SISSA ISAS

SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI
INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

A Geometric Perspective on Protein Structures and Heteropolymer Models

Thesis submitted for the degree of

Doctor Philosophiæ

Candidate:
Antonio Trovato

Supervisor:
Prof. Amos Maritan

October 2000

Contents

Introduction

The explosive development of molecular biology has revealed fundamental similarities in the mechanisms by which all cells operate in any living being.

Two fundamental classes of biopolymers exist in living organisms, proteins and nucleic acids, RNA and DNA. They are linear heterogeneous polymeric chains, assembled by means of successive repetition along the chain of different basic structural units. A restricted set of twenty aminoacids constitute the universal building blocks of the polypeptide chains which form proteins, while RNA and DNA molecules are each constructed from just four type of nucleotides.

Proteins virtually control and regulate all properties characterizing an organism, whereas the genetic hereditary information, mostly consisting in instructions for the expression and the assembling of proteins, is encoded in the linear sequence of nucleotides of one strand of DNA double helix.

Proteins are responsible for determining the shape and the structure of the cell, for storing and transportation of ions, electrons, atoms and molecules, for signalling and communications within different cells. Proteins serve as the main instruments of molecular recognition and enzymatic catalysis, and are crucial components of muscles and other systems for converting chemical energy into mechanical energy. They also provide the architectural structure of the materials that form hair, nails, feathers, tendons, and bones of animals.

The crucial factor which allows the amazingly almost perfect accomplishment of such a series of complex and often interwoven tasks is the *functional specificity* of proteins. Each different protein, that is each different sequence of aminoacids which may be expressed in a cell under genetic control, carries out its own specific biological function. This remarkable ability of proteins resides into a well-defined three-dimensional conformation, uniquely characteristic for each protein sequence [1], called the *native state,* which determines how the protein will interact with other molecules and respond to the environment conditions.

Genetic information stored in the nucleotide sequence of DNA can be copied and transmitted by simply replicating one DNA strand, thanks to the base-pairing mechanism. The expression of a gene, that is a portion of DNA sequence codifying for a specific protein, involves a rather complicated process, mediated by enzymes and by different types of RNA molecules, which culminates in the synthesis of the protein on a ribosome. The aminoacid sequence making up the synthetized protein is specified by the nucleotide sequence of the messenger-RNA through *codons*, different triplets of nucleotides which encode specific aminoacids.

The functional specificity of proteins, due to their unique three-dimensional structure, and the reproducibility of genetic information contained in the linear nucleotide DNA sequence are the two crucial properties essential for evolution. In this respect, complementary roles are played by *linear sequences* (nucleotide and aminoacid sequences can be considered as informationally equivalent) and *three-dimensional structures*, since they represent respectively the *genotype*, the hereditary information transmitted to successive generations, and the *phenotype*, the expression of the genetic information on which natural selection operates.

A crucial step in the whole process of genotypic expression is the *folding* of proteins into their native conformations, when the mono-dimensional information stored in the aminoacid sequence is converted into the three-dimensional structure which determines the specific biochemical functions of that protein.

Protein folding is a pure physical-chemical process, since most small globular proteins are able to fold *spontaneously* ‘in vitro’, without any assistance from cellular machinery, as was first shown in a famous experiment by Anfinsen [2]. Denatured unfolded proteins, which are biologically not active because of elevated temperature or not appropriated pH/salt conditions, restore their functionality upon return to the proper conditions in solution. Folded native conformations of globular proteins are compact, in order to shield most of the hydrophobic residues in the core of the folded structure, leaving most of the polar and charged side chains in contact with water molecules on the outer surface of the protein [3, 4]. Anfinsen experiment provides the main reason for the statistical mechanics approach to the study of protein folding, since it can be interpreted by assuming the native state conformation of a protein to be the free energy minimum of the system composed by the polypetide chain and the solvent molecules. The folding of larger proteins is instead often facilitated by ‘molecular chaperones’, which prevent improper protein aggregation [5].

Two fundamental questions regarding protein folding can be posed, according to whether sequences or structures are looked for:

- Given a sequence, is it possible to predict the structure which that sequence will fold into (folding problem)?
- Given a target structure, *design* the sequence which will efficiently fold into that structure (inverse folding or design problem).

Both issues are of paramount conceptual and practical importance. An accurate method for protein structure prediction, which is still lacking, would avoid the long and expensive X-ray cristallography or nuclear magnetic resonance experiments currently needed to determine proteins structures. This will be the more and more needed, as a greater amount of DNA sequence information is provided by large-scale genomic sequencing, and requires the determination of the functions of the encoded proteins. On the other hand, an efficient design method would be invaluable in the preparation of new drugs and viral inhibitors.

The design of novel protein sequences mimic the selection process carried out by natural evolution on existing proteins. Genetic information is exploited efficiently only if the protein aminoacid sequence is able to fold into its native structure in a *fast* and *reproducible* way. The evolutionary pressure consequently exerted should have selected those sequences and structures (genotype and phenotype) having the proteinlike features necessary to the accomplishment of specific biological functions.

What makes protein sequences and structures different from random ones? Is there any simple physical principle underlying this selection process, or are they merely an arbitrary and random outcome of evolution [6]? If a selection principle exists, did it operate in sequence or in structure space? Clearly, any answer to this questions is likely to shed light into the understanding of the folding process.

The issue of *sequence selection for a given target native structure* has been intensively studied in the last decade [7], and a selection principle has emerged, known as *minimal frustration* [8, 9]. A typical random sequence of aminoacids is very frustrated, meaning that there is no state where all interactions are optimized simultaneously. Chain connectivity is indeed constraining two consecutive monomers with possibly opposite affinities to remain in the same environment. Frustration results in a very rugged energy landscape, containing a large number of deep local minima which closely compete with the ground state, the putative native structure. Random frustrated sequences can easily get trapped in a low-energy state and thus are not good folders [10].

The special protein sequences selected by natural evolution are instead minimally frustrated. This imposes a relatively smooth funnel shape to the energy landscape, with a high energy gap between the ground state and the first excited state structurally unrelated to it, thus making easier the reaching of the native state [11]. The so-called energy-landscape theory easily explains the major experimental findings of recent years about folding kinetics of small (50-100 aminoacids) globular proteins [12, 13, 14, 15, 16, 17].

Many different proteins are known to have a *common fold*, that is they fold into approximately the same structure [18]. Common folds occur even for proteins with different biological functions [19], and the total number of known folds is extremely small when compared with the number of known protein sequences [20]. Then, it seems natural to conjecture that structures played a more important role than sequences in the evolutionary process, but the issue of *structure selection* has remained so far more elusive.

Protein structures form indeed a very special class among all possible compact conformations of a polypeptide chain [21]. Several folding patterns recur repeatedly in portions of different protein molecules, which are known as secondary structures [22]. Two motifs are particularly common, β-sheets and α-helices, both characterized by an extremely high degree of geometrical regularity [23]. The elegant deduction by Pauling *et al.* [24] of the correct geometry (3.6 residues per turn) of the α-helix structure, by simply minimizing the energy of intra-chain hydrogen bonds would imply that energetic considerations are needed in order to explain the emergence of secondary structures.

Nevertheless, very recent works have increasingly emphasized the role of the *topology* of the native state [25]. Folding rates and mechanisms observed experimentally are largely determined by purely geometrical considerations [26, 27, 28], without invoking knowledge of detailed chemistry, and new selection principles in structure space have been proposed, which lead to the emergence of secondary structures in coarse-grained reduced representations of protein molecules [29, 30].

Motivated by these recent developments, a major part of this thesis is devoted to the investigation of which polymer chain conformations in the *three-dimensional continuum space* are selected from a simple *geometrical variational principle*. We look for those *best packing* chain conformations which may be wrapped up in the smallest possible volume.

On one hand, this is intended to verify the possible emergence of secondary structures when looking for chain conformations having a *maximum geometrical accessibility*, a property which seems to characterize the native structures of natural proteins with respect to random conformations having the same compactness [29]. On the other hand, the best packing problem for strings, which we study in this thesis, is a natural and fascinating generalization of the classic best packing problem for hard spheres, first posed by Kepler [31].

Maximally accessible structures are intuitively related to best packing, since in both cases ‘space occupation’ is optimized. This is more clearly formalized (chapter 2), by introducing the concept of *thickness*, which we borrow from knot theory [32, 33]. The thickness of a curve is defined as the *maximum* critical radius of a tube uniformly swollen around the curve, above which the tube cannot grow anymore due to either local bending of the curve or to the spatial proximity of different portions of the curve. Best packing structures can be indeed characterized as having *maximum thickness* among those sharing the *same fixed compactness*.

We thus look for optimal chain conformations having maximum thickness, and perform numerical simulations of discretized strings of beads in the three-dimensional continuum space, by means of the simulated annealing technique, a classic method when dealing with the optimization of a high number of degrees of freedom [34]. Due to the huge conformational space which a chain has to sample, our simulations remain feasible and reliable just up to chain lengths of a few tens of beads.

A crucial issue is to determine how compactness is enforced. There are two ways of controlling the compactness of a polymer chain that are currently used in the literature [35]. The first one, geometrical in nature, makes use of the *gyration radius* of the chain, that is the average monomer distance from the chain centre of mass. The second method is more physical, and involves the counting of *close contacts* between different non-consecutive beads, where the introduction of a cutoff length defining close contacts is needed. The smaller the gyration radius, or the more the close contacts, the more compact the chain.

We have used both methods, and the optimal structures that we find, do indeed vary depending on them. When the gyration radius of the chain is constrained (chapter 3), optimal structures in an intermediate compactness regime are *portion of helices*, and *saddle-shaped* conformations are found to be close competitors for optimality. If the same constraint is enforced *locally* (chapter 4), optimal conformations are long perfect *helices* with many turns, characterized by a particular pitch/radius ratio. Amazingly, the *same* geometry is observed in helices appearing in naturally occurring proteins [36]. When the minimum number of close contacts is constrained (chapter 5), the optimal conformations which we find are *planar zigzag hairpin shapes*, much resembling the geometry of β-sheets appearing in proteins, for a *low* number of contacts, and again *helical shapes* for a *higher* number of contacts.

To summarize, despite the complex atomic chemistry associated with the hydrogen bond and the covalent bonds along the polypeptide chain, a simple geometrical variational principle, optimal packing, seems to select different among the repeatedly occurring structural motifs in natural proteins.

In the second part of this thesis, we study a heteropolymer model in the presence of an interface.

The study of heteropolymer models, that is of the properties of a *typical random sequence*, may be considered as a backstage preparation, which is however indispensable for a proper understanding of the main role played by the peculiar features differentiating random and designed protein sequences, as we have discussed above. Heteropolymer models also provide a fascinating theoretical framework by themselves, in the wider context of the statistical mechanics of disordered systems [37, 38, 39]. Moreover, apart from heteropolymeric biomolecules, there is an enormous class of materials usually referred to as *copolymers*, consisting of polymer chains of two types of monomer that may appear in the chain in blocks of each type. The widest class of periodic block copolymers with techonological applications are the *diblock* copolymers [40, 41]. As a particular case, they are used as reenforcement agents at interfaces separating two immiscible polymeric fluids [42]. Interestingly, random copolymers have been found to be more effective than diblock copolymers in their reenforcement effect [43].

Partially motivated by this, we have studied a random heteropolymer model in the presence of an interface separating a polar from a non-polar solvent. Our main motivation is however a first attempt in the construction of a heteropolymer model, where one is trying to set the basis for a description of the properties of *membrane* proteins, within the framework discussed above.

A chain composed of *hydrophobic* and *hydrophilic* (polar or charged) components in a polar (aqueous) solvent evolves toward conformations where the hydrophobic part are buried in order to avoid water, whereas the polar part is mainly exposed to the solvent This is what commonly happens to globular proteins and makes them soluble in aqueous solutions. However other proteins, for example structural proteins, are almost insoluble under physiological conditions and prefer to form aggregates. Many of the proteins which are insoluble in water are segregated into cell membranes, which have a lipid bilayer structure. Membrane proteins have a biological importance at least as great as those which are water soluble. Structure determination of membrane proteins is an experimental task even more complex and difficult than for globular proteins. Only very few (less than fifteen) membrane protein structures have been resolved, most of them just very recently. Correspondingly, also a proper theoretical modeling is just in its early stages [44, 45].

The simplest theoretical approach to the above problems has been proposed by Garel *et al.* [46]. A chain with random hydrophobic-hydrophilic charges is considered in the presence of an interface separating a polar from a non polar solvent. In the case of membrane proteins the finite layer of lipidic environment is modeled as an infinite semi-space. Though a quite rough approximation, this is the simplest attempt in capturing the relevant features due to the competition of different selective effects.

Since the stimulating paper by Garel *et al.*, this model has attracted much attention, mostly in very recent years. It has been studied with a variety of different techniques; molecular dynamics [47], Monte Carlo simulations [48, 49], renormalization group computation [50], different variational approaches [51, 52, 53], dynamical treatment of the quenched disorder [54], scaling approaches [55, 56, 57]. In all cases a *localization transition* is found, from a high temperature region, where the chain is delocalized, to a low temperature region, where the chain is localized at the interface, trying to accomodate both kind of monomers in their preferred side. The transition temperature diverges as the neutrality of the chain is approached.

In chapter 6, we firstly perform a *Gaussian variational computation* in replica space for an ideal hydrophobic-hydrophilic chain in the continuum three-dimensional space [58]. Within this approach, we confirm the same general picture as above and we give an explicit expression for the *localization length* of the chain at all temperatures. We also address the problem of *replica symmetry breaking*, and we find the replica-symmetric solution to be unstable. Secondly, we obtain *exact bounds* for the quenched free energy of an analogous lattice model [59]. The bounds allow one to prove exactly the above scenario, that is a *neutral* random polymer is localized near the interface at any temperature, whereas a *non-neutral* chain is shown to undergo a delocalization transition at a finite temperature. These results are valid for *both ideal and self-avoiding* chains, and also for a quite general *a priori* probability distribution for both independent and correlated hydrophobic charges. The latter case is the relevant ones for protein sequences, which have been designed by natural evolution.

Chapter 1

Proteins and heteropolymers

The remarkable molecular unity in the living world is underlying the amazing diversity that we see at the macroscopic level. All living organisms, from the smallest viruses to plants and whales, are similar at the molecular level. They all use the same twenty aminoacids in their proteins, and the same nucleotides in their DNA and RNA. Proteins in bacteria and in men are characterized by the same basic properties, which we will describe in section 1.1.

The fundamental fact about proteins, from both a biological and a physical point of view, is their ability to attain a specific three-dimensional *folded* conformation, the *native state*, in phisiological conditions [5].

Correct *folding* into the native state is indeed essential for proteins in order to accomplish their biological functions. The latters depend on the direct physical interaction of proteins with other molecules, which in turn relies on the specific three-dimensional conformation attained in the native state. The *specificity* of such interactions is of crucial importance; in the crowded interior of a cell, each protein must interact only with the appropriate molecules, and not with any of the others that are present, often in extremely high concentration.

On the other hand, protein folding is a remarkable physical process. As we will see in subsection 1.1.3, protein molecules are rather flexible, and have thus at their disposal a huge conformational space, with a number of accessible conformations which increases exponentially with the number of aminoacids, becoming soon astronomically large. As was first pointed out by Levinthal [60], it is thus quite puzzling that a protein be able to fold *always* into the native state in very short times, ranging from $10^{-3}s$ to $10^{-1}s$, without being trapped in an endless search.

Moreover, after the classical works of Anson and Anfinsen on the *reversibility* of protein unfolding [61, 2], it was commonly recognized that the folding of the protein polypetide chain is a *thermodynamically driven* process, and that the unique native structure of proteins represents the energetic balance of various types of interactions between protein groups, and between these groups and the surrounding medium, usually water. In other words, the native state is believed to be the thermodynamic *ground state* of the system composed of the protein chain and the solvent molecules. This provides a firm ground for a statistical mechanics approach to protein folding [62].

As we will see in section 1.2, a sequence of randomly chosen aminoacids has physical properties dramatically different from natural protein sequences. Protein sequences are believed to have been selected precisely to fold rapidly and correctly into their native three-dimensional structure [9]. A principle underlying the selection process has emerged, *minimal frustration*, according to which the selected sequences have an optimized energy landscape allowing fast and correct folding [8, 15]. A typical random sequence is instead very *frustrated* and remains easily trapped in metastable states, and thus does not fold in a fast and reproducible way [10]. The properties of *heteropolymers* closely resemble those of *spin glasses*, which are the reference model in the statistical mechanics of disordered systems [63]. As we will see, the same analytical tools, for example the so-called *replica trick*, can be applied [64, 65, 66].

Protein structures are also very peculiar among all possible conformations attainable by a polypeptide chain. Almost all natural proteins exhibit extremely regular folding patterns known as *secondary structures* [22]. The main subject of this thesis is an investigation of whether and how a selection process may have occurred also for protein structures, on the basis of *purely geometrical properties*. In section 1.3, we will address this point in a preliminary way, by reviewing recent works in the literature, that have already investigated the possible role of *packing effects* and other geometrical factors, such as native state geometrical accessibility and fast folding, in the formation of secondary structures. Finally, we will introduce and motivate our approach based on *optimal packing*, that we will investigate in the following chapters.

1.1 Properties of proteins

Since the discovery and the resolution of many protein structures, it has become customary to distinguish several levels of organization in the structure.

The *primary structure* is the chemical sequence of aminoacids along the polypeptide chain. We will briefly describe the associated chemistry in subsection 1.1.1.

Local ordered motifs occurring in most of known protein structures are known as *secondary structures*. They were first predicted by Pauling and Corey on the basis of energy considerations [24, 67, 68, 69, 70]. The different secondary structures occurring in natural proteins are described in subsection 1.1.3.

The compact packing of secondary structures determines the unique full three-dimensional native conformation of a biologically active protein. It is also referred to as the *tertiary structure*. It is the result of the delicate tuning of various kinds of physical interaction, occurring between different atoms of the protein chain and between the latters and the solvent molecules. Such interactions will be briefly described in subsection 1.1.2.

Most natural proteins in solution have roughly spherical shapes, and thus are usually referred to as *globular* proteins. Large proteins exist, composed of smaller parts called *domains*, that is compact globular regions, separated by a few aminoacids. The domain arrangement with respect to one another is called the *quaternary structure*.

Proteins normally exist in solution or embedded in membranes. We will devote subsection 1.1.4 to a brief description of the peculiar features of membrane proteins, since we will study related heteropolymer models in chapter 6.

The physical properties of globular proteins in the folded native conformation do not change, or change very little, when the environment is altered by changes in temperature, pH, or pressure. Upon increasing temperature or pressure, or varying pH, a point is eventually reached at which there is a sudden drastic change, *denaturation*, invariably associated with a loss of biological functions.

Depending on chemical conditions, there exist two different denatured phases. In the *coil* phase, the polypeptide chain has no definite shape, being characterized by a large conformational entropy and a highly increased size, as opposed to the unique conformation attained in the compact native state. The coil phase may be well described as the swollen phase of a homopolymer in a good solvent [71, 72]. In the *molten globule* phase, the polypeptide chain is slightly less compact than in the native state, but it still does not have a well defined structure, being characterized by a finite large conformational entropy. The average content of secondary structure is similar to that of the native state, but not necessarily in the same position. The molten globule phase can be described as the collapsed phase of a homopolymer in a bad solvent [71, 72].

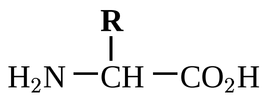
The unfolding of most small proteins is spontaneously reversible ‘in vitro’, as was first shown by Anfinsen, showing that it is an equilibrium thermodynamic process [73]. Thermodynamic analysis has revealed that the unfolding transition can often be a two-state cooperative phenomenon, with only the native fully folded and the denatured fully unfolded states present. Partially unfolded structures are unstable relative to both states. With abuse of language, the folding transition is called *first order*. Exceptions to two-state transitions usually occur in large multidomain proteins, in which the domains undergo two-state transitions independently. Such large proteins need also the aid of *molecular chaperones* in order to fold correctly ‘in vivo’ and avoid improper aggregation of the ‘wrong’ domains [74].

As a result of the complex thermodynamic properties of water, some small proteins are observed to unfold upon temperature decreasing, undergoing the so-called *cold denaturation* [75].

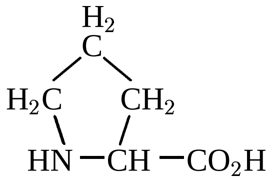
1.1.1 Chemistry of proteins

Despite the enormous variety of their biological functions, proteins are a relatively homogeneous class of molecules. From a physico-chemical point of view, proteins are heteropolymers, made up of different species of *aminoacids*, which can be chosen from twenty different species.

The generic chemical structure of the aminoacids occurring in natural proteins is:



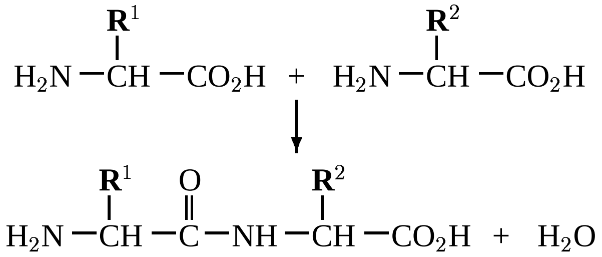
where H_2N is the amino group and CO_2H is the acidic group. The twenty aminoacids differ only in the chemical structure of the *side chain* **R**, except for *prolyne*, whose side chain is bonded also to the nitrogen atom to form an imino group:



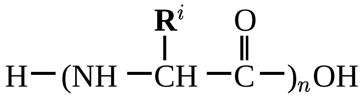
The central carbon atom to which the side chain is bonded is called *α-carbon*.

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 side chain, contains both a carbon aromatic ring and an indole ring, with one nitrogen atom. 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.

Proteins are formed by polycondensation of different aminoacids. The chemical binding of two aminoacids produces the *peptide bond*, with release of a water molecule, as illustrated below:



Each natural protein is characterized by an exact sequence of aminoacids, with a length ranging from approximately 50, for small globular proteins, to 3000, for complex multi-domain proteins. All aminoacids forming a specific protein are linked by means of peptide bonds to form a linear *polypeptide chain*, which may be considered as a weakly branched polymer, due to the presence of the side chains. The basic repeating unit of the polypeptide chain is referred to as an aminoacid *residue*:



The polypeptide *backbone* consists of the repetition of the basic unit common to all residues, that is of the three atoms N, C^{α}, and C', as they are usually indicated. As we will see, a popular coarse-graining method in protein modeling consists in representing the polypeptide backbone by retaining only the α-carbon atoms, which are the hinges of the backbone.

1.1.2 Interactions in proteins

All physical interactions occurring between the various atoms present in the system, that is the atoms composing the polypeptide chain and the solvent molecules, are Coulomb electrostatic interactions at the microscopic level. Such a microscopic approach has only very recently been tackled within a fully *ab initio* quantum molecular dynamics method. Simulation of even very small peptides, however, are computationally very intensive, and the study of a whole protein still requires the introduction of semi-empirical classical interactions at a macroscopic level, which can then be included in more traditional methodologies, such as energy minimization, force-field molecular dynamics, Monte Carlo simulations.

Such macroscopic interactions are usually divided into *covalent* and *non-covalent* interactions, according to their typical energy scale. The chemical structure of proteins, reviewed in the previous subsection, is determined by *covalent bonds*. Covalent interactions, which may comprise also *sulfur bridges* forming between the sulfur atoms of the cysteine residues, involve a typical energy ranging from 50 kCal/mole to 150 kCal/mole. At room temperature $K_B T \simeq 0.6$ kCal/mole, implying that covalent interactions simply freeze the corresponding degrees of freedom at their minimum energy value.

The complex three-dimensional structure of the folded native state is instead the result of the delicate interplay of *non-covalent* 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. 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 thermodynamical 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 ionized 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 participate in electrostatic interactions, since atoms with different electronegativities cause the electron density to be localized, 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:

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

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 first term in equation (1.1) models the *repulsion* that eventually takes place whenever two atoms approach each other, due to overlapping of electron orbitals. The repulsive energy rises so steeply, that it is common practice to model individual atoms as *hard impenetrable spheres*, characterized by an excluded volume defined by the *Van der Waals radius*. The minimum allowed distance between two atoms is thus the sum of their respective Van der Waals radii, if they are not covalently bonded. When using the potential (1.1), the sum of the Van der Waals radii may be defined as the value r_* such that $E\left(r_*\right)=0$. 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. Typical values of Van der Waals radii are 1.55 Å for nitrogen, and 1.75 Å for carbon, the two species occuring in the polypeptide backbone.

The second term in equation (1.1) models the weak short-range *attraction* felt by all atoms and molecules, even in the absence of ionized or partial charged groups, as a result of transient induced polarization effects. The optimal distance for the interaction of two atoms, corresponding to the mininum of the potential (1.1), is usually 0.3-0.5 Å greater than the sum of their Van der Waals radii.

A *hydrogen bond* 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. Hydrogen bonds in proteins most frequently involve the carboxilic C=O and amino N–H groups of the polypeptide backbone, and the $H \cdots O$ distance is most often 1.9-2.0 Å. They are believed to be responsible for the energetic stabilization of secondary structure patterns, such as helices and sheets [24, 70]. 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.

Folded conformations of proteins usually occur in a liquid-water environment or in membranes. In spite of water's biological importance, it is not one of the best understood liquids. 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. The interactions with water are not so favourable, however, for nonpolar groups, basically because they cannot participate in the hydrogen bonding, which appears to be very important in liquid water [76]. This causes a much more favourable effective interaction among nonpolar groups, which has come to be known as the *hydrophobic effect* [3, 77], than would be the case in other solvents. The *hydrophobicity* of a molecule is defined as the free energy of transfer from water to a nonpolar liquid (the more hydrophobic molecules have the more negative hydrophobicities). The strength of the hydrophobic effect has the unusual property of decreasing at lower temperature [78], due to the increasing tendency of water to form ordered structures. The complex thermodynamics caused by water ordering effects is believed to be responsible for the *cold denaturation* transition which is observed for some proteins [75].

The hydrophobic effect is believed to be the driving force leading to the collapse of globular proteins in compact folded conformations [3, 4]. *Hydrophobic* nonpolar groups will be buried inside the globule, in order to get shielded from contact with water molecules, whereas *hydrophilic,* that is charged and polar, groups will be on the outer surface, in contact with water. Nevertheless, by looking at structures of real protein, 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 [79].

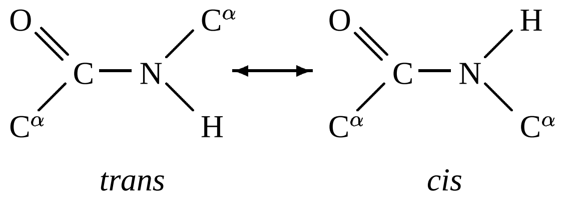
The hydrophobicities of the single side chains have been measured experimentally in a variety of ways [80]. The classification of different residues according to their hydrophobocity is at the basis of several popular models, as those that we will study in chapter 6, where only two types of residues, hydrophobic and hydrophilic, are considered [81, 82, 83, 79, 84, 85].

1.1.3 Protein conformations

Following the separation in energy scales described in the previous subsection, two types of degrees of freedom can be distinguished in proteins. Covalent bonds, including the peptide bond, and *valence angles* between two covalent bonds, are *hard* degrees of freedom. They are very rigid at room temperatures, and their values are fixed by the laws of chemistry (or of quantum mechanics). The *torsion angles* along the backbone chain and the side chains are *soft* degrees of freedom, since they can easily fluctuate at room temperature. They are responsible for the conformational *flexibility* of both the polypeptide backbone and the side chains. The torsion angles, or *dihedral* angles, are the rotation angles about covalent bonds of one portion of the chain (see appendix A for a mathematical definition of the dihedral angles in the case of α-carbon reduced representation).

Rotation about the N–C$^\alpha$ bond of the peptide backbone is denoted by ϕ, rotation about the C$^\alpha$–C' bond by ψ, and that about the peptide bond C'–N by ω. Specification of ϕ, ψ, and ω, for all residues of the polypeptide chain completely determines the backbone conformation. Different rotations about single bonds are intrinsically equivalent, and the relative preference for each particular torsion angle is determined by the energetics of the *noncovalent* interactions among the atoms, and of the atoms with the solvent molecules.

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.



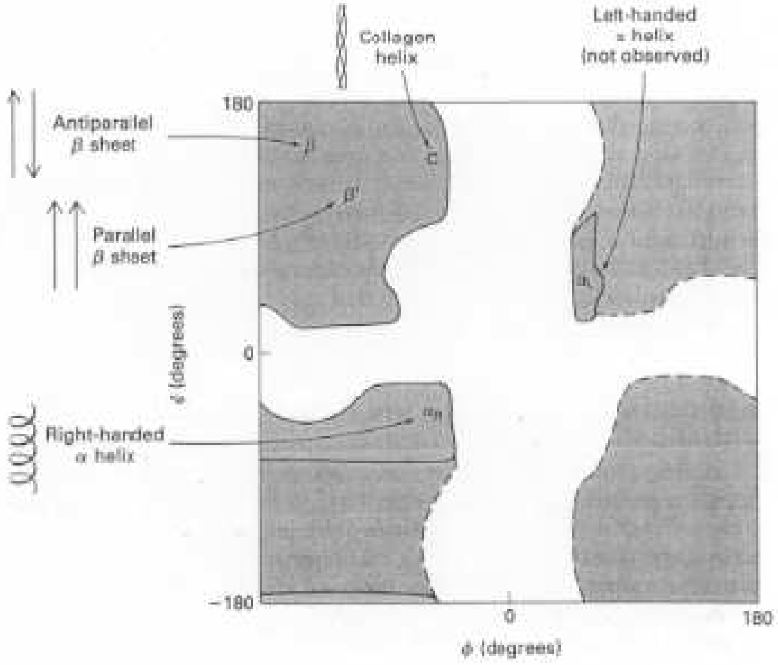


Figure 1.1: Ramachandran plot showing allowed values of the torsion angles ϕ and ψ for alanyne residues (region contoured by solid lines). Additional conformations are accessible to glycine (contoured by dashed lines) because it has a very small side chain. The typical values of the torsion angles corresponding to the different secondary structures are shown.

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 atoms of adjacent residues is fixed to the value of 3.80 Å. When the residue following the peptide bond is prolyne, 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. A part from prolyne exception, ω is fixed, whereas ϕ and ψ, which may in principle assume all possible values, are constrained geometrically because of steric clashes between atoms far apart along the chain.

In a simplified description, it seems natural to consider all the hard degrees of freedom, corresponding to covalent interactions, as frozen, and take into account only the soft degrees of freedom, that is ϕ and ψ, corresponding to non-covalent interactions.

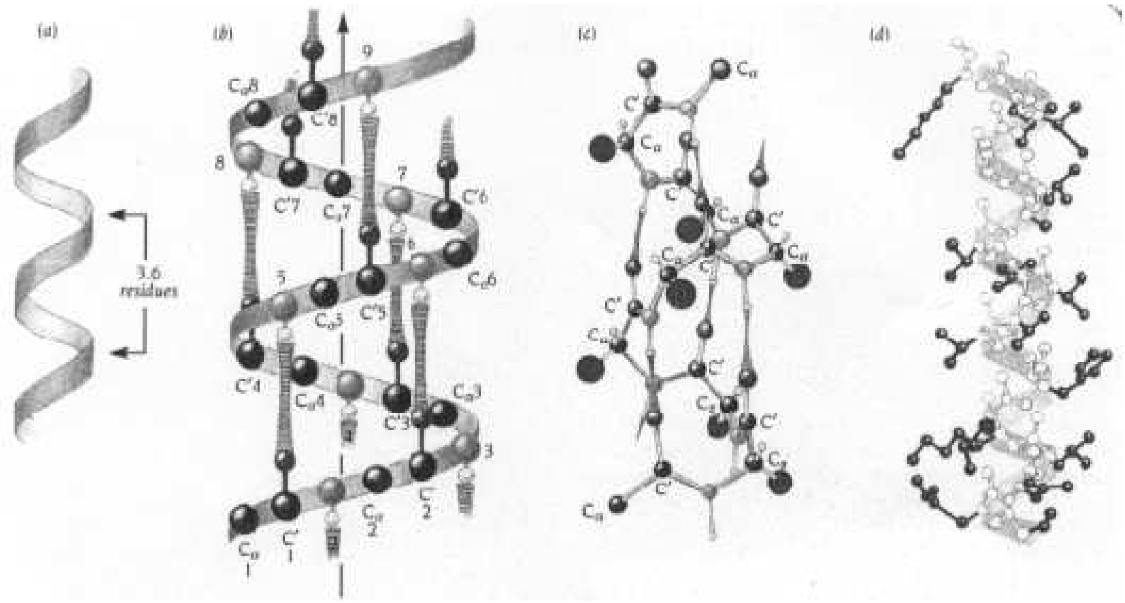


Figure 1.2: (a) Idealized diagram of the backbone path in an α-helix. (b) The same as (a) but with approximate positions for the backbone atoms and hydrogen bonds included. (c) Schematic diagram with the correct position of all backbone atoms. Big dark circles represent side chains. (d) A ball-and-stick model of one α-helix in myoglobin.

The allowed values of ϕ and ψ are usually shown on a two-dimensional map in the ϕ-ψ plane, the so-called *Ramachandran plot*, as in figure 1.1.

The secondary motifs, helices and sheets, which repeatedly occur in the structures of natural proteins, leave their own specific signature in the Ramachandran plot. Each motif is characterized by ordered local geometry which results in particular values of the torsion angles. In figure 1.1, are indicated the different regions of the Ramachandran plot corresponding to the three types of secondary structures that we are going to describe in the remainder of this section; right-handed α-helices, parallel and antiparallel β-sheets, and the collagen triple helix. Left-handed α-helices are not observed, since the side chains would be too close to the backbone.

The right-handed α*-helix* is the best known and most easily recognized of the secondary structures (see figure 1.2). The backbone carbonyl oxigen of each residue forms a nearly straight hydrogen-bond with the backbone amino group of the fourth residue 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.

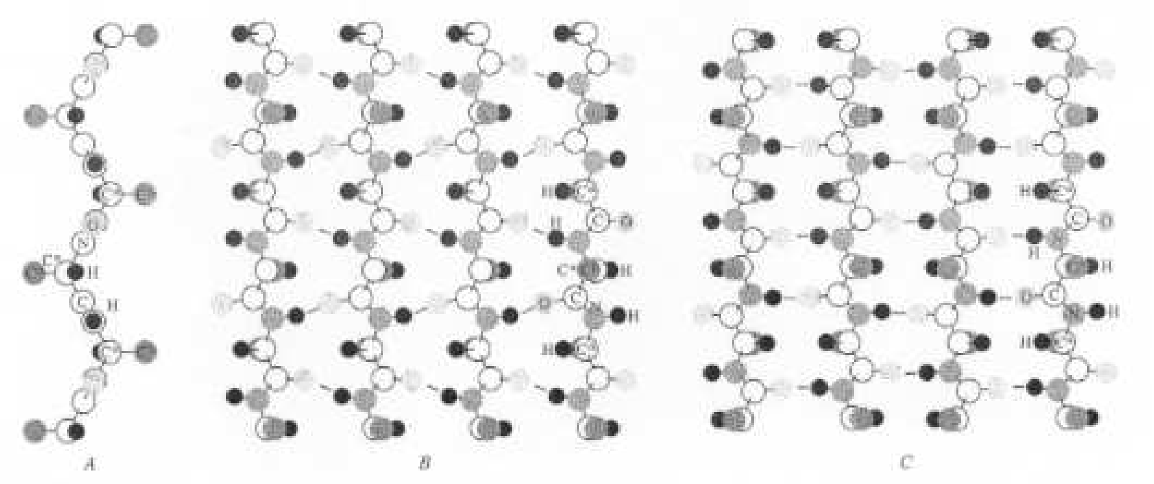


Figure 1.3: (A) A single β-strand. (B) A planar parallel β-sheet. (C) A planar antiparallel β-sheet. Note the different geometric patterns of hydrogen bonds in (B) and (C). The horizontal direction in (A) is othogonal to the plane represented in (B) and (C).

The second most regular and identifiable secondary structure is the β-*sheet* (see figure 1.3). The basic unit is the β-*strand*, a planar zig-zag conformation with the side chains alternatively projected in opposite directions. It may be considered a special type of helix with 2.0 residues per turn. A single β-strand is not stable, because no interactions are present among the atoms. The β-strand conformation is stabilized only when two or more strand are assembled into a β-sheet, a planar structure where hydrogen bonds are formed between the peptide groups on adjacent β-strands. Side chains from adjacent residues of the same strand protrude from opposite sides of the sheet and do not interact with each other. Side chains from neighbouring residues of adjacent strands are projected instead into the same side, and thus interact significantly. Adjacent β-strand can be either *parallel* or *antiparallel*, and the resulting geometry varies slightly. In antiparallel sheets, all hydrogen bonds are parallel to each other, whereas in parallel sheets they are arranged in two different alternating directions.

Most *fibrous* proteins play structural roles and have regular, extended structures that represent a level of complexity somewhat intermediate between pure secondary structure and the tertiary structures of globular proteins. Due to their general insolubility they are more diffucult to characterize experimentally. One of the best understood related structures is *collagen triple helix*. (see figure 1.4). Collagen is the main

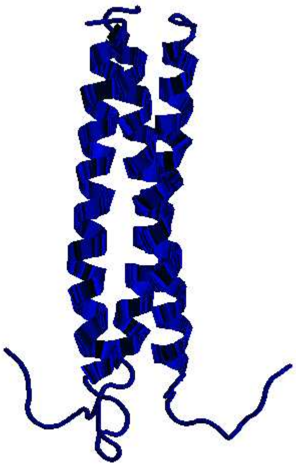


Figure 1.4: Cartoon representation of the collagen triple helix, from the structure *1aq5* taken from the Protein Data Bank. In chapter 4, we will measure the degree of optimal packing of the same protein.

constituent of higher animal frameworks, such as bones, tendons, skins, cartilage, and membrane supporting tissues. There exist only a few distinct types of collagen polypeptide chains. They are all charachterized by the repetition of glycin every three residues of their aminoacid sequence. Three different chains are coiled together, each with a slightly twisted, left-handed helical conformation, whose typical torsion angles can be inferred from figure 1.1. The three helices are wound around each other to form a right-handed super-helix, and are stabilized 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 *different* from the α-helix. It has opposite handedness, different torsion angles, and 3.0 residues per turn.

1.1.4 Membrane proteins

A plasma membrane encloses every cell, mantaining the essential differences between its contents and the enviroment. All biological membranes have a common general structure. They are assemblies of lipid and protein molecules held together mainly by non-covalent interactions.

Lipids are *amphiphilic* (or amphiphatic) molecules, that is they have a small hydrophilic polar head group, and two long hydrophobic nonpolar tails. As a result of their amphiphilic nature, lipid molecules spontaneously form bilayers in aqueous solution, so that the hydrophobic tails are buried in the interior and the polar heads are exposed to water.

The lipid molecules in cell membranes are arranged as a continuous double layer about 50 Å thick. The *lipid bilayer* provides the basic structure of the membrane, and serves as a relatively impermeable barrier to the passage of the most water-soluble molecules. The protein molecules, usually 'dissolved' in the lipid bilayer, mediate most of the other functions of the membrane, serving as specific receptors, enzymes, transport proteins, and so on.

Most membrane proteins extend across the lipid bilayer, and pass through the whole extension of the membrane one or more times (*integral* membrane proteins). Other proteins associated with membranes do not extend into the nonpolar interior of the lipid bilayer at all, but are bound to one or other face of the membrane by means of covalent or non-covalent links (*non-integral* membrane proteins). In the following we will refer only to integral membrane proteins.

Membrane proteins are amphiphilic. Their sequence is characterized by groups of hydrophobic residues that pass through the membrane and interact favourably with the nonpolar interior of the lipid bilayer, and by groups of hydrophilic residues that are exposed to water on both sides of the membranes. Consequently membrane proteins are soluble neither in aqueous solution nor in nonpolar solvents.

Membrane proteins are very hard to crystallize, and consequently very few structures have been resolved with high resolution. The available structures are far less diverse than those of globular proteins. This is a consequence of the requirement that membrane proteins must either bind to one leaflet of, or completely span, the lipid bilayer, while at the same time maintaining contact with the surrounding aqueous medium. Peripheral binding to one side of the bilayer can be mediated by amphiphilic α-helices oriented parallel to the bilayer surface, and membrane-spanning structures can be built either from *β-barrels* or from *bundles* of trasmembrane α-helices. Large extra-membraneous domains do not seem to be much influenced by their attachment to the bilayer, and can essentially be viewed as tethered globular proteins [86].

In the simplified model studied in chapter 6, we will investigate how the behaviour of a random sequence of hydrophilic and hydrophobic residues varies, in the presence of both a polar and a nonpolar medium, when changing the average hydrophobicity of the chain.

The structural differences between membrane and globular proteins, however, do not only result from their different enviroments, but also reflect fundamental differences in the mechanisms responsible for their folding and assembly. Membrane proteins are indeed first synthesized and exposed to aqueous surrounding, and only later inserted in the lipid bilayer [87].

1.2 Random heteropolymers

The physical properties of typical random sequences may be described within *random heteropolymer* models [7]. They were introduced, more than ten years ago, as a phenomenological approach to protein folding, based on the analogy of a protein with a spin glass system, the model of reference in the field of the statistical mechanics of disordered systems [63]. Nevertheless, it was soon realized that random sequences are not sufficiently proteinlike; rather, they have dramatically different physical properties.

In a seminal work [8], Bringelson and Wolynes postulated the applicability of the Random Energy Model (REM), a simplified spin-glass model introduced by Derrida [88], to the case of random heteropolymers. This amounts to assume that the energy values of different conformations can be considered as independent equally distributed random variables over the possible realizations of disorder, that is over the possible aminoacid sequences.

The simple assumption of the statistical independence of energies in different states implies remarkable consequences. The energy spectrum consists of an upper quasicontinuous part, which is independent of disorder, and a few discrete low energy levels that are placed very individually for each realization of disorder. The ground state for typical realizations is of order $\sqrt{N}$ below the edge of the continuum spectrum, where N is the number of residues. All discrete levels are of order $\sqrt{N}$ from each other for typical realizations. A *freezing* temperature exists, above which the system explores the high-entropy continuous part of the energy spectrum. Below the freezing temperature, the heteropolymer is confined with vanishing entropy into discrete individual states.

The spin-glass freezing transition to a phase dominated by very few conformations was then found to describe successfully heteropolymer behaviour in more microscopic models [65, 64], even by means of the same analytical tools used in spin-glass mean field theory [66], that is computation of averages over the quenched disorder with the replica trick, and the emergence of replica-symmetry breaking as a signature of the onset of the freezing transition, as we will see in subsection 1.2.1. A typical random sequence indeed freezes in a ground state conformation with vanishing entropy, but the ground state is also random and thus not specific. Moreover, many low-energy states exist closely competing with the ground state and typically unrelated to it. The folding kinetics is thus slow and unreliable, getting easily trapped in such states. These features are typical of disordered systems and are a consequence of its *frustration*, that is of the fact that the degrees of freedom cannot be optimized simultaneously [63]. In the case of proteins and heteropolymers, frustration is due to chain connectivity constraining two consecutive monomers with opposite affinities to remain in the same environment.

What makes protein sequences different from random sequences? Real proteins are believed to be the result of a biased evolutionary search for those particular sequences which are able to fold rapidly and reversibly into the sequence-specific native state, a condition needed in order to accomplish their biological functions. In other words, special sequences have been selected which overcome the hindrance of frustration in their search for the native state, thus leading to emergence of an optimization principle, ‘minimal frustration’, underlying sequence selection [8]. Minimal frustration causes the energy landscape to attain a *funnel* shape, which implies a high energy gap between the native ground state and the first excited states structurally unrelated to it. The native state is thus thermodynamically stable and the aminoacid sequence folds into it in a fast and reproducible way [11].

1.2.1 Analytic approach and replica-trick

We now discuss some of the key ingredients that are commonly used in analytical studies of heteropolymer models in the continuum space $\mathbb{R}^d$. In order to keep our discussion the more general as possible, we do not refer to any specific heteropolymer model.

The polymeric component, the *chain constraint*, is usually implemented in the model by considering a harmonic potential between successive monomers along the chain:

$$\beta \mathcal{H}_{ch} = \frac{d}{2l^2} \sum_{i=1}^{N} \left(\mathbf{R}_{i+1} - \mathbf{R}_i\right)^2 \ . \tag{1.2}$$

Monomers along the chain are labelled by $0 \leq i \leq N$ and $\mathbf{R}_i$ is the position of monomer i in the d-dimensional space, $\beta = \frac{1}{T}$ is the inverse temperature (if the Boltzmann constant is set equal to unity), l^2 is the mean square length of the polymer bond length. Note that the chain constraint is considered as an *entropic* effect. Then, one in general may want to consider some *homopolymeric* interaction $\mathcal{H}_{hom}\left(\{\mathbf{R}_i\}\right)$ where usually repulsive interactions accounting for the excluded volume effect are considered.

The partition function $Z\left(\{\zeta_i\}\right)$ of a generic heteropolymer model, for a given disorder configuration $\{\zeta_i\}$, is:

$$Z\left(\{\zeta_i\}\right) = \int \prod_{i=1}^{N} \mathrm{d}^d R_i \exp\left[-\beta\mathcal{H}_{ch} - \beta\mathcal{H}_{hom} - \beta\mathcal{H}_{dis}\left(\{\zeta_i\}\right)\right] \ . \tag{1.3}$$

We have denoted by $\{\zeta_i\}$ the set of random variables entering the interaction term $\mathcal{H}_{dis}\left(\{\zeta_i\}\right)$. For example, $\{\zeta_i\}$ could be the set of hydrophobic charges ζ_i characterizing the hydrophobicities of each monomer (negative values correspond to hydrophobic residues and positive values to hydrophilic ones) [46, 89]. The random variables $\{\zeta_i\}$ are thought to be picked up according to a probability distribution $\mathcal{P}\left[\{\zeta_i\}\right]$. If the random variables are independently equally distributed, $\mathcal{P}\left[\{\zeta_i\}\right] = \prod_i P\left(\zeta_i\right)$.

Since the sequence of monomers is fixed by highly stable covalent bonds, one is usually interested in the case of *quenched* frozen random variables. The quenched free energy $\beta F_q = -\int \prod_i \mathrm{d}\zeta_i \mathcal{P}\left[\{\zeta_i\}\right] \ln\left[Z\left(\{\zeta_i\}\right)\right]$ is usually computed by means of the *replica trick* [63]:

$$\beta F_q = -\lim_{n\to 0} \frac{\overline{Z^n} - 1}{n} \ . \tag{1.4}$$

The bar denotes the average over random variables, and $\overline{Z^n} = \int \prod_i \mathrm{d}\zeta_i \mathcal{P}\left[\{\zeta_i\}\right] Z^n\left(\{\zeta_i\}\right)$ is the average over disorder of the partition function of n replicas of the system, which are labelled by a, all having the *same* disorder configuration $\{\zeta_i\}$:

$$\overline{Z^n} = \int \prod_{i=1}^{N} \prod_{a=1}^{n} \mathrm{d}^d R_i^a \exp\Bigg[- \sum_a \beta\mathcal{H}_{ch}\left(\{\mathbf{R}_i\}^a\right) - \sum_a \beta\mathcal{H}_{hom}\left(\{\mathbf{R}_i\}^a\right) - \\ - \beta\mathcal{H}^*_{dis}\left(\{\mathbf{R}_i^a\}\right)\Bigg] . \tag{1.5}$$

The average over randomness yields an effective interaction energy $\mathcal{H}^*_{dis}$ which introduces coupling between different replicas. Note that $\mathcal{H}^*_{dis}$ has also an implicit dependence on temperature. Then, $\overline{Z^n}$ is usually estimated within a variational approach in the spirit of mean-field approximation.

By using different variational ansatz for the structure of the solution in replica space, one can check whether replica-symmetry breaking is occurring or not [90, 91]. In close analogy with spin glass mean field theory, replica-symmetry breaking signals the onset of the freezing transition. Freezing is associated with the *ergodic decomposition* of the conformation space in different energy valleys separated by high barriers. When the chain ‘freezes’ in one of these valleys it is locked within it and looses much of its conformational entropy. In which valley, that is in which conformation, the chain freezes, is chosen basically in a random way, since all competing conformations have almost the same energy.

1.3 Packing and geometrical properties in protein structures

Virtually all known protein structures exhibit locally ordered patterns characterized by a high degree of geometrical regularity [22], the secondary structures that we have described in subsection 1.1.3.

But striking regularities also occur in the geometry of protein global structures [21]. Two proteins are said to have a common fold if they have the same major secondary structures in the same arrangement with the same topological connections [18]. Common folds occur even for proteins with different biological functions [19]. The total number of known folds is only of the order of thousands [20], an extremely small number compared with both the number of known protein sequences and the number of all possible conformations, which is increasing exponentially with the number of aminoacids.

Also the number of known sequences is a negligible fraction of all possible sequences which may be created by choosing randomly aminoacids. Nevertheless, the reduction factor operated by evolutionary selection is much higher for structures. Moreover, evolution acts directly on structures, since they control the specific biochemical functions of proteins. It thus would seem that structure selection could have played a role more crucial than sequence selection.

Several approaches have been proposed in order to investigate this point. On one hand, it has been suggested that the requirement of minimum energy alone would explain the presence of both highly symmetrical proteinlike structures and few distinctly shaped folds, since a structure with the lowest energy for a given sequence is likely to stabilize more sequences [92, 93, 94]. Moreover, the elegant deduction by Pauling *et al.* [24] of the correct geometry (3.6 residues per turn) of the α-helix structure by simply minimizing the energy of intra-chain hydrogen bonds would imply that energetic considerations are indispensable in explaining the emergence of secondary structures. It is however under current debate, whether hydrogen bonds between different groups of the protein chain are stronger or weaker than hydrogen bonds between chain groups and water molecules [95, 96].

Successive studies have suggested that purely energetic considerations are not the whole story, and proposed for native conformations the selection mechanism of *high designability* [84, 97, 98, 99]. In lattice models computation, a few number of higly designable structures emerge with a number of sequences successfully folding into them much larger than the average. These special structures are more thermodynamically stable than other structures, thus yielding more efficient folding. They also exhibit the peculiar geometric feature of being far away from neighboring competing structures in the multi-dimensional structure space.

On the other hand, a purely geometrical approach had been already put forward in order to explain structure selection, without invoking detailed chemistry and energetics [82, 100]. On the basis of exhaustive enumerations in lattice models, it had been firstly hypothesized that compactness by itself is sufficient to drive secondary structure formation [101, 102]. Nevertheless, this revealed itself as an artifact of lattice structural order [103, 95, 104], since a generic compact polypetide chain in the continuum space was shown to account for only a small secondary structure content

[105].

Very recent works have raised new interest in the geometrical characterization of protein structures. It has been observed, that natural folds of proteins have a much larger density of nearby structures than randomly generated conformation with the same compactness, after a coarse-graining and a discretization of the conformational degrees of freedom [29]. This results in an exceedingly large geometrical accessibility, which provides a large basin of attraction for the native state, thus favouring funnel formation in the energy landscape when sequences are folded into that structure. Moreover, simulations in a reduced representation space have shown that, optimizing the density of alternative conformations leads to the emergence of regular patterns in the reduced space, which are reminiscent of secondary structures [29].

It seems intuitive to assume that such geometrical properties be connected to the folding rate; the more accessible the native structure, the faster the folding. A dynamical variational principle has been proposed [30], according to which, among all possible native conformations, a protein backbone will attain only those that are optimal, under the action of evolutionary pressure favouring fast folding. Simulations on a coarse-grained discretized model show that fast folding requirement leads to the emergence of helical order in compact structures [30].

Apart from direct investigations of the principles underlying structure selection, many different recent studies show that folding rates and mechanisms observed experimentally are largely determined by the topology of the native state [25]. This provides a further clue that native structure selection played a crucial role during evolution.

An important feature common to many of the coarse-grained approaches discussed above, that we also use in our approach, is to model the polypeptide chain by retaining only the α-carbon atoms of the backbone chain [106]. Clearly, the complete neglecting of the side chains may affect drastically the polypepdtide chain behaviour. Nevertheless, secondary structures are stabilized by interactions between different backbone peptide groups. Moreover, it is conceivable that the maximum geometrical accessibility of the native *backbone* conformation is just the result of seeking the conformational freedom necessary to accomodate the side chains. In this respect, it is interesting to note that, whereas helix-promoting energetic factors involve the polypeptide backbone, and are therefore common to all residues (except glycine and prolyne), side chain steric factors seem to predispose segments of the chain towards either α-helix or β-sheet formation [107]. Some side chains lose sufficient conformational entropy when the backbone is helical, that they push the corresponding residues towards a β-sheet conformation. Note that the conformational entropy is nothing else than the geometrical accessibility.

Motivated by this recent developments, we thus search for a precise mathematical formulation in the continuum three-dimensional space of the intuitive notion of the ‘number of nearby structures’ to a given backbone α-carbon atoms conformation. The problem is now purely geometrical in nature, so that in the following we will speak of the conformations of a chain of beads. We would then search for the chain conformation maximizing the number of nearby structures, as natural folds of proteins seem to do. The number of nearby structures is in itself a rather vague concept in the continuum; how is it possible to measure the density of states?

Following again intuition, one may think that the number of nearby structures to a given one is simply the *free volume* available around that conformation, such that this one can be moved within it, before its topology is drastically changed.

A related notion, *thickness*, has been already introduced in the context of knot theory, in order to characterize the ‘ideal shape’ of a given knot topology [108, 32, 33]. The thickness associated with a given chain conformation is the maximum allowed radius of a tube of uniform radius, which is inflated around the chain. The tube cannot grow anymore, when 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 [32]. In our language, the available free volume around a chain conformation simply becomes the volume of the tube inflated around the chain, when the tube radius is the thickness.

We will then borrow the notion of thickness and look for the conformations of polymer chains in the continuum space maximizing the thickness, within some defined compactness constraint. As we will see in the next chapter, this can be considered as a *optimal packing* problem for strings, generalizing the classic dense-packing problem for hard spheres, first posed by Kepler [31].

Chapter 2

The packing problem for strings

The characterization of particular geometrical structures and shapes satisfying variational principles has been long studied since the very beginning of mathematics history. By variational principle we mean the requirement that some functional of the spatial arrangements of the geometric elements under consideration, a set of discrete points, one or more continuous curves, be minimized, or equivalently maximized.

The variational principle whose enforcing is the subject of this chapter and of the following ones, involves the determination of structures in the ordinary euclidean three-dimensional space. People in everyday life face the problem of packing objects, where the relevant issue is to optimize the way available volume is employed, finding for example the maximum density arrangement which is attainable for the objects under consideration. Our aim is to investigate the possibility that some of the structures more frequently present in biological macromolecules have been selected, just because they optimize volume occupation by having the largest possible conformational entropy, as explained in chapter 1.

In this chapter we formalize the generalization of the close-packing problem to the case of a chain of beads connected with each other as pearls on a necklace. From a mathematical point of view it is convenient to consider the limit in which the discrete chain of beads becomes a continuous string.

In section 2.1, we briefly discuss the standard hard-sphere packing problem, which was first formulated by Kepler [109]. In section 2.2, we outline and discuss the main issues that one has to face when dealing with the packing problem for a thick impenetrable tube. Fortunately, useful tools have already been developed in the context of knot theory [108, 32, 110]. In section 2.3, we introduce and define the fundamental concept of *thickness* and of the *global radius of curvature* function [33], in the case of continuous curves. In section 2.4, we extend both definitions to the case of discrete curves. Note that, besides being the obvious reference for numerical simulations, this last case is also the conceptually relevant one when dealing with biological macromolecules.

2.1 Optimal packing of hard spheres

The problem of the close packing of hard impenetrable spheres was posed almost four centuries ago by Kepler. It is in fact the simplest non trivial way[1] to address the packing issue, since one is dealing with identical objects having the higher symmetry degree. Yet, such a problem is of obvious practical importance - just think about oranges in a grocery or cannonballs in a weapon storage - and has attracted much interest in the scientific community [111], culminating in its recent rigorous mathematical solution, just two years ago [112, 113, 114]. Of course, common people well knew the solution by using it everyday!

The three-dimensional arrangement of spheres solving the close-packing problem can be more easily understood starting from the solution of the analogous two-dimensional problem [31]. As any billiard player well knows, the centres of a set of impenetrable hards disks in their best packing configuration form a triangular lattice. The general solution of the three-dimensional hard spheres packing problem is obtained by superposing one on top of the other two-dimensional layers in which the centres of the spheres are occupying the sites of a triangular lattice. There is an infinite number of degenerate best packing configurations, since when placing one layer on top of the other only half of the ‘holes’ of the bottom layer are filled by spheres of the top layer, thus leaving the possibility of choosing how the holes may be filled each time a new layer is added. One particular sequence of choices result in the face-centred cubic lattice, which emerges as the best solution, for entropic reasons, when temperature comes into play [115]. The free energy of the *fcc* structure has been indeed found to be slightly less than the free energy of the other regular best packing arrangement, the hexagonal close-packed lattice.

[1] The close packing of identical cubes is trivial!

The study of Kepler's problem has had important applications in such fundamental physical phenomena as crystallization and melting of condensed matter [116]. In fact, in many atomic and molecular systems pairwise interactions are characterized by a very strong repulsion at short distances and a weaker attraction at longer distances. The former characteristics gives rise to the so-called excluded volume effect, and is well approximated by considering atoms (or molecules) as impenetrable hard spheres. The attractive tail at long distances originates from Van der Waals forces and is in some cases much weaker than the hard core repulsion, e.g. for rare gases. It is thus not a surprise that many of them crystallize in *fcc* structure, as predicted by Kepler's solution of the close-packed hard sphere problem.

2.2 Optimal packing of a thick tube

As we have discussed in chapter 1, we would like to investigate the possibility that some of the structures or of the structural motifs more commonly occurring in natural proteins may have been selected by natural evolution due to their *geometrical* properties. As suggested by recent studies [29] on the density of states of real protein structures, when compared with artificial compact decoy structures, our aim is to formalize the intuitive notion of free volume of a chain conformation and to identify which is the structure having the maximum available free volume.

A natural framework is provided by the generalization of the close-packed hard sphere problem discussed in the previous section to the case in which pairs of spheres are consecutively linked together as beads on a string. Which are the close-packed configurations of such a structure?

The close-packed hard sphere problem may be restated in an alternative manner, more convenient for numerical implementation, 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, $d = \min_{i<j} |\vec{r}_i - \vec{r}_j|$, being as large as possible [117]. More precisely, the optimal packing configuration of N hard impenetrable spheres of radius R, which minimizes the side L of the smallest cube containing the N spheres, is the same that is obtained when d is maximized for a set of N points contained in a cube of side 1, when neglecting boundary effects. The attained maximum value is just $d_{max} = \frac{R}{L}$. It is notable that the resulting arrangement is strongly influenced by boundary effects [117]. Nevertheless 'bulk' optimal arrangements, such as the *fcc* lattice, exhibit trans-

lational invariance in that, far from the boundaries, the local environment is the same for all points.

It would be highly convenient to restate in an analogous way the best packing problem for chains, but the very presence of the chain raises two different issues. Firstly, one has to define the notion of a ‘minimum contact length’ related to a chain configuration, in a different way than simply the minimum distance between any pair of points. The packing problem then amounts to search for the chain structure maximizing this minimum distance when placed in a finite volume or in the presence of some other compactness constraint. Secondly, the presence of the chain provides a preferential direction, thus breaking the translation invariance present in the simpler hard spheres packing problem. One has therefore to handle with great care all issues regarding the presence of boundary effects and the emergence of ‘bulk’ behaviour, especially when dealing with the choice of suitable compactness constraints.

The first issue has had already been approached in the context of knot theory [108]. The *ideal* form of a knot with a given topology and assembled from a tube of uniform width has been defined by Katritch *et al*. [32, 110] as the tube configuration having the highest ratio of volume to surface area, or in other words the thickest tube of fixed length that can be tied into a given knot topology. It is notable that the above definition closely matches the intuitive notion of the optimization of volume occupation which we are interested in, whereas the tube thickness is just the ‘minimum contact length’ that we are looking for. As we will see, *triplet* of beads have to be considered in order to compute the thickness of a string [33]. Interestingly, ideal knots turns out to be related to the time-averaged shapes of knotted DNA molecules in solution [118, 119, 120].

We generalize this approach to the generic case of an open curve, thus removing both the constraints of curve closing and of fixed knotted topology. We will search for the optimal shapes of a tube of fixed length N maximizing its thickness. Note that curve closing is an implicit compactness constraint, so that explicit constraints are necessary when dealing with open curves, such as for examples placing the tube in a finite volume. In the following sections we will formalize the problem which we are then going to tackle numerically, beginning from the case of a continuous curve. We will proceed much along the general guidelines provided by Gonzalez and Maddocks [33], who also studied knotted curves, and introduced the fundamental notion of *global* radius of curvature, as an extension of the local radius of curvature concept. They also showed how to generalize the notion of tube thickness to the case of a *discrete* curve, providing thus a natural tool for our numerical simulations.

2.3 Global radius of curvature and thickness

Any smooth, non-self-intersecting curve can be thickened into a smooth, non-self-intersecting tube of constant radius centred on the curve. If the curve is a straight line, there is no upper bound on the tube radius, but for non-straight curves, there is a critical radius above which the tube either ceases to be smooth or exhibits self-contacts. This critical radius is an intrinsic property of the curve called its *thickness* or *normal injectivity radius*.

In mathematical language, a *curve* $\mathcal{C}$ is a continuous three-dimensional vector function $\vec{g}\left(s\right)$ of a real variable s with $0 \leq s \leq N$. We will assume the curve $\mathcal{C}$ to be *smooth*, that is the function $\vec{g}\left(s\right)$ is continuously differentiable to any order and the tangent vector $\dot{\vec{g}}\left(s\right)$ is nonzero for all s, *open*, i.e. in general $\vec{g}\left(N\right) \neq \vec{g}\left(0\right)$, and *simple*, that is with no self-intersections, $\vec{g}\left(s_1\right) = \vec{g}\left(s_2\right)$ only when $s_1 = s_2$.

A smooth solid tube of constant radius η centred on $\mathcal{C}$ may be defined, for sufficiently small η, as the union, over all points $\vec{x}$ on $\mathcal{C}$, of all the circular disks of radius η centred at $\vec{x}$ and contained in the normal plane to $\mathcal{C}$ at $\vec{x}$. The thickness of the curve is the critical radius $\eta_* \left[\mathcal{C}\right]$ above which, for non-straight curves, the tube either ceases to be smooth or exhibits self-contacts. It is simple to show [121] that the tube becomes *locally* singular when its radius η becomes equal to the local radius of curvature $\rho_L\left(\vec{x}\right)$ of $\mathcal{C}$ at some point $\vec{x}$. [2] On the other hand, the occurrence of self-contacts between different portions of the tube is a *non-local* effect, which can be shown [121] to be related to the presence of two non-adjacent points of *closest approach* on $\mathcal{C}$. For such a pair of distinct points $\vec{x}$, $\vec{y}$, the vector $\vec{x} - \vec{y}$ is orthogonal to the tangent vectors to $\mathcal{C}$ at both $\vec{x}$ and $\vec{y}$. The tube thickness may thus be written as

$$\eta_* \left[\mathcal{C}\right] = \min \left\{ \min_{\vec{x} \in \mathcal{C}} \rho_L\left(\vec{x}\right), \min_{\vec{x}, \vec{y} \in \Omega} \left|\vec{x} - \vec{y}\right| / 2 \right\} , \tag{2.1}$$

where Ω is the set of all pairs of points of closest approach on $\mathcal{C}$, defined as above, and $\left|\vec{x} - \vec{y}\right|$ is the Euclidean distance between the points $\vec{x}$ and $\vec{y}$. In words, the thickness

[2] The local radius of curvature is simply the radius of the circle which locally best approximates the curve. We will give an explicite definition of the local radius of curvature function just below.

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* [33] which we are going to introduce allows to capture simultaneously both possibilities.

The definition of global radius of curvature is based on the elementary fact that any three non-collinear points $\vec{x}$, $\vec{y}$, and $\vec{z}$ in three-dimensional space define a unique circle (the *circumcircle*). The radius of this circle (the *circumradius*) can be written as

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

where $\mathcal{A}\left(\vec{x},\vec{y},\vec{z}\right)$ is the area of the triangle with vertices $\vec{x}$, $\vec{y}$, $\vec{z}$. When the points $\vec{x}$, $\vec{y}$, and $\vec{z}$ are distinct but collinear the circumcircle degenerates into a straight line, and we assigne a value of infinity to the corresponding circumradius $r\left(\vec{x},\vec{y},\vec{z}\right)$.

When $\vec{x}=\vec{g}\left(s\right)$, $\vec{y}=\vec{g}\left(\sigma\right)$, and $\vec{z}=\vec{g}\left(\tau\right)$ are points on a simple, smooth curve $\mathcal{C}$, in the double limit $\sigma,\tau\to s$ we recover the usual definition of the standard *local radius of curvature* $\rho_L\left(\vec{x}\right)$ of the curve $\mathcal{C}$ at the point $\vec{x}=\vec{g}\left(s\right)$:

$$\lim_{\sigma,\tau\to s} r\left(\vec{x},\vec{y},\vec{z}\right)\equiv r\left(\vec{x},\vec{x},\vec{x}\right)=\rho_L\left(\vec{x}\right)\ . \tag{2.3}$$

Following [33], we introduce the notion of the *global radius of curvature* $\rho_G\left(\vec{x}\right)$ at each point $\vec{x}$ of $\mathcal{C}$:

$$\rho_G\left(\vec{x}\right)\equiv\min_{\vec{y},\vec{z}\in\mathcal{C}} r\left(\vec{x},\vec{y},\vec{z}\right)\ . \tag{2.4}$$

The function ρ_G can be interpreted as a generalization, indeed a globalization, of the standard local radius of curvature function. It immediately follows from definition that global radius is bounded by local radius:

$$\rho_G\left(\vec{x}\right)\leq\rho_L\left(\vec{x}\right)\quad\forall\ \ \vec{x}\in\mathcal{C}\ . \tag{2.5}$$

The global radius is also infinite for all points when $\mathcal{C}$ is a straight line, just as with local radius.

The optimality condition associated with the minimization in eq. (2.4) implies [33] that the global radius $\rho_G\left(\vec{x}\right)$ may be either the local radius of curvature $\rho_L\left(\vec{x}\right)$, or the strictly smaller radius of a circle containing $\vec{x}$ and another distinct point $\vec{y}$ at which the circle is tangent. Thus, to determine $\rho_G\left(\vec{x}\right)$, one need consider only the minimization in eq. (2.5) with the restriction $\vec{y}=\vec{z}$.

As we will see below, the thickness $\eta_* [\mathcal{C}]$ of any smooth and simple curve $\mathcal{C}$ is simply the minimum global radius of curvature over all points $\vec{x}$ of the curve:

$$\eta_* [\mathcal{C}] = \Delta [\mathcal{C}] \equiv \min_{\vec{x} \in \mathcal{C}} \rho_G (\vec{x}) \ , \tag{2.6}$$

or, in other words, the minimum value of the circumradius function $r (\vec{x}, \vec{y}, \vec{z})$ over all triplets of points. According to this last definition, it is possible to demonstrate [33] that the minimum global radius $\Delta [\mathcal{C}]$ is either the minimum local radius of curvature or the strictly smaller radius of a sphere that contains no portion of the curve in its interior and is tangent to the curve at two diametrically opposite points $\vec{x}$ and $\vec{y}$. At such points the symmetry property $r (\vec{x}, \vec{x}, \vec{y}) = r (\vec{x}, \vec{y}, \vec{y})$ holds, and the vector $\vec{x} - \vec{y}$ is orthogonal to the tangent vectors to $\mathcal{C}$ at both $\vec{x}$ and $\vec{y}$. This is just the condition required when searching for couple of points of closest approach of $\mathcal{C}$, and shows that the minimum global radius of curvature is indeed the thickness.

The reformulation of the notion of thickness (compare equations (2.1) and (2.6)) has both conceptual and practical importance. The definition of global radius of curvature allows to characterize thickness in eq. (2.6) in a 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.

We are interested in characterizing the properties of the optimal curve having maximum thickness within some class of curves satisfying suitable compactness conditions. More precisely, let $\mathcal{R}$ denote the set of all simple, smooth curves $\mathcal{C}$ with fixed length $N > 0$ satisfying some definite compactness constraint (in the following we will consider for example curves with a maximum allowed gyration radius G). Our focus in the following chapters will be in determining the ideal optimal shapes $\mathcal{C}^*$ in $\mathcal{R}$ satisfying:

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

This definition yields precisely the optimal curves around which the thickest tube of fixed length may be inflated, compatibly with the specified compactness constraint, and corresponds to the intuitive notion of the optimization of volume occupation, as discussed in the previous section. 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 $\mathcal{C}^*$ 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 [33]. 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 [33] cannot be easily generalized to the case of a generic compactness constraint. Nevertheless, we have found this property to be verified in *all* our simulations.

2.4 Discrete curves

A discrete curve $\mathcal{C}_n$ is an ordered set of distinct points $\{\vec{x}_0, \ldots, \vec{x}_n\}$ in three-dimensional space. To any discrete curve $\mathcal{C}_n$ one can associate a continuous, piecewise linear curve $\mathcal{C}$ by connecting $\vec{x}_0$ to $\vec{x}_1$ with a straight line and so on. Extensions of definitions introduced in the previous section are straightforward [33]. The global radius of curvature ρ_G at each point $\vec{x}_i$ of $\mathcal{C}_n$ is

$$\rho_G\left(\vec{x}_i\right) \equiv \min_{\substack{0\leq j<k\leq n \\ j\neq i\neq k}} r\left(\vec{x}_i, \vec{x}_j, \vec{x}_k\right) \ . \tag{2.8}$$

That is, $\rho_G\left(\vec{x}_i\right)$ is the radius of the smallest circle containing $\vec{x}_i$ and two other distinct points $\vec{x}_j$ and $\vec{x}_k$. The local radius of curvature of a discrete curve is simply the radius of the circle containing three consecutive points:

$$\rho_L\left(\vec{x}_i\right) \equiv r\left(\vec{x}_{i-1}, \vec{x}_i, \vec{x}_{i+1}\right) \ . \tag{2.9}$$

We introduce a third kind of radius of curvature function, which we call *non-local* radius of curvature:

$$\rho_{NL}\left(\vec{x}_i\right) \equiv \min_{\substack{0\leq j<k\leq n \\ j\neq i\neq k \\ k-j+|k-i|+|j-i|>4}} r\left(\vec{x}_i, \vec{x}_j, \vec{x}_k\right) \ . \tag{2.10}$$

That is $\rho_{NL}\left(\vec{x}_i\right)$ is the radius of the smallest circle containing $\vec{x}_i$ and two other distinct points $\vec{x}_j$ and $\vec{x}_k$, *with exclusion* of the case of i, j, k being *consecutive*, when $|k - j| +$

$|k-i|+|j-i| = 4$. In this way $\rho_{NL}(\vec{x}_i)$ keeps track of purely *non-local* effects, and measures the proximity of different portions of the chain. Finally, the thickness $\Delta[\mathcal{C}_n]$ is the radius of the smallest circle containing three distinct points of $\mathcal{C}_n$:

$$\Delta[\mathcal{C}_n] \equiv \min_{0\le i\le n} \rho_G(\vec{x}_i) \ . \tag{2.11}$$

Let $\mathcal{R}_n$ denote the set of all discrete curves $\mathcal{C}_n$ with the property that the associated piecewise linear curve $\mathcal{C}$ is simple and of a prescribed length $N>0$, and satisfying some definite compactness constraint as for a continuous curve. In order to easily implement the condition of fixed chain length N, we will consider as belonging to $\mathcal{R}_n$ only curves having a fixed constant distance between consecutive points [3]. For simplicity, assume $|\vec{x}_i - \vec{x}_{i-1}| = 1$, for $1\le i\le n$, so that $N=n$. We furtherly restrict curves in $\mathcal{R}_n$, by imposing a self-avoidance constraint, that is a hard-core repulsion $|\vec{r}_i - \vec{r}_j| \ge R_s$ for all pairs of different non-consecutive beads i,j ($i\ne j$, $i\ne j+1$, $i\ne j-1$). In the limit of continuous curves, self-avoidance is not necessary, since it is a natural consequence of thickness maximization. Nevertheless, when dealing with discrete curves, as we will, the self-avoidance constraint has to be enforced in order to avoid trapping in self-intersecting structures or convergence to pathological optimal shapes. Moreover, we are dealing with open curves, so that the end-to-end distance $R_e = |\vec{x}_0 - \vec{x}_n|$ is a free parameter. In the limit $R_e \to 0$, the case of closed, possibly knotted, curves is recovered, which is in itself an interesting mathematical topic. Nevertheless, from a physical point of view it is more sensible to require self-avoidance also between the first and the last bead of the chain, $R_e > R_s$. As we will see, in some regimes this constraint may be crucial in selecting a type of optimal shape over the other.

Having defined $\mathcal{R}_n$ in the discrete case, we would like to identify those ideal curves $\mathcal{C}_n^*$ belonging to it, satisfying

$$\Delta[\mathcal{C}_n^*] = \sup_{\mathcal{C}_n\in\mathcal{R}_n} \Delta[\mathcal{C}_n] \ . \tag{2.12}$$

Just as in the smooth case, a necessary condition can be derived, implied by eq. (2.12), that any ideal closed curve with fixed knotted topology must satisfy [33]. A discrete closed curve can be ideal only if its global radius of curvature function is *constant* and *minimal* on every curved sequence of points of the curve.

[3] In future investigations it would be interesting to relax this condition while still keeping the overall string length fixed.

As we will see, all ideal shapes resulting from our simulations will satisfy this condition, *without* the presence of non-curved sequences of points, exhibiting thus a *constant* global radius of curvature along all points of the chain. Moreover, the ideal discrete curves that we have obtained will show up an even stricter characterization. For the moment, we note that what happens for ideal discrete curves closely resembles the properties of best-packed hard spheres configurations. In the latter case, one maximizes the minimum distance between any pair of points by finding the optimal configuration in which each pair of neighbouring spheres is at a *constant* distance, thus recovering translational invariance when boundary effects are neglected. Stated otherwise, for each sphere the minimum distance between itself and all other spheres is constant. This is the exact analog of the global radius of curvature function being constant for each point of the chain. We can thus interpret the global radius as the generalization of the notion of minimum distance between one sphere and all of the others, when these are connected along a chain. In this case full translational invariance is of course not present anymore, but a sort of tranlational invariance along the chain is still present, since the global radius is constant for all points of the chain.

To conclude this section, note that the presence of the chain implies that *three-body* effects have to be considered. It is possible to further generalize the notion of thickness and define what could be natural candidates for a three-body interaction term [33]:

$$U_{p,3}\left[\mathcal{C}_n\right] \equiv \left[\sum_{0\leq i<j<k\leq n} \frac{1}{\left[r\left(\vec{x}_i,\vec{x}_j,\vec{x}_k\right)\right]^p}\right]^{1/p} . \tag{2.13}$$

The circumradius function $r\left(\vec{x}_i,\vec{x}_j,\vec{x}_k\right)$ (2.2) is the simplest way to assign a *unique* distance characterizing the three points $\vec{x}_i,\vec{x}_j,\vec{x}_k$. The connection with thickness is obtained when considering the limit $p\to\infty$:

$$\lim_{p\to\infty} U_{p,3}\left[\mathcal{C}_n\right] = \frac{1}{\Delta\left[\mathcal{C}_n\right]} . \tag{2.14}$$

To *maximize* the thickness $\Delta\left[\mathcal{C}_n\right]$ of a discrete curve $\mathcal{C}_n$ is somewhat equivalent to *minimize* the *repulsive* potential $U_{p,3}\left[\mathcal{C}_n\right]$. Again, the trivial answer in the absence of any compactness constraint is the straight line. Note, however, that the thickness (2.11) is a *global* quantity, whereas the three-body potential (2.13) is, for finite p, the sum of *local* terms.

Chapter 3

Global Compactness

In this chapter we present the result of numerical simulations performed within the simulated annealing scheme in order to find the optimal shapes of discrete chains (2.12), that is the structures having maximum thickness (2.11), in the presence of *global* compactness constraints.

The optimization problem which we are facing is highly non-trivial, since we are dealing with a large number of continuous variables and the functional to optimize is a global quantity involving the computation of the circumradius function (2.2) for *all* triplets of points. Moreover, the compactness constraint is likely to cause the ‘thickness landscape’ in the configuration space to be very rugged and characterized by the presence of many local maxima closely competing with the global one. Due to this reasons, simulations of discrete chains turned out to be feasible and reliable only for chains with a few tens of beads. It is important to note that we are also interested in the determination of local optimal shapes, especially when these are close competitors of the global one, since the structure selection based on best packing properties that we are investigating, may have been loose, when present at all.

A classic method in combinatorial optimization (finding the minimum of a given functional depending on many parameters) is *simulated annealing*, which in principle allows to avoid trapping in local minima [34]. In section 3.1, we briefly discuss in some detail the procedure which we utilized, based on a simple Metropolis updating rule for Monte Carlo dynamics [122].

In section 3.2, we present and discuss the results obtained when global compactness is implemented by constraining the maximum allowed gyration radius of the chain G. Analogous results can be obtained when placing the chain in a finite volume, namely a cube of side L. In both cases, two distinct families of strings, helices and saddles, appear [36]. The two families are close competitors for optimality and different boundary conditions, also involving the end-to-end distance of the chain, may stabilize one over the other.

In section 3.3, we analyze the scaling properties with increasing chain length of the optimal shapes found in our simulations. We will show that, as it must be, they have the same scaling behaviour of *collapsed* polymer chains. In section 3.4, we discuss some important features of the ideal shapes that we find, with particular respect to the global and non-local radius of curvature functions. The global radius is always constant along the chain, whereas the local and the non-local radius are pushed towards equality by the optimality requirement.

3.1 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 (the opposite of the thickness of a discrete chain in our case) 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 [34]. 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 [123, 124]. The standard Metropolis algorithm [122] 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 $p\left(\Delta E\right) = \exp\left[-\Delta E/T\right]$, 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 appendix A, we give a detailed description of the set of moves which we used in our Monte Carlo dynamics for discrete chains. The unnormalized Boltzmann weight which we want to sample, for discrete curves $\mathcal{C}_n$ in $\mathcal{R}_n$, is

$$p\left[\mathcal{C}_n\right] = \exp\left(\Delta\left[\mathcal{C}_n\right]/T\right) \ , \tag{3.1}$$

where $\Delta\left[\mathcal{C}_n\right]$ is the thickness (2.11) of the curve $\mathcal{C}_n$. 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

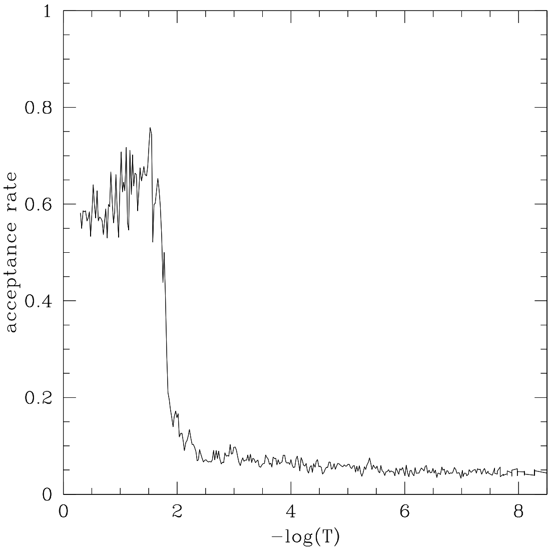


Figure 3.1: Acceptance rate versus fictitious temperature in logarithmic scale for a typical simulated annealing run. The high temperature acceptance rate is not unity because of unavoidable rejections due to compactness and self-avoidance constraints. The low temperature acceptance rate is not zero because we are dealing with continuous variables and arbitrarily small moves may be always accepted.

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 (see figure 3.1), 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.

To conclude this section, we note that the optimal chain conformations, which we will show in this and in the following chapters, have been 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.

3.2 Gyration radius constraint

In this section, we present the result of simulated annealing runs in the case in which compactness has been enforced by constraining the maximum allowed gyration radius of the chain G. The gyration radius R_G of a discrete chain in a conformation $\mathcal{C}_n$ is the mean bead distance from the chain center of mass:

$$R_Gl\left[\mathcal{C}_n\right] = \frac{1}{n+1}\sum_{i=0}^{n}\left|\vec{x}_i - \vec{x}_{cm}\right| \ , \quad \vec{x}_{cm} = \frac{1}{n+1}\sum_{i=0}^{n}\vec{x}_i \ ; \tag{3.2}$$

the gyration radius is usually employed in polymer physics as a typical characteristic length of the chain. It can also be measured by means of light scattering or neutron diffraction experiments [71, 72].

We then consider discrete chains having a fixed constant distance $a = 1$ between consecutive beads, with the constraints of a hard-core repulsion at a distance R_s between different non-consecutive beads, as explained in section 2.4, and with a gyration radius $R_G \leq G$. In practice, a penalty term

$$\mathcal{H}_p\left[\mathcal{C}_n\right] = \begin{cases} A_p\left\{R_G\left[\mathcal{C}_n\right] - G\right\}^2 & R_G\left[\mathcal{C}_n\right] > G \\ 0 & R_G\left[\mathcal{C}_n\right] < G \end{cases} \ , \tag{3.3}$$

with $A_p \gg 1$, has been added to the thickness $\Delta\left[\mathcal{C}_n\right]$ in order to get the fictitious Hamiltonian

$$\mathcal{H}\left[\mathcal{C}_n\right] = -\Delta\left[\mathcal{C}_n\right] + \mathcal{H}_p\left[\mathcal{C}_n\right] \ . \tag{3.4}$$

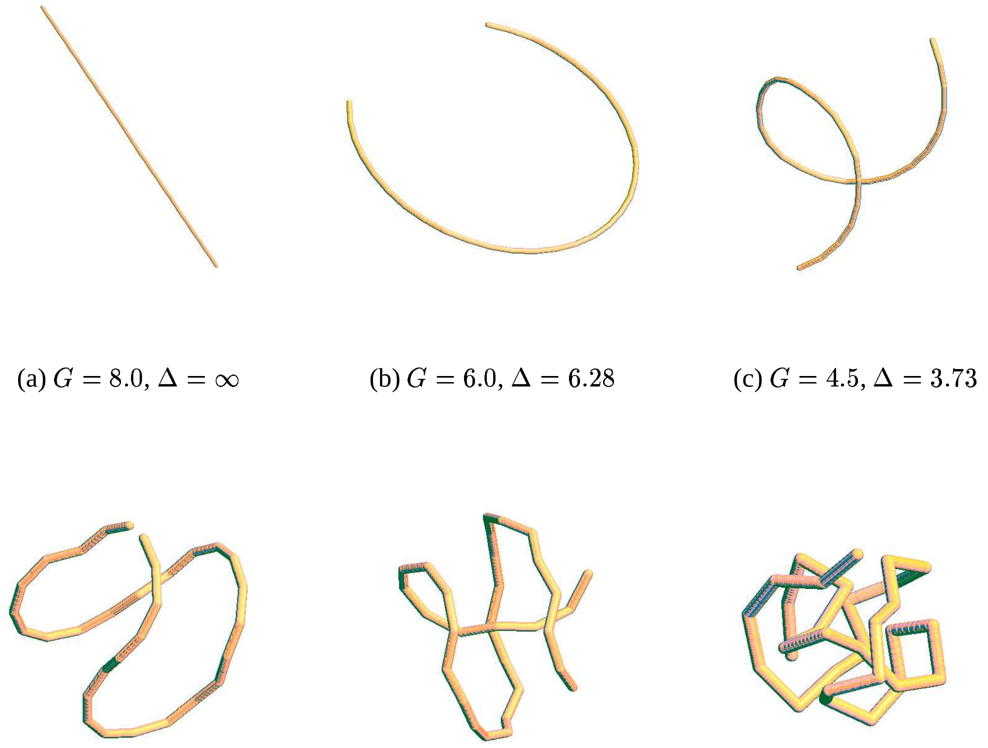
The Boltzmann weight that we have sampled in our simulation is correspondingly:

$$p\left[\mathcal{C}_n\right] = \exp\left(-\mathcal{H}\left[\mathcal{C}_n\right]/T\right) \ . \tag{3.5}$$

We performed simulations for $n = 15, 30, 45$, with G ranging in the interval $[1.5, n/4]$ [1]. As expected, the radius of gyration of the optimal configurations is just G; requiring maximum thickness is equivalent to reduce compactness as much as possible.

The optimal shapes resulting from our simulations can be summarized as in figure 3.2; they have been obtained with $R_s = 1.0$, in bond length units. For $G \gtrsim n/4$ the

[1] For $G \gtrsim n/4$ we expect the constraint on R_G to be uneffective, since in the absence of any compactness constraint the optimal shape is a straight line, for which $R_G \simeq n/4$.



(a) $G = 8.0$, $\Delta = \infty$

(b) $G = 6.0$, $\Delta = 6.28$

(c) $G = 4.5$, $\Delta = 3.73$

(d) $G = 3.0$, $\Delta = 1.86$

(e) $G = 2.5$, $\Delta = 1.39$

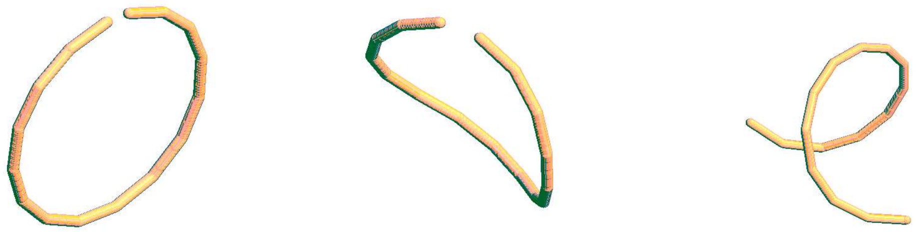
(f) $G = 1.5$, $\Delta = 0.69$

Figure 3.2: Optimal shapes with thickness Δ obtained by constraining strings of length $n = 30$ to have a radius of gyration less than G. The gyration radius of optimal shapes is just G. The excluded volume distance is $R_s = 1.0$.

optimal shape is indeed the *straight line*; then, upon decreasing G, optimal shapes become, in succession, *circle arcs* more and more curled, until $G \simeq \frac{n}{2\pi}$ [2], then *helices*, and eventually *saddles*. In the intermediate regime, $n/10 \lesssim G \lesssim n/4$, helices and saddles are close competitors for optimality, and the role played by tuning the excluded volume distance R_s may be crucial, especially because of the consequent constraint $R_e > R_s$ on the end-to-end distance R_e.

We report in figure 3.3 a detailed analysis of this effect, in the particular case

[2] If $R_s = 1.0$ the optimal shape remains a circle until the $n + 1$ points of the chain are the vertices of a regular polygon inscribed in the circle, implying thus $\Delta = R_G = \frac{1}{2}\frac{1}{\sin[\pi/(n+1)]} \simeq \frac{n+1}{2\pi}$.



(a) $R_s = 0.1$, $\Delta = 2.50$, $R_e = 0.61$

(b) $R_s = 1.0$, $\Delta = 2.35$, $R_e = R_s$

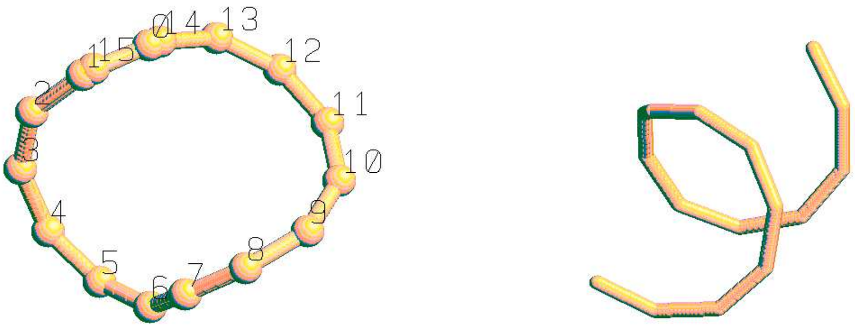
(c) $R_s = 1.6$, $\Delta = 2.20$, $R_e = 5.81$

Figure 3.3: Optimal shapes with thickness Δ obtained by constraining strings of length $n = 15$ to have a radius of gyration less than $G = 2.5$, with different values of the self-avoidance distance R_s.

$n = 15, G = 2.5$, which is just at the threshold value $G/n \simeq 0.16 \simeq 1/2\pi$ above which the optimal shape is a circle arc. For $R_s = 0.1$, the optimal shape is indeed still an almost closed circle ($R_e = 0.61$), whereas for $R_s = 1.0$ the optimal shape is a saddle-like slightly distorted circle ($R_e = R_s = 1.0$). For $R_s = 1.6$, the optimal shape eventually becomes a helix ($R_e = 5.81$). Note that for the saddle $R_e = R_s$, unlike the other two cases, showing that, in this particular case, the optimal saddle shape is *not stable* with respect to R_s variation. The circle and the helix shapes are instead stable in this respect; the circle is the globally optimal configuration, but it requires a very close contact between the ends of the chain, which may be not physical, due to hard-core repulsion. Moreover, both ends of a protein chain have polar groups preferring to be exposed to water. It is thus very unlikely to find the two extremities in close contact between them.

A similar analysis in the case $n = 15, G = 2.0$ is shown in figure 3.4. For $R_s = 0.1$, the optimal shape is a saddle ($R_e = 0.83$), but the configuration is *pathological*, since it is not a solution of the problem that we have posed of maximally inflating a tube around the string [3]. This is caused by the discreteness of the chain, combined with the low value of the self-avoidance distance R_s, as discussed in section 2.4. For

[3] It is easily seen in figure 3.4(a) that due to the exceedingly close contact of bead pairs $(0, 14)$, $(1, 15)$, the two extremities of the swelling tube would immediately overlap.



(a) $R_s = 0.1$, $\Delta = 1.65$, $R_e = 0.83$

(b) $R_s = 1.0$, $\Delta = 1.53$, $R_e = 5.24$

Figure 3.4: Optimal shapes with thickness Δ obtained by constraining strings of length $n = 15$ to have a radius of gyration less than $G = 2.0$, with different values of the self-avoidance distance R_s. In (a) beads have been numbered and represented as spheres in order to highlight the exceedingly close contact between the pairs of beads $(0, 14)$ and $(1, 15)$.

$R_s = 1.0$, or for higher values, the optimal shape is again a helix ($R_e = 5.24$).

For lower values of the maximum allowed radius of gyration ($G \simeq 0.1n$), saddles are the globally optimal shapes in any case. When G is furtherly decreased, we come to a very compact phase, where many closely competing minima are present, and practically every simulated annealing run yields a different optimal configuration (see an example in figure 3.2(e) and 3.2(f)).

Results similar to those presented in this section can be obtained when compactness is enforced by constraining the chain to remain into a finite volume, e.g. a cube of side L. Again, the trivial straight line, which is optimal for high values of L/n, is curled into a circle arc when L/n is decreased. At variance with the radius of gyration case, as L/n is furtherly lowered, first saddles are selected and then helices [36].

3.3 Scaling of optimal shapes

In this section we analyze the scaling properties of ideal shapes and thickness at varying n and G. As plotted in figure 3.5, the optimal shapes that we have obtained within the simulated annealing procedure exhibit a nice data collapse, when dividing both the thickness Δ and the gyration radius G by the number of beads $n+1$. Our finding can be summarized in the scaling behaviour:

$$\frac{\Delta}{n+1} \simeq d\left(\frac{G}{n+1}\right) \ . \tag{3.6}$$

This is a simple rescaling effect, due to the fact that we are approaching the limit of a continuous string, when the number of beads is increased. Since we keep constant the unit distance between consecutive bead along the chain, doubling the length of the string and all the other relevant lengths, Δ and G, is equivalent to the usual procedure of keeping constant the length of the string, and thus Δ and G, while increasing the number of beads by halving the unit bond distance.

Note that the data collapse is obtained when dividing the thickness Δ by the number of beads $n+1$, and not by the chain length n, as would seem more natural. On one hand, the self-avoidance constraint $R_s = 1$ yields an effective chain length $n+1$. On the other hand, when the radius of gyration is computed, a single discrete bead represents a portion of the corresponding continuum string centred on it. This also leads in a natural way to an effective chain length $n+1$.

Taking into account this factor, the scaling function $d\,(x)$ plotted in figure 3.5 shows how the thickness/length ratio $\hat{\Delta}/\hat{L}$ of a *continuous string* varies as a function of the gyration radius/length ratio $\hat{G}/\hat{L}$:

$$\frac{\hat{\Delta}}{\hat{L}} \simeq d\left(\frac{\hat{G}}{\hat{L}}\right) \ . \tag{3.7}$$

The good data collapse of our data confirms ‘a posteriori’ that our simulated annealing procedure yields consistent results at different chain lengths.

In figure 3.6, we plot in logarithmic scale the scaling function $d\,(x)$ for $x < \frac{1}{2\pi}$, that is below the transition from the circle arc to the helix regime (see the caption of figure 3.5). For all chain lengths considered, different linear regression fits yield the

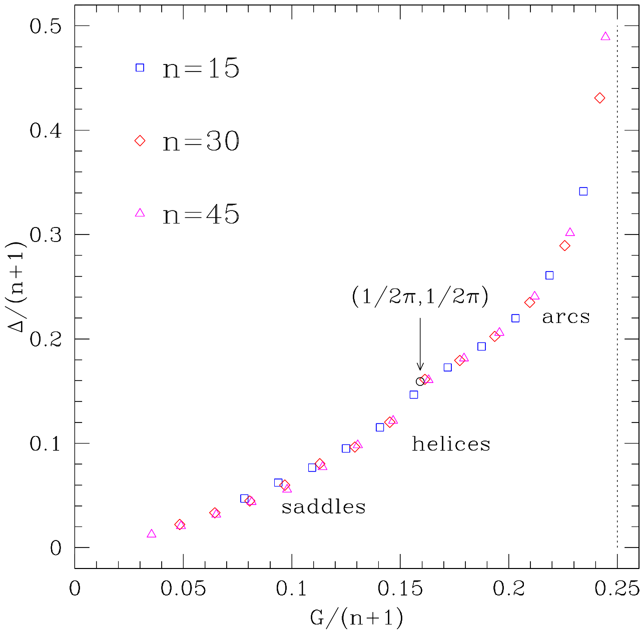


Figure 3.5: Thickness Δ of optimal shapes at varying gyration radius G, for different chain lengths $n = 15, 30, 45$. The data collapse is obtained when dividing by the number of beads $n+1$, as explained in the text. Note that $\Delta \to \infty$ as $G \to 0.25$, in rescaled variables, when the optimal shape is approaching the straight line. The circle at $(\Delta = 1/2\pi, G = 1/2\pi)$ marks the transition between the circle arc and the helix regime.

power law:

$$\frac{\hat{\Delta}}{\hat{L}} \simeq \left(\frac{\hat{G}}{\hat{L}}\right)^{\psi}, \tag{3.8}$$

where the exponent ψ is roughly 1.5.

The exponent ψ can be easily related to the usual correlation length exponent ν which controls the scaling $\xi \simeq N^{\nu}$ of a typical length ξ of the chain, when the number of beads N (the chain length $\hat{L}$ in the limit of a continuum string) is increased [71, 72, 125] . The polymer chain is said to be in the *collapsed* phase when $\nu = 1/3$ ($\nu = 1/d$ in generic dimension) since its *fractal dimension* $1/\nu$ is the same as for ordinary three-dimensional space-filling objects.

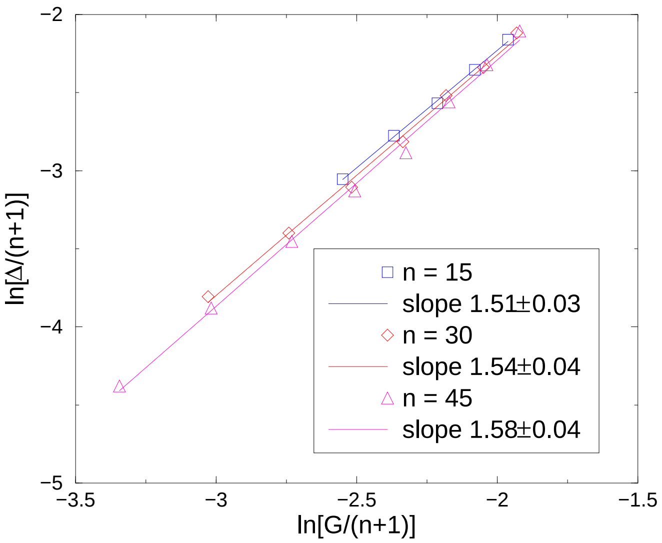


Figure 3.6: Log-log plot of the thickness Δ versus the gyration radius G, for different chain lengths $n = 15, 30, 45$, after proper rescaling by the number of beads $n + 1$. Only $G < 1/2\pi$ values, below the circle arc-helix transition, have been plotted. The lines drawn in the figure are the result of separate linear regression fits for each considered value of the chain length. The slopes and the corresponding errors are reported in the inset.

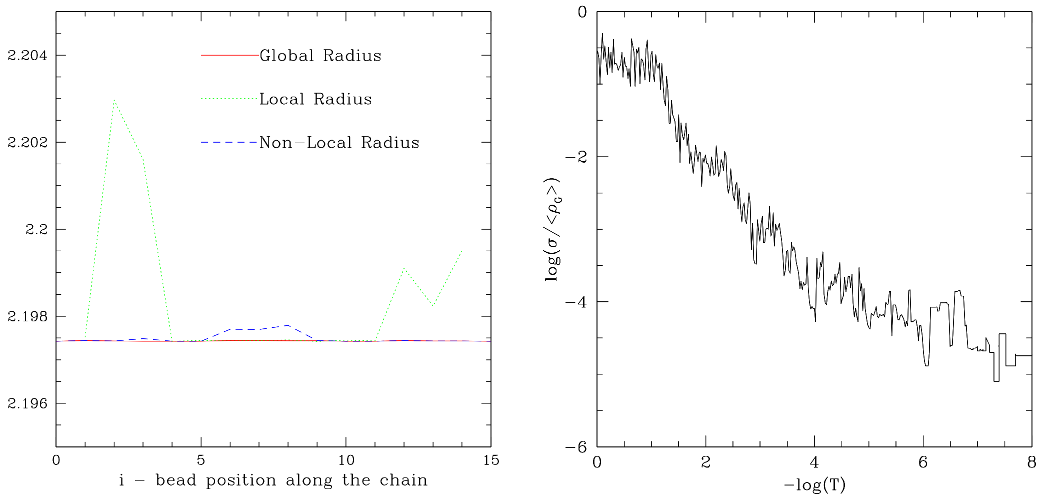
Equation (3.8) can be rewritten:

$$\frac{\hat{G}}{\hat{\Delta}} \simeq \left(\frac{\hat{L}}{\hat{\Delta}}\right)^{1-\frac{1}{\psi}} , \tag{3.9}$$

showing that the thermal exponent controlling how the gyration radius scales with the length of the string is

$$\nu = 1 - \frac{1}{\psi} . \tag{3.10}$$

Note that it is crucial that the string has an excluded volume, due to the finite thickness Δ. Inserting the value $\psi = 3/2$ that results from our data, we get precisely the value $\nu = 1/3$ characterizing the scaling of collapsed chains. Note that the thickness computed in our simulations is always underestimated, in principle, with respect to the 'true' value. This underestimation is more likely to grow, when increasing the chain length and investigating more compact regimes. Thus, it is reasonable to expect that the estimate $\psi = 1.58$ for $n = 45$ (see figure 3.6) may converge to the 'collapsed' $\psi = 3/2$ value as slower annealing simulations are performed, and to conclude that



(a) $n = 15, G = 2.5$

(b)

Figure 3.7: (a) Values of the global, local, and non-local radius of curvature functions along the chain for the optimal configuration showns in figure 3.3(c). Note the scale on the vertical axis. (b) Ratio of the variance σ of the global radius of curvature function along the chain and its average value $\langle \rho_G \rangle$, as a function of the annealing temperature T of a typical run, in logarithmic scale.

optimal best packing conformations exhibit the scaling behaviour typical of collapsed chains, as expected.

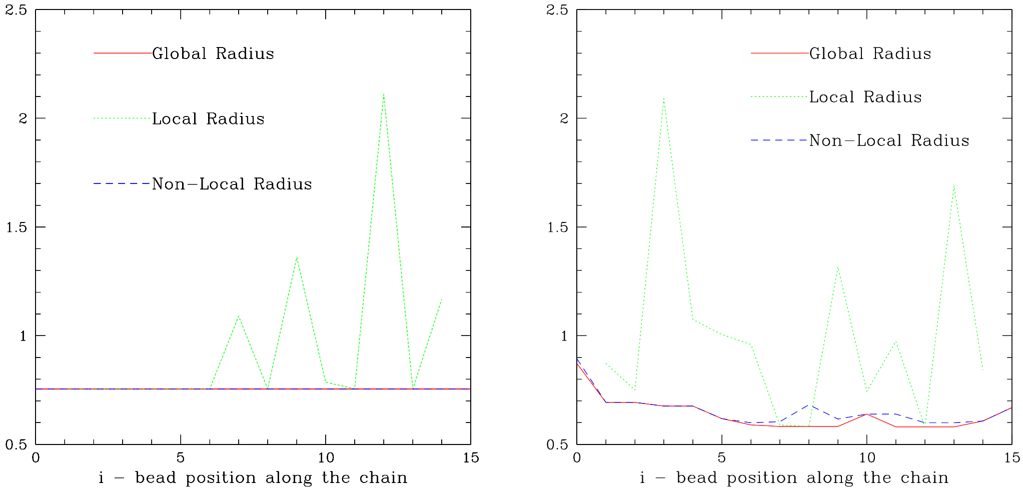
3.4 Local and non-local effects: Optimal degeneracy

In this section we characterize furtherly the properties of optimal shapes, describing in which respect they are *different* from typical random collapsed chain conformations.

In figure 3.7(a) we plot the local, non-local, and global radius of curvature function along the chain for a typical optimal configuration at intermediate values of G (the helical shape of figure 3.3(c) has been used, but qualitative results are the same for other optimal shapes, too, such as saddles and circle arcs).

As predicted by Gonzalez and Maddocks for closed knotted curves [33], the global radius of curvature function of optimal discrete curves is constant along the chain[4]. This is a robust and general feature of our simulations; as shown in figure

[4] We never came across the case of optimal curves having non-curved portions, as happens for the



(a) $n = 15, G = 1.25$

(b) $n = 15, G = 1.25$

Figure 3.8: Values of the global, local, and non-local radius of curvature functions along the chain for a typical compact (a) optimal configuration (b) non optimal configuration. In (a) the non-local radius is always equal to the local radius, at the scale of the figure.

3.7(b), the variance $\sigma = \sqrt{\frac{1}{n+1} \sum_i \left[\rho_G\left(\vec{x}_i\right) - \langle \rho_G \rangle\right]^2}$ of the global radius of curvature function along the chain around its average $\langle \rho_G \rangle = \frac{1}{n+1} \sum_i \rho_G\left(\vec{x}_i\right)$, always decreases, when lowering the fictitious temperature during simulated annealing.

But figure 3.7 shows more; the local and the non local-radius of curvature, which are in general greater than the global radius, are *both* equal to it for optimal shapes in an intermediate compact regime, within a 1% precision. It is easy to see from definitions in section 2.4 that the minimum of the local and the non-local radius is just the global radius, apart when the global radius at a given point is attained for a triplet of consecutive points not centred in that point. This is signalling a characteristic degeneracy which seems to be implied by thickness maximization; the circumradius function is maximum for triplets of both consecutive and non-consecutive points. This feature is more easily understood when thinking about the tube swelling uniformly around the curve. For optimal shapes the tube *at the same time* ceases to be smooth *and* exhibits self-contacts. Optimality requires local and non-local effects to be on

knotted curves studied in references [32, 110, 33].

the same footing.

When moving to lower G values in a more compact regime, the global radius of optimal shapes is still constant along the chain, but both local and non-local radius are not constant anymore. Nevertheless, the degeneracy between local and non-local effects referred to above is still present for some beads of the chain, even if not for all of them (see figure 3.8(a)). It is instructive to perform the same analysis for a compact, non optimal configuration, for which one simply requires the gyration radius to be less than G, with no condition on thickness. As is shown in figure 3.8(b), the global radius is not constant, and is almost always equal to the non-local radius, whereas the local radius is greatly varying. In typical compact configurations, non-local entanglement effects are the relevant ones. The optimality condition would ‘disentangle’ the chain as much as possible until also local bending effects come into play. In a very compact regime, the chain gets frustrated in its search for optimality, and the degeneracy between local and non-local effects takes place only along some portions of the chain.

Chapter 4

Local Compactness

In the previous chapter we have tried to approach the best packing problem for strings, by enforcing compactness as a *global* overall feature of the chain. As a result, we have found many different types of optimal shapes with increasing compactness. In an intermediate regime, regular shapes, such as circle arcs, helices, and saddles, appear, but optimal shapes with very high compactness are highly entangled and much less regular. The best packing of strings seems to produce much more varying and complex structures than the simple regular close-packed arrangement, *fcc* lattice, solving the hard sphere problem. The bulk solution of the hard sphere packing problem, obtained in the limit of infinite system, is homogeneous and isotropic, that is invariant for discrete translations and rotations. Stated otherwise, invariance for isotropic scale transformations at each point of the system holds.

Considering for the moment the case of a continuous string, the presence of the chain defines instead a preferential direction, the tangent to the curve, which in general varies along the chain, thus breaking both the isotropy and the homogeneity of the system. When compactness is enforced using global constraints, such as the gyration radius of the whole chain, our findings show that the optimal shape of the string depends on the value of the ratio of the effective string length $n + 1$ 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 would be much more appealing. Does such a solution exist? If it does, how could we detect it? Intuitively, we expect that for an infinite system such a solution should be able to recover at least at a macroscopic level the homogeneity and the isotropy lost because of the presence of the chain at a microscopic level. Of course boundary effects would enormously affect the emergence of a bulk solution [117]. In this respect, note that the gyration radius constraint affects at the same time all beads of the chain on the same footing. In a sense, the chain has no bulk, and all beads are on the boundary. In order to overcome this problem, we have to enforce compactness *locally*.

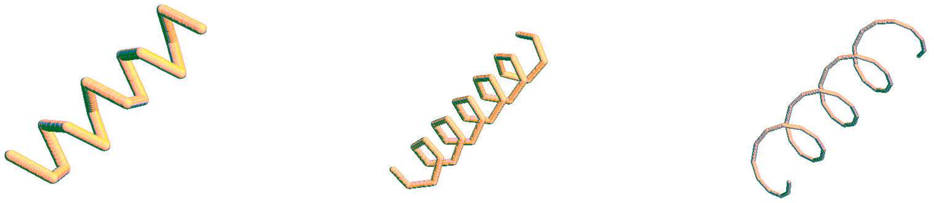
In section 4.1, we present the results that we have obtained, when using only local compactness conditions. For a broad class of local constraints, the optimal shapes are perfect helices, exhibiting translational invariance along the chain, which is the natural symmetry of the system. We thus conclude that *the helical shape is locally the bulk solution to the best packing problem for strings* [36]. Moreover, the optimality condition selects helices with a particular pitch-radius ratio. In section 4.2, we analyze the structures of naturally occurring proteins taken from the Protein Data Bank, and show 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 [36].

Clearly long helices with many turns are not globally compact object. In section 4.3, we discuss a possible way to characterize the compact bulk best-packing conformation. We require the system to be homogeneous and compact, by imposing a *constant* minimum density of beads at each point. This involves the use of compactness constraints *at all scales* along the chain, and it should result in recovering the scale invariance typical of homogeneous and isotropic systems. Simulations are very hard, but preliminary findings show that optimal configurations emerging from this approach are hierarchical structures with again helices as basic elements, in which a long helix is wound to form a superhelix structure and so on.

4.1 Optimal local packing selects helices

In this section we report the optimal shapes obtained when compactness constraints are enforced only locally. More specifically, we have used constraints of the following type:

$$R_G(i, i+j-1) \leq G_j \qquad \text{for} \quad 0 \leq i \leq N-j+1 \; , \tag{4.1}$$



(a) $n = 15, j = 4, G_j = 0.7$ (b) $n = 30, j = 6, G_j = 1.0$ (c) $n = 45, j = 8, G_j = 1.7$

Figure 4.1: Optimal helical shapes obtained with local compactness constraints as in equation (4.1). The resulting values of the thickness are (a) $\Delta = 0.68$ (b) $\Delta = 0.92$ (c) $\Delta = 1.88$. We have used $R_s = 0.5$ for the excluded volume distance. The emergence of helices does not depend on R_s values, except that the formation of too tight helices is forbidden. As in the global compactness case, constraints (4.1) are satisfied as equalities.

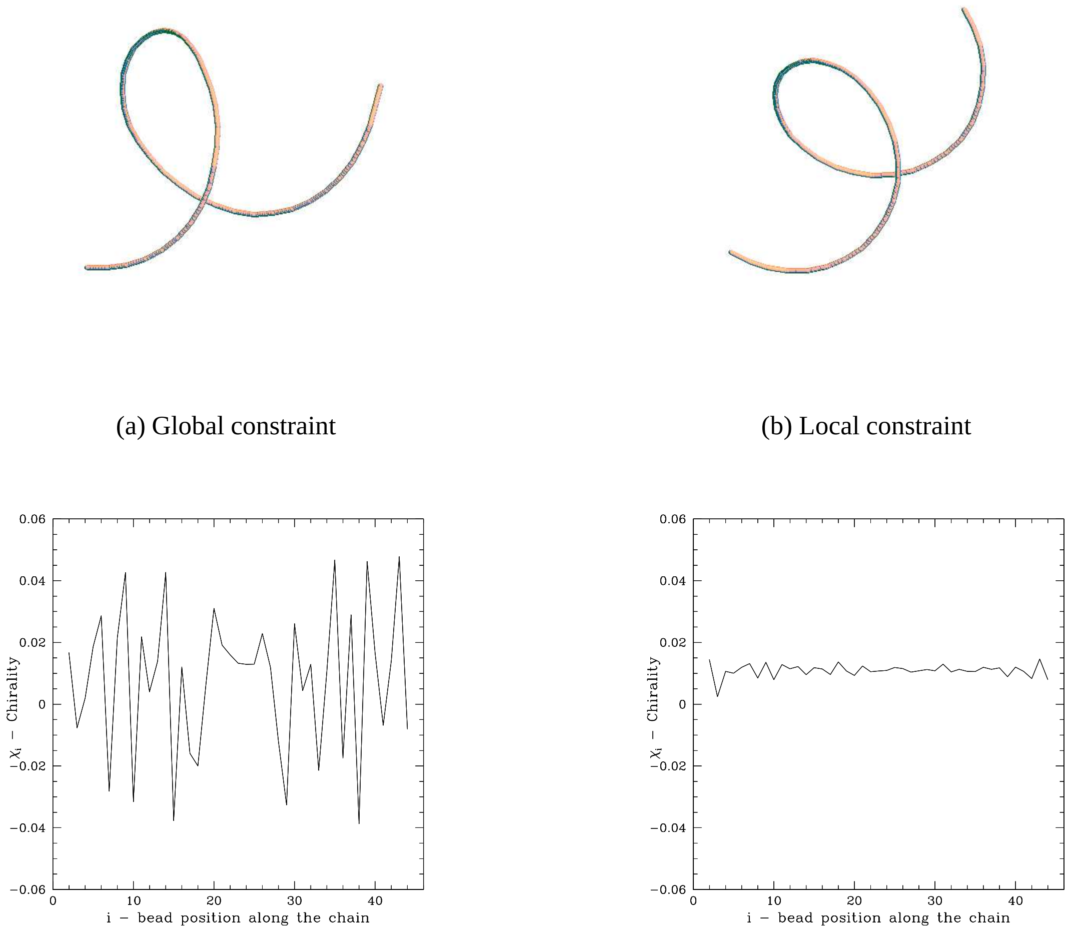
where

$$R_G(l,k) = \frac{1}{k-l+1}\sum_{i=l}^{k} |\vec{x}_i - \vec{x}_{cm}(l,k)| \ , \quad \vec{x}_{cm}(l,k) = \frac{1}{k-l+1}\sum_{i=l}^{k} \vec{x}_i \ , \tag{4.2}$$

is the gyration radius of the set of $k-l+1$ beads $l, l+1, \cdots, k$, with $l < k$. In other words, we are constraining the gyration radius of all sets of j consecutive beads along the chain to be less than G_j, where j is a small number, say $j \lesssim 10$, in order to have a truly local constraint. As in the case of global constraints, conditions (4.1) are enforced by introducing several penalty terms, similar to (3.3), in the fictitious Hamiltonian $\mathcal{H}(\mathcal{C}_n)$ leading to the Boltzmann weight (3.5), that we sample in our simulations.

By varying j and G_j within the following range of values, $4 \leq j \leq 10$ and $0.15 \lesssim G_j/j \lesssim 0.25$, we find always the same optimal configuration, that is a perfect helix, as shown in figure 4.1. Again, similar results can be obtained when local compactness is enforced by constraining each set of j consecutive beads to stay in a finite volume, for example a cube of side L or a sphere of radius R. The number of beads per turn of the optimal helices varies with j and G_j, as shown in figure 4.1.

Optimal helices found by requiring local compactness are extremely regular; much



(a) Global constraint

(b) Local constraint

(c) $G = 6.75$, $\Delta = 5.60$

(d) $j = 10$, $G_j = 2.45$, $\Delta = 5.14$

Figure 4.2: Examples of optimal helical shapes found by imposing (a) global ((b) local) compactness for a chain length $n = 45$. The corresponding chiralities χ_i are plotted below as a function of bead position i along the chain.

more than those obtained with global constraints in the previous chapter. In figure 4.2, chirality χ_i is plotted along the chain in both cases, where

$$\chi_i = \vec{u}_i \cdot (\vec{u}_{i-1} \wedge \vec{u}_{i-2}) \ , \quad \vec{u}_i = \vec{x}_i - \vec{x}_{i-1} \ , \tag{4.3}$$

Chirality fluctuations in ‘global’ helices are much greater than in ‘local’ ones. Note

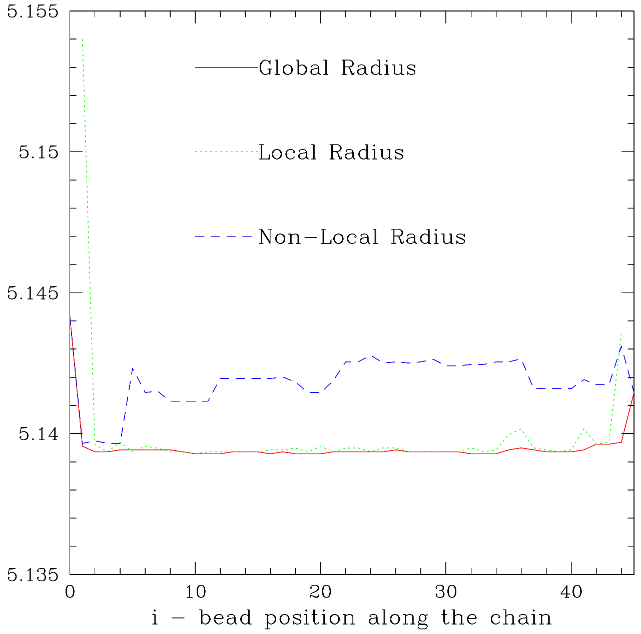


Figure 4.3: Plot of the local, non-local, and global radius of curvature function along the chain for the optimal helix shown in figure 4.2(b). Note the scale on the vertical axis.

that, in the latter case, deviations from the otherwise extremely regular pattern are present at both ends of the chain. As discussed in the introduction to this chapter, we interpret these deviations as boundary effects, and the helical pattern as the bulk local solution to the best-packing problem for strings.

Note that thickness maximization does not imply any ‘a priori’ violation of chirality reversal invariance, and indeed the optimal helices that we find may be left-handed (figure 4.1(a)), as well as right-handed (figure 4.1(b)). In fact, it may happen that the annealing procedure gets trapped in non-optimal configurations where two helices having opposite handedness are connected together. The emergence of long regular helices with many turns is thus the consequence of requiring best packing with local compactness, and is not an artefact of possible biases introduced in our simulations, favouring chain conformations having chirality with a definite sign.

In figure 4.3 the local, non-local and global radius of curvature function along the chain of an optimal helix are plotted. The degeneracy of local and non-local radius, observed also for optimal shapes with global compactness conditions, is now realized within less than 0.1%, apart again from possible deviations at both ends of the chain.

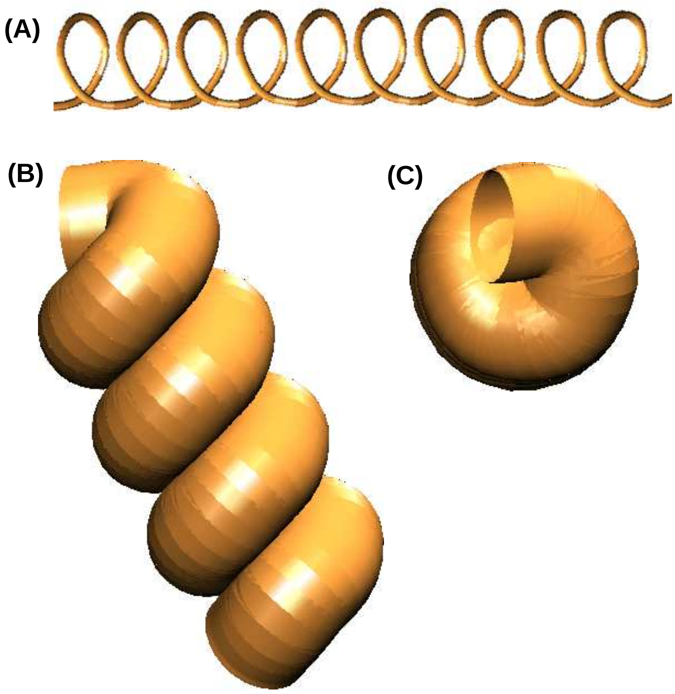


Figure 4.4: Optimal helix with local constraints; $n = 66$, $j = 6$, $G_j = 1.0$. **(A)**, Bare skeleton of the optimal helix connecting the discrete beads. **(B)**, **(C)**, side and top views of the same helix inflated to its thickness by means of the VMD visualization program.

The geometrical meaning is the same discussed in the previous chapter. Optimal packing requires that the tube which may be inflated around the helix stops swelling because local singularities and self-intersections between consecutive turns occur at the same time. This has an immediate visualization in that, as shown in figure 4.4, when the tube is inflated to its thickness around the optimal helix, there is no free space left either between consecutive turns of the helix or in the plane perpendicular to the helix axis. The occupation of three-dimensional space has been thus optimized, as was our aim.

When dealing with continuous curves, the equality of local and non-local radius determines a particular value c^* of the ratio $c = p/r$ of the pitch, p, and the radius, r, of the circle projected by the helix on a plane perpendicular to its axis. Simple analytic calculations, reported in appendix B, yield the value $c^* = 2.512$ for the pitch-radius ratio of optimal helices [36]. This is a special critical value separating two different regimes. If $c > c^*$, the local radius is smaller than the non-local radius and the tube would stop swelling because of local singularities, leaving free space between consecutive turns. If $c < c^*$, the non-local radius is smaller than the local radius, and the tube would stop swelling due to self-intersections between consecutive turns, leaving free space along the axis of the helix[1].

For a discrete string, the special critical value c^* of the pitch/radius ratio depends on the discretization level. In order to check whether a discrete chain configuration is optimal or not, the correct quantity which has to be controlled is the ratio

$$f\left(\vec{x}_i\right) \equiv \frac{\rho_{NL}\left(\vec{x}_i\right)}{\rho_L\left(\vec{x}_i\right)} \ , \tag{4.4}$$

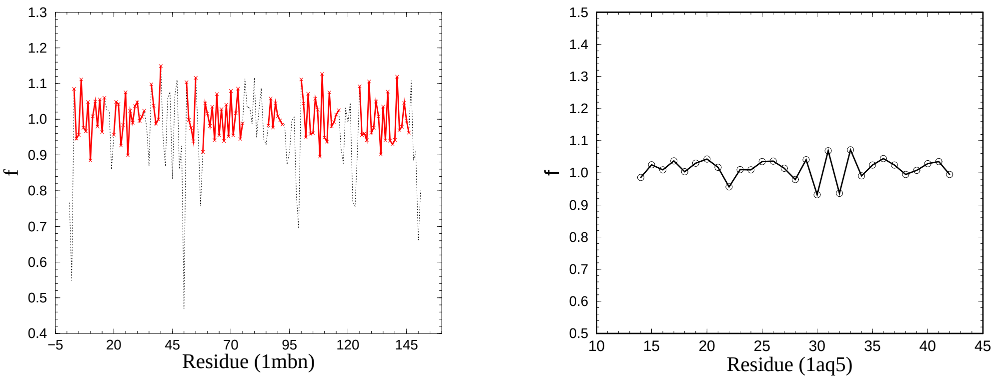
defined as a function of bead position i along the chain, with average value $\overline{f} = \sum_{i=0}^{n} \frac{1}{n+1} f\left(\vec{x}_i\right)$. In this way $f = 1$ just at the transition described above, whereas $f > 1$ in the 'local' regime and $f < 1$ in the 'non-local' regime. Optimal helices found with local compactness conditions have a typical average value $\overline{f} = 1$ within 0.1%. In the next section, we will compute $f\left(\vec{x}_i\right)$ and $\overline{f}$ for naturally occurring proteins extracted from the Protein Data Bank.

4.2 Optimal helices in proteins

As explained in chapter 1, in nature there are many istances of the appearance of helices. For example, many biopolymers, such as proteins and enzymes, have backbones which frequently form helical motifs, as seen in section 1.1.3. It has been argued that the emergence of such motifs in proteins is the result of the evolutionary pressure exerted by nature in the selection of native state structures that are able to house sequences of amino acids which fold reproducibly and rapidly [30], and are characterized by a high degree of thermodynamic stability [83]. Furthermore, because of the interaction of the amino acids with the solvent, globular proteins attain compact shapes in their folded states [4].

It is then natural to measure the shape of these helices and asses if they are optimal in the sense described in this thesis. In figure 4.5, we plot the function $f\left(\vec{x}_i\right)$ defined in equation (4.4) for the discrete chains formed by the backbone C^α atoms

[1] The free space along the axis does not appear immediately as $c \lesssim c^*$, due to the presence of an 'inner hole' in the tube at $c = c^*$. If the helically wrapped tube were hollow, it would be possible to see light through it as in a spyglass [126]. Therefore, as $c \lesssim c^*$, first the inner hole disappears and then free space appears along the axis.



(a) Myoglobin

(b) Collagen

Figure 4.5: Plot of f values as a function of sequence position for (a) sperm whale myoglobin (b) a single collagen helix. In (a) the solid thick lines are the portions of the polypeptide chain forming α-helices. In (b) we considered the residues 14-41; the same plot for each of the three collagen helices would simply superimpose.

of the two protein structures $1mbn$ and $1aq5$ in the Protein Data Bank. Sperm whale myoglobin ($1mbn$) was the first three-dimensional protein structure to be resolved in 1958. It stores oxygen by reversibly binding it at a bound heme group. Myoglobin structure is characterized by the presence of 8 different portions of the peptidic backbone forming α-helices, with slight variations of the parameters controlling the helix geometry, that is number of residues per turn, angular rotation per residue, translation per residue. The helical portion of $1mbn$ structure are shown as thick solid lines in figure 4.5(a). It can be seen that f values are swinging just around the optimal unity value in correspondance of helices, whereas C^{α} not corresponding to α-helices have f-values quite less than unity. Of course one cannot expect $f = 1$ for each single C^{α}, because of residue specificity. By measuring $\overline{f}$ for different α-helices found in different unrelated proteins, we get the average value $\langle\overline{f}\rangle = 1.03 \pm 0.01$.

We then perform the same analysis for the backbone C^{α} atoms of a single collagen helix $1aq5$, a protein present in the chicken cartilage matrix tissue. What is important for us, collagen helices have a rather different geometry from average α-helices, e.g. a different number of residues per turn (3.0 instead of 3.6). Nevertheless, as shown in figure 4.5(b), f values are again oscillating closely around the optimal unity value, hinting 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 [36].

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 that secondary structures in natural proteins have a much larger configurational entropy than other compact conformations [29]. This uniformity in the free volume distribution seems to be an essential feature because the requirement of a maximum packing of backbone sites by itself does not lead to secondary structure formation [104, 105].

4.3 Compactness at all scales and hierarchical optimal structures

In section 4.1, we have seen that optimal helices are *locally* the bulk solution to the best packing problem for strings, with no dependence on the kind of constraint used and on boundary effects. Still, helices are not compact object, so we are left with the problem of which are the compact configurations being *globally* the bulk solution to the best packing problem. A natural generalization of the use of local compactness condition would be to enforce compactness constraints over *all* length scales, and we may expect as a result the emergence of a hierarchical structure with helices as basic elements.

A fixed compactness degree μ^* over all *macroscopic* length scales can be obtained by imposing the following conditions on the *microscopic* local density of beads $\mu_l(\vec{r})$:

$$\mu(\vec{r}_0, d) \geq \mu^* \qquad \text{for all} \quad d, \vec{r}_0 \ , \tag{4.5}$$

where

$$\mu(\vec{r}_0, d) = \int_{|\vec{r}-\vec{r}_0|<d} \mathrm{d^d r} \frac{\mu_\mathrm{l}(\vec{\mathrm{r}})}{\frac{4}{3}\pi \mathrm{d}^3} \ , \quad \mu_\mathrm{l}(\vec{\mathrm{r}}) = \sum_{\mathrm{i}=0}^{\mathrm{n}} \delta(\vec{\mathrm{r}} - \vec{\mathrm{x}}_\mathrm{i}) \ . \tag{4.6}$$

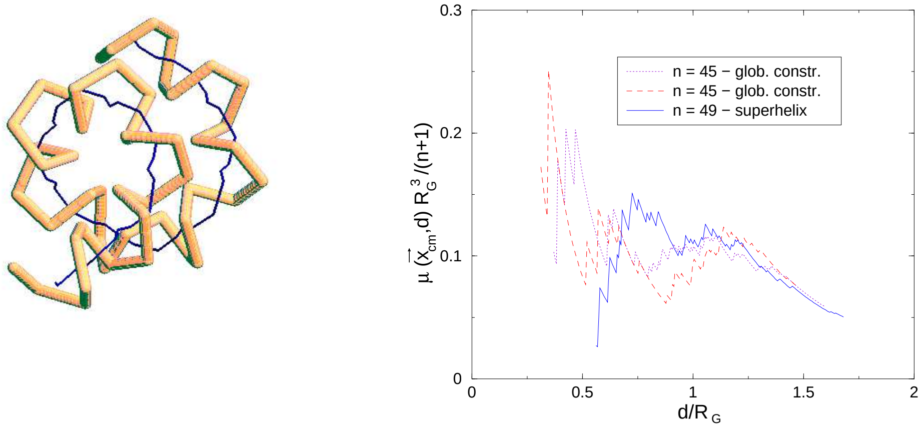
In other words, we would require that the macroscopic average density of beads, computed within a sphere of radius d centred in $\vec{r}_0$, be at least μ^* for all spheres with a macroscopic radius $d \gg 1$. Besides ensuring a finite density of monomers, the enforcement of constraints (4.5) should also more closely match the notion of the optimization of space occupation. On one hand, it is intuitive that the latter should indeed imply the smoothing of density fluctuations, thus recovering the feature of a scale invariant, homogeneous and isotropic system. On the other hand, the optimality condition of maximum thickness, together with conditions (4.5), should 'swell' the chain, just to recover $\mu(\vec{r}_0, d) \simeq \mu^*$, $\forall\ \vec{r}_0, d$, in the optimal configuration.

In practice, it is not possible to utilize conditions (4.5), since one has to face the presence of surface effects. The constraint should indeed not be enforced, when the sphere of radius d is not entirely contained within the 'core' region having a finite density of beads, but we are not able to determine 'a priori' whether this happens or not.

We proceed along a different route, by noting that, if the chain is in the collapsed phase characterized by a finite density of beads, different chain pieces of intermediate size can be thought of as different small chains in the so-called molten chains regime. In the molten chains regime, many different chains are entangled together in a highly compact fashion. It is well known that, due to screening effects, self-avoiding molten chains behave effectively like ideal chains, so that the gyration radius of a N-step chain scales as $R_G \sim N^{1/2}$ [72]. We will thus enforce analogous constraints on the gyration radius of all set of consecutive j beads, for *different* values of j, $j_{min} \leq j \leq j_{max}$:

$$R_G\,(i, i+j-1) \leq G_{j_{min}} \left(\frac{j}{j_{min}}\right)^{1/2} \quad , \tag{4.7}$$

where $G_{j_{min}}$ is the maximum gyration radius allowed for the smallest number, $j = j_{min}$, of consecutive beads that we constrain. Note that in this way, the 'ideal' scaling exponent $1/2$ could be exploited beyond its range of validity, both in the $j \lesssim j_{min}$ and in the $j \gtrsim j_{max}$ regime. The ideal walk scaling, $R_G\,(i, i+j-1) \sim j^{1/2}$, should indeed hold only in the intermediate regime $1 \ll j \ll N$, for chain pieces long but short with respect to the total chain length, so that different pieces are not too correlated and may be considered as belonging to different chains. For very short chain pieces, $j \gtrsim 1$, shorter than the persistence length, consecutive steps are still correlated with each other and $R_G\,(i, i+j-1) \sim j$. For very long chain pieces, $j \lesssim N$, the usual scaling behaviour of a collapsed chain is recovered, $R_G\,(i, i+j-1) \sim j^{1/3}$. One should thus carefully choose j_{min} and j_{max}. In any case, we will check 'a posteriori' that $\mu(\vec{r}_0, d) \simeq \mu^*$, $\forall\ \vec{r}_0, d$, is indeed verified. Again, conditions (4.7)



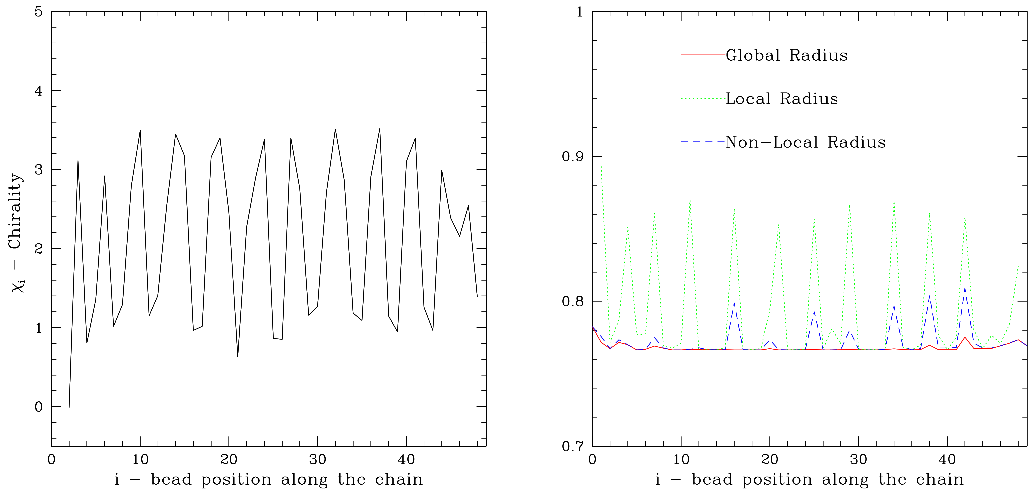
(a) $n = 49$, $j_{min} = 4$, $j_{max} = 25$, $\Delta = 0.77$

(b)

Figure 4.6: (a) Optimal super-helix hierarchical structure obtained with compactness conditions enforced at all scales according to equation (4.7). The helical ‘core’ is shown, computed as the succession of the centres of the circumcircles passing through the triplets of points corresponding to the values of the global radius of curvature along the chain. We have used $R_s = 0.5$ for the excluded volume distance. The emergence of optimal structures does not depend on R_s values, except that the formation of too tight helices is forbidden. Constraints (4.7) are satisfied as equalities within 1÷2%, on average. (b) Plot of the macroscopic density $\mu(\vec{r}_0, d)$, as in equation (4.6), computed within spheres centred at the centre of mass of the chain $\vec{r}_0 = \vec{x}_{cm}$. Three different conformations are considered; the optimal superhelix (a), and two optimal shapes obtained with global compactness conditions, as in chapter 3, with gyration radius $G = 2.25$, $G = 1.63$, respectively. All data has been rescaled in order to yield proper comparison, by using the gyration radius of the whole chain R_G and the chain length n as in axis labels.

are enforced by introducing several penalty terms, similar to (3.3), in the fictitious Hamiltonian $\mathcal{H}(\mathcal{C}_n)$ leading to the Boltzmann weight (3.5), that we sample in our simulations.

The implementation of a very high number of conditions, as in (4.7), for long enough chains is of course a formidable numerical task. We report here the emergence in many simulated annealing runs of hierarchical optimal strucures with helices as basic elements, as shown in figure 4.6(a). Note that the handedness of helices at subsequent levels of the ierarchy is the same. The check that condition (4.5) is indeed



(a) (b)

Figure 4.7: (a) Chirality χ_i as a function of bead position i along the chain for the optimal super-helix conformation shown in figure 4.6(a). (b) Plot of the local, non-local and global radius of curvature function along the chain for the same conformation.

realized as an equality is shown in figure 4.6(b), where $\mu(\vec{r}_0, d)$ is plotted for $\vec{r}_0 = \vec{x}_{cm}$ as a function of d, and compared with the same plot for compact optimal shapes obtained with global compactness conditions in the previous chapter. Since we have considered not too long chains, the result is reasonably good, and shows that the super-helix hieararchical structure yields indeed a more constant density of beads than the other shapes with global constraints. Note, in particular, that the density peak at low d-values is not present anymore.

In figure 4.7 the chirality and the local, non-local and global radius of curvature function along the chain of a optimal super-helix are plotted. As observed in section 3.4, the optimality requirement of degeneracy between local and non-local effects is frustrated by entanglement effects in a very compact regime. The local and the non-local radius of curvature are equal only along some portions of the chain. As opposed to optimal shapes obtained with global compactness constraint, for a super-helix optimal structure, these optimal portions repeat themselves in an *ordered periodic* fashion.

On the basis of these results, it would be very intriguing to conclude that superhelical hierarchical structures are the solution of the best packing problem for strings. Nevertheless, the result here reported are only preliminary, and further work is needed, both numerical and analytical in order to get access to longer chains and better describe the super-helix structure.

Chapter 5

Optimal packing with contact interactions

In the preceding chapters, we have studied which kinds of optimal shapes emerge when compactness is enforced using condition involving the gyration radius, or of the whole chain only, as in chapter 3, or of short chain pieces, as in section 4.1, or of chain pieces of different lengths, as in section 4.3. The gyration radius of a polymer chain is currently used, also in experiments, as a typical carachteristic length of the chain and as an index of its compactness degree [71, 72, 125]. On the other hand, the gyration radius is, in a way, a mathematical abstract concept, which is not likely to appear directly in any physical interaction.

An alternative way to characterize the compactness of a polymer chain is to look at the number of ‘close’ contacts between pairs of different non-consecutive beads. Note that one needs to introduce a typical cut-off distance below which two beads are considered in contact. This approach is more appealing from a physical point of view, since the compactness of globular proteins is due to the interaction of the amino-acids with the solvent, and these are usually modeled as effective pairwise interaction between different residues. Moreover, different recent studies of simple statistical mechanics protein models have shown that many features of folding dynamics are simply encoded in the native state topology, that is in the native *contact map* specifying which close contacts are present between pairs of different non-consecutive beads [25, 30].

In this chapter we report the optimal shapes obtained by maximizing thickness with the constraint of a fixed minimum number of close contacts between pairs of different non-consecutive beads. In section 5.1, we consider the case of a *sharp* definition of close contact, that is two beads are considered in contact if their distance is less than a cut-off distance b. This is equivalent to associate an interaction energy to each couple of beads, characterized by an attractive square well potential of width b, and then to ask for thickness maximization with the constraint of a fixed maximum energy. In section 5.2, we consider the case of a *smooth* definition of close contact, with an attractive long range potential energy between pairs of different beads. Similar results are obtained in both cases, that is the optimal shapes are higly ordered planar hairpin structures, closely resembling the β-sheets appearing in protein structures, when a *low* number of contacts is required, and again helical shapes when a *higher* number of contacts is enforced.

In section 5.3, we pursue furtherly the idea of inducing compactness by constraining the pairwise interaction energy between different beads, and we relax at the same time the condition of thickness maximization. As discussed in section 2.4, the thickness of a discrete chain, defined as the minimum circumradius of all triplets of beads, may be approximated by a sum of local repulsive three-body energy terms (2.13) over all triplets [33]. Thickness maximization would thus become equivalent to the minimization of a repulsive three-body energy. What happens if the repulsive three-body term is hindered by the presence of a two-body attractive energy term? We have looked for the ground state conformations of discrete chains when the two-body attractive energy is the usual $6-12$ Lennard-Jones potential (the repulsive part of L-J potential plays the role of self-avoidance). By varying the ratio of the amplitudes of the two-body attractive and three-body repulsive energy terms, different interesting shapes are found. In an intermediate regime, helices are again found as the ground state structures.

5.1 Sharp contacts

In this section we report the result of simulated annealing runs, in which the thickness of discrete curves is maximized, with the constraint of having a minimum allowed number of contacts n_c^*. The number of contacts $n_c\left[\mathcal{C}_n\right]$ is defined as

$$n_c\left[\mathcal{C}_n\right] \equiv \sum_{i<j-2} C_b\left(\left|\vec{x}_i - \vec{x}_j\right|\right) \ , \tag{5.1}$$

where the sum does not run on consecutive and next-consecutive pairs of beads in order to avoid the counting of trivial contacts, and

$$C_b\left(r\right) = \begin{cases} 1 & r \leq b \\ 0 & r \geq b \end{cases} \tag{5.2}$$

defines a close contact between two beads whose distance is less than the cut-off distance b.

Alternatively, one can think of $C_b\left(r\right)$ as being the opposite of a square well potential energy, so that $n_c\left[\mathcal{C}_n\right]$ plays the role of the opposite of an attractive two-body energy term. In this view, it is clear that compactness is induced by constraining the chain to have a maximum allowed *negative* repulsive energy. As done in previous cases, the constraint $n_c\left[\mathcal{C}_n\right] \geq n_c^*$ is enforced by introducing a penalty term, similar to (3.3), in the fictitious Hamiltonian $\mathcal{H}\left[\mathcal{C}_n\right]$ leading to the Boltzmann weight (3.5), that we sample in our simulations.

Note that the constraint on the number of contacts is global, in the same way as the constraint on the gyration radius of the whole chain used in chapter 3, since it involves a quantity which is defined for all beads. But, unlike the gyration radius, the number of contacts (5.2) is a sum of *local* terms. On one hand, this should allow to capture the best of both global and local approaches. On the other hand, it is much easier to get trapped in local minima during the annealing procedure, since high barriers in the energy landscape are easily created between different configurations having the same number of contacts, but being topologically very unrelated. Because of this, we are limited to small chain lengths; almost all the result we report in this section are for chain lengths of $n = 14$, with a few preliminary results for $n = 29$. Nevertheless, as we will see, the complexity of the energy landscape reflects itself in a variety of interesting results.

In figure 5.1 we report the optimal shapes found with a chain length $n = 14$, and low values of the minimum allowed number of contacts n_c^*, namely $n_c^* = 1, 2, 3$. The cut-off distance is $b = 1.75$, in bond length units, for all the figures which we will report in this section, but we have obtained similar results for different values of b in

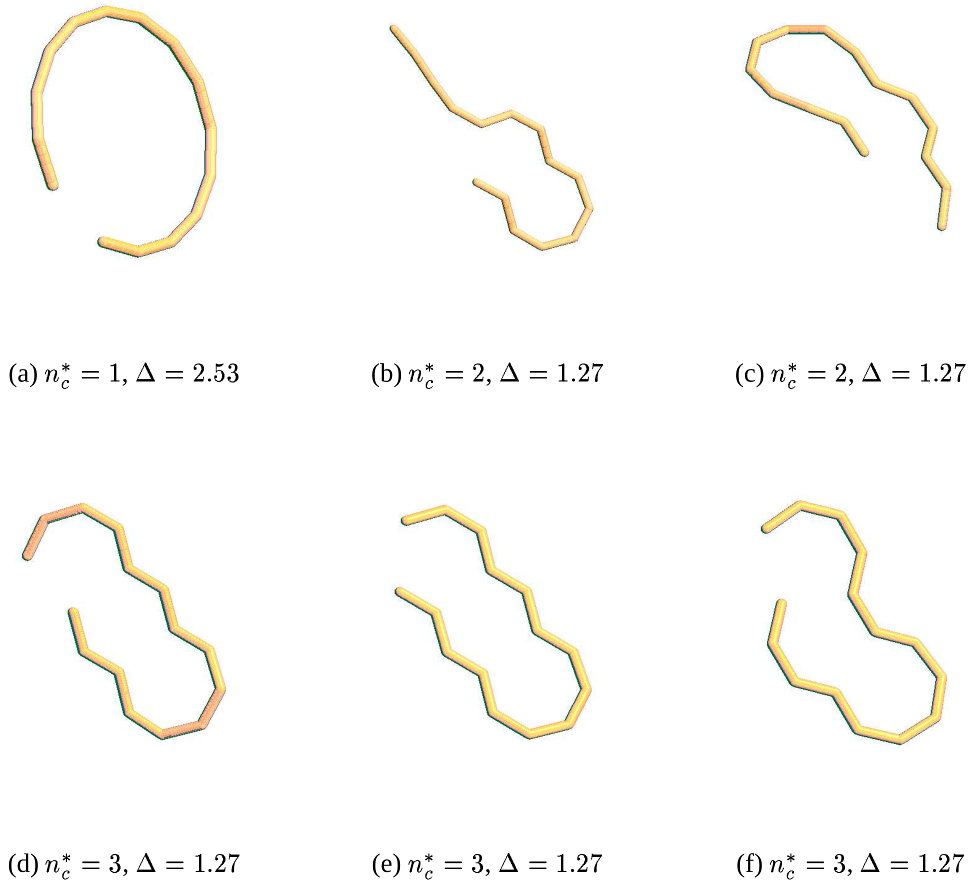

(a) $n_c^* = 1$, $\Delta = 2.53$
(b) $n_c^* = 2$, $\Delta = 1.27$
(c) $n_c^* = 2$, $\Delta = 1.27$
(d) $n_c^* = 3$, $\Delta = 1.27$
(e) $n_c^* = 3$, $\Delta = 1.27$
(f) $n_c^* = 3$, $\Delta = 1.27$

Figure 5.1: Optimal *planar* shapes with low number of contacts. The conformations shown have thickness Δ, and have been obtained by constraining strings of length $n = 14$ to have at least n_c^* 'sharp' close contacts, that is n_c^* pairs of beads less distant than $b = 1.75$. All lengths are in bond length units.

the range $1.2 \leq b \leq 1.9$. We have also enforced self-avoidance, in order to avoid trapping in pathological conformations, as explained in section 3.2, by constraining all distances between different non-consecutive beads to be greater than the excluded volume distance $R_s = 1.0$.

As expected, if $n_c^* = 1$ the optimal shape is an almost closed circle with the ends of the chain in close contact between them. Note that since thickness maximization is swelling the chain, the close contacts present in optimal configurations should be realized just at the borderline value b, at least for n_c^* relatively small in order to avoid very compact and frustrated regimes. This is indeed what we find in all cases, moreover confirming that the self-avoidance constraint does not affect the shapes of optimal conformations, since $R_s < b$.

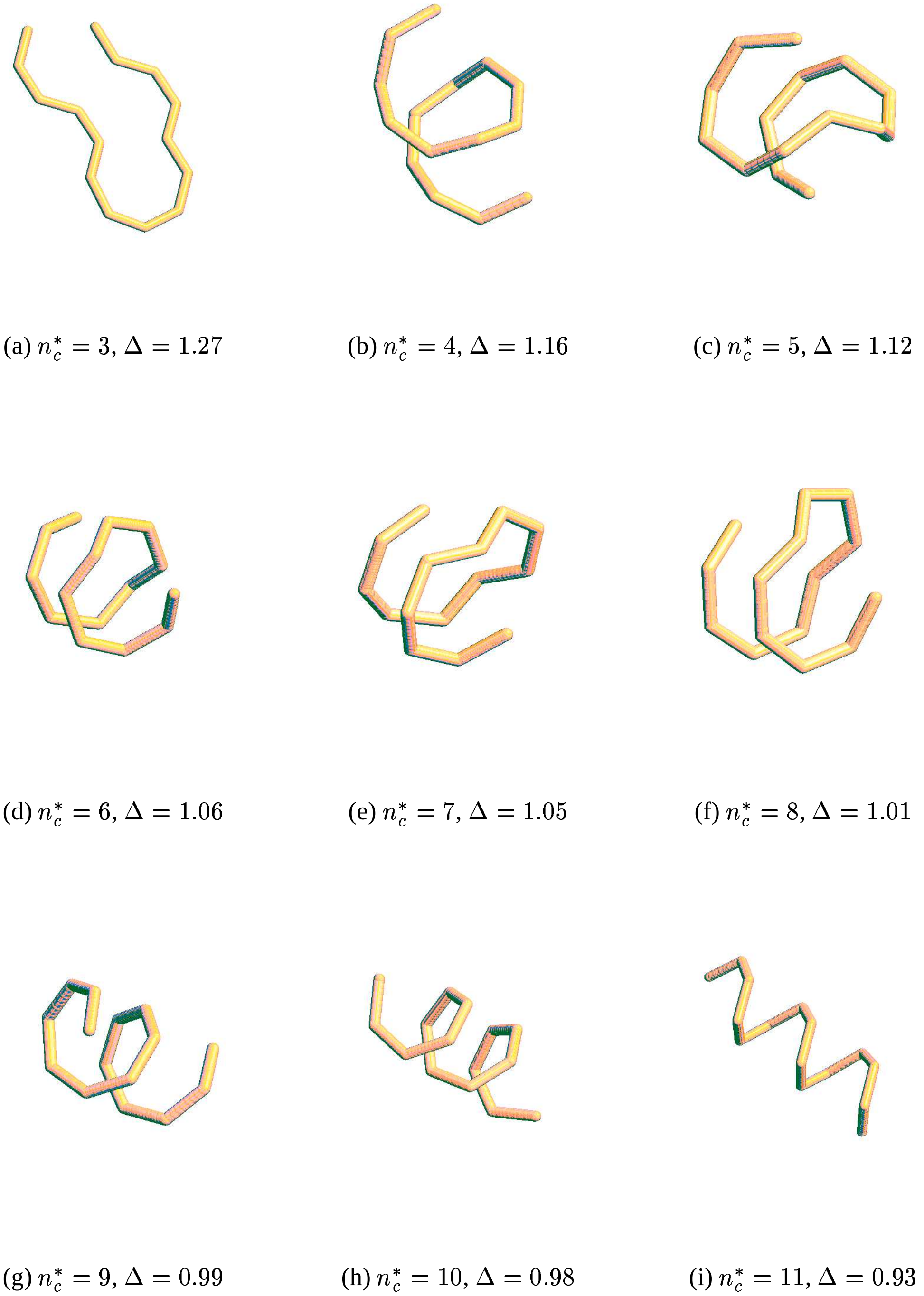
If $n_c^* = 2$, we find many different degenerate optimal shapes having the same thickness. As shown in figure 5.1 for two such conformations, all of them share the same local planar structure in correspondence of the two contacts which have to be formed. The rest of the chain is free, as two contacts do not put sufficient restriction on the overall structure of the chain.

This instead happens for $n_c^* = 3$, when only a little degeneracy is left and four different, but nearly similar, optimal planar hairpin-like structures are found, which are shown in figures 5.1 and 5.2, having the same thickness as the $n_c^* = 2$ optimal structures. All of these structures are perfectly planar, with the common feature of having the two ends of the chain in close contact between them, as for the circular $n_c^* = 1$ shape. Both strands of the hairpin proceed in a zig-zag coupled way which closely resembles the analogous planar order of β-sheets appearing in protein structures, as seen in section 1.1.3.

The geometry of β-sheets in natural proteins is of course much more complicated, since the presence of the side chains causes the polypeptide backbone to zig-zag in the direction orthogonal to the plane containing the hydrogen bonds between adjacent β-strands (see figure 1.3). Nevertheless strikingly similar geometrical features are also present in our optimal hairpin conformations.

Apart from the trivial degeneracy due to the different possible choices in placing the close contact between the opposite ends of the chain (compare (d) and (e) in figure 5.1), two intrinsically different hairpin structures are present. The first structure (see figure 5.1(e)) is characterized by *parallel* ‘contact’ vectors, joining the pairs of beads in close contact between them, as is the case for hydrogen bonds in antiparallel β-sheets. The second hairpin structure (see figure 5.1(f)) has ‘contact’ vectors *alternating* in two different directions, as hydrogen bonds in parallel β-sheets. Moreover, the two strands of optimal hairpins zig-zag in a parallel or antiparallel way, correspondingly.

In figure 5.2, we show the optimal shapes found with higher values of the minimum allowed number of contacts n_c^*, namely $4 \leq n_c^* \leq 11$, again for a chain of length $n = 14$. As soon as $n_c^* \geq 4$, the planarity of optimal structures is lost, and helical shapes appear, though not perfect. For $n_c^* = 11$ the optimal perfect helix with



(a) $n_c^* = 3$, $\Delta = 1.27$

(b) $n_c^* = 4$, $\Delta = 1.16$

(c) $n_c^* = 5$, $\Delta = 1.12$

(d) $n_c^* = 6$, $\Delta = 1.06$

(e) $n_c^* = 7$, $\Delta = 1.05$

(f) $n_c^* = 8$, $\Delta = 1.01$

(g) $n_c^* = 9$, $\Delta = 0.99$

(h) $n_c^* = 10$, $\Delta = 0.98$

(i) $n_c^* = 11$, $\Delta = 0.93$

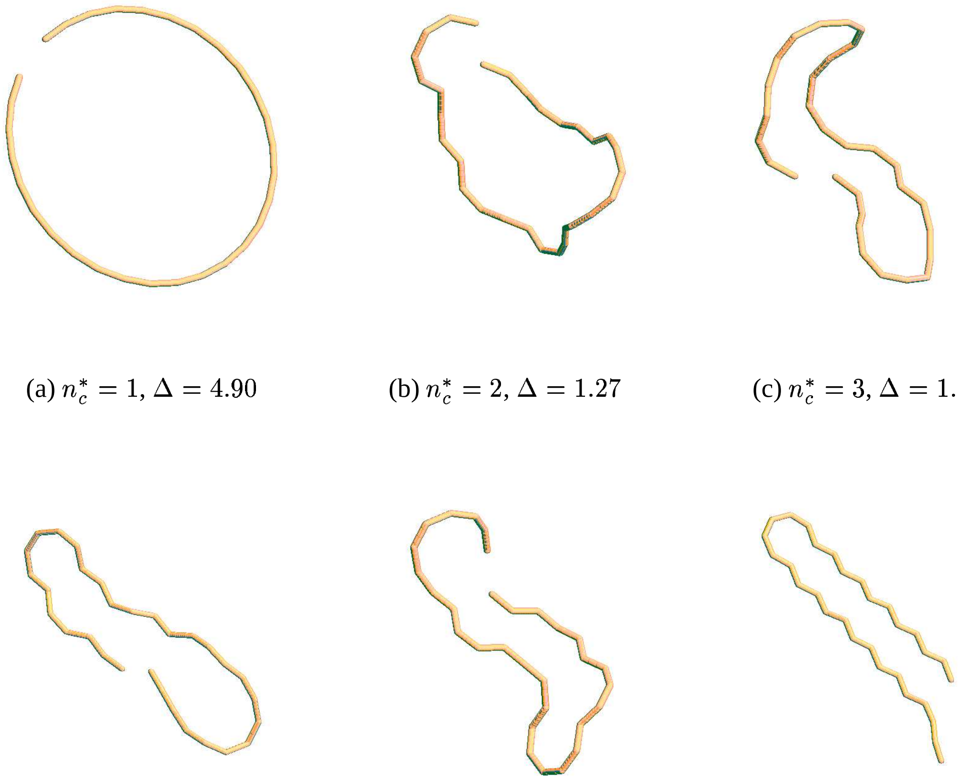
Figure 5.2: Evolution of optimal conformations from planar hairpin structures to helical shapes, when increasing the minimum number of contacts from $n_c^* = 3$ to $n_c^* = 11$. The conformations shown have thickness Δ, and have been obtained by constraining strings of length $n = 14$ to have at least n_c^* 'sharp' close contacts, that is n_c^* pairs of beads less distant than $b = 1.75$. All lengths are in bond length units.

complete degeneracy between local and non-local effects is recovered, that has been already found by imposing local compactness conditions, as in section 4.1.

The transition from a very low compactness regime with planar optimal configurations to a higher compactness regime with helical optimal structures has been also found in chapter 3, when using a global compactness condition involving the gyration radius. Nevertheless, in the latter case only circular structures have been found, whereas in the present case also hairpin optimal structures show up. For all optimal shapes shown in figures 5.1 and 5.2, the global radius of curvature function is constant along the chain. The degeneracy of the local and the non-local radius characterizing the optimal shapes found with the gyration radius constraint is instead not present, in general, when compactness is enforced using the number of contacts, apart from the case $n_c^* = 1$. Further investigations are needed to clarify this point, by taking into account also the possible role played by the fine tuning of the cut-off distance b, in allowing the realization of the optimal degeneracy between the local and the non-local radius. We remark again that the overall features of the optimal shapes shown in figures 5.1,5.2, such as the presence of ‘parallel’ and ‘antiparallel’ hairpins and the emergence of helical shapes when the number of minimum allowed contacts is increased, do not change when b is varied.

The increasing difficulty in obtaining reliable results does not allow us to proceed easily in the study of higher compactness regimes or of longer chain lengths. We only report in figure 5.3 preliminary results for optimal shapes found with a chain length $n = 29$, and low values of the minimum allowed number of contacts $n_c^* \leq 6$. We get the same results as for the shorter chain case $n = 14$. That is, if $n_c^* = 1$ the optimal configuration is an almost closed circle with the two chain extremities in close contact at a distance b, whereas if $n_c^* \geq 2$ we find locally planar optimal configurations, all having the same thickness, which are less and less degenerate and closer and closer to a perfectly planar closed hairpin structure as n_c^* increases. Note that in order to have a perfect closed hairpin one has to tune the chain length. In this sense, $n = 14$ is a lucky case, whereas $n = 29$ is not. In this latter case, for $n_c^* = 6$ there are still some beads left free to move in the optimal hairpin-forming structure, whereas if $n_c^* = 7$ planarity is lost.

Note that locally planar hairpin-forming optimal shapes have *all* the same thickness $\Delta = 1.27$, even for different chain lengths (see also figure 5.1). This confirms that hairpin-like structures are built up by simply repeating in a modular way the same



(a) $n_c^* = 1$, $\Delta = 4.90$ (b) $n_c^* = 2$, $\Delta = 1.27$ (c) $n_c^* = 3$, $\Delta = 1.27$

(d) $n_c^* = 4$, $\Delta = 1.27$ (e) $n_c^* = 5$, $\Delta = 1.27$ (f) $n_c^* = 6$, $\Delta = 1.27$

Figure 5.3: Locally planar hairpin-forming structures with increasing minimum number of sharp contacts n_c^* for a chain length $n = 29$. For $2 \leq n_c^* \leq 6$, all optimal conformations have the same thickness $\Delta = 1.27$ as the planar hairpin-like structures obtained with a shorter chain length $n = 14$.

locally optimal structures which is required in order to form *two* contacts. As can be seen in the different figures, this structures are planar polygons, inscribed in a circle in order to maximize thickness, with two missing sides in order to account for the close contacts to be formed. The thickness of the whole conformation is thus determined locally by the thickness of such modular structures. By varying the number of contacts one controls if it is possible to form the planar hairpin conformation, and in the affirmative case the degeneracy of the allowed optimal conformations sharing the same thickness. The higher the number of contacts, the less degenerate are the possible optimal structures, until for a high enough number of contacts planarity is lost.

5.2 Smooth contacts

In this section we report the results of simulated annealing runs, in which the thickness of discrete curves is maximized, with the constraint of having a minimum allowed number of contacts n_c^*, defined in a *smooth* way, as opposed to the *sharp* definition considered in the previous section. The number of contacts $n_c\left[\mathcal{C}_n\right]$ is defined in a similar way to (5.1):

$$n_c\left[\mathcal{C}_n\right] \equiv \sum_{i<j-2} C_s\left(\left|\vec{x}_i - \vec{x}_j\right|\right) \ , \tag{5.3}$$

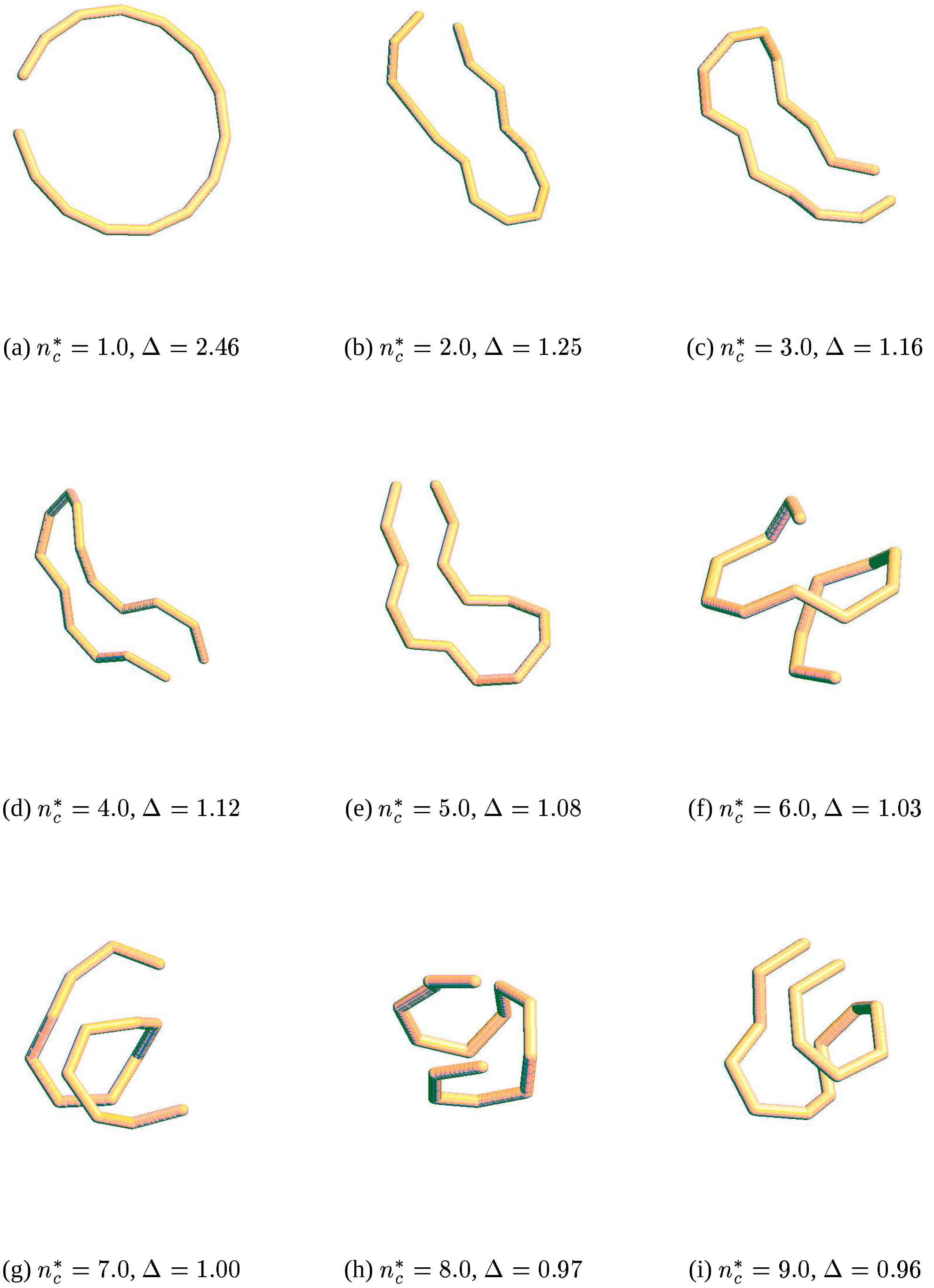
where the sum again does not run on consecutive and next-consecutive pairs of beads in order to avoid the counting of trivial contacts, and now

$$C_s\left(r\right) = 1 - \tanh\left[\left(r - R_s\right)/s\right] \qquad \text{for } \ r \geq R_s \ , \tag{5.4}$$

defines contacts in a smooth way, R_s being the self-avoidance distance of minimum allowed contact between two different non-consecutive beads. Note that $C_s\left(R_s\right) = 1$, so that one should think of $b \sim R_s$, in order to compare with the sharp contact case discussed in the previous section.

It is more natural to think of $C_s\left(r\right)$ as being the opposite of a potential energy, with a hard repulsive wall at $r = R_s$, and a long range attractive tail decaying exponentially with a characteristic length s. In this view, $n_c\left[\mathcal{C}_n\right]$ plays the role of the opposite of an attractive two-body energy term, and again compactness is induced by constraining the chain to have a maximum allowed *negative* repulsive energy. The constraint $n_c\left[\mathcal{C}_n\right] \geq n_c^*$ is enforced, as usual, by introducing a penalty term, similar to (3.3), in the fictitious Hamiltonian $\mathcal{H}\left[\mathcal{C}_n\right]$.

In figure 5.4, we report the optimal shapes obtained for a chain of length $n = 14$, with $R_s = 1.25$ and $s = 0.5$, both in bond length units. Again, similar optimal shapes are obtained if R_s and s are varied within a sensible range. It can be seen that the results are qualitatively the same as in the case of the sharp definition of contacts considered in the previous section. As the minimum allowed number of contacts n_c^* is increased, one finds first circles, then hairpin-like and eventually helical structures. The hairpin-like optimal structures are no more perfectly planar, but they are still



(a) $n_c^* = 1.0$, $\Delta = 2.46$

(b) $n_c^* = 2.0$, $\Delta = 1.25$

(c) $n_c^* = 3.0$, $\Delta = 1.16$

(d) $n_c^* = 4.0$, $\Delta = 1.12$

(e) $n_c^* = 5.0$, $\Delta = 1.08$

(f) $n_c^* = 6.0$, $\Delta = 1.03$

(g) $n_c^* = 7.0$, $\Delta = 1.00$

(h) $n_c^* = 8.0$, $\Delta = 0.97$

(i) $n_c^* = 9.0$, $\Delta = 0.96$

Figure 5.4: Optimal conformations with thickness Δ obtained by constraining strings of length $n = 14$ to have at least n_c^* 'smooth' close contacts, as defined in equation (5.3), with $R_s = 1.25$, $s = 0.5$, both in bond length units. The minimum number of contacts increases from $n_c^* = 1.0$ to $n_c^* = 9.0$.

found at relatively high values $n_c^* = 4, 5$, for which they are instead not found in the correspondent sharp contact case.

At variance with the sharp contacts case, for $n_c^* = 8, 9$ saddle-like optimal structures are found. They are similar to those found in chapter 3, when compactness is enforced by constraining the gyration radius of the chain. Since the displacement of a single bead is now more likely to influence its interaction energy with a higher number of other beads, the smoothing of the definition of close contacts emphasizes the globality of the number of contacts constraint, thus making it more similar to the gyration radius constraint.

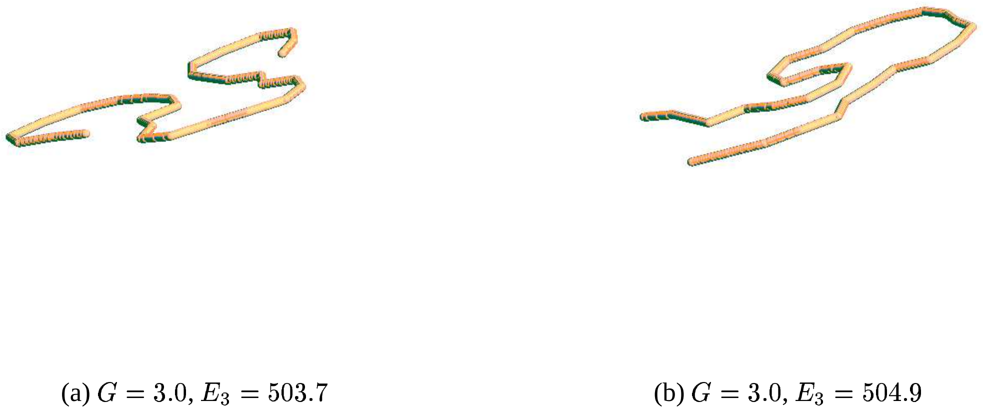
5.3 Lennard-Jones interactions

In this section we adopt a different point of view in our search for optimal structures. Firstly, we relax the condition of thickness maximization, by shifting from the evaluation of the minimum of the circumradius function over all triplet of beads to that of a sum over all triplets of suitable three-body interaction terms, as discussed at the end of section 2.4. Secondly, the three-body interaction term is not maximized within some fixed compactness constraint; rather, we search for the ground state structures of a total energy term being the sum of a *three-body repulsive* term, mimicking the optimality condition, and a *two-body attractive* term ensuring the compactness of the chain.

As shown in section 2.4, the computation of the thickness of a discrete chain is based on three-body effects, being the minimum over all triplets of beads of the circumradius function (2.2) [33]. Different repulsive three-body potentials (2.13) can be introduced, parametrized by the value p of the exponent weighting the single terms in the sum over all triplets [33]. In the limit $p \to \infty$, the definition of thickness is recovered, as from equation (2.14). In this section, we will consider the three-body interaction term with $p = 2$, which we then write in the following way:

$$E_3\left[\mathcal{C}_n\right] \equiv \sum_{0 \le i<j<k \le n} \frac{1}{\left[r\left(\vec{x}_i, \vec{x}_j, \vec{x}_k\right)\right]^2} \,. \tag{5.5}$$

We note again that the energy term E_3 is a repulsive term, having thus the trivial straight line as its ground state structure, in the absence of any compactness constraint or of any other competing attractive term.



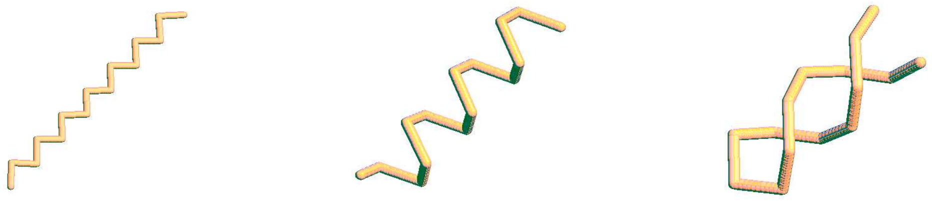
(a) $G = 3.0$, $E_3 = 503.7$

(b) $G = 3.0$, $E_3 = 504.9$

Figure 5.5: Ground state structures minimizing the repulsive three-body energy E_3 defined in equation (5.5) for a chain of length $n = 29$, with the constraint of having a maximum allowed gyration radius G.

To begin, we first search for structures minimizing E_3 when subject to some fixed compactness constraint. We report in figure 5.5 two of such ground state structures when compactness is enforced by constraining the gyration radius of the chain $R_G\left[\mathcal{C}_n\right]$ to be lower than G. The result is surprising, in that perfectly planar structures are obtained. We remark that no 'a priori' shift towards planarity is present in the structure of the interaction term (5.5). Moreover, the planarity of ground state structures has been found to be peculiar to the $p = 2$ case, while is not present anymore for $p = 3$, or for $p = 1$. To our knowledge, this is the first case in which it is shown that a three-body potential has planar ground state configurations and it surely deserves further investigation. The only similar result present in the literature [127] states that in the analogous three-body problem for n free points, that is without the constraint of forming a chain, they arrange themselves along straight lines in their ground state structure, as is indeed natural. It is remarkable that the introduction of the chain constraint *and* of a compactness constraint[1] leads again to an effective reduction of dimensionality, when searching for ground state structures of three-body potential.

[1] Without any compactness constraint, the ground state structure would be again a straight line, also for a chain of beads.



(a) $a = 2 \cdot 10^{-3}$, $E_t = -0.053$

(b) $a = 1 \cdot 10^{-3}$, $E_t = -0.157$

(c) $a = 8 \cdot 10^{-4}$, $E_t = -0.186$

Figure 5.6: Ground state structures minimizing the total energy E_t defined in equation (5.6) for a chain of length $n = 14$, at varying ratio a between the strength of the repulsive and the attractive energy terms. In all cases $b = 0.2$. The double helical shape in (c) is a *local* energy minimum; the ground state structure for $a = 8 \cdot 10^{-4}$ is again a helix as in (b) with an energy $E_t = -0.204$.

We now change the way in which the compactness of the chain is ensured. We exploit furtherly the idea, that we have developed in the previous sections, of enforcing compactness by suitable constraints on the minimum allowed number of close contacts and thus on the maximum allowed value of a correspondingly defined two-body repulsive energy. We introduce a direct competition between the two energy terms, three-body repulsive and two-body attractive, and look for the ground state structures of the total energy interaction term E_t defined as the sum of the two competing terms:

$$E_t\left[\mathcal{C}_n\right] \equiv aE_3\left[\mathcal{C}_n\right] + E_2\left[\mathcal{C}_n\right] \ . \tag{5.6}$$

Different regimes are expected depending on the value of the ratio a of the amplitudes of the two terms. For the two-body attractive term we consider the well known 6-12 Lennard-Jones interaction potential:

$$E_2\left[\mathcal{C}_n\right] \equiv \sum_{0 \leq i < j \leq n} U_{LJ}\left(\left|\vec{x}_i - \vec{x}_j\right|\right) \ , \tag{5.7}$$

where

$$U_{LJ}\left(r\right) \equiv \frac{1}{r^{12}} - \frac{b}{r^6} \ . \tag{5.8}$$

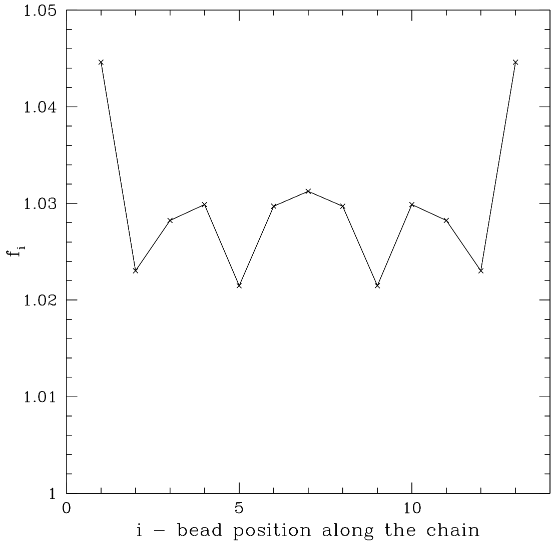
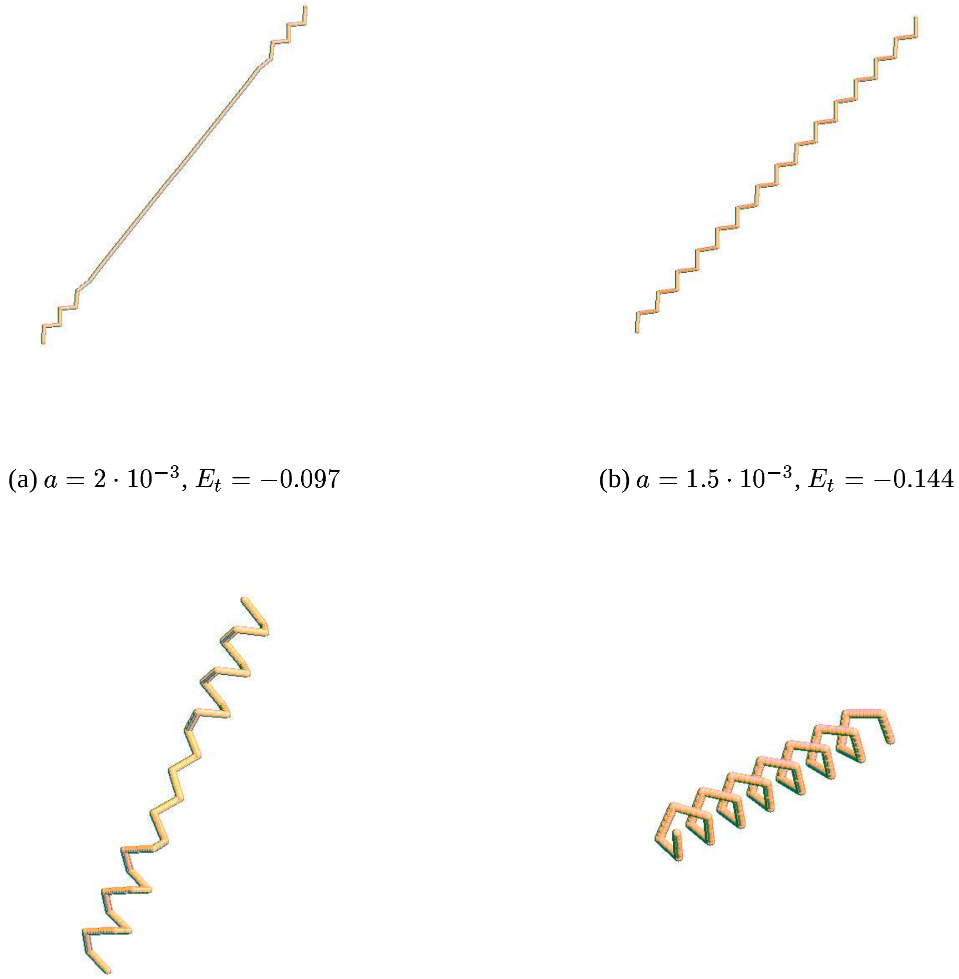


Figure 5.7: Plot of the ratio $f_i = f(\vec{x}_i)$ between the local and the non-local radius of curvature, defined in equation (4.4), as a funtion of bead position along the chain, for the ground state helical structure shown in figure 5.6(b). Note the scale on the vertical axis.

The repulsive term $1/r^{12}$ has the same role as the self-avoidance constraint which we have used until now. Note that for the potential energy (5.8) we can locate the position of the almost hard repulsive wall at $R_s = b^{-1/6}$, and the position of the minimum of the potential at $R_0 = (b/2)^{-1/6} = 2^{1/6} R_s$. In our simulations we considered the case $b = 0.2$, so that $R_s = 1.31$, and $R_0 = 1.47$.

In figure 5.6 we report the ground state structures of discrete curves $\mathcal{C}_n$, which minimize the total interaction energy term $E_t[\mathcal{C}_n]$, for different values a of the ratio of the amplitudes of the two competing energy terms, in the case of a chain of length $n = 14$. If a is sufficiently high, the repulsive term dominates, and the ground state configuration is a straight line. When a is decreased, a first transition occurs to a regime characterized by the emergence of linear zig-zag rods as minimum energy structures. If a is furtherly decreased, helices become the dominating structures. We also report the existence of a local minimum configuration, where our annealing procedure has got frequently trapped, having the intriguing shape of a double helix.

As in section 4.1, we emphasize that helices emerge in our simulations with both possible handedness. No ‘a priori’ shift towards chirality with a definite sign is present, and it is remarkable that the competition of a two-body attractive term and



(a) $a = 2 \cdot 10^{-3}$, $E_t = -0.097$

(b) $a = 1.5 \cdot 10^{-3}$, $E_t = -0.144$

(c) $a = 8 \cdot 10^{-4}$, $E_t = -0.335$

(d) $a = 5 \cdot 10^{-4}$, $E_t = -0.535$

Figure 5.8: Ground state structures minimizing the total energy E_t defined in equation (5.6) for a chain of length $n = 29$, at varying ratio a between the strength of the repulsive and the attractive energy terms. In all cases $b = 0.2$.

a three-body repulsive term causes the emergence of ordered helical structures.

Having relaxed the condition of thickness maximization, it is not surprising to observe that the global radius of curvature is no more constant along the chain for ground state configurations. It is however worthy to note that for helical minimum energy structures the ratio of the local and the non-local radius of curvature is still close to unity, within a few percent, as shown in figure 5.7. We expect this value to approach exactly 1 in the $p \to \infty$ limit.

In figure 5.8 we report the ground state structure minimizing the total interaction energy term $E_t\left[\mathcal{C}_n\right]$, in the case of a chain of length $n = 29$. The importance of considering an energy interaction term being the sum of local terms is now clearly seen. The basic structural motifs are the same already seen, that is straight lines, zig-zag rods and helices as a is more and more decreased, but there are intermediate regimes where two different motifs can coexist in the same minimum energy configuration.

The high compactness regime of very low a values is under current investigation. As is usually the case, the more compact the chain is, the more difficult and long simulations are [123, 124].

Chapter 6

Localization transition of heteropolymers at interfaces

Random heteropolymers are often studied in connection with the behavior of complex macromolecules such as proteins [8, 65, 64]. From a theoretical point of view, random heteropolymers are easier to study, since one can use many tools introduced in the framework of the statistical mechanics of disordered systems. In the case of proteins, it is believed [73, 4] that the main force driving the protein towards the folded state in physiological conditions is the interaction of the amino-acids with the polar solvent. In the folded state, hydrophobic (non-polar) residues are buried in the chain interior, whereas hydrophilic (polar or charged) ones are predominantly confined at the surface.

Interesting questions arises when an interface separates a polar and a non-polar solvent. The latter may be thought to model the lipidic environment inside the cell membrane. It is indeed known that membrane proteins have a higher concentration of hydrophobic amino-acids than proteins in solution [5]. Moreover, random heteropolymers at interfaces have also a technological relevance, due to their effectiveness in the reinforcing of interfaces between two immiscible polymers [42, 43].

In section 6.1, we will compute the quenched free energy (1.4) for a very simple model of a hydrophobic-hydrophilic chain in the presence of an interface separating a polar and a nonpolar solvent, by means of a Gassian variational approach in the replica space.

No excluded volume interaction is present in the model ($\mathcal{H}_{hom} = 0$ within the formalism introduced in subsection 1.2.1), and the interaction disorder-depending term $\mathcal{H}_{dis}$ does not contain self-interaction terms coupling the position of different monomers. There is simply a selective effect due to the preference of each residue towards a specific solvent, according to its hydrophobicity. Frustration is induced by chain connectivity, since two adjacent monomers along the chain may have opposite hydrophobicity. At sufficiently low temperature, the chain has to remain localized at the interface, in order to minimize the energy term and place as much residues as possible in their own preferred solvent. At high temperature the entropic loss due to localization takes over, and the chain might delocalize. Within our variational approach, we characterize the *localization* transition occurring at some intermediate temperature, by explicitly computing the localization length at all temperatures. If the chain has an overall neutral hydrophobic charge the localization transition temperature is infinity, and the chain gets localized at all temperatures.

In section 6.2, we will derive exact bounds on the quenched free energy of an analogous lattice model. This allows us to prove exactly the existence of the localization transition of a non-neutral chain in a variety of cases, namely for ideal and self-avoiding chain and also in the case of *correlated* hydrophobicities. The latter is the case of interest for natural protein sequences, which have been selected by natural evolution.

6.1 Gaussian variational approach in replica space

In this section we study within the Gaussian variational approach a very simple model of an ideal random hydrophilic-hydrophobic chain in presence of an interface separating a polar solvent (e.g. water) and a non-polar one (e.g. oil). The model has been firstly introduced by Garel *et al*[46]. In their work it is shown, with Imry-Ma type arguments [128] and some analytical observations, that the quenched randomness in the sequence of hydrophobic charges causes a localization transition at sufficiently low temperature. The transition temperature becomes infinite in the case of a neutral chain, i.e. a neutral chain is always localized. The case of self-avoiding neutral chains at an asymmetric selective interface has been also studied with scaling arguments [55, 56] and Monte-Carlo simulations [48, 49], yielding a similar scenario.

We have employed a Gaussian variational approach for the calculation of the replicated partition function, in the same spirit as in references [66, 90, 129]. Within this approach it should be possible to cope with the possible presence of many metastable states separated by high energy barriers [91], via the usual Parisi ansatz for replica symmetry breaking (RSB) [63]. On the other hand, the choice of a quadratic trial Hamiltonian does not provide the exponential decay of the monomer density which has been found both in homopolymer adsorption at interfaces [130] and in Monte Carlo simulations of the model considered here [49]. Nevertheless, we think the capability to explore correctly replica space, typical of the Gaussian approach, broadly makes up for this shortcoming. This is also justified ‘a posteriori’ by our results.

Within a replica-symmetric ansatz, we obtain explicitly the temperature dependent localization length for a neutral chain. Our result correctly reproduces the expected scaling behavior in the limit of high temperatures, whereas in the limit of low temperatures it shows a non trivial dependence on the statistical segment length of the chain, the only microscopic length in the problem. For a non neutral chain we obtain the expected localization transition, which turns out to be continuous. Within a one step RSB ansatz, we show the replica-symmetric solution to be unstable, which is usually interpreted as a signal of a phase space decomposition in many energy valleys [63].

6.1.1 Computation of the variational free energy

The partition function of an ideal chain in d dimensions [71], in the presence of an interface at $x = -x_0$, is:

$$Z\left\{\zeta_i\right\} = \mathrm{Tr}_{\{\mathbf{R}_i\}} \exp\left[-\frac{d}{2l^2}\sum_{i=1}^{N}\left(\mathbf{R}_i - \mathbf{R}_{i-1}\right)^2 + \beta\sum_{i=0}^{N}\zeta_i \,\mathrm{sgn}\left(x_i + x_0\right)\right] , \qquad (6.1)$$

where $\beta = 1 \;/\; K_B T$ is the inverse temperature and $\mathbf{R}_i = \left(x_i, \mathbf{R}_{i||}\right)$ $(0 \leq i \leq N)$ are the positions of the monomers, $\mathbf{R}_{i||}$ being the coordinates parallel to the interface. The trace operation is simply a multiple integral over all possible positions of the $N + 1$ monomers. The chain connectivity is implemented as usual with a harmonic potential, l being the typical length of the chain bond. The hydrophobic charge ζ_i is assigned to the i-th monomer, and determines its interaction energy with the solvent. If the monomer is on the polar solvent side (say $x > -x_0$), it takes the energy $-\zeta_i$, whereas if it is on the non-polar solvent side ($x < -x_0$), it takes the energy ζ_i. Thus, when $\zeta_i > 0$ ($\zeta_i < 0$), the monomer is hydrophilic (hydrophobic) and prefers the right side (the left side). The hydrophobic charges ζ_i are identically distributed independent random Gaussian variables, with average ζ_0 and variance Δ, $P\left(\zeta_i\right) = \exp\left[-\left(\zeta_i-\zeta_0\right)^2/\,2\Delta^2\right]\,/\sqrt{2\pi\Delta^2}$; $\zeta_0 = 0$ corresponds to a neutral chain. Any self-interaction between monomers, such as self-avoidance, is neglected in this simple approach. Hence, integration over $\mathbf{R}_{i||}$ can be carried out immediately and the problem becomes one-dimensional.

We are interested in computing the quenched free energy per monomer $f_q = -1/\left(N\beta\right)\ \overline{\ln Z\left\{\zeta_i\right\}}$, where $\bar{\cdot}$ denotes the average over disorder. We introduce n replicas of the chain for a given disorder realization $\{\zeta_i\}$ in order to employ the usual replica trick: $f_q = \lim_{n\to 0}\left(\overline{Z^n}-1\right)/n$. Thus, we deal with an effective one-dimensional homopolymer hamiltonian

$$
\begin{aligned}
\mathcal{H}_n = \ & -\ \frac{\beta^2\Delta^2}{2}\sum_{i=0}^{N}\left(\sum_{a=1}^{n}\ \mathrm{sgn}(x_{i,a}+x_0)\right)^2 - \\
& -\ \sum_{a=1}^{n}\left[\beta\zeta_0\sum_{i=0}^{N}\ \mathrm{sgn}\left(x_{i,a}+x_0\right)-\frac{d}{2l^2}\sum_{i=1}^{N}\left(x_{i,a}-x_{i-1,a}\right)^2\right]\ . \qquad (6.2)
\end{aligned}
$$

The Gaussian variational approach consists in approximating $\mathcal{H}_n$ with a trial function $\mathcal{H}_t$ which is quadratic in the monomer coordinates $x_{i,a}$, for integer n. This provides the bound $\ln\overline{Z_n} \geq \ln Z_t + \langle\mathcal{H}_t-\mathcal{H}_n\rangle_{\mathcal{H}_t}$, where $Z_t = \mathrm{Tr}_{\{x_{i,a}\}}\exp\left[-\mathcal{H}_t\right]$. For simplicity we consider the case of a ring polymer ($x_{0,a} = x_{N,a}$), in such a way to exploit the translation invariance along the chain, and we introduce the corresponding Fourier coordinate k ($\tilde{x}_a\left(k\right) = \sum_{i=1}^{N}x_{i,a}\exp\left(-\,\mathrm{i}\,ik\right)$). In the thermodynamic limit ($N \gg 1$) we get:

$$
\begin{aligned}
\mathcal{H}_t &= \frac{1}{2}\int_{-\pi}^{\pi}\frac{\mathrm{d}k}{2\pi}\sum_{a,b}\tilde{x}_a\left(k\right)\Lambda_{ab}\left(k\right)\tilde{x}_b^*\left(k\right)\ , \\
\Lambda_{ab}\left(k\right) &= \left[\frac{2d}{l^2}\left(1-\cos k\right)+\mu\right]\delta_{ab}-\sigma_{ab}\left(1-\delta_{ab}\right)\ , \qquad (6.3)
\end{aligned}
$$

where the variational parameters μ and σ_{ab} have been introduced, associated with single replicas and replica pairs, respectively.

Let us consider now macroscopic quantities such as the local density of a-monomers, $\rho_a\left(x\right) = \sum_i\delta\left(x_{i,a}-x\right)$, and the so called overlap of replicas a and b, $q_{ab}\left(x,x'\right) =$

$\sum_i \delta\left(x_{i,a}-x\right)\delta\left(x_{i,b}-x'\right)$, with $a<b$. Not surprisingly, the quadratic ansatz (6.3) yields a Gaussian structure for the average of both of them:

$$\begin{aligned} \left\langle \rho_a\left(x\right)\right\rangle_{\mathcal{H}_t} &= N\,\exp\left[-\frac{x^2}{2\lambda_0}\right] / \left(2\pi\lambda_0\right)^{1/2}\ , && (6.4)\\ \left\langle q_{ab}\left(x,x'\right)\right\rangle_{\mathcal{H}_t} &= N\,\frac{\exp\left[-\frac{1}{4}\frac{(x+x')^2}{\lambda_0+\lambda_{ab}}-\frac{1}{4}\frac{(x-x')^2}{\lambda_0-\lambda_{ab}}\right]}{2\pi\left[\left(\lambda_0+\lambda_{ab}\right)\left(\lambda_0-\lambda_{ab}\right)\right]^{1/2}}\ , && (6.5)\end{aligned}$$

where $\lambda_0=\int_{-\pi}^{\pi}\mathrm{d}k\left[\Lambda^{-1}\left(k\right)\right]_{aa}/2\pi$ and $\lambda_{ab}=\int_{-\pi}^{\pi}\mathrm{d}k\left[\Lambda^{-1}\left(k\right)\right]_{ab}/2\pi$. The 'relative' characteristic length $\sqrt{\lambda_0-\lambda_{ab}}$ directly measures the overlap degree of replicas a and b. Indeed, there is a very low probability of finding the monomers of replicas a and b at positions more distant than $\sqrt{\lambda_0-\lambda_{ab}}$. Thus, the larger the overlap length, the less the two replicas are interacting. In the extreme cases: $\lambda_{ab}=0$, $q_{ab}\left(x,x'\right)=\frac{1}{N}\rho_a\left(x\right)\rho_b\left(x'\right)$, the overlap length is the same as the single chain characteristic length, and replicas are not interacting at all. If instead $\lambda_{ab}=\lambda_0$, $q_{ab}\left(x,x'\right)=\rho_a\left(x\right)\delta\left(x-x'\right)$, the overlap length vanishes, and the two replicas are interacting as strong as possible.

From now on, we will consider λ_0 and λ_{ab} as the variational parameters[1], with respect to which the free energy has to be stationary. The interface breaks the translation invariance, so that the position of the center of mass of the chain is a relevant degree of freedom. In the following we will consider the chain center of mass fixed in the origin, according to (6.4), with the interface moving at the variable position $-x_0$, which then becomes another variational parameter.

Within the replica-symmetric ansatz, $\lambda_{ab}=\lambda\ \forall\left(a,b\right)$, the quenched free energy becomes:

$$\begin{aligned} \beta f_q &= -\frac{1}{2}\left\{\ln\left(\frac{2\pi l^2}{d}\right)+h\left[\frac{2d}{l^2}\lambda_0\left(1-v\right)\right]+\frac{2d}{l^2}\lambda_0 v h'\left[\frac{2d}{l^2}\lambda_0\left(1-v\right)\right]+\right.\\ &+\left. 2\beta\zeta_0\,\mathrm{erf}\left(\frac{\tilde{x}_0}{2}\right)+\left(\beta\Delta\right)^2\left[1-\epsilon\left(v,\tilde{x}_0\right)\right]\right\}\ , && (6.6)\end{aligned}$$

[1] The explicit dependence of (λ_0,λ_{ab}) on (μ,σ_{ab}) can be worked out easily in the specific cases we are going to consider. In the replica-symmetric case (i.e. $\sigma_{ab}=\sigma\ \forall\left(a,b\right)$) we have $\lambda=\frac{\sigma\left(\mu+\sigma+2d/l^2\right)}{\left[(\mu+\sigma)(\mu+\sigma+4d/l^2)\right]^{3/2}}$ and $\lambda_0=\lambda+\frac{1}{(\mu+\sigma)(\mu+\sigma+4d/l^2)}$. Similar, but more complicated, equations hold for the case of one step RSB.

where $\tilde{x}_0 = x_0 \left(2/\lambda_0\right)^{1/2}$, $v = \lambda/\lambda_0$, $h\left(A\right) = \sqrt{A^2+1} - A - \ln\left(\frac{1+\sqrt{A^2+1}}{A}\right)$, $h'\left(A\right) = \mathrm{d}h/\mathrm{d}A$, and $\epsilon\left(v, \tilde{x}_0\right) = \int \mathrm{d}x_1 \mathrm{d}x_2 \frac{\exp\left[-\left(x_1^2+x_2^2\right)/2\right]}{2\pi}$ sgn $\left(x_1\sqrt{1+v} + x_2\sqrt{1-v} + \tilde{x}_0\right)$ sgn $\left(x_1\sqrt{1+v} - x_2\sqrt{1-v} + \tilde{x}_0\right)$. The chain is localized if the characteristic length $\sqrt{\lambda_0}$ and the chain center of mass $\tilde{x}_0$ remain finite in the thermodynamic limit $N \to \infty$. The free energy has to be minimized with respect to the variational parameters $\tilde{x}_0$ and λ_0, which are single-replica quantities, and maximized with respect to v, which is a two-replicas parameter [63].

6.1.2 Neutral chain

Let us consider firstly the case of a neutral chain ($\zeta_0 = 0$). The quenched free energy is invariant for the transformation $\tilde{x}_0 \to -\tilde{x}_0$, and it can be easily seen that the minimum of the free energy with respect to $\tilde{x}_0$ is attained for $\tilde{x}_{0,\mathrm{min}} = 0$. Minimization with respect to λ_0, and maximization with respect to v, yield then $\lambda_{0,\mathrm{min}} = l^2/\left(2d\sqrt{2v_{\mathrm{max}}-1}\right)$, where $1/2 \leq v_{\mathrm{max}} \leq 1$ is the unique solution of the equation $\left[\left(1+v\right)\left(2v-1\right)\right]^{1/2} = 2\beta^2\Delta^2\left(1-v\right)^{3/2}/\pi$. Thus, at any finite temperature the chain is localized, in agreement with heuristic arguments [46].

We analyze now the asymptotic behavior of our solution at high and low temperatures. At high temperature ($\beta\Delta \ll 1$) we find:

$$v_{\mathrm{max}} \simeq \frac{1}{2} + \frac{1}{6\pi^2}\left(\beta\Delta\right)^4 \ ; \ \lambda_{0,\mathrm{min}} \simeq \frac{l^2}{2d}\sqrt{3}\pi\left(\beta\Delta\right)^{-2} \ . \tag{6.7}$$

The characteristic length of the chain in the direction orthogonal to the interface is $R_\perp = \sqrt{\lambda_0}$. We have recovered the scaling relation $R_\perp \sim \left(\beta\Delta\right)^{-2\nu}$ coming from the Imry-Ma argument, with the exponent $\nu = 1/2$ of an ideal chain. The argument is expected to fail at low temperature, where it is not possible to neglect anymore the role of the microscopic length l. In the limit of low temperature ($\beta\Delta \gg 1$), we indeed find:

$$\begin{aligned} v_{\mathrm{max}} &\simeq 1 - \left(\frac{\pi^2}{2}\right)^{1/3}\left(\frac{1}{\beta\Delta}\right)^{4/3} \ ; \\ \lambda_{0,\mathrm{min}} &\simeq \frac{l^2}{2d}\left[1 + \left(\frac{\pi^2}{2}\right)^{1/3}\left(\frac{1}{\beta\Delta}\right)^{4/3}\right] \ . \end{aligned} \tag{6.8}$$

At zero temperature, the characteristic length $R_\perp = l/\sqrt{2d}$ is of course of the same order as the microscopic length l. Increasing the temperature results in a power law correction with the non trivial exponent $4/3$.

6.1.3 Non-neutral chain

We consider now the more general case of a non neutral chain ($\zeta_0 \neq 0$). In this case one gets $\lambda_{0,\min} = l^2/\left(2d\sqrt{2v_{\max}-1}\right)$, as in the neutral case, and the following coupled equations for $v_{\max}$ and $\tilde{x}_{0,\min}$:

$$\begin{aligned} \left[\frac{(1+v)\,(2v-1)}{(1-v)^3}\right]^{1/2} &= \frac{2\beta^2\Delta^2}{\pi}\exp\left(-\frac{1}{2}\frac{\tilde{x}_0^2}{1+v}\right) , & (6.9) \\ \frac{\beta\Delta^2}{\zeta_0}\,\mathrm{erf}\left(\frac{\tilde{x}_0}{2}\sqrt{\frac{1-v}{1+v}}\right) &= 1 \,. & (6.10) \end{aligned}$$

There are two different regimes. The high temperature regime, $\beta\Delta^2/\left|\zeta_0\right| < 1$, where equation (6.10) does not admit a solution. The variational paramaters then read $v_{\max} = 1/2$, $\lambda_{0,\min} = \infty$, $\tilde{x}_{0,\min} = \ \mathrm{sgn}\left(\zeta_0\right)\infty$, and the chain is delocalized. The low temperature regime, $\beta\Delta^2/\left|\zeta_0\right| > 1$, where a solution exists with $v_{\max} > 1/2$, $\lambda_{0,\min} < \infty$, $\left|\tilde{x}_{0,\min}\right| < \infty$, and the chain is localized at the interface, though the center of mass is not coincident with the interface itself ($\tilde{x}_{0,\min} \neq 0$). Therefore, lowering the temperature the chain undergoes a localization transition at the critical temperature $\beta_c = \left|\zeta_0\right|/\Delta^2$. This is a continuous transition: if $\beta \searrow \beta_c$ the chain center of mass $\tilde{x}_0$ moves continuously to infinity.

The free energy is stationary at $v_{\max} \geq 1/2$, both for neutral and non neutral chains, and in the delocalized phase $v_{\max} = 1/2$. As noticed previously, since $v = \lambda/\lambda_0 \neq 0$, this means that replicas are interacting in a sensible way, even in the delocalized phase. This is in contrast with the Hartree approximation, which assumes instead no interaction between different replicas. Such approximation has been used in a new variational approach based on a non-stationary Green function, which results in an asymmetric ground state even in the case of a neutral chain [51].

6.1.4 Replica symmetry breaking

In order to ascertain the stability of the replica-symmetric solution, we have computed the quenched free energy within the so-called one step RSB ansatz. We divide the n replicas in n/m groups, each containing m replicas, in such a way that $\lambda_{ab} = \lambda_1$ if replicas a and b belong to the same group, and $\lambda_{ab} = \lambda_2$ otherwise. Defining $v = \lambda_1/\lambda_0$ and $w = \lambda_2/\lambda_0$, the quenched free energy reads:

$$\begin{aligned} \beta f_q &= -\frac{1}{2}\left\{ \ln\left(\frac{2\pi l^2}{d}\right) + \frac{1}{m}h\left[B\right] - \frac{1-m}{m}h\left[A\right] + \right. \\ &+ \frac{2d}{l^2}\lambda_0 w h'\left[B\right] + 2\beta\zeta_0 \operatorname{erf}\left(\frac{\tilde{x}_0}{2}\right) + \\ &+ \left. (\beta\Delta)^2\left[1-(1-m)\,\epsilon\left(v,\tilde{x}_0\right) - m\epsilon\left(w,\tilde{x}_0\right)\right]\right\}, \end{aligned} \tag{6.11}$$

where $A = 2d\lambda_0\left(1-v\right)/l^2$ and $B = 2d\lambda_0\left[m\left(1-w\right) + \left(1-m\right)\left(1-v\right)\right]/l^2$. This free energy has to be maximized with respect to v, w, m, and minimized with respect to $\lambda_0, \tilde{x}_0$. Performing the change of variables $(v,w) \to (s,t)$, with $s = v - w$ and $t = (1-m)\,v + mw$, the equation $\partial f_q/\partial s = 0$ has always the solution $s = 0$, which is the replica-symmetric solution, whereas $\partial f_q/\partial t \mid_{s=0} = 0$ gives equation (6.9). The stability of this solution depends on whether it is actually a maximum of the free energy, i.e. depends on the sign of

$$\begin{aligned} \left.\frac{\partial^2 f_q}{\partial s^2}\right|_{s=0} &= \frac{m\left(1-m\right)}{2\beta}\left\{\left(\frac{2d}{l^2}\lambda_0\right)^2 h''\left[\frac{2d}{l^2}\lambda_0\left(1-t\right)\right] + \right. \\ &+ \left. (\beta\Delta)^2\ \epsilon''\left(t,\tilde{x}_0\right)\right\}. \end{aligned} \tag{6.12}$$

Notice that $h''\left[A\right] < 0$ (the 'entropy' contribution) and $\epsilon''\left(t,\tilde{x}_0\right) = \partial^2\epsilon/\partial t^2 > 0$ (the energy contribution). At high temperature entropy is dominant, $\partial^2 f_q/\partial s^2|_{s=0} < 0$, and $s = 0$ is in fact a maximum. But at low temperature the energy contribution prevails, $\partial^2 f_q/\partial s^2|_{s=0} > 0$ and $s = 0$ becomes a minimum, signalling that the replica-symmetric solution is unstable. Therefore, within the Gaussian variational approach, we find a glass transition to a phase characterized by the presence of many metastable states separated by high energy barriers, according to the usual interpretation of RSB [63]. If the transition takes place in the way sketched before, i.e. if it is continuous, the transition temperature is given by the equation $\partial^2 f_q/\partial s^2|_{s=0} = 0$. In principle, however, it could be possible to have a discontinuous transition if another local maximum is present which provides a better extremum for the free energy when the replica-symmetric solution is still stable.

In the case of a neutral chain ($\zeta_0 = 0, \tilde{x}_0 = 0$), we can easily compute the temperature β_t at which the replica-symmetric solution becomes unstable. The solution of equations $\partial^2 f_q/\partial s^2|_{s=0} = 0$ and $\partial f_q/\partial t \;|_{s=0} = 0$ yields $v_t = 1/\sqrt{2}$, $\beta_t^2 = \pi\sqrt{2+5/\sqrt{2}}/\Delta^2$, and $\lambda_{0,t} = (l^2/2d)\sqrt{1+\sqrt{2}}$ for the square of the characteristic length at the transition.

The occurrence of RSB in such a simple model, i.e. in a model without any self-interaction between different monomers, is higly surprising and counter-intuitive. In fact, the energy term $-\beta\sum_{i=0}^{N}\zeta_i \;\mathrm{sgn}\,(x_i)$ alone does not provide any frustration. Each monomer simply chooses its preferred side and frustation arises from the competiton between energy and chain connectivity, which is an entropic effect. Thus, at sufficiently low temperature we would expect no frustration at all.

6.2 Exact bounds on the free energy

In this section we study a lattice discretized version of the model introduced by Garel *et al.* The nodes of an N links chain occupy the sites $\vec{r}_i = (x_{i1}, \ldots, x_{id})$, $i = 0, \ldots, N$ of a d-dimensional hypercubic lattice. A flat interface passing through the origin and perpendicular to the $\vec{u} = (1, \ldots, 1)$ direction separates a polar (e.g. water), on the $\vec{u}\cdot\vec{r} > 0$ side, from a nonpolar (e.g. oil or air), on the $\vec{u}\cdot\vec{r} < 0$ side, solvent. The i-th monomer interacts with the solvent through its hydrophilic charge $q_i > 0$ ($q_i < 0$ means that it is hydrophobic) and contributes to the energy with a term $-q_i\mathrm{sgn}(\vec{u}\cdot\vec{r}_i)$ [2]. For simplicity we can associate the charge to the link between adjacent positions on the chain instead that to the single monomer.

[2] The case where the energy is $-\lambda_+ q_i$ and $\lambda_- q_i$ ($\lambda_{+,-} > 0$) when $\vec{u}\cdot\vec{r}_i > 0$ and $\vec{u}\cdot\vec{r}_i < 0$ respectively is readily reduced to the one treated here apart from an additive constant and a redefinition of the 'charges' $q_i' = \frac{\lambda_+ + \lambda_-}{2} q_i$.

The partition function of the model for a chain W starting at position $\vec{r}$ is

$$\mathcal{Z}(\vec{r},\{q_i\}) = \sum_{W:\vec{r}\to .} \exp\left\{\beta \sum_{i=1}^{N} q_i \mathrm{sgn}(\vec{u}\cdot\vec{r}_i)\right\}, \tag{6.13}$$

where $\beta^{-1} = k_B T$. If one sums over non-interacting (ideal) chains, then the lattice version of the model introduced in Ref. [46] is recovered. We will consider also the more physical case where steric interaction among monomers does not allow for multiple occupancy of lattice nodes, studied for a particular case in Ref. [48].

The free energy density of the system reads $f(\vec{r},\beta) = -\frac{1}{\beta N}\overline{\ln \mathcal{Z}(\vec{r},\{q_i\})}$, where $\overline{\cdots}$ denotes the quenched average over the distribution of the charges $\{q_i\}$. In the following we will assume that $\{q_i\}$ are independent random variables having a Gaussian distribution of the form

$$P(q_i) = \frac{1}{\sqrt{2\pi\Delta^2}} \exp\left[-\frac{(q_i - q_0)^2}{2\Delta^2}\right]. \tag{6.14}$$

More general cases will be treated at the end. In particular considering charges not independently distributed is of interest for *designed* sequences as happens for real proteins.

We will show that for a neutral chain ($q_0 = 0$) $f(0,\beta) < f(\vec{r},\beta)$ with $|\vec{r}| \geq N$, in the large N limit, and for all β. The same holds also for $q_0 \neq 0$ if $\beta > \beta_{upper}(q_0,\Delta)$ with $\beta_{upper} \to 0$ if $\frac{q_0}{\Delta} \to 0$. This implies that the chain is localized around the interface at any temperature if $q_0 = 0$, and at sufficently low temperature if $q_0 \neq 0$. The proof is rigorous for the ideal chain, whereas for the self-avoiding case only a mild and well accepted hypothesis on the asymptotic behavior of the entropy is needed. When $q_0 \neq 0$ a rigorous lower bound on the free energy for both the ideal and the self-avoiding chain allows to determine a $\beta_{lower}(q_0,\Delta)$ below which the chain is delocalized.

We will first consider the ideal chain case and then explain the modifications necessary to extend the results to self-avoiding chains.

6.2.1 Ideal chain

For clarity we derive the bounds in the $d = 1$ case. The general case does not contain any further difficulty[3]. Let us first consider initial positions far from the interface in

[3] Notice that in the off-lattice model of Ref. [46] the entropy has only a harmonic term (Edwards functional) and thus the d-dimensional case reduces trivially to the $d = 1$ case. In the lattice model

the favorable solvent, $x \geq N$ if $q_0 > 0$ or $x \leq -N$ if $q_0 < 0$. Under these assumptions all chains remain in the same side, implying that $\mathrm{sgn}\,(x_i) = 1$ (or $\mathrm{sgn}\,(x_i) = -1$ respectively) for all i. Upon averaging over the charge distribution, we obtain the free energy density of a walk in the favorable solvent:

$$f^* = -\frac{1}{\beta} \ln 2 - |q_0|. \tag{6.15}$$

We give an upper bound to the free energy as follows. Consider only chains made up of blobs of k steps, with k even. Bringing a blob in its globally favored side leads to an energy contribution of the form $H_{blob\,j} = -\left|\sum_{i=1}^{k} q_{k(j-1)+i}\right|$, so that

$$\mathcal{Z}(x = 0, \{q_i\}) \geq (C_k)^{\frac{N}{k}} \exp \left\{ \beta \sum_{j=1}^{\frac{N}{k}} \left| \sum_{i=1}^{k} q_{k(j-1)+i} \right| \right\}, \tag{6.16}$$

where C_k is the number of chains starting and ending in the origin and remaining in the same side. In the one-dimensional case it is easy to exactly determine C_k. It turns out $C_k = \dfrac{1}{\frac{k}{2}+1} \begin{pmatrix} k \\ \frac{k}{2} \end{pmatrix}$, so that, by using the Stirling's formula, the asymptotic result $C_k \sim 2^k k^{-\frac{3}{2}}$ is found (in d dimensions $C_k \sim (2d)^k k^{-\frac{d+2}{2}}$ [131]). The upper bound on the free energy is then:

$$f(0, \beta) \leq -\frac{1}{\beta} \frac{\ln C_k}{k} - \frac{1}{k} \overline{\left| \sum_{i=1}^{k} q_i \right|}, \tag{6.17}$$

and, by using Eq. (6.14), we obtain:

$$\begin{aligned} \Delta f &= f(0, \beta) - f^* \leq h_{q_0}(k, \beta) \equiv \\ &\equiv \frac{1}{\beta} \left[\ln 2 - \frac{\ln C_k}{k} \right] - |q_0| G \left(\frac{\sqrt{k}|q_0|}{\sqrt{2}\Delta} \right), \end{aligned} \tag{6.18}$$

where the scaling function G is given by

$$G(x) = \frac{1}{\sqrt{\pi}} \frac{1}{x} \exp\left(-x^2\right) - [1 - \mathrm{erf}(x)]\,. \tag{6.19}$$

$G(x)$ is a positive decreasing monotonic function for positive arguments.

considered here this is no more true.

We consider separately the neutral and the $q_0 \neq 0$ cases. In the neutral case it turns out that the chain is always localized at the interface. In fact, if $q_0 = 0$ the last two terms in Eq.(6.18) vanish and we are left with

$$\Delta f \leq h_0(k,\beta) = \frac{1}{\beta}\left[\ln 2 - \frac{\ln C_k}{k}\right] - \sqrt{\frac{2}{\pi}}\Delta\frac{1}{\sqrt{k}}. \tag{6.20}$$

It is easy to see that for any β there exists a value $k(\beta)$ such that $h_0(k,\beta) < 0$ for $k > k(\beta)$. At high temperature $k(\beta) \sim [\ln(\beta\Delta)]^2 (\beta\Delta)^{-2}$. This shows that at any temperature a neutral random chain is always adsorbed by the interface.

In the non-neutral case with $k = 2$, one has $\Delta f < 0$ if

$$\beta > \beta_{upper} = \frac{\ln 2}{q_0 G(\frac{q_0}{\Delta})}. \tag{6.21}$$

The limit $\lim_{k\to\infty} h_{q_0}(k,\beta) = 0_+$ does not allow to deduce the existence of a negative minimum in k, so that the previous argument, showing that the neutral chain is always localized, does not hold for $q_0 \neq 0$. Equation (6.21) proves localization at sufficiently low temperatures. For $q_0 \ll \Delta$, it yields $\beta_{upper} = \frac{\sqrt{\pi}\ln 2}{\Delta}$, and in the opposite regime $\Delta \ll q_0$, $\beta_{upper} = 2\sqrt{\pi}\ln 2 \exp\left(\frac{q_0^2}{\Delta^2}\right)\frac{q_0^2}{\Delta^3}$. In the limit $|q_0|/\Delta \ll 1$ we can give a better estimate for β_{upper}, such that $\beta_{upper} \to 0$ as $q_0 \to 0$, by considering a larger blob size $k = 2x_0^2\left(\Delta/|q_0|\right)^2$, where $x_0 \gg |q_0|/\Delta$ is fixed:

$$\beta_{upper} = \frac{3}{2x_0^2 G(x_0)} \ln\left(\sqrt{2}x_0\Delta/|q_0|\right)\frac{|q_0|}{\Delta^2}. \tag{6.22}$$

We now look for a lower bound on f for all chain initial positions. If, for example, $q_0 > 0$, the preferred side is the right one. Consider then the starting point at $x = N - k$ ($0 < k \leq N$) and let g_E, with E some subset of the last k steps of the walk ($0 < |E| \leq k$ with $|E|$ the number of elements in E), be the number of walks having the steps belonging to E in the unfavourable side. Upon defining $G_k = \sum_E g_E$ ($G_k < 2^k$) one has

$$\begin{aligned}\mathcal{Z}(x = N-k, \{q_i\}) &= \left(2^N - G_k\right)\exp\left[\beta\sum_{i=1}^{N} q_i\right] + \\ &+ \exp\left[\beta\sum_{i=1}^{N} q_i\right]\sum_E g_E \exp\left[-2\beta\sum_{i\in E} q_i\right], \end{aligned}\tag{6.23}$$

so that, by using the inequality $\ln \overline{x} \geq \overline{\ln x}$ and averaging over the charge distribution, one obtains

$$f(N-k) \geq f^* - \frac{1}{\beta N} \ln \left[1 + \sum_E g_E \frac{a(\Delta, q_o, \beta)^{|E|} - 1}{2^N} \right], \tag{6.24}$$

with $a(\Delta, q_o, \beta) = \exp\left[2\Delta^2\beta^2 - 2\beta q_0\right]$. This equation and its analogous in the $q_0 < 0$ case show that there is a delocalization temperature

$$\beta_{lower} = \frac{|q_0|}{\Delta^2} \tag{6.25}$$

such that, if $\beta < \beta_{lower}$, $f(|N-k|) \geq f^*$ and the chain delocalizes. This argument can be extended to the d-dimensional case, in which $1 \leq G_k < (2d)^k$, $\mathcal{Z}_d^*(q_i) = (2d)^N \exp\left(\beta \sum_{=1}^N q_i\right)$ and $f_d^* = -\frac{1}{\beta} \ln 2d - |q_0|$.

The bounds we have proved above allow to conclude that there is a critical value β_c such that for values of β smaller than β_c the chain is delocalized in the favorable solvent, while for larger values it is adsorbed by the interface, with the estimates $\beta_{lower} < \beta_c < \beta_{upper}$ (it is easy to verify that $\beta_{lower} < \beta_{upper}$). The lower bound (6.25) and the upper bound (6.22), in the limit $|q_0|/\Delta \ll 1$, show the same behavior found by using both an Imry-Ma type argument [46] and variational approaches [51, 58].

6.2.2 Self-avoiding chain

All the results shown for a random chain can be readily generalized for a self-avoiding chain. Namely, a neutral chain is localized at all temperatures, whereas a non-neutral chain undergoes a localization transition at some critical temperature β_c.

The delocalization temperature β_{lower} can be derived exactly in the same way, since the division of walks into classes according to the number of steps made in the unfavorable solvent does not depend on the self-avoidance constraint.

The upper bounds on the free energy, which allow to prove chain localization, requires instead some refinements with respect to the previous case. While the energy term is computed in the same way as before, the entropy term is different. Firstly, the connective constant ($\kappa = 2d$ for a random walk in d dimensions) is different. We recall that the existence of the connective constant, $\kappa = \lim_{N\to\infty} \ln S_N / N$, for self-avoiding walks (SAW) has been rigorously established [132] (S_N is the total number of N-steps SAW starting from the same site). The subleading correction of the form

$S_N \simeq \kappa^N N^{\gamma-1}$ is widely agreed upon, although not rigorously proved. Secondly we introduce the notion of *loop*, following e.g. [133], and consider only walks made up of N/k blobs, each blob being a k-loop, in such a way that different blobs can be embedded independently, as well as for a random chain. A N-loop is a N-steps SAW, starting and ending on the interface, which always remains in the same half-space, with the further condition $x_{01} \leq x_{i1} < x_{N1}\ \forall i$. It has been proved ([134]) that the free energy density of loops is the same as for SAW, $\kappa_l \equiv \lim_{N\to\infty} \ln L_N / N = \kappa$, where L_N is the number of N-loops. The subleading correction is usually assumed in the same form as for the number of SAW:

$$L_N \simeq \kappa^N N^{\gamma_s - 1}\ . \tag{6.26}$$

These considerations are sufficient to generalize the previous results to the self-avoiding case, yielding the following bounds for the critical temperature:

$$\frac{|q_0|}{\Delta^2} \leq \beta_c \leq \frac{\ln \kappa}{|q_0| G(\frac{\sqrt{2}|q_0|}{\Delta})}\ , \tag{6.27}$$

which do not depend on the assumption (6.26) and is therefore rigorous. Again, in the limit $|q_0|/\Delta \ll 1$ a better estimate β_{upper} can be derived by using Eq. (6.26):

$$\beta_{upper} = \frac{1-\gamma_s}{x_0^2 G(x_0)} \ln\left(\sqrt{2} x_0 \Delta / |q_0|\right) \frac{|q_0|}{\Delta^2}. \tag{6.28}$$

6.2.3 Generic probability distribution

Up to now we have considered the hydrophobic charges as independently distributed Gaussian random variables. Actually, the results we have proved do not depend on this assumption. We will briefly sketch this in a few cases [135].

The argument showing localization at any temperature for a neutral chain holds true, both for random and self-avoiding chains, if $\overline{\left|\sum_{i=1}^k q_i\right|} \simeq \sqrt{k}$ as $k \to \infty$. The central limit theorem ensures this for independent random variables having a generic probability distribution with finite variance and null mean. In the non-neutral case, the existence of a delocalization transition can be proved e.g. for a bimodal distribution. This corresponds to the more realistic case of two kinds of monomers, one hydrophilic and the other hydrophobic. We thus consider the generic bimodal distribution[4]:

$$P\left(q_i\right) = \alpha\delta\left(q_i - q_+\right) + (1-\alpha)\,\delta\left(q_i + q_-\right) \ , \tag{6.29}$$

with $q_+, q_- > 0$. The probability distribution (6.29) has three independent parameters, and fixing the average charge $q_0 = \alpha\left(q_+ + q_-\right) - q_-$ and the variance $\Delta = \sqrt{\alpha\left(1-\alpha\right)}\left(q_+ + q_-\right)$ we are left with one free parameter. It is interesting to report the delocalization temperature β_{lower}, which provides a good estimate for the critical temperature in the previous cases:

$$\beta^{bim}_{lower} = \frac{\sqrt{\alpha\left(1-\alpha\right)}}{2\Delta}\ln\left[1+\frac{q_0}{\sqrt{\alpha\left(1-\alpha\right)}\Delta - \left(1-\alpha\right)q_0}\right] \ . \tag{6.30}$$

Notice that in the limit of nearly neutral chain ($q_0 \ll \Delta$) we get $\beta^d_c \simeq \frac{q_0}{2\Delta^2}$, which is the same function of q_0 and Δ as in the Gaussian case, suggesting the existence of a universal behavior. In the limit of nearly homogeneous chain ($\Delta \ll q_0$ which implies $\alpha \simeq 0$ or $\alpha \simeq 1$) instead $\beta^{bim}_{lower} = \frac{1}{2(q_+ + q_-)}\ln\left[\frac{\alpha}{1-\alpha}\frac{q_+}{q_-}\right]$ diverges logarithmically in contrast with the Gaussian case.

We consider now the case in which the hydrophobic charges $\{q_i\}$ are not independent random variables, but are Gaussianly distributed with $\overline{q_i} = q_0 \ \forall i$, $\overline{q_i q_j} - \overline{q_i}\ \overline{q_j} = M^{-1}_{ij}$. We assume $M^{-1}_{ii} = \Delta^2 \ \forall i$, in analogy with the non-correlated case, and also translational invariance along the chain for the correlation matrix: $M_{ij} = b(|i-j|)$. One can prove that if long range correlations decay exponentially or even algebraically the neutral chain is again localized at all temperatures. In fact, by assuming an algebraic decay, $b\left(r\right) \simeq r^{-\eta}$, it turns out that $\overline{\left|\sum_{i=1}^k q_i\right|} \simeq k^{\delta/2}$ with $\delta = \min\left(\eta, 1\right)$. Only if correlations are so strong that they do not vanish along the chain ($\eta = 0$), the chain does not localize at all temperatures.

In the non neutral case the existence of the transition can be proved also in this case. For example the estimate of the delocalization temperature is

$$\beta^{corr}_{lower} = \min_E\left\{\frac{kq_0}{\sum_{i,j\in E} M^{-1}_{ij}}\right\} \ . \tag{6.31}$$

If the charges are positively correlated ($M^{-1}_{ij} > 0$ for $i \neq j$), chain localization is more favored than in the non-correlated case, whereas if charges are anti-correlated ($M^{-1}_{ij} < 0$ for $i \neq j$) it is less favored.

[4] Some previous exact results for the bimodal distribution and only for ideal chains have been already obtained [136, 137].

6.2.4 Asymmetric interface potentials

Finally we extend our demonstrations to the random AB-copolymers studied by Sommer *et al.* [48]. Their model corresponds to consider the following Hamiltonian:

$$\mathcal{H} = \sum_i |q_i|[\lambda\theta(q_i)\theta(-\vec{u}\cdot\vec{r}_i) + \theta(-q_i)\theta(\vec{u}\cdot\vec{r}_i)], \qquad (6.32)$$

with the charges distributed according to Eq. (6.29) with $\alpha = 1/2$ and $q_+ = q_-$ ($|\lambda - 1|$ measures the potential asymmetry). Such an AB-copolymer (Eq. (6.32)) is equivalent to a non-neutral chain in symmetric potentials ($\lambda = 1$), a case that we have already discussed. We have proved the existence of a delocalization transition for a neutral chain also in the Gaussian case. For both distributions, the delocalization temperature shows the behavior $\beta_{lower} \simeq \frac{|\lambda-1|}{\Delta}$ in the limit of nearly symmetric potentials ($\lambda \simeq 1$), in agreement with the scaling law and the numerical results found in [48]. On the contrary, in the highly asymmetric cases (small and large λ) different asymptotic behaviors for β_{lower} occur [135].

Conclusions and perspectives

The major part of this thesis has been devoted to the investigation of which polymer chain conformations are selected in the continuum three-dimensional space, when the property of *best packing* is required. We have performed numerical simulations of a chain of discrete beads, for chain lengths up to a few tens of beads, within the simulated annealing procedure. We have searched for the chain conformation maximizing the *thickness*, that is the radius of the largest impenetrable tube which can be inflated uniformly around the chain [108, 32, 33].

Our main result is the selection, by means of this simple geometrical variational principle, of two higly regular and ordered geometrical motifs, that closely resemble α-helices and β-sheets, the secondary patterns appearing in the structures of natural proteins [36]. This is the more notable, in that we have stripped away all possible details from our model, that is detailed chemistry, energetic interactions, the presence of the side chains, the specificity of different residues, and the role of the solvent. Note also that the emergence of structures in the continuum space, characterized by a high degree of geometrical order, is definitely not an artifact, as could be instead the case in lattice models [104, 95].

Optimal best-packing conformations varies depending on which way the *compactness* of the chain is enforced. We have addressed the issue in two different ways, by using two different parameters driving the compactness of the chain; the gyration radius of the chain and the number of close contacts between different non-consecutive beads [102, 35].

The gyration radius constraint is purely geometrical in nature, and allows the limit of *continuous chain* to be approached in a natural way. The ‘number of contacts’ constraint is more physical, and can be seen as an energy constraint involving a *two-body attractive* interaction underlying the collapse of the chain. Thickness maximization is instead formally equivalent to the minimization of a *three-body repulsive* interaction energy [33]. Three-body effects are indeed necessary in order to introduce the notion of thickness for a continuous string, but two-body interactions between different beads only make sense for a *discrete* chain.

When using the gyration radius constraint, we have found repeatedly in our simulations that the *local* solution to the best packing problem for strings is a perfect helical shape. Selected helices have a particular value of the pitch-radius ratio, reflecting a peculiar degeneracy between local and non-local geometrical effects. The same degeneracy is observed to hold, within a few percent, in two different types of helical motifs appearing in natural proteins, the α-helix and the collagen triple helix [36].

The requirement of *global* compactness with the same constraint also leads to the emergence of helical optimal shapes in an intermediate regime. *Saddle*-like optimal conformations take over with increasing compactness, but are in general very close competitors against helices. Saddles can be considered as topological defects connecting different helical portions with opposite chirality. They are indeed observed, in the case of helical bacteria flagella that flip their chirality, in order to reverse the direction of motion [138].

When compactness is increased furtherly, optimal shapes are very entangled and do not show any apparent regular pattern. We have discussed a possible way of extracting a globally compact *regular* best packing conformation. Simulations are hard, but preliminary results indicate that such a solution could be a hierarchical structure composed by a long helix wound to form a super-helix.

Interestingly, nature seems to have adopted the same recipe, in order to solve the probably most spectacular problem of packing, the tight arrangement of chromosomal DNA in the cell nucleus. Chromosomes are very long molecules encoding genetic information of the whole organism. The length of chromosomal DNA molecules, when uncoiled, ranges in the order of cm, while the size of the cell nucleus is typically of some μm. The packaging of DNA in chromosomes is a very complex phenomenon, involving the presence of binding proteins, such as histones. The basic structure of this tight packaging is however believed to be a hierarchy of helices at different successive levels. At the lower level stands the double stranded DNA helical structure, which is then wrapped around histones to form the nuclesomes. Nucleosomes are the unit particles of chromatin, which is believed to be in its turn helically arranged within the so-called solenoidal model of the 30-nm fiber [139, 140, 141]. Whether the structural organization is still helical at higher level is not known, but there is some evidence supporting this conclusion [86].

When using the ‘number of contacts’ constraint, we find that optimal shapes are planar hairpin-like conformations if a *low* number of contacts is required, and again helices if the number of required contacts is *higher*. The ordered planar geometry of optimal hairpins closely resemble the β-sheets appearing in natural proteins. Moreover, we find *two* different kinds of hairpin structures, degenerate from the point of view of best packing. They reproduce the same patterns formed by hydrogen bonds between adjacent β-strands of a parallel β-sheet, in one case, and of an antiparallel β-sheet, in the other case.

Helices are also found as the ground state structures of a total interaction energy term being the sum of a usual pairwise Lennard-Jones contribution and a three-body repulsive one, which mimicks the condition of thickness maximization. We also have found a low energy structure having the shape of a double helix. In this case, as in all previous ones, the occurrence of helices is not driven by any preference for chirality with a definite sign, and we indeed find optimal helices with opposite handedness.

We deem our findings interesting by themselves, in particular the fact that helices emerge in a natural way, as minimal energy conformations of a three-body interaction, as we emphasize again, having no preference for a definite chirality. Note that we have verified the emergence of helical optimal shapes in practically all the different ways in which we have approached the best packing problem. Hairpin structures instead are found only when a low number of contacts is required, and do not seem to be extendable in the limit of continuous string.

Whether the best packing variational principle captures or not some of the features of the real evolutionary optimization process which has selected natural proteins, is of course a quite different issue. There are some arguments hinting that this could be the case. Optimal packing seems intuitively related to the property of large geometrical accessibility which has been shown to characterize the folds of natural proteins [29]. This in turn is, again intuitively, connected to the property of fast folding. Interestingly, the requirement of fast folding leads to the selection of helical conformations, too [30].

Of course further investigation is needed, in the direction of both increasing the chain length accessible to our simulations and generalizing the best packing principle to more refined models, for example with the presence of the side chains. It is anyhow remarkable that simple structures as helices and hairpins, which has been selected by natural evolution, can be also characterized as being extremal in a geometrical sense. Note that the DNA double helix structure has also been suggested to involve optimally packed tubes [142].

In the second part of this thesis, we have studied the localization of a hydrophobic-hydrophilic chain at a flat selective interface, separating a polar and a non-polar solvent [58, 59].

Within the Gaussian variational method and a replica-symmetric ansatz, an explicit calculation of the relevant physical quantities has been performed, in the case of an ideal random chain [58]. At high temperature the results agree with the predictions based on Imry-Ma argument [46, 55], whereas at low temperature a non trivial dependence on the microscopic bond length is present. Furthermore, the replica-symmetric solution has been shown to be unstable within a one step replica-symmetry breaking (RSB) scheme [63]. This would imply a breaking of the ergodicity of the system, in contrast with simple intuitive arguments. The Gaussian variational approach is believed to be effective in describing correctly the physics of disordered frustrated systems [66, 90, 129, 91]. We deem therefore an interesting issue to try to evidentiate RSB with Monte Carlo simulations, or to understand why the Gaussian variational method fails in predicting it, if this should be the case. Interestingly, RSB has been reported to occur also when treating the quenched disorder within a dynamical approach.

We also have proved exactly the occurrence of the localization transition in a related lattice model [59]. A neutral chain is localized at all temperatures, whereas a non-neutral chain delocalizes at a finite temperature. The result is quite general and holds for ideal and self-avoiding chains, Gaussian and bimodal distribution with independent and correlated charges. Furthermore, our lower bounds for the transition temperature confirm previous estimates.

Different generalizations of this simple model are possible, in the spirit of constructing a heteropolymer model describing membrane proteins. Firstly, one could consider a more complicate but realistic geometry, with a finite-width slab of non-polar ‘lipidic’ environment separating two polar semi-spaces. Secondly, explicit self-interactions between different monomers could be considered.

Appendix A

Monte Carlo moves

The configuration space we want to sample is the set $\mathcal{R}_n$ of discrete chains $\{\vec{x}_0, \dots, \vec{x}_n\}$ in the continuum three-dimensional space, having a fixed constant distance a between consecutive beads, a hard-core repulsion at a distance R_s between different non-consecutive beads, and with some further compactness constraint as defined in section 2.4. Chain configurations can be expressed equivalently in terms of the angles $\theta_2, \theta_3\phi_3, \dots, \theta_n, \phi_n$, with

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

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

where $\vec{u}_i = \vec{x}_i - \vec{x}_{i-1}$, and $\vec{v}_i = \vec{u}_i \wedge \vec{u}_{i-1}$. 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 1.1). The above correspondence between angles and coordinates becomes one-to-one, when fixing $\vec{x}_0 = (0,0,0)$, $\vec{x}_1 = (0,0,a)$, $\vec{x}_2 \cdot \hat{e}_j = 0$, 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 ($\hat{e}_j = (0,1,0)$).

We have employed four different moves in our Monte Carlo dynamics, 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

a 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 [123, 124].

• *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 $\vec{x}_j - \vec{x}_i$. 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.

• *Multi-pivot move*
This move works directly in angle representation. It can be seen as a composition of many different simple pivot moves performed in succession. We simply update *each* of the relevant $(n-1)$ $(n-2)$ angles with a probability ρ_a. Angles are updated in the same way as in the pivot move, with different rules for θ and ϕ angles. Multi-pivot moves should rearrange drastically the whole shape of the chain, thus ensuring a more efficient search in configuration space.

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, ρ_a, $\Delta\phi_m$, $\Delta\theta_m$, 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.

Appendix B

Optimal Helices

We now derive the particular value c^* of the pitch/radius ratio c of an optimal helix, that is a helix having the local radius of curvature equal to the non-local radius.

The parametric equation of a helix is

$$\vec{x}(t) = (\cos t, \sin t, vt) ; \tag{B.1}$$

where the pitch/radius ratio is $c = 2\pi v$. The tangent and the acceleration vectors are:

$$\dot{\vec{x}}(t) = (-\sin t, \cos t, v) \ ; \quad \ddot{\vec{x}}(t) = (-\cos t, -\sin t, 0) \ . \tag{B.2}$$

Since $\dot{\vec{x}}(t) \cdot \ddot{\vec{x}}(t) = 0$, the local radius of curvature is simply given by

$$\rho_L(t) = \frac{\left|\dot{\vec{x}}(t)\right|^2}{\left|\ddot{\vec{x}}(t)\right|} = 1 + v^2 \tag{B.3}$$

for all points.

The non-local radius for a continuous curve may be defined in the following way. Fix a point $\vec{A} = \vec{x}(t)$ on the curve, and compute the distance $d(s,t) = \left|\vec{B} - \vec{A}\right|$ from a second point $\vec{B} = \vec{x}(s)$ moving along the curve as a function of s. The non-local radius is then

$$\rho_{NL}(t) = \frac{1}{2} \min_{s \neq t} \{d(s,t)\} \ , \tag{B.4}$$

where it is required that $\frac{\partial d(s,t)}{\partial s} = 0$ at some $s^* \neq t$. Note that the non-local radius needs not to exist and is in principle a varying function of t, when the curve is not invariant for traslations along it.

The helix is invariant for traslations along the curve, and we can choose $t = 0$ so that

$$d^2\left[s\right] = 2\left(1 - \cos s\right) + v^2 s^2 \ . \tag{B.5}$$

The extremality condition is

$$\sin s + s v^2 = 0 \ ; \tag{B.6}$$

equation (B.6) has always the solution $s = 0$, which we are not interested in, and has no other solution for sufficiently high v. If v is decreased, more and more new solutions appear, two at a time, the smaller a maximum and the greater a minimum, corresponding to the increasing packing of helix turns. We are interested in the minimum s^* corresponding to $\vec{A}$ and $\vec{B}$ staying on two consecutive turns, that is $\pi < s^* < 2\pi$. For sufficiently low v, equation (B.6) then defines the implicit function $s^*\left(v\right)$, and one has

$$\rho_{NL}\left(t\right) = \frac{1}{2} d\left[s^*\left(v\right)\right] \tag{B.7}$$

for all point of the helix. In the limit $v \ll 1$, one has $s^* \simeq 2\pi$ and thus $\rho_{NL} \simeq \pi v = \frac{c}{2}$, where $c = 2\pi v$ is the pitch of the helix.

The particular value $v = v^*$, for which the local and the non-local radius of curvature are equal, is then defined by

$$\frac{1}{2} d\left[s^*\left(v^*\right)\right] = 1 + \left(v^*\right)^2 \ . \tag{B.8}$$

If $v > v^*$, the local radius is smaller than the non-local radius and a tube swelling around the helix would stop increasing due to local singularities. If $v < v^*$, the non-local radius is smaller than the local radius, and the tube would stop swelling due to self-intersection between different turns. At $v = v^*$ the two effects occur at the same time, and the corresponding pitch/radius ratio is $c^* = 2\pi v^* = 2.512$.

Acknowledgments

I would like to warmly tkank Amos, firstly for giving me the opportunity of working on very interesting and fascinating topics. Secondly, but not less important, I have really enjoyed his friendly attitude and his enthusiasm in ever proposing new ideas and interests. I hope I will be able to work with him again in the future.

Flavio deserves a special thank, since I still owe him (due to lack of time!) adequate acknowledgments for his fundamental support during my degree thesis. He has always been an importante reference for me also these years.

During my Ph.D., I have had the pleasure to collaborate with Cristian, Davide, Maria Pia, Jort, and Jayanth. I am grateful to all of them for their patience in bearing me. Each of them taught me something important.

I also enjoyed interesting discussions with both the ‘old’ people, Stefano, Cecilia, Claudio, and the ‘new entries’ Andrea, Gianni, Gianluca, Fabio, and Sandro, in Amos’s group. They all contributed to my nice stay at SISSA. I especially thank Cecilia for having tried to tech me swimming! I also acknowledge interesting and useful discussions with Delos, Enzo, and Attilio.

I would like to thank all people in the CM sector and in SISSA, for creating a unique atmosphere, and above all the efficiency and kindness of the secretaries.

I am particularly grateful to my family and to all of my friends for their continuous support along these years.

Some of them has been already cited, and I want to especially thank Federico, Alessandro, Sonia, Stella, Cristina, Mario, Paola, Ugo, and Giovanni. Without them, life would have been much more boring.

Special thanks go to all ‘raga’ in Rovigo, to my colleagues at the civil service, and to the eno-football team in Padova, without mentioning the glorious SISSA team. I have also appreciated the patience of Emma, Gianluca, and Mary in having tried their best to keep in touch with me, despite me.

Finally, I would like to remember Lando and Emanuele.

Bibliography

[1] A. E. Mirsky, and L. Pauling, *Proc. Natl. Acad. Sci. USA* **22**, 439 (1936).

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

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

[4] K. A. Dill, *Biochem.* **29**, 7133 (1990).

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

[6] J. Monod, *Chance and Necessity: an Essay on the Natural Philosophy of Modern Biology*, (Knopf, New York, 1971).

[7] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, *Rev. Mod. Phys.* **72**, 259 (2000).

[8] J. Bryngelson and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **84**, 7524 (1987).

[9] P. G. Wolynes, and W. A. Eaton, *Phis. World*, 39 (Sept. 1999).

[10] J. Bryngelson and P. G. Wolynes, *Biopolymers* **30**, 171 (1990).

[11] J. N. Onuchic, Z. Luthey-Shulten, and P. G. Wolynes, *Annual Rev. Physical Chem.* **48**, 545 (1997).

[12] R. L. Baldwin, and G. D. Rose, *Trends Biochem. Sci.* **24**, 26 (1999).

[13] S. E. Jackson, *Fold. Des.* **3**, R81 (1998).

[14] A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, (Freeman, New York, 1999).

[15] J. Bryngelson, J. N. Onuchic, J. N. Socci, and P. G. Wolynes, *Proteins: Struct. Funct. Gen.* **21**, 167 (1995).

[16] H. S. Chan, and K. A. Dill, *Proteins: Struct. Funct. Gen.* **30**, 2 (1998).

[17] C. M. Dobson, and M. Karplus, *Curr. Opin. Struct. Biol.* **9**, 92 (1999).

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

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

[20] C. Chothia, *Nature* **357**, 543 (1992).

[21] J. S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981).

[22] M. Levitt, and C. Chothia, *Nature* **261**, 552 (1976).

[23] C. Branden, and J. Tooze, *Introduction to Protein Structure* (Garland Publishing, Inc., New York & London, 1989).

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

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

[26] K. W. Plaxco, K. T. Simons, and D. Baker, *J. Mol. Biol.* **277**, 985 (1998).

[27] E. Alm, and D. Baker, *Curr. Opin. Struct. Biol.* **9**, 189 (1999).

[28] J. N. Onuchic, H. Nymeyer, A. E. Garcia, J. Chaine, and N. D. Socci, *Adv. Protein Chem.* **53**, 87 (2000).

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

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

[31] N. J. A. Sloane, *Nature* **395**, 435 (1998).

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

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

[34] S. Kirkpatrick, C. D. Gelatt, Jr. , and M. P. Vecchi, *Science* **220**, 671 (1983).

[35] F. M. Richards, and W. A. Lim, *Q. Rev. Biophys.* **26**, 423 (1994).

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

[37] T. Garel, H. Orland, and D. Thirumalai, "Analytical Theories of Protein Folding", in *New developments in theoretical studies of proteins*, edited by R. Elber (World Scientific, Singapore, 1996).

[38] C. D. Sfatos, and E. I. Shacknovich, *Phys. Rep.* **288**, 77 (1997).

[39] T. Garel, H. Orland, and E. Pitard, "Protein Folding and Heteropolymers", in *Spin Glasses and Random Fields*, edited by A. P. Young (World Scientific, Singapore, 1997).

[40] L. Leibler, *Macromolecules* **13**, 1602 (1980).

[41] F. S. Bates, and G. H. Fredrickson, *Annu. Rev. Phys. Chem.* **41**, 525 (1990).

[42] H. R. Brown, V. R. Deline and P. F. Green, *Nature* **341**, 221 (1989).

[43] C.-A. Dai, B. J. Dair, K. H. Dai, C. K. Ober, E. J. Kramer, C.-Y. Hui and L. W. Jelinsky, *Phys. Rev. Lett.* **73**, 2472 (1994).

[44] R. Bonaccini and F. Seno, *Phys. Rev.* **E60**, 7290 (1999).

[45] E. Orlandini, F. Seno, J. R. Banavar, A. Laio and A. Maritan *Deciphering the folding kinetics of transmembrane helical proteins*, Preprint University of Padova (2000).

[46] T. Garel, D. A. Huse, S. Leibler and H. Orland, *Europhys. Lett.* **8**, 9 (1989).

[47] C. Yeung, A. C. Balazs, and D. Jasnow, *Macromolecules* **25**, 1357 (1992).

[48] J.-U. Sommer, G. Peng and A. Blumen, *J. Chem. Phys.* **105**, 8376 (1996).

[49] G. Peng, J.-U. Sommer and A. Blumen, *Phys. Rev.* **E53**, 5509 (1996).

[50] C. Monthus, *Eur. Phys. J.* **B13**, 111 (2000).

[51] S. Stepanow, J.-U. Sommer and I. Y. Erukhimovich, *Phys. Rev. Lett.* **81**, 4412 (1998).

[52] X. Châtellier, and J.-F. Joanny, *Eur. Phys. J.* **E1**, 9 (2000).

[53] Z. Y. Chen, *J. Chem. Phys.* **112**, 8665 (2000).

[54] V. Ganesan, and H. Brenner, *Europhys. Lett.* **46**, 43 (1999).

[55] J.-U. Sommer and M. Daoud, *Europhys. Lett.* **32**, 407 (1995).

[56] J.-U. Sommer, G. Peng and A. Blumen, *J. Phys. II France* **6**, 1061 (1996).

[57] Z. Y. Chen, *J. Chem. Phys.* **111**, 5603 (1999).

[58] A. Trovato, and A. Maritan, *Europhys. Lett.* **46**, 301 (1999).

[59] A. Maritan, M. P. Riva, and A. Trovato, *J. Phys.* **A32**, L275 (1999).

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

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

[62] K. A. Dill, *Science* **250**, 297 (1990).

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

[64] E. I. Shakhnovich and A. M. Gutin, *Biophys. Chem.* **34**, 187 (1989).

[65] T. Garel and H. Orland, *Europhys. Lett.* **6**, 307 (1988).

[66] E. I. Shakhnovich and A. M. Gutin, *J. Phys.* **A22**, 1647 (1989).

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

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

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

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

[71] J. des Cloiseaux and G. Jannink, *Polymers in Solution: their Modelling and Structure*, (Clarendon Press, Oxford, 1990).

[72] P. G. de Gennes, *Scaling Concept in Polymer Physics*, (Cornell University Press, Ithaca, New York, 1979).

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

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

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

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

[77] P. L. Privalov, and S. J. Gill, *Adv. Protein Chem.* **39**, 191 (1989).

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

[79] 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).

[80] G. I. Makhatadze, and P. L. Privalov, *J. Mol. Biol.* **213**, 375 (1990).

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

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

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

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

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

[86] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. D. Watson, *Molecular Biology if the Cell*, (Garland Publishing, Inc., New York & London, 1989).

[87] G. von Heijne, *Prog. Biophys. Mol. Biol.* **66**, 113 (1996).

[88] B. Derrida, *J. Phys. A: Math. Gen.* **14**, L5 (1981).

[89] T. Garel, L. Leibler and H. Orland, *J. Phys. II France* **4**, 2139 (1994).

[90] M. Mézard and G. Parisi, *J. Phys. I France* **1**, 809 (1991).

[91] A. Engel, *Nucl. Phys* **B410**, 617 (1993).

[92] A. V. Finkelstein, and O. B. Ptitsyn, *Prog. Biophys. Mol. Biol.* **50**, 171 (1987).

[93] A. V. Finkelstein, A. M. Gutin, and A. Y. Badretdinov, *FEBS Lett.* **325**, 23 (1993).

[94] A. V. Finkelstein, A. M. Gutin, and A. Y. Badretdinov, *Subcell. Biochem.* **24**, 1 (1995).

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

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

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

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

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

[100] H. S. Chan, and K. A. Dill, *J. Chem. Phys.* **90**, 492 (1989).

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

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

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

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

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

[106] G. D. Rose, and J. P. Seltzer, *J. Mol. Biol.* **113**, 153 (1977).

[107] R. Aurora, T. P. Creamer, R. Srinivasan, and G. D. Rose, *J. Bio. Chem.* **272**, 1413 (1997).

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

[109] W. Barlow, *Nature* **29**, 186 (1883).

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

[111] C. A. Rogers, *Proc. Lond. Math. Soc.* **8**, 609 (1958).

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

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

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

[115] L. V. Woodcock, *Nature* **385**, 141 (1997).

[116] R. Car, *Nature* **385**, 115 (1997).

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

[118] A. Stasiak, V. Katritch, J. Bednar, D. Michoud, and J. Dubochet, *Nature* **384**, 122 (1996).

[119] G. Buck, *Nature* **392**, 238 (1998).

[120] J. Cantarella, R. B. Kusner, and J. M. Sullivan, *Nature* **392**, 237 (1998).

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

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

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

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

[125] C. Vanderzande, *Lattice Models of Polymers* (Cambridge University Press, Cambridge, 1998).

[126] P. Pieranski, private communication.

[127] G. Date, P. K. Ghosh, and M. V. N. Murthy, *Phys. Rev. Lett.* **81**, 3051 (1998).

[128] Y. Imry and S.-K. Ma, *Phys. Rev. Lett.* **35**, 1399 (1975).

[129] A. Moskalenko, Yu. A. Kuznetsov and K. A. Dawson, *Europhys. Lett.* **40** (2), 135 (1997).

[130] P. G. de Gennes, *Rep. Progr. Phys.* **32**, 187 (1969).

[131] B. D. Hughes, *Random Walks and Random Environments*, Vol. I, Clarendon Press, Oxford, 1995.

[132] J. M. Hammersley, *Proc. Camb. Phil. Soc.* **53**, 642 (1957).

[133] S. G. Whittington, *J. Phys. A: Math. Gen.* **31**, 3769 (1998).

[134] J. M. Hammersley, G. M. Torrie, and S. G. Whittington, *J. Phys. A: Math. Gen.* **15**, 539 (1982).

[135] A. Trovato, M. P. Riva, and A. Maritan, in preparation.

[136] Ya. G. Sinai, *Theory Prob. Appl.* **38**, 382 (1993).

[137] E. Bolthausen and F. den Hollander, *Ann. Prob.* **25**, 1334 (1997).

[138] R. E. Goldstein, A. Goriely, G. Huber, and C. W. Wolgemuth, *Phys. Rev. Lett.* **84**, 1631 (2000).

[139] T. C. Bishop, and J. E. Hearst, *J. Phys. Chem.* **B102**, 6433 (1998).

[140] J.-R. Daban, and A. Bermudez, *Biochemistry* **37**, 4299 (1998).

[141] J. Bednar, R. A. Horowitz, S. A. Grigoriev, L. M. Carruthers, J. C. Hansen, A. J. Koster, and C. L. Woodcock, *Proc. Natl. Acad. Sci. USA* **95**, 14173 (1998).

[142] A. Stasiak, and J. H. Maddocks, *Nature* **406**, 406 (2000).