomain ones represent an exception since they need the aid of *molecular chaperones* in order to fold correctly 'in vivo' and avoid improper aggregation of the 'wrong' domains [39].

1.2 Chemistry of proteins

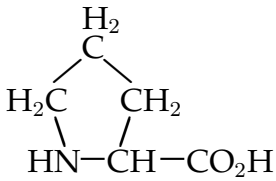
In this section and in the following ones we are going to describe the protein macromolecule according to all its levels of organisation. We start illustrating the chemical properties of the amino acid subunits and how they are linked together in the polypeptide chain. The macromolecule chemistry determines the physical interactions affecting the system and leading the polypeptide chain to adopt its compact conformation in aqueous solvent.

1.2.1 Aminoacids: the building-blocks of proteins

By a chemical point of view proteins are really complex macromolecules since they are composed by 20 different types of monomers, the amino acids. The generic chemical structure of amino acids is



where H_2N is the amino group and CO_2H is the acidic group. The twenty amino acids differ only in the chemical structure of the *side chain* **R**, that is bonded to the central carbon atom, the α*-carbon*. The only exception is represented by *proline*, whose side chain is bonded also to the nitrogen atom to form an imino group:

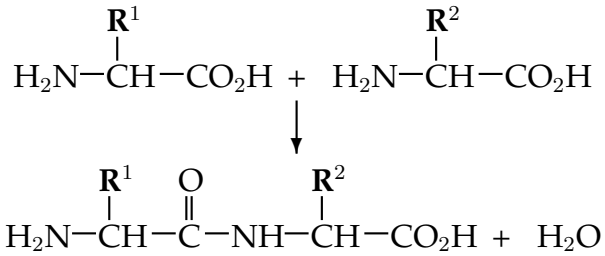


The chemical composition of side chains varies in a considerable way. Glycin, the lightest side chain, consists of only a hydrogen atom, whereas tryptophan, the heaviest, contains both a carbon aromatic ring and an indole ring, with one nitrogen atom. Some side chains moreover may be partially charged at physiological Ph. The atoms occurring more frequently in the side chains are hydrogen, oxygen, nitrogen, and carbon, but a sulfur atom is also present in

two side chains, methionine and cysteine. Except for glycin, the central α-carbon atom is asymmetric. In all known natural proteins, the α-carbons have the same chirality, being all left-handed.

1.2.2 Peptide bond and polypeptide chain

The chemical binding of amino acids produces the *peptide bond*, with release of a water molecule as illustrated below.



By the effect of this process the polypeptide chain is formed: it has the shape of a long chain, the *backbone*, composed by the repetition of the three atoms N, C^{α}, C', with the side chains branching outward. It is characterised by a number of amino acid residues (the remaining parts of amino acids after peptide bond formation) ranging from approximately 50, for small globular proteins, to 3000, for complex multi-domain proteins. The residue sequence along the chain, the primary structure, determines the chemical character of each protein.

The peptide bond is determined by covalent interactions, the same interactions that hold the atoms of a single amino acid together. Typical energies of the covalent interactions range from 50 kCal/mole to 150 kCal/mole, implying that at room temperature ($K_B T \simeq 0.6$ kCal/mole) the corresponding degrees of freedom are frozen at their minimum energy value. In particular the length of covalent bond and the valence angles between two covalent bonds are very rigid, and their values are fixed by the laws of chemistry. They are called 'hard' degrees of freedom in contrast with the 'soft' degrees of freedom represented by the torsion angles along the backbone chain and the side chains that can easily fluctuate at room temperature. The torsion angles, or dihedral angles, are the rotation angles about covalent bonds of one portion of the chain with respect to the other, and are responsible for the conformational flexibility of

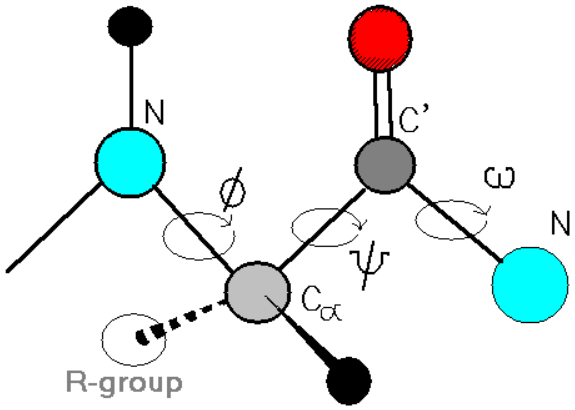
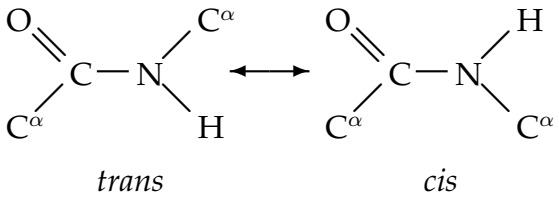


Figure 1.3: Schematic representation of the polypeptide chain, in which the three torsional angles ϕ, ψ, and ω are shown.

both the polypeptide backbone and the side chains. Rotation about the N–C^{α} bond of the peptide backbone is denoted by ϕ, rotation about the C^{α}–C' bond by ψ, and that about the peptide bond C'–N by ω, the three angles are shown in Fig. 1.3. Specification of ϕ, ψ, and ω, for all residues of the polypeptide chain completely determines the backbone conformation.

The peptide bond have partial double-bonded character, due to resonance between a single-bonded and a double-bonded isomer. Rotation about the peptide bond is thus restricted, and the atoms of the polypeptide backbone between two successive α-carbons have a strong tendency to be coplanar, as shown below.



trans *cis*

Two planar conformations are possible for the peptide bond, but the *trans* form is highly energetically favoured, because in the *cis* form the side chains of neighbouring residues are in too close proximity. In the *trans* conformation, the distance between corresponding C_{α} atoms of adjacent residues is fixed to

the value of 3.80 Å. When the residue following the peptide bond is proline, the double-bond character is lost, and there are small deviations from planarity of either the *trans* and the *cis* form, which in turn are almost equally energetically favoured.

1.2.3 Interactions in proteins

The complex three-dimensional conformation of the polypeptide chain, the tertiary structure, is the result of the delicate interplay of interactions between atoms which are far apart along the polypeptide chain but may come into close spatial contact, and between atoms and the solvent molecules. At a microscopic level all the physical interactions affecting the system are Coulomb electrostatic and can be described by the quantum-mechanical laws. The introduction of semi-empirical classical interactions nevertheless is necessary both to reduce calculation costs and to have a clearer view of the different mechanisms driving the protein behaviour.

The analysis of the different kind of mesoscopic interactions affecting the system is complicated by the strong diversity of amino acids chemical character. As we have seen in previous section the chemical structure of proteins is the product of covalent interactions, that are characterised by high energy scales compared with the thermal fluctuations at room temperature. With the exception of the *sulfur bridges*, forming between the sulfur atoms of the cysteine residues, all the interactions between amino acids far apart along the polypeptide chain are non-covalent. The typical energy scale of non-covalent interactions ranges from 1 to 5 kCal/mole. The associated degrees of freedom are thermally excited at room temperature, and are thus responsible for the folding and all the observed thermodynamic properties of proteins. Non-covalent interactions between different atoms of the protein-solvent system are usually divided into electrostatic forces, Van der Waals interaction, and hydrogen bond interactions.

Electrostatic interactions between ionised or partially charged atoms may be described according to the Coulomb's law for point charges in a homogeneous dielectric medium. A molecule need not to have a net charge to par-

ticipate in electrostatic interactions, since atoms with different electronegativities cause the electron density to be localised, and are thus assigned a partial charge. A correct description at the microscopic level should take into account that charges are not point-like at short distances, and the solvent is an inhomogeneous dielectric medium at the molecular level.

Van der Waals interactions between pairs of atoms are often represented by an energy potential as a function of their distance r, which is usually taken in the Lennard-Jones form:

$$E(r) = \frac{C_n}{r^n} - \frac{C_6}{r^6} \qquad (n > 6) \,, \tag{1.1}$$

where C_n and C_6 are constant. The most common potential has $n = 12$, since $12 = 2 \cdot 6$ makes it computationally efficient. The repulsive energy in the first term of Eq. 1.1, due to overlapping electronic orbitals, rises so steeply, that it is common practice to model individual atoms as hard impenetrable spheres, characterised by the *Van der Waals radius*. The excluded volume of the hard spheres of all the polypeptide chain atoms determines the *steric constraints* that play a crucial role in selecting the protein possible conformations. Typical values of Van der Waals radii are 1.55 Å for nitrogen, and 1.75 Å for carbon, the two species occurring in the polypeptide backbone. If the two atoms are covalently bonded, the inter-atomic distance is shorter, since covalent bonding implies sharing of electron orbitals. The Van der Waals radius of a given atom depends also on the way the atom is covalently bonded to other atoms. The attractive part of the potential, determined by the transient induced polarisation effect, is short-range and weak. The optimal distance for the interaction of two atoms, corresponding to the minimum of the potential 1.1, is usually 0.3-0.5 Å greater than the sum of their Van der Waals radii.

The **hydrogen bond** plays an important role in protein structural stability, since is responsible for the energetic stabilisation of secondary structure patterns [9, 50] that will be described in detail in the next section. It occurs when two electronegative atoms compete for the same hydrogen atom. The hydrogen atom is formally bonded covalently to one of the atoms, the donor, but it also interacts favourably with the other, the acceptor. In the most common configuration the three bonded atoms are collinear. Responsible of secondary

structure stabilisation are the *peptide hydrogen bonds*, involving the carboxilic C=O and amino N–H groups of the polypeptide backbone; the H $\cdots$ O distance in this case is most often 1.9-2.0 Å. The interaction responsible for the formation of hydrogen bonds can be introduced explicitly by means of different potentials mimicking how the strength of the interaction varies with departures from linearity. It is now quite accepted, however, that hydrogen bonds are just a result of the combination of Coulomb and dipolar Van der Waals interactions.

It is important to remember that the interactions listed above are strongly affected by the protein surrounding. In the case of globular protein, folding and biological activities occur in aqueous environment and the water solution has a great influence in these processes. Water is a dipolar molecule, and thus has strong interactions with charged or dipolar groups, or hydrogen bond acceptors and donors, so that the forces occurring among such groups in vacuum are greatly diminished. Moreover the dipolar character of water molecule is responsible for the **hydrophobic effect**, an effective interaction considered one of the most important driving force leading to the folding of polypeptide chain in a compact conformation. The interactions with water are not so favourable for non-polar side chains of some amino acids, basically because they can not participate in the hydrogen bonding, which appears to be very important in liquid water [40]. As a consequence they tend to bury themselves inside the globule, in order to get shielded from contact with water molecules. Owing to their behaviour these amino acids are called hydrophobic in contrast with the hydrophilic ones, characterised by polar side chains. This effect does not determine a strict rule: by looking at structures of real proteins, there is a substantial probability, $\sim 35\%$, to find a hydrophobic residue on the surface of a protein, or to find a hydrophilic one buried inside [41]. It rather represents the mechanism determining a overall protein collapse. Such important interaction is the main ingredient of simplified models [41, 42, 43, 44, 45, 46], and will be the subject of section 2.3 where a recent contribute by our group to this topic will be discussed.

1.3 Regularity of the protein conformation

In response to the high variety characterising the protein chemistry the three dimensional conformation of the native state is quite complex and diversified. In comparison with the simple double-helix of DNA molecule, the myoglobin structure, the first protein native structure experimentally observed, appeared to its discoverer completely lacking in order at the first time. Nevertheless a subsequent and more careful analysis pointed out the presence of recurrent conformations characterised by regular motifs such as helices and sheets. They are called secondary structures and represent the basic constituents of all the native conformations. Each of them is characterised by a definite geometry which we are going to describe in the following. At the moment we want to point out the general character of these structures: they represent the product of factors that are independent by the amino acid sequence specificity.

Two historical works explained secondary structure origin starting from different basis. Pauling and Corey [9, 47, 48, 49, 50] predicted the geometry of helices and sheets, studying the conformations optimising peptide hydrogen bond formation. The independent work of Ramachadran [10], on the other hand, analysed the steric constraints due to the backbone atoms, obtaining the so called *Ramachandran Plot*.

We want to briefly introduce the Ramachandran idea. Remembering section 1.2.2, it seems natural to consider, in a simplified description of the polypeptide chain, all the hard degrees of freedom, corresponding to covalent interactions, as frozen, and take into account only the soft degrees of freedom, that are ω, ϕ and ψ, corresponding to non-covalent interactions. A part from proline exception, ω is fixed, whereas ϕ and ψ, that may in principle assume all possible values, are limited by the steric constraints. Ramachandran studied the values of (ϕ, ψ) consistent with such constraints. Being ω fixed, the chain's backbone is essentially made of rigid planar peptide units. This fact suggested him to study the mathematical problem of rotation of two rigid planes containing interacting hard spheres that must avoid bumping against each other. The results are represented in the Ramachadran Plot (shown in Fig. 1.4) and are in agreement with the experimental observations. It is remarkable that with the

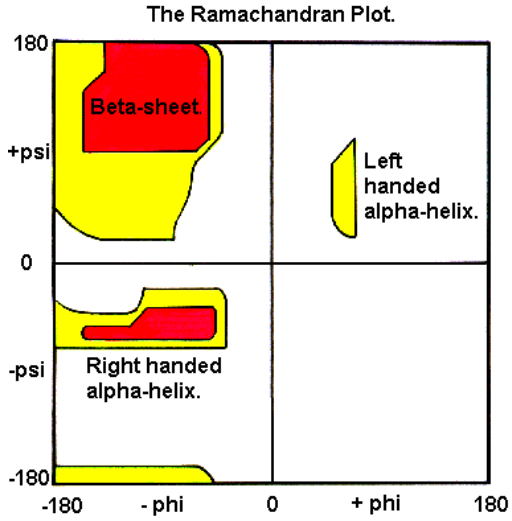


Figure 1.4: Ramachandran plot showing allowed values of the torsion angles ϕ and ψ for all the residues with the exception of proline (region in red). Additional conformations are accessible to glycine (region in yellow) because it has a very small side chain. The typical values of the torsion angles corresponding to the different secondary structures are shown.The values corresponding to the collagen triple helix are not indicated in the figure, they belong to the red region on the top left of the diagram.

exception of glycine, which can occupy a pretty large region of the conformational space (ϕ, ψ) because of its smallest side chain (only a hydrogen atom), all the other amino acids behaves more or less in the same manner.

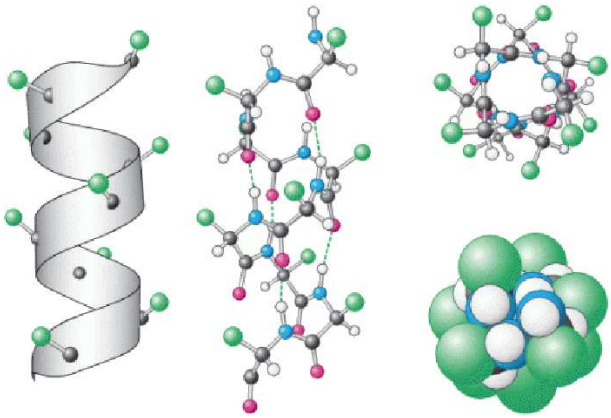
The secondary motifs, helices and sheets, leave their own specific signature in the Ramachandran plot. Each motif is characterised by ordered local geometry which results in repeating values of the torsion angles. In figure 1.4, are indicated the different regions of the Ramachandran plot corresponding to the types of secondary structures that we are going to describe in the remainder of this section: collagen triple helix, right-handed α-helices and parallel and antiparallel β-sheets. Left-handed α-helices are observed only rarely, since the side chains are very close to the backbone.

1.3.1 The secondary structures

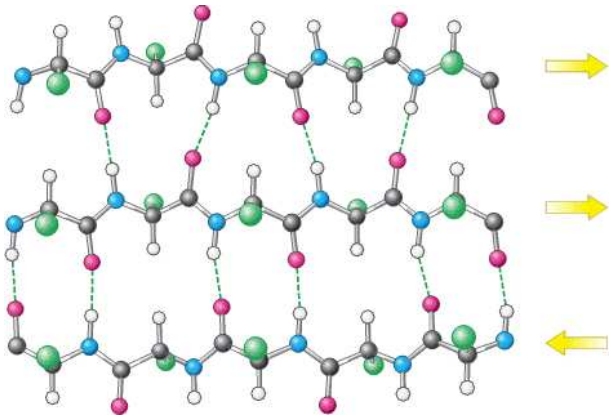
The right-handed α-**helix** is the best known and most easily recognised of the secondary structures (see figure 1.5a). The backbone carbonyl oxygen of each residue forms a nearly straight hydrogen-bond with the backbone amino group of the fourth subsequent residues along the chain. This results in the α-helix having 3.6 residues per turn. The side chains project outward and do not interfere with the helical backbone. Among secondary structures, conformations similar to α-helix are found: they include the 3_{10}-helix ($i+3 \rightarrow i$ hydrogen bonding) and the π-helix ($i + 5 \rightarrow i$ hydrogen bonding). These alternative helices are relatively rare.

The second most regular and identifiable secondary structure is the β-**sheet** (see figure 1.5b). The basic unit is the *β-strand*, a planar zig-zag conformation with the side chains alternatively projected in opposite directions. A single β-strand is not stable, because no interactions are present among intra-strand atoms. The β-strand conformation is stabilised only when two or more strands are assembled into a β-sheet, a planar structure where hydrogen bonds are formed between the peptide groups of adjacent β-strands. The side chains are perpendicular to the plane of the sheet. While side chains from adjacent residues of the same strand do not interact with each other, side chains from neighbouring residues of adjacent strands are projected into the same side, and thus interact significantly. Adjacent β-strands can be either *parallel* or *antiparallel*, and the resulting geometry varies slightly. In antiparallel arrangement, all hydrogen bonds are parallel to each other, whereas in parallel arrangement they are oriented in two different alternating directions.

As discussed in the first section fibrous protein conformations are not well characterised experimentally. Their regular, extended conformations represent a level of complexity somewhat intermediate between pure secondary structures and the tertiary structures of globular proteins. One of the best understood related conformation is **collagen triple helix** (see figure 1.6). There exist only a few distinct types of collagen polypeptide chains. They are all characterised by the repetition of glycin every three residues of their amino acid sequence. Three different chains are coiled together, each with a slightly



(a)



(b)

Figure 1.5: Schematic representation of secondary structures. Fig. a: Different α-helix views. The α-helix has 3.6 residues per turn, corresponding to a pitch of approximately 5.4 Å. Its length can vary from $4-5$ to ~ 40 residues. The left of the figure represent the idealised diagram of the backbone path, showing the side chains pointing outward. In the middle a schematic diagram with the positions of all backbone atoms is reported. Dashed lines represent hydrogen bonds. In the right of the figure two different top views. Fig. b: representation of a *mixed* β-sheet, with the central strand aligned parallel with the upper one and antiparallel with the lower one, as pointed out by the yellow arrows. The dashed lines show the hydrogen bonds, having two different patterns between parallel and antiparallel arrangements. A strand length typically varies from 5 to 10 residues.

twisted, left-handed helical conformation. The three helices are wound around each other to form a right-handed super-helix, and are stabilised by hydrogen bonds that form between peptide groups of different chains according to a regular pattern. The geometry of collagen helix, apart the slight twisting, is differ-

ent from the α-helix. It has opposite handedness, different torsion angles, and 3.0 residues per turn.

1.3.2 Tertiary and quaternary structure

The tertiary structure represents the overall three-dimensional conformation of the polypeptide chain. It is formed by the assembly of secondary structures connected together by loop regions, segments of the polypeptide chain with irregular shape and varying length. The resulting packing is usually much tighter in the interior of the protein producing a solid-like core and a fluid-like surface [103].

Many large proteins consist of several structural domains and hence they are called *multidomain* proteins. The way different domains are arranged within the same multidomain protein defines the quaternary structure. A domain is a stable and independent unit in the protein structure, that in many cases folds separately. Its length can not exceed 500 residues with an average of approximately 100 residues [52]. In multidomain proteins, each domain may fulfil its own function independently, or in a concerted manner with its neighbours.

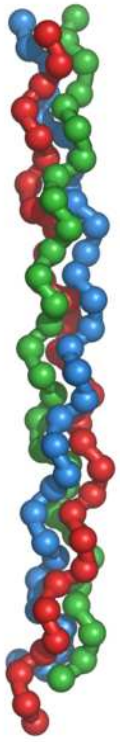


Figure 1.6: Schematic representation of the collagen triple helix.

The subdivision in domains presents many advantages. In comparison with a unique large domain protein, the structure is more stable and the folding times are accelerated, each domain being able to individually fold. The domain formation appears to be the optimal solution for a large protein to bury its hydrophobic residues while keeping the hydrophilic residues at the surface [53]. Finally, the subdivision in domain increases the protein flexibility [54].

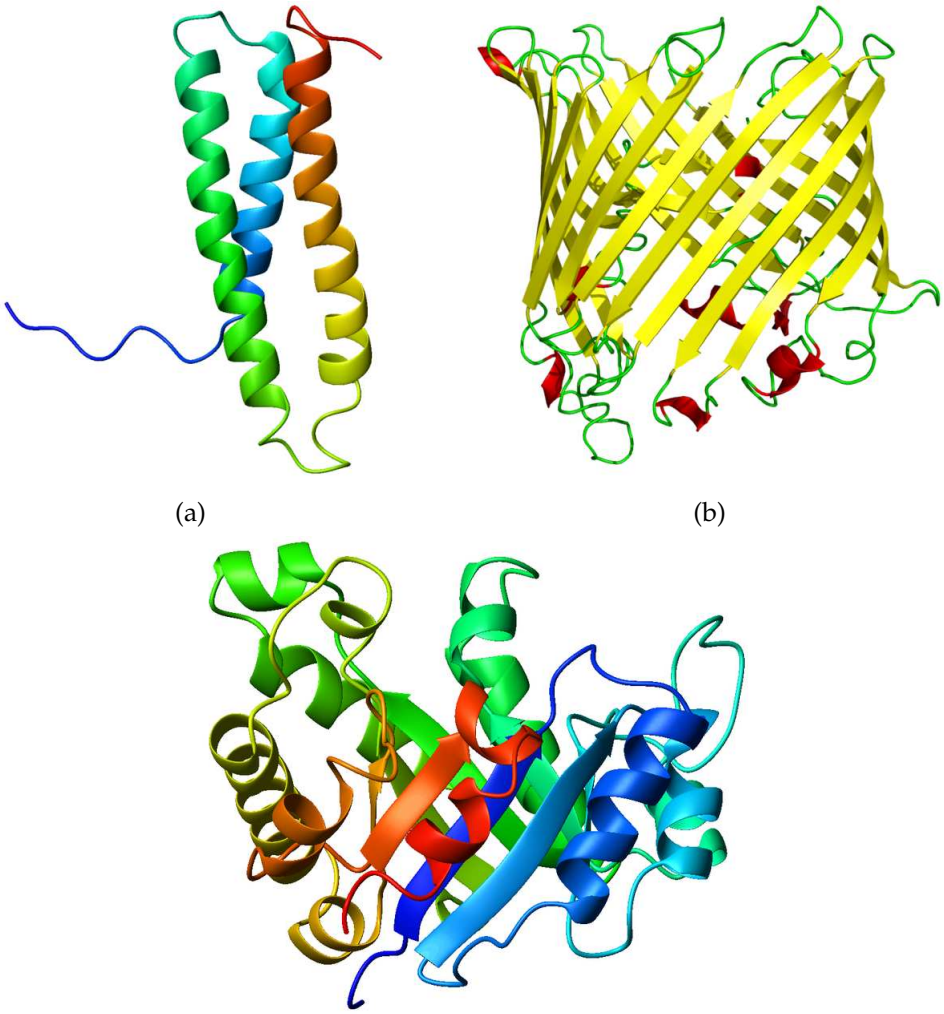
Two domains are said to have a common fold if they have the same major secondary structures in the same arrangement with the same topological connections [55]. One fold may appear in a variety of different proteins [11]. For instance, many enzyme catalysing completely unrelated reactions, have

the same α/β-barrel conformation (see Fig. 1.7c). Indeed it seems that the evolutionary mechanism has created more and more complex protein structures simply by assembling the pre-existent conformations: in fact many domain folds in eukaryotic multidomain proteins can be found as independent proteins in prokaryotes.

The Structural Classification of Proteins (SCOP) database [55, 56] is a largely manual classification of protein structural domains based on similarities of their three-dimensional structures. According to SCOP classification, domains are usually divided in four many categories based on the secondary structural content. The all-α domains (Fig. 1.7a) have a core built exclusively from α-helices. The all-β domains (Fig. 1.7b) are formed of antiparallel β-sheets, usually two sheets packed against each other. The α/β domains (the class including the α/β-barrel shown in Fig. 1.7c) are made from a combination of $\beta-\alpha-\beta$ motifs that predominantly form a parallel β-sheet surrounded by α-helices. Finally the $\alpha+\beta$ domains are a mixture of all-α and all-β motifs. A different data base of domain classification, CATH Protein Structure Classification [57], is a semi-automatic classification and shares many broad features with SCOP.

Molecular evolution gives rise to families of related proteins with similar sequences and structures. However, sequence similarities can be extremely low between proteins that share the same structure. There are currently about 45000 experimentally determined protein structures deposited within the Protein Data Bank (PDB) [58]. However this set contains a lot of identical or very similar structures [12]. Some fold families are highly populated and are referred to as *super-folds* [59]. The most populated is the α/β-barrel domain.

Some proteins are composed by different polypeptide chains that fold separately in their three-dimensional structure and then aggregate with each other forming an agglomerate; this is another example of protein quaternary structure. An example of quaternary structure is represented by hemoglobin shown in Fig. 1.1.



(a) (b)

(c)

Figure 1.7: Three examples of very populated domains each of them respectively belonging to the first three categories according to the SCOP classification. Fig a: a helix bundle domain, belonging to the all-α class. Fig. b: a β-barrel, example of all-β domains. Fig c: a α/β-barrel, belonging to the α/β category. In order to point out the conformation regularity, the secondary structure are represented by ribbons.

1.4 The challenge of protein folding

The unique three-dimensional structure in which the protein is biologically active is the result of the folding process occurring after (or in some cases during) the biosynthesis. Although many works in the last decades have enlightened important aspects of this process, it remains not completely understood

at present. However, its deep comprehension is of fundamental importance, not only because it would be an actual step-forward in our basic knowledge, but also for its practical utility. Each protein carries out a menu of functions according to its native state conformation, hence the knowledge of the latter is very useful to understand the role of the protein inside the cell. The experimental determination of the three dimensional structure is often very difficult and expensive; on the other hand, the amino acid sequence can be determined in a quite simpler manner, and the number of proteins whose sequence is known is extremely large, $\sim 10^7$ [60]. For this reason many efforts are devoted to the development of techniques which predict the three-dimensional conformation based solely on the knowledge of the primary structure.

In order to reproduce the folding process and to predict the native state conformation, methods borrowed from distinct areas of research, are employed, according to the actual goal. To predict the three-dimensional conformation associated to a known sequence, bio-informatics techniques are the most effective: a powerful method is *sequence alignment*, a way of arranging the primary sequences of different proteins (usually the unannotated protein taken together with annotated ones) to identify regions of similarity that may be a sign of functional, structural, or evolutionary relationships. Despite the efficacy of these semi-automatic methods, however, other studies of practical interest, such as prediction of folding routes (i.e. intermediate structures along folding) and identification of key residues (i.e. residues involved in binding), require a deep comprehension of the folding mechanism. The physical approach can be very powerful for these theoretical investigations.

In this section we want to introduce same known facts and some open issues regarding the folding process. We want also to give a brief overview of the physical approaches to the matter.

1.4.1 Thermodynamics of the protein folding transition

The amino acid sequence of each protein contains the information that specifies the native structure. As discussed in section 1.1, the Anfinsen experiments showed that small proteins are able to correctly fold spontaneously in vitro;

larger multidomain ones need otherwise the action of other proteins called chaperones. However what happens in vivo is not so clear. The process of folding often begins co-translationally, so that one extremity of the protein chain begins to fold while the other portion of the protein is still being synthesised by the ribosome. Moreover the cellular environment is very crowded with other proteins and macromolecules: as shown by recent experiments [61, 62], such a crowding seems to have only a slight influence on folding; the matter however is a current subject of debate.

Despite this open issues, many facts about the folding of small single-domain proteins are well established. The physical properties of most of them in the folded native conformation do not change, or change very little, when the environment is altered by changes in temperature, pH, or pressure, until a threshold is overcome. When this point is reached the *protein denaturation* occurs, a sudden complete unfolding of the protein, invariably associated with a loss of biological functions. The unfolding transition is a two-state cooperative phenomenon (see Fig. 1.8), with only the native fully folded and the denatured fully unfolded states present. Partially unfolded structures are unstable relative to both states. With abuse of language, being far from the thermodynamic limit, the folding transition is called *first order*.

As a consequence of the two-state character of folding, a solution of proteins can be well described as a chemical equilibrium of two different compounds, the native state proteins (N) and the completely unfolded ones (U). The equilibrium constant $K_{eq} = [N]/[U]$ gives us the difference in free energy ΔG_{U-N} upon the folding conformational change: $\Delta G_{U-N} = -kT \ln K_{eq}$, where k represents the Boltzmann constant and T the absolute temperature. This has allowed to estimate ΔG_{U-N}, for many proteins obtaining values ranging between -5 to -15 Kcal/mol. It is a surprisingly small value in comparison to the typical energies of hydrogen bonds and other non covalent interactions ($\sim 1 - 5$ Kcal/mol). A rough explanation of this fact, can be obtained observing that every scale energies involved in folding are large, but the various contributions have opposite signs and are comparable in magnitude and so they cancel each other. It is shown by many works in literature [63, 64, 65] that upon folding the net difference in energy due to hydrogen bonds formation is

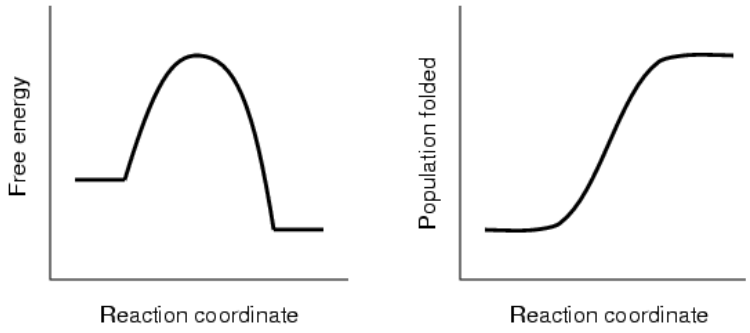


Figure 1.8: (a) Sketch of the free energy associated to the protein folding of small globular proteins as a function of the reaction coordinate – abstract one-dimensional coordinate which represents progress along a reaction pathway. The folding transition is a typical two-state process, since the free energy presents only two energy minima separated by a single energetic barrier. The minima correspond to the completely unfolded state and the natively folded one. (b) Schematic plot of the folded protein fraction. This characteristic plot represents the signature of a highly cooperative phenomenon. Each of the interactions affecting the protein molecule (electrostatic, Van der Waals, hydrogen bond) is weak nd contributes very little to the stability of the protein structure, yet the very large number of these interactions act together cooperatively to impart high stability to the protein fold.

almost zero: if the protein native state has a large energy gain by forming intramolecule hydrogen bonds, it has on the other hand a comparable energy loss because it is unable to form hydrogen bonds with water molecules. Actually whether the peptide hydrogen bonds stabilise or otherwise opposes folding is currently controversial [66]. In a so fine tuned system a small effect can make the difference.

Since the beginning the kinetics of folding has represented another fascinating puzzle. The time scales of the single-step folding process of small proteins range from $10^{-3}s$ to $10^{-1}s$ [8]. A paradox, pointing out how these time scales are short, was formulated by Levinthal [67]. Assuming for simplicity that each amino acid can choose only between three different configurations per torsional angle, then a chain of 100 residues has to explore $9^{100} \sim 10^{95}$ conformations in order to find its equilibrium state. The bond rotation speed can not be less than $10^{-13}s$, implying that the universe would end before the chain could encounter the native conformation via an unguided search. It is then

clear that the folding can not occur through a random sampling of the conformational space, evidence that determined the effort spent understanding the *folding pathways* towards the native state [68], that means to identify for each protein the intermediates that define and direct the folding.

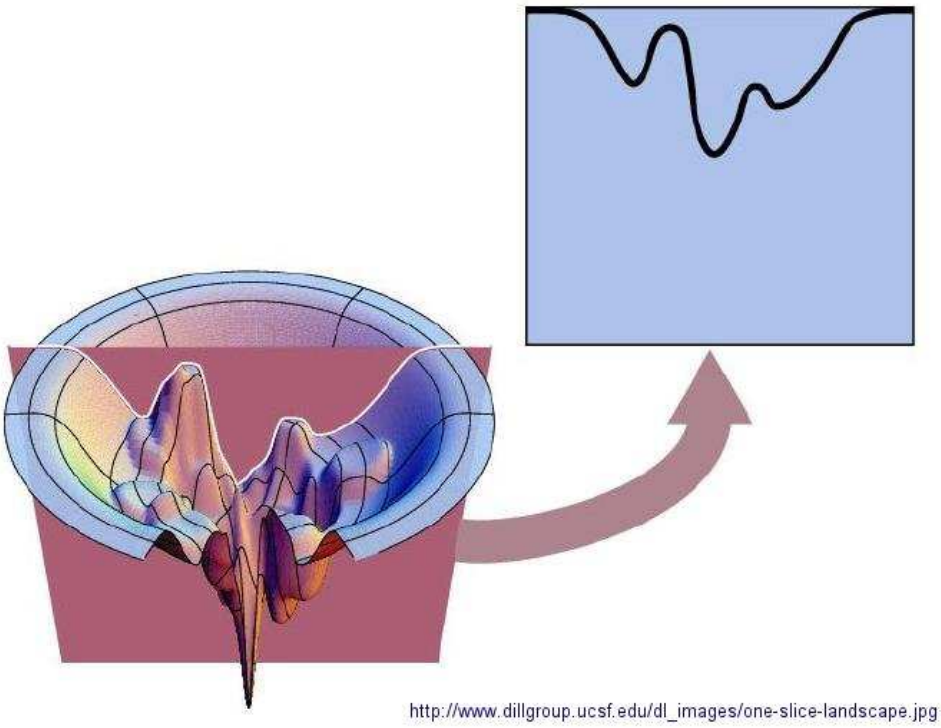


Figure 1.9: Sketch of the rugged funnel energy landscape. In order to give an explanation to the rapidity and reproducibility of folding, Bringelson and Wolynes hypothesised that the energy profile among the conformational space has the shape of a rugged funnel.

In the late 1980's an alternative viewpoint emerged [14, 69, 70, 71]; according to it what is most important for understanding the folding process is a global overview of the protein's energy surface, the *energy landscape*. There is no single pathway but a multiplicity of folding routes, since globally the energy landscape resembles a funnel driving the protein throughout folding. The folding landscape is necessarily rugged because the polypeptide chains

sample many conformations during their motion and can make inappropriate contacts between residues. Nevertheless because native contacts and local conformation energies are more stabilising than their non-native competitors, there is a smooth overall slope of the energy landscape towards the native state. This intuition leads to the idea of *rugged funnel energy landscape* shown in Fig. 1.9.

In order to understand the connection between amino acid sequence and folding kinetics, several studies involved concepts borrowed by the statistical mechanics of disordered systems. The paradigm of this branch of statistical mechanics is represented by the *spin glass* model[1] [72, 7], in which the concept of *disorder* was introduced for describing more realistic systems, such as impure materials. According to the main idea of the model, the high complexity of such systems can be conveniently described including a stochastic disorder in the Hamiltonian, and treating the energy in a first approximation as a random variable. A feature typical of disordered systems is their *frustration*: for the presence of conflicting forces and geometric impositions, the system can not satisfy all the constraints (geometric and energetic) at the same time. As a consequence of this fact a unique ground state does not exist, and the system behaviour is governed by the *rugged energy landscape*, that is characterised by a huge number of minima with roughly the same energy, separated by high barriers. A modest change in the system configuration gives rise to a large change in energy, and low energy states exist that have very different configurations but are close in energy.

Given the high complexity of the protein molecule, determined by the amino acid heterogeneity and the complex chemistry of the protein backbone, a treatment of the protein chain as a disordered system looks like quite reasonable. The *random heteropolymer model* was then introduced [14]. Nevertheless, this framework can not be completely adapted to the case of protein molecules. In a situation as the one depicted by the rugged energy landscape, the folding kinetics would be slow and unreliable, getting easily trapped in one of the many non-native compact conformations representing an energy minimum.

[1] A spin glass is a magnetic system where ferromagnetic and anti ferromagnetic bonds are randomly distributed.

To overcome this tangle, Bringelson and Wolynes proposed that protein-like sequences were selected by evolution according to a principle of *minimal frustration*. The protein amino acid sequences are not random, as supported by the fact that the number of sequences found in nature is extremely small compared with the possible combinations of the twenty amino acids[2]. Therefore minimally frustrated folded structures are the product of well-designed sequences, which are eligible to promote rapid and reproducible folding and avoid glassy behaviour. Given a sequence of amino acids, with all the attendant details of the side chains and the surrounding water, one obtains a funnel-like landscape (Fig. 1.9) with the minimum corresponding to its native state structure.

The aforementioned considerations stress the importance of the primary structure in determining the energy landscape and hence the native conformation. Nevertheless the tie between the amino acid sequence and the native state shape is not so strict. It is proved by experiments [73, 74, 75] that, in some cases, even a mutation of the 60% of amino acid sequence can leave both the native state and the folding kinetics unchanged. However, in other cases a change in a single amino acid can alter completely the protein shape and functionality: usually the most important amino acids to stabilise the structure are the ones of the protein hydrophobic core. Moreover, as observed in the previous section, the correspondence between sequences and native state topologies is many-to-one [11]. If the protein sequences are very few compared with the possible combination of the twenty amino acids, the different topologies are even less. Two different proteins can accomplish different functions, though they are characterised by the same conformation, simply by having different amino acids in the surface: for instance two enzymes with equal shape and different amino acids in the active site catalyse completely different reactions.

[2]If we consider, for instance, a polypeptide chain with 250 residues we have $20^{250} \sim 10^{325}$ possible sequences, that is a very large number compared with the number of sequences found in nature [60].

1.4.2 Theoretical approaches to the protein folding

The Anfinsen experiments provided a firm ground for a physical approach to the theoretical investigations on protein folding process. The native configuration is the thermodynamic equilibrium state of the solvent-protein system, for this reason molecular dynamics simulations represent the most suitable tool in order to study its properties. Nevertheless, at present days a fully *ab initio* quantum molecular dynamics method is too much computationally expensive to be employed even in the case of small proteins. For this reason a careful analysis become necessary in order to simplify the system reducing the huge number of degrees of freedom associated to it. In section 1.2.3 we have described the most important mesoscopic interactions affecting the system in a classical semi-empirical view proper of most molecular dynamic methods, but both the choice of the potential expressions and the tuning of parameters could result very complicated.

Chemical details are important in order to understand how a given protein perform its proper function. But if our aim is to study general properties of the overall class of protein molecules, a coarse grained approach proper of statistical mechanics is often very useful. The polypeptide chain can then be represented as a self avoiding walk on a lattice or as the so called "string and bead" model, a freely jointed chain out of lattice with an hard core interaction responsible for the self avoidance. Each amino acid is then reduced to its C_α atom, neglecting the side chains and other backbone atoms. However the attractive interactions between monomers can, according to the simplification degree that the model wants to reach, take into account the amino acid specificity by means of effective potentials. The HP model [76, 77], for instance, considers only a distinction between hydrophobic (H) and polar (P) amino acids: the energy term $V_{H,H}$, representing the interaction between two hydrophobic residues, is negative, while the other two terms, $V_{H,P}$ and $V_{P,P}$, are larger. In this manner the strong attraction imposed between hydrophobic residues expresses their tendency to bury themselves into the protein core, taking into account the solvent effect in the simplest possible way, without considering it explicitly.

By means of these simplified models many obscure aspects have been investigated: first above all the origin of the remarkable regularity of secondary structures. The many-to-one correspondence between sequences and folds, moreover, raises the issue of the role of sequence in selecting native state conformations. Energetic considerations, such as intra-backbone hydrogen bond formation, of course cover an important role in selecting native conformations. In the usual string and bead polymer model, nevertheless, the number of low energy structures is huge compared with the number of folds, and low energy does not necessarily imply thermodynamic stability. For this reason, in the latter approach, the requirement of minimum energy alone could not explain the presence of both highly symmetrical protein-like structures and few distinctly shaped folds.

A first work trying to explain the secondary structure presence without invoking any chemical and energetic detail is due to Chan and Dill. On the basis of exhaustive enumerations in lattice models, they proposed that compactness by itself, a purely geometrical requirement, is sufficient to drive secondary structure formation [78, 79]. Nevertheless, this revealed itself as an artifact of lattice structural order [80, 63, 81], since a generic compact polypeptide chain in the continuum space was shown to account for only a small secondary structure content [82].

More recent works suggested a selection mechanism based on *high designability* for native conformations [45, 83, 85, 84]. In lattice models, a small number of highly designable structures emerge with a number of sequences successfully folding into them much larger than the average. The structure having this property should be particularly stable against mutation and more thermodynamically stable than other structures, thus yielding more efficient folding. An idea emerging in this works, and upheld also by successive studies [86, 87], is that among all possible energetically favoured configurations the protein backbone will attain those that are optimal, under the action of evolutionary pressure favouring *geometrical accessibility*.

As proved by recent experiments, the topology of a protein native state appears to determine the major features of its folding energy landscape [15]. The folding rate and the distribution of structures in the transition states ensemble,

results to be determined by the native state topology, while, at the same time, they are only weakly affected by sequence mutations that leave the topology unchanged.

Purely geometrical considerations seem to be important in order to understand the special properties of protein native state conformations. The *tube-like polymer* model that will be the subject of the following chapter, is the development of the aforementioned intuitions and enlightens the importance of the intrinsic symmetry of the polypeptide chain in selecting the regular protein-like motifs.

Chapter 2

The tube-like polymer model

In this chapter we want to introduce the very simplified model of the *tube-like polymer*; its aim is to give some insight into the open issues concerning the high regularity and reproducibility of protein conformations. As previously discussed in section 1.4.2, many indications suggest that pure geometrical criterions should play a crucial role in selecting the protein-like structures. According to the main idea, underlined in previous works [86, 87], the large basin of attraction towards the native state can be a consequence of distinctive qualities not just of the amino acid sequences, but rather of the native state conformations themselves: the funnel formation in the energy landscape may be determined by the property of the protein structures of being easily geometrically accessible.

The importance of geometrical requirement in determining the equilibrium conformations of a system is evident in crystal lattices. The exigency of ideal packing and periodicity determines a priori the menu of all possible lattice conformations: the Russian mathematician Fyodorov first catalogued the 230 space groups corresponding to the discrete symmetry operations including translation, reflection and rotation; the different space groups determine all the possible molecule arrangements in a crystal, independently of the kind of forces acting between the molecules. This gives us a natural explanation of the high regularity of crystals and of the many-to-one correspondence between chemical compounds and lattice structures: the lattice conformations are predetermined by symmetry principles and any chemical compound has only to

pick out among the possible structures the most favourable depending on the specific interactions.

Following intuition one may suppose, hazarding a comparison between protein native states and crystal structures, that in order to satisfy the requirements common to all proteins, few conformations are selected among all possible compact structures. A limited number of peculiar conformations could be able to respond to the energetic and geometrical exigency of proteins backbone, warranting a fast and reproducible folding. This could be the origin of the ubiquity of secondary structures in proteins.

In order to clarify these ideas we can try to reformulate the concept of 'geometric accessibility', that is rather vague and difficult to deal with in the continuum space. We can observe that it should be connected with the *free volume* available around the conformation such that the protein backbone can be moved within it without the native state topology is drastically changed.

Wishing to address this issue within a simplified model we need a mathematical tool to catch this property. For this purpose we can borrow the notion of *thickness* introduced in the context of knot theory [88, 89, 90]. The thickness associated with a smooth curve is the radius of the larger tube allowed with uniform circular radius and smooth surface, which can be inflated around the curve. Therefore to put our ideas in a concrete form we identify then the available free volume around a polypeptide chain conformation with the volume of the tube inflated around the corresponding curve. This allows us to study the variational principle of finding the *ideal shape* of a curve subject to compaction, that is the curve conformation associated with the maximum thickness under some defined compactness constraint. This problem has the equivalent formulation of finding the most compact shape of an impenetrable tube with fixed thickness. In such a way it becomes a problem of optimal packing, a generalisation of the simpler (but not really trivial at all) issue of finding the closed packing of identical impenetrable spheres, whose solution, the FCC lattice, was conjectured almost four centuries ago by Kepler, and has reached the mathematical demonstration only few years ago [91, 92, 93].

The results of this basic model, that we are going to analyse in this chapter, are remarkable: an impenetrable tube subject to compaction leads, without any

additional ingredient, to a limited menu of *marginally compact* motifs [17], characterised by high regular shape. In this set helices and double helices emerge with conformational parameters very closed to the respective ones of helices of proteins and double helix of DNA [16, 94]. This supports our hypothesis that a simple geometric optimisation underlies the selection of elementary motifs shape in proteins and biomolecules in general.

A determinant role in this framework is covered by the intrinsic symmetry of the polypeptide chain. The fact that it is modelled as a tube instead of a chain of spheres represents the main feature of this model. As mentioned in section 1.4.2, studies conducted on 'string and bead' models have demonstrated that compaction by itself is unable to reproduce regular conformations like secondary structures. This is because the spheres, the 'string and bead' model building blocks, does not represent the suitable object to address the uniaxial symmetry of the amino acid chain. In the same manner as crystals, whose macroscopic regularity reflects the symmetry of the building blocks compounding it, helices and sheets, the motifs of protein conformations, are the macroscopic expression of the intrinsic anisotropy of the polypeptide chain.

In order to give an overview of the studies conducted in these few years on the tube-like polymer model we will start introducing its basic ingredient, the thickness, whose mathematical formulation [90], both for smooth curves and for discrete chains, will be the subject of section 2.1. In the same section we will illustrate also the analysis on the ideal shape of a curve subject to compactness constraint [16, 95] (different compactness constraints were tested and their different results will be discussed) and we will describe in detail the first remarkable result: the *ideal helix*, having the good property of representing the tube helical conformation without 'holes', was obtained as the ideal shape of a tube with local compaction constraints; its importance stems from the fact that its pitch to radius ratio is very closed to the α-helix one. Following the first encouraging results, the *tube-like polymer model* was formulated [17, 96, 19], and will be mentioned in section 2.2. The ground state conformations of a discrete chain subject to the thickness constraint, and with self-attracting monomers, were examined. Tuning the parameters of the system the marginally com-

pact 'phase' emerges in between the two typical phases of polymers, compact and swollen. The peculiar features characterising this phase make it the ideal setting for describing biopolymer conformational properties. The idea of *pre-sculpted energy landscape* is suggested by these results [20]. It is able to gather all the intuitions given in this paragraph: instead of each sequence shaping its own free energy landscape, the overarching principles of geometry and symmetry determine the menu of possible folds that sequences can choose from. Finally a significant development of the model will be the object of section 2.3 [21]. A simple attractive interaction is introduced that is able to catch the protein tendency to collapse induced by the hydrophobic effect. This is simply proportional to the area of the buried tube regions, hidden from solvent particles. We argue that the thickness, on one hand, represents the suitable tool for catching the steric constraints of the polypeptide chain, while the buried area, on the other hand, represents an efficacious and physically motivated way in order to describe a self-interacting polymer.

2.1 Optimal packing of a thick tube

The notion of thickness was first introduced in the mathematical context of knot theory with the purpose of studying the problem of the ideal shape of a knot, formulated by Katritch *et al*. [89, 97]. Given a closed curve in the three-dimensional space, we can think to enclose it with a tube having a uniform circular radius. Inflating the tube a limit is reached when the tube can not grow anymore, since it either ceases to be smooth, due to local bending of the chain, or exhibits self-intersections, due to the proximity of two different portions of the chain. The tube radius associated to this extremal limit defines the thickness of the closed curve. Finding the ideal shape of a knot simply means to seek the curve with maximum thickness among the class of closed curves with fixed length associated to a given knot topology.

This is actually a problem of optimisation of volume occupation. Indeed reversing the perspective it is equivalent to looking for the shape that a tube with fixed thickness would take when tied in a given knot in the most com-

pact manner, that means in this case to use the minimum tube length possible. Similarly two equivalent formulations can be pointed out also for the simpler analogous issue: the problem of closest packing of an identical hard spheres set can be alternatively formulated as the determination of the arrangement of a set of points in a given volume that results in the minimum distance between any pair of points, $d = \min_{i<j} |\mathbf{x}_i - \mathbf{x}_j|$, being as large as possible [98]. Notice that in this case we are interested in the bulk optimal arrangement, neglecting the boundary effects.

Our aim is to generalise the notion of ideal shape of a knot, to the class of open curves: we replace the restriction of being closed and hence with fixed topology with an explicit compaction constraint, for instance a confinement in a finite volume. Therefore among the class of open curves with fixed length and a defined compaction constraint, we search for the ones maximising their thickness. Purpose of these preliminary studies is the search of a first confirmation of the intuitions illustrated in the beginning of this chapter: the secondary structures can be selected by an optimisation of volume occupation requirement, and therefore they could just represent the ideal shapes of a chain, among all the energetically favoured, hence compact, conformations.

In this section we will introduce the analysis done on the ideal shape of a compact curve, describing the simulation carried out by Maritan, Trovato *et al*. [16, 95]. To face this problem a mathematical formulation of thickness is necessary and will be the subject of the first subsection. We will proceed along the general guidelines provided by Gonzalez and Maddocks [90], introducing the notion, immediately related to the thickness, of *global radius of curvature*, an extension of the local radius of curvature concept. First defined for smooth curves, it can be easily adapted to the case of discrete chains, providing thus a natural tool for numerical implementation.

2.1.1 Global radius of curvature

In mathematical language, a curve Γ is a continuous three-dimensional vector function $\mathbf{R}(s)$ of a real variable s with $0 \leq s \leq L$. We will assume Γ to be smooth and 'simple': the first condition implies that the function $\mathbf{R}(s)$

is continuously differentiable to any order and the tangent vector $\mathbf{R}'(s)$ is nonzero for all s, while the second one means that it has no self-intersections, $\mathbf{R}(s_1) = \mathbf{R}(s_2) \Leftrightarrow s_1 = s_2$. We do not impose the condition of closure, hence in general we will have $\mathbf{R}(L) \neq \mathbf{R}(0)$.

A smooth solid tube of constant radius δ centred on Γ may be defined, for sufficiently small δ, as the union, over all points $\mathbf{x} \in \Gamma$, of all the circular disks of radius δ centred at $\mathbf{x}$ and contained in the normal plane to Γ at $\mathbf{x}$. For non-straight curves, the thickness of the curve, $\Delta[\Gamma]$, is the critical radius above which the tube either ceases to be smooth or exhibits self-contacts. It is simple to show [99] that the tube becomes *locally* singular when its radius δ becomes equal to the local radius of curvature $R_{LC}(\mathbf{x})$ of Γ at some point $\mathbf{x}$[1]. On the other hand, the occurrence of self-contacts between different portions of the tube is a *non-local* effect. It can be demonstrated [99] that it happens when two non-adjacent points on Γ of *closest approach* have a distance less then 2δ. A couple of points, $\mathbf{x}$ and $\mathbf{y}$, of closest approach are defined by having the vector $\mathbf{x} - \mathbf{y}$ orthogonal to the tangent vectors to Γ at both $\mathbf{x}$ and $\mathbf{y}$. Indicating with Ω the set of all pairs of points of closest approach the tube thickness can then be formalised as

$$\Delta[\Gamma] = \min\left\{\min_{\mathbf{x}\in\Gamma} R_{LC}(\mathbf{x}),\ \frac{1}{2}\min_{\mathbf{x},\mathbf{y}\in\Omega}|\mathbf{x}-\mathbf{y}|\right\}. \tag{2.1}$$

In words, the thickness is either the minimum local radius of curvature or half of the minimum distance of closest approach, whichever is smaller. The notion of global radius of curvature [90] which we are going to introduce allows to capture simultaneously both possibilities.

Any three non-collinear points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ in three-dimensional space determine a unique circle. The radius of this circle defines thus a scalar function associated to each tern of points, $r(\mathbf{x},\mathbf{y},\mathbf{z})$. We call it *circumradius*, and we can easily calculate it using the formula

$$r(\mathbf{x},\mathbf{y},\mathbf{z}) = \frac{|\mathbf{x}-\mathbf{y}|\,|\mathbf{x}-\mathbf{z}|\,|\mathbf{y}-\mathbf{z}|}{4\mathcal{A}(\mathbf{x},\mathbf{y},\mathbf{z})}, \tag{2.2}$$

[1]The local radius of curvature is simply the radius of the circle which locally best approximates the curve.

where $\mathcal{A}(\mathbf{x},\mathbf{y},\mathbf{z})$ is the area of the triangle with vertices $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$. When the points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ are distinct but collinear the circumcircle degenerates into a straight line, and we assigne a value of infinity to the corresponding circumradius $r(\mathbf{x},\mathbf{y},\mathbf{z})$. Consider now a tern of points belonging to Γ, $\mathbf{x} = \mathbf{R}(s_1)$, $\mathbf{y} = \mathbf{R}(s_2)$ and $\mathbf{z} = \mathbf{R}(s_3)$. By definition, in the double limit $s_2, s_3 \to s_1$, the circumradius associated to that tern become the local radius of curvature

$$\lim_{s_1,s_2\to s_3} r(\mathbf{x},\mathbf{y},\mathbf{z}) \equiv r(\mathbf{x},\mathbf{x},\mathbf{x}) = R_{LC}(\mathbf{x}) \ . \tag{2.3}$$

Following [90], we can then generalise the concept of local radius of curvature, introducing the notion of the *global radius of curvature* $R_{GC}(\mathbf{x})$ at each point :

$$\forall \mathbf{x} \in \Gamma, \quad R_{GC}(\mathbf{x}) \equiv \min_{\mathbf{y},\mathbf{z}\in\Gamma} r(\mathbf{x},\mathbf{y},\mathbf{z}) \ . \tag{2.4}$$

It immediately follows from definition that global radius is bounded by local radius:

$$R_{GC}(\mathbf{x}) \leq R_{LC}(\mathbf{x}) \quad \forall \ \mathbf{x} \in \Gamma \ . \tag{2.5}$$

Indeed, in the case of smooth and simple curves, it can be proved [90] that the optimality condition associated with the minimization in Eq. 2.4 implies that the global radius $R_{GC}(\mathbf{x})$ may be either the local radius of curvature $R_{LC}(\mathbf{x})$, or the strictly smaller radius of a circle containing $\mathbf{x}$ and another distinct point $\mathbf{y}$ at which the circle is tangent. Thus, to determine $R_{GC}(\mathbf{x})$, one need consider only the minimization in Eq. 2.5 with the restriction $\mathbf{y} = \mathbf{z}$. In the trivial case of a straight line, the global radius is infinite for all points.

It is a quite remarkable result of Gonzalez and Maddocks the proof that, for any smooth simple curve Γ, the minimum global radius of curvature over all points $\mathbf{x} \in \Gamma$ is exactly its thickness $\Delta[\Gamma]$:

$$\Delta[\Gamma] = \min_{\mathbf{x}\in\Gamma} R_{GC}(\mathbf{x}) \ . \tag{2.6}$$

We will not go in depth with the demonstration of Eq. 2.6, that can be found in [90]. We want to highlight both the conceptual and the practical importance of such a reformulation of the thickness notion. The connection between thickness and global radius of curvature allows to characterise thickness in a vary simple way, by taking into account at the same time both local and non-local

effects, which are instead explicitely distinguished in Eq. 2.1. The definition of thickness as the minimum value of the circumradius function over all triplets of points is also more apt to numerical implementation, since in computer simulation one is naturally dealing with discrete chains.

2.1.2 Ideal shape of a chain subject to global constraints

The fundamental tool of global radius of curvature allows us to face the optimisation problem of ideal shape of a compact curve. It can be formalised in this way: let $\mathcal{R}$ denote the set of all simple, smooth curves Γ with fixed length $L > 0$ satisfying a given compactness constraint, the ideal shapes Γ^* in $\mathcal{R}$ is then the curve satisfying the condition:

$$\Delta\left[\Gamma^*\right] = \sup_{\Gamma \in \mathcal{R}} \Delta\left[\Gamma\right] \ . \tag{2.7}$$

Note that Eq. 2.7 immediately yields the trivial result, in agreement with intuition, that the optimal shape in the absence of any compactness condition is the straight line, which has infinite thickness.

The existence of an ideal smooth shape Γ^* achieving the supremum in Eq. 2.7 has not been demonstrated yet, even in the simpler case of closed curves with fixed knotted topology. In this latter case it is possible to derive a necessary condition, implied by Eq. 2.7, that any smooth ideal shape must satisfy [90]. A smooth closed curve can be ideal only if its global radius of curvature function is constant and minimal on every curved segment of the curve. The proof given in [90] can not be easily generalized to the case of a generic compactness constraint. Nevertheless, this property was found to be numerically verified.

we have found this property to be verified in all our simulations.

Before describing the simulations we need to generalise the notions just introduced to the case of a discrete curve. A discrete curve Γ_N is an ordered set of distinct points $\{\mathbf{x}_0, \ldots, \mathbf{x}_N\}$ in three-dimensional space. To any discrete curve Γ_N one can associate a continuous, piecewise linear curve Γ by connecting $\mathbf{x}_0$ to $\mathbf{x}_1$ with a straight line and so on. For simplicity we will consider the particular case in which all the bond distances between successive points are

fixed to the same value, b, such that the chain has a length $L = b(N+1)$. Extensions of definitions introduced in the previous section are straightforward [90]. The local radius of curvature is simply the radius of the circle containing three consecutive points:

$$R_{LC}(\mathbf{x}_i) \equiv r(\mathbf{x}_{i-1}, \mathbf{x}_i, \mathbf{x}_{i+1}) \ . \tag{2.8}$$

For a discrete curve is well defined also the *non-local* radius of curvature:

$$R_{NLC}(\mathbf{x}_i) \equiv \min_{\substack{0\le j<k\le N \\ j\neq i\neq k \\ k-j+|k-i|+|j-i|>4}} r(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) \ , \tag{2.9}$$

a quantity being able to keep track of purely non-local effects measuring the proximity of different portions of the chain. Therefore the global radius of curvature simply combines both the quantities: at each point $\mathbf{x}_i$ of Γ_N it is defined as

$$R_{GC}(\mathbf{x}_i) \equiv \min_{\substack{0\le j<k\le N \\ j\neq i\neq k}} r(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) \ . \tag{2.10}$$

Also in this case the quantity $R_{GC}(\mathbf{x}_i)$ can be put in relation to the thickness, $\Delta[\Gamma_N]$, the last being the radius of the smallest circle containing three distinct points of Γ_N:

$$\Delta[\Gamma_N] \equiv \min_{0\le i\le N} R_{GC}(\mathbf{x}_i) \ . \tag{2.11}$$

The functional $\Delta[\Gamma_N]$, defined among the set of all discrete curves with fixed b, N and a given compaction constraint, can easily optimised numerically, using the *simulated annealing* method [100]. This method was extensively employed in many analysis carried out in this thesis, and it is described in appendix A. When dealing with discrete curves a pairwise self-avoidance constraint need to be imposed in order to avoid trapping in self-intersecting structures or convergence to pathological optimal shapes. Therefore the curves have then to satisfy the condition $|\mathbf{x}_i - \mathbf{x}_j| \geq R_0$ for all pairs of different non-consecutive beads, where R_0 was chosen slightly greater then b.

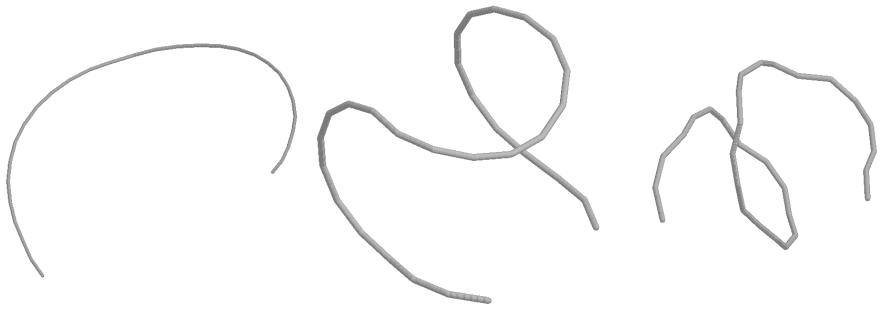
In the first simulations [95, 16] *global* compactness was imposed in two different ways obtaining the same results: in the first case the whole chain was confined in a cube with side l, while in the second case the gyration radius of

the chain was enforced to be less than a fixed value, G. The gyration radius R_G of a discrete chain in a conformation Γ_N is the quadratic mean bead distance from the chain center of mass:

$$R_G^2\left[\Gamma_N\right] = \frac{1}{N+1}\sum_{i=0}^{N}\left(\mathbf{x}_i - \mathbf{x}_{cm}\right)^2 \ , \quad \mathbf{x}_{cm} = \frac{1}{N+1}\sum_{i=0}^{N}\mathbf{x}_i \ ; \tag{2.12}$$

it is usually employed in polymer physics as a typical characteristic length of the chain.

(a) $G = 6.0$, $\Delta = 6.42$ (b) $G = 4.5$, $\Delta = 3.82$ (c) $G = 3.0$, $\Delta = 1.93$



(d) $l = 22.0$, $\Delta = 6.11$ (e) $l = 9.5$, $\Delta = 2.3$ (f) $l = 8.1$, $\Delta = 1.75$

Figure 2.1: Examples of optimal conformations. The chains are made up of $N = 30$ equally spaced points, the spacing between neighbouring points, b, is defined to be 1 unit. In the top row the optimal shapes obtained by constraining the string with a radius of gyration less than G are shown. The optimal shapes displayed in the bottom row otherwise were obtained by confining the string within a cube of side l.

The results obtained from simulations can be summarised as follow: if the chain has enough space (large values of G or l) its optimal conformation is, as expected, a straight line; for very low values of G (l), on the other hand, the conformations maximising the thickness are disordered globules completely

lacking in regularity. The optimal shapes obtained, varying the parameters l and G in the intermediate regime, are shown in Fig. 2.1. For both the methods, few very regular conformations appear, helices and saddles, that are close competitors for optimality.

2.1.3 Ideal shape of chain subject to local constraints: the ideal helix

The optimal conformations obtained with global constraints resulted strongly affected by boundary effects: they depend on the value of the ratio of the chain length over the typical length controlling the size of the chain. This changeable scenario is somewhat unsatisfactory, and the existence of a single regular best packing chain conformation, the *bulk* conformation, would be much more appealing. To overcome this problem, studies were carried out on the ideal shape of a chain in the limit of infinite length, enforcing the compactness *locally*. Specifically, in the simulations, the compactness was imposed [95] on each set of j consecutive beads, either forcing it to stay in a finite volume (i.e. a cube of side l_j), or fixing an upper limit, G_j, to its gyration radius.

For both the methods, by varying the parameters (j, l_j, G_j) within a broad range of values, we find always the same optimal configuration, that is the *ideal helix*, as shown in figure 2.2. It is an extremely regular helix since it is characterised by a degeneracy of local and non-local radius of curvature, which result along all the curve equal within 0.1%[2]. The geometrical meaning of this feature can be understood when thinking about the tube swelling uniformly around the curve. This equality implies that the tube at the same time ceases to be smooth and exhibits self-contacts: indeed there is no free space left either between consecutive turns of the helix or in the plane perpendicular to the helix axis. Optimality requires local and non-local effects to be on the same footing, being in this way the occupation of three-dimensional space optimised, as was our aim.

A particular value of the pitch to radius ratio, c, characterises the helix hav-

[2]The same degeneracy characterises also optimal shape with global compactness condition. But in this case the equality is only within a 1% of precision.

ing this property: simple analytic calculations, done for the continuous curve, yield the value $c \equiv c^* = 2.512\ldots$. This is a special critical value separating two different regimes: if $c > c^*$, the tube would stop swelling because of local singularities, leaving free space between consecutive turns; if $c < c^*$, the tube would stop swelling due to self-intersections between consecutive turns, leaving free space along the axis of the helix.

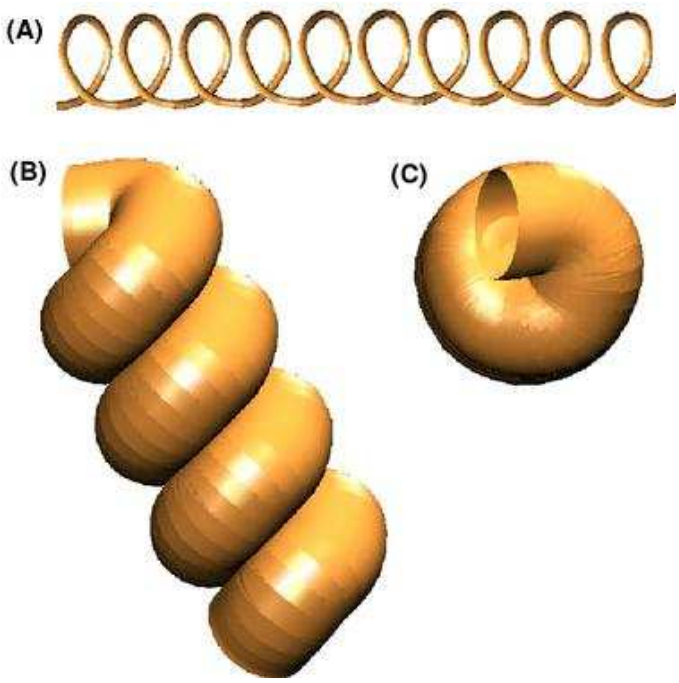
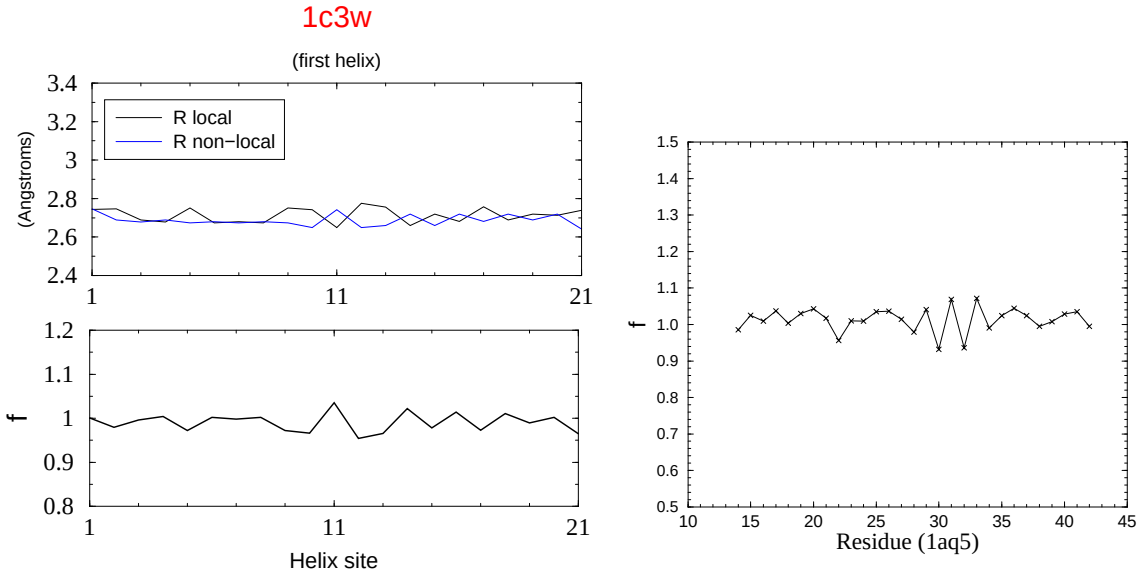


Figure 2.2: Optimal helix with local constraints; $N = 66$, $j = 6$, $G_j = 1.0$. (A) Bare skeleton of the optimal helix connecting the discrete beads. (B), (C) Side and top views of the same helix inflated to its thickness.

The ideal helix represents a remarkable result: the helical conformation exhibits translational invariance along the chain, expressing in this way the natural symmetry of the system; no boudary effects influence the conformaion, it represents then a bulk-like solution to the best packing problem for chains [16]. Moreover, the structures of naturally occurring proteins taken from the Protein Data Bank were analysed, finding that the two different kinds of helical motifs, the α-helix and the collagen helix, appearing most frequently in protein structures, share the same peculiar geometry as optimal best-packing helices.



(a) (b)

Figure 2.3: (a) Analysis of the first α-helix of bacteriorhodopsin (pdb code 1c3w): in the top the values of the local and non-local radii of curvature as a function of sequence position, in the bottom the respective values of f. (b) Plot of f for a single collagen helix (pdb code $1aq5$): we considered the residues 14-41; the same plot for each of the three collagen helices would simply superimpose.

In figure 2.3, we plot the function $f(\mathbf{x}_i)$ defined as

$$f(\mathbf{x}_i) \equiv \frac{R_{NLC}(\mathbf{x}_i)}{R_{LC}(\mathbf{x}_i)} , \tag{2.13}$$

for the discrete chains formed by the backbone C_α atoms of the helical segment of two different protein structures in Protein Data Bank: the first α-helix of bacteriorhodopsin (a remarkable α-helix), and a single collagen helix of the protein $1aq5$, present in the chicken cartilage matrix tissue. Despite the two helices are characterised by a rather different geometry, e.g. a different number of residues per turn (3.0 instead of 3.6), both of them have the ratio $f(\mathbf{x}_i)$ oscillating closely around the optimal unity value. This hints that, despite the complex atomic chemistry associated with the hydrogen bonds and the covalent bonds along the backbone, helices in proteins satisfy optimal packing constraints [16]. This result implies that the backbone sites in protein helices have an associated free volume distributed more uniformly than in any other

conformation with the same density. This is consistent with the observation we put forward, that secondary structures in natural proteins have a much larger configurational entropy than other compact conformations [86].

2.2 The thickness as an ingredient of a biopolymer model

The global radius of curvature provides us an effective tool to describe polymers with non-zero thickness. Starting by the worm-like chain model, where the polymer is reduced to a continuous chain, we can catch the impenetrable tube of radius Δ inflated around the curve simply imposing a *three-body hard core potential* among all the terns of points, $(\mathbf{x}, \mathbf{y}, \mathbf{z},)$, belonging to the chain:

$$\mathcal{V}_3(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \begin{cases} +\infty & \text{for} \quad r(\mathbf{x}, \mathbf{y}, \mathbf{z}) < \Delta \\ 0 & \text{otherwise} \end{cases}, \tag{2.14}$$

where $r(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is the circumradius function. The thickness is then the natural generalisation to lines of the hard-core potential catching the spherical excluded volume surrounding point particles, when interacting each other with a strong repulsion.

When polymers are modelled with continuous curves, a difficulty arises trying to describe their self-interaction through the standard pairwise potential. The latter involves a two-points distance that is unable to distinguish whether two neighbouring points are adjacent along the chain or otherwise belonging to distant curve regions approaching each other because of a curve bending. The standard hard-core repulsive interaction imposed among the curve points would necessarily lead to singular contributions to the Hamiltonian of the system, hence the energy would be infinity for all the chain conformations. In order to impose a self-avoidance upon a curve without improper singular terms, the Dirac delta function can be employed, adding to the Hamiltonian a potential energy

$$E(\Gamma) = \int_0^L \int_0^L \delta\left(\mathbf{R}(s_1) - \mathbf{R}(s_2)\right) \, d\Gamma_{s_1} d\Gamma_{s_2}, \tag{2.15}$$

where Γ is a chain conformation and $d\Gamma_s$ is the element of arc length at position s along the curve. But this model does not capture effects of non-zero thickness. The three-body hard core potential defined by the global radius of curvature, on the contrary, is free of singularities and at the same time provides an intrinsic length scale appropriate for modelling the thickness of a physical system. It is indeed able to capture in the proper way both the local and non-local effects: we must remember that when the three points of expression 2.14 become coincident it does not arise any singularity, we recover instead the local curvature radius of the curve.

To sum up, a two-body function is the proper tool to model the hard core repulsion of point particles, zero-dimensional objects. Likewise the self-avoidance of a thick line, a one-dimensional object, requires a three-body function. This analogy hints at an empirical rule, that is in fact followed by the two-dimensional counterpart too: as described in reference [102], a self-avoiding surface, $\mathcal{S}$, can effectively be described using a four-point potential, involving the radius of the unique sphere passing to one point $\mathbf{x} \in \mathcal{S}$ and tangent to $\mathcal{S}$ in an another point $\mathbf{y}$[3].

Despite its good mathematical properties, one should observe that thickness is not a priori necessary to describe the polypeptide chain, because of the discreteness of the latter. Being it a collection of atoms bound together, a string and bead model, which does not involve any singularity problem, should be adequate to catch its geometrical features. It is a fact nevertheless that the preliminary results obtained studying the ideal shape of a compact curve are remarkable: a tube, when subject to compaction, behaves in a qualitatively different way compared to a chain of spheres, and the peculiar aspects of the former in respect to the latter seem to be the missing ingredients in order to properly describe the polypeptide chain.

To further investigate this point, we can study a very basic model [18, 19, 96], *the tube-like polymer model*. Given a chain with a fixed thickness, we impose an attractive interaction between the different parts of the chain, and we seek for the conformations of minimum energy. In this way the issue of finding

[3]It is a particular four-point function since the tangent plane to $\mathcal{S}$ at $\mathbf{x}$ may be constructed through a limit of three points.

the ideal shape of a knot is reversed: instead of maximising the thickness of a chain subject to compaction we seek for the most compact conformation of a chain subject to a thickness constraint. We have observed in previous section that the two problem are in principle equivalent. Anyway, as we are going to see, we will be able in this manner to highlight the peculiar features carried by the thickness, when this ingredient is added to the standard string and bead model.

2.2.1 Compact conformations of a tube-like polymer

Consider a polymer modelled as the discrete chain described in the previous section: a set of N points tethered together with fixed bond distance, b. Let choose for the self-interaction of the chain the Hamiltonian

$$\mathcal{H}(\Gamma) = \sum_{0 \le j < i-1 \le N} \mathcal{V}_2(\mathbf{x}_i, \mathbf{x}_j) + \sum_{0 \le i < j < k \le N} \mathcal{V}_3(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) . \tag{2.16}$$

where V_3 is the three-body hard core potential aforementioned, while V_2 is simply a square-wall potential,

$$\mathcal{V}_2(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} +\infty & \text{for} \quad |\mathbf{x}_i - \mathbf{x}_j| \le 2R_0 \\ -\epsilon & \text{for} \quad 2R_0 < |\mathbf{x}_i - \mathbf{x}_j| \le R_1 \\ 0 & \text{otherwise} \end{cases} . \tag{2.17}$$

Remember that in the case of a discrete chain, as mentioned in section 2.1.2, a pairwise hard-core potential must be added to the thickness constraint, in order to avoid trapping in pathological structures. For simplicity we can choose $\epsilon = 1$, in this case the energy of a polymer configuration is proportional to the number of contacts, that is the number of monomer pairs that are in a distance closer than R_1. With this simplified Hamiltonian the minimum energy conformations of the chain varying the ratio Δ/R_1 can be studied. Careful numerical simulations, through the simulated annealing method, have been done [18] for short chains (several values of N between 10 and 20 were tested obtaining the same behaviour), and the results are summarised in Fig. 2.4. Large values of thickness, with respect to R_1, prevent the tube to form contacts, so the chain is in a swollen configuration. For small values of Δ/R_1 on

the other hand, the chain collapses in very compact conformation, forming many contacts. This regime is characterised by a large number of minimum energy conformations, feature typical of a purely string and bead model. Indeed if $\Delta/R_1 << 1$, the thickness effects are negligible and we recover the behaviour of a chain of spheres interacting only with the second term of the 2.16: a disordered liquid-like phase is reached when freezing such a polymer at low temperatures [103], preceding a solid-like phase, that presents a crystalline order. For $\Delta/R_1 \simeq 1$ finally the thickness strongly affect the system determining a peculiar character. We find in this regime a few number of very regular ground state conformations: helices, saddles and planar hairpins. The large degeneracy of the compact regime is strongly reduced and the conformations are characterised by a few number of contacts, hence we shall refer to this 'phase' as *marginally compact*.

Thickness selects in the marginally compact phase conformations with remarkable geometrical features. Both helices and planar hairpin, obtained in the simulations, catch the microscopic symmetry of the basic element compounding the system, the tube-like chain. The anisotropy of the latter indeed determines in its compact state the breakdown of the discrete translational and rotational invariance typical of the best-packed conformations of a spheres set, arousing rather uniaxial and biaxial arrangements.

In the following subsection we will give a detailed comparison between the conformations obtained by the tube-like polymer model and the motifs forming the building block of proteins. The similarity is strong, and this is quite surprising because the model neglects any sequence specificity. This indicates that the intrinsic anisotropy of the polypeptide chain covers a fundamental role in selecting protein conformations. Indeed we can have a confirmation of this idea observing that many chemical properties of this polymer enhance its local uniaxial symmetry. The partial double-bounded nature of the peptide bond is responsible of the high local bending rigidity of the chain. Moreover since the Van der Waals radius is reduced when two atoms are covalently bound together, when modelling the polypeptide chain with a string of spheres, we should impose that each of them overlaps with its neighbours. The tube is the simplest geometry well fitting the shape of this object.

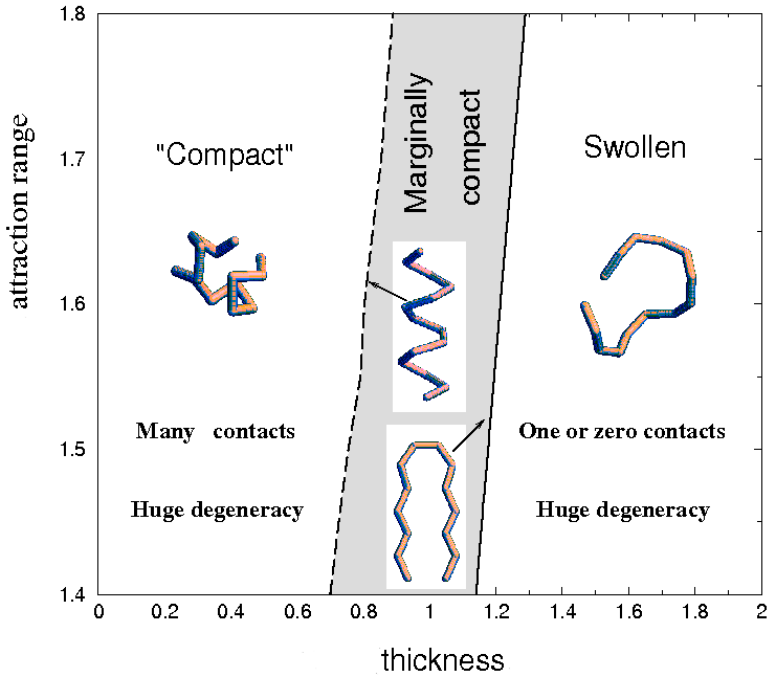


Figure 2.4: Sketch of the different regimes, their characteristics and the associated ground state structures, found for short tubes in the (Δ, R_1) plane. It summarises the results obtained in the simulations described in ref. [18]. The value of R_0 has been set to 0.55 where b represents the unit of length.

The effective thickness of the polypeptide chain is determined not only by local rigidity but also by non-local effects. As already mentioned, in order to accommodate side chains enough free space has to be left around the backbone when reaching the compact conformation. Besides, the attractive interaction between distant segments of the chain approaching each other , is optimised when the two segments lie side by side[4], which strongly reminds the behaviour

[4]It is a feature of thick chains in general (see references [96, 19]). Anyway the requirement of parallelism is particularly strong in the case of peptide hydrogen bond. Quite rigid geometrical constraints characterise this interaction as they can be observed through an analysis on the PDB [20].

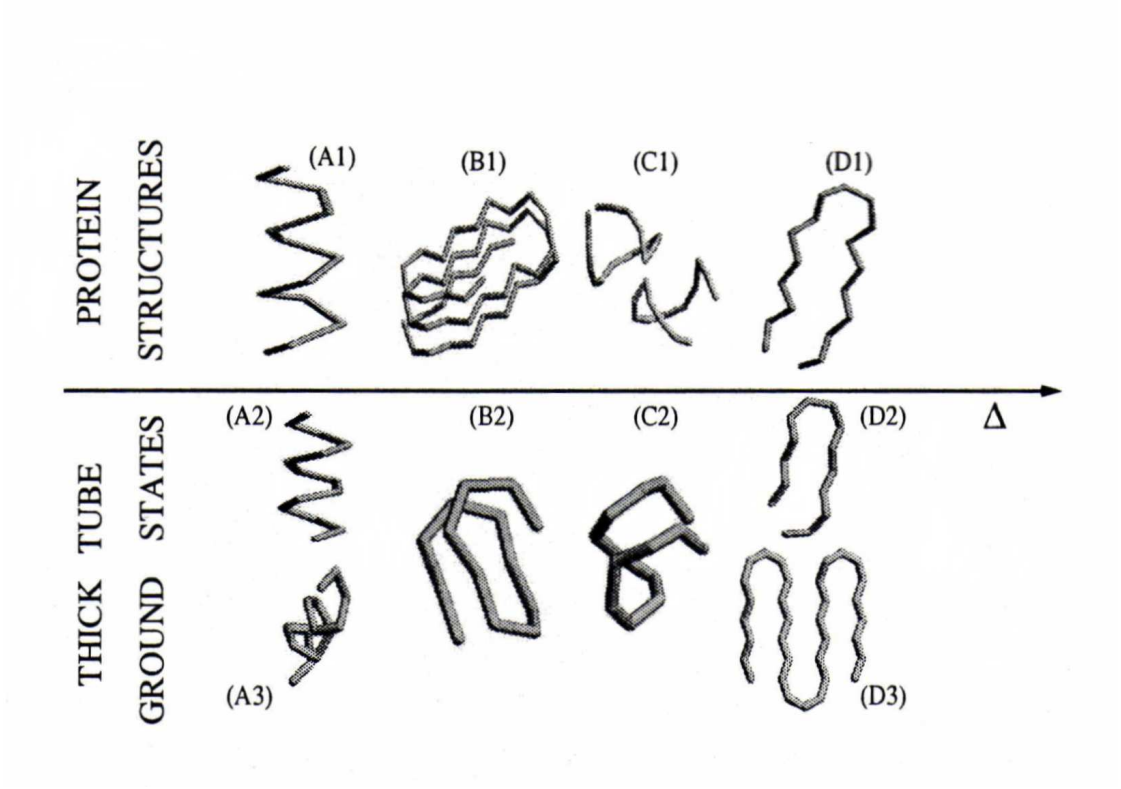


Figure 2.5: Comparison between building blocks of biomolecules (in the top row) and ground state structures associated with the marginally compact phase of a short tube (bottom). The Hamiltonian of the system is given by expression 2.16, where $b = 1$ and $R_0 = 0.55$. For the results shown here Δ ranges over the values corresponding to the marginally compact phase (the dashed region in Fig. 2.4), while R_1 is fixed to 1.6. The structure (A1) is an α-helix of a naturally occurring protein, while (A2) and (A3) are the helices obtained in the calculations for $\Delta \simeq 0.8$: (A2) has a regular contact map whereas (A3) is a distorted helix in which the distance between successive atoms along the helical axis is not constant but has period 2. Helices analogous to (A3) occur, e.g., in the HMG protein NHP6a [104] with pdb code 1CG7. The structure (B1) is a helix of strands found experimentally in Zinc metalloprotease [106] (pdb code: 1KAP), whereas (B2) shows the corresponding structure ($\Delta \simeq 0.88$) obtained in computer simulations. The structure (C1) shows the "kissing" hairpins of RNA [105] and (C2) the corresponding conformation obtained in simulations with $\Delta \simeq 0.95$. Finally (D1) and (D2) are two instances of quasi-planar hairpins. The former is found in many experimental structures while the second is a typical conformation found in simulations when $\Delta > 0.98$. The sheet-like structure (D3) is obtained for a longer tube.

of a tube upon compaction.

The work of Ramachandran has demonstrated that the excluded volume ef-

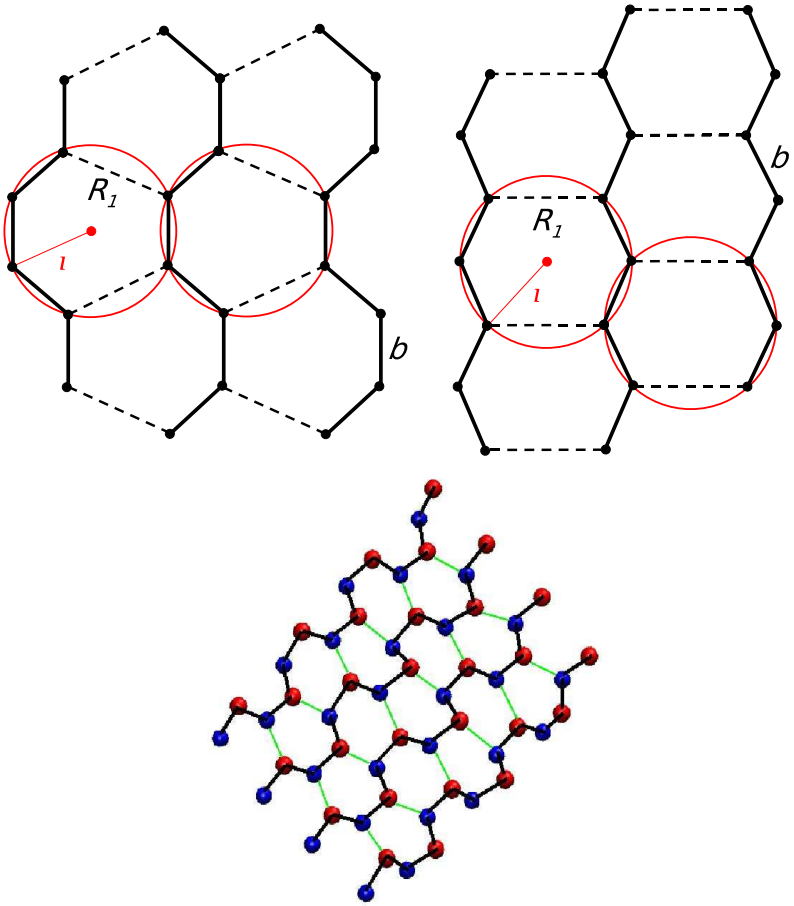


Figure 2.6: (a) Idealised sketch of the conformations found through numerical simulations with three-distinct chains [21]. The interaction among the monomers is given by the three-bodies potential V_3 of Eq. 2.14 and the square-wall potential V_2 of Eq. 2.17 acting between non-consecutive beads. We have fixed as usual $b = 1$, $R_0 = 0.55$ and $R_1 = 1.6$. The tubes was placed inside a hard-wall spherical box of radius 9.0; we have verified that the walls of the box do not influence the conformations shown. Two planar arrangements were found for $\Delta \simeq 1.10$, both of them forming the same number of contacts. These two conformations satisfy the property of ideal conformation, having the local radius of curvature equal of the non-local one. The appearance of planar zigzagging structures is a distinctive feature of the discrete nature of the chain. In the continuum limit the three chains are unable to aggregate in the case $\Delta > R_1/2$, because two points belonging to different chains can not come closer than 2Δ. Discrete chains otherwise, through these planar motifs are able to form contacts despite the large value of the thickness. It is possible to demonstrate analytically [18] that the perfect planarity is required in order to reach the best packed arrangement in the limit $\Delta \simeq 1.2124$. (b) Schematic representation of a parallel β-sheets. The resemblance between model and experiments is stronger, when identifying the interaction centres in the right with both the $N - H$ and $C - O$ bonds in the left, instead of the C_α atoms as usual.

fects due to the backbone atoms are able by themselves to provide a rationale for to the ubiquity of helices and sheets in protein native states. The rotational angles allowed, as we have stressed in section 1.3, are independent by the specific amino acid (with the exception of glycine and proline), which determines the general character of these important constraints. We can argue that, in order to construct a simplified model for the polypeptide chain, the thickness represents the key ingredient eligible to catch the steric effects in the proper manner.

2.2.2 Secondary motifs of the marginally compact phase

Slight changes were tested in the tube-like model described in the previous section [95]. Simulations were performed varying, for instance, the attractive potential in the Hamiltonian 2.16 (i.e. replacing the square-wall potential with a Lennard-Jones interaction), or considering longer chains. Anyway the tube behaviour has resulted strongly resistant to these variations. In particular the existence of the marginally compact phase seems to represent a peculiar characteristic of the thickness: it presents the same features and include more or less the same conformations for most of the model versions considered.

Let us now enter in details with the comparison between the conformations found in the marginally compact phase and the elementary motifs in biomolecules. A close similarity was found between the tube-like polymer ground states and the secondary structures in proteins; other marginally compact conformations moreover bear a qualitative resemblance to experimental structures of RNA and DNA.

A summary of the results obtained by the simulations, described in previous section, and the respective biopolymers motifs, is reported in Fig. 2.5. Increasing the value of Δ/R_1 first helices and then planar hairpins appear. In the middle other conformation occur such as saddles, irregular helix and helix made up of strands.

The helix depicted in Fig. 2.5 (A1), is the ideal helix described in section 2.1.3. As already mentioned it has the same pitch to radius ratio of the α-helix, and thus represents the structure that mostly fits the experimental observa-

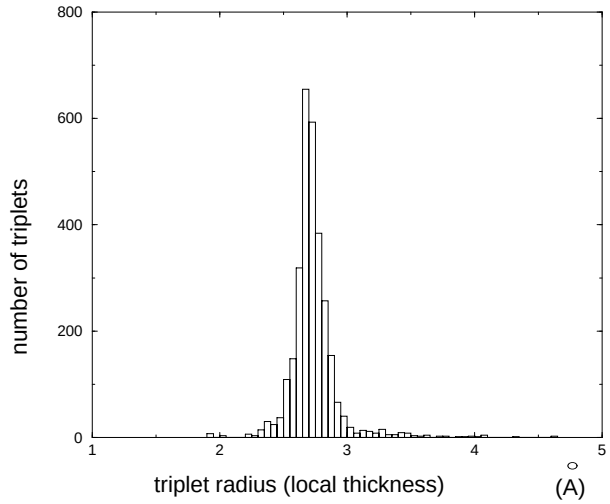


Figure 2.7: Distribution of the local thickness of the native structures of 30 proteins [96]. Reducing the polypeptide chain to a discrete chain of C_α atoms the global radius of curvature was computed for each monomer. The peaked distribution shows that it is a good approximation to consider a protein as a tube of uniform thickness of around 0.27 nm.

tions. The planar hairpin represents a noteworthy result too. Its strong similarity with the β-hairpin [5] has encouraged simulations with longer chains in order to seek for planar conformations able to reproduce β-sheets. The result is reported in Fig. 2.5 (D3), but extensively studies were done on a system composed with more short chains, since a long chain is more difficult to handle with numerical simulations[6]. In Fig. 2.6 we report the conformations obtained with three distinct chains: the tendency to form planar structures is strong in that particular range of parameters which is a remarkable result. These structures nevertheless can not be completely superimposed with β-sheets. The C_α atoms form in β-sheets a zigzag across the plane of the strands, whilst the planar structures of our tube, are characterised by a zigzag lying on the plane. The resemblance nevertheless is stressed by the fact that two different types of planar conformations were found. Each of them is characterised by a peculiar

[5] The β-hairpin consists of two strands that are adjacent in primary structure oriented in an antiparallel arrangement and linked by a short loop of two to five amino acids

[6] As described in appendix A a typical difficulty of the simulated annealing method is to explore extensively all the conformational space of the system without being trapped in locally stable states which does not represent the real ground state. The longer the chain the bigger this problem.

pattern of contacts that can be related respectively with the two arrangements in β-sheets, parallel and antiparallel.

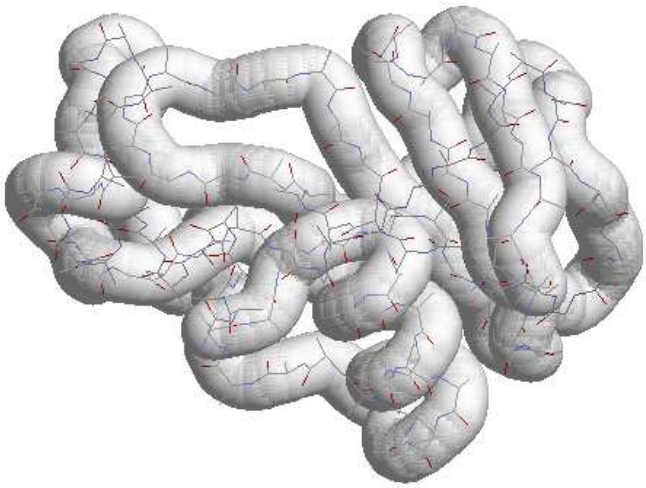


Figure 2.8: An all atoms representation of a segment of protein backbone (thin coloured lines) together with its corresponding backbone 'tube'. The tube is not inflated up to its maximum thickness for convenience of visualisation.

The ability of the tube-like polymer model to remind bio-molecules motifs is not restricted to proteins. The comparison between the saddle and the "kissing hairpin" of RNA is shown in Fig. 2.5 (C1) and (C2). Moreover, in our simulations, a double helix was frequently recovered as the ground state too. It presents the same property of optimal volume occupation of the single helix, being its local and non-local radius equal and minimum in each point of the chain. Analytical studies on the ideal double-helix continuous analogous have yielded the value of 2π for its pitch to radius ratio, which is within 4% of the value in the double-helix of DNA.

In order to have a physical estimation of the typical length scales in proteins, the thickness of the polypeptide chain was measured [96]. Many protein structures in the PDB were analysed, and the global radius of curvature defined by Eq. 2.11 was computed for each C_α atom, giving us the local thickness of the polymer. The histogram in Fig. 2.7 summarises the results and show that the protein backbone can be view as a tube with roughly uniform thickness of around 2.7 Å. A picture of this tube is reported in Fig. 2.8: the tube surrounds all the backbone atoms and also many side chains. Then evaluating the range of attraction between interacting residues, we can compare the experimental quantities with the parameters values corresponding to the marginally compact phase. We take as the physical counterpart of the square-wall potential, the peptide hydrogen bond, because it is the most important short range attraction affecting the protein. The typical distance between the C_α atoms of two residues involved in this interaction gives us a range of 6 Å. This demonstrates that the model matches the experiments not only in the qualitative resemblance among the respective motifs but also in the magnitude of physical quantities involved.

2.2.3 Pre-sculpted energy landscape of the protein folding process

In the marginally compact phase many conformations forming the building blocks of biomolecules naturally emerge as a product of a very basic model that does not include any sequence specificity. In the previous subsection we have analysed in detail the geometrical features of the tube-polymer ground states and we have compared them with experimental structures stored in the PDB. The good match found, anyway, is not the only virtue of the model: new insights in the comprehension of the protein folding issue can be obtained by a careful analysis of these results.

A remarkable aspect of this model is the strong reduction of degeneracy in the marginally compact phase with respect to the compact one. The thickness in this particular regime, near the edge of compactness, selects a few number of ground state conformations whose peculiar geometrical properties reflect

the main features of biopolymers motifs. This suggests that symmetry and geometry determine the limited menu of folded conformations that a protein can choose for its native state structure.

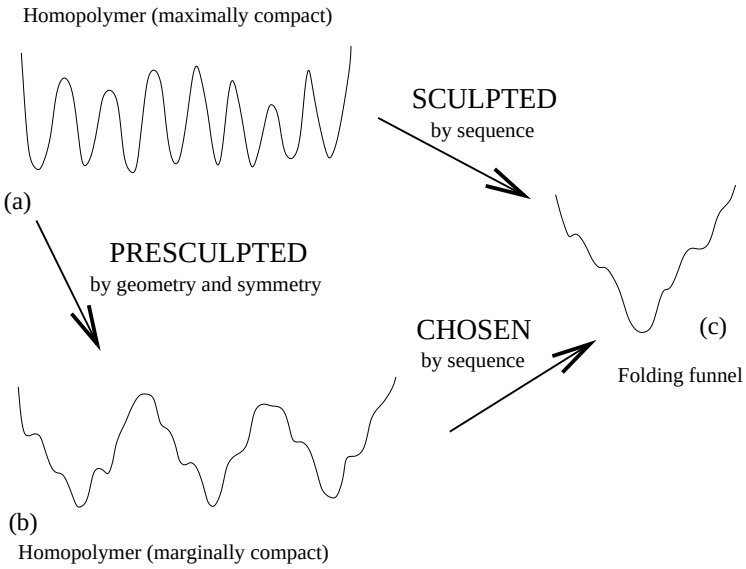


Figure 2.9: Simplified one dimensional sketches of energy landscape. (a) Rugged energy landscape for a string and bead homopolymer chain. (b) Pre-Sculpted energy landscape for a homopolymer chain in the marginally compact phase. The number of minima is greatly reduced and the width of their basin increased by the introduction of geometrical constraints. (c) Funnel energy landscape for a protein sequence. As folding proceeds from the top to the bottom of the funnel, its width, a measure of the entropy of the chain, decreases cooperatively with the energy gain. Such a distinctive feature, crucial for fast and reproducible folding, is proper of protein-like sequence according to the idea of Bryngelson and Wolynes. In the present view otherwise, funnel-like properties already result from considerations of geometry and symmetry in the marginally compact phase (b), making the role of the sequence the stabilisation of one of the pre-sculpted funnels.

This idea culminates in an innovative hypothesis [17, 20] able to give an explanation of the rapidity and high reproducibility characterising the folding process. Let us try for a while to describe the protein behaviour through a string and bead model. At low temperatures a string and bead homopolymer presents many distinct maximally compact ground state conformations with roughly the same energy, separated by high energy barriers[7]. Recalling section

[7]the degeneracy of ground state energies would be exact in the case of both lattice models

1.4.1, this situation can be depicted with a rugged energy landscape (Fig. 2.9a). The role played by sequence heterogeneity should then be to break the degeneracy of maximally compact conformations, leading to a unique ground state which, of course, depends on the amino acid sequence. Nevertheless, we have already observed (sect. 1.4.1) that for a typical random sequence, the energy landscape is still very rugged and it is virtually the same as in Fig.2.9a. Bryngelson and Wolynes [14] suggested that there is a principle of minimal frustration at work for well-designed sequences in which there is a nice fit between a given sequence and its native state structure carving out a funnel-like landscape [69, 70, 71] which promotes rapid folding and avoids the glassy behavior (Fig. 2.9c). In this scenario each protein is characterized by its own landscape. The protein sequence is all-important and the protein folding problem, besides becoming tremendously complex, needs to be attacked on a protein-by-protein basis.

In contrast, the tube-polymer model show a deep underlying unity in the proteins behaviour. Indeed the high reduction in degeneracy observed in our simulations hints that the gross features of the protein energy landscape could result from the amino acid unspecific common properties share by all proteins. We can imagine that the energy landscape of a homopolymer is *pre-sculpted* by general considerations of geometry and symmetry, and presents the profile sketched in Fig. 2.9b. It presents broad minima corresponding to putative native structures each of them characterised by a funnel-like behaviour. In this framework the role of the amino acid sequence is to favour the appropriate native state structure over the menu of predetermined ground states leading to an energy landscape conducive to rapid and reproducible folding of that particular protein.

The idea of pre-sculpted energy landscape is very appealing, as it enlightens the relation between primary structures and native state conformations. In particular a convincing explanation of the many-to-one character of this correspondence is given: among a menu of conformations predetermined by common geometrical requirements, each sequence selects the most suitable con-

and off-lattice models with discontinuous square-well potentials

formation for itself. It is hence recovered the parallelism with crystalline structures that has represented the starting hint of these studies.

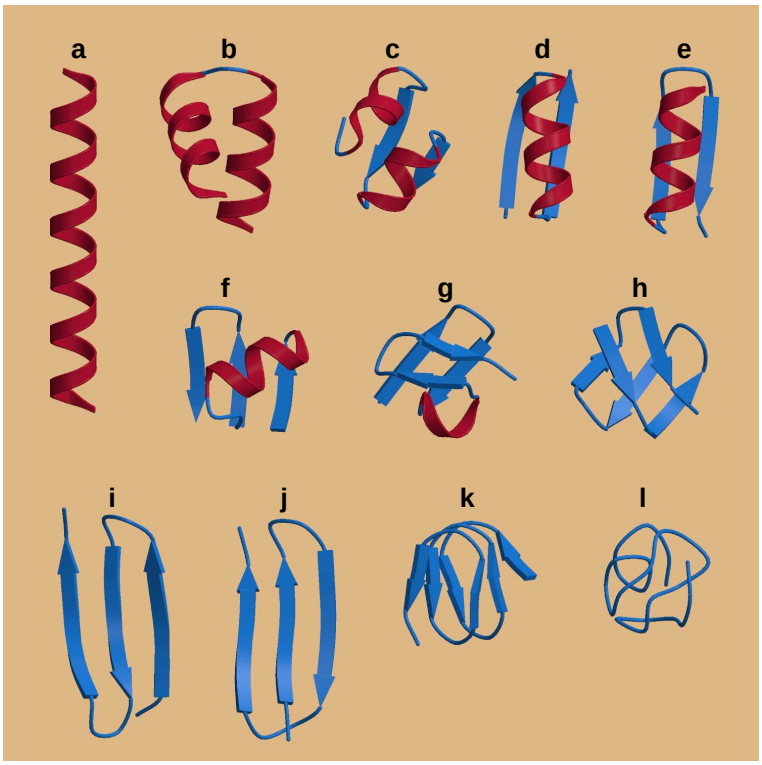


Figure 2.10: Overview of the minimum energy conformations obtained with numerical simulations of the refined model described in [17, 20]. The different conformations are obtained varying the parameters of the Hamiltonian. Most of them have their own counterpart in supersecondary structures of proteins. We refer the reader to [17, 20], for the model details.

A slightly more sophisticate model was developed in order to investigate this idea [20]. More ingredients were added to the basic tube-like model in order to make it a little bit more realistic, without renouncing anyway to its homopolymeric behaviour. The square-wall pairwise interaction is still present, whose role is to mimic the overall tendency towards compaction induced by the hydrophobic effect. A further short-range attraction is then considered; it has the peculiar property of having the interaction strength dependent on the orientation of the bond direction with respect to the two chain segments involved in the binding. This rule wants to catch the peculiar features of hydrogen bond. Moreover the thickness constraint was enhanced, stressing the

importance of local bending rigidity over the non-local effect. For the details of the model we refer to [17, 20], here we want to briefly outline the main result: through numerical simulations, many conformations were found as local minima of the energy function, in the case of long polymers ($N > 20$). They are very akin to super-secondary structures[8] since they are combinations of helices and sheets. Notice that with the basic version of the model only secondary structure motifs can be reproduced, while, enriching the homopolymer Hamiltonian with few ingredients, large portions of the protein structures are naturally obtained. The key point is that all the conformations found have comparable energy in the homopolymer case. But an amino acid sequence can be 'designed' in order to select each of these local minima as the ground state of the system, enhancing it in comparison with the others. In the present model only a distinction between amino acid hydrophobic and polar was considered according to the HP polymer, described in section 1.4.2. In summary, the simple geometrical and energetic ingredients of the model (i.e. hydrogen bond, steric constraints and a hydrophobic character assigned to the whole polymer) pre-sculpt the energy landscape, selecting through general principles a reduced number of structures. Well designed sequences are, in this framework, responsible of the stabilisation of the chosen pre-sculpted funnels.

2.3 Computation of the buried surface: a model for the hydrophobic effect

The extensive analysis done on the tube-polymer model have persuaded us that the thickness is the suitable tool in order to catch the steric hindrance of the protein backbone, in a qualitative simplified model. In this section we turn our attention to another mechanism that plays a crucial role in the folding process, the hydrophobic effect. Our purpose is to give a satisfactory modelling of this interaction in order to include its effects in our model [21, 22].

[8]The super-secondary structure represents a level of organization intermediate between secondary and tertiary structure. Super-secondary structure elements are combinations, recurring in many protein native states, of few secondary motifs.

The importance of hydrophobic effect in driving protein folding was theorised first by Kauzmann in the 1950s [23]. The details of the mechanism are still undisclosed, owing to the complexity of the water by itself. In principle, the non solubility of non-polar residues in aqueous solution determines their tendency to aggregate with each other, leading the protein to a rapid collapse. The behaviour of non-polar solutes in water is however quite peculiar. In fact, it is quite accepted [108, 109] that the origin of hydrophobic effect is essentially *entropic*. Water molecules prefer to form hydrogen bonds between themselves rather than 'waste' hydrogen bonds by pointing donor and acceptor groups towards the non-polar side chains. This determines the formation of an ordered cage of water molecules surrounding the non-polar side chains, which, with respect to bulk water, results in an enthalpy gain on one hand, but in an entropy loss on the other: at room temperature the latter overwhelms the former. The change of this delicate balance on decreasing the temperature gives rise to a decreasing of the hydrophobic interaction strength [24, 107], which is believed to be responsible for the *cold denaturation* transition which is observed for some proteins [108].

The complexity of this process requires detailed studies to be fully understood, since the peculiar geometry of water molecules has to be taken into account. Anyway, our focus is not the mechanism by itself rather its effects on the whole protein. It is a fact that, as already mentioned in section 1.2.3, the hydrophobic interaction does not determine a strict rule being high [41] the probability of finding a non-polar side chain on the protein surface and vice versa. Moreover, according to very recent works [110], the protein backbone should have an effectively hydrophobic character, because it prefers to form peptide hydrogen bonds instead of peptide-water ones[9]. For this reason, in the first instance, we can simplify the hydrophobic effect to an overall propensity of the protein molecule to bury as much surface as possible.

As a second step, we could consider a more refined version of the model. The side chains of amino acids stick out in a direction approximately opposite

[9]Whether the intra-chain hydrogen bond formation is responsible of the protein stability or on the contrary opposes it, has represented a matter of controversy during the half-past century and still remain not completely understood [66].

to the bending direction of the protein backbone [111] yielding an effectively *mixed hydrophobic* tube. In other words, certain parts of the hydrophobic backbone, determined by the instantaneous chain conformation, are already protected from the solvent by the side chains (we can consider them hydrophilic for simplicity) whereas the rest of the backbone needs to shield itself from the solvent by means of the compaction process.

In the following subsections we will illustrate the mathematical formalism for describing the surface of a tube centred on a continuous curve. Then, we will model the water as a collection of hard spheres, and we will introduce a formal rule for discerning if a portion of the tube surface can be touched by one sphere. The effective interaction describing both the uniformly hydrophobic tube and the mixed hydrophobic one, will descend then straightforwardly. We will conclude this section with the results of numerical simulations.

2.3.1 Mathematical description of the tube surface

In this section we briefly recall the basic mathematical expressions from differential geometry and local theory of curves [112], in order to describe the surface of a tube inflated around a continuous curve.

Let define a curve Γ, described by the arclength parameterization $\mathbf{R}(s)$. In dealing with such curved objects, it proves convenient to introduce suitable curvilinear coordinates to perform the calculations in a more efficient way; this is achieved by introducing a particular Frenet frame of unit vectors $\{\hat{\mathbf{T}}(s), \hat{\mathbf{N}}(s), \hat{\mathbf{B}}(s)\}$ for the tangent, normal and binormal respectively, as follows

$$\begin{aligned} \hat{\mathbf{T}}(s) &= \mathbf{R}'(s) \\ \hat{\mathbf{N}}(s) &= \frac{\hat{\mathbf{T}}'(s)}{|\hat{\mathbf{T}}'(s)|} \\ \hat{\mathbf{B}}(s) &= \hat{\mathbf{T}}(s) \times \hat{\mathbf{N}}(s), \end{aligned} \tag{2.18}$$

where the prime denotes the derivative with respect to the argument. In this

unit frame the Frenet-Serret equations are satisfied

$$\begin{aligned} \frac{\partial \hat{\mathbf{T}}(s)}{\partial s} &= \kappa(s)\,\hat{\mathbf{N}}(s) \\ \frac{\partial \hat{\mathbf{N}}(s)}{\partial s} &= -\kappa(s)\,\hat{\mathbf{T}}(s) + \tau(s)\,\hat{\mathbf{B}}(s) \\ \frac{\partial \hat{\mathbf{B}}(s)}{\partial s} &= -\tau(s)\,\hat{\mathbf{N}}(s), \end{aligned} \tag{2.19}$$

which automatically define the curvature $\kappa(s)$ and the torsion $\tau(s)$ from the first and the last equations of 2.19. Note that it is conventional to choose $\kappa(s)$ as positive by absorbing the sign in the direction of the normal vector $\hat{\mathbf{N}}(s)$.

Let us now impose the thickness constraint to the curve. Each tern of points belonging to the curve is hence subject to the potential energy given by Eq. 2.14. A generic point on the surface can then be described by the expression

$$\mathbf{r}_\Delta(s,\theta) = \mathbf{R}(s) + \Delta\left(\hat{\mathbf{N}}(s)\cos\theta + \hat{\mathbf{B}}(s)\sin\theta\right), \tag{2.20}$$

where θ is an azimuthal angle running from 0 to 2π. To compute the area of portions of the tube surface we can then integrate in the new coordinates (s,θ). The surface element is given by $J(s,\theta)\,dsd\theta$, where J is the Jacobian of the transformations $\mathbf{r}(s,\theta)$

$$J^2 \equiv \begin{vmatrix} (\partial_s \mathbf{r}_\Delta)^2 & \partial_s \mathbf{r}_\Delta \partial_\theta \mathbf{r}_\Delta \\ \partial_\theta \mathbf{r}_\Delta \partial_s \mathbf{r}_\Delta & (\partial_\theta \mathbf{r}_\Delta)^2 \end{vmatrix}. \tag{2.21}$$

The calculation of J^2 descend directly from the relations 2.19, obtaining

$$J^2(s,\theta) = \Delta^2(1-\Delta\kappa(s)\cos\theta)^2. \tag{2.22}$$

We remind that the thickness constraint determines a boundary on the local curvature radius, that from now on will be simply indicated by $R_c(s)$: $R_c(s) \equiv 1/\kappa(s) \geq \Delta$, $\forall s$, that implies $1-\Delta\kappa(s)\cos\theta \geq 0$, and

$$J(s,\theta) = \Delta(1-\Delta\kappa(s)\cos\theta). \tag{2.23}$$

The total area of the tube can then be computed by the integral

$$\Sigma(\Gamma) = \int_0^L \int_0^{2\pi} J(s,\theta)\,ds\,d\theta, \tag{2.24}$$

which give us the trivial result $2\pi\Delta L$.

2.3.2 Conformations adopted by a hydrophobic polymer

Having set up the mathematical apparatus, our aim is to compute the area of the buried portions of the tube, which can not be reached by the solvent, or vice versa the contact surface, composed by the tube regions that can be touched by the solvent particles. We model the solvent particles as spheres of radius ϵ. Then a point $\mathbf{r}_\Delta(s,\theta)$ on the tube surface can not be touched by the spherical particles if is satisfied the following inequality, which involves the point $\mathbf{r}_{\Delta+\epsilon}(s,\theta)$ on the surface of the tube with thickness $\Delta+\epsilon$, having the same centreline. The relation is

$$B_\epsilon(s,\theta) \equiv \min_{s'} |\mathbf{r}_{\Delta+\epsilon}(s,\theta) - \mathbf{R}(s')| < \Delta + \epsilon. \tag{2.25}$$

We can now introduce the Heaviside function

$$\vartheta(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x < 0 \end{cases}, \tag{2.26}$$

which helps us to write an expression for the buried area:

$$\Sigma_B(\Gamma) = \Delta \int_0^L \int_0^{2\pi} (1 - \Delta\kappa(s)\cos\theta) \cdot \vartheta\left[\Delta + \epsilon - B_\epsilon(s,\theta)\right] \, ds \, d\theta. \tag{2.27}$$

The mixed hydrophobic tube can be represented with two types of surface regions, each characterised by a different degree of hydrophobicity. Locally, the inner side of the tube, approximately in the direction of the normal vector, is taken to be hydrophobic, and the outer side hydrophilic. The following expression gives the effective interaction, which can describe both the completely hydrophobic and the mixed hydrophobic tube upon varying the parameter c,

$$\begin{aligned} V_{H.E.}(\Gamma) \;=\; & -\Delta \int_0^L \int_0^{2\pi} (1 - \Delta\kappa(s)\cos\theta) \cdot \{\vartheta\left[\Delta + \epsilon - B_\epsilon(s,\theta)\right] - \\ & -c\,\vartheta\left[\theta_1 - |\pi - \theta|\right] \cdot (1 - \vartheta\left[\Delta + \epsilon - B_\epsilon\right])\} \, d\theta \, ds \end{aligned} \tag{2.28}$$

where θ_1 defines half the angular width of a region centred around $\theta = \pi$ and c is a measure of the coupling between this region and the solvent. The region is hydrophilic when $c > 1$, while $c = 0$ corresponds to the case of uniformly hydrophobic tube.

The effective interaction described by Eq. 2.28 is more physically motivated than the simple square-wall potential considered in section 2.2, it moreover provides us a well defined energy function for a self-attracting curve, since it is free of singularity in the continuum limit. For this reason it represents an actual step-forward in our coarse grained modellisation of biopolymers.

Simulations, on short tubes subject to the potential energy 2.28, have been performed in order to study their minimum energy conformations. Once again the simulated annealing method (appendix A) is the most suitable tool for our purposes, but Monte Carlo at constant temperature has been also utilised in some cases. We considered a discretised representation with N segments separated by a distance $b = \Delta/2$. The continuum limit is obtained when $b << \Delta$; we have verified that our results are substantially the same on reducing the value of b down to $\Delta/3$. Also multiple chains systems were simulated; in this case, the tubes are placed inside a hard-wall cubic box of side 40Δ and we have verified that the walls of the box do not influence the conformations shown. The results are summarised in Fig. 2.11. The conformations (a-e) were obtained minimising the potential energy 2.28, with a single uniformly hydrophobic tube ($c = 0$) of various lengths and for different solvent molecule radius ϵ; the parameters values are reported in the caption of the figure. The conformations (f-h) have been obtained with multiple hydrophobic tubes ($c = 0$), using Monte Carlo simulations at constant temperature. Finally the conformations shown in figures 2.11 (i-l) represent the minimum energy states of a multiple chains system with a mixed hydrophobicity potential ($c = 5$). Two different solvent radius ϵ, of respectively 0.1Δ and 0.2Δ, were chosen.

The hydrophobic interaction, that in this coarse grained model is simply an energy gain proportional to the area of both the hydrophobic buried regions and the hydrophilic exposed ones (if $c > 1$), selects conformations akin to biomolecules motifs: indeed secondary structures of proteins and double helix of DNA were found. Therefore we should conclude that the solvent radius in this range of values tunes the system in the marginally compact regime. It is noteworthy that, also in the continuum limit, structures really close to β-sheets are obtained by the effect of a mixed hydrophobic interaction. Notice that the planar conformations shown in Fig. 2.11 presents a zigzag crossing

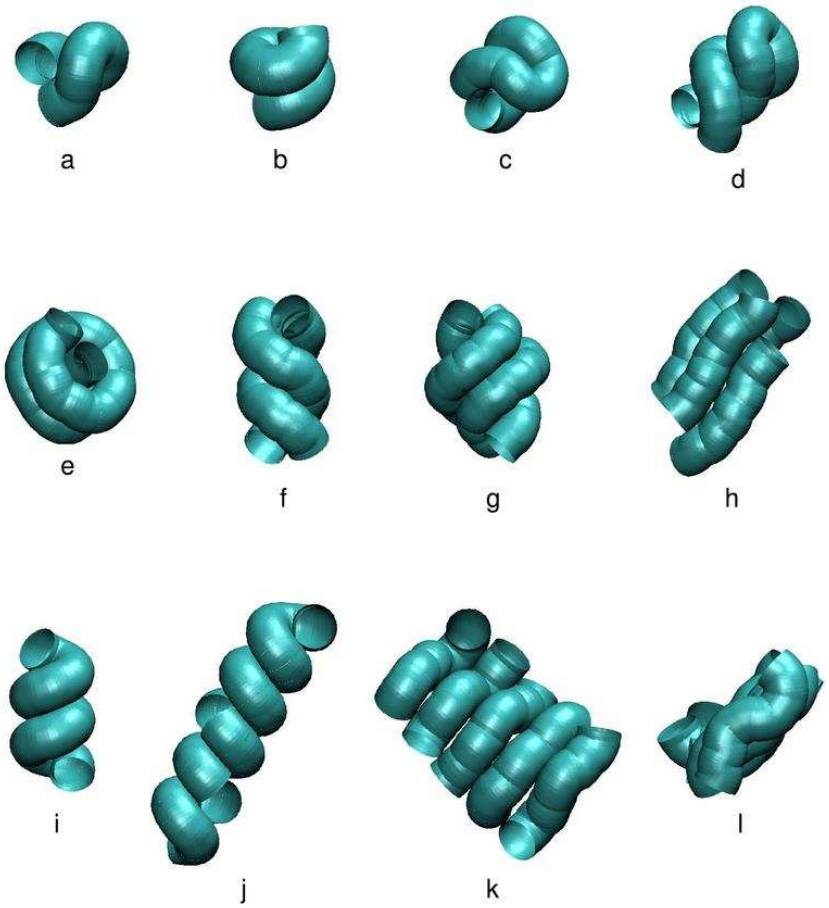


Figure 2.11: Results of simulation with one or more tubes subject to the effective energy described by Eq. 2.28. The conformations are obtained using Metropolis Monte Carlo simulations by annealing or by long simulations at constant temperature. (a-e) Conformations of single hydrophobic tubes ($c = 0$): a) $N = 20$ and $\epsilon = 0.1\Delta$, b) $N = 20$ and $\epsilon = 0.2\Delta$, c) $N = 30$ and $\epsilon = 0.1\Delta$, d) $N = 40$ and $\epsilon = 0.1\Delta$, e) $N = 50$ and $\epsilon = 0.2\Delta$. (f-h) Optimal conformations of multiple hydrophobic tubes ($c = 0$) of length $N = 20$ and for $\epsilon = 0.1\Delta$ obtained in long simulations at constant temperatures: f) two tubes at a low temperature, g) three tubes at a low temperature, h) four tubes at an intermediate temperature. (i-l) Conformations of mixed hydrophobicity tubes ($c = 5$); $\epsilon = \Delta/2$ for all cases. i) A single helix of length $N = 30$ and $\theta_1 = 15^o$ obtained by slow annealing (one obtains the same conformation for $\theta_1 = 30^o$ or 45^o). j) A stack of 4 helices of length $N = 15$ and $\theta_1 = 45^o$ obtained by slow annealing. k-l) Two views of a planar sheet arrangement of 5 chains of length $N = 15$ and $\theta_1 = 30^o$ obtained in a constant temperature simulation.

perpendicularly the sheet plane like as β-sheets when only the C_α atoms are considered.

Chapter 3

Depletion interaction for a tube-like polymer in a solvent

The interaction between a polymer and the physiological solution that surrounds it, is the result of different physical processes. In the previous chapter we have studied an attractive potential able to model in a very simple way the hydrophobic effect for a tube-like polymer immersed in a solvent of hard spheres. But despite the importance of the hydrophobic effect other forces can play a crucial role in determining the strong tendency towards compaction that affects protein folding.

In this chapter we want to analyse a purely entropic effect that characterises the crowded systems such as microemulsion and colloidal solutions: the tendency to maximise the entropy is the origin of *depletion* forces that cause *ordering* phase transitions, transitions from an apparently less ordered to an apparently more ordered state, such as flocculation and other agglomeration phenomena. The suggestion by recent works [25, 113] that a similar effect can determine a collapse phase transition for a single polymer in solution, justifies our interest in this mechanism.

The existence of ordering transitions has been known for a long time [27, 28, 114]. Possibly the first example was discussed by Onsager [115] who showed that a fluid of thin hard rods, of lengths much bigger that their diameters, undergoes an isotropic-nematic transition from a disordered (isotropic) to a more ordered (nematic) state, upon compression. This is because the loss in ori-

entational entropy is overwhelmed by the gain in translational entropy. The flocculation in colloidal solution is a similar effect. In the experimental studies of colloidal systems, where macromolecules are dispersed in a solution of non-absorbing polymers, it was observed experimentally that a small change in polymers concentration induces an effective attraction between the colloidal particles leading to a phase transition. This fact was explained by the two independent works of Asakura and Oosawa [29, 30], and Vrij [116]. They considered a system composed by two large spherical particles immersed in a solution of small ones. In this simple picture it is easy to show how the tendency of the solvent to maximise its entropy induces a short range attraction between the large spheres, since in this manner *they decrease their excluded volume*.

Essentially all applications of the entropically driven phase transitions carried out in the past were on the effective interactions between degrees of freedom of large bodies mediated by the presence of small ones. However it was recently proposed [25] that a similar mechanism could drive effective interactions between different parts of a *single* polymer thus favouring the transition from an extended to a compact configuration. Adapting the Asakura-Oosawa scenario to the case of a tube-polymer, Snir and Kamien in their papers [25, 113], inspired by the results of Maritan *et al.* [16, 17], suggested that the tube might spontaneously bend and twist to form a helical configuration, in order to minimise its excluded volume. This could explain the ubiquity of helix formation in nature since virtually all biological activities occur into an environment crowded by colloidal almost spherical particles. While it is undeniable the appeal of this simple scenario, in the passage from spheres to tubes an important issue should be addressed. The Asakura-Oosawa-Vrij mechanism favours the increase of the *solvent* entropy and not the total entropy. While in the case of spheres the difference is irrelevant, as spheres haven't inner degrees of freedom, the same mechanism applied to the tube would lead to a collapsed phase of the tube, thus yielding a corresponding reduction of the entropy of the polymer. Therefore, in order to drive the tube into a compact configuration by pure entropic forces, the reduction in the tube entropy should not exceed the gain in solvent entropy.

In this chapter after a brief introduction to the Asakura-Oosawa-Vrij model

(section 3.1) we shall study its generalisation [22] to the case of a tube-polymer immersed in a solution of spherical particles. We provide an analytical formula to compute the excluded volume of a tube in every configuration. In section 3.2 we will introduce this formula illustrating with simple examples same issues arising from it. The section 3.3 will focus on the analysis of a tube in a helical configuration: computing the tube excluded volume as a function of the helix pitch to radius ratio for different solvent diameters we looked for the 'optimal helix', able to minimise its excluded volume. The results we obtained are in complete agreement with those reported in previous works [25, 117]. To conclude this study we performed numerical simulations in order to find, without any a priori assumption, which is the configuration of a tube that minimises its excluded volume. The simulated annealing is again the most suitable method for our purposes. We will report the results in section 3.4.

3.1 The Asakura-Oosawa-Vrij model

Experiments early in the last century showed that colloidal systems composed of dispersions of macromolecules and non-absorbing polymers exhibit interesting agglomeration phenomena. A series of experiments, performed mainly by groups in Utrecht, Bristol and Edinburgh in the 1980s and early 1990s [118, 119, 120], confirmed that adding sufficient non-absorbing polymers to a suspension of colloids can cause phase separation into two fluid phases, one rich in colloid and the other poor.

It is well established that certain colloidal suspensions behaves as hard spheres. This limit can be reached experimentally, for example, by dressing particles in layers that reduce the Van der Waals attractive force and the coulombic interactions that otherwise would affect the system. Polymers too can be approximated to hard spheres, because when they are immersed in a good solvent they are in a swollen configuration, and they interact essentially as inert impenetrable objects, whose shape can considered, in first approximation, spherical given their small dimensions with respect to the colloids.

In the absence of attracting interactions the origin of these aggregation phe-

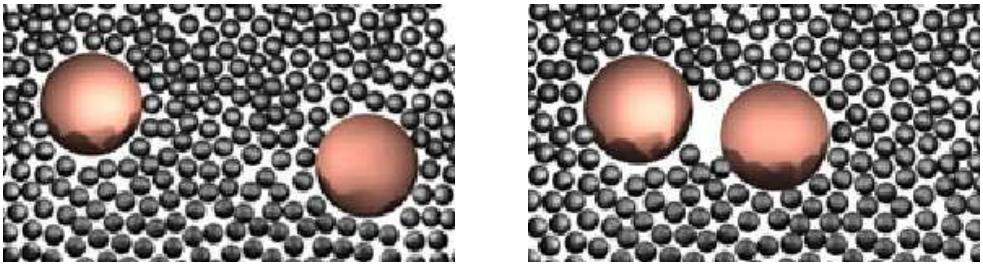
nomena is then quite mysterious and indeed it wasn't understood until the paper of Asakura and Oosawa and the independent studies of Vrij [29, 30, 116]. Their paradigm is a simple system composed of a mixture of large and small hard spheres. As a first approximation they assume that polymers are *ideal*, meaning that they behaves as small hard-spheres when interacting with a large hard-sphere but penetrable when interacting with a small one. The interaction potential of the system is then given by

$$\begin{aligned} \mathcal{H}(\mathbf{R}_1...\mathbf{R}_{N_L}, \mathbf{r}_1...\mathbf{r}_{N_s}) &= \mathcal{H}_L(\mathbf{R}_1...\mathbf{R}_{N_L}) + \mathcal{H}_{L,s}(\mathbf{R}_1...\mathbf{R}_{N_L}, \mathbf{r}_1...\mathbf{r}_{N_s}) \\ &= \sum_{a<b} \mathcal{V}_L(\mathbf{R}_a, \mathbf{R}_b) + \sum_{a,i} \mathcal{V}_{s,L}(\mathbf{R}_a, \mathbf{r}_i), \end{aligned} \tag{3.1}$$

where, as a consequence of the ideal polymer condition, the solvent-solvent interaction, $\mathcal{V}_s(\mathbf{r}_i, \mathbf{r}_j)$, is simply identically zero, and the otherinteraction potentials are

$$\begin{aligned} \mathcal{V}_L(\mathbf{R}_a, \mathbf{R}_b) &= \begin{cases} +\infty & \text{for } |\mathbf{R}_a - \mathbf{R}_b| < d_L \\ 0 & \text{otherwise} \end{cases} \\ \mathcal{V}_{s,L}(\mathbf{R}_a, \mathbf{r}_i) &= \begin{cases} +\infty & \text{for } |\mathbf{R}_a - \mathbf{r}_i| < \frac{1}{2}(d_L + d_s) \\ 0 & \text{otherwise} \end{cases}, \end{aligned} \tag{3.2}$$

here N is the particles number and d their diameter, the subscripts s and L refer respectively to small and large spheres.



(a) (b)

Figure 3.1: Representation of the depletion effect. When the two large spheres are distant (Fig. a) they are subject to an isotropic pressure. If, otherwise, the two spheres are close together (Fig. b), the depletion between them produces a net pressure pushing the large spheres one toward the other.

Asakura, Oosawa and Vrij proved that in this extremely simple system an effective attractive interaction between the large particle can be induced. An intuitive picture can explain the origin of this attraction and is shown in Fig. 3.1. When the separation between two large particles is sufficiently large the distribution of small particles around each big particle is uniform with spherical symmetry. Otherwise, when the two colloids are close one to the other the density of small particle in the space between them decreases, this reduction is referred to as *depletion*. The difference in density produces an unbalanced osmotic pressure pushing the large spheres one toward the other.

Another explanation of the phenomenon arises from purely entropic consideration. The excluded volume to the centres of small particles due to the presence of a colloid is a sphere of diameter $d_L + d_s$. If two colloids a and b come one close to the other their excluded volumes overlap increasing the free volume for the polymers and hence their entropy. Given $R = |\mathbf{R}_a - \mathbf{R}_b|$, the difference in free energy $\Delta F(R) = F(R) - F(R \to \infty)$ is a simple function of the total excluded volume, $V_{exc}(R)$, and is easily computed in this framework (see e.g. [121]). Starting with the canonical partition function of the system

$$\mathcal{Z}_{N_s}(T,V) = \int_V \frac{d\mathbf{r}^{N_s}}{N_s!\Lambda_s^{3N_s}} \frac{d\mathbf{R}_1\, d\mathbf{R}_2}{2\Lambda_L^6}\, e^{-\beta(\mathcal{H}_L+\mathcal{H}_{L,s})}, \tag{3.3}$$

(where Λ is the thermal de Broglie wavelength) and assuming the condition $d_s << d_L$ is satisfied, we can integrate out the microscopic degrees of freedom associated to the small spheres, obtaining

$$\mathcal{Z}_{N_s}(T,V) \;\; = \;\; \int_V \frac{d\mathbf{R}_1\, d\mathbf{R}_2}{2\Lambda_L^6}\, e^{-\beta(\mathcal{H}_L+F(R))}, \tag{3.4}$$

with

$$e^{-\beta F(R)} = \frac{1}{N_s!}\left(\frac{V - V_{exc}(R)}{\Lambda_s^3}\right)^{N_s}. \tag{3.5}$$

This gives us, after few approximations

$$\Delta F(R) = -N_s kT \frac{V_{exc,\infty} - V_{exc}(R)}{V - V_{exc,\infty}}, \tag{3.6}$$

where $V_{exc,\infty} = \frac{\pi}{3}(d_L+d_s)^3$, is twice the excluded region of one sphere. The idea is illustrated in Fig. 3.2: when the colloids come close each other, the excluded

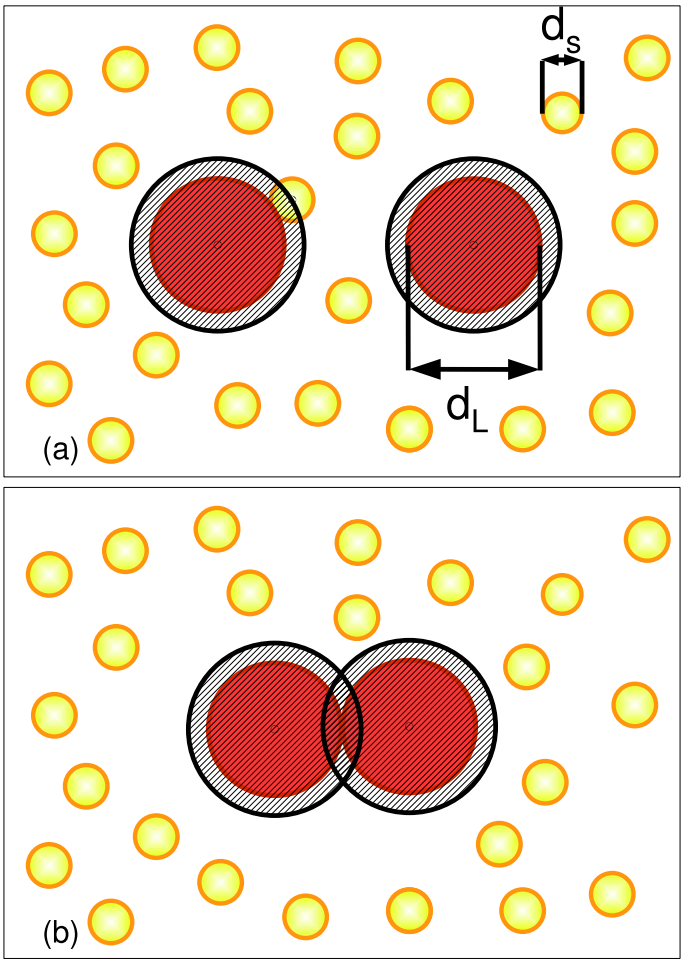


Figure 3.2: Schematic picture of the Asakura Oosawa model. The volume forbidden to the small sphere centres (coiled polymer) due to the colloid presence is a sphere of diameter $d_L + d_s$, see the two shaded regions in Fig. a. When the two large spheres come close one to the other the overlap of the two excluded volumes (the lens region in Fig. b) increases the free volume for the small particle.

volume has decreased by an amount $V_{exc,\infty} - V_{exc}(R)$, and hence the available volume to small particles has increased by the same amount, thus leading to an increase of the polymer entropy which is measured by $F(R) - F(R \to \infty)$. This gives rise to an effective attractive potential, $W(R) = \Delta F(R)$, between the

two spheres which to the leading order reads

$$W(R) = n_S kT \left(V_{exc}(R) - V_{exc,\infty}\right) = -n_S kT V_{Ov}(R), \tag{3.7}$$

where $n_S = N_S/V \approx N_S/(V - V_{exc,\infty})$, and $V_{Ov}(R)$ is the overlap volume, the volume of the lens region in Fig. 3.2. As we are going to see, the relation $V_{Ov}(R) = V_{exc,\infty} - V_{exc}(R)$, that is trivial in this case, does not hold anymore in the case of a self-interacting polymer. The depth of the potential wall and its range are, in this simple approximation, regulated respectively by the concentration and the size of polymers.

More detailed studies followed these simple calculations. In order to understand how this interaction affects a system composed by more than two colloids, the partition function in Eq. 3.3 can be evaluated using the cluster expansion [122]. We obtain then a many-body attractive potential with the two-body term of Eq. 3.7 representing the first order approximation.

3.2 An Asakura-Oosawa model for a tube-like polymer

The effects of crowding on protein folding, and on biological processes in general, represent a current subject of study. In particular it was recently proposed [62, 123] that depletion forces could affect protein folding and many other phenomena occurring inside the cell, such as amyloid formation, stability of actin and genome looping. In these works the role of crowding agents is covered by macromolecules such as proteins, whose diameter is of the order of $10 - 20$ Å. However the importance of depletion interaction on a finer scale too, was recently recognised [25, 26]. The tendency of water to gain translational entropy gives rise to a force, which, beside the highly specific interactions of steric and chemical nature, is omnipresent in biological systems. This entropic effect could play an important role in driving the protein to fold in its native conformation. In particular, it was suggested [25, 26] that as a consequence of this phenomenon secondary structures emerge, being, among the compact conformations, the ones eligible to maximising solvent translational entropy.

In order to exclusively investigate this entropic effect we will model the solvent molecules as hard spheres, we will neglect all the soft interaction (e.g. electrostatic and Van der Waals interactions), and we will study the simplest approximation represented by the Asakura-Oosawa-Vrij scenario, following in this way the suggestion of Snir and Kamien [25, 113]. In place of large spheres we will consider a single tube-like polymer and we will investigate the possibility of secondary structures formation in this framework. Actually the Asakura-Oosawa model represents a very rough approximation and more refined techniques (such as integral equation theory) were employed in order to study this effect [26]. However it is quite astonishing the ability of this basic model to account for secondary motifs formation, as it will be shown by the numerical analysis we are going to illustrate.

Originally the Asakura-Oosawa model involved an effective interaction between distinct large bodies mediated by the presence of small ones. Therefore the innovation of Snir and Kamien idea consists in considering the effective interactions among different parts of a single body (a polymer in this case) thereby inducing a conformational change. Hence the competition between the two entropic components, due to the solvent and to the polymer inner degrees of freedom, have to be taken into account in order to study the equilibrium configuration. Anyway, the generalisation of the Asakura-Oosawa picture to the tube-polymer case is quite straightforward, despite this non trivial issue. Consider a solvent of hard spheres with radius ϵ and an impenetrable-tube polymer with thickness Δ and length L, if the tube is in a straight configuration the region forbidden to the small spheres centres is a tube of thickness $\Delta+\epsilon$, so the excluded volume is $V_{exc,S} = \pi(\Delta+\epsilon)^2 L$ – see Fig. 3.3 for a schematic representation. When otherwise the tube is in a compact configuration different parts of the excluded region may overlap, so that in such a case the total excluded volume decreases determining an increase in the solvent entropy. This phenomenon produces an effective self-attraction between different parts of the polymer, that could drive the polymer towards a compact configuration.

Consider a curve, Γ, with arclength parametrisation $R(s)$. Reminding section 2.2, it represents the centreline of an impenetrable tube when it is subject

to the Hamiltonian $\mathcal{H}_T(\Gamma)$:

$$\mathcal{H}_T\left(\Gamma\right) = \int_0^L \int_0^L \int_0^L \mathcal{V}_3\left[\mathbf{R}(s_1), \mathbf{R}(s_2), \mathbf{R}(s_3)\right] \, d\Gamma_{s_1} d\Gamma_{s_2} d\Gamma_{s_3}, \tag{3.8}$$

where $d\Gamma_s$ is the element of arc length, and $V_3(\mathbf{x}, \mathbf{y}, \mathbf{z})$, is the three-bodies hard core potential given in Eq. 2.14. Recalling Eq. 3.7 the depletion self-interaction, has then the expression,

$$W\left(\Gamma\right) = n_S kT \left(V_{exc}\left(\Gamma\right) - V_{exc,S}\right). \tag{3.9}$$

Notice that the balance between this attractive interaction and the polymer entropy is regulated by the partition function

$$\mathcal{Z}(T, n_s) \propto \sum_{\{\Gamma\}} e^{-\beta \mathcal{H}_T(\Gamma) - n_S\left(V_{exc}(\Gamma) - V_{exc,S}\right)}, \tag{3.10}$$

where we can see again that the strength of the tendency towards compaction is determined by the solvent density.

A first step of the analysis is the search of the tube *optimal conformations*, able to minimise its excluded volume. Snir and Kamien in their paper studied only helical conformations, computing numerically their excluded volume as a function of the helix pitch to radius ratio. In this manner they determined the optimal helix, as a function of the spheres radius. The assumption that the helix, compared with all possible compact conformations of a tube, is the configuration of minimum excluded volume, is reasonable and motivated by the studies of the ideal packing of a tube reported in chapter 2. It necessitates nevertheless a confirmation, i.e. the minimisation with any a priori assumption of the functional given in Eq. 3.9. A general formula to compute $V_{exc}\left(\Gamma\right)$ is our fundamental contribution to the topic [22], and will be the subject of this section. It presents, with respect to the Asakura-Oosawa case, innovative aspects that we are going to illustrate by means of some examples.

It has to be reminded that, in the present case, the competition between depletion effects and polymer entropy or other polymer self-interactions will be neglected.

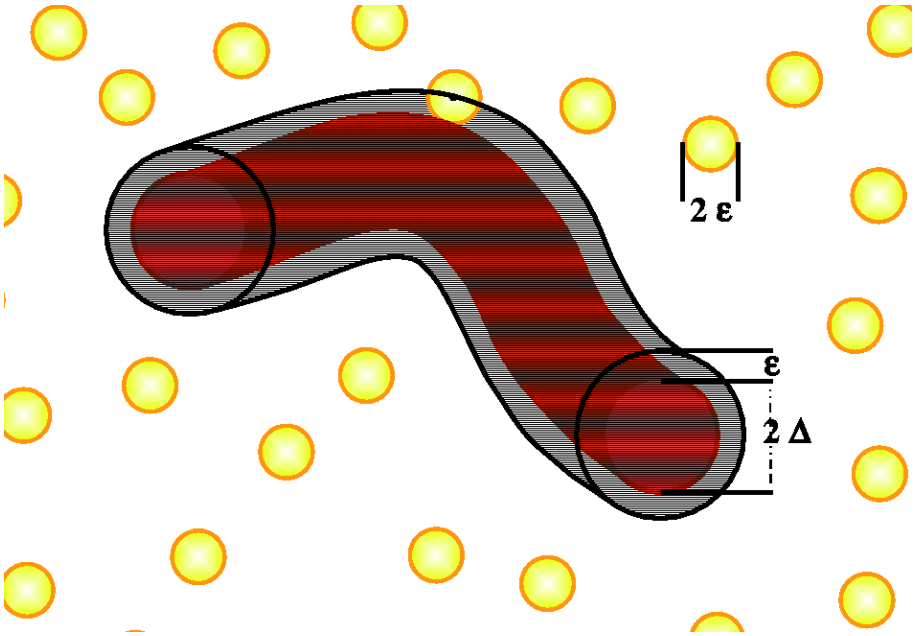


Figure 3.3: Sketch of an impenetrable tube immersed in a solvent of hard spheres. The excluded volume to the small spheres centres is a 'soft' tube of thickness $\Delta + \epsilon$, where Δ is the tube thickness and ϵ the solvent radius.

3.2.1 Computation of the excluded volume

Consider an impenetrable tube with thickness Δ, centreline Γ, and parameterisation $\mathbf{R}(s)$. The excluded region to the spheres centres determined by its presence is a tube with the same centreline and thickness $\Delta_\epsilon = \Delta + \epsilon$, characterised by the fact that the external layer large ϵ is 'soft': it is able to compenetrate with the other parts of the tube. A point $\mathbf{r}$ inside this region can be identified by the transformation of coordinates

$$\mathbf{r}\,(s,\theta,\rho) \;=\; \mathbf{R}(s) + \rho\left[\cos\theta\;\hat{\mathbf{N}}(s) + \sin\theta\;\hat{\mathbf{B}}(s)\right], \tag{3.11}$$

defined in the domain $\mathcal{D} = [0, L]\times[0, 2\pi]\times[0, \Delta_\epsilon]$. It represents the analogous of Eq. 2.20, with the only difference of the coordinate ρ, the radial distance from the centreline, that in the buried area computations was fixed to Δ. Volume calculations in the new set of coordinates, requires the Jacobian of $\mathbf{r}\,(s,\theta,\rho)$, that results

$$J\,(s,\theta,\rho) \;=\; \rho|1 - \kappa\,(s)\,\rho\cos\theta| = \rho\,|\chi\,(s,\theta,\rho)|\,, \tag{3.12}$$

where $k(s)$ is the curvature of Γ. We have renamed the argument of the above absolute value since it will play an important role in our discussion.

If the tube configuration is such that there is no overlapping among different parts of the excluded region (in the following we will call swollen all such conformations), its volume can be computed simply by the integral

$$V_{exc,S}(\Gamma) \;=\; \int_{\mathcal{D}} \rho\,|\chi\,(s,\theta,\rho)|\;\, d\rho\; d\theta\; ds. \tag{3.13}$$

that yields $\pi\Delta_\epsilon^2 L$ independently of the swollen configuration we are considering: since a centreline region with $k(s) > \Delta_\epsilon^{-1}$ produces an overlap, swollen configurations have to satisfy the condition $k(s) < \Delta_\epsilon^{-1}$ for every s, that implies $\chi\,(s,\rho,\theta) > 0$ in all the domain, the integration becomes then elementary.

The expression 3.13 is not correct anymore when the tube is in a compact configuration and hence there are some overlapping parts of the excluded region, since in this case the coordinate transformation, $\mathbf{r}(s,\theta,\rho)$, is not a bijective map. To see how we can compute the volume in this case we argue as follows. Take a point $\mathbf{r}(s,\theta,\rho)$ belonging to the overlap region. This means that it belongs to more than one circular section associated to different points of the tube axis. It implies that integrating in all the domain $\mathcal{D}$, we add many times to the total amount the same volume element. In order to correct this overcounting it is sufficient to include in Eq. 3.13 the correct weight for each $\mathbf{r}(s,\theta,\rho)$. Defining $n(s,\theta,\rho)$ to be the number of sections $\mathbf{r}(s,\theta,\rho)$ belong to, then this function provides the weight we are looking for. Noting that $n(s,\theta,\rho) \geq 1$ since at the very least the point belongs to one section, the excluded volume of this compact phase is then given by

$$V_{exc,C} \;=\; \int_{\mathcal{D}} \rho\,|\chi(s,\theta,\rho)|\,\frac{1}{n(s,\theta,\rho)}\; d\rho\; d\theta\; ds. \tag{3.14}$$

In the absence of overlap, $n(s,\theta,\rho) \equiv 1$ and $\chi(s,\theta,\rho) > 0$ in all $\mathcal{D}$ and Eq. 3.14 reduces to the swollen result.

Let us now explain how to evaluate the multiplicity of intersections, $n(s,\theta,\rho)$. The point $\mathbf{r}(s,\theta,\rho)$ belongs to the tube section centred in a given centreline point $\mathbf{R}(s')$ if the distance vector $\mathbf{d}_{s,\theta,\rho}(s') = \mathbf{r}(s,\theta,\rho) - \mathbf{R}(s')$ is perpendicular to $\hat{\mathbf{T}}(s')$ and has magnitude $d_{s,\theta,\rho}(s')$ less than Δ_ϵ. Note that the first condition

is equivalent to the extremalisation requirement $\partial d_{s,\theta,\rho}(s')/\partial s' = 0$. Hence we can define $n(s,\theta,\rho)$ as the number of $s' \in [0, L]$ that satisfies the two conditions

$$\frac{\partial d_{s,\theta,\rho}\,(s')}{\partial s'} = 0 \qquad\qquad d_{s,\theta,\rho}\,(s') \leq \Delta_\epsilon. \tag{3.15}$$

Following this mathematical recipe the excluded volume computation of a tube in every configuration is possible. With the exception of same simple cases, nevertheless, this formula is very difficult to handle in analytical computations, but it is rather suitable for numerical implementation.

In order to better compare the expression for the depletion interaction among a tube polymer (Eq. 3.9), with the Asakura-Oosawa case, it is convenient to rewrite Eq. 3.14 in a more transparent way, by introducing also in this case the overlap volume, $V_{ov}(\Gamma)$, defined as the volume of the soft tube parts in which the excluded region overlaps with itself, that is parts such that $n(s,\theta,\rho) > 1$. Let us therefore evaluate the difference in excluded volume between compact and swollen configuration

$$\begin{aligned} V_{exc,C}(\Gamma) - V_{exc,S} \;\; &= \;\; \int_{\mathcal{D}} \rho|\chi(\rho,\theta,s)|\frac{1}{n\,(s,\rho,\theta)}\; d\rho\; d\theta\; ds\; - \\ &\quad - \int_{\mathcal{D}} \rho\chi(\rho,\theta,s)\; d\rho\; d\theta\; ds. \end{aligned} \tag{3.16}$$

We recover now the Heaviside function, $\vartheta(x)$, defined by Eq. 2.26, and we observe that the following identity is satisfied

$$|x| - x = 2\,|x|\,\vartheta\,(-x)\,. \tag{3.17}$$

This helps us to rewrite

$$\begin{aligned} V_{exc,C}(\Gamma) - V_{exc,S} \;\; &= \;\; -\int_{\mathcal{D}} \rho|\chi(\rho,\theta,s)|\frac{n\,(\rho,\theta,s) - 1}{n\,(\rho,\theta,s)}\; d\rho\; d\theta\; ds\; + \\ &\quad +2\int_{\mathcal{D}} \rho|\chi\,(\rho,\theta,s)\,|\,\vartheta\,[-\chi\,(\rho,\theta,s)]\,, \end{aligned} \tag{3.18}$$

obtaining a meaningful rearrangement of Eq. 3.16: the first term of second side, indeed, represents the overlap volume, while the second integral yields a contribution associated with highly curved regions of the centreline, i.e. with

$k(s) > \Delta_{\epsilon}^{-1}$. Hence defining

$$V_{Ov} = \int_{\mathcal{D}} \rho|\chi(\rho,\theta,s)|\frac{n(\rho,\theta,s)-1}{n(\rho,\theta,s)}\, d\rho\, d\theta\, ds \tag{3.19}$$

$$V_{\chi<0} = \int_{\mathcal{D}} \rho|\chi(\rho,\theta,s)|\,\vartheta\left[-\chi(\rho,\theta,s)\right], \tag{3.20}$$

we can summarise

$$\Delta V(\Gamma) = V_{exc,C}(\Gamma) - V_{exc,S} = -V_{ov} + 2V_{\chi<0}. \tag{3.21}$$

Therefore we have found that the difference between the excluded volumes in the swollen and in the compact phase, is *not* simply equal to V_{Ov} as in the Asakura-Oosawa case. This difference stems by the fact that whilst the two spheres are rigid, the polymer undergoes a conformational change upon folding. As a consequence of this fact a compact conformation is not always advantageous with respect to the swollen one: the overlap decreases the excluded volume, but on the other hand the regions with high curvature increases it. In the following subsections we will illustrate this effect with two examples. We have studied two simple cases in which $\Delta V(\Gamma)$ can be analytically evaluated, in order to clarify the issues just discussed.

3.2.2 The coil-torus transition

Whereas the mathematical origin of the term $V_{\chi<0}$ is clear, its physical meaning of the quantity may result rather obscure. In order to better understand this point we now address a simple example where all calculations can be carried out analytically.

Consider a toroidal tube of thickness Δ and radius $R = L/2\pi$, whose centreline lies in the $\{\hat{\mathbf{e}}_x, \hat{\mathbf{e}}_y\}$ plane. Its parametric equation reads

$$\mathbf{R}(s) = R\cos\left(\frac{s}{R}\right)\,\hat{\mathbf{e}}_x + R\sin\left(\frac{s}{R}\right)\,\hat{\mathbf{e}}_y, \tag{3.22}$$

where the condition $R > \Delta$ have to be satisfied. The Frenet unit vectors are

$$\begin{aligned}
\hat{\mathbf{T}}(s) &= -\sin\left(\frac{s}{R}\right)\,\hat{\mathbf{e}}_x + \cos\left(\frac{s}{R}\right)\,\hat{\mathbf{e}}_y \\
\hat{\mathbf{N}}(s) &= -\cos\left(\frac{s}{R}\right)\,\hat{\mathbf{e}}_x - \sin\left(\frac{s}{R}\right)\,\hat{\mathbf{e}}_y \\
\hat{\mathbf{B}}(s) &\equiv \hat{\mathbf{e}}_z
\end{aligned} \tag{3.23}$$

Because of the toroidal symmetry, is then easy to see that both the curvature and the torsion, appearing in the Frenet-Serret equations 2.19, are independent of s and are

$$\kappa(s) = \frac{1}{R_c(s)} \equiv \frac{1}{R} \qquad\qquad \tau\,(s) \equiv 0 \tag{3.24}$$

We now insert this torus in a solution containing particles of radius ϵ and compute the volume of the excluded region, that is the region of space forbidden to the centre of the solvent spheres, generated by this 'compact' configuration. Clearly no changes will occur in the excluded volume compared with the swollen results as long as $R > \Delta_\epsilon$. The opposite situation, $R < \Delta_\epsilon$, is depicted in Fig. 3.4b. We see that the excluded volume produced by the torus (the dashed region in figure) is the solid generated by a rotation around the z axis of the part of the circle with radius Δ_ϵ contained in the right-half plane, and hence its calculation can be directly done using the Guldinus theorem[1]. An even simpler method consists in using the parameterisation $\{s, \theta, \rho\}$ introduced in the previous paragraph. The translational invariance along s simplify the integration since the integrand is a function only of (θ, ρ). Hence, upon defining $\theta_c = \arcsin\sqrt{1 - R^2/\Delta_\epsilon^2}$ the angular coordinate of the yellow sphere centre (see Fig. 3.4b), we have

$$\begin{aligned} V_{exc,C} \;=\; & 2\pi R \int_{\theta_c}^{2\pi-\theta_c} d\theta \int_0^{\Delta_\epsilon} d\rho\,\rho \left|1 - \frac{\rho\cos\theta}{R}\right| + \\ & +2\pi R \int_{-\theta_c}^{\theta_c} d\theta \int_0^{\rho_z(\theta)} d\rho\,\rho \left|1 - \frac{\rho\cos\theta}{R}\right| \end{aligned} \tag{3.25}$$

where $\rho_z(\theta) = R/\cos\theta$ is the radial distance, at an angle θ, of the z axis. As it is easy to see, both absolute values in 3.25 can be removed and the calculation is straightforward. We then obtain for the difference $\Delta V = V_{exc,C} - V_{exc,S} = V_{exc,C} - 2\pi^2 R\Delta_\epsilon^2$, expressed in function of $\eta = R/\Delta_\epsilon$,

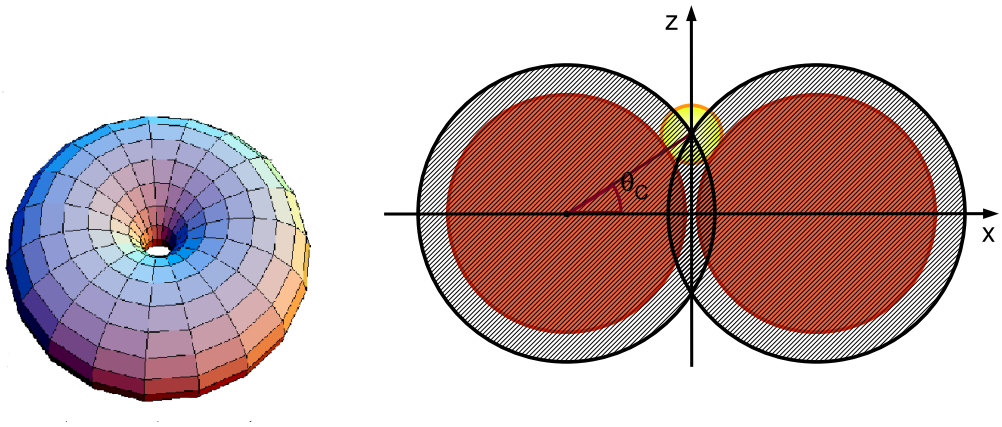
$$\Delta V = \frac{4}{3}\pi\Delta_\epsilon^3\,\Phi(\eta), \tag{3.26}$$

[1]The Guldinus theorem (also knows as the Pappus's centroid theorem) states that the volume of a solid of revolution, generated by rotating a plane figure F about an external axis, is equal to the product of the area of F and the distance travelled by its geometric centroid.

with

$$\Phi(\eta) \equiv \left(1 + \frac{1}{2}\eta^2\right)\sqrt{1-\eta^2} - \frac{3}{2}\eta \arcsin\sqrt{1-\eta^2}. \tag{3.27}$$

A rapid analysis of the function $\Phi(\eta)$ show that is always *positive* in the region $\Delta/\Delta_\epsilon \leq \eta \leq 1$. We can conclude that a tube bent in a toroidal configuration, tight enough to overlap in its interior, increases its excluded volume, meaning that the swollen configuration is entropically *more* favourable compared with the compact one! To understand this counterintuitive result let's compute separately the overlap volume, V_{Ov} and the $V_{\chi<0}$ term.



(a) (b)

Figure 3.4: Fig. *a*: Picture of a tight torus. Fig. *b*: Schematic representation of a torus section in a plane passing through the torus axis of symmetry (the z axis in the picture). The excluded region is dashed and the lens of overlap stands out. With respect to the section in the left, the inner boundary of the overlap region is the left arc delimiting the lens, characterised by the radial distance $\rho_{min}(\theta)$ given in Eq. 3.28. The outer boundary has simply $\rho(\theta) \equiv \Delta_\epsilon$.

In order to compute V_{Ov}, we note that, again in view of the toroidal symmetry, the overlap region is given by rotation of the lens area shown in Fig. 3.4b. This intuitive picture can be confirmed by the evaluation of the function $n(s,\theta,\rho)$: it results identically 2 for any (s,θ,ρ) belonging to the domain $\mathcal{D}_{Ov} = [0, 2\pi R] \times [-\theta_c, \theta_c] \times [\rho_{min}(\theta), \Delta_\epsilon]$, while we have $n(s,\rho,\theta) \equiv 1$ in the complementary set[2]. The radial distance $\rho_{min}(\theta)$ represents the inner boundary of

[2]The points of the excluded region belonging to z axis represent an exception since they be-

the lens and is given by the expression

$$\rho_{min}(\theta) \;=\; 2R\cos\theta - \sqrt{\Delta_\epsilon^2 - 4R^2\sin^2\theta}. \tag{3.28}$$

Indeed, as plotted in Fig. 3.5, the overlap conditions 3.15 in the domain $\mathcal{D}_{Ov}$, always yield two solutions (one maximum and one minimum) corresponding to the two aforementioned sections. The overlap volume can be readily obtained from

$$V_{Ov} \;=\; 2\pi R\int_{-\theta_c}^{+\theta_c} d\theta \int_{\rho_{min}(\theta)}^{\Delta_\epsilon} d\rho\,\frac{1}{2}\rho\left|1-\frac{\rho\cos\theta}{R}\right|. \tag{3.29}$$

Let's now evaluate the term $V_{\chi<0}$. The condition $\chi(\theta,\rho) = 1-\rho\cos\theta/R < 0$ become $\rho > R/\cos\theta$, thus the integration domain is simply $[-\theta_c,\theta_c]\times[\rho_z(\theta),\Delta_\epsilon]$, obtaining the integral

$$V_{\chi<0} \;=\; 2\pi R\int_{-\theta_c}^{\theta_c} d\theta \int_{\rho_z(\theta)}^{\Delta_\epsilon} d\rho\,\rho\left|1-\frac{\rho\cos\theta}{R}\right|. \tag{3.30}$$

A direct calculation then confirms what can be deduced by the picture

$$V_{Ov} = V_{\chi<0} = \frac{4}{3}\pi\Delta_\epsilon^3\,\Phi(\eta) \qquad \Rightarrow \tag{3.31}$$

$$\Delta V = -V_{Ov} + 2V_{\chi<0} = \frac{4}{3}\pi\Delta_\epsilon^3\,\Phi(\eta),$$

where we remark that $\Phi(\eta)$ is always positive. This result stems from the fact that the calculation of volume based on Eqns. 3.19 and 3.20, always yield inflated or deflated volume elements depending on whether the normal unit vector $\hat{\rho} = \cos\theta\;\hat{\mathbf{N}}(s) + \sin\theta\;\hat{\mathbf{B}}(s)$ is pointing inside or outside the curve[3]. In the absence of overlap, the two effects balance against each other yielding the object volume as a overall result. In the case of the torus, however, the geometry of the system prevents the solvent particles to fit inside the torus, thus yielding an unbalanced inflated volume outside which is responsible of the result 3.26.

long to infinite sections. We neglect this pathology because they form a region of null volume.

[3]The unit vector $\hat{\rho}$ points inside (outside) the curvature if its scalar product with the principal normal vector $\hat{\mathbf{N}}$ is positive (negative).

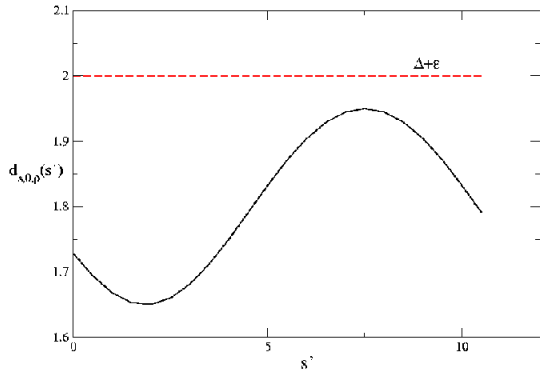


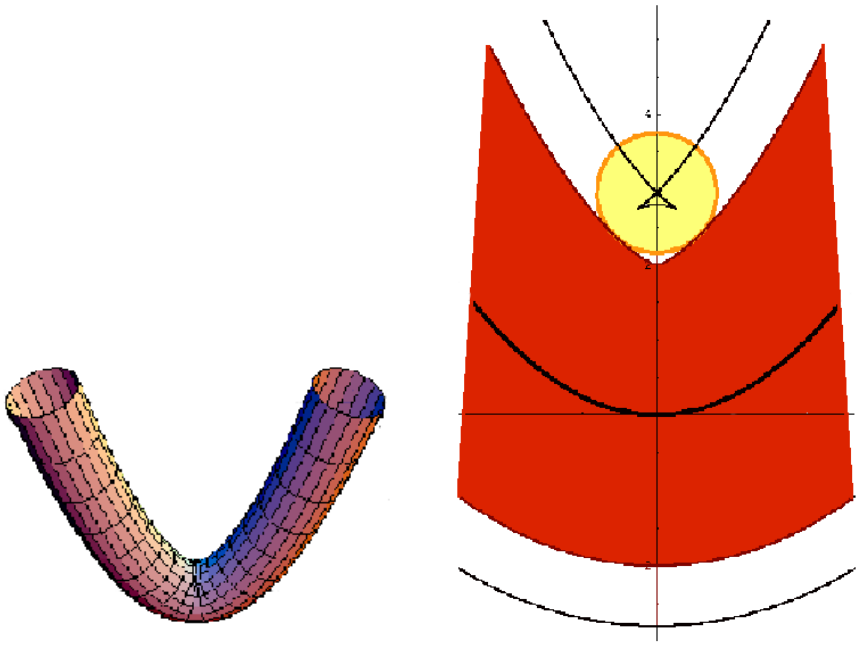
Figure 3.5: Plot of $d_{s,\theta,\rho}(s')$ computed numerically for a point with $(s,\theta,\rho) = (7.5, 0, 1.95)$ inside a torus characterised by $R = 1.8$ and $\Delta = 1.5$ and $\epsilon = 0.5$. The function $d_{s,\theta,\rho}(s')$ is however easy to calculate analytically and a trivial derivation shows that it presents two extrema in correspondence of s and $s + \pi$.

3.2.3 The coil-parabolic transition for a tube

In the previous example we have seen how a particular shape of the tube (a torus) was yielding a result which contrasted with the picture based on the Asakura-Oosawa-Vrij scenario, and we have clarified the origin of this peculiar effect. On the other hand, one might truthfully expect that there exist geometries, other than a helix, where a mechanism akin to the one suggested by Snir and Kamien [25] applies: the tube spontaneously deform and bend, compatibly with the hard core excluded volume, because in doing so the excluded volume is reduced.

In this second example we illustrate the case of a parabolic-shaped tube as depicted in Fig. 3.6. An analytical calculation for this system proves to be hard for the most general configuration. On the other hand, it is very simple the same calculation of the two-dimensional counterpart. Is is an instructive exercise to analyse the two different contributes V_{Ov} and $V_{\chi<0}$: we will show that in this case their balance make the tube bending entropically advantageous.

We put the tube centreline in the $\{\hat{\mathrm{e}}_x, \hat{\mathrm{e}}_y\}$ plane, and we express it through



(a) (b)

Figure 3.6: The parabolic tube analysed in this section. (a) Three dimensional picture of the tube. (b) Schematic representation of the two dimensional problem. The red region represents the section of the parabolic tube in the $\{\hat{\mathbf{e}}_x, \hat{\mathbf{e}}_y\}$ plane, its centreline being the simplest parabola $y = ax^2$. The black line that surround the tube both in the top and in the bottom represents the boundary of the excluded region to the centre of the solvent particle in yellow. Since the parabola is tight enough, the excluded region overlaps: the overlap region is the small triangle crossing the y axis.

the relation

$$\mathbf{R}(\lambda) = \lambda \hat{\mathbf{e}_x} + a\lambda^2 \hat{\mathbf{e}}_y \tag{3.32}$$

since it is more handy than the arclength parameterisation. Here the value of a gives a measure of the tightness of the parabola. The Frenet unit vectors defined in Eq. 2.19 are

$$\begin{aligned} \mathbf{T}\hat{(\lambda)} &= \frac{\hat{\mathbf{e}_x} + 2a\lambda\hat{\mathbf{e}}_y}{s'(\lambda)} \\ \mathbf{N}\hat{(\lambda)} &= \frac{-2a\lambda\hat{\mathbf{e}_x} + \hat{\mathbf{e}}_y}{s'(\lambda)} \\ \mathbf{B}\hat{(\lambda)} &= \hat{\mathbf{e}}_z \end{aligned} \tag{3.33}$$

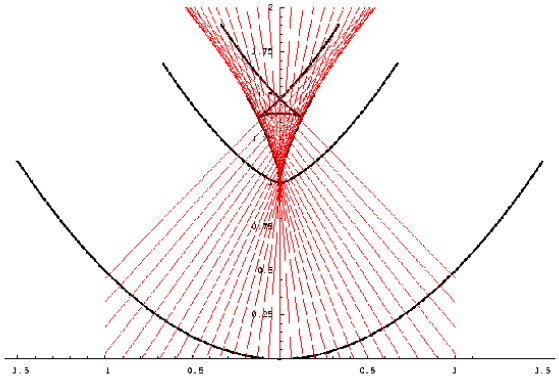


Figure 3.7: Bundle of normal lines to the parabola. The evolute represents the boundary between the region in which the normals do not intersect each other, and region in which each point is the intersection of three distinct normals.

where we have introduced the derivative

$$s^{'}(\lambda) = \sqrt{1+4a^2\lambda^2} \tag{3.34}$$

Likewise, we can compute the curvature and torsion from 2.19 and find after few manipulations

$$\begin{aligned} \kappa(\lambda) \equiv \frac{1}{R_c(\lambda)} &= \frac{2a}{s^{'3}(\lambda)} \\ \tau(\lambda) &= 0 \end{aligned} \tag{3.35}$$

These equations imply that the relation $a < 1/(2\Delta)$ have to be satisfied so that $R_c(\lambda) > \Delta$, $\forall\lambda$, and the tube does not compenetrate. A point belonging to the two-dimensional region of excluded volume produced by the tube can be then identified by

$$\mathbf{r}(\lambda,\rho) = \mathbf{R}(\lambda) + \rho\hat{\mathbf{N}}(\lambda) \tag{3.36}$$

with $-\Delta_\epsilon \leq \rho \leq \Delta_\epsilon$ [4]. We will be interested in configurations verifying the condition $1/(2\Delta_\epsilon) < a$ corresponding to a non-vanishing overlap of the volume.

[4]It is equivalent to take $0 \leq \rho \leq \Delta_\epsilon$ and θ varying between 0 and π, the angular values corresponding to the $\{\hat{\mathbf{e}}_x, \hat{\mathbf{e}}_y\}$ plane.

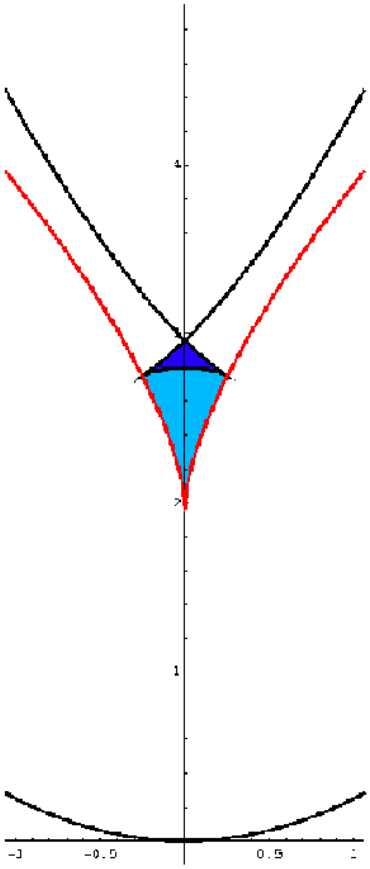


Figure 3.8: Representation of the overlap region. The evolute (the red line) delimits the region of overlap. The azure region is characterised by $n(\lambda,\rho)=3$ and $\chi(\lambda,\rho)<0$, while the blue region is characterised by $n(\lambda,\rho)=2$. Hence according to Eq. 3.21 we have $A_{exc,C}-A_{exc,S}=-A_2$.

In order to make contact with previous analysis, as described in Eq. 3.15, consider a fixed point $\mathbf{r}(\lambda,\rho)$, with cartesian coordinate $(x_{\mathbf{r}},y_{\mathbf{r}})$, and its distance $\mathbf{d}_{\lambda,\rho}(\lambda')=\mathbf{r}(\lambda,\rho)-\mathbf{R}(\lambda')$, that is

$$\mathbf{d}_{\lambda,\rho}(\lambda') \;=\; (x_{\mathbf{r}}-\lambda')\,\hat{\mathbf{e}}_x+\left(y_{\mathbf{r}}-a\lambda'^2\right)\hat{\mathbf{e}}_y \tag{3.37}$$

Upon imposing that the scalar product $\mathbf{d}_{\lambda,\rho}(\lambda')\,\cdot\,\mathbf{T}(\lambda')$ vanishes one obtains a cubic equation

$$(x_{\mathbf{r}}-\lambda')+2a\lambda'\left(y_{\mathbf{r}}-a\lambda'^2\right) \;=\; 0. \tag{3.38}$$

Depending on the values of $(x_{\mathbf{r}}, y_{\mathbf{r}})$, i.e. the regions of the $\{\hat{\mathbf{e}}_x, \hat{\mathbf{e}}_y\}$ plane $\mathbf{r}(\lambda, \rho)$ belong to, this equation can have one, two or three solutions. The condition for this equation to have exactly two real roots is given by the system

$$\begin{cases} \Psi\left(x_{\mathbf{r}}, y_{\mathbf{r}}, \lambda'\right) = \left(x_{\mathbf{r}} - \lambda'\right) + 2a\lambda'\left(y_{\mathbf{r}} - a\lambda'^2\right) = 0 \\ \partial_{\lambda'}\Psi\left(x_{\mathbf{r}}, y_{\mathbf{r}}, \lambda'\right) = 1 - 2ay_{\mathbf{r}} + 6a^2\lambda'^2 = 0 \end{cases} \tag{3.39}$$

that, by definition, gives the evolute of the parabola [5]. This is the curve separating regions of the plane with only one normal to the parabola axis (below the evolute) and regions with three normals (above the evolute) as depicted in Fig. 3.7. We now have all necessary tools to compute A_{Ov} and $A_{\chi<0}$, the two-dimensional analogue of Eqns. 3.19 and 3.20. The Fig. 3.8 illustrates the situation. Below the evolute $n(\lambda, \rho) = 1$, while over the evolute two regions are distinguishable: the azure region is characterised by $n(\lambda, \rho) = 3$ while the blue ones has $n(\lambda, \rho) = 2$ – indeed there are three λ' in this last region so that $\mathbf{d}_{\lambda,\rho}(\lambda')$ is normal to the parabola, but only two of them satisfy the second condition of 3.15, $d_{\lambda,\rho}(\lambda') < \Delta_\epsilon$. Hence being A_3 and A_2 the areas of the regions with $n(\lambda, \rho)$ equal to 3 and 2 respectively and remembering that $n(\lambda, \rho) - 1$ represents the multiplicity of overlap we have

$$A_{Ov} = 2A_3 + A_2, \tag{3.40}$$

where A_3 and A_2 are easy to integrate.

The value of $A_{\chi<0}$ can be determined noting that in the region above the parabola the expression of $\chi(\rho, \lambda)$ results $1 - \rho/R_c(\lambda)$, meaning that $A_{\chi<0}$ includes points with $\rho > R_c(\lambda)$. Hence the boundary of this region is given again by the evolute (see footnote 5) implying that, remembering the trivial constraint $\rho < \Delta_\epsilon$, the region with $\chi(\rho, \lambda) < 0$ coincides with the azure area. This leads to $A_{\chi<0} = A_3$, and hence

$$A_{exc,C} - A_{exc,S} \;=\; -A_{ov} + 2A_{\chi<0} = -A_2 \tag{3.41}$$

which is a negative quantity as expected.

[5] The evolute of a curve $\Gamma \in \mathbb{R}^2$ is the locus of the curvature centres of Γ. Being $\mathbf{R}(\lambda)$ the parametrisation of Γ, the evolute has the property of being the envelope of the family $\Psi(x, y, \lambda)$ of normals to Γ, that means satisfying the two condition of Eq. 3.39: $\Psi\left(x, y, \lambda\right) = 0$ and $\partial_\lambda \Psi\left(x, y, \lambda\right) = 0$.

3.3 The coil-helix transition for a tube

In the previous section we have introduced a method to compute the excluded volume produced by a tube. We have moreover presented two simple examples with the only aim of clarifying the issues related to it. Once tested the formula, we are ready to face the physical problem of adapting the Asakura-Oosawa scenario to the case of a single polymer embedded in solution. Neglecting the contribution to the equilibrium configuration of the system given by the polymer conformational entropy we start with studying the configurations minimising the depletion interaction. This means finding the *optimal* conformation of the tube-polymer, able to minimise the excluded volume.

This optimisation problem was tackled in the most general case by means of the simulated annealing method, and will be the subject of the following section. In this section we assume instead that the helix is the optimal conformation of a tube. This idea has a strong appeal since it is able to explain the ubiquity of this conformation in biomolecules. The assumption is however motivated by the property of maximal packing, illustrated in chapter 2, characterising helices and the magic helix in particular. Therefore we will analyse in detail this conformation in order to find which is the helix that minimise its excluded volume.

In the first part of this section the mathematical apparatus necessary for the helix description is introduced. In the second part we report our results, summarised in the plot of the optimal helix pitch to radius ratio in function of the solvent radius. The results obtained are in perfect agreement with the Snir and Kamien ones.

3.3.1 mathematical apparatus for the helix description

Given a helical tube of radius Δ and length L, its axis can be described in cartesian coordinates $\{\hat{\mathbf{e}}_x, \hat{\mathbf{e}}_y, \hat{\mathbf{e}}_z\}$ in parametric form using an angle $0 \leq t \leq 2\pi n$, n being the number of turns of the helix, by a vector

$$\mathbf{R}(t) = R\cos t\, \hat{\mathbf{e}}_x + R\sin t\, \hat{\mathbf{e}}_y + \frac{P}{2\pi} t\, \hat{\mathbf{e}}_z \tag{3.42}$$

where R and P are the radius and the pitch of the helix, respectively. The Frenet frame unit vectors are readily obtained from Eq. 2.19

$$\begin{aligned}\hat{\mathbf{T}}(t) &= \frac{2\pi}{\Lambda}\left[-R\sin t\,\hat{\mathbf{e}}_x + R\cos t\,\hat{\mathbf{e}}_y + \frac{P}{2\pi}\,\hat{\mathbf{e}}_z\right] \\ \hat{\mathbf{N}}(t) &= -\cos t\,\hat{\mathbf{e}}_x - \sin t\,\hat{\mathbf{e}}_y \\ \hat{\mathbf{B}}(t) &= \frac{2\pi}{\Lambda}\left[\frac{P}{2\pi}\sin t\,\hat{\mathbf{e}}_x - \frac{P}{2\pi}\cos t\,\hat{\mathbf{e}}_y + R\,\hat{\mathbf{e}}_z\right]\end{aligned} \tag{3.43}$$

where we have introduced the quantity

$$\Lambda = \sqrt{P^2 + 4\pi^2 R^2}. \tag{3.44}$$

The curvature and torsion are

$$\begin{aligned}\kappa(t) &\equiv \frac{4\pi^2 R}{\Lambda^2} \\ \tau(t) &\equiv \frac{2\pi P}{\Lambda^2},\end{aligned} \tag{3.45}$$

note that they are both independent of t. Finally the length per unit of turn is simply

$$\frac{L}{n} = \Lambda \tag{3.46}$$

The requirement for the helix to satisfy self-avoidance and have non-vanishing overlap of the excluded volume leads [25] to a range of possible values in the (R, P) plane as depicted in Fig.3.9. Note that the two boundary curves, $R_c = \Delta$ and $d = 2\Delta$, d being the minimum distance between successive turns, intersect at a point (R^*, P^*) such that the ratio $c^* = P^*/R^*$, corresponds to the magic helix value $2.512\ldots$, fully analysed in chapter 2. Imposing the same conditions with regard to the helix representing the forbidden region, i.e. $R_c = \Delta_\epsilon$ and $d = 2\Delta_\epsilon$, we obtain the boundary over which no overlap occurs and the excluded volume is simply the swollen one, $\pi\Delta_\epsilon^2 L$. So the light grey region shown in Fig. 3.9 highlights the range of values we are interested in.

3.3.2 Excluded and overlap volumes of a helix

We seek now the optimal helix pitch to radius ratio in varying the solvent size ϵ. This topic was already discussed in Ref. [25] but, as will be elaborated

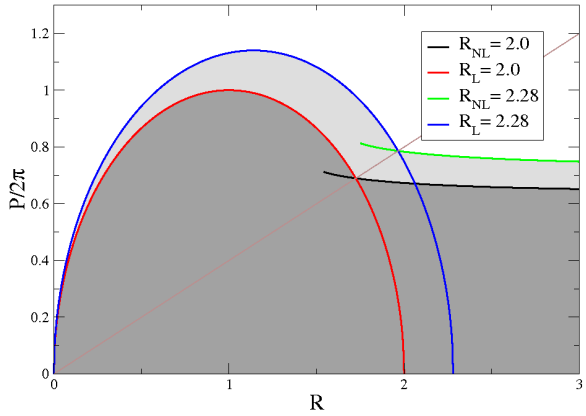


Figure 3.9: Diagram of the different regimes according to the (R, P) values. The dark grey region corresponds to values not consistent with the thickness constraint, while the light gray region represents helices with an overlapping excluded region. The gray straight line has equation $P = c^* R$.

below, our reanalysis will uncover some nuances which were missing in their discussion.

We start with studying the 'bulk' helix, whose length approaches infinity, in order to neglect boundary effects. This requirement simplify our calculations since we can take advantage of the helix screw symmetry, and compute only the overlap between a helix section and the whole curve, obtaining in this manner an excluded volume per unit length.

We know from Eq. 3.21 that in order to minimise the excluded volume we have to maximise the overlap avoiding, as much as possible, to make turns with $R_c < \Delta_\epsilon$. For a fixed radius R the maximum overlap is obtained when successive turn are lying one on the top of the other. On varying R and, simultaneously, taking the pitch such that the last condition is satisfied (this means to choose the parameters (R, P) along the black curve of Fig. 3.9), we compute the excluded volume numerically, implementing our method summarised in Eq. 3.14 and 3.15 after a fine discretisation of the variables s, θ and ρ. We sum the contributions given by each element of volume centred in $\mathbf{r}(\theta_k, \rho_n)$ after

computing the integer function $n(\theta_k, \rho_n)$, where the coordinate along the curve disappears for the translational invariance just mentioned. After finding $R(\epsilon)$, associated to the optimal helix for a given ϵ, we have plotted the corresponding value of $c(\epsilon)$. The plot is reported in Fig. 3.10: we can see that for small ϵ, the resulting $c(\epsilon)$ corresponds to the magic value $2.512\ldots$, while after the critical threshold $\epsilon \cong 0.08$, there is a quite sharp transition to a regime in which $c(\epsilon)$ decreases almost linearly. These results are identical to those obtained by Snir and Kamien [25].

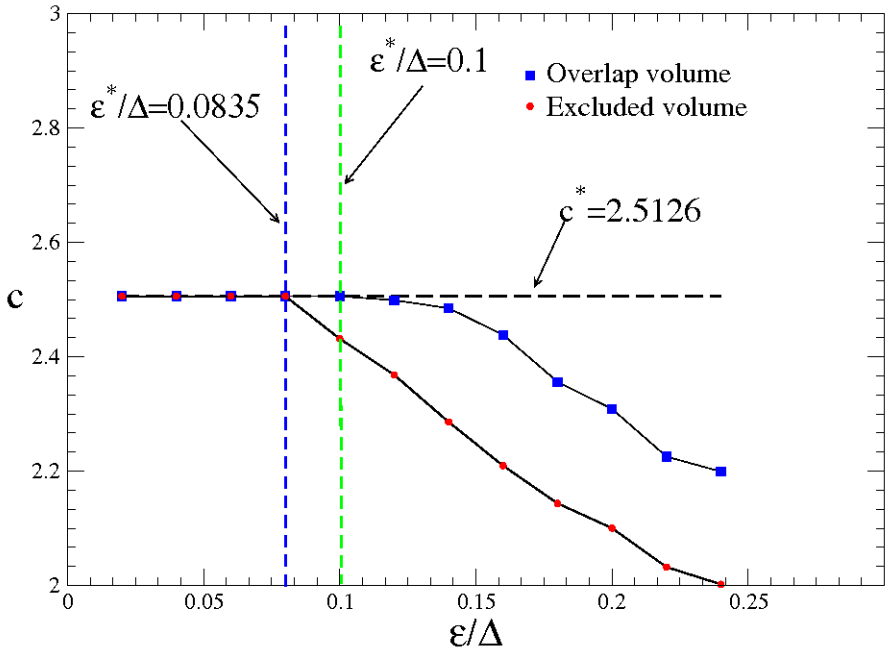
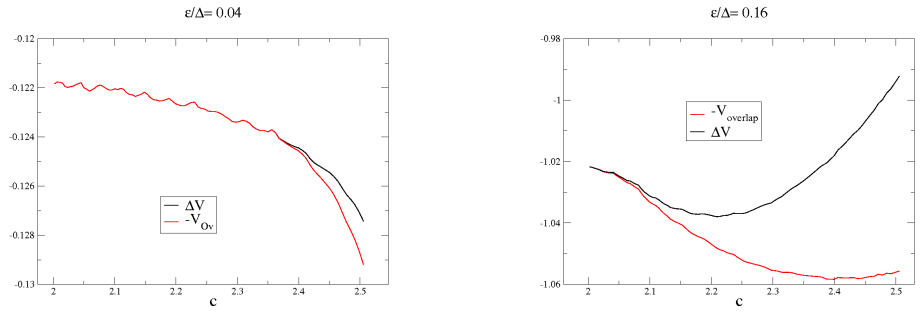


Figure 3.10: Plot of the pitch to radius ratios corresponding to the helices minimising the excluded volume (red circles) and maximising the overlap (blue square). The magic helix minimises the excluded volume for small ϵ/Δ, then after a threshold value the optimal c decreases almost linearly. For the case of overlap volume instead the curve remains close to c^* for larger values of ϵ/Δ.

In view of Eq. 3.21 we expect $\Delta V = V_{exc,C} - V_{exc,S}$ and V_{ov} *not* to coincide,

the difference being, in fact, given by $-2V_{\chi<0}$. To evaluate in this case the contribution due to this last term, we have plotted the two quantity ΔV and V_{Ov} as a function of c as depicted in Fig. 3.11 for two different values of ϵ/Δ. Fluctuations appearing in the plots are due to the numerical discretisations. As expected we find $\Delta V < 0$ for every values of c, so the excluded volume of the helix is, in any case, smaller than the corresponding straight tube one, independently of the ratio ϵ/Δ. This last quantity influences however the ΔV trend. When $\epsilon/\Delta = 0.04$ (Fig. 3.11a) it has a minimum in correspondence of c^* and an increase of R leads to an increase of ΔV. In the case $\epsilon/\Delta = 0.16$, on the other hand, the helix minimising the excluded volume is no longer the ideal helix and ΔV decreases with R until it reaches a minimum. In both cases the difference between ΔV and V_{Ov} is greater for small R (i.e. larger c) because this is the regime where the effect of $V_{\chi<0}$ is more effective.



(a) (b)

Figure 3.11: Plot of the $\Delta V(c)$ and $V_{Ov}(c)$ for ϵ/Δ equal to 0.04 (Fig. a) and 0.16 (Fig. b).

The importance of the $V_{\chi<0}$ term is highlighted in Fig. 3.10 too, where beside the plot of $c(\epsilon)$ corresponding to minimum ΔV the same calculation relative to the overlap volume only is reported. We note how the optimal value of c remains closer to c^* for higher ϵ/Δ with respect to the previous case. The helix is forced to unwind itself in order to decrease its curvature and reduce the contribution of $V_{\chi<0}$, losing in this way overlap volume.

It is quite unexpected otherwise that the magic helix does not maximise the overlap volume for every ϵ/Δ. To clarify this point, we have checked the θ dependence of the overlap volume V_{ov}. This is reported in Ref. 3.12, for

the case of $\epsilon/\Delta = 0.18$, where the contributions $V_{ov}(|\theta| < \theta_0)$ and $V_{ov}(\theta_0)\delta\theta_0$ are displayed as a function of θ_0 for both the magic helix c^* and the value c maximising V_{ov}. The behaviour of $V_{ov}(\theta_0)\delta\theta_0$ shows us that for small θ_0, that is in the proximity of the helix axis, the contribute to the overlap is quite small, while for $\theta_0 \approx \frac{\pi}{2}$ becomes the dominant contribution, corresponding to the overlap between two successive turns lying one at the top of the other. This latter increases as R increases (i.e. c decreases) tending to its maximum value for $R \to \infty$, in a configuration that approximately corresponds to two torus on the top of one other. So, for ϵ not too small, it is advantageous for the helix to gain overlap with successive turns reducing a bit the central one.

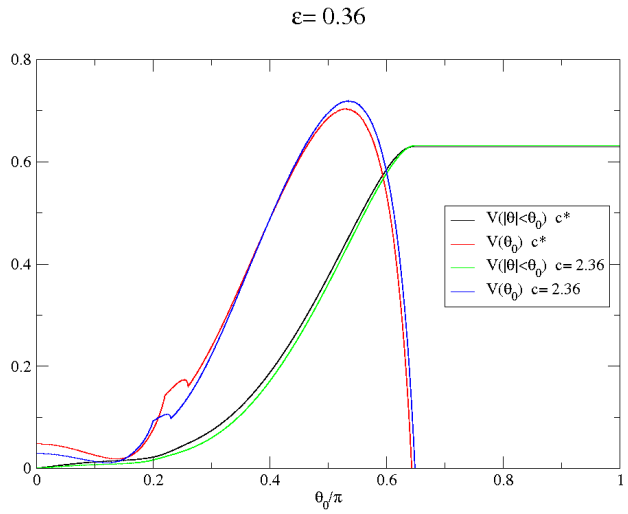


Figure 3.12: Plot of $V_{ov}(|\theta| < \theta_0)$ and $V_{ov}(\theta_0)\delta\theta_0$ for $\epsilon/\Delta = 0.18$, in the two cases $c = c^*$, and $c \cong 2.36$ corresponding to the optimal helix. $\delta\theta_0$ has been set equal to the discretisation amplitude $\theta_{k+1} - \theta_k$.

To conclude our analysis of the bulk helix, it is instructive to study how regions of different multiplicity $n(\rho, \theta)$ are distributed in the section of the tube with radius Δ_ϵ. This is done in Figure 3.13 where we have pinned down regions with three (in orange), two (in yellow), and one (in blank) sections (these are in fact the only possible values obtained) for a given ϵ and c. We can note how the regions with θ near $\frac{\pi}{2}$ and $\frac{3\pi}{2}$ are characterised by having $n = 2$, as one

could expect on intuitive basis, while the region with θ near zero, in the inner side of the helix has $n = 3$.

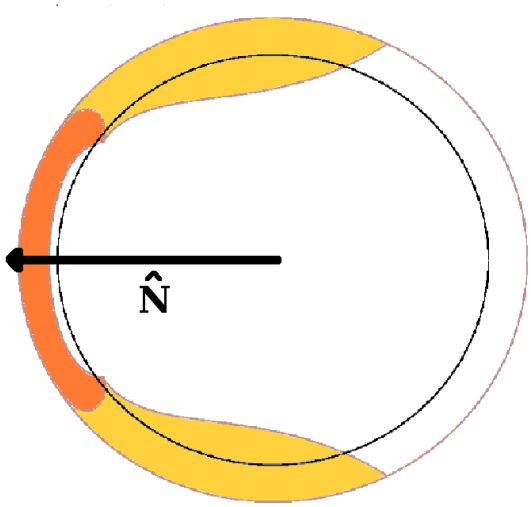


Figure 3.13: A section of the soft tube with thickness Δ_ϵ forming the excluded region. The black circle corresponds to hard core tube. The regions of overlap with relative multiplicity are highlight with different colour: the yellow and orange regions corresponds to $n = 2$ and $n = 3$ respectively.

Our method allows us also to study finite size polymer without the approximation of translational invariance. Therefore we were able to study the role played by the finiteness of the chain in a helical conformation. In Fig. 3.14 the same plot of Fig. 3.10 is shown for helices of different length compared with the infinite helix: $c(\epsilon)$ corresponds to helix minimising the overall excluded volume. We can see that the finite size indeed affects the optimal conformation. For the smallest length tested ($L = 10\Delta$) the optimal helix remains close to c^* for a large range of values, whereas this range decreases (approaching to the infinity case) with the increasing of the helix length.

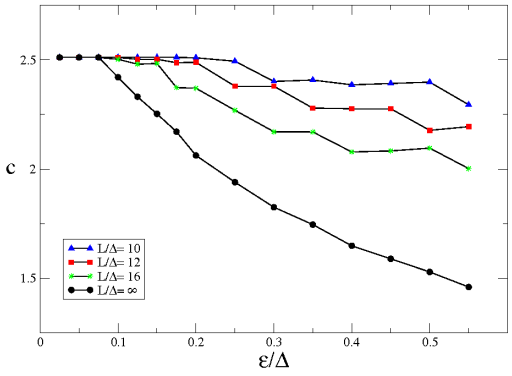


Figure 3.14: Plot of pitch to radius ratio of the optimal helix for different length L. The fluctuations in the plot are due to the numerical discretisations.

3.4 The optimal conformation of a tube-like polymer

In this section we want to conclude our investigations on the depletion effect for a tube-polymer in a solution, facing the problem of finding the optimal conformation of a finite length tube among the menu of all possible compact conformations. For this purpose we performed numerical simulations using the simulating annealing method, described in appendix A. The discretisation chosen for these computations is quite coarse: the tube centreline is described as a chain of N points with fixed bond length, $b = \Delta/2$; while the azimuthal angle and the radial distance are described by 200 points spanning the respective intervals $[0, 2\pi]$ and $[\Delta - \epsilon, \Delta + \epsilon]$, the last one being the range of value that can be affected by an overlap or that can be characterised by $\chi < 0$[6]. Preliminary analysis shows that such a discretisation does not affect the polymer behaviour in the case of large sphere radius, since a finer discretisation leads to the same results. Otherwise the structures obtained for small values of ϵ, as

[6]In truth far by the continuum limit (i.e. $\epsilon << \Delta$, as we will see) overlapping regions can be found also for $\rho < \Delta - \epsilon$. Therefor we have taken care, in this case, to check also the interval $[0, \Delta - \epsilon]$.

we will be explained, are clearly affected by the finite bond length. This allows us to study the role played by a discrete versus continuum descriptions of the tube, that, we recall, wants to model an intrinsically discrete object such as the polypeptide chain.

ϵ/Δ	$L = 10\Delta$	$L = 20\Delta$
0.005	0.049	0.053
0.025	0.061	0.067
0.050	0.079	0.098
0.100	0.12	0.18
0.200	0.24	0.39
0.500	0.77	1.3

Table 3.1: Values of $\Delta V/(L\Delta^2)$ related to the optimal structures, for the tested values of ϵ/Δ and L/Δ.

We have considered two different polymer lengths, $L = 10\Delta$ and $L = 20\Delta$ and several values of ϵ/Δ for each: 0.005, 0.025, 0.05, 0.10, 0.20 and 0.50. The results are summarised in Fig. 3.15, where the optimal structures obtained on varying the parameters are shown; the values of $\Delta V = V_{exc,S} - V_{exc}$ reached are instead reported in Tab. 3.1. In Fig. 3.15a the conformations with $L = 10\Delta$ are displayed. We can see that two different 'phases' emerge on varying ϵ: at small values of the parameter, a *planar phase*, in which the ground state has a typical planar hairpin structure, is found; the large ϵ regime, otherwise, is characterised by a *helicoidal phase* in which helices and saddles dominate. The two different regimes are separated by a *crossover* region characterised by the coexistence of all these structures. The crucial role played by the discrete nature of the polymer appears clearly for small ϵ, since the planar hairpin is able to reach an high overlap volume thanks to the special zigzag typical of this conformation.

A more complex scenario, shown in Fig. 3.15b, characterise the results with $L = 20\Delta$. The helicoidal phase in this case consists in a more varied collection of structures: double helices, saddles and irregular helices, with turns of differ-

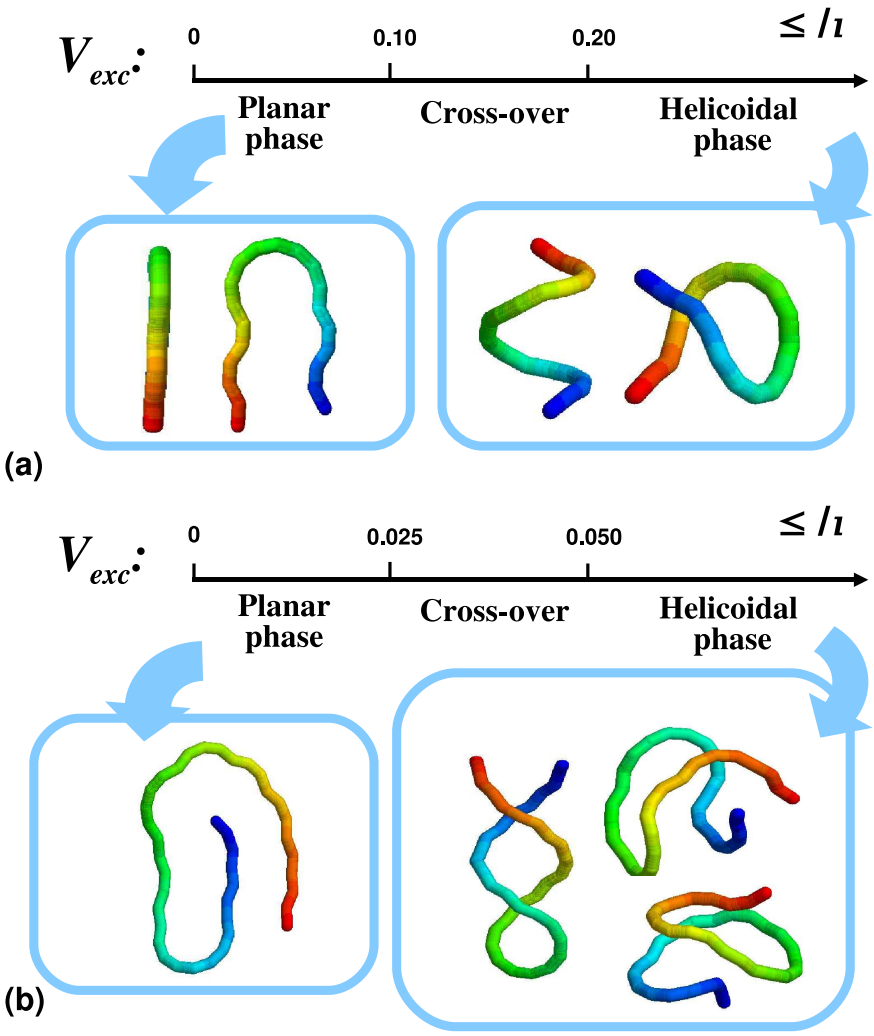


Figure 3.15: Conformations adopted by tubes of length (a) $L = 10\Delta$ and (b) $L = 20\Delta$ subject to the excluded volume prescriptions for promoting compaction. The values of ϵ/Δ are 0.005, 0.025, 0.05, 0.10, 0.20 and 0.50. The resulting conformations for the shorter tube are either saddles or helices for $\epsilon/\Delta \gtrsim 0.20$, while for smaller ϵ, after a crossover phase, the hairpin becomes the ground state – the planarity of the structure is highlighted. For the longer tube, the planar phase consist of β-sheet-like structures and the helicoidal phase is characterised by double helices, saddles and irregular helices. All these conformations are akin to those found in Ref. [17, 18].

ent lengths. Notice how by increasing the length the planar hairpin evolves in a β-sheet-like conformation. For both lengths our simulations suggest, in the crossover region, the existence of an energy barrier between the two classes of conformations, hairpin versus helix or saddle.

We can then conclude our analysis on the effect of the depletion interaction of a tube-polymer. Our results show that the depletion interaction in the same manner as the buried area, discussed in section 2.3, leads to structures similar to secondary motifs. The fact that both interactions lead to the same motifs confirm how these geometries have a common and fundamental origin which transcends the details of the model.

Chapter 4

Conclusion and perspective: the Hadwiger theorem

In the introduction on protein folding given in this thesis, the various classical semi-empirical interactions promoting the folding process have been introduced. Despite their importance, recognised since many decades, the role of each of them in the complex interplay governing the protein behaviour is, at the present time, not completely clear yet and still represents matter of discussion [24, 26, 66]. In particular the fact that interactions within the protein chain are so strongly influenced by the aqueous solvent, further complicates the attempt of identifying the principal driving forces of folding. For this reason also a simplified model should include an appropriate description of water surrounding the protein in order to reach a full comprehension of the delicate balance of the different mechanisms involved.

The solvent effects on the protein molecule have been the central focus of this thesis. Our aim was to provide a very basic model eligible to investigate, beside the peculiar properties of the polypeptide chain (i.e. steric hidrance, ability to form peptide hydrogen bonds), aspects proper of the protein-solvent interaction, following the conviction that both protein and solvent must be modelled at the same level.

In the previous chapters we have discussed two very basic interactions for a tube-like polymer, each of them introduced to capture a particular solvent effect. In chapter 2, we have analysed an energy function proportional to the area

of the buried regions of the tube surface, in order to model the hydrophobic effect. In chapter 3, on the other hand a depletion interaction was illustrated: the tendency of water molecules to maximise their translational entropy, was captured with an effective interaction proportional to the excluded volume produced by compact conformations of the tube.

Both the hydrophobic effect and depletion forces, can be efficaciously described with geometrical measures. Both of them have entropic origins and contribute to the solvation free energy of aqueous solution. The latter is defined as the difference between the free energy of the solvent in which the solute is immersed, and the free energy of the pure solvent, the solute being the protein with a prescribed configuration. It represents the proper function in order to analyse the connection between solvent and the protein molecule. But, despite its importance relatively little is known about the effective way of calculating this free energy difference. In order to obtain an accurate evaluation of it, each water molecule should be taken into account explicitly, like molecular dynamic simulations do in many cases. These calculations, nevertheless, are limited to small solute molecules and infeasible for large polyatomic molecules like proteins.

We want to conclude this thesis with a brief sketch of an alternative approach that has emerged recently [32, 33]. A *morphometric method* [31, 124] was applied in order to calculate the solvation free energy of a protein in a given structure, having specified the solvent-solvent and solvent-molecule interactions [34, 125]. Through this method the thermodynamic properties of the system are investigated from a geometric point of view: the starting assumption is that the free energy of the solvent can be efficaciously expressed, on a mesoscopic scale, as a simple linear combination of four fundamental geometrical measures. For its extreme simplicity and the importance given to the geometric characterisation of the system, the method seems to represent the natural extension of our studies. Indeed, as we are going to describe, the two self interactions (excluded volume and buried area) introduced in this thesis on physical ground, become in this framework simply two terms of the same functional. For these reasons an application of the morphometric method to the tube-polymer case is an immediate follow-up of the work illustrated in

this thesis.

In the following sections we will outline the main ideas underling the morphometric method, and we will discuss its possible applications to the case of a tube-like polymer.

4.1 The morphometric approach

In the presence of complex spatial structures, the exigency of characterising the morphology of the system arises, by reducing the spatial information to a finite number of relevant terms. Many systems considered by statistical physics, are characterised by complex patterns which exhibit an enormous amount of informations that need to be codified. We can cite, for example, a porous media[1] in which the percolation behaviour (an important quantity for practical purposes) is related to the morphology of its stochastic spatial structure. Complex fluids, such as microemulsions[2] and large scale distributions of the galaxies in the universe are some of many other examples of such systems.

The mathematical theory of *integral geometry*, developed some decades ago [31, 124, 128], provides a systematic formalisation of a class of measures called *Minkowski functionals*. The good properties of these functionals make them suitable morphological measures, i.e. measures eligible to carry quantitative informations on the shape of spatial homogeneous domains; they represent also the basis for the formulation of the geometric probability, powerful theory for investigations on complex systems. We do not want to go in depth with the treatment of this mathematical theory, since a complete introduction to the subject would require a whole thesis by itself. We then refer the reader to [32, 124] for the details, and here we will limit ourselves to define the basic concepts, focusing our attention to the physical application we are interested in.

[1] A solid material characterised by pores of different shape and size distributed in its interior is called porous media. Generally one observe a threshold in the volume density of the pores, the percolation threshold, above which the system supports water flow.

[2] Microemulsions are water-oil emulsions stabilised by amphiphilic molecules which decrease the surface tension between oil and water.

Define $\mathcal{P}$ the class of all subsets of $\mathbb{R}^n$ that are convex or that can be written as finite unions of convex sets. The term Minkowski functional indicates each geometrical measure defined on $\mathcal{P}$, that is characterised by the following properties:

i. Additivity Given two distinct domains A and B of $\mathcal{P}$, the measure of the union, $\mathcal{M}(A \cup B)$, have to satisfy the additivity relation: $\mathcal{M}(A \cup B) = \mathcal{M}(A) + \mathcal{M}(B) - \mathcal{M}(A \cap B)$.

ii. Motion invariance Let g a transformation belonging to the group of motions (i.e. rotations and translations), the action of g on A being denoted by gA, then the invariance $\mathcal{M}(gA) = \mathcal{M}(A)$ must be satisfied.

iii. Continuity For every sequence of sets $A_n \to A$ for $n \to \infty$, we require that $\mathcal{M}(A_n) \to \mathcal{M}(A)$; this condition is required only for A_n and A being convex sets.

The importance of this class of measures stems from their good quality of completeness. The property is known as Hadwiger theorem [31, 128] and states that each Minkowski functional in $\mathbb{R}^n$ can be written as a linear combination of $n+1$ Minkowski functionals, $\mathcal{M}_\nu$, $\nu = 0, \dots n$, where $\mathcal{M}_\nu$ has dimensionality $n - \nu$, i.e. it is identically zero on d-dimensional sets, embedded in $\mathbb{R}^n$, with $d < n - \nu$. The definition of these fundamental functionals starts by setting the measure $\mathcal{M}_n$ to be the Euler characteristic. This is the simplest possible functional, since it is defined to be 1 for each non-empty convex domain, 0 for the void set and its value for each finite union of convex set is computed through the additivity relation. It can be proved that this definition is equivalent to the one given in algebraic topology[3]. Each $\mathcal{M}_\nu$ can then be constructed starting from $\mathcal{M}_n$ by means of the Haar measure [32, 124]. Without going in depth with the details, we immediately report the result of practical interest. Given a domain, $A \subseteq \mathbb{R}^3$, bounded by a smooth surface, the four fundamental Minkowski functionals become simply the volume $V(A)$, the surface $\Sigma(A)$,

[3]The Euler characteristic, χ, is a topological invariant. Given a set A in the three-dimensional space, $\chi(A)$ can be computed through the useful relation: $\chi(A) = n_{cc} - n_t + n_c$, where n_{cc} is the number of connected components of A, n_t is the number of handles and n_c is the number of cavities.

the integrated mean curvature, $C(A) = \int_{\partial A} \frac{1}{2}(\varkappa_1 + \varkappa_2)\, d\Sigma$, and the integrated Gaussian curvature[4], $X(A) = \int_{\partial A} \varkappa_1 \varkappa_2\, d\Sigma$, where $\varkappa_1$ and $\varkappa_2$ are the local principal curvatures[5] of the surface ∂A. When the domain has edges and corners the last two functional are not defined anymore, in this case we can consider the sequence A_n of smooth domains approximating A, and compute the limits $\mathcal{M}_1(A) = \lim_{n\to\infty} C(A_n)$ and $\mathcal{M}_0(A) = \lim_{n\to\infty} X(A_n)$.

To summarise, the Hadwiger theorem asserts that each measure defined on $\mathcal{P}$ in $\mathbb{R}^3$, and satisfying the three aforementioned properties, can be written as a linear combination of the four basic measures listed above. These property of completeness is of crucial importance, since it gets the Minkowski functionals handy tools for an effective characterisation of systems in many fields of statistical physics.

Applications of integral geometry to various problems, such as porous media or microemulsions can be found in Ref. [32, 126, 127]. Here we focus our attention on an application of integral geometry to thermodynamics that has been the subject of recent theoretical investigations [33, 34, 117, 125]. Under some restrictions a thermodynamic function can be expressed in a simple and efficacious manner by means of morphological measures. Consider the case of a fluid confined in a box. In the thermodynamic limit the free energy of the fluid is an extensive quantity, thus it is simply proportional to the volume of the system. This fundamental ansatz, however, does not holds anymore when finite size effects are important: for a fluid bounded by a container its free energy depends on the shape of the container in a potentially complicated manner. The integral geometry provides a very powerful method for describing such a system: the free energy of the fluid, indeed, can be expressed in an

[4]We note that, as demonstrated by the *Gauss Bonnet Theorem* of differential geometry, the Gaussian curvature of an orientable compact surface ∂A is simply equal to $2\pi\chi(\partial A)$. Since the general relation [128] $\chi(\partial A) = (1 - (-1)^n)\chi(A)$ is satisfied, we recover the former assertion that $\mathcal{M}_n$ is (unless a multiplicative factor) the Euler characteristic.

[5]Consider the bundle of normal planes to ∂A in a given point $\mathbf{x}$; the intersection between the planes and the surface defines a bundle of curves in $\mathbb{R}^3$, the *normal curves* at $\mathbf{x}$. It is possible to demonstrate that the functional, which associates to each normal curve its curvature $\varkappa$ at $\mathbf{x}$, has a maximum and a minimum: these are the two local principal curvatures, $\varkappa_1$ and $\varkappa_2$. For a complete treatment of the subject we refer to [112, 129].

easy and, at the same time, accurate manner by means of the Minkowski functionals. It was verified by means of numerical simulations with a thorough comparison with other methods, such as 3D integral-equation theory and density functional theory, that the free energy, μ, of a fluid immersed in a container S is given by the linear combination

$$\mu(S) = -pV(S) + \sigma\Sigma(S) + \kappa C(S) + \bar{\kappa}X(S). \tag{4.1}$$

After computing once and for all the coefficients, p, σ, κ and $\bar{\kappa}$, for example in the case in which S has a simple geometry, the free energy of the fluid in contact with a complexly shaped wall can be immediately evaluated computing only the four elementary functionals. The coefficients, indeed, depend on the fluid-fluid and fluid-wall interactions and are given by physical quantities, since they are respectively the pressure, the surface tension, and the bending rigidities.

It is important to remark that the Hadwiger theorem does not hold if the free energy does not satisfy the three fundamental properties of Minkowski functionals. This happens in the case of critical phenomena, or if long ranged interactions are considered or if wetting or drying phenomena occur at the wall, since intrinsic lengths in such systems have a macroscopic size. Despite these restrictions, however, the theory has allowed rapid and accurate computations of the thermodynamic quantities of a confined fluid [33].

Encouraged by this results a biological application was considered [34, 117, 125]. The framework just introduced can straightforwardly be adapted to the case of a protein immersed in solution, simply considering the protein molecule, in a given configuration, as part of the solvent wall. The solvation free energy can then obtained by Eq. 4.1: after fixing the solvent-solvent and solvent-protein interaction potentials the coefficients can be computed for the case of a simple geometry and then, for each protein configuration, the solvation free energy descends immediately by the evaluation of the four basic functionals. Notice that this requires the relaxation of the condition that the morphological domain is composed by a finite union of convex sets; accurate numerical simulations have shown that this does not invalidate the reliability of the method.

In Ref. [34, 125], a protein was represented as the union of the Van der Waals spheres of its atoms. Then a menu of compact conformations obtained with all-atom molecular dynamic simulations was compared with the real native state of the protein, computing for each structure the solvation free energy with the morphological method. The results obtained show that the native state minimises the solvation free energy of the system, which confirm the importance of this quantity in determining the stability of the protein conformation and, at the same time, reinforces the usefulness of the morphometric approach.

4.2 Solvent effects for a tube-like polymer

In order to model in a simplified way two important effects arising from the interaction between the protein molecule and the solvent, the hydrophobic effect and the depletion interaction, two different effective potentials were introduced for the tube-like polymer case, respectively proportional to the excluded volume and to the buried area of the tube. A method was provided in order to compute each of the two potentials for a tube in every configuration, when it is immersed in a solution of hard spheres. Within the morphological approach we find that the two measures, defined for different purposes, can be combined in the same framework, because they turn out to be connected with the first two terms of the morphological formulation of the solvation free energy.

An adaption of the morphometric approach to the case of a tube-like polymer, was considered first in Ref. [117]. The helical configuration was the subject of that work: the solvation free energy was evaluated in varying the helical parameters in order to find the more stable structure. On the other hand the mathematical recipes illustrated in this thesis for the excluded volume and the buried area allow a more general treatment of the subject since a tube in every configuration can be considered. A merging between the tube-polymer model and the morphometric approach becomes the natural following-up of the line of research described in this thesis.

Consider an impenetrable tube folded in a compact conformation and im-

mersed in a sea of hard spheres, representing the aqueous solution. The tube can be seen as the impenetrable wall confining the hard spheres. Therefore, in order to evaluate the solvation free energy through Eq. 4.1, the four measures we need to compute are the excluded volume $V_{exc}(\Gamma)$ to the sphere centres due to the tube presence, the surface of this excluded region $\Sigma_{exp}(\Gamma)$, that we will call exposed surface, and the integrated mean and Gaussian curvature on this surface, $C(\Gamma)$ and $X(\Gamma)$. The first measure is exactly the functional introduced in the previous chapter. The second one, on the other hand, is akin to the contact surface, the complementary of the buried area; it has the only difference that instead of the surface of the tube, we now consider the surface of the excluded region, the 'soft' tube with thickness $\Delta_\epsilon \equiv \Delta + \epsilon$, where ϵ is the sphere radius and Δ the tube thickness (see 3.2.1). The criterion expressed with Eq. 2.25 in order to distinguish if a region of the tube is buried holds also in this case; it can be demonstrated moreover that, in analogy with the excluded volume, this criterion can be formulated by means of the multiplicity function[6] $n(s, \theta, \rho)$ defined by the conditions 3.15. We remind the expression of the volume element of the tube, $dV = \rho|1 - \kappa(s)\,\rho\cos\theta|\; d\rho\; d\theta\; ds$, and the domain of integration $\mathcal{D} = [0, L] \times [0, 2\pi] \times [0, \Delta_\epsilon]$. Then the excluded volume and exposed surface can be written as

$$\begin{aligned} V_{exc}(\Gamma) &= \int_{\mathcal{D}} dV \frac{1}{n(s, \theta, \rho)} \\ \Sigma_{exp}(\Gamma) &= \int_{\mathcal{D}} dV \delta(\rho - \Delta_\epsilon)\vartheta\left[\frac{3}{2} - n(s, \theta, \Delta_\epsilon)\right]. \end{aligned} \tag{4.2}$$

The computation of the last two terms of Eq. 4.1 is less immediate. The evaluation of $C(\Gamma)$ and $X(\Gamma)$ in the smooth regions of the surface is straightforward [129]; nevertheless the computation of the contributions given by edges and corners which arises when the soft tube overlap is not trivial at all. Therefore, in order to apply the theorem to a tube-like polymer, the evaluation of these curvature terms is required and represents at the present a work in progress. This will allow us to study the equilibrium configurations of the sys-

[6] It is possible to show that the equivalence between the condition 2.25 and the ones expressed by means of the multiplicity function is exact when two hemispheres are attached to both the ends of the tube.

tem as a function of the thermodynamic parameters entering Eq 4.1. Beside this analytical effort, however, we are employing numerical simulation in order to study the thermodynamics at finite temperature of the system affected only by the first two terms of the solvation free energy.

The morphometric approach provides a very simplified way to model the solvent effects. The method is eligible to be employed both in an all-atom and in coarse grained representations of the protein molecule. For this reason we retain that it could represent an important ingredient for the tube-like polymer model.

Appendix A

Simulated annealing

The basic idea of simulated annealing is to search for the minimum of a cost function depending on a large number of variables in the same way as a physical system with a large number of degrees of freedom reaches the ground state of minimum internal energy in the limit of low temperature. One may think, for example, of crystalline solids which are formed by several atoms. In practical contexts, low temperature is not a sufficient condition for finding ground states of matter. Experiments that determine the low-temperature state of a material are done by careful annealing, first melting the substance, then lowering the temperature slowly, and spending a long time at temperatures in the vicinity of the freezing point. If this is not done carefully, the resulting crystal will have many defects, or the substance may form a glass, with no crystalline order and only metastable locally optimal structures.

Coming back to the original optimization problem, one may apply the same scheme by using the cost function in place of the internal energy and then introducing a fictitious temperature, which is simply a control parameter in the same unit as the cost function. The simulated annealing procedure consists in carrying out numerical simulation by gradually decreasing the temperature from high values to lower ones, until the system has frozen in some configuration and no further changes occur. For the procedure to succeed in finding the correct globally optimal state, it is crucial that the system be in thermodynamic equilibrium along all the cooling route [100]. This ensures that the system eventually escapes any possible trapping in local minima. Of course,

this is strictly true only for an exceedingly slow cooling rate, and designing cooling schedules to minimize the time needed to reach a 'good' solution is a non-trivial task. Moreover, when the exact solution is not known 'a priori', one can never be sure that the obtained numerical solution is the optimal one.

A natural and efficient way to simulate systems in thermodynamic equilibrium at a given temperature is by means of Monte Carlo stochastic dynamics in the configuration space [130, 131]. The standard Metropolis algorithm [132] allows one to generate a stochastic process which samples randomly the configuration space with a probability proportional to the correct Boltzmann weight at the desired temperature. The basic step consists in proposing an updating of the current system configuration, based on some predetermined set of possible moves, and to accept it or not according to the Metropolis rejection test. That is, if the proposed move lowers the cost function, accept it, and if it raises the cost function, accept it only with a probability $p(\Delta E) = \exp[-\Delta E/T]$, where ΔE is the resulting change in the cost function E, and T the fictitious temperature. By repeating this procedure many times, the desired Boltzmann distribution gets to be sampled by the generated stochastic process. Again, in practice this may be true only for unfeasibly long times, and a crucial issue is to devise efficient moves to be used in the updating process. Moreover, the correct Boltzmann distribution is recovered only if the set of employed moves is ergodic, that is if each configuration can eventually be reached from each other configuration. In the following section we will describe the mostly used moves in our simulations.

The fictitious temperature T of the simulated annealing procedure has to be decreased according to an efficient cooling schedule. We used a rather common recipe, based on monitoring the acceptance rate along the cooling route, that is the fraction of the accepted moves over the tried ones. The cooling shedule depends on two different factors; the rate at which temperature is lowered, and the number of moves which are tried at a given temperature. Both parameters can be varied during the annealing procedure. At high temperatures one expects an acceptance rate very close to unity, since nearly every move is likely to be accepted. On the other hand, at low temperature, that is after the system has frozen in some configuration, the acceptance rate is practically zero. The

crucial temperature values corresponds to the often sharp crossover region, where the acceptance rate starts decreasing from unity and then drops to zero, meaning that the system is doing the crucial steps in order to reach the optimal configuration. The cooling rate has then to be slowed, and the maximum number of tried moves increased, correspondingly.

We note that the optimal chain conformations are usually obtained in different simulated annealing runs starting from random unrelated initial configurations. Nevertheless, since we are dealing with numerical simulations, it cannot be taken for granted that we hit on locally, rather than globally, optimal shapes.

A.1 Monte Carlo moves

In all the simulations described during this thesis, the configuration space we want to sample is the set $\mathcal{R}$ of discrete chains $\{\mathbf{x}_0, \ldots, \mathbf{x}_N\}$ in the continuum three-dimensional space, having a fixed constant distance b between consecutive beads, and a hard-core repulsion at a distance R_0 between different non-consecutive beads. Chain configurations can be expressed equivalently in terms of the angles $\theta_2, \theta_3\phi_3, \ldots, \theta_N, \phi_N$, with

$$\cos(\theta_i) = \frac{\mathbf{u}_i \cdot \mathbf{u}_{i-1}}{|\mathbf{u}_i|\,|\mathbf{u}_{i-1}|}\,, \tag{A.1}$$

$$\phi_i = \frac{\pi}{2} - \mathrm{sgn}\{(\mathbf{v}_i \wedge \mathbf{v}_{i-1}) \cdot \mathbf{u}_{i-1}\} \arccos\left[\frac{\mathbf{v}_i \cdot \mathbf{v}_{i-1}}{|\mathbf{v}_i|\,|\mathbf{v}_{i-1}|}\right]\,, \tag{A.2}$$

where $\mathbf{u}_i = \mathbf{x}_i - \mathbf{x}_{i-1}$, and $\mathbf{v}_i = \mathbf{u}_i \wedge \mathbf{u}_{i-1}$. The angles θ_i are the angles between consecutive bonds along the chain, that is the *valence angles*, whereas ϕ_i are basically the *torsion angles* between successive planes formed by pairs of consecutive bonds, which are usually plotted in the *Ramachandran plot* (see section 1.3). The above correspondence between angles and coordinates becomes one-to-one, when fixing $\mathbf{x}_0 = (0,0,0)$, $\mathbf{x}_1 = (0,0,b)$, $\mathbf{x}_2 \cdot \hat{e}_j = 0$, which amounts to define 'reduced' chain configurations by exploiting symmetries; translation of the first and the second bead, and rotation of the rest of chain around the segment connecting the first two beads ($\hat{e}_j = (0,1,0)$).

In the simulations described in this thesis, three different moves was usually employed, one of them involving coordinate representation and the other three involving angle representation. All kinds of moves do not violate the constraint of fixed constant distance b between consecutive beads along the chain, but we have to check whether the proposed updated configuration satisfies other constraints which we want to enforce, such as compactness, end-to-end distance or self-avoidance. When the trial configuration violates any of these constraints, it is rejected before submitting it to the Metropolis test. We now list the moves which we used and discuss their local/global character, with respect to the number of beads involved in the move, as is customarily in polymer physics [130, 131].

• *Crankshaft move*
Select randomly two beads i, j ($i < j$), such that $j - i \leq n_c + 2$, with $n_c \ll n$ (in most cases we have used $n_c = 5$). Then, rotate beads $i+1, \ldots, j-1$ of an angle $\Delta\phi_c$ around the axis $\mathbf{x}_j - \mathbf{x}_i$. The angle $\Delta\phi_c$ is chosen randomly with a uniform probability distribution in the interval $[-\Delta\phi_m/2, \Delta\phi_m/2]$. This is a *local* move, since only n_c beads are involved, and it has to be carried out in coordinate representation.

• *Reptation move*
This kind of move is also known as *slithering-snake* move. It consists in deleting a bead from one end of the chain and appending a new bead at the other end. Which of the two ends the bead is removed from is chosen randomly each time the move is tried. The orientation in which the new bead is appended is also chosen randomly. In the angle representation introduced above reptation amounts, in one case, to reshuffle angles, $\theta_i = \theta_{i+1}$ for $2 \leq i \leq N-1$, and $\phi_i = \phi_{i+1}$ for $3 \leq i \leq N-1$, and to assign a new value to θ_N and ϕ_N. In the other case, $\theta_i = \theta_{i-1}$ for $3 \leq i \leq N$, and $\phi_i = \phi_{i-1}$ for $4 \leq i \leq N-1$, and a new value is assigned to θ_2 and ϕ_3. In both cases, new values for angles are picked up randomly with a uniform probability distribution in the interval $[0, \theta_e]$ ($[0, 2\pi]$) for θ (ϕ) angles. Reptation is a *bilocal* move, since it alters two disjoint small groups of consecutive beads of the chain.

• *Pivot move*
Select randomly one bead i, with $1 \leq i \leq N-1$ as the pivot point, and then rotate the part of the chain subsequent to the pivot point while keeping fixed the rest of the chain, using the pivot point as the origin. In the angle representation, this is simply carried out by updating or θ_{i+1}, or ϕ_{i+1} (if $i = 1$ only θ_2 is updated). Which angle is to be updated is again chosen randomly. The updating rule is $\theta_{i+1} = \theta_{i+1} + \Delta\theta_p$ ($\phi_{i+1} = \phi_{i+1} + \Delta\phi_p$), where $\Delta\theta_p$ ($\Delta\phi_p$) is selected randomly with a uniform probability distribution in the interval $[-\Delta\theta_m/2, \Delta\theta_m/2]$ ($[-\Delta\phi_m/2, \Delta\phi_m/2]$). Pivot moves are *global* ones, since they involve a rearrangement of a macroscopic portion of the chain.

Quite often in our simulations a multiple chains system was considered. In this case at each step one chain is randomly selected and one of the moves listed above is performed. In order to make the sampling of the conformational space more efficient a further move, consisting in rigidly translating one chain in respect to the others, can be employed. In this case the translation vector is randomly chosen.

To conclude, we note that the efficiency of Monte Carlo dynamics may depend crucially on tuning the different control parameters which we have introduced, n_c, $\Delta\phi_m$, $\Delta\theta_m$, θ_e, possibly considering them as non-constant functions of the simulated annealing temperature. Efficiency also depends on the relative frequency of the different kinds of moves which we use.

Acknowledgments

I would like to warmly thank Amos that has introduced me to physical research. I have greatly appreciated his enthusiasm and his frank and spontaneous attitude, and I have been always fascinated by his ability to look at problems from different perspectives in oder to find the clearest way of formulating them.

A special thank to Antonio, because during this years he has followed me step by step in my research activity and he has always behaved as a true friend. I have learned a lot from his careful and patient approach to research problems.

I am very grateful to Achille. With his precious contribute and his important suggestions has allowed the realisation of the work illustrated in this thesis.

I would like to thank all the people that have supported me during my Ph.D. and more or less directly have provided opportunities of improve my attitude to research. In particular Luca, Antonio, Dario, Alessandra and Boris for their patience to answer my frequent and sudden questions about everything. Elena, who, in these last months, has been close to me with her encouragements and her happiness despite the distance, and has spent exciting evening reading my chapter 3. I think she is my best friend. My friends Francesco and Lino, who have been as a second family for me. I will never forget all the dinners at the chinese restaurant, and the endless discussions about quantum mechanical behaviour of Chinese communism, fine-tuning of intelligence in critical phenomena (with particular regard to scaling exponents of Tetris), Life, the Universe and Everything ... I am not sure I have disclosed the Mining of Life, but chinese food was very good. Finally, my deepest thank goes to Andrea for having bear me during these years and also for other thousands of

reasons, which are too much and too important to explain them here, and so I won't say anything at all, and I'll offer him a beer!

Bibliography

[1] A. B. Kolomeisky, M. E. Fisher, *Annu. Rev. Phys. Chem.* **58**, 675 (2007).

[2] H. De Jong, *J. Comp. Biol.* **9**, 67 (2002).

[3] J. J. Tyson, K. C. Chen, and B. Novak, *Curr. Opin. Cell Biol.* **15**, 221 (2003).

[4] T. E. Creighton, *Proteins: Structure and Molecular Properties*, W. H. Freeman and Company, New York (1993).

[5] C. B. Anfinsen, *Science* **181**, 223 (1973).

[6] C. Clementi, *Curr. Opin. Struct. Biol.* **17**, 1 (2007).

[7] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[8] J. K. Myers, T. G. Oas *Annu. Rev. Biochem.* **71**, 783 (2002).

[9] L. Pauling, R. B. Corey, and H. R. Branson, *Proc. Natl. Acad. Sci. USA* **37**, 205 (1951).

[10] G. N. Ramachandran, C. Ramakrishnan and V. Sasisekharan, *J. Mol. Biol.* **7**, 95 (1963).

[11] B. W. Matthews, *Annu. Rev. Biochem.* **62**, 139 (1993).

[12] C. Chotia, *Nature* **357**, 543 (1992).

[13] J. L. king, and T. H. Jukes, *Science* **164**, 788 (1969.

[14] J. D. Bryngelson, P. G. Wolynes, *P. Natl. Acad. Sci. USA* **94**, 7524 (1987).

[15] D. Baker, *Nature* **405**, 39 (2000).

[16] A. Maritan, C. Micheletti, A. Trovato, and J. R. Banavar, *Nature* **406**, 287 (2000).

[17] J. R. Banavar, T. X. Hoang, A. Maritan, F. Seno, A. Trovato, *Phys. Rev. E* **70**, 041905 (2004).

[18] D. Marenduzzo, A. Flammini, A. Trovato, J. R. Banavar, A. Maritan, *J. Polym. Sci. B Pol. Phys.* **43**, 650 (2005).

[19] J. R. Banavar, A. Flammini, D. Marenduzzo, A. Maritan, and A. Trovato, *J. Phys.: Cond. Matt.* **15**, S1787 (2003).

[20] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, *Proc. Natl. Acad. Sci. USA*, **21**, 7960 (2004).

[21] J. R. Banavar, T. X. Hoang, J. H. Maddocks, A. Maritan, C. Poletto, A. Stasiak, and A. Trovato, *Proc. Natl. Acad. Sci. USA*, **104** (2007).

[22] C. Poletto, A. Giacometti, A. Trovato, J. R. Banavar, and A. Maritan, *Physic Review E* **77**, 061804 (2008).

[23] W. Kauzmann, *Adv. Protein Chem.* **14**, 1 (1959).

[24] K. A. Dill, *Biochemistry* **29**, 7133 (1990).

[25] Y. Snir, and R. D. Kamien, *Science* **307**, 1067 (2005).

[26] Y. Harano, M. Kinoshita, *Chem. Phys. Lett.* **399**, 342 (2004).

[27] C. N. Likos, *Physics Reports* **348**, 267 (2001).

[28] J. M. Brader, R. Evans and M. Schmidt, *Molec. Phys.* **101**, 3349 (2003).

[29] S. Asakura, and F. Oosawa, *J. Chem. Phys.* **22**, 1255 (1954).

[30] S. Asakura, and F. Oosawa, *J. Pol. Sci.* **33**, 183 (1958).

[31] H. Hadwiger, *Vorlesungen über Inhalt, Oberflche und Isoperimetrie* Springer, Berlin, Germany (1957).

[32] K. Mecke, *Int. J. Mod. Phys. B* **12**, 861 (1998).

[33] P. M. König, R. Roth, and K. R. Mecke, *Phys. Rev. Lett.*, **93**, 160601 (2004).

[34] R. Roth, Y. Harano, M. Kinoshita, *Phys. Rev. Lett.* **97**, 078101 (2006).

[35] C. Branden, J. Tooze, *Introduction to Protein Structure*, Garland Publishing, New York (1991).

[36] H. J. Dyson, and P. E. Wright, *Nat. Rev. Mol. Cell. Biol.* **6**, 197 (2005).

[37] C. B. Anfinsen, *J. Biol. Chem.* **221**, 405 (1956).

[38] M. L. Anson, *Adv. Protein Chem.* **2**, 361 (1945).

[39] A. L. Horwich, *Proc. Natl. Acad. Sci. USA* **96**, 11033 (1999).

[40] C. Chothia, *Nature* **248**, 338 (1974).

[41] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, *Protein Sci.* **4**, 561 (1995).

[42] K. F. Lau, and K. A. Dill, *Macromolecules* **22**, 3986 (1989).

[43] H. S. Chan, and K. A. Dill, *Macromolecules* **22**, 4559 (1989).

[44] A. Sali, E. Shakhnovich, and M. Karplus, *Nature* **369**, 248 (1994).

[45] H. Li, R. Helling, C. Tang, and N. S. Wingreen, *Science* **273**, 666 (1996).

[46] C. Micheletti, J. R. Banavar, A. Maritan, and F. Seno, *Phys. Rev. Lett.* **80**, 5683 (1998).

[47] L. Pauling, and R. B. Corey, *Proc. Natl. Acad. Sci. USA* **37**, 235 (1951).

[48] L. Pauling, and R. B. Corey, *Proc. Natl. Acad. Sci. USA* **37**, 251 (1951).

[49] L. Pauling, and R. B. Corey, *Proc. Natl. Acad. Sci. USA* **37**, 272 (1951).

[50] L. Pauling, and R. B. Corey, *Proc. Natl. Acad. Sci. USA* **37**, 729 (1951).

[51] Y. Ramachandran, D. Vitkup and M. Karplus, *J. Mol. Biol.* **285**, 1371 (1999).

[52] M. A. Savageau, *Proc. Natl. Acad. Sci. USA* **83**, 1198 (1986).

[53] R. A. George and J. Heringa *J. Mol. Biol.* **316**, 839 (2002).

[54] M. Gerstein, A.M. Lesk and C. Chothia *Biochemistry* **33**, 6739 (1994).

[55] A. G. Murzin, S. E. Brenner, T. J. P. Brenner, and C. Chothia, *J. Mol. Biol.* **247**, 536 (1995).

[56] http://scop.mrc-lmb.cam.ac.uk/scop/

[57] http://www.cathdb.info/index.html

[58] http://www.pdb.org

[59] C. a. Orengo, D. T. Jones and J. M. Thornton *Nature* **372**, 631 (1994).

[60] *UniParc sequence archive*, http://www.uniprot.org/uniparc/

[61] B. van den Berg, R. Wain, C. M Dobson, R. J. Ellis, *EMBO J.* **19**, 15, 3870 (2000).

[62] M. S. Cheung, D. Klimov, D. Thirumalai, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 4753 (2005)

[63] N. G. Hunt, L. M. Gregoret, and F. E. Cohen, *J. Mol. Biol.* **241**, 214 (1994).

[64] G. I. Makhatadze, and P. L. Privalov, *Adv. Protein Chem.* **47**, 307 (1995).

[65] R. L. Baldwin, *J. Biol. Chem.* **278**, 17581 (2003).

[66] G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan, *Proc. Natl. Acad. Sci. USA* **103**, 16623 (2006).

[67] C. Levinthal, *J. Chem. Phys* **65**, 44 (1968).

[68] T. E. Creighton, *Biochem. J.* **270**, 1 (1990).

[69] P. E. Leopold, M. Montal, and J. N. Onuchic, *Proc. Natl. Acad. Sci. USA* **89**, 8271 (1992).

[70] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, *Science* **267**, 1619 (1995).

[71] K. A. Dill and H. S. Chan, *Nat. Struct. Biol.* **4**, 10 (1997).

[72] D. Sherrington, S. Kirkpatrick, *Phys. Rev. Lett.***35**, 1792 (1975).

[73] V. Villegas, J. C. Martinez, F. X. Aviles, L. Serrano, *J. Mol.Biol.* **283**, 1027 (1998).

[74] J. C. Martinez, L. Serrano, *Nat. Struct.Biol.* **6**, 1010 (1999).

[75] D. S. Riddle et al., *Nat. Struct.Biol.* **6**, 1016 (1999).

[76] K. F. Lau, K. A. Dill, *Macormolecules* **22**, 3986 (1989)

[77] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, M. H. Hecht, *Science* **262**, 1680 (1993).

[78] H. S. Chan, and K. A. Dill, *J. Chem. Phys.* **92**, 3118 (1990).

[79] H. S. Chan, and K. A. Dill, *Proc. Natl. Acad. Sci. USA* **87**, 6388 (1990).

[80] L. M. Gregoret, and F. E. Cohen, *J. Mol. Biol.* **219**, 109 (1991).

[81] D. P. Yee, H. S. Chan, T. F. Havel, and K. A. Dill, *J. Mol. Biol.* **241**, 557 (1994).

[82] N. D. Socci, W. S. Bialek, and J. N. Onuchic, *Phis. Rev.* **E49**, 3440 (1994).

[83] H. Li, C. Tang, and N. S. Wingreen, *Proc. Natl. Acad. Sci. USA* **95**, 4987 (1998).

[84] T. Wang, J. Miller, N. S. Wingreen, C. Tang, and K. A. Dill, cond-mat/0006372 (2000).

[85] N. E. G. Buchler, and R. A. Goldstein, *Proteins: Struct. Funct. Gen.* **34**, 113 (1999)

[86] C. Micheletti, J. R. Banavar, A. Maritan, and F. Seno, *Phys. Rev. Lett.* **82**, 3372 (1999).

[87] A. Maritan, C. Micheletti, and J. R. Banavar, *Phys. Rev. Lett.* **84**, 3009 (2000).

[88] G. Buck, and J. Orloff, *Topol. Appl.* **61**, 205 (1995).

[89] V. Katritch, J. Bednar, D. Michoud, R. G. Scharein, J. Dubochet, and A. Stasiak, *Nature* **384**, 142 (1996).

[90] O. Gonzalez, and J. H. Maddocks, *Proc. Natl. Acad. Sci. USA* **96**, 4769 (1999).

[91] T. C. Hales, *Discrete Computational Geom.* **17**, 1 (1997).

[92] T. C. Hales, *Discrete Computational Geom.* **18**, 135 (1997).

[93] B. Cipra, *Science* **281**, 1267 (1998).

[94] A. Stasiak, J. H. Maddocks, *Nature* **406**, 251 (2000).

[95] A. Trovato, *Ph. D. Thesis*, Università SISSA (2000).

[96] J. R. Banavar, A. Flammini, D. Marenduzzo, A. Maritan, and A. Trovato, *ComPlexUs*, 1:4 (2003).

[97] V. Katritch, W. K. Olson, P. Pieranski, J. Dubochet, and A. Stasiak, *Nature* **388**, 148 (1997).

[98] I. Stewart, *Sci. Am.* **278**, 80, (Feb. 1998).

[99] R. A. Litherland, J. Simon, O. Durumeric, and E. Rawdon, Vol. 91, *Topology and Its Applications*, (Elsevier Science, Amsterdam, 1999).

[100] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, *Science* **20**, 671 (1983).

[101] M. Tinkham, *Introduction to Superconductivity*, McGraw-Hill, New York (1996).

[102] J. R. Banavar, O. Gonzalez, J. H. Maddocks, A. Maritan, *J. Stat. Phys.*, **110**, 35 (2003).

[103] Y. Zhou, C. K. Hall, and M. Karplus, *Phys. Rev. Lett*, **77**, 2822 (1996).

[104] F. H. T. Allain, M. Yen, J. E. Masse, P. Schultze, T. Dieckmann, R. C. Johnson, and J. Feigon, *Embo J.* **18**, 2563 (1999).

[105] K. Y. Chang, and I. Tinoco, *J. Mol. Biol.* **269**, 52 (1997).

[106] U. Baumann, S. Wu, K. M. Flaherty, and D. B. Mckay, *Embo J.*

[107] R. L. Baldwin, *Proc. Natl. Acad. Sci. USA* **83**, 8069 (1986).

[108] P. L. Privalov, *Annu. Rev. Biophys. Biophys. Chem.* **18**, 47 (1989).

[109] C. Tanford, *The Hydrophobic Effect*, 2nd ed., J. Wiley, New York (1980).

[110] A. Möglich, K. Joder, and T. Kiefhaber, *Proc. Natl. Acad. Sci. USA* **103**, 12394 (2006).

[111] B. Park, and M. Levitt, *J Mol Biol* **258**, 367 (1996).

[112] H.S.M. Coxeter, *Introduction to Geometry*, J. Wiley, New york (1989).

[113] Y. Snir, and R. D. Kamien, *Phys. Rev. E* **75**, 051114 (2007).

[114] D. Frenkel, *Physica A* **263**, 26 (1999).

[115] L. Onsager, *Ann. N. Y. Acad. Sci.* **51**, 627 (1949).

[116] A. Vrij, *Pure Appl. Chem.* **48**, 471 (1976).

[117] H. Hansen-Goos, R. Roth, K. Mecke, and S. Dietrich, *Phys. Rev. Lett.* **99**, 128101 (2007).

[118] P. N. Pusey, 1991, *Liquids, Freezing and Glass Transition*, J. P. Hansen, D. Levesque and J. Zinn-Justin (Amsterdam: Elsevier Science), Chap. 10.

[119] E. J. Meijer and D. Frenkel, *J. chem. Phys.*, **100**, 6873, (1994).

[120] H. N. W. Lekkerkerker, W. C. K. Poon, P. N. Pusey, A. Stroobants, and P. B. Warren, *Europhys. Lett.*, **20**, 559, (1992).

[121] J. L. Barrat, and J. P. Hansen, *Basic Concepts for Simple and Complex Liquids* Cambridge University Press, Cambridge (2003).

[122] M. Dijkstra, J. M. Brader and R. Evans, *J. Phys.: Condens. Matter* **11**, 10079 (1999).

[123] D. Marenduzzo, K. Finan, and P. R. Cook, *The Journal of Cell Biology* **175**, 681 (2006).

[124] L. A. Santaló, *Integral Geometry and Geometric Probability*, Addison-Wesley, Reading, MA, (1976).

[125] Y. Harano, R. Roth, M. Kinoshita, *Chem. Phys. Lett.* **432**, 275 (2006).

[126] K. R. Mecke, Th. Buchert, and H. Wagner, *Astron. Astrophys.* **288**, 697 (1994).

[127] C. N. Likos, K. R. Mecke, and H. Wagner, *J. Chem. Phys.*, **102**, 9350 (1995).

[128] G. Rota, Lezioni Lincee held at the Scuola Normale Superiore, Pisa, (1986).

[129] B. O'Neill, *Elementary differential geometry*, Academic Press, New York (1966).

[130] A. D. Sokal, *Nuclear Physics* **B47**, 172 (1996).

[131] K. Binder, *Monte Carlo and molecular dynamics simulations in polymer science*, (Oxford University Press, 1995).

[132] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).