SISSA ISAS

Scuola Internazionale Superiore di Studi Avanzati
International School for Advanced Studies

Phases of Polymers and Biopolymers

Thesis submitted for the degree of

Doctor Philosophiæ

Candidate:
Davide Marenduzzo

Supervisor:
Prof. Amos Maritan

October 2002

Contents

Introduction

Molecular biology and biochemistry have traditionally constituted an enormous reservoir of interesting and important problems for polymer physics and statistical mechanics. The two fundamental molecules of life are proteins and nucleic acids, such as DNA and RNA. Their natural forms are far from being featureless, they instead display a high degree of internal order. DNA is made up of two regular intertwining helices[1] of opposite chirality. Proteins (and also RNA) are known to fold reproducibly[2] into native states which, thanks to nuclear magnetic resonance experiments, are now known to be made up of building blocks, or 'secondary structure', with high symmetry. For proteins these secondary structures are the well-known alpha helices and beta sheets[3]. Explaining the origin of the optimal shapes attained by biomolecules, as well as describing even with a rough accuracy the folding process is a long-standing question which can be only claimed to have been solved partially at present day.

A second key question for biophysicists and molecular biologists is how these molecules of life function, and further in what way their $3D$-structure is related to their biological role in the cell, i.e. 'in vivo'. In this respect, the push is towards an understanding of the thermodynamic, physical and mechanical behaviour of the biopolymers as factors such as temperature and solvent composition are changed and as external forces are exerted on them either 'in vivo' by molecular motors or by means of sophisticated machinery in the laboratory.

Studying the physics and the response to external disturbances of one isolated polymer, which is a nano- or micro-sized object, is by no means an easy task for experimentalists. To this purpose, a completely new field, at the border between physics and technology, has had to be started and developed very rapidly in the last decade: that of *single molecule experiments*[4].

These single molecule techniques have enabled experimentalists to monitor with

high sensitivity instruments the response of individual biomolecules to external stresses and to manipulation in general. Since many years ago, the force versus elongation characteristic curves of single and double stranded DNA have stimulated theoretical research on models of polymer elasticity, culminating in the work of Ref. [5] in which it was realized that the *worm-like chain* model accounts for all the raw features observed. More recently, other puzzles have arisen as soon as techniques like atomic force microscopes (AFM)[6], optical and magnetic tweezers[7], and soft microneedles[8], have allowed to gather more and more measurements and data. These experiments revealed and characterized the overstretching regime[9] in pulled polymers, the unzipping transition in a double-stranded DNA[7] and in RNA[10]. They also substantiate the theoretical prediction of a Rayleigh instability when stretching a collapsed polymer, a protein or DNA molecule [11, 12, 13, 14].

On the one hand, these experiments have provided data against which to test the prediction of various polymer physics models. At the same time, these studies, whether theoretical or experimental, are expected to have an important impact on biology as well, because the mechanical performance of proteins, nucleic acids and protein motors is often a fundamental feature of their biological function.

One instance of potentially high biological interest comes from DNA unzipping. The double helical structure of DNA contains the genetic information in the core of the double helix, preventing in this way easy access (and accidental damage) by proteins to the genetic code. Since the elucidation of DNA structure, it has become clear that DNA replication and DNA transcription into messenger RNA necessitate the unwinding, or unzipping, of the two paired strands. Indeed it is known that DNA replication is a *correlated* process involving many proteins and other molecules[15] working at different points in space and time. Understanding the nature and origin of this correlation is in fact a major motivation for statistical mechanics models of DNA unzipping. It has been demonstrated[16, 17] that the force induced unzipping of DNA is a genuine phase transition, different from its thermal melting denaturation. It was hypothesized[16] that the initiation of replication at the origins along the DNA, *e.g*, by dnaA for E.Coli[15, 18] or by the "origin recognition complex" (ORC) in eukaryotes[19] is like this unzipping near the phase transition point (with dnaA or ORC acting as the force-inducing agent) and the resulting correlation during unzipping leads to the co-operativity required for replication. Furthermore, a sound investigation of DNA replication in vitro requires an understanding of the coupling between the opening of the strands and the subsequent events during the process. This has motivated the introduction of coarse grained models which allow the study of the unzipping transition and to characterize the dynamical behaviour, and the role of the

solvent.

It should however be recalled that originally DNA unzipping experiments were ideated with a somewhat more 'practical' purpose. The 'static' force which is required to open a heterogeneous DNA, whose base pair content (i.e. the CG and AT content, where A,T,C and G denote the well-known four bases of DNA) is not known a priori, has been shown [7, 20] to contain a high degree of internal structure. Here 'structure' means mostly the oscillations which are observed in the force versus displacement curves recorded experimentally. This 'structure' is related to the local content of CG base pairs, which are more tightly bound by hydrogen than are AT bonds, in the DNA sequence. In principle, one could hope of even 'sequencing' a whole DNA by simply pulling it apart, which would constitute a dramatic improvement over the present-day sequencing procedure which is time and money consuming. This aim and (up to now unfulfilled) hope has brought up a very lively theoretical debate. In the end, complete sequencing in present day experiments is unattainable due to fundamental difficulties inherent in the measurements, according to Ref. [21], though it should be possible to get coarse-grained information on the local CG vs AT base pair content over patches of ten base pairs roughly. This would still be of relevance for sequence determination, in case we want *quick* and *coarse* information on a DNA composition: in many cases it is not necessary to have the whole base-by-base sequencing to understand a particular issue related to that molecule. Also the possibility of detecting *point mutations* has been discussed theoretically based on exactly solvable models[22], as well as tested thoroughly experimentally[7]. This seems to be more promising on a practical level than complete sequencing. Quite a few other interesting experiments, such as that proposed in Ref. [23] in order to measure dynamical effects in the force versus elongation curves, have been suggested in recent years by theorists working in the field.

A second biologically interesting phenomena is *protein mechanical unfolding*. By applying forces to *proteins*, they can be deformed and eventually completely unfolded. The unfolding dynamical pathways gives information on the shape of the free energy surface in the neighbourhood of the native state, which is of importance for theory and experiments. Moreover, many proteins are designed in such a way that their native folded states are able to withstand forces without disrupting: stretching experiments are expected to clarify the relation between $3D$ native state shape and biological function. The giant titin protein, also known as connectin, which is found in muscles is a well-known and well-studied example of a polypeptide whose mechanical properties are essential for its biological role. The passive tension developed in muscle sarcomers when stretched is largely due to the rubberlike properties of titin.

Stretching experiments[11] have revealed that the unfolding of this protein occurs via different 'steps' in each of which a distinct immogluboline domain is stretched at a time. These steps are clearly marked by peaks in the force versus extension experimental curve.

We now step back to the first mentioned, crucial and still open problem, namely that of *protein folding*. This in fact is one of the 'hottest' topic in molecular biology, or at least that for which the biggest number of contributions from physicists have come and in which the most lively debate is present nowadays between different groups who have rather different views on this complicated and fascinating issue.

Protein folding is the process through which a protein, starting from a swollen or extended configuration, reaches its native (ground) state. It is a pure physical-chemical process, since many 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 Anfisen[2]. Denatured unfolded proteins, which are biologically not active because of non physiological conditions (temperature, pH), or because they have been mechanically opened (see above) restore their biological functionality by returning to the native state. In other words, protein can fold and unfold reproducibly 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. Anfinsen's experiment provides the main reason for the statistical mechanics approach to the protein folding problem. Indeed it can be interpreted by assuming the native or folded state of a protein to be the free energy minimum of a system composed by the polypeptide chain and the solvent molecules. The folding of very large proteins is instead often facilitated by 'chaperones', which prevent improper protein aggregation (think e.g. of the example of prion aggregation in the bovine spongiform encephalopathy (BSE) disease, colloquially known as the 'mad cow' disease).

Since Anfisen's experiment, a number of conceptual questions and real enigmas have been posed by the folding of proteins. First, folding is thought to be driven mainly by hydrophobic interactions which make so that the protein constitutes a core where water cannot penetrate. Hydrophobic interactions alone however seem to be insufficient to study the problem even in a very poor approximation, because a native state is usually non-degenerate, whereas the compact phase of a polymer under the action of a two-body attractive potential, which should mimic the hydrophobicity rather well, has a high degeneracy. Moreover, protein folds have a *hierarchical* degree of

order[24, 25]: chains of the order of 100 aminoacids forms *secondary* motifs, which are helices and sheets which are then put together almost unspoiled in order to form a *tertiary* structure. This are single domain proteins. Longer chains have another hierarchical level where many independent 'domains' aggregate to form a quaternary structure. On the other hand a polymer in its collapsed phase is featureless[26]. Heteropolymers with pairwise interaction potentials are expected to have glassy behaviour and thus the folding process from an extended conformation to the native state would be dynamically rather difficult, as many local minima have to be overcome before coming to the target state. On the other hand, proteins usually fold in a fast and reproducible way[27]. In other words, it is thought that the free energy surface has a rather large attraction basin near the native state (indeed the form of the free energy is speculated to be that of a funnel[28], with high propensity towards the native state), whereas heteropolymer models tend to have a 'golf-like' rugged energy landscape with a ground state which is non-degenerate but also with many competing local minima.

Moreover, given that there are 20 aminoacids and that a typical protein length is ~ 100 residues, there are in principle an enormous number of protein sequences of aminoacids. However, many proteins, which have widely different sequences (more than 70% of difference in the sequences) fold in the same, or nearly so, native state[29]. This has been to some extent substantiated also by calculations done with toy models on the lattice[30, 31]. It has been found that the map sequence to structure is many-to-one, not one-to-one, and that there were very few goal structures which were selected as the ground state by many different sequences. Another surprising property of proteins which is rather hard to account for is that they are very flexible in order to perform a wide array of functions. In addition, some aminoacids which are known to better (i.e. sterically) fit in an alpha helix, in some conditions are accomodated in other secondary structures (a beta sheet or a loop), so that the selection mechanism of structure should be versatile.

Plan of the thesis

In this thesis we develop coarse grained models aiming at understanding physical problems arising from phase transitions which occur at the single molecule level. The thesis will consist of two parts, grossly related to and motivated by the two subjects dealt with above. In the first half, we will focus on critical phenomena in stretching experiments, namely in DNA unzipping and polymer stretching in a bad solvent. In the second part, we will develop a model of thick polymers, with the goal of under-

standing the origin of the protein folds and the physics underlying the folding 'transition', as well as with the hope of shedding some light on some of the fundamental questions highlighted in this Introduction.

In the first part of the thesis we will introduce a simple model of self-avoiding walks for DNA unzipping. In this way we can map out the phase diagram in the force vs. temperature plane. This reveals the present of an interesting cold unzipping transition. We then go on to study the dynamics of this coarse grained model. The main result which we will discuss is that the unzipping dynamics below the melting temperature obeys different scaling laws with respect to the opening above thermal denaturation, which is governed by temperature induced fluctuating bubbles.

Motivated by this and by recent results from other theoretical groups, we move on to study the relation to DNA unzipping of the stretching of a homopolymer below the theta point. Though also in this case a cold unzipping is present in the phase diagram, this situation is richer from the theoretical point of view because the physics depends crucially on dimension: the underlying phase transition indeed is second order in two dimensions and first order in three. This is shown to be intimately linked to the failure of mean field in this phenomena, unlike for DNA unzipping. In particular, the globule unfolds via a series (hierarchy) of minima. In two dimensions they survive in the thermodynamic limit whereas if the dimension, d, is greater than 2, there is a crossover and for very long polymers the intermediate minima disappear. We deem it intriguing that an intermediate step in this minima hierarchy for polymers of finite length in the three-dimensional case is a regular mathematical helix, followed by a zig-zag structure. This is found to be general and almost independent of the interaction potential details. It suggests that a helix, one of the well-known protein secondary structure, is a natural choice for the ground state of a hydrophobic protein which has to withstand an effective pulling force.

In the second part, we will follow the inverse route and ask for a minimal model which is able to account for the basic aspects of folding. By this, we mean a model which contains a suitable potential which has as its ground state a protein-like structure and which can account for the known thermodynamical properties of the folding transition. The existing potential which are able to do that[32] are usually constructed 'ad hoc' *from knowledge of the native state*. We stress that our procedure here is completely different and the model which we propose should be built up starting from minimal assumptions. Our main result is the following. If we throw away the usual view of a polymer as a sequence of hard spheres tethered together by a chain (see also Chapter 1) and substitute it with the notion of a flexible tube with a given thickness, then upon compaction our 'thick polymer' or 'tube' will display a rich

secondary structure with protein-like helices and sheets, in sharp contrast with the degenerate and messy crumpled collapsed phase which is found with a conventional bead-and-link or bead-and-spring homopolymer model. Sheets and helices show up as the polymer gets thinner and passes from the swollen to the compact phase. In this sense the most interesting regime is a 'twilight' zone which consists of tubes which are at the edge of the compact phase, and we thus identify them as 'marginally compact strucures'. Note the analogy with the result on stretching, in which the helices were in the same way the 'last compact' structures or the 'first extended' ones when the polymer is being unwinded by a force.

After this property of ground states is discussed, we proceed to characterize the thermodynamics of a flexible thick polymer with attraction. The resulting phase diagram is shown to have many of the properties which are usually required from protein effective models, namely for thin polymers there is a second order collapse transition (θ collapse) followed, as the temperature is lowered, by a first order transition to a semicrystalline phase where the compact phase orders forming long strands all aligned preferentially along some direction. For thicker polymers the transition to this latter phase occurs directly from the swollen phase, upon lowering T, through a first order transition resembling the folding transition of short proteins.

Chapter 1

Background and Methodology

In this Chapter we introduce some of the basic terminology, background information and methodology, both analytical and numerical, which will be used throughout this thesis.

Most biomolecules (DNA, RNA and proteins) are polymers, i.e. they consist of a linear chain made up of repeating structural units. In the case of DNA, these units are the four bases cytosine (the shorthand is C), guanine (G), adenine (A), and timine (T). For proteins, the repeating units which form the *polypeptide chain* are the twenty aminoacids.

We will thus start with a simple introduction on mathematical models which aim to describe a polymer. Then we will give a brief overview of the techniques used in this thesis in order to treat these problems. The presentation is in increasing order of problem difficulty: first we deal with analytical techniques, then with exact numerical approaches, last with numerical simulations which are nowadays[33] an indispensable tools for systems which are not amenable to exact analysis (such as a three-dimensional self-avoiding walk with interaction[34] or the thick polymer study in Part 2 of this thesis).

1.1 Biological background

1.1.1 Properties of DNA

The DNA molecules in each cell of an organism contain all genetic information necessary to ensure the normal development and function of that organism. This genetic information is encoded in the precise linear sequence of the nucleotide bases from which DNA is built.

DNA is a *double-stranded* molecule, i.e. it is made up of two strands paired to one another by means of hydrogen bonds. There are four such nucleotide bases which build each one of the two strands of DNA: adenine (A), cytosine (C), guanine (G), and thymine (T). Bases A and G are classified as *purine*, C and T as *pyrimidenes*. The bases are capable of forming hydrogen bonds among them, but in a selective way: A can only couple to T and C to G, so in DNA there are two possible *base pairs* or base pairing interactions: AT or TA, CG or GC. The C-G base pair (bp) involves the formation of three hydrogen bonds, which makes it a stronger interaction than the A-T bp, which involves two H bonds. The typical strength of one H bond is $2 - 5$ KCal/mol. The two DNA strands are *complementary* to each other: the bases in one strand are capable of making H bonds with the corresponding bases in the other one (e.g. one strand is ATACCGG and the other one is TATGGCC). At room temperature, all base pairs are intact: quantitatively, the probability that a base pair breaks due to thermal fluctuations is of the order of 10^{-6}. However, by increasing the temperature, T, this probability gradually increases and *bubbles* forms, i.e. regions where the two DNA strands have their first and last bp's joined and the rest unpaired. Ultimately, we get to a value, called melting temperature or T_m, in which the number of intact base pairs drops abruptly and DNA is in the denatured form, i.e. it is no longer a double-stranded polymer but the two strands are better represented as single-stranded chains virtually not interacting with each other. Experimental spectroscopic results on DNA denaturation are very well-established[35]: the absorbance of light at 260 nm is monitored as it is proportional to the number of broken bp's. If room T is restored, DNA reversibly recovers its native shape of double- stranded polymer.

We now come to the structure(s) of the DNA molecule. Early diffraction photographs of such DNA fibers taken by Franklin and Wilkins in London and interpreted by Watson and Crick in Cambridge revealed two types of DNA: *A-DNA* and *B-DNA*. The B-DNA form is obtained when DNA is fully hydrated as it is 'in vivo'. A-DNA is obtained under dehydrated non-physiological conditions.

Both A-DNA and B-DNA have the familiar *local* shape (also called *secondary structure* in analogy with the nomenclature commonly adopted for proteins, see below) of a double helix, or better of a right-handed helical staircase (see Fig. 1.1). The rails are two parallel phosphate-sugar chains, which are helix of opposite handedness (Fig. 1.1), and the rungs are purine-pyrimidine base pairs. In A-DNA there are an average of 10.9 base pairs per turn of the helix, which corresponds to an average helical-twist angle of 33.1° from one base pair to the next. The spacing along the helix axis from one base pair to the next is 0.29 nm. In B-DNA these values are 10 base pairs, 35.9° and 0.34 nm, respectively. There are, however, considerable variations

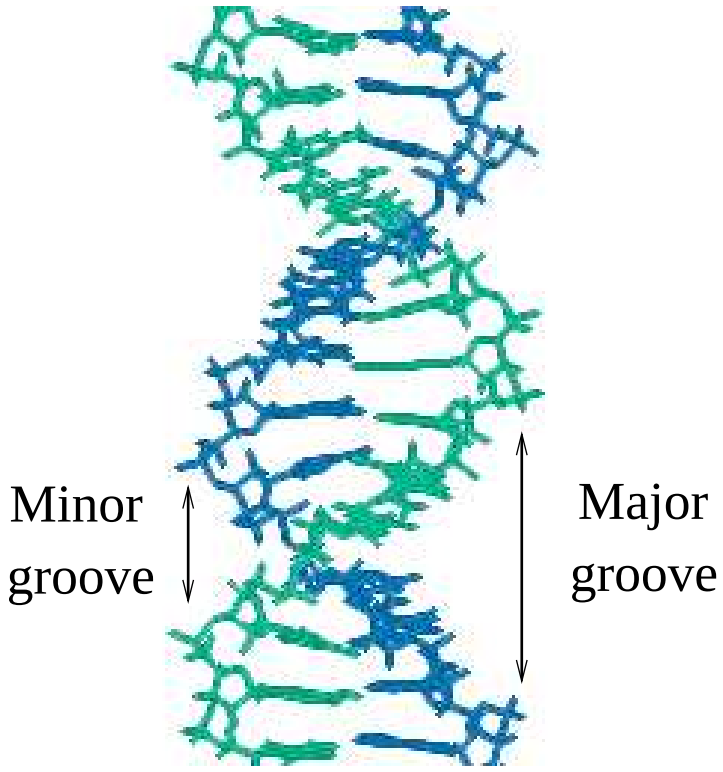


Figure 1.1: The DNA double-helix. We have highlighted the major and minor grooves. All atoms in the two strands are shown in the picture.

in individual twist angles from the average values, and these variations are larger in A-DNA than in B-DNA. Note also that the base pairs are attached asymmetrically to the backbone. Due to this, one *groove* (i.e. separating distance) between the strands is wider than the other. These are called the *major* and the *minor* groove. Both grooves provide opportunities for base-specific interactions, but the major groove is better suited for that task and more often observed as the primary binding site for proteins.

Though DNA is a double-stranded polymer, it is often useful and enough for the needed accuracy of description, to think of it as a single polymer. The relevant physical properties of DNA as a polymer are known since long time ago: DNA is a linear molecule, its diameter is about 2 nm (2.5 if it is hydrated), whereas if stretched out its length can reach many millimeters (this length is referred to as the *contour length* of the molecule). This means that concentrated solutions of DNA can be pulled into fibers in which the long thin DNA molecules are oriented with their long axes parallel. DNA is relatively stiff, it has a high resistance to bending, which is quantified by its *persistence length*, which is around 150 base pairs (bps), or (see above) roughly

50 nm.

Naked DNA is strongly self-repelling, mainly due to the electrostatic repulsion between the charges along its backbone. In a solution, in presence of positive multivalent counterions, though, it can collapse into compact structures with very peculiar shapes. In presence of a low concentration of counter-ions, DNA condenses into spools or toroids of a radius of the order of its persistence length. For very long DNA molecules, or for high counterion concentration, on the other hand, it will form a regular phase with long segments of DNA straight and parallel to each other in a way which resembles the nematic phase of liquid crystals. Though these properties will not be our concern, we will come back to the picture of DNA as a single polymer (neglecting the double-strandedness) when we deal with thick polymers, by introducing a model which in principle is expected to describe well DNA when neglecting its secondary structure. The model, which we introduce in Chapter 4, should be thought of as an alternative to the usual polymer model of DNA in which the molecule is modeled as a *stiff*, rather than a *thick* polymer.

1.1.2 Properties of proteins

It is customary to distinguish several levels of organization in protein structure.

The *primary structure* is the chemical sequence of aminoacids along the polypeptide chain.

Local ordered motifs occurring in most known proteins are called *secondary structures*. They were first predicted theoretically by Pauling and Corey on the basis of energetic considerations[3, 36, 37]. The different secondary structures are alpha helix, beta strands and loops.

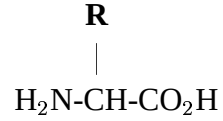
The compact packing of secondary structures determines the unique full three-dimensional native conformation of a biologically active protein. It is also referred to as the *tertiary structure*. It is the result of delicate tuning of various kinds of physical interactions occurring between different atoms of the protein chain and between these and solvent molecules.

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

We will now briefly analyze the physico-chemical and structural basic properties of proteins at each level of hierarchical organization of their structure.

From a physico-chemical point of view, proteins are heteropolymers, made up of

different 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. The central carbon atom to which the side chain is bonded is called the α-carbon, or C_α. The other C atom is also referred to as C' atom.

The chemical composition of side chains varies significantly from one aminoacid to another. Glycin is the lightest one and its side chain consists simply of one hydrogen atom. Triptophan is the heaviest one and its side chain contains both an aromatic ring and an indole ring, with one nitrogen atom. The most frequent atoms in side chains are H, C, N and O, but a sulphur atom is also present in two side chains, methionine and cysteine. Except for glycine, the central C_α atom is asymmetric. In all known *natural* proteins, the C_α's have *left-handed* chirality.

Proteins are formed by polycondensation of different aminoacids. The chemical binding of two aminoacids produces the *peptide bond*, with release of a water molecule. The peptide bond links the two CH group attached to the side chains of two neighbouring amminoacids via an NH group. All aminoacids forming a protein are then linked via the peptide bond. Proteins are thus called polypeptide chains in which the basic repeating unit is the *aminoacid residue*, which is the form the aminoacid takes after polycondensation, when it is embedded in a chain. Length in proteins range from ~ 50 aminoacids, for small globular proteins, to ~ 3000, for complicated multi-domain ones. In order to describe the protein secondary structure, it is convenient to divide the polypeptide chain into *peptide units* (these are the repeating units in the polymer) which go from one C_α atom to the successive one in the chain (see also models such as that in Part 2 of this thesis, in which the protein backbone is schematized by the C_α atoms only). Each C_α atom thus, except from the first and the last one, belongs to two such units. The reason for dividing the chain in this way is that all the atoms in such a unit are fixed in a plane with the bonds and the bond angles very nearly the same in all units in all proteins. Note that the peptide units do not involve the side chains. The distance between two successive C_α atoms is 0.38 nm roughly. Since the peptide units are effectively rigid groups that are linked into a chain by covalent bonds at the C_α atoms, the only degrees of freedom they have are

rotations around these bonds. Each unit can rotate around two such bonds: the C_α-C' and the N-C_α bonds. By convention the angle of rotation around the N-C_α bond is called ϕ, and the angle around the C_α-C' bond from the same C_α atom is called ψ. In this way each aminoacid residue is associated with two conformational angles ϕ and ψ. Since these are the only degrees of freedom, the conformation of the whole main chain (i.e. without the side chains) of the polypeptide is completely determined when the ϕ and ψ angles for each aminoacid are known.

All physical interactions occurring between the atoms in 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 with a fully *ab initio* quantum molecular dynamics method. Simulations of even very small peptides are, however, computationally very expensive, 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 is determined by covalent bonds. Covalent interactions involve a typical energy ranging from 50 KCal/mol to 150 KCal/mol. At room temperature $k_BT \sim 0.6$ KCal/mol (k_B is the Boltzmann constant), thus covalent interactions for all practical purposes freeze the corresponding degrees of freedom to their minimum energy value.

The secondary structure and the complex three-dimensional structure of the folded native state is on the other hand the result of the 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 solvent molecules. The characteristic energy scale of non-covalent interactions ranges from 1 to 5 KCal/mol. The associated degrees of freedom are thermally excited at room temperature, and are thus responsible for the folding and the thermodynamics of proteins. Non-covalent interactions between different atoms of the protein-solvent system are usually divided into electrostatic forces, van der Waals interactions, and hydrogen bond interactions.

An easy way to present secondary structures in proteins is via the Ramachandran plot [38] (Fig. 1.2). Many combinations of the conformational angles (ϕ and ψ defined above) are not allowed because of steric collisions between the side chains and the main chain. Ramachandran plots show allowed combinations of the conformational angles ϕ and ψ. Shaded areas show the sterically allowed regions, which correspond to the regular motifs corresponding to secondary structures in proteins.

Three types are highlighted: α helices, β sheets and collagen helix. Left-handed helices are not observed, since the side chains would be too close to the backbone.

The right-handed alpha helix (shown in Fig. 1.3) is found when a stretch of consecutive residues all have the ϕ, ψ angle pair approximately -60° and -50° (see Fig. 1.2). The α helix has 3.6 residues per turn, which corresponds to a rise of 0.54 nm per turn (0.15 nm per residue) along the helix axis. There are hydrogen bonds between $C' = O$ of the n-th residue with NH of the $n+4$-th. Thus all NH and $C' = O$ groups are joined with hydrogen bonds except the first NH and the last $C' = O$ groups at the ends of the α helix. As a consequence, the ends of α helices are polar and are almost always at the surface of protein molecules.

The second most regular and identifiable secondary structure is the β sheet (see Fig. 1.4). 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 kind of helix with 2 residues per turn. A single β strand is not stable, as no interactions are present among the atoms. The β-strand is stabilized only when two or more strands 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 chain from neighbouring residues of adjacent strands are projected instead into the same side, and thus interact significantly. Adjacent β-strands can be either parallel or antiparallel, and the resulting geometry varies slightly. In antiparallel sheets, all hydrogen bonds are parallel to each other, whereas in parallel sheets they are arranged in two different alternating directions.

We finish our survey of protein structures by mentioning two examples of tertiary or better supersecondary motifs recurring in the tertiary structures of proteins (see Ref. [24]): one is the helix-loop-helix or calcium binding motif, typical in all α proteins (i.e. proteins with only α helices in their secondary structure content) and the $\beta - \alpha - \beta$ motif. The calcium binding motif is symbolized by a right hand with forefinger and thumb up. The first helix runs from the tip to the base of the 'forefinger', the flexed 'middle finger' corresponds to the turn or loop which binds calcium, and the second helix runs up to the end of the 'thumb'. In the $\beta - \alpha - \beta$ motif, two adjacent parallel β strands are connected by a helix from the end of the first strand to the beginning of the second one. The two strands lie in the same plane.

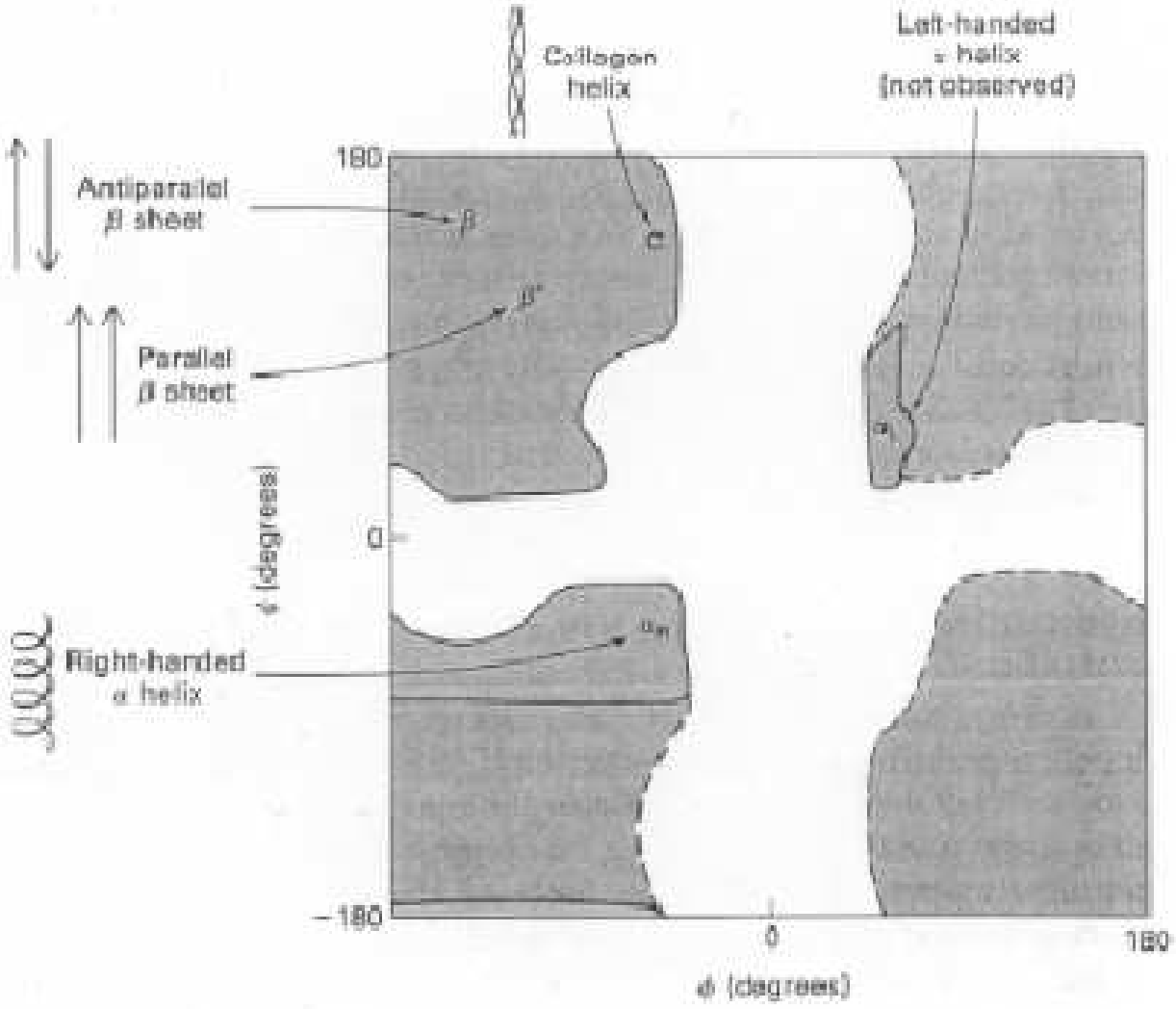


Figure 1.2: Ramachandran plot showing allowed values of the torsion angles ϕ and ψ for alanine residues (region contoured by solid lines). Additional conformations are accessible to glycine (contoured by dashed lines) because it has a very small side chain. The typical values of the torsion angles corresponding to the different secondary structures are shown.

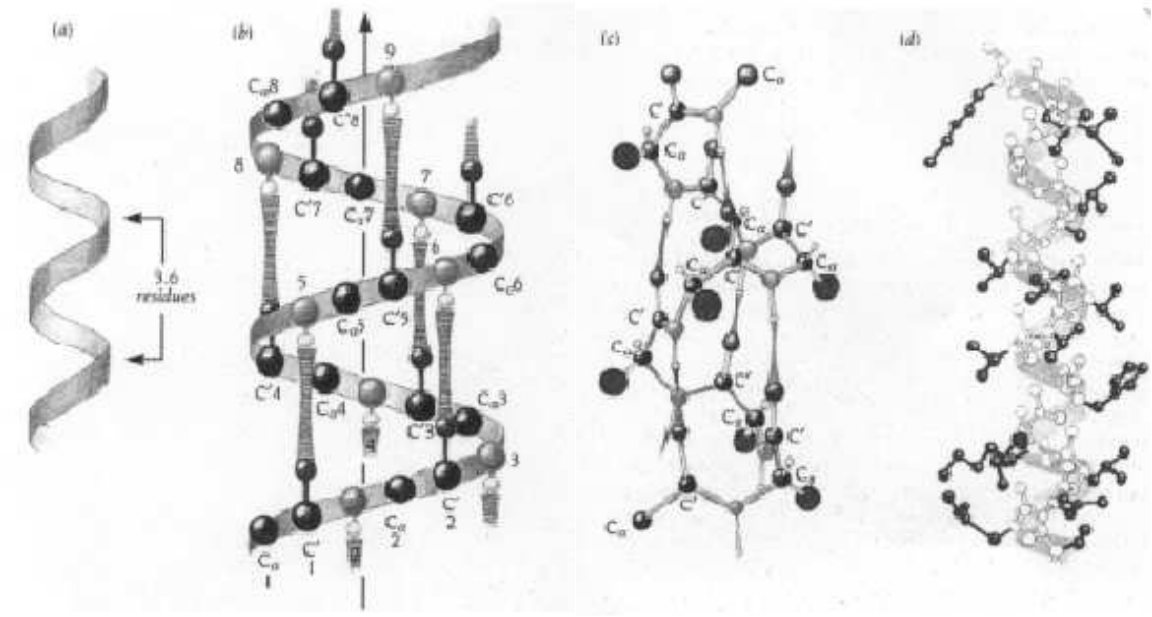


Figure 1.3: (a) Idealized diagram of the backbone path in an α helix. (b) The same as (a) but with approximate positions for the backbone atoms and hydrogen bonds included. (c) Schematic diagram with the correct position of all backbone atoms. Big dark circles represent side chains. (d) A ball-and-stick model of one α helix in myoglobin.

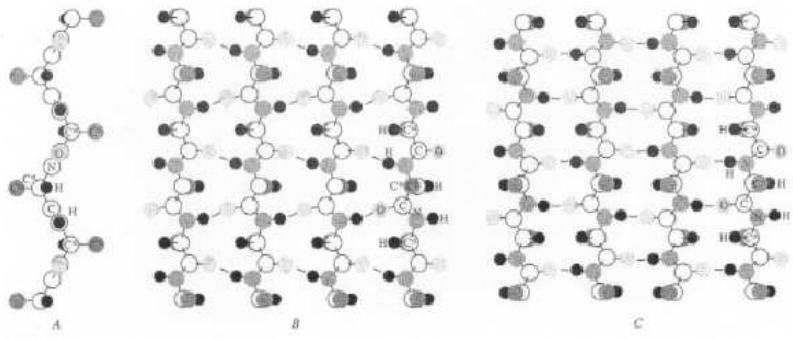


Figure 1.4: (A) A single β strand. (B) A planar parallel β sheet. (C) A planar antiparallel β sheet. The horizontal direction in (A) is orthogonal to the plane represented in (B) and (C).

1.2 Polymer models

As we have seen, both DNA and proteins are polymers made up of different repeating units, even though their physical properties are rather different. Flexible, semiflexible and rigid polymers can take up many configurations by the rotation of chemical bonds. It is therefore meaningful to describe a polymer with the methods of statistical physics. This is what we are going to do from now on in this Section.

1.2.1 Ideal polymers: the freely jointed chain

Let us start from the simplest possible model: a chain consisting of N links, each connecting two beads and of length l_0, which we will take in the following to be equal to 1 without loss of generality (every length will be scaled by l_0). Such a model is called *the freely-jointed chain* (FJC) [39, 40, 41]. A configuration of the FJC when the dimension, d is e.g. 3, is the set of the $(N+1)$ position vectors $\{\vec{R}_i\}_{i=0,\ldots,N}$ of the beads. Equivalently, it can be characterized by the set of N links $\{\vec{R}_{i-1,i}\}_{i=1,\ldots,N}$, with $\vec{R}_{i-1,i} \equiv \vec{R}_i - \vec{R}_{i-1}$. In the FJC the link vectors $\vec{R}_{i-1,i}$ are independent of one another, so that the weight given to a configuration $\{\vec{R}_i\}_{i=0,\ldots,N}$ is:

$$P(\{\vec{R}_i\}_{i=0,\ldots,N}) = \prod_{i=1}^{N} \frac{\delta(|\vec{R}_{i-1,i}| - 1)}{4\pi}, \tag{1.1}$$

where δ denotes the usual Dirac delta function and 4π is the corresponding normalization factor, to ensure that the integral of $P(\{\vec{R}_i\}_{i=0,\ldots,N})$ over the configuration space is 1.

To characterize the size of a polymer, we may consider the end-to-end distance, denoted here by $\vec{R}$ and defined as $\vec{R} \equiv \vec{R}_N - \vec{R}_0 = \sum_{i=1}^{N} R_{i-1,i}$. If $\langle\rangle$ denotes the average with respect to the probability measure in Eq. 1.1, then it is immediate to realize that $\langle\vec{R}\rangle = \vec{0}$. Nevertheless, $\langle\vec{R}^2\rangle$ is finite and can be used as a characteristic length of a polymer. The same quantity is in general used in polymer physics to define[40, 41] a critical exponent ν of the chain in the following way:

$$\langle\vec{R}^2\rangle \sim N^{2\nu} \qquad \text{as} \qquad N \to \infty. \tag{1.2}$$

By making the explicit calculations, one gets $\nu = \frac{1}{2}$ for the FJC. Other common attributes of polymer which identify its size are the *gyration radius* (see the definition in Chapter 6) and the *contour length*, which is equal to Nl_0.

Another important quantity in a polymer is the *persistence length*. It can be defined in terms of the average dot products of two links which are far apart along the

chain, i.e.:

$$B(i,j) \equiv \langle \vec{R}_{i,i+1} \cdot \vec{R}_{i+j,i+j+1} \rangle. \tag{1.3}$$

In usual models, one has $B(i,j) = B(|i-j|) \sim \exp \frac{|i-j|}{\xi}$, and the quantity ξ, which is the typical range of decay of $B(|i-j|)$, and which typically depends on the interactions between beads on the polymer, is usually called the persistence length of the polymer chain, $\{\vec{R}_i\}_{i=0,\ldots,N}$.

Let us now consider a hyper-cubic lattice in d spatial dimensions, $\mathbf{Z^d}$. On the lattice, the links can have only $2d$ directions and the lattice counterpart of the FJC is the well-known *random walk*. An N-link random walk (RW) is a set of points $\{\vec{R}_i\}_{i=0,\ldots,N}$, with $\vec{R}_i \in \mathbf{Z^d}\ \forall i = 0,\ldots,N$, and with the property that $\vec{R}_i$ and $\vec{R}_{i+1}$ are nearest neighbour on $\mathbf{Z^d}$ for any $i = 0,\ldots,N-1$. We will consider in Chapter 2 a simplified model of DNA unzipping with RWs directed along the lattice diagonal. This will have the virtue of being exactly solvable while retaining some of the important features of the general solutions, which makes it a rather ideal and very instructive case to be treated.

The FJC and the RW are useful model as most of their properties can be calculated exactly and explicitly (as for example the average extension versus force curve, which is nothing but the Langevin function[40, 42]). However, both neglect two important features of real polymers: the first is semi-flexibility (many real polymers like DNA are *stiff*), the second is self avoidance.

We will only briefly discuss how to take care of stiffness here, and discuss more in detail the excluded volume interaction in the next paragraph. This is because in this thesis we will fully implement excluded volume in our models, but we will not usually treat stiffness, unless in some cases to show that our results are not dependent on considering the polymer fully flexible. Indeed, the thick polymers introduced in Part 2 do have an intrinsic *stiffness* which arise because of the *thickness* and has a role in their physics. But we will defer a more thorough discussion on this to Chapter 4.

In order to study semi-flexible polymers, the FJC has to be substituted by the worm-like chain (WLC) (see e.g. [5]). In its simplest version, the WLC is a FJC with an energetic penalty against sharp bends, H_{WLC}, of the form:

$$H_{WLC}(\{\vec{R}_i\}_{i=0,\ldots,N}) \propto \sum_{i=1}^{N-1} \vec{R}_{i-1,i} \cdot \vec{R}_{i,i+1}. \tag{1.4}$$

Very shortly, while the WLC is still ideal in the sense that $\nu = \frac{1}{2}$, its force versus extension curve behaves differently than the FJC for high forces[5], whereas for low forces the behaviour is identical.

1.2.2 Excluded volume interaction: the self-avoiding chain

In real polymers, contrarily to the ideal case described above, interactions among the polymer links are not limited to a few neighbours along the chain, but are in principle highly non-local, as links which are distant along the chain do interact if they come sufficiently close together. Such an interaction is called 'long range' as opposed to the 'short range' interactions of the ideal case. Note here that the term long range interaction stands for steric effects, van der Waals attractions, and also other specific interactions mediated by solvent molecules. However at large length scales statistical mechanics suggests that the details of the interaction should not matter and the physical picture should be universal. In particular, there should be two universality classes which the potential belongs to. We can discuss them jointly by introducing a commonly used two-body potential which acts between beads in $\vec{R}_i$ and in $\vec{R}_j$ respectively ($i \neq j$):

$$V(R_{ij}) = \frac{A_1}{R_{ij}^{12}} - \frac{A_2}{R_{ij}^{6}} \tag{1.5}$$

where $R_{ij} \equiv |\vec{R}_i - \vec{R}_j|$. The square well potential considered in Chapters 5 and 6 is of course expected to stay in the same universality class of the Lennard-Jones potential, thus to be interchangeable with it.

The first universality class is that of purely repulsive interactions ($A_1 > 0$, $A_2 = 0$), in which attractions are not present: this is the self-avoiding chain in the continuum space (e.g. in $\mathbf{R}^3$). A common choice is $A_1 = 4\epsilon\sigma$, where ϵ and σ fix the energy and distance scales respectively. A self-avoiding walk (SAW) on the lattice $\mathbf{Z}^d$, on the other hand, is similarly defined as a random walk whose configurations $\{\vec{R}\}_{i=1,\ldots,N}$ are such that no two beads can share the same lattice site, i.e. $|\vec{R}_i - \vec{R}_j| \geq 1 \; \forall i, j$: $i \neq j$, $i, j = 1, \ldots, N$. Both on the lattice and in the continuum, a self-avoiding chain is in a different universality class than the FJC ($V(R_{ij}) = 0$), as in this case $\nu = 0.588\ldots$ (this value is known via numerical simulations [43]).

The second class of potentials is one which couples a hard core at short distances with an attractive tail for large distances: in Eq. 1.5 one would put $A_1, A_2 > 0$ (a common choice is $A_1 = 4\epsilon\sigma$, $A_2 = -4\epsilon\sigma$). This can be straightforwardly generalized to SAWs on the lattice, giving rise to the famous lattice model for the Θ transition [39, 40, 41, 44, 45, 46]. Here the polymer + solvent system is in one of three states: compact polymer (poor solvent), where $\nu = \frac{1}{d}$, for high enough ϵ; swollen polymer (good solvent) for low ϵ ($\nu = 0.588\ldots$), and Θ solvent at the transition 'temperature',

correspondent to an energy scale ϵ_Θ. In a Θ solvent, the value of ν, ν_θ is given by:

$$\nu_\theta = \begin{cases} \frac{4}{7} & \text{if } d = 2 \\ \frac{1}{2} & \text{otherwise} \end{cases} \tag{1.6}$$

1.2.3 Some analytical and numerical methods for polymers

We here introduce and sketch the basic analytical and 'numerical' techniques used in the following. By 'numerical' we mean that the only numerics involved can be done in a computer with almost arbitrary precision: they are meant to be contrasted to the numerical simulations which we will deal with in the next Section.

A very powerful method for studying ideal polymers such as e.g. the RW is the *generating function technique*. It can be viewed formally either as a Laplace transform (discrete for RWs on the lattice) in mathematical terms or as a grand partition function in statistical terms. Let us consider a concrete example. Suppose that we want to calculate the return probability for a d-dimensional RW to its starting point. If $P_N(\vec{x})$ denotes the number of walks that after N steps reach the site $\vec{x}$, we have that $P_0(\vec{x}) = \delta_{\vec{x},0}$, where δ is the Kronecker δ. Besides, P satisfies the following recursion relation:

$$P_{N+1}(\vec{x}) = \sum_{i=1}^{d} \left(P_N(\vec{x} + \vec{e}_i) + P_N(\vec{x} - \vec{e}_i) \right), \tag{1.7}$$

where $\vec{e}_i$ denote the Euclidean versors. The generating function $P(\vec{x}, z)$, corresponding to $P_N(\vec{x})$, is defined as:

$$P(\vec{x}, z) = \sum_{N=0}^{\infty} z^N P_N(\vec{x}), \tag{1.8}$$

and z is called the *fugacity*. If we make the ansatz (satisfied in all physical cases) that:

$$P_N(\vec{x}) \sim z_c^{-N} N^{-\gamma}, \tag{1.9}$$

valid for big N and $z_c < 1$, and where γ is an entropic exponent resulting in a subleading correction in the free energy. We easily realize that $P(\vec{x}, z)$ has a singularity for $z = z_c$. The inverse of the critical fugacity, z_c^{-1}, is also known as the *connectivity* of the RW. From the Tauberian theorem[47] one can relate the behaviour of the generating function near z_c to the entropic exponent in the canonical ensemble. We thus aim at computing $P(\vec{0}, z)$ near $z = z_c$, this will yield in turn via the Tauberian theorem the number of RW's with N step and with the last step in the origin, namely

$P_N(\vec{0})$. In order to proceed, we define the Fourier transforms of $P_N(\vec{x})$ and $P(\vec{x},z)$,

$$P_N(\vec{q}) = \sum_{\vec{x}} e^{i\vec{q}\cdot\vec{x}} P_N(\vec{x}), \tag{1.10}$$

$$P(\vec{q},z) = \sum_{\vec{x}} e^{i\vec{q}\cdot\vec{x}} P(\vec{x},z). \tag{1.11}$$

From Eq. 1.8, and from the definition of inverse Fourier transform, one easily gets:

$$P(\vec{0},z) = \int \frac{d\vec{q}}{2\pi^2} \frac{1}{1-2z\sum_{i=1}^{d}\cos q_i}. \tag{1.12}$$

The singularity in the integral is obtained for $z = z_c = 1/(2d)$. In $d = 1$ for example one obtains:

$$P(0,z) \sim (z_c - z)^{-1/2} \tag{1.13}$$

which means by the Tauberian theorem that:

$$P_N(0) \sim (2)^N N^{-1/2}. \tag{1.14}$$

As a consequence one easily realizes that the sum $\sum_N P_N(0)$ is divergent in $d = 1$: this physically means that the walker will surely return to the origin sooner or later. As $\gamma = d/2$ for generic d, one easily finds that the walker has a finite probability of escaping from the origin only if $d > 2$. We will use the generating function technique in Chapter 2 in a case with interactions in order to solve a simplified version of our model for DNA unzipping.

Let us now proceed to the case of a SAW on the lattice in $d = 2$. A simple question would be to find its *connective constant* μ_{SAW}, i.e. to find the number μ_{SAW} which satisfies the asymptotic equivalence $c_N \sim {\mu_{SAW}}^N$ for $N \to \infty$, where c_N is the number of N-step SAWs. Using the terminology just introduced, the inverse of the connective constant is the critical fugacity of the generating function associated to c_N. Another quantity of common interest is the critical exponent ν for the $2d$ SAW. These are not accessible via elementary methods. One way to find an approximation to them is to use a *real space renormalization group* procedure. If $c_N(\vec{x})$ is the number of N-step SAW's starting at the origin and ending at $\vec{x}$, its corresponding generating function is:

$$G(\vec{x},z) = \sum_{N=0}^{\infty} c_N(\vec{x}) \tag{1.15}$$

Though we do not make the derivation here, scaling theory requires the above quantity to have the following functional form:

$$G(\vec{x},z) \sim \frac{1}{|\vec{x}|^{d+2-\eta}} F\left(\frac{|\vec{x}|}{\xi(z)}\right), \tag{1.16}$$

where η is an exponent called the anomalous dimension of the SAW and ξ is called the correlation length. Now suppose there exists a mapping in the system from z to $z_l'(z)$ (which can be shown to correspond physically to rescaling the lengths in the system by a factor l) such that:

$$G(\vec{x}, z) = l^{-d+2+\eta} G(\frac{\vec{x}}{l}, z_l'(z)). \tag{1.17}$$

After proving (which we do not do here) that such a mapping exists, then it follows that:

$$\xi(z_l'(z)) = \frac{\xi(z)}{l}. \tag{1.18}$$

At criticality, from scaling theory (see e.g. Ref. [40]) we know that ξ must diverge like $(z_c - z)^{-\nu}$, so that one gets that, if the mapping is regular, it must have a fixed point in z_c, i.e. $z_l'(z_c) = z_c$. Moreover, near z_c ($z = z_c + \delta z$):

$$\xi(z_c - \delta z) = l\xi\left(z_c - \left(\frac{\partial z_l'}{\partial z}\right)_{z=z_c} \delta z\right) \sim (z_c - z)^{-\nu}, \tag{1.19}$$

so it is necessary that $\left(\frac{\partial z_l'}{\partial z}\right)_{z=z_c} = l^{1/\nu}$. We now come to the specific case of the $2d$ SAW. The mapping z_l' corresponds to the $l = 2$ rescaling corresponding to the 2×2 cell renormalization shown in Fig. 1.5.

The equation of this mapping reads:

$$z_2' = z^2 + 2z^3 + z^4. \tag{1.20}$$

The fixed point of Eq. 1.20 is $0.46\ldots$, which gives $\mu_{SAW} = 1/z_c \sim 2.17\ldots$ (the exact value is $2.638\ldots$[47]). The exponent ν is found via

$$\left(\frac{\partial z_l'}{\partial z}\right)_{z=z_c} \sim 2^{\nu} \tag{1.21}$$

and $\nu \sim 0.72\ldots$ (in good agreement with the exact value, $3/4$). More accurate mappings can be found by considering 3×3 or larger cells. If we want to consider interactions between the SAW sites, besides, it proved necessary to study the simultaneous renormalization of two 2×2 cells. Apart from these details, the procedure is completely analogous to the simpler case just discussed. This will be used in order to rationalize the thermodynamics of the polymer stretching in a poor solvent in Chapter 3. Though it was introduced by now a long time ago, we will see that the real space renormalization group proves very useful there, in that it can account for the

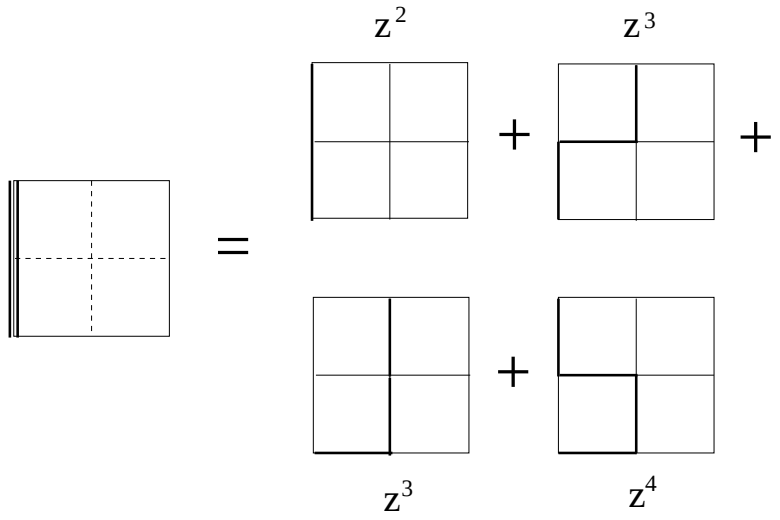


Figure 1.5: Schematic view of the rescaling leading to Eq. 1.20. The double line indicates a step in the renormalized 2×2 cell. Note that the walks corresponding to this particular renormalized step must start from the lower left of the 2×2 cell and end at the upper left corner. This explains why only the walks on the right contribute to the renormalization of the step on the left.

fact that in $d = 2$ the unfolding transition is second order and becomes first order in $d = 3$. This very fact was under debate before and the reason for that behaviour was not known.

We mention that another very useful method in order to deal with self-avoiding walks is the transfer matrix which we will use in Chapter 3. We defer a brief introduction to this method there, when it will be used. A more detailed understanding of it is not needed in order to appreciate the results given in this thesis.

1.3 Monte-Carlo numerical simulations

We now come to the numerical Monte-Carlo techniques that we have made use of in this thesis (see also Ref. [33] for more details). Monte-Carlo simulations are typically performed with two different goals: first, to find the ground state of a given system (optimization problem); second, to calculate ensemble averages through random sampling (equilibrium problem). Since the purposes are different, the methods are also in general different. In this Section we will explain one common optimization Monte-Carlo algorithm, *simulated annealing*, and one equilibration technique,

parallel tempering (following the presentation of Ref.[34]).

1.3.1 Simulated annealing

In 1983, Kirkpatrick and coworkers[48] proposed a method of using a Metropolis Monte Carlo simulation to find the lowest energy (most stable) of a system, which is very commonly used in practical optimization problem such as the one we intend to solve in Chapters 3 and 5: this method goes under the name of 'simulated annealing'. The basic idea of simulated annealing is to search for a minimum of an energy function in the same way as a real physical system whose equilibrium properties are governed by that energy function reaches the ground state of minimal energy in the limit of low temperature. One may think for example of crystalline solids which are formed by several atoms. In practical context, 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 indeed by following a careful 'annealing' schedule: first the substance is made to melt, than the temperature is lowered very slowly, spending a lot of time in the neighbourhood 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 structure.

A simulated annealing optimization proceeds via a similar scheme. First, the internal energy is replaced by the cost function to be minimized. Then, a fictitious temperature is introduced. The fictitious temperature, T, is a control parameter in the same unit as the cost function, E. Then the procedure consists in carrying out numerical simulations by starting at a high temperature, then gradually decreasing it from high values to lower ones, until the system has frozen in a configuration and no further changes occur. For the procedure to be successful, it is in principle necessary that the system be in thermodynamic equilibrium during all the cooling schedule. This ensures that the system can overcome energy barriers in the free energy landscape and as T is low enough, it falls into the ground state attraction basin and gives the optimal configuration. This is of course ideal, because it would imply that we can cool the system down at an infinitesimally slow schedule.

A natural and efficient way to simulate systems in thermodynamic equilibrium at a given T is via Monte-Carlo stochastic dynamics in configuration space[43]. The standard Metropolis algorithm[49] allows one to generate a stochastic process, more often referred to as a *Markov chain*, which samples randomly the configuration space with a probability proportional to the correct Boltzmann weight at the desired T. The unnormalized Boltzmann weight of a configuration of the system, which we indicate

by C, at the temperature T at which we want to sample is simply:

$$P(C) \propto \exp\left(\beta E(C)\right) \tag{1.22}$$

where $\beta \equiv \frac{1}{T}$ is the inverse sampling temperature. The basic step in Monte-Carlo dynamical simulations is to propose an update of the current system configuration, C, into a trial configuration, which we call C'. This updating is done according to a set of predetermined moves. In the case of our interest, in which the system is a polymer chain, the dynamical moves are listed and described in Appendix A. Afterwards, the move is accepted or rejected according to the Metropolis test. This well-known algorithm consists in generating a random number, r, evenly distributed from 0 to 1 and in comparing it with the quantity:

$$P(C \to C') = \min\left(1, \exp\left(\beta(E(C) - E(C'))\right)\right). \tag{1.23}$$

If $r < P(C \to C')$ the move is accepted and the system goes into configuration C', otherwise it is rejected, and it remains in configuration C. This means that all the moves which decrease the cost function E are automatically accepted. In the first phase of the annealing, when T is high enough and the 'freezing' transition is quite far, a relatively large percentage of the random steps that result in an increase in the energy will also be accepted. After a sufficient number of Monte Carlo steps, or attempts, T is decreased. The Metropolis Monte Carlo simulation is then continued. This process is repeated until the final T is reached. As T gets lower, it will be more and more unlikely to accept move which cause an increase in E.

The way in which the temperature is decreased is known as the cooling schedule. In practice, the cooling schedule which is predominantly used is a proportional one in which the new T is a constant K times the old T, where $K \lesssim 1$ (typically between 0.99 and 1). Ideally, one could try to devise an optimal way to find the annealing schedule, as the decrease rate and the number of Monte-Carlo steps per T can be varied during the numerical simulation. In general, an important parameter to monitor during the annealing is the number of accepted moves per T. At high T, this number will be very high as nearly all the moves which do not violate physical constraints (such as self-avoidance in a polymer) are basically accepted, whereas for low T almost all moves can be rejected. There will be a 'critical region' in which the acceptance rate drops rather abruptly. This is where the biggest number of moves should be accumulated, as here the system is doing the crucial moves in order to choose its optimal low T configuration. Another aspect is to determine how the amplitude of the Monte-Carlo trial moves should depend on T. While at high T large moves are in general better so as to enhance the portion of phase space sampled, at lower T,

they get useless because drastic changes are no longer possible and the system only needs to rearrange its configuration in order to improve its energy via small steps. In this way the acceptance rate is reasonable even at low T. One should also note, when choosing the number of steps per T, that, while a pure Metropolis Monte Carlo simulation (done e.g. in order to find canonical averages at a given T or β) attempts to reproduce the correct Boltzmann distribution at a given temperature, a simulated annealing optimization only needs to be run, at a given T, long enough to explore the regions of search space that should be reasonably populated. This allows for a reduction in the number of Monte Carlo steps at each T, but the balance between the maximum step size and the number of Monte Carlo steps is often difficult to achieve, and depends very much on the characteristics of the search space or energy landscape.

1.3.2 Parallel tempering (multiple Markov chain method)

A rather frequent difficulty in Monte Carlo simulations of an equilibrium process is encountered at low T, where the chain moves only little in the configuration space and sampling becomes therefore inefficient (i.e. convergence to the Boltzmann distribution probability is slow). A widely used way to mitigate this difficulty is to use the *parallel tempering* algorithm[34].

Suppose we want to sample the configuration of a system at some value T which is rather low, so that convergence is very slow; but that we know that convergence is fast at some other values $\beta'^{-1} \equiv T' > \beta^{-1} \equiv T$. The idea is to select a set of values $T' = T_1 > T_2 > \ldots > T_m = T$ to interpolate between T' and T so that there is considerable overlap in the distributions of two neighbouring 'replica' of the system, at T_l and at T_{l+1} for $l = 1, \ldots, m$. The Markov chains at $T_1, \ldots, T_m$ are evolved in parallel for a specified number of time steps. Afterwards, an adjacent pair T_l, T_{l+1} of temperature values is chosen randomly, uniformly from the $m-1$ adjacent pairs of 'replicas', and, as a trial move, the configurations at these T values are swapped. Let C_l and C_{l+1} represent the configurations of the replicas at $T_l \equiv {\beta_l}^{-1}$ and $T_{l+1} \equiv \beta_{l+1}^{-1}$ respectively, and let $P_C(\beta_l)$ stand for the probability that the system is in the state C at inverse temperature β_l: then, the probability that the trial swapping between the states is accepted is:

$$\begin{aligned} \pi(l \longleftrightarrow l+1) &= \min\left(1, \frac{P_{C_{l+1}}(\beta_l) P_{C_l}(\beta_{l+1})}{P_{C_l}(\beta_l) P_{C_{l+1}}(\beta_{l+1})}\right) \\ &= \min\left(1, \exp\left(-(\beta_l - \beta_{l+1})(E_{l+1} - E_l)\right)\right). \end{aligned} \tag{1.24}$$

If each Markov chain is ergodic, so is also the composite Markov chain.

The advantages of the method over a traditional set of unrelated Monte-Carlo simulations at different values of T is apparent. In traditional calculations, the configuration of the system at low T tends to have a high probability of getting trapped in a local minimum or metastable configurations. This creates a two-fold difficulty: on the one hand, this minimum being in general only metastable, we will have a poor characterization of the true ground state of the system; secondly, as it is very difficult for the system to sample configurations which have little overlap with that of the minimum it is trapped into, we will get a very biased and unreliable estimate for the ensemble average we want to compute, as they will be calculated on the basis of the knowledge of a restricted portion of the configuration space. Note that this is not a question of waiting for longer simulations, as the time that it takes to the system in order to cross any finite free energy barrier at low enough T gets so high that to it can be considered infinite for all practical purposes. On the other hand, the swapping moves proposed during parallel tempering are drastic, in that we exchange configurations coming from different 'histories'. So the chance of getting trapped in non-optimal local minima is significantly lower, provided that there are high enough temperatures (low enough β's) in the set of interpolating β values so that for those T's overcoming energy barrier is not an issue. At the same time we can enhance the sampling efficiency as configurations coming from different dynamical evolutions will have a much lower overlap with each other with respect to configurations coming from the *same* dynamical evolution. In technical terms, this means that via parallel tempering we enhance the mobility of the chain, and at the same time we reduce the 'correlation time' τ, in the region of low T. The 'correlation time' of a quantity X in a Monte-Carlo run at a given β or more generally in a Markov chain is defined as the time τ after which the ensemble average (i.e. average over distinct realizations of the Markov chain) $\langle X(t)X(0)\rangle$ becomes smaller than $\frac{1}{e}$.

From a practical point of view, parallel tempering is almost always necessary when wanting to sample a polymer in a compact phase (which occurs at low T), in order to get meaningful values for ensemble averages, especially when dealing with technically 'difficult' quantities such as the specific heat which is defined in terms of the energy fluctuation (see in particular Chapter 6 where we use parallel tempering extensively in order to find the phase diagram of a thick polymer).

PART 1: From compact to swollen

I knew a man who grabbed a cat by the tail and learned 40 per cent more about cats than the man who didn't.

Mark Twain (19th century), Carlos Bustamante (2000)

Chapter 2

Statics and dynamics of DNA unzipping

We have given the motivations for the study of DNA unzipping (single molecule experiments and DNA replication) in the Introduction and we have quickly introduced the elementary biological notions on DNA in Chapter 1.

In this Chapter, we will begin by analyzing an ideal single molecule DNA unzipping experiment. Then, we will introduce a polymer physics model to study the force-induced unzipping transition. This requires in general self-avoiding walks, which makes the treatment rather cumbersome and unattackable with analytical methods. A simpler version of the model is then presented. This is amenable to exact solutions and will be used as a guide as it retains the basic features of the complete model, both for statics and dynamics. We then comment on the differences between the two descriptions and discuss the role of disorder. We finally discuss the possible fate of sequencing ideas based on unzipping experiments.

2.1 An ideal unzipping experiment

In Fig. 2.1a we show a typical DNA unzipping experimental setup, while in Fig. 2.1b we show the corresponding result, some force versus extension experimental and theoretical curves. For other possible setups and further details on experiments see Refs. [50, 51, 52, 53, 54, 55].

The setup shown in Fig. 2.1a refers to a soft microneedle experiment. Other possible setups would involve laser tweezers or atomic force microscopes. With the setup shown, experiments have been performed with DNA coming from phage λ, of

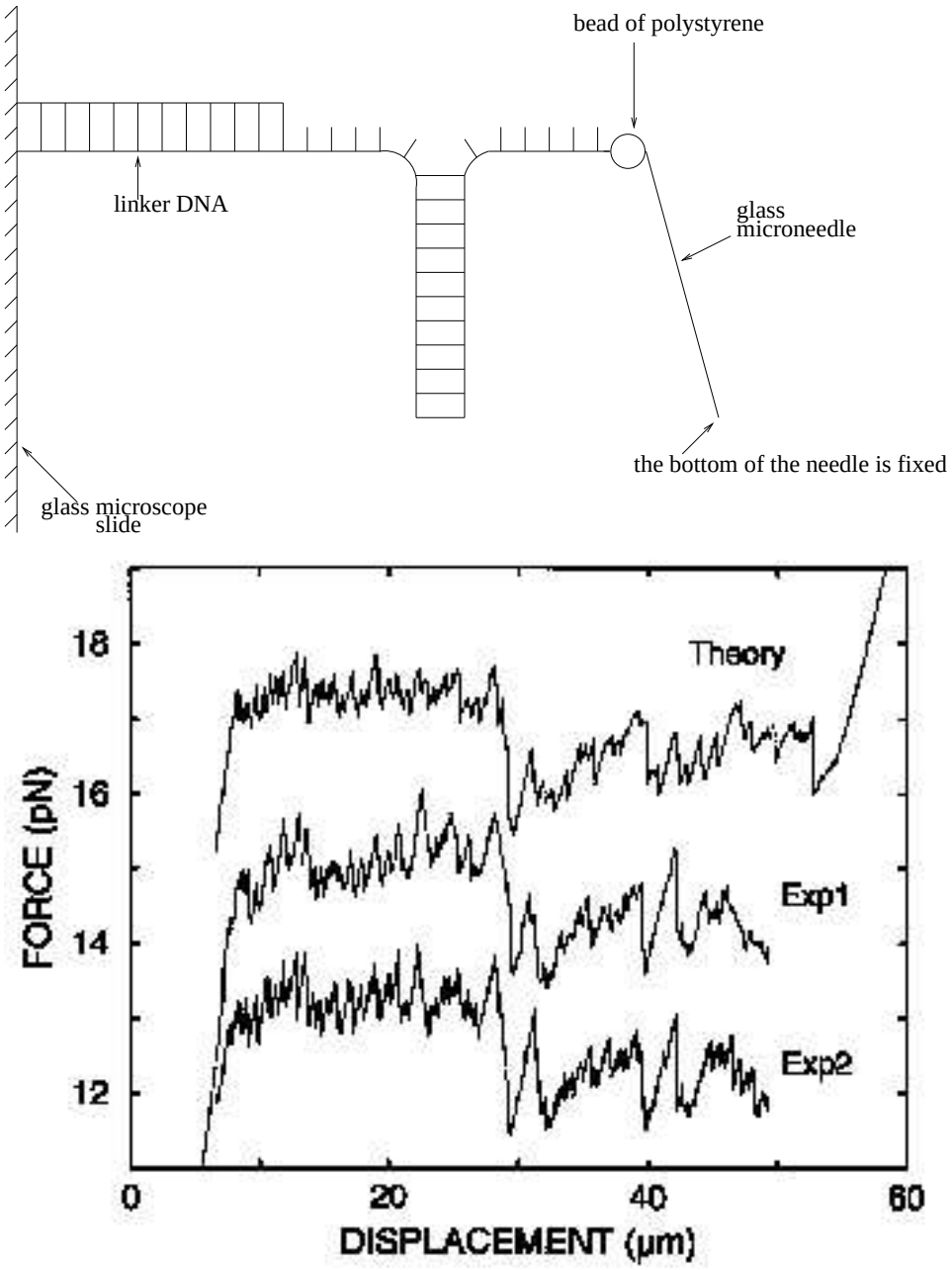


Figure 2.1: a) (Top) Schematical view of the force measurements in the DNA unzipping experiment of Ref. [20] described in the text. b) (Bottom) Force versus elongation (x) curves for a λ phage DNA obtained in Ref. [20]. The bottom and middle curves come from experiments performed with a displacement velocity of the piezo translational stage of respectively 40 and 200 nm/s. The middle curve is shifted upwards by 2 pN for clarity. The third curve, shifted by 4 pN upwards, is obtained in Ref. [20] with a theoretical calculation. Note that experiments are stopped before the DNA molecule is completely unzipped.

contour length of several micrometers ($\sim 4.85 \cdot 10^4$ base pairs), at room temperature, and in physiological conditions (physiological pH, i.e. 7). Note that in physiological conditions a typical DNA melting temperature is $T_m \sim 370\ K$, so that in present experiments T/T_m is around 0.8. However, T can be changed roughly from a few degrees Celsius to the melting temperature, and T_m itself can be modified by acting on the pH of the buffer solution (the maximum melting temperature being around 450 K). This means that present-day values of T/T_m could in principle range between 0.58 and 1 roughly. This ratio T/T_m is of interest to us because it is the parameter which enables ideally our results to be compared with experimental possibilities. A lot of emphasis is usually put on the fact that the persistence length of double-stranded (ds) DNA (roughly 150 base pairs, see Chapter 1) is from 30 to 50 times higher than the persistence length of the single-stranded DNA, which is usually a few base pairs. Though in our simple models we neglect this, as the gross features of our results are unaffected by such features, we will discuss the *quantitative* role of dsDNA stiffness afterwards.

In the guide experiment that we consider, the two strands of the dsDNA at one end are attached via a linker DNA arm to a glass microscope slide, and to a microscopic polystyrene bead (at the other end the strands are joined). The tip of a glass microneedle is attached to the bead and serves as a force lever to measure the tension exerted by the strands. In this experiment, the open end distance is controlled, as the microscope slide is moved at a nearly constant velocity by using a piezo translational stage. The slide velocity v is of the order of 20 to 400 nm/s. This means that every second around 60 to 1200 bases of DNA are opened. In order for an equilibrium calculation to reproduce as much as possible the data, it is necessary that this velocity be as low as possible, because non-equilibrium and dynamical effects are expected to influence strongly fast unzipping measurements. What is then measured is the average force that is necessary to keep the ends of the two strands at a mutual distance x, with $x \sim vt$, and t is time. The resulting characteristic curves are force versus elongation curves. A change in the lever deflection below $0.1\ \mu m$ can be confidently resolved. As the lever stiffness used in Ref. [20] is 1.7 pN$/\mu m$, the force resolution is better than 0.17 pN. If we look at the force versus elongation curves in Fig. 2.1b (note that the curves in the figure are translated for clarity), we note that these curves are constituted by a rugged plateau followed by a sharp increase (which is known to be present but is not shown in the two experimental curves). The average force in the plateau is around 12 pN, but there are excursions and fluctuations about $3 - 4$ pN wide. The fluctuations in the force are mainly related to changes in the local sequence in the part of DNA which is unzipped when the two open ends are at the distance read in the

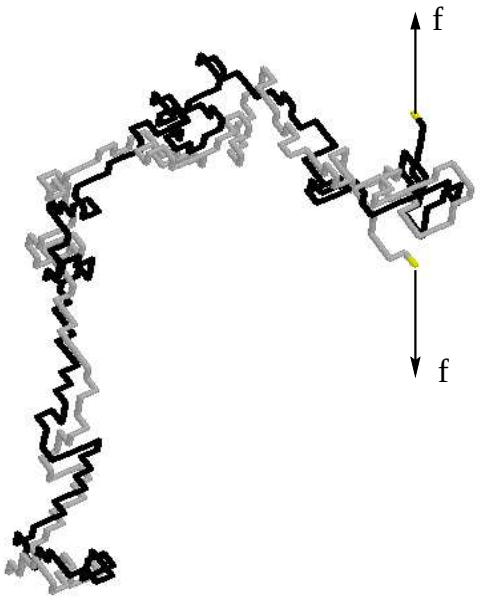


Figure 2.2: One configuration for our SAW model for the unzipping of DNA in $d = 3$.

abscissa. The force needed to open a DNA with only CG or GC hydrogen bonds is around 15 pN, whereas in order to open a DNA with only AT or TA base pairs one would need a force or around 10 pN. So in principle fluctuations of about 1.5% can be nowadays detected in average force measurements.

2.2 A model for DNA unzipping

Our wish here is to define a set of minimal models, in the spirit of statistical mechanics, for which on one hand the unzipping transition can be characterized with high accuracy, and on the other hand such that the results will apply also to more detailed approaches. On the basis of the equilibrium solution, a detailed study of the dynamics not relying on any mean-field approximation can be performed.

Our lattice model is an extensions of the famous Poland and Sheraga model [56, 57, 58, 59]. Quite a few papers related to ours have recently been written both on the unzipping [16, 17, 60, 61, 62, 63, 64, 65, 66, 67, 68] and on the thermal denaturation [69, 70, 71, 72, 73, 74] of DNA. In the case of Ref.[56] the strands are taken

as random walkers with base pair interaction between corresponding monomers, i.e. *wrong base pairing* is forbidden. This means that the $i-$th monomer in the first strand interacts via base pair interaction only with the $i-$th monomer in the second strand, gaining an energy $-\epsilon$ ($\epsilon > 0$) if the two occupied sites are nearest neighbours on the lattice, and nothing otherwise. In our model there is also a force $\vec{f}$ (whose magnitude we call f) pulling the open end of the walks (tied at the other end). Our models are different from that of Ref. [56], as we incorporate the self and mutual-avoidance (hard-core repulsion) of the strands (see Fig. 2.2 where a typical configuration of the two filaments of DNA in our model is shown). This ensures that the zero-force melting temperature T_m takes a finite value also in $d = 2$ unlike in Ref. [56]. Contrarily to what happens when describing the bubble statistics in the denaturation transition [69, 70], it will become apparent that apart from this self-avoidance is not too crucial a property to understand the unzipping transition. In formulas, the configurations of DNA in the canonical ensemble where the number of base pairs, N, is kept fixed is labelled by the possible configurations of two self-avoiding walks, Γ_1 and Γ_2, on the hypercubic lattice, $\mathbf{Z}^d$. The chain Γ_p, with $p = 1, 2$ is specified by the position on $\mathbf{Z}^d$ of its N sites, i.e. $\{\vec{R}_i^p\}_{i=0,\ldots,N-1}$. The energy of a DNA configuration, (Γ_1, Γ_2) is called $H(\Gamma_1, \Gamma_2)$ and explicitly reads:

$$H(\Gamma_1, \Gamma_2) = -\epsilon \left(\sum_{i=0}^{N} \delta\Big(|\vec{R}_i^1 - \Gamma_i^2|, 1\Big) \right) - \ f\hat{e}_z \cdot \left(\vec{R_N}^{\,1} - \vec{R}_N^2 \right), \tag{2.1}$$

where $\delta(a, b)$ is the Kronecker delta and $|x - y|$ is the Euclidean distance between x and y in lattice units. Note that we have taken, without loss of generality, $\vec{f}$ directed along the z axis, whose versor is $\hat{e}_z$.

Two different types of models can be defined at this level:

(i) The "Y-model" in which the two strands are zipped together up to a bifurcation point, *i.e.* the only allowed configurations are those which have the first $N - m$ monomers bound and the remaining m separated as in a Y.

(ii) The "b-model" in which configurations with "bubbles" are allowed. In real situations there are single-strand binding (SSB) proteins which bind to the opened strands to prevent rejoining. This justifies the "Y-model" where the bubble formation (rejoining) is completely suppressed. However in vitro studies need not contain the SSB proteins and bubbles (denatured regions) can form anywhere along the strands due to thermal fluctuations. This is realized in the "b-model" (see e.g. Figs. 2.2 and 2.3).

Throughout all the calculations $\epsilon = +1$ and therefore all the thermodynamic properties of the systems depend on the temperature $T = \beta^{-1}$ and on the force f. The

free energy per monomer $\mathcal{F}_N(\beta, f)$, is

$$\mathcal{F}_N(T, f) = -\frac{1}{\beta N} \log \left(\sum \exp\left(-\beta H_N\right) \right) \tag{2.2}$$

where the sum is over all the possible pairs of N-step SAW's Γ^1 and Γ^2. The phase diagram can be studied by looking at the average fraction number of contacts[56] Θ and at the z-component of the average end point separation $\langle r^z \rangle$ defined as:

$$\Theta = \frac{\langle N_c \rangle}{N} = -\frac{\partial}{\partial \epsilon} \mathcal{F}_N, \quad \langle r^z \rangle \sim \frac{\partial}{\partial f} \left(\mathcal{F}_N N \right). \tag{2.3}$$

In the denaturated regime Θ should vanish whereas it should be nonzero in the zipped phase. To compute averages, for fixed values of f and N, Monte-Carlo simulations of the model have been performed by a Metropolis based hybrid algorithm[75, 76] that consists of local and global (pivot moves) deformations attempted on each strand of the polymer. In addition a move that tries to interchange the position of pairs of zipped (or unzipped) portions of the double chain had been considered.

While, as it is defined, the b-model with complete self-avoidance in $d = 3$ can be approached only by means of numerical schemes, we can define a simplified model, in which mutual avoidance is still worked out exactly, but self-avoidance is only accounted for partially. In this model the two strands of DNA are modelled by two directed self and mutually avoiding walks (see next Section and Fig. 2.3). This simplified model has the virtue of being solvable exactly. The main features of the SAW model introduced here will be shown to be present also in the simplified model. hence we will thus consider it as our guide model in order to derive results explicitly. Discussions will follow after each of the results to assert their limits of validity. Results for the SAW model will be given only when different than those for the simplified model.

So in the next Section we will find the exact phase diagram in the directed walk case. We will find that the equilibrium phase diagram displays a re-entrant region at low T: for a finite range of forces the molecule gets unzipped by *decreasing* the temperature. A simple and general physical reason for this counterintuitive feature of our results will be given. The dynamics of both the b- and the Y-models in the various phases and on the phase boundary are then studied, in the next section, by starting from a non-equilibrium bound state as the initial condition. By using Monte-Carlo dynamics, we find that in all regimes above or on the phase boundary, the time evolutions of the order parameters follow dynamical scaling laws. The basic features of both statics and dynamics are maintained for any d. Exceptions occur for the b-model and are noted below.

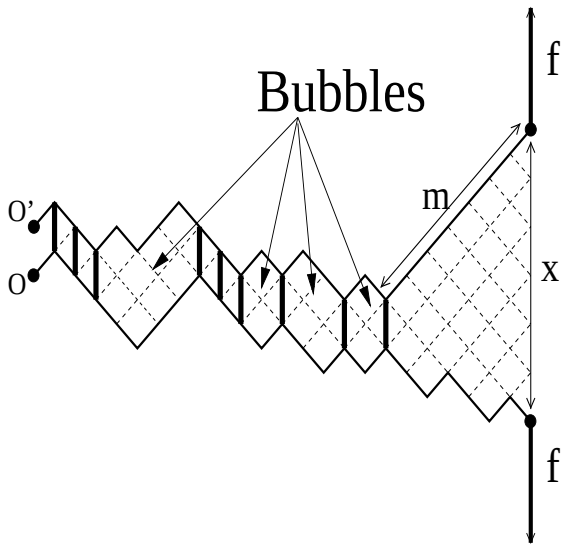


Figure 2.3: The figure shows a typical configuration of the two strands in the directed walk model. Denaturation bubbles are shown, together with the definition of the quantity m (number of liberated monomers) and x (open end distance).

2.2.1 Statics

We now derive the solution for the model with directed self-avoiding walks shown in Fig. 2.3. We consider the $d = 2$ case for simplicity. Generalizations to $d > 2$ are easy and will be discussed below. In $d = 2$ on the square lattice (see Fig. 2.3), the two walks are forced to follow the positive direction of the diagonal axis $(1, 1)$ (*i.e.* the coordinate along this direction always increases). The force acts along the transverse $(-1, 1)$ direction (the x-direction). By measuring the x separation in unit of the elementary square diagonal, we say as in Eq. 2.1 that two complementary monomers are in contact when $x = 1$ (recall that a binding energy -1 is gained for each contact). Due to the geometrical properties of the lattice, all these contacts contributing to the energy involve monomers labelled by the same base pair index, as one would require for base pairing in DNA.

The partition function of the two n-step chains, one starting at $(0, 1)$ and the other at $(1, 0)$, at temperature $T = \beta^{-1}$ (with the Boltzmann constant $k_B = 1$) and under a force f is $\mathcal{Z}_n(\beta, f) = \sum_{x \geq 1} d_n(x) \exp(\beta f x)$, where $d_n(x)$ represents the fixed distance partition function, *i.e.*, the sum over all interacting pairs of directed chains whose last monomers are at distance x. The model can be asymptotically solved

by locating the singularity closest to the origin of its related generating (or grand partition) function[69, 77, 78] (see also Chapter 1):

$$\mathcal{G}(\mathsf{z},\beta,f)=\sum_{n=0}^{\infty}\mathsf{z}^n\mathcal{Z}_n(\beta,f), \tag{2.4}$$

where z is the fugacity.

The quantity $d_n(x)$ in the b-model is seen to obey the recursion relation:

$$d_{n+1}(x)=\left[2d_n(x)+d_n(x+1)+d_n(x-1)\right]\left[(e^{\beta}-1)\delta_{x,1}+1\right], \tag{2.5}$$

together with the boundary conditions:

$$d_n(0)=0 \qquad \forall n, \tag{2.6}$$

$$d_0(x)=\delta_{x,1}. \tag{2.7}$$

In order to find the two competing singularities of the grand partition function, it proves useful to define the generating function of $d_n(x)$, i.e.:

$$d(x,z)\equiv\sum_{n=0}^{\infty}d_n(x)z^n \tag{2.8}$$

From Eq. 2.5, one finds that $d(x,z)$ obeys the following equations:

$$d(x,z)=\delta_{x,1}+z\left[(e^{\beta}-1)\delta_{x,1}+1\right]\left[2d(x,z)+d(x+1,z)+d(x-1,z)\right]. \tag{2.9}$$

One can solve Eq. 2.9, by the following ansatz on $d(x,z)$:

$$d(x,z)=A(z)\lambda(z)^x, \tag{2.10}$$

with $A(z)$ and $\lambda(z)$ to be determined starting from Eq. 2.9. One obtains two conditions from Eq. 2.9, respectively when $x=1$ and when $x>1$. These allow one to solve for $A(z)$ and $\lambda(z)$, finding in this way:

$$\lambda(z)=\frac{1-2z-\sqrt{1-4z}}{2z} \tag{2.11}$$

$$A(z)=\frac{1}{z\left[1-ze^{\beta}\left(2+\lambda(z)\right)\right]}. \tag{2.12}$$

One can then insert the explicit form of $d(x,z)$ into Eq. 2.4 and find the explicit value of the grand partition function, $\mathcal{G}(\mathsf{z},\beta,f)$. In particular one can find the two competing singularities of $\mathcal{G}(\mathsf{z},\beta,f)$, which allows an asymptotic solution of the directed walks model. The first such singularity reads:

$$\mathsf{z}_1=\sqrt{1-e^{-\beta}}-1+e^{-\beta}, \tag{2.13}$$

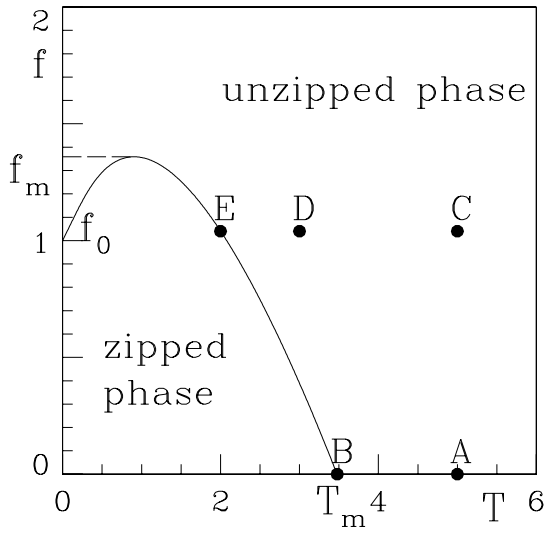


Figure 2.4: Phase diagram of the exactly solvable directed walk model. It is also a good schematic phase diagram for the more complicated SAW model, in that the general features are retained.

and depends only on β, while the second one is found to be:

$$z_2 = \frac{1}{2\left[1+\cosh(\beta f)\right]} \tag{2.14}$$

and also depends on the external force.

When the smallest singularity is the latter, the molecule is in the unzipped phase, otherwise it is in the zipped phase. By equating these two singularities we get the critical line, $f_c(T)$, separating the two phases. For the b-model one obtains $f_c(T) = T\cosh^{-1}\left[p(\beta)-1\right]$, with $p(\beta) = 1/(2z_1)$ (see Fig. 2.4). For the Y model, we get a similar $f_c(T)$ but with $p(\beta) = e^{\beta}$.

For both the models, the critical force at $T = 0$ is $f_0 = 1$. There is a maximum threshold $f_{\rm m}$ such that for $f > f_{\rm m}$ the system is always in the unzipped state. For the b-model, $T_m = \frac{1}{\log 4/3}$ and $f_{\rm m} = 1.358806...$, while $T_m = \frac{1}{\log 2}$ and $f_{\rm m} = 1.282143...$ for the Y-model. The average fraction of contacts is zero when $f > f_c(T)$, but non-zero for $f < f_c(T)$.

Two other quantities of interest are the average open-end-separation $\langle x\rangle = \frac{\partial \ln Z_N}{\partial \beta f}$, and the average number of "liberated" monomers (*i.e.* from the last contact to the end), m (See Fig.1a.). Both these quantities take nonzero values only for $f > f_c(T)$.

The transition line is first order everywhere for both the models *except for the b-model at* $f_c(T) = 0$ in $2 \leq d \leq 5$: this can be seen by noticing that the average number of contacts on the transition line is finite in all but these cases. For $f_{\rm m} > f > f_0$ the usual zipping transition is present but if the temperature is still lowered the two strands cleave again through a "cold unzipping".From the free energy of the exactly solvable model, one finds that the average number of liberated monomers, m, and x both diverge near criticality like[16]:

$$\langle x \rangle \sim \langle m \rangle \sim \frac{1}{f_c - f}, \tag{2.15}$$

so suggesting that $\langle m \rangle / N$ and $\langle x \rangle / N$ are scaling functions of the argument $N(f_c - f)$.

The above models and exact results are useful to understand reentrance in a more general and model independent way. Let us consider a dominant low T configuration of DNA where only the first $N - m$ monomers are zipped and the remaining are completely stretched by the force f.

The free energy of this dominant configuration is simply:

$$\mathcal{F}(m, N) = -\left[afm + (N - m)\right] - \left[(N - m)cT\right], \tag{2.16}$$

where c is the entropy per bounded unit and $a = 1$ in our model. The minimum of the free energy switches from $m = 0$ (completely zipped) to $m = N$ (completely unzipped) when the force crosses the critical threshold $f_c = (1 + cT)/a$: this is the phase boundary at low T which shows the emergence of reentrance. This argument is intuitively appealing and clarifies that reentrance is due to the entropy loss that the DNA strands undergo during the unzipping at low T: in this regime they denature from a double-stranded self-avoiding walk with extensive entropy into two single-stranded stretched walk which at low T basically have entropy 0. At higher T, on the other hand, the two single strands gain mobility and after a crossover T their entropy exceeds that of the single double stranded self-avoiding walk. At the same time, the simplicity of the argument suggests that reentrance is model independent being due to the entropy loss discussed above. When comparing with realistic DNA experiments, however, we should point out that the value of c is far from being universal, rather it depends on the stiffness of the double stranded DNA, which in order to mimic the real situation should be much higher than that of the single-stranded DNA. This fact might confine reentrance to regimes which are not accessible experimentally: In other words the effect is theoretically present and robust to model details, while the ratio T/T_m up to which it persists is extremely sensitive to all details. To give an idea of the typical values, if we call $T_{\rm max}$ the temperature at which $f_c(T)$ attains its maximum value,

we have $T_{\max}/T_m \sim 0.2 - 0.3$ according to the relative stiffness of single-stranded and double-stranded DNAs with the directed walk model, while the ratio is slightly higher for the undirected SAW model in Fig. 2.2 (around $0.3 - 0.4$). This argument thus suggests that with a proper choice of solvent and double stranded polymers, this reentrance should be observable.

Note that in the continuum approximation of Refs.[16, 17, 61] this effect is not found since the validity of those continuum models does not extend to the low temperature regime. This is due to the fact that two gaussian chains at high force poorly represent two DNA strands under large tension, because the Gaussian chain bonds are infinitely extendible. Two FJC's (see Chapter 1) are a better choice and they do show the correct low T behaviour, namely that f_c increases with T as $T \to 0$. Remarkably, the reentrance character is even more enhanced and one finds $f_c(T) - f_c(0) \sim -T \log T$ for $T \to 0$. The most realistic model would however consider each of the two strands as a semiflexible polymer, which could be done by modelling each of the strands by a worm-like chain. On the one hand, it is clear from what we have said that the critical force will again rise at low T. Though one cannot solve this model unless with drastic mean field approximations, it can be found that the exact low T behaviour for two WLC's is different than that of the two FJC's: this time one obtains a still steeper increase and $f_c(T) - f_c(0) \sim T^{1/2}$.

2.2.2 Dynamics

We now consider the dynamics of the b-model and the Y-model. In both cases, we start from a non-equilibrium initial condition with the two chains zipped in a zig-zag configuration (as at zero temperature), and let the system evolve at a temperature T and under a force f, with T and f chosen so that the equilibrium state is either on or above the phase boundary. Let us first consider the two-dimensional case. The five regimes considered are marked A-E in Fig. 1b. Numerically, a Monte-Carlo dynamics with one-bead local move (discretized Rouse model) is used to monitor the time evolution of m and x previously introduced. We accept or reject the moves according to a probability given by the Boltzmann factor as in the standard Metropolis algorithm. We reject configurations where the strands cross. In the Y-model, recombination of the unzipped part is not allowed.

In all cases we find the following dynamical scaling laws to hold good[79]

$$m(t) = N^{d_1} G_m\left(\frac{t}{N^{z_1}}\right) \sim t^{\theta_1},\ (\theta_1 = \frac{d_1}{z_1}), \tag{2.17}$$

$$x(t) - x(0) = N^{d_2} G_x\left(\frac{t}{N^{z_2}}\right) \sim t^{\theta_2},\ (\theta_2 = \frac{d_2}{z_2}), \tag{2.18}$$

(omitting the average signs for simplicity of notation). Here N is the length of each chain and $G_{m,x}$ are two scaling functions. Eqs. (2.17,2.18) also define the exponents $d_{1,2}$ and $z_{1,2}$ for the two variables with $\theta_{1,2}$ describing the early time evolution away from saturations. Note that $d_{1,2}$ can be obtained through equilibrium considerations as one requires $m(t) \sim N^{d_1}$ and $x(t) \sim N^{d_2}$ for $t \to \infty$. The crossover to the equilibrium behavior is described by the "dynamic" exponents $z_{1,2}$.

The exponents are obtained from simulations by collapsing (see Fig. 2.5) the Monte-Carlo data according to Eqs.(2.17,2.18)[80]. The results are summarized in Table I.

The dynamics displays two different time scales mirrored in the difference of the exponents $z_{1,2}$ in regimes A and C. One time scale ($t \sim N^{z_1}$) quantifies the time necessary for the unbinding (or unzipping) of the bases while the other ($t \sim N^{z_2}$) gives the time needed to open (and to stretch whenever $f \neq 0$) the two chains up to their equilibrium open-end-separation. At $T < T_m$ the two processes are virtually the same, because the unbinding (or unzipping) is dragged by the stretching. However, above T_m, the processes decouple and the unbinding gets faster, being controlled by the temperature, yielding $z_1 < z_2$. Furthermore, in the numerical calculations we found large sample-to-sample fluctuations, thereby requiring a huge number of runs ($\sim 10^4$) to reduce statistical fluctuations in $m(t)$ and $x(t)$. This is due to the long time correlation that exists in the system, which keeps samples with different initial histories far apart for any t. Apart from these general considerations, it is possible to explain the exponents for the Y-model found numerically.

Regime A: $f = 0, T > T_m$: Above the critical temperature the dominance of the entropy implies that at every time step one base pair breaks, yielding a linear behavior with $\theta_1 = 1$ and $d_1 = 1$. Also $x(t)$ tends to increase, up to its equilibrium value $\sqrt{N}$ (a fact reflected in the upward derivative at $x = 1$ of the equilibrium probability distribution). This suggests that the dynamics of this quantity should be in the same universality class of the one-dimensional Kardar-Parisi-Zhang equation[79], and so $\theta_2 = 1/3$, $d_2 = 1/2$, and $z_2 = 3/2$.

Regime B: $f = 0, T = T_m$: In this regime $\theta_1 = \frac{1}{2} = z_1^{-1}$ because at criticality the probabilities to increase and to decrease m are expected to be equal, so that m

performs, roughly speaking, a random walk in time with reflecting boundaries at $m = 0, N$. Also for the open-end-separation, steps toward larger or smaller values of x are equally probable, and therefore the time evolution of x should be in the universality class of the one-dimensional Edward-Wilkinson equation[79]. Hence, $\theta_2 = \frac{1}{4}$ and $z_2 = 2$.

Regime C: $f > 0$,$T > T_m$: The strands tend to stretch along the direction of the pulling force. However, once we have pulled the two chains up to an open-end-separation x, to increase x further by one unit we first need to move all the stretched part, which would take a time typically of order x. In other words, one has $x(t+t_0) \sim x(t) + \frac{t_0}{x}$ and we expect the dynamical exponent θ_2 to be $\frac{1}{2}$ ($\theta_1 = 1$ as before and $d_{1,2} = 1$).

Regime D: $T < T_m$,$f > f_c(T)$: Here the only microscopic mechanism for opening the fork is through the applied force: the strands must stretch completely in the vicinity of the bifurcation point and only at this point will the fork liberate one more monomer. Thus, $x \sim m$, and, using arguments as done for regime C, $\theta_{1,2} = \frac{1}{2}$, $d_{1,2} = 1$. In ref.[60] it was found, in a mean field approach for a model resembling our Y-model, that the time necessary to unzip the two strands completely is of the order of N^2. This is consistent with our analysis, but works only in this regime.

Regime E: $T < T_m$,$f = f_c(T)$: On the phase boundary, one expects that the cost for unzipping and zipping is the same (the equilibrium probability distribution of having m monomers unzipped or an open-end-separation equal to x is flat), so that $x(t+t_0) = x(t) \pm \frac{t_0}{x}$ with equal probability. Therefore, the open-end-separation makes a random walk in the rescaled time $\frac{t}{x}$ so that $x \sim (\frac{t}{x})^{\frac{1}{2}}$ implying $\theta_{1,2} = \frac{1}{3}$. Moreover, $d_{1,2} = 1$ since at coexistence there is a finite fraction of liberated monomers. Another way of obtaining $\theta_1 = \frac{1}{3}$ is to demand that a kink liberated at the fork needs to diffuse out of the end before the next one is released. In other words, the rate of change of m is determined by the diffusion of a kink over a distance m. The latter time-scale being of order m^{-2}, we expect $dm/dt \sim m^{-2}$ which gives $m \sim t^{\frac{1}{3}}$.

Turning to the b-model with bubbles, the dynamical exponents in regimes B, E and D with $T \leq T_m$ are the same as found in the Y-model. This establishes that at $T < T_m$ not only for statics, as we saw previously, but also for the dynamics, bubbles are not relevant in the scaling properties. At $T \geq T_m$, on the other hand, the opening of bubbles heavily affects the base unpairing process, unlike the Y-model case where bubbles are forbidden[1]. The length of the unzipped part in the present case now can change by $\pm l(t)$, where $l(t)$ is the typical length of bubbles, and the motion of the fork point can by no means be approximated by a simple random walk (and so θ_1

[1]Regime B is a borderline case: bubbles heavily affect the evaluation of m but do not change θ_1.

$Regime$	d_1	z_1	θ_1	d_2	z_2	θ_2
$A:Y$	1	1	1	$1/2$	$3/2$	$1/3$
$A:b$	1	$4/3$	$3/4$	$1/2$	$3/2$	$1/3$
$B:Y,b$	1	2	$1/2$	$1/2$	2	$1/4$
$C:Y$	1	1	1	1	2	$1/2$
$D:Y,b$	1	2	$1/2$	1	2	$1/2$
$E:Y,b$	1	3	$1/3$	1	3	$1/3$

Table 2.1: "Dynamic" and equilibrium exponents for the Y-model (Y) and the b-model with bubbles (b) as defined in Eqs.(2.17, 2.18). The regimes A,B,C,D and E are those shown in Fig.1.

changes as shown in Table I). The quantity $x(t)$ instead has a dynamics in the b-model similar to the fork case and indeed θ_2 is the same for both the models in all regimes. We show in Fig. 2.5 the collapse leading to $\theta_1 = \frac{3}{4}$ in regime A.

An important question is the dependence of our results on dimensionality (let us consider only θ_1 for simplicity). As for the Y-model, the arguments we gave above for θ_1 for regimes A-E suggest that there be no d-dependence. For the b-model, instead, the d-independence should be true only in regimes D and E for $T < T_m$, where the Y-model gives the exact result; at $T \geq T_m$, on the other hand, bubbles play a dominant role and so we expect a dependence on d. We confirmed this picture with some calculations on a simpler model which should be in the same universality class of the one under study: that of a single random walk, pinned at the origin by an attractive interaction and subject to a stretching external force. In this system m is defined as the number of monomers from the last visit to the origin to the end of the walk. For $T > T_m$, in regimes A and C, our calculations show that the exponent θ_1 increases as dimension increases, and we conjecture that in these regimes:

$$\theta_1 = \frac{3}{4} \qquad \mathrm{d} = 1, \tag{2.19}$$

$$\theta_1 = \frac{1}{2} \text{ with log } corrections \qquad \mathrm{d} = 2, \tag{2.20}$$

$$\theta_1 = \frac{d}{4} \qquad \mathrm{d} > 2. \tag{2.21}$$

Just at criticality at zero force, instead, we find that the exponent θ_1 is very close to $\frac{1}{2}$ in any d. The emerging picture of robust results for $T < T_m$ and model-dependent dynamics for $T \geq T_m$ would be preserved even if, in the original models, the direct-

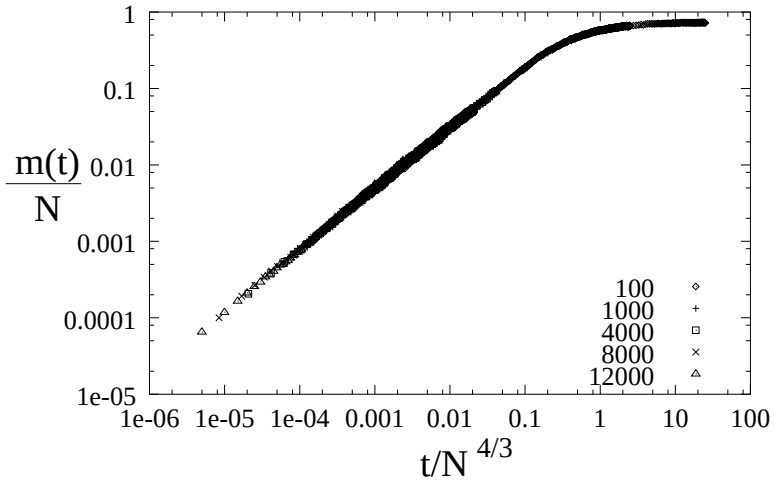


Figure 2.5: Plot of $\frac{m(t)}{N}$ vs $\frac{t}{N^{\frac{4}{3}}}$ for various values of N (regime A, b-model): the collapse of all curves indicates that $d_1 = 1$, $z_1 = \frac{4}{3}$ and $\theta_1 = \frac{3}{4}$.

edness constraint is relaxed (as for the statics). In particular in regimes A and C with SAWs we get that θ_1 is slightly bigger than 1, whereas in $d = 2$ we find $\theta_1 \sim 2/3$, but in that case local moves are known not to be ergodic (the chains may get trapped) even for a single SAW so this could render the evaluation of the exponents less easy.

The arguments presented so far could be generalized to add other ingredients of dynamics also. An example is the nonlocal effects in dynamics, as is the case *e.g.* in the Zimm model. In ref.[60], the author suggests that, in the regime we call D, in the mean field approximation, non-locality could speed up the opening so that $\theta_{1,2}$ would be $\frac{1}{1+u}$ instead of $\frac{1}{2}$ (u is the exponent characterizing the length dependence of the mobility as defined, e.g., for the Zimm model, see [60] for the notation and [41] for a review). In regime E, we can combine our arguments with the same reasoning to get $\theta_1 = \theta_2 = \frac{1}{2+u}$. As we expect, the Rouse model results are obtained with $u = 1$.

2.3 Unzipping random DNAs: the effect of disorder

The above analysis can be extended to binary disorder in the contact potential, which is of course a realistic feature to be included in the model, *i.e.* the energy of the contact of the i-th base may be $\epsilon_i = \epsilon + \Delta_i$, Δ_i being a random variable with binary distribution (equal to $\pm\Delta$) and zero mean. The Y-model offers a good starting point for the study of the effects of heterogeneity, because the critical line of the quenched model can be proved to be the same as that of the pure model with energy ϵ. By quenched

disorder we mean that in principle the free energy should be obtained as follows: first it should be calculated as $\log Z$, if Z is the partition function, for all realizations of the binary disorder, and only at this level can the average over the sequence distribution probability be taken. This is contrasted to the *annealed* average, in which the averaging is done at the level of the partition function, so that the logarithm of the average of Z is taken instead of the average of the logarithm of Z. Of course *quenched* averages are the most difficult to perform, whereas the annealed system is virtually equivalent in its properties to the pure system.

A model analogous to our Y model was proposed in Ref. [17], where the scaling analysis of Eq. 2.15 was generalized to consider sequence quenched disorder. It was found that Eq. 2.15 generalizes to:

$$\langle m \rangle \sim \langle x \rangle \sim \frac{1}{(f_c - f)^2}, \tag{2.22}$$

as $f \to f_c^+$. This can be understood as done in Ref. [17, 65] in a simple way. If we call $\mathcal{F}(m)$ the free energy density of a configuration with m (out of N) unzipped monomers, then within the Y model (we expect this to be true also in the b-model sufficiently below the melting region) one has:

$$-N\beta\mathcal{F}(m) = \sum_{i=1}^{N-m} \frac{\Delta_{N-m}}{T} + (N-m)\frac{\epsilon}{T} + m\log(1+\cosh(\frac{f}{T})) + N\log 2 \tag{2.23}$$

where we take Δ_i's to be uncorrelated. Now, the partition function Z is found via $Z = \int dm \exp -\left(\frac{\mathcal{F}}{T}\right)$. One can see that in $\mathcal{F}(m)$ m appears in the combination $m(f_c - f)$, near $f \sim f_c$. So if $\Delta_i = 0$ for every i we recover Eq. 2.15. Now, with sequence heterogeneity, the typical random contribution to $\mathcal{F}(m)$ is of order $\sqrt{\Delta m}$; the random part exceeds the average contribution $m(f_c - f)$ for $m \stackrel{\sim}{<} \frac{\Delta}{(f_c-f)^2}$, which is a length scale which diverges more rapidly than $1/(f_c - f)$, suggesting that there should be a crossover to the scaling in Eq. 2.22 for $f_c - f \sim \Delta/T$.

We now come to the role of randomness on the dynamics. We expect on general grounds that $P(m,t)$, the probability of having m monomers unzipped at time t, will obey a master equation with transition rates $W_{\pm,m}$ depending on the realization of disorder, *i.e.* we expect that $W_{\pm,m} \propto \min\left\{e^{\mp\beta N \frac{d\mathcal{F}(m)}{dm}}, 1\right\}$, where from Eq. 2.23 $N\beta\frac{d\mathcal{F}(m)}{dm} = +\frac{\Delta_{N-m}}{T} + \frac{\epsilon}{T} - \log(1+\cosh(\frac{f}{T}))$ [2], which can be seen to consist of a

[2] At $f = 0$, $N\beta\frac{d\mathcal{F}(m)}{dm}$ displays also a subleading (for large m) entropic contribution which in $d = 2$ is $-\frac{1}{2m}$. This should not affect the mapping proposed in the text except perhaps just at the critical temperature T_m.

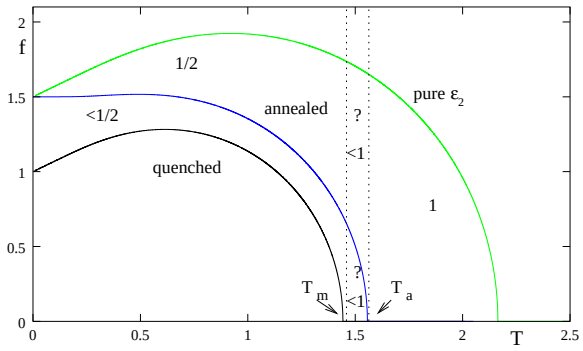


Figure 2.6: Plots of the phase diagram for homo-DNA (top) made up of the more attractive base ($\epsilon_2 > \epsilon_1$), and for the annealed (middle) and quenched (bottom) disordered models. The numbers refer to the values of the exponent θ_1 in the various regions of the phase diagram, which we believe to hold on the basis of the mapping onto Temkin's model. Note that above the annealed critical line at $T < T_m$ and everywhere at $T > T_a$ the pure system results are recovered. Interestingly, the annealed line does not show the re-entrant behavior found in Section II and in the quenched system. "?" indicates the regime where we do not have numerical evidence in support of our results. We took $\epsilon_2 = 1.5$ and $\epsilon_1 = 0.5$ in the calculations.

zero-mean random "noise" ($\frac{\Delta_{N-m}}{T}$) plus a "bias". We thus believe that the dynamics of the Y-model can be mapped onto Temkin's model of a random walker in a random environment [47] (provided that at $T < T_m$ we rescale time as described above in regimes D and E), so that, following Ref. [81], there will be a region around the critical line, at $T < T_m$, in which $m, x \ll t^{\frac{1}{2}}$ (and just on the critical line at $f \neq 0$ one has $m, x \sim \left(\log\left(\frac{t}{x}\right)\right)^2$). The quantities m, x are now quenched averages over realizations of disorder.

The scenario we propose is sketched in Fig. 2.6. The curve bounding the region where, at $T < T_m$, disorder should be important for the dynamics has been found by applying the criterion of Ref. [81] to the system with the above transition rates. This curve coincides with the critical line of the static model with annealed disordered. Runs at very low T suggest that as $T \to 0$, for a given realization of disorder and with any force between ϵ_1 and ϵ_2 (the two binding energies), the unzipping will take place as far as the first more attractive base pair is found. Besides, the probability that the monomer of index m_0 will ever be unzipped decays exponentially: $P(m_0) \propto 2^{-m_0}$. This is consistent with the mapping suggested above.

So far, we have considered the effects of randomness on scaling properties: the fact that base pair heterogeneity can in some instances change the critical behaviour

of this single molecule phase transition is certainly interesting for a theoretical physicists, and can also lead to new experiments aiming at confirming the values of all the exponents predicted here. We should not forget, however, what the original aim of biophysicists was when proposing single molecule DNA unzipping experiments: that of sequencing a DNA by looking at the features of the force versus displacement curves. In order to make our models apter to study the theoretical feasibility of this project, it would be necessary to work in the ensemble conjugate to the fixed force ensemble in which we are now considering. This would create some technical problems because Monte-Carlo simulations with constraints are very hard to perform. Moreover, the two ensembles are equivalent in the thermodynamic limit, but for finite DNAs with single realizations of disorder (i.e. no quenched averaging) the force vs. x curves are highly different[22, 65]. A good basis to start working on the feasibility of sequencing DNA is to consider the b-model in Fig. 2.3 for which exact transfer matrix methods can be used (see Fig. 2.7). This model was used in Ref. [22] to verify the feasibility of a) mutation detection, b) sequence determination from the unzipping force on a theoretical level, i.e. neglecting other difficulties coming from the experimental setup as those discussed in Ref. [21]. The analysis in Ref. [22] was also confined to the case in which dynamics effects (important in the analysis in Ref. [23]) are not an issue. The numbers coming from this model suggest that the detectability of a point mutation would be possible in principle with present day technology if the point mutation is sufficiently near the point we are unzipping at, whereas if it is far apart the point we start the unzipping, the entropy of the single strands renders the identification not reliable. Other factors limiting resolution are chain length and temperature or pH. The situation for sequencing is very different: even neglecting the apparatus, the conditions for complete sequencing with present-day reachable resolution suggested that it would be possible in principle only for unrealistically short chains or low temperatures. Nevertheless, in very beautiful experiments such as that in Refs. [20, 7] cited above, coarse-grained information about the local sequence (CG-vs-AT content), are possible, and base pairing of patches of ten bases or so might be recognizable in the near future.

* * * * * * * * * * * * *

We close this Chapter by highlighting a possible relevancy to the biology of DNA replication of our results on the unzipping dynamics. The observed power-law and scaling behaviours in Eqs.(2.17,2.18), imply that in our model there is scale invariance. As it is generally accepted[79], this could be due to the presence of strong temperature driven fluctuations or of long range time correlations (necessary to build

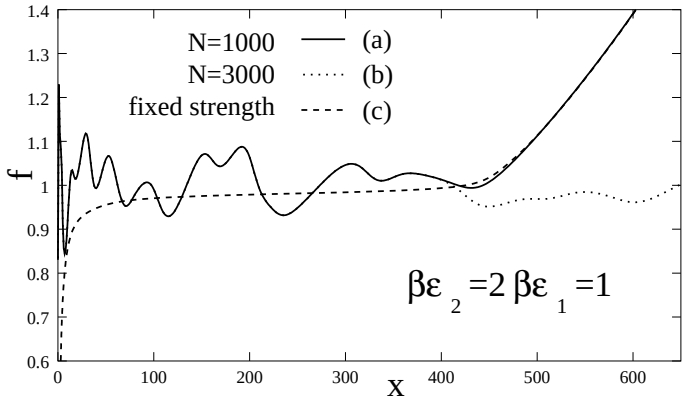


Figure 2.7: The force vs stretching distance curves for heterogeneous DNAs (b-model). The sequences are chosen randomly but both share the same sequence (1000 ϵ's) from the open, pulled end. For the fixed stretch ensemble, curves (a) and (b), the pattern is identical over a region of x. Curve (c) is the fixed force ensemble phase coexistence curve with finite-size effect.

up a cooperative mechanism) in the unzipping. In a replication process, it is expected that fluctuations in unzipping, after its initiation at the origin (Step I), would affect the binding of the next set of proteins (Step II). A proper theory, so far lacking, needs to couple these two steps. We note that the difference in the dynamical scaling at the thermal denaturation point and on the unzipping phase boundary can in principle provide a natural selection mechanism on the basis of which Step II might not initiate a replication after an occasional zero-force fluctuation induced bubble formation, but would rather wait for the correct force-induced unzipping to start. However, at this point, this remains speculative.

Chapter 3

Stretching of a polymer below the Θ point

In this Chapter, we study a physical phenomena which intuitively looked very similar to DNA unzipping, namely the force-induced unfolding of proteins or homopolymeric collapsed chains (below the Θ point, or in a bad solvent). Also the first treatments of these subjects (see Refs. [82, 83, 84]) were based on a mean field like description and a first order phase transition was predicted together with a general phenomenology (e.g. d independence) similar in many aspects to that detailed in the preceding Chapter. We show here that mean field is for several aspects a poor approximation to the present phenomena. Other more experiment oriented theoretical papers have also been written recently[85, 86].

We begin the Chapter by discussing the experimental and theoretical knowledge from existing literature and the open questions. The following Sections are then mainly focused on two theoretical models aimed at understanding the unfolding behaviour of polymers in a poor[40, 87] solvent (globules), i.e. below the theta temperature, T_θ. One is the familiar model for θ transition in $d = 2$ on the lattice, where a lot is known when there is no force. The main reason for restricting to the lattice is the wealth of exact numerical methods which can be used in this case. The other model is that of a self-avoiding chain in the three-dimensional continuum space subject to a homopolymeric attractive potential and to a stretching force. For both the models, we will first study the ground states, then go into the thermodynamic properties. Finally, we discuss the behaviour of heteropolymers, which better mimic protein stretching experiments, and we comment on the dynamical behaviour, highlighting which differences and similarities with DNA unzipping we expect here.

3.1 Experimental and theoretical knowledge

Unlike for DNA unzipping, the response to stretch of homopolymers in a poor solvent is relevant also to real experiments. For example in case one wants to describe the stretching of collapsed DNA molecules[13], and of other polymeric globules such as those studied in the experiment of Ref.[14], one can restrict a realistic model to one in which the interactions between monomers are all the same. In order to study protein unfolding, on the other hand, such as the problems arising from titin stretching[11] the heteropolymeric nature of polypeptides is vital and has to be taken into account. This will be discussed at the end of the Chapter.

There is a vast number of experimental results in the literature. In particular, various force vs elongation (f vs x) curves have been recorded in experiments studying stretching phenomena in a poor solvent [13, 14]. We stress once more (see also preceding Chapter) that in a typical experiment the end-to-end distance is controlled, and the average force needed to do so is measured, so that the natural ensemble in which to do calculations would be the fixed stretch ensemble.

In most cases a curve consisting of three distinct regimes (namely first linear response for low stretches, then a characteristic plateau for intermediate stretch[13, 14], and finally a sharp increase when the contour length of the polymer is approached) have been observed (see e.g. the DNA chain response in Fig. 3.1a). In a few examples, for somewhat shorter DNA molecules, a stick-release pattern with hysteresis has been found (see Fig. 3.1b, curve B). Other, still different curves, with only the linear response followed by the plateau, have also been reported, but occurred less frequently (see Ref. [13] for details).

The first observation is in good agreement with the mean field theory proposed in Ref.[82], and the plateau strongly suggests the presence of a first order phase transition (see also the recent measurements in [14]). In the plateau region, the typical configuration in the thermodynamic limit in the experiment ensemble is constituted by two stretched chains of variable extension with a globule in the middle.

Our results suggest that there might be more than one possible shape for the f vs. x curves according to d, the spatial dimension and to polymer length, so that a different behaviour occurs when mean field theory[1] is qualitatively incorrect.

It is to be noted that the study of these models has a remarkable interest even on a purely theoretical ground. First, the numerical study recently performed in Ref.

[1] Note that we use 'mean field theory' in a broader sense, including approaches which consider the relative stability of the globule and the coil only (see also [84]). Mean field theories should be all equivalent as regards universality.

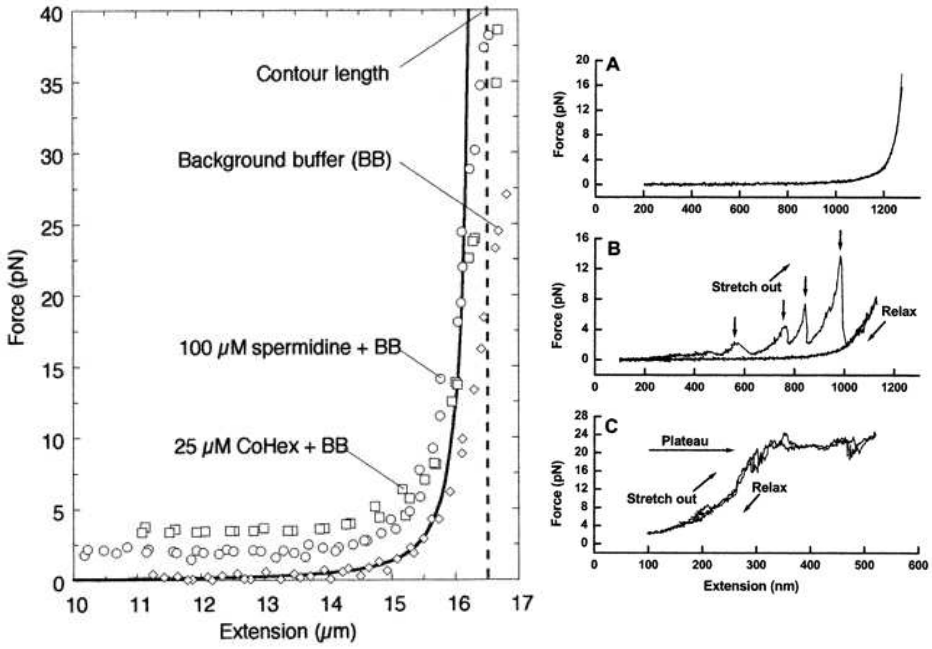


Figure 3.1: (Left) Response of single λ phage DNA molecules to an applied force with condensing concentrations of the trivalent cations CoHex and spermidine, taken from Ref. [13]. The unstretched molecule is thus in its collapsed state, and one can see that the force vs. extension curves display a plateau for $x \sim 11 - 14$ μM in good agreement with mean field theory. The curve for the uncondensed molecule (background buffer) is also shown for comparison. (Right) Curves obtained when stretching plasmid-length DNA. The spermidine concentration for curves B and C is > 200 μM, from Ref. [13]. (A) Uncondensed molecule force-extension curves (WLC like). Force-extension curves for condensed DNA molecules show two distinct behaviors. (B) Most of the curves show a stick-release pattern. During the relaxing phase, the DNA shows no stick-release behavior, but the F-x curve displays hysteresis. (C) A few DNA tethers show a force plateau behavior during both stretch and release. The F-x curve displays a plateau at a force of ~ 20 pN. This figure has been taken from Ref. [13].

[88] has suggested the possibility that the force-induced unfolding transition is second order in $d = 2$ and first order in $d = 3$. This has been confirmed to some extent in a study of the model on a hierarchical lattice with fractal dimension two[89]. The $d = 2$ case is important as it is below the upper critical dimension for Θ collapse and mean field predictions may well be incorrect. A thorough analysis and a clear physical mechanism underlying the difference of the nature of the transition as d changes are the first theoretically minded purpose of this Chapter.

Second, the mean field analysis of Ref.[84] has suggested there could be a re-entrant region in the phase diagram for low temperature similar to what happens for DNA unzipping as was explained in Chapter 2. However the exact results in[89] prove mean field is not valid in $d = 2$. Then, our other aim here is to describe theoretically the unfolding transition of globules not relying on the mean field approximation.

In what follows, we first characterize the evolution of the ground states of a finite polymer as the pulling force increases. This will help to understand the thermodynamics. In particular, we compute the phase diagram in the temperature–force (T, f)-plane in $d = 2$ on the lattice, where we can use the transfer matrix method (Fig. 3.4) together with exact enumerations (Fig. 3.2).

3.2 Ground states in $d = 2$ and in $d = 3$

We begin by considering a self-avoiding walk (SAW) on the square lattice with *fixed* origin. The model partition function (generating function) in the canonical ensemble in which T and f, the stretching force, are fixed is:

$$Z_N(f, T) = \sum_{\mathcal{C}} e^{-\beta H(\mathcal{C})} = \sum_{b,\mathbf{R}} w_{o\mathbf{R}}(N, b) e^{\beta(b\epsilon + f R_x)} \tag{3.1}$$

where N is the number of monomers (including the origin) of the SAW, $H(\mathcal{C})$ (referred to a configuration $\mathcal{C}$, i.e. to a SAW) is the energy of a SAW, b is the number of pairs of neighboring occupied sites not adjacent along the chain, $\beta \equiv \frac{1}{T}$ is as usual the inverse temperature, $w_{o\mathbf{R}}(N, b)$ is the number of configurations of a SAW with fixed origin o and end-to-end distance $\mathbf{R}$, of length N and b contacts, R_x is the projection along the force direction (x axis) of $\mathbf{R}$. We take as usual both the Boltzmann constant and ϵ equal to one.

When T is low, one may look for the ground states among the rectangles of sides L_x and L_y that are completely covered by the SAW (in other words such that $L_x L_y = N$, where we neglect the small effects arising when this rectangle cannot

be constructed with both $L_{x,y}$ integers). The energy of this rectangular Hamiltonian walk with a non zero f is:

$$-H(N \equiv L_x L_y, L_x) = N - L_x - \frac{N}{L_x} + 1 + f(L_x - 1). \tag{3.2}$$

The minimum of $H(N, L_x)$ for given N with respect to L_x yields the most stable configuration for various values of f. The minimum occurs for an f-dependent value of $L_x \equiv L_x^0(N, f)$, namely:

$$L_x^0(N, f) = \sqrt{\frac{N}{1-f}}. \tag{3.3}$$

For any $f < 1$ one thus has a compact configuration. However, when the critical value $f = 1$ (for $T = 0$) is reached all integer values of L_x from N (stretched coil) to $L_x \sim N^{1/2}$ (compact globule) become degenerate for large N. Note that this does not hold in $d = 3$ where it is well known that there is a Rayleigh instability in the thermodynamic limit[14, 82, 88]. This can be seen by comparing the globule energy - which in $d = 3$ is $2N$ in the large N limit - with the energy of a parallelepiped, with elongation along $\vec{f}$ equal to L_x and with edges $L_y = L_z$ in the perpendicular plane. The force above which (for $T = 0$) the parallepiped is a better ground state than the three-dimensional globule is $2L_y$ ($L_y \ll N^{1/3}$), and there is no longer any degeneracy at the critical force $f_c = 2$ (at $T = 0$).

In Fig. 3.2a we sketch the situation in $d = 2$. The minima hierarchy, shown in the shaded area in the top panel, affects the low T region of the $< x >$ (average elongation) vs. f curves for finite length (bottom panel). However, only one transition survives in the large N limit and represents a true phase transition (as represented in Fig. 3.2a by the shaded wedge ending in just one point in the $N^{-1} = 0$ axis). Similarly, when T is raised the multi-step character of the $< x >$ vs. f curves is lost due to fluctuations which blur the ground state dominance in the partition. In Fig. 3.2b we show the analogous picture for a $3d-$model discussed below. Short chains in $d = 2$ and 3 behave similarly whereas in the infinite length limit the $d = 3$-case shows an abrupt unfolding transition.

Let us now discuss the $d = 3$ case. The model we used is the freely jointed chain (FJC) (see e.g. [87, 41, 40]) in the continuum (off lattice). The FJC is subject to a compacting pairwise attractive potential between non-consecutive beads and to a stretching force $\vec{f}$ at the extrema. The pairwise potential is chosen to be an asymmetric square well with a hard core radius, $2R_{hc}$, which acts between non-consecutive beads along the chain, and an attraction range $R_1 > 2R_{hc}$, i.e. the distance up to

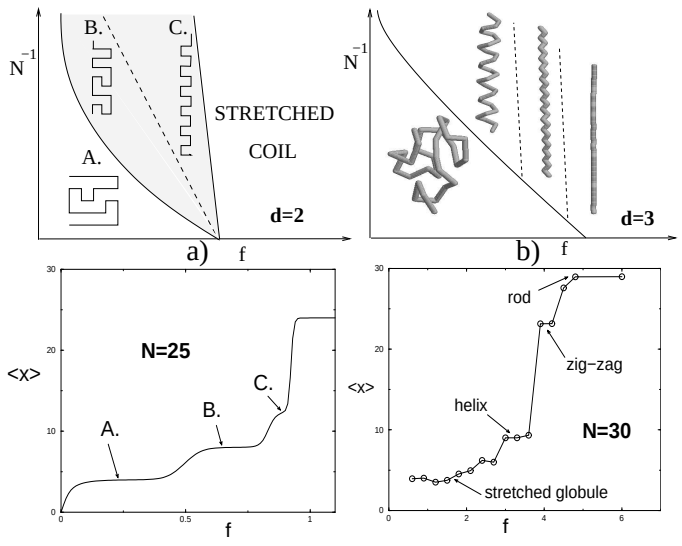


Figure 3.2: a) $d = 2$ Schematic diagram of minima for polymers of different sizes (top) and low $T < x >$ vs. f curve for $N = 25$ (bottom) found with exact enumerations in $d = 2$. b) d=3: Same as in a), except that the $< x >$ vs. f curve (bottom) is for $N = 30$ and is found by simulated annealing.

which the particles interact. We have checked that the results reported in the following do not appreciably depend on the two-body potential details. The parameters we have used to generate the configurations shown in Fig. 3.2b (top panel) are $R_{hc} = 0.6$, $R_1 = 1.6$ where the unit length is the distance between successive beads along the chain.

The ground states of short chains (up to $N = 30$ bead long) has been determined by performing simulated annealing employing Monte-Carlo dynamical simulations. The FJC is evolved dynamically by means of three sets of moves: the pivot, reptation and crankshaft moves (see Appendix A and Ref. [43]). We lower T during the simulation according to a standard annealing schedule as detailed in Chapter 1. We found evidence also in this $d = 3$ case that the unfolding of a finite length homopolymer proceeds in a multistep fashion. The collapsed globule first orients itself along the pulling force as soon as there is a nonzero f. At larger f the globule is slightly elongated (much less than in the $d = 2$ case) and after this a helix forms followed by a zig-zag curve and finally by a stretched coil. This succession of minima, shown in fig. 3.2b is intriguing for a two-fold reason: firstly because it suggests that helices, one of the well known building block of proteins, come out rather naturally as one of a few stable minima of a homopolymer in a poor solvent subject to a finite stretching

force; secondly because the unfolding transition of a finite polymer in this model appears to be markedly different from a globule-to-coil (two-state) transition. The mean field picture of an all or none transition is recovered for infinite length. The situation is depicted in Fig. 3.2b. Since in $d = 3$ the transition is first order, the mean field treatment is correct in the thermodynamic limit.

The fact that helices become better ground states than compact globules could be easily verifiable in experiments made in the fixed force ensemble. These can now be confidently performed, with not more substantial difficulties than the more conventional ones, performed in the fixed stretch ensemble. Indeed the stick-release pattern reported in Ref.[13] (Fig. 3.1b) shows several different peaks in the f vs. x curve, and is compatible with a 'multi-step transition' in which the globule undergoes more than one conformational changes (signalled by the peaks) during the unfolding. Intriguingly, we recall that this pattern is reported for smaller polymers, a fact which would be in agreements with our findings; whereas for large N this effect is much less important and the curves have a single (dominant) plateau. In order to give this observation a more quantitative basis, we have taken 30 points from the data in Fig. 3.1b (curve B) and calculated numerically the free energy as a function of x, $\mathcal{F}(x)$, by integrating the average force $f(x)$, according to the well-known thermodynamic relation $f(x) = \frac{1}{\beta}\frac{\partial \mathcal{F}(x)}{\partial x}$. The resulting free energy is shown in Fig. 3.3 and shows regions of different convexity. From the free energy we can calculate the average elongation vs. f curve, which we would obtain in our calculations. We do not know what value of β is appropriate for the case at hand, but one can clearly see that when β is high and entropy does not play too big a role, the $\langle x \rangle$ vs f curve displays a distinct multi-step behaviour. Though the agreement is only qualitative and the free energy obtained in the experiment might not correspond to the equilibrium free energy but might rather have dynamical effects in it, we consider the agreement with our picture remarkable. In passing, we note that this also shows that the experimental fixed stretch ensemble contains more information than the fixed force ensemble, as it would not have been possible to re-obtain the fixed stretch ensemble force vs. x curve from our fixed force results. This is completely analogous to what was observed in the previous Chapter.

Helices appear as ground states for a potential consisting of a force term and a hydrophobic contact potential because they are both elongated and offer a good shielding from the outside solvent to the monomers lying in its interior. Note also that helices arise as optimal states of tubes of non-zero thickness subject to compaction (see Ref. [90] and the second Part of this thesis). If a worm-like chain is used instead of the FJC, the picture should not change, because the elongated states such as the helices and the zig-zag are stiffer than the collapsed globule (sharp corners between

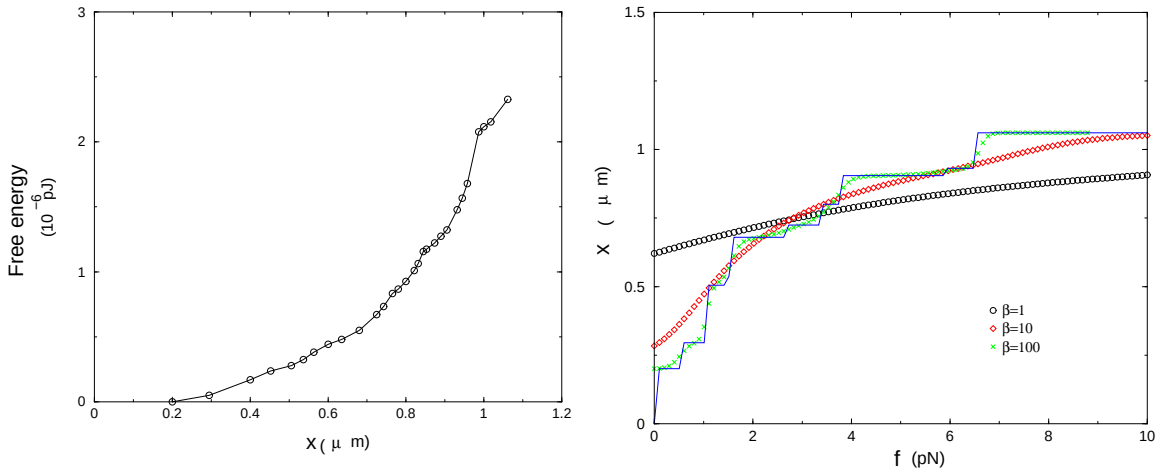


Figure 3.3: a) (Left) Plot of the free energy $\mathcal{F}(x)$ as a function of the stretch for the data of Fig. 3.1 b, curve B. b) (Right) Plot of $\langle x \rangle$ (with same data) for various values of β. For high β entropy is not important and the multi-step character inherent in the experimental situation shows up. The solid line refers to the case $\beta \to \infty$.

successive monomers are necessary in this state) and so should be even more favoured as f is increased.

3.3 Thermodynamic properties

We now turn to the thermodynamic behaviour of the SAW model on a square lattice. We use the transfer matrix (TM) technique, following Refs.[91, 92, 93].

Let us introduce briefly the principal features of the TM approach: the partition function of a polymer of N sites is given by Eq. (3.1), with $f = 0$. In the thermodynamic limit ($N \to \infty$) we expect that $Z_N(T) \sim [\mu(T)]^N$, then the free energy per monomer $\mathcal{F}$ is simply, $\mathcal{F} = -T \log \mu(T)$. It is more convenient[91] to introduce the following *generating function* (z is the step fugacity)

$$g_{o\mathbf{R}} = \sum_{N,b} z^N e^{\beta b} w_{o\mathbf{R}}(N, b) \tag{3.4}$$

It is known that for $z < z^c(T) = 1/\mu(T)$, the inverse SAW connectivity, $g_{o\mathbf{R}} \sim \exp(-R_x/\xi(z,T))$, where $\xi(z,T)$ is the *correlation length* and R_x is the projection of $\mathbf{R}$ along x. We study the stretching of an interacting SAW in a strip of finite size L along y and infinite length along x. It is possible to define[91] an L-dependent

correlation length $\xi_L(z,T)$ *via* the formula:

$$\xi_L(z,T) = -\frac{1}{\log \lambda_L(z,T)}, \tag{3.5}$$

where $\lambda_L(z,T)$ is the *largest* eigenvalue of the transfer matrix, that equals 1 at $z = z_L^c(T)$. We apply the *phenomenological renormalization*, to find successive estimates for $z_c(T) = \lim_{L\to\infty} z_L^c(T)$. Including the force via Eq. (3.1), the equation for the critical force $f_c(T)$ is then ideally found via:

$$f_c = -T \lim_{z\to z_c(T)^-} \lim_{L\to\infty} \log \lambda_L. \tag{3.6}$$

The order of the limits in Eq. (3.6) and a correct choice of the boundary conditions (see below) are crucial. For example if we take $\lim_{L\to\infty} \lim_{z\to z_c(T)^-} \log \lambda_L$ one would trivially get $f_c = 0$ so that at any non-zero force the polymer would be stretched. Similarly, from Ref. [40] one might be tempted to generalize the scaling behaviour $\langle x\rangle \sim f^{1/\nu-1}$ to collapsed polymers finding in this way again there is no finite force collapsed phase. But this is not true for infinitely long collapsed polymers, as seen e.g. in the experiments[2].

In Fig. 3.4 the *phase diagram* for the stretched interacting SAW is shown. With the TM, a right choice of the *boundary conditions* is needed[91, 92, 93]. We have used both *periodic* (PBC) and *free boundary conditions* (FBC). PBC have been employed to get the best estimate of $z_c(T)$ through phenomenological renormalization. This value is then used with FBC, to find the correct L-dependent critical eigenvalue $\lambda_L(z_c(T),T)$. Finally, adopting the extrapolation algorithm of [95], $\lim_{L\to\infty} \lambda_L(z_c(T),T)$ is obtained which, through Eq. (3.6), allows to get the phase diagram. As expected, in the case of FBC, there are oscillations in data going from odd L to even L. As usual in this context, a separated analysis of even and odd L data was necessary for obtaining a better convergence (see Fig. 3.4). One point on the transition line obtained previously in [88], is recovered here.

One can get an approximate description of the transition if one requires that the globule and coil phases coexist. The globule free energy is easily estimated in terms

[2] In Ref. [94] it was also suggested that there is no transition as force increases, in a different model, that of a directed self-avoiding walk in $d = 2$. This model is solvable with high precision with the procedure used here for the undirected case. In our view, the result in Ref. [94] is due to numerical errors which prevent the author from realizing that the $f = 0$ singularity persists up to some $f_c > 0$. Indeed we have made extensive transfer matrix calculation (data not shown in the thesis) which fully support our view that the critical fugacity is $z_c(f = 0)$ and decreases only below $f_c(T)$, which can be found analytically.

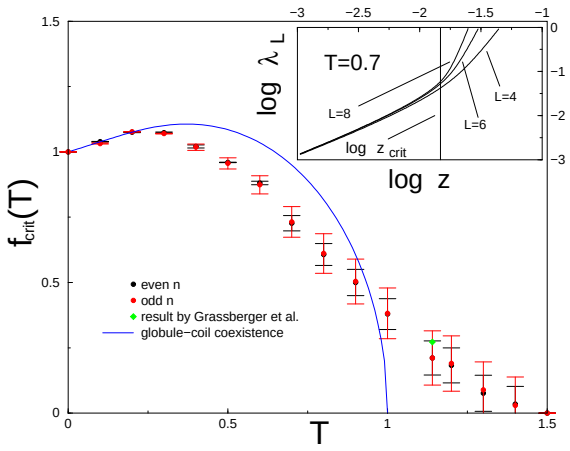


Figure 3.4: Phase diagram for the stretching of a SAW on a $2d$-lattice, obtained with the TM technique. Inset: plot of $\log\lambda$ vs. $\log z$ for $T = 0.7$.

of Hamiltonian walks[87]. On a square lattice the energy is simply given by minus the length of the polymer whereas the entropy is given in terms of the number of Hamiltonian walks which grows exponentially with N[87]. Thus the globule free energy per monomer is $\mathcal{F}_g = -1 - T\log(4/e)$ where we have used the accurate mean field estimate of the entropy as given in [96]. The coil free energy $\mathcal{F}_c$ is approximated by that of an unconstrained random walk in presence of a pulling force and contacts are neclected. We thus get $\mathcal{F}_c = -\log(2(1+\cosh(\beta f)))$. At coexistence one finds $f_c(T) = T\cosh^{-1}(2\exp(1/T-1)-1)$ (the continuos curve in Fig. 3.4).

We note that $f_c(0) = 1$ is the exact result and at low T the phase diagram displays a reentrant region. As $T \to T_\theta$, f approaches 0 rather smoothly. We can give an argument to predict the value of the exponent a defined via:

$$f_c(T) \sim (T_\Theta - T)^a \qquad T \to T_\Theta. \tag{3.7}$$

The free energy $\mathcal{F}$ behaves, near the point $(f = 0, T = T_\Theta)$, as:

$$\mathcal{F} \sim f^{1/\nu_\Theta} \qquad T = T_\Theta, f \to 0^+ \tag{3.8}$$

$$\mathcal{F} \sim (T_\Theta - T)^{2-\alpha=1/\phi_\Theta}, \tag{3.9}$$

where $\nu_\Theta, \phi_\Theta, \alpha$ are the critical exponents characterizing the Θ collapse. By matching the two scaling laws one gets $a = \frac{\nu_\Theta}{\phi_\Theta} = \frac{4/7}{3/7} = \frac{4}{3}$, a value compatible with our data.

Within the TM approach one can also infer the order of the transition. To do this we observe that, if for $z \to z_c(T)^-$:

$$\lambda(z_c(T), T) - \lambda(z, T) \sim (z_c(T) - z)^\Delta, \tag{3.10}$$

then $\Delta < 1$ ($\Delta = 1$) means a second (first) order transition, as $< x > \sim (f - f_c)^{\frac{1}{\Delta}-1}$ for $f \sim f_c$. Our data at not too low T are compatible with a second order transition (inset of Fig. 3.4), with $\Delta \sim 0.6$, though its evaluation is difficult with the data we have. By noting that for an infinitely long stretched polymer with $\langle x \rangle / N \to 0$ de Gennes's scaling must hold, so that $\langle x \rangle / N \sim (f - f_c)^{1/\nu(f_c)-1}$, we obtain that Δ is the critical exponent of the polymer at the critical force below the Θ point. This is particularly interesting because it implies that Δ is a new critical index of a polymer, to be added to the ones already known. It is most easily seen as a generalization of ν_Θ at non-zero force: at $T = T_\Theta$ the polymer is neither compact nor swollen, here at f_c below T_Θ it is neither open nor compact (as from our numerics we get $\Delta > 0.5$). Of course for $d > 2$ or at $T = 0$ one has coexistence due to the first order transition and thus $\Delta = 1$.

Inspired by the encouraging result obtained by exact renormalization group (RG) on the Sierpinski lattices in Ref. [89], we now give an argument which aims at explaining in a theoretical way the reason of the change in character of the transition from $d = 2$ to $d = 3$. We can obtain the approximate RG equation in $d-$ dimensional lattices, in a way similar to the one outlined in Chapter 1, though here the situation is more cumbersome, having to deal with an interacting SAW. We defer to Ref. [98] for the detailed derivation which here we skip. The RG recursions relations can be written for the generating function representing polymers traversing a hypercube of linear size 1 once, A, and 2^{d-1}-times, B (Fig. 3). The terms A and B represent parts of the chain which are in the coil and globular state respectively. The recursion relations can be calculated as in Ref.[97] or by enumerating the SAWs on 2×2 or $2 \times 2 \times 2$ cells as in [98]. To leading order in A at $0 \neq T \ll T_\theta$

$$A' = A^2 + \alpha(d) A^2 B^2, \quad B' = \gamma(d) B^{2^d}, \tag{3.11}$$

where $\alpha(d)$ and $\gamma(d)$ are $d-$dependent constants. There are three fixed points in the flux in Eq. 3.11: $A = 0$, $B = B^* = \gamma(d)^{-\frac{1}{2^d-1}}$ corresponds to the globular phase, $(A, B) = (1, 0)$ to the coil phase, while the last non-trivial fixed point $A = A^* = \frac{1}{1+\alpha(d)(B^*)^2}$, $B = B^*$ characterizes the unfolding transition. Wee display in Fig. 3.6 the flux of the equations in Eq. 3.11 in $d = 2$.

The value of the term $\alpha(d)$ affects the behaviour of the RSRG flux near the fixed point (A^*, B^*). One can see[3] that $\alpha(d) \neq 0$ if $d = 2$ and is 0 in $d > 2$. When $\alpha(d) \neq 0$ (i.e. in $d = 2$), the RG flux is smooth near (A^*, B^*) and the critical fugacity

[3]These facts can be proved exactly in a family of Sierpinski gaskets (see [89] for an example), where d is the fractal dimension. On conventional lattices, an approximate RG treatment as in [98] gives qualitatively the same behaviour.

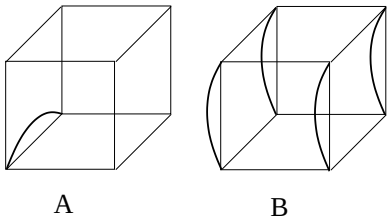


Figure 3.5: Representation of the generating functions A and B in the text (for $d = 3$) at a generic level of iteration in the RG. Thick lines represent portions of the SAW that enter through a vertex and go out from another one.

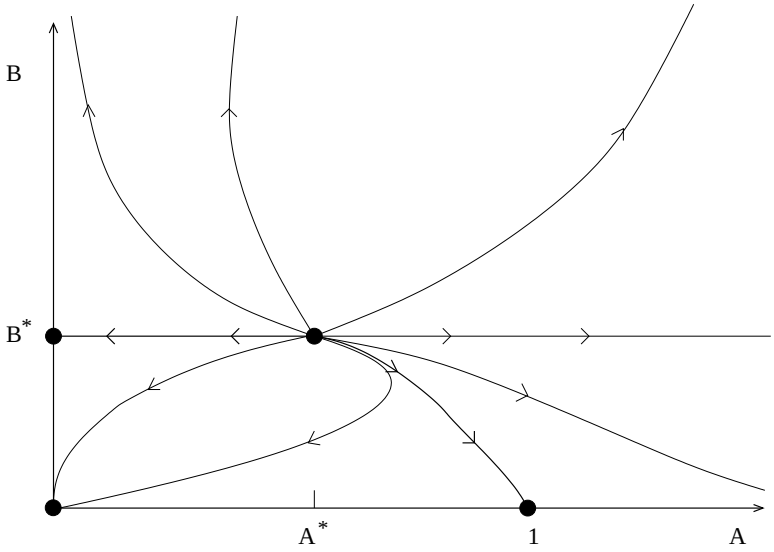


Figure 3.6: Plot of the flux lines resulting from the simplified RG equations in Eq. 3.11

near $f = f_c(T)$ behaves as $z_c(f_c(T), T) - z_c(f, T) \sim (f - f_c(T))^2$ signalling a second order transition with $\Delta = 1/2$ in Eq.3.10. On the other hand when $\alpha(d) = 0$ (i.e. $d > 2$), the transition is first order and two-state like. The presence of the mixed term in Eq. (3.11) is crucial and enhances the entropy of the coil phase since it contributes to A'. Consequently, the $d = 2$ two-state approximation in Fig. 3.4 gives a transition line which is higher than the numerical result for $0 \neq T \ll T_\theta$. The entropy gain in the stretched coil, as $T \to 0$, is hampered as it costs a finite surface energy (dominant as $T \to 0$) to change locally an elongated globular region into a coil and vice-versa. This is why the solid curve in Fig. 2 matches our numerics as $T \to 0$.

In particular, by taking into account all the hierarchy of stretched states (shaded region in Fig. 3.2a), we find that the $T \to 0$ limit of the critical force is given by

$f_c(T) \sim \min_{L>0}\left(1 + TL\log\frac{\mu_{HAW}}{\mu_{HAW;L}}\right)$, where μ_{HAW} and $\mu_{HAW;L}$ are the connective constants of the Hamiltonian walks without constraints and confined in a strip of length L respectively. Our $T \to 0$ limit in Fig. 3.4 takes into account the terms up to $L = 8$ and so is strictly speaking an upper bound to the real $T \to 0$ slope.

3.4 Sequence specificity, proteins & dynamics

Up to this point, the discussion was valid for *homopolymers* such as the DNA molecules of Fig. 3.1. If we consider real proteins, then this is no longer a good approximation and one needs to introduce heteropolymer models. The simplest would be one in which monomers can be either hydrophopic (H) or polar (P) and the energy of two neighbouring monomers can take three different values according to whether the sites in contact are HH, HP or PP. This is known as the 'HP' model and is a well respected model for proteins on the lattice. For more than two (say n) kind of 'aminoacids' in the protein sequences, the contact energy is best viewed as a matrix (with $n(n+1)/2$ different entries). Irrespective of the model we choose, we observe that as $T \to 0$ all proteins should reach their folded state, which is unique (up to trivial symmetry transformations). This means that, at least in our lattice models, there should be no reentrance as there is no entropy loss in the unfolding. In continuum models, e.g. for the FJC or WLC, on the other hand, one has nevertheless an entropy loss in stretching a string and reentrance should be present although relegated to very low values of T (see the corresponding cases in the previous Chapter). The finite T behaviour is expected to be richer. Interestingly, as we will also see in the second Part of this thesis, there are basically two model behaviour for the folding of proteins. Some proteins have a 'folding' transition to the unique ground state which is coincident with the Θ collapse, whereas others below some temperature, T_θ form a collapsed globule with very little resemblance to the native state, and only at some lower temperature, T_F, they go into the unique folded state. For this second kind of proteins, the unfolding behaviour will depend on T: when T is between T_F and T_Θ we expect multi-step behaviour to occur for finite samples as for homopolymers, whereas below T_F one gets an all-or-none transition. In $d = 2$, the order of the transition might also change, being likely to be second order for $T_F < T < T_\Theta$, and first order for $T < T_F$.

This picture is confirmed to some extent by exact enumerations which have been performed on $16-$mers with unique ground state: for low T there is no reentrance and the transition is basically two-state (data not shown). Also the numerics in Ref. [85] is consistent with this picture, showing a first-order like transition and no re-entrancy.

In Ref. [84] this was given as an example of re-entrancy in protein models, however both the formula cited by the authors and the phase diagram show this is not the case. There is indeed a small dip in the force vs T critical line, which is however referred to a finite sample, and is most probably due to the trivial symmetry degeneracy of the ground state discussed above, which does not cause entropy to be non-zero at low T.

We close this Chapter by mentioning an interesting point which deserves further work. The unfolding dynamics of a homopolymer at infinite size in $d = 3$ should closely resemble, as regards the scaling laws, the dynamics of homo-DNA unzipping, in that e.g. for $f > f_c$ one has $\langle x(t) \rangle \sim t^{1/2}$, while at criticality the dynamic exponent should get $1/3$. For finite samples, however, the $1/2$ behaviour will be recovered only for strong enough forces, above those which yield the helix as the ground state. So in a sense the finite size samples might curiously more closely follow the heterogeneous case for DNA unzipping. Further work is needed to test this hypothesis. As a successive step one could search for the differences in protein unfolding dynamics, which most likely will display a closer similarity to the homogeneous case for infinite samples as for statics.

****

In this Chapter, we have discussed the unfolding transition of a homopolymer under the action of an external pulling force in $d = 2$ on the lattice and in $d = 3$ off lattice. A ground state analysis shows that for finite length polymers, the unfolding is not abrupt, rather it occurs via a multi step sequence of states. These are more elongated than the globule and make more contacts than a coil. In $d = 3$ helices arise naturally as ground states at intermediate forces. For infinite polymers, on the other hand, the situation is different: in $d = 3$ the intermediate ground states disappear due to the Rayleigh instability, and the transition is effectively two-state, whereas in $d = 2$ they survive in the thermodynamic limit. Indeed, from the transfer matrix results, it is apparent that the mean field hypothesis is incorrect in $d = 2$ even at rather low T: indeed the transition is second order as also found in Ref. [88] and not first order[84] although the prediction of the re-entrant region[84] agrees with the TM results. Furthermore a renormalization group based argument is in agreement with this picture.

PART 2: From swollen to compact

Some molecules are so long and skinny that they act like strands of spaghetti. In this experiment we will see how the long, skinny molecules called polymers can sometimes behave the same way.

Kathleen A. Carrado, in ’ChemShorts for Kids from 1993’.

Chapter 4

A model for thick polymers

4.1 Introduction

The outline of the second Part of the thesis is as follows. In this Chapter, starting from a review of Edwards's well-known polymer model, we present a description of thick polymers that is well defined in the continuum, and we introduce the discrete model with which we make the numerical calculations. This involves three-body constraints as well as two-body interactions in the potential. Then, in Chapter 5 we describe the numerical results that we have obtained when looking for the minima (through simulated annealing) of a short thick polymer under compaction. We show the marginally compact structures that are found when the polymer thickness is tuned with the attraction and compare them with naturally occurring secondary structure in proteins and other biopolymers. In Chapter 6, we study the physics at non-zero temperature of a thick polymer: we do so first by applying the mean field approximation to a chain of coins (first section), as it is more tractable analytically than is a polymer with three-body constraints; then by performing Monte-Carlo simulations in the original model (last section). Technical details are deferred to the Appendixes.

4.2 From Edwards's model to thick polymers

We shall consider here the case in which the interacting particles are restricted to lie on, or close to, a curve embedded in the d dimensional Euclidean space R^d. This system is widely studied in polymer science [40, 41, 99, 100]. Generalization to 2-dimensional manifolds describing random surfaces or membranes have also been recently studied (see [101, 102] and references therein). One of the paradigm of

polymer science is the celebrated Edwards's model which is the continuum version of the FJC with excluded volume interaction introduced in Chapter 1. It describes a polymer in terms of a continuous curve, $\vec{r}(s)$, with an effective energy given by

$$H(\vec{r}) = \frac{1}{2}\int_0^L \dot{\vec{r}}(s)^2 ds + \frac{v_2}{6}\int_0^L\int_0^L \delta(\vec{r}(s)-\vec{r}(s'))dsds' + \\ + \frac{v_3}{90}\int_0^L\int_0^L\int_0^L \delta(\vec{r}(s)-\vec{r}(s'))\delta(\vec{r}(s)-\vec{r}(s''))dsds'ds'' + \cdots \quad (4.1)$$

The first term in the above equation takes into account the chain entropy. It is simply due to the central limit theorem. Indeed one can derive it in the case of a non interacting chain of beads tethered together by a potential $u(r_{j,j+1})$, with $r_{m,n} = \|\vec{r}_n - \vec{r}_m\|$, acting between adjacent beads and keeping them at a typical mean square distance a^2. This means that the probability to find the j-th step joining the positions $\vec{r}_j$ and $\vec{r}_{j+1}$ is proportional to $\exp\{u(\vec{r}_{j,j+1})\}$. Using the central limit theorem one easily derives that the probability to find the first beads at $\vec{r}_0$ and the k-th at $\vec{r}_k$ is well approximated by

$$< \delta(\vec{r}_0 - \vec{r}_k) > \propto \exp\{-\frac{d}{2a^2k}(\vec{r}_0 - \vec{r}_k)^2\} \quad (4.2)$$

for large k. Considering now n pieces of such k-step chains, the probability to have a 'coarse grained' chain with given positions $\vec{r}_{ik}$ for $i = 1, 2, ..., n$ is given by

$$P(\vec{r}_0, \vec{r}_1, \ldots, \vec{r}_n) \propto \exp\{-\frac{d}{2a^2k}\sum_{i=1}^{n}(\vec{r}_{(i-1)k} - \vec{r}_{ik})^2\}. \quad (4.3)$$

In the continuum limit one defines $\epsilon = ka^2/d$ and the total 'length' of the 'coarse grained' chain, $L = n\epsilon$, is kept fixed while $n \to \infty$ (this implies that $a \to 0$). Thus the previous equation represents the definition of the formal expression of eq.(4.1), in the case of $v_i = 0, i = 2, 3, \ldots$, since

$$\lim_{\epsilon\to 0}\sum_{i=1}^{n}(\vec{r}(t_{i-1}) - \vec{r}(t_i))^2/a = \int_0^L \dot{\vec{r}}(s)^2 ds \quad (4.4)$$

where $\vec{r}_{ik} = \vec{r}(t_i)$.
In the self interacting case one again starts from a string of beads as before, where the total energy is given by,

$$H_{chain} = \sum_j u(r_{j,j+1}) + \sum_{m<n} V(r_{m,n}) \quad (4.5)$$

with $V(r)$ being a two-body interaction between any pair of beads at a distance r. A typical shape could be the one of a Lennard Jones potential with both repulsive and attractive part (e.g. $6-12$ shape as in Chapter 1). The connection between eqs.(4.1) and (4.5), i.e. the continuum limit model in eq.(4.5), is made through the virial expansion. The v_m's in eq.(4.1)represent effective m-body interactions for the continuum chain, $\vec{r}(s)$, and they depend on the temperature, T. At high (low) T $v_2 > 0$ ($v_2 < 0$) whereas $v_3 > 0$. The three-body term is strictly necessary only in the low T region in order to stabilize the system. The model defined by eq.(4.1) is widely used in the literature and it has been shown to have a precise meaning in perturbation theory using renormalization group techniques (see [100] and the original references therein) [1]. It is possible to give a heuristic explanation of why the continuum version of eq.(4.5) can only give rise to interaction terms which are the singular (δ-like) potentials present in eq.(4.1). Suppose the potential $V(r)$ in eq.(4.5) has a repulsive part which increases as r decreases. Then in order to take the continuum limit the bead density along the chain increases, because the u potential constrains the successive beads to reduce their mutual distance ($a \to 0$). Consequently the number of bead-pairs laying at a distance corresponding to the repulsive region of the V potential increases since the chain constrains nearby (not only adjacent ones) beads to approach each other as a decreases. This would imply that the continuum chain has an infinite energy and an infinite rigidity. This is avoided only if simultaneously to $a \to 0$ also the repulsive region of the interaction potential V in eq.(4.5) shrinks leading to singular δ -like potentials as in eq.(4.1).

In the following we want to consider models of self-avoiding curves that are non-singular, and which include an explicit thickness length scale. In mathematical terms, starting from a continuous curve $\vec{R}(s)$, $0 \leq s \leq N$, which we will also call γ, and which we assume to be smooth and continuously differentiable to any order, we can define a smooth solid tube or thick polymer centered in γ and of thickness R_0, and

[1] One can formally map the model in eq. 4.1 into a field theory of an n-component field, $\vec{\phi}(\vec{x})$. In the grand canonical formulation with a fugacity per unit length, r, the Hamiltonian of the field theory is given by

$$H_n(\{\phi\}) = \int d^dx \{\frac{1}{2}\partial_\mu \vec{\phi}\partial_\mu \vec{\phi} + r\vec{\phi}^2 + \sum_m \frac{v_m}{m!}(\vec{\phi}^2)^m\} \tag{4.6}$$

Correlation functions for the corresponding polymer model are calculated from the $\vec{\phi}$ correlation in the $n \to 0$ limit [103, 104]. Renormalizability of such field theory is well known and it guarantees that from perturbative expansion finite results can be extracted which depend only on a finite number of parameters.

we call this object $\tau_{R_0}(\gamma)$. The 'tube' $\tau_{R_0}(\gamma)$ is then the union, over all points $\vec{r}$ in γ, of all the circular disks of radius R_0 centered at $\vec{r}$ and contained in the plane normal to γ at the point $\vec{r}$.

In particular, in order to describe a thick polymer, as we will demonstrate below, we will have to replace the singular delta-function potential above by a smooth potential $V(\cdot)$ (such as e.g. a Lennard-Jones potential) that is dependent upon a single scalar variable $\cdot$. Our continuum models will retain the following two desirable features of their discrete version: i) all finite-energy configurations of the manifold are non-self-intersecting, with a positive distance of closest approach (see below), and ii) the explicit form of the potential $V(\cdot)$ provides a characteristic length scale for the distance of closest approach of the manifold. The difficulty in constructing such models is that the standard choice of taking the argument $\cdot$ of the potential to be the Euclidean pairwise distance cannot satisfy our two desiderata when the underlying system is continuous. The basic idea beyond this is simple. For a generic one-dimensional curve such as that depicted in Figure 1, one wishes to penalize true points of closest approach between distinct parts of the curve, *without* penalizing adjacent points from the same part of the curve. The pairwise Euclidean distance simply cannot distinguish between these two cases; in other words, it cannot distinguish between proximity of points that is forced by continuity of the manifold in *any* configuration, and the real phenomenon of interest, namely proximity of points due to large scale geometrical deformation in *some* configurations. One can show[105, 106, 107] that the above two requirements can be achieved simply by taking the argument $\cdot$ to be a quantity other than the Euclidean distance between two points.

We now argue what the argument $\cdot$ should be for a thick polymer. A tube of thickness R_0 is different from a continuous polymer of Edwards's kind because it is not enough that $\vec{R}(s) \neq \vec{R}(s')$ unless $s = s'$ for a curve γ to constitute an acceptable configuration (when inflated) for the corresponding tube of thickness R_0, $\tau_{R_0}(\gamma)$. For this to happen, $\tau_{R_0}(\gamma)$ must also neither have local sharp bends nor non-local self-intersection. A convenient criterion to check whether γ is an allowed 'axis' for $\tau_{R_0}(\gamma)$ turns out (computationally) to be the following: given γ, we inflate it until it is so thick to have self-intersection, at that point we stop and we call $\Delta(\gamma)$ the thickness of the maximally inflated tube, before self-intersections appear. If $\Delta(\gamma) > R_0$, then the curve γ corresponds to a valid configuration of a tube with thickness R_0.

Therefore, we need a working procedure in order to calculate $\Delta(\gamma)$. It is rather simple to see[108] that the thick polymer $\tau_\eta(\gamma)$ becomes locally singular (i.e. self-intersecting) when η is bigger than the *local radius of curvature* of γ at any point $\vec{r}$

in γ. The local radius of curvature of γ in $\vec{r} \in \gamma$, $\rho_L(\vec{r})$, is the radius of the circle which best approximates γ. On the other hand, $\tau_{R_0}(\gamma)$ may also become singular if two portions of it which pertain to values of $\vec{r}(s), \vec{r}(s')$ correspondent to points which are distant along γ, come into close contact with each other. This is a non-local effect, which can be shown[108] to be related to the presence of two non-adjacent points of *closest approach* on γ. A pair of points $(\vec{r}, \vec{r}')$ are said to belong to the set (which we will call Ω) of all pairs of points of closest approach on γ, if $\vec{r}, \vec{r}' \in \gamma$ and if $\vec{r} - \vec{r}'$ is orthogonal to the tangent vectors to γ at both $\vec{r}$ and $\vec{r}'$. One can convince with some thought that it is sufficient to consider these two cases in order to rule out self-intersecting configurations of a thick polymer. As a consequence, the thickness of the maximally inflated tube around γ, $\Delta(\gamma)$, reads:

$$\Delta(\gamma) = \min \left\{ \min_{\vec{r} \in \gamma} \rho_L(\vec{r}), \min_{\vec{r}, \vec{r}' \in \Omega} \frac{|\vec{r} - \vec{r}'|}{2} \right\}. \tag{4.7}$$

In other words, the thickness of the maximally inflated tube is either the minimal value of the local radius of curvature or half the minimum distance of closest approach, whichever is smaller. We can introduce a quantity which encompasses both cases, namely the *global radius of curvature* of γ at $\vec{r}$, $\rho_G(\vec{r})$. The global radius of curvature of a curve γ in $\vec{r}$ is defined in terms of the *triplet radii* $R(\vec{r}, \vec{r}', \vec{r}'')$, with $\vec{r}, \vec{r}', \vec{r}''$ non-coincident and belonging to the curve γ. The quantity $R(\vec{r}, \vec{r}', \vec{r}'')$ (see also Fig. 4.1) is nothing but the radius of the (unique) circle which goes through $\vec{r}, \vec{r}', \vec{r}''$. Note that when $\vec{r}', \vec{r}'' \to \vec{r}$ the triplet radius goes by definition to the local radius of curvature $\rho_L(\vec{r})$. On the other hand, when $\vec{r}'' \to \vec{r}'$, the circle defining the triplet radius is one which contains $\vec{r}$ and is tangent to γ in $\vec{r}'$. We now define the global radius of curvature in $\vec{r} \in \gamma$ as:

$$\rho_G(\vec{r}) = \min_{\vec{r} \in \gamma} \rho_G(\vec{r}). \tag{4.8}$$

Now, the thickness of the maximally inflated tube is given by:

$$\Delta(\gamma) = \min_{\vec{r} \in \gamma} \rho_G(\vec{r}) = \min_{\vec{r}, \vec{r}', \vec{r}'' \in \gamma} R(\vec{r}, \vec{r}', \vec{r}''). \tag{4.9}$$

In order to prove this one first has to realize that the extremality condition in Eq. 4.8 implies that the global radius of curvature is either the local radius of curvature or (if it is smaller) the radius of a circle passing through $\vec{r}$ and tangent in another point $\vec{r}'$. Second, as we have to minimize over $\vec{r}$ as well, the quantity $\min_{\vec{r} \in \gamma} \rho_G(\vec{r}) = \min_{\vec{r}, \vec{r}', \vec{r}'' \in \gamma} R(\vec{r}, \vec{r}', \vec{r}'')$ is either the minimum local radius of curvature, or (if it is smaller) the radius of a sphere which contains no portion of the curve in its interior

and is tangent to the curve at both portions at two diametrically opposite points $\vec{r}$ and $\vec{r}'$. At such points the symmetry property $R(\vec{r},\vec{r}',\vec{r}') = R(\vec{r},\vec{r},\vec{r}')$ holds and then $\vec{r}$ and $\vec{r}'$ belong to Ω. By definition this means that the minimum over $\vec{r}$ of the global radius of curvature of γ at $\vec{r}$ is $\Delta(\gamma)$.

Thus a suitable potential to describe a thick polymer in the continuum may be a hard core constraint on the triplet radii, which forbids any of them (and hence the minimum, thus $\Delta(\gamma)$) from being smaller than R_0. In this way we naturally introduce a three-body term in the continuous potential, which here has a well-defined physical meaning. In general, we may introduce different functions of the triplet radii (such as Lennard-Jones potentials as a function of the above defined $R(\vec{r},\vec{r},\vec{r}')$, $\rho_{G,L}(\vec{r})$, $\Delta(\gamma)$ etc.) as polymer potentials with which to make calculations, in case we want to favour a given value either of the thickness, or of the local or global radii of curvature, or of each of the triplet radii. In any case, the scalar argument $\cdot$ that we were looking for in order to describe thick tubes is the triplet radius (see Fig. 4.1).

When the argument $\cdot$ of $V(\cdot)$ depends as is here the case upon more than simply two points, for example triplets or quadruplets of points (which pertain to the case of surfaces [107], which are not our concern here), we shall refer to $V(\cdot)$ as a many-body or multi-point potential. The use of many-body potentials is an essential ingredient in the models of continuous systems that we propose, and they should not be viewed as a higher-order correction to two-body or pairwise potentials, as is the case in Edwards's model. Indeed our proposal for continuous models is to *replace* pairwise self-interaction potentials, which must be singular, with many-body self-interaction potentials, which need not be singular.

Knotted polymer problems have recently attracted considerable interest among physicists[109, 110, 111]. A feature of our treatment, which is appealing, is that it is the first continuum formulation to our knowledge which allows one to study systems within a knot class, i.e. with a fixed number of knots. A drawback of Edwards's formulation, indeed, as of any two-body potential (with singular potentials like in eq. (4.1)) formulation, is that the self-intersection of the chain is possible though it costs some energetic penalty, so that one can change at will the knotting number and the linking number of the polymer. This is not what usually occurs realistically in the first place. Secondly, many physical problems require estimates on entropic exponents and weights of polymer configurations with a fixed linking number. With this formalism these would become tractable at least in principle from a theoretical point of view, whereas at present they are manageable only numerically.

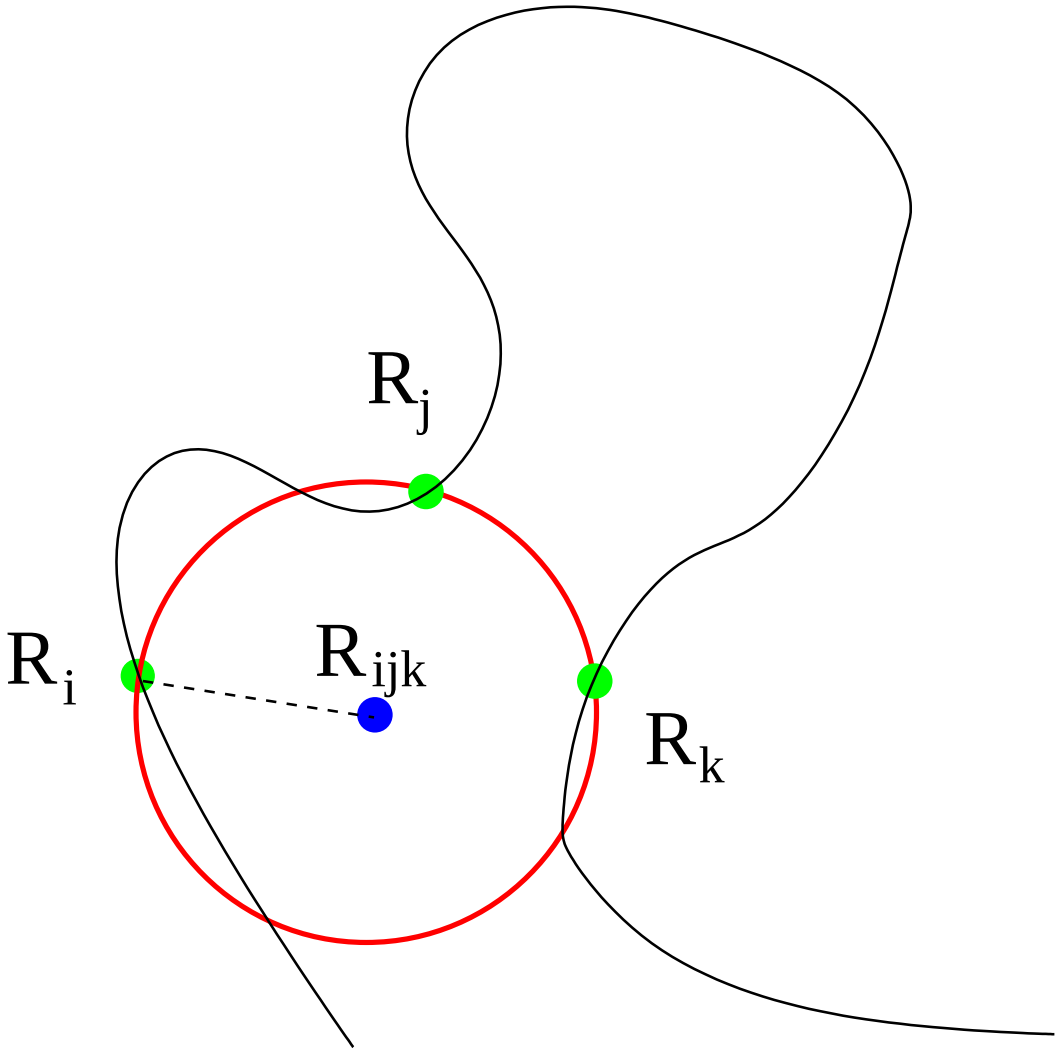
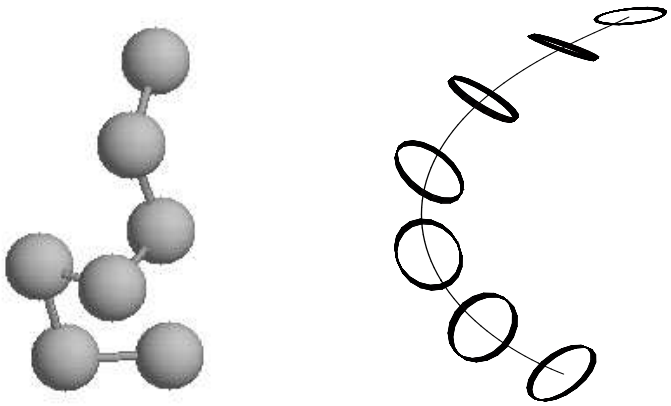


Figure 4.1: This is one of the triplets in the curve which are used to calculate its thickness.



CHAIN OF SPHERES CHAIN OF COINS

Figure 4.2: In this figure we contrast the model of a chain of coins, which should model a discrete thick polymer, to the usual chain of spheres tethered together.

4.3 Discrete thick polymers

We now come to the case of our interest, in which a spaghetto or thick polymer (tube) is meant to mimic the backbone of a protein or of other biopolymers. Due to this, the model we want to consider deals with discrete chains rather than continuous as in the preceding Section.

The discrete model we will consider in the numerical calculation is the following. Two beads at positions $\vec{r}_i$ and $\vec{r}_j$ interact through the two-body potential:

$$V(r_{i,j}) = \begin{cases} \infty & \text{if } r_{i,j} < 2R_{h.c.} \\ -1 & \text{if } 2R_{h.c.} < r_{i,j} < R_1 \\ 0 & \text{if } R_1 < r_{i,j} \end{cases} \tag{4.10}$$

so that R_1 is the attraction range (1.6 in what follows) and $R_{h.c.}$ is the hard core radius, which we take to be 0.6 in our calculations.

Since this is a discrete model, the thickness is no longer uniquely defined. One way to capture it would be the discretization of the definition of $\tau_{R_0}(\gamma)$, i.e. would be to consider at each bead of the chain a 'hard coin' and to constrain the physical configuration of the system to be such that coins attached to distinct beads do not compenetrate. In Fig.4.2 we show how a chain of coins looks like and oppose it to the usual chain of spheres. An alternative choice to capture the non-zero thickness [105, 107] of the discrete chain is to discretize Eq. 4.9, so that we disallow in our

model conformations for which the quantity (discretized $\Delta(\gamma)$):

$$\Delta \equiv \min_{i\neq j\neq k} R_{i,j,k}, \tag{4.11}$$

where $R_{i,j,k}$ is the radius of the circle going through the centers of the beads i, j and k (Fig. 4.1), is smaller than R_0.

These two implementation of the thickness are strictly equal only in the continuous limit, but we expect them to give the same results physically also for discrete chains. For this reason we will choose the most convenient one from time to time. In particular, in the Monte-Carlo simulations of Chapter 6 it is more convenient to use the second procedure, whereas when applying mean field it is easier to deal with the chain of coins.

At this point, let us recall that for us thick polymers are meant to represent a protein C_α backbone better than a conventional chain of spheres. We have described in Chapter 1 the elementary biological facts about proteins. In intuitive terms, we expect the thickness to be due to the steric constraints given by the side chains, which are neglected when putting a hard core radius only on the C_α atoms. Such a fact will also cause the emergence of an anisotropy in the system of which there is no trace in the conventional treatment.

However, this is only heuristics and a more quantitative answer to the question 'Is a protein backbone modellable by a thick polymer ?' must be given. In order to move in this direction, we analyzed the previously defined $\rho_G(\vec{r})$ (suitably discretized as done above for $\Delta(\gamma)$)– which gives a measure of the local thickness at the residue in $\vec{r}$ in the chain – for every residue of 30 unrelated protein native structures taken from the protein data bank ([112]). The result is shown in Fig. 4.3. It is apparent from it that the distribution of $\rho_G(\vec{r})$, or equivalently of local thicknesses, is sharply peaked around a central value, which is roughly $2.7\ \mathring{A}$. This lends credibility, as a suitable zero-th order approximation, to our model in which a protein is approximated by a tube of uniform thickness. Of course a more realistic model would have to take into account fluctuations and hence dishomogeneity in the local thickness of the polymer and also heterogeneity in the two body potential, as pertaining to the different kinds of aminoacids found in nature. We will see in the next Chapter, however, that our model is detailed enough to fully account for protein secondary structures.

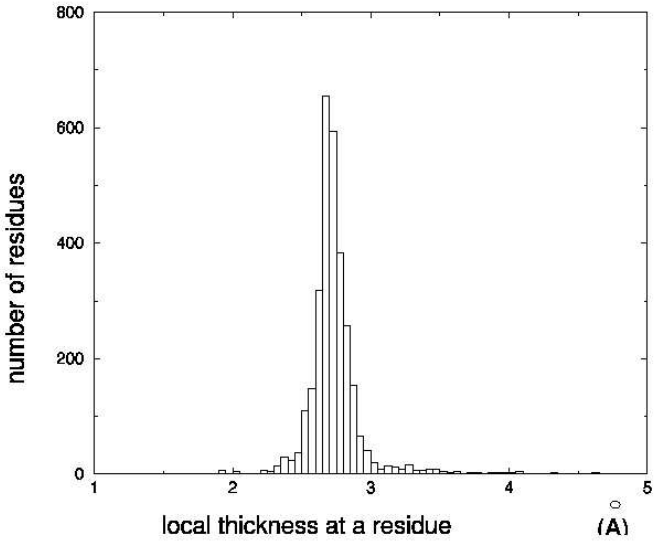


Figure 4.3: Probability distribution of the local thickness per residue for all aminoacids of 30 unrelated protein native structures taken from the protein data bank.

4.4 How stiff is a thick polymer ?

Usually, it is believed that suitable models for biopolymers such as proteins and, most notably, for DNA must take into account the fact that the biomolecule is not fully flexible, but it rather has an intrinsic persistence length or equivalently an intrinsic stiffness, which is an adjustable free parameter in the model (example models are the semiflexible polymer or the WLC defined in Chapter 1). Here we do not introduce a stiffness; however it is clear from intuition that the thickness of the polymer specifies also a non-zero stiffness. Here we want to predict how much this stiffness will be.

Let us thus consider a thick polymer of thickness R_0, and look for its persistence length, ξ (of course the bigger ξ is, the stiffer the polymer is). To get a prediction on ξ, we consider only the simple case in which there is no self-interaction and in which the polymer is in the swollen phase, so that non-local triplets (i.e. other than the triplets such as $\vec{r}_{i,i+1,i+2}$) do not feel the constraint imposed by the thickness. The thickness constraint on neighbouring links $\vec{t}_i \equiv \vec{r}_{i+1} - \vec{r}_i$, $\vec{t}_{i+1} \equiv \vec{r}_{i+2} - \vec{r}_{i+1}$, instead, is the following:

$$\vec{t}_i \cdot \vec{t}_{i+1} \geq 1 - \frac{1}{2R_0^2} \qquad R_0 \geq 1/2 \tag{4.12}$$

$$\vec{t}_i \cdot \vec{t}_{i+1} \geq -1 \qquad \text{otherwise} \tag{4.13}$$

for every $i = 1, \ldots, N-2$. Let us now calculate the average of $\vec{t}_n \cdot \vec{t}_m$, $n, m = 1, \ldots, N-1$, $n > m$ taken with the portion of the chain from bead n to $m-1$ fixed. We obtain:

$$\langle \vec{t}_n \rangle_{\vec{t}_m, \ldots, \vec{t}_{n-1}} = \overline{\cos\theta} \vec{t}_{n-1}, \tag{4.14}$$

where $\bar{\cdot}$ denotes average over the allowed values of $\vec{t}_n \cdot \vec{t}_{n+1} \equiv \cos\theta$. So one gets:

$$\langle \vec{t}_n \rangle_{\vec{t}_m, \ldots, \vec{t}_{n-1}} = \left(1/2 + \left(1 - \frac{1}{2R_0^2} \right) /2 \right) \vec{t}_{n-1}. \tag{4.15}$$

Taking the dot product with $\vec{t}_m$ and averaging over the portion of the chain that before was kept fixed we get:

$$\langle \vec{t}_n \cdot \vec{t}_m \rangle = \left(1 - \frac{1}{4R_0^2} \right) \langle \vec{t}_{n-1} \cdot \vec{t}_m \rangle. \tag{4.16}$$

Eq. 4.16 is a recursion relation to be solved with the initial condition $\langle \vec{t}_n \cdot \vec{t}_n \rangle = 1$. The solution is:

$$\begin{aligned} \langle \vec{t}_n \cdot \vec{t}_m \rangle &= \left(1 - \frac{1}{4R_0^2} \right)^{|n-m|} \\ &= \exp - \left(|n-m| \log \left(\frac{1}{1 - (2R_0)^{-2}} \right) \right), \end{aligned} \tag{4.17}$$

so that the persistence length of a non-interacting thick polymer from Eq. 4.17 is:

$$\xi(R_0) = -\frac{1}{\log\left(1 - (2R_0)^{-2}\right)}. \tag{4.18}$$

It is interesting to apply Eq. 4.18 to some realistic cases. If we want to model a single stranded DNA (ssDNA) as a thick polymer with one bead per base pair, the value of R_0 pertaining to this case would be $\sim 0.7 - 0.8$ (the diameter is roughly one bp, the number given is influenced by discreteness). Consequently, the persistence length is $\sim 1.5 - 3$ bp according to the number chosen for R_0. This is in good agreement with the experimental data for ssDNA[35]. If on the other hand we want to model double stranded, ds, DNA by a thick polymer, we must recall (see Chapter 1) that its diameter is 2.5 nm, so that its thickness R_0 measured in bp's is slightly more than 4 (discreteness effects here change the value less than for ssDNA). From Eq. 4.18, we obtain that for dsDNA ξ would be ~ 70 bp's, which is roughly half of the correct value[35]. Despite this, we deem it interesting that with *no* adjustable parameters we obtain that ξ is 1 to 2 orders of magnitude bigger in dsDNA with respect to ssDNA. Discrepancy might be due to the fact that dsDNA has an internal double-helical structure stabilized by H bonds which enhances its stiffness which here we neglect.

Chapter 5

Ground states of a short thick polymer

In this Chapter we discuss how the ground state of a short thick polymer evolves when its thickness, R_0, and the attraction range, R_1, are varied continuously. If we had only hard spheres not tethered together into a chain, the analogous task would be to look for the optimal cluster obtained by varying the attraction range (see Appendix A). In that case we know that when R_1 is just above the hard core radius the optimal configuration is the face-centered cubic lattice, which is employed in many real physical solid system. So our task can be rephrased as the search for the analogous of the face-centered-cubic lattice (and the other lattices which are stabilized by different values of R_1) for a thick polymer, i.e. for its 'crystalline' state. In this calculation, for simplicity we will fix a value of R_1 and see what happens on varying R_0 (no substantial differences are found when R_0 changes and the arising picture is as in Fig. 5.1). This is equivalent to varying the ratio R_0/R_1, i.e. the ratio between the tube thickness and the attraction range. We will call this quantity X.

When the above defined quantity X is very large compared to 1, the tube is so fat that it is unable to benefit from the attractive interactions. The constraints of the three body interaction dominate (the pairwise interaction plays no role) and one then obtains a swollen phase – all self-avoiding conformations of the tube are equally likely and the vast majority of them is not effective in filling the space in the core of the structure (globular proteins fold in order to squeeze out water from the core, which houses the hydrophobic amino acids, see Introduction). On the other hand, for a tube with a very small X compared to 1, one obtains many conformations, leading to an energy landscape studded with numerous multiple minima. We thus expect a transition to occur for intermediate value of X, when the thickness is comparable with the

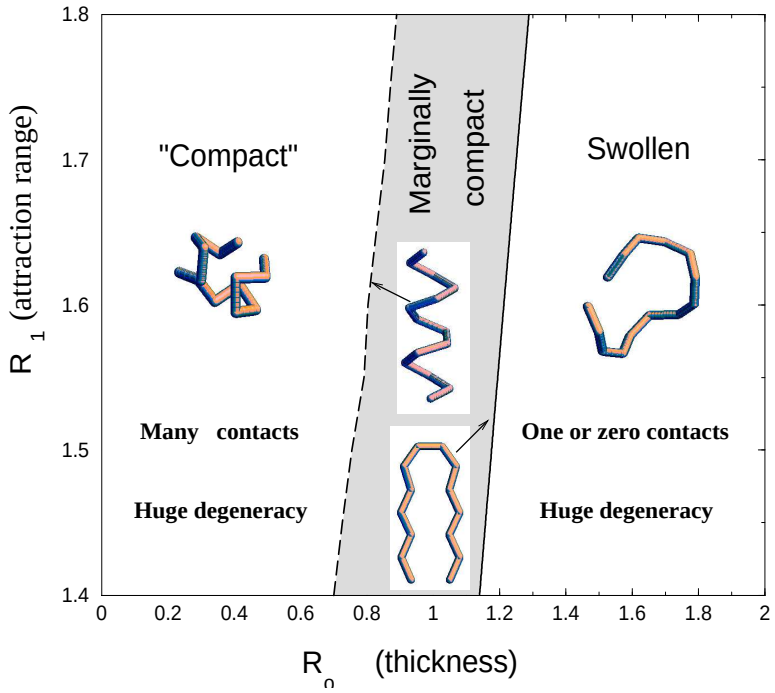


Figure 5.1: Succession of minima as the thickness, R_0, and the attraction range, R_1 are varied continuously. This is valid for short chains. Longer chains will be treated separately below and in the next Chapter.

attraction range (see Fig. 5.1). We call the resulting conformations marginally compact structures because they make a non neglectable number of contacts but lie at the edge of the compact phase: a rather small increase in the tube thickness is enough to change these into swollen configurations unable to benefit from the attracting potential. This situation is conceptually similar to the one encountered in Chapter 3, where marginally compact structures arose when stretching a finite size globule, for a force slightly below the critical force needed to unfold it completely.

5.1 Results and discussion

Our goal is to study the ground state conformations on varying the tube thickness. For a short tube, made up of N balls, subject to pair-wise attractive interactions, let us denote by $N_c(N, R_0)$ the maximum number of contacts (each gaining an energy of -1 according to Eq. (4.10)) that can be made respecting the hard-core and the three body constraints. Quite generally, one would expect that, for a fixed N, N_c is a decreasing function of R_0 and that the decrease occurs in discrete steps corresponding to the inability to form the same number of contacts as before. Physically, of course, for a given contact energy (or equivalently number of contacts), one would choose the largest possible thickness in order to provide as much internal wiggle room (for the side-chains) within the tube as possible.

We have carried out Monte-Carlo simulations by using the standard simulated annealing algorithm (see Chapter 1) in order to find the ground state of the thick polymer. As we have already mentioned, it is known that this procedure may cause the system to 'trap' into metastable minima. This is avoided in two ways: first by repeating the simulation many times, changing the algorithm parameters (such as the amplitude of our dynamical moves), second by comparing the annealing performance with that of a recently proposed algorithm [113], originally proposed in order to find the density of states of the polymer. The results we present in this Section do not depend on the algorithm.

We have made simulations for several values of N and present the scenario for $N = 14$, which is representative for values of N between 6 and 20. A schematic "minima diagram" is shown in Figure 5.1, while Figure 5.2 is a plot of N_c (or equivalently the negative of the ground state energy) as a function of the tube thickness. For small R_0 one gets a highly degenerate phase with N_c saturated at a value of 45 for $N = 14$ – the conformations that the tube adopts depend rather strongly on the details of the pair-wise potential. For R_0 between ~ 0.75 and 0.8, there is an energy plateau in which the degeneracy is greatly reduced and helices are the ground states. Furthermore, for the tube with the largest thickness in this plateau, one obtains a specific helix as the unique ground state (see upper conformation in "Marginally compact phase" of Fig. 5.1 and Fig. 5.3 (A2)). For R_0 between 0.8 and 0.98, several classes of conformations including saddles (which are planar hairpin conformations distorted into three dimensional structures) (see Fig. 5.3 (C2)), generalized helices (in which the distance between successive balls along the helical axis is not constant but is periodic) (see Fig. 5.3 (A3)), helices made up of strands (see Fig. 5.3 (B2)) and other more disordered conformations compete.

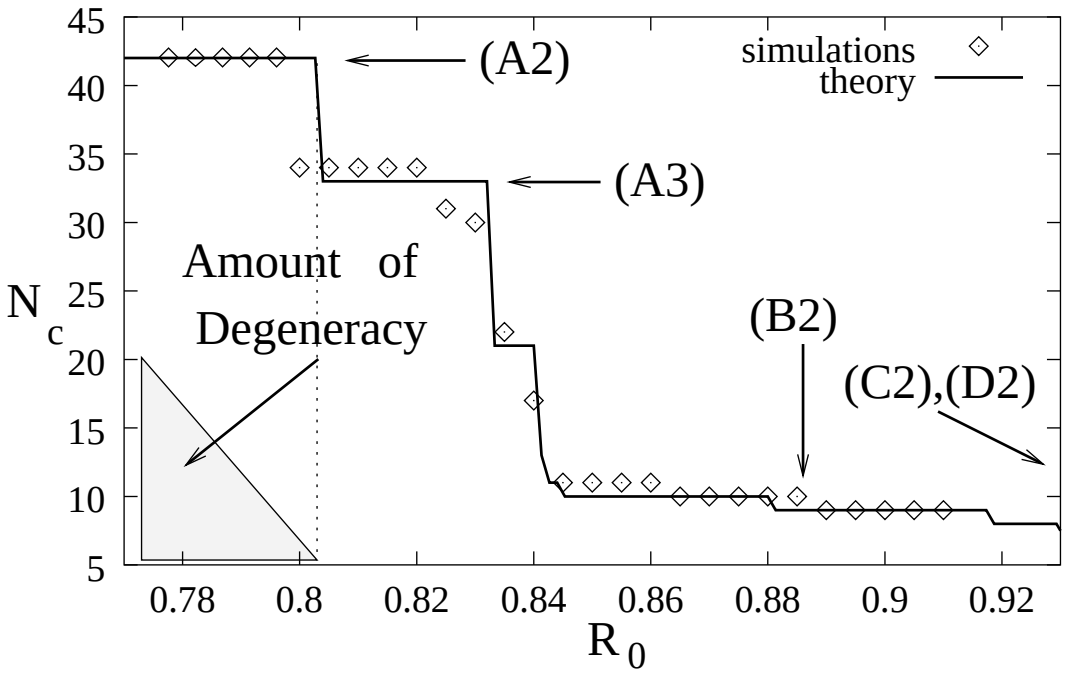


Figure 5.2: Plot of the number of contacts, N_c, vs. R_0 for a short thick tube ($N = 14$). The points refer to simulations, the solid line to the numerical exact minimization in the space of regular and generalized helices (see the text, below). The amount of degeneracy is minimum at the rightmost edge of each step. It is pictorially represented in the first such step by the height of the shaded triangle.

At the end of each plateau in Figure 5.2 (there are as many as three major plateaus in this range of tube thickness, and each plateau comprises several distinct sub-steps or smaller plateaus), we find an ordered unique ground state (see some examples in Fig. 5.3, second row). (It is interesting to note here the fact that helices and sheet-like conformations emerge as the conformations of choice at the end of some of the major plateaus for other parametrizations of the potential energy of interaction as well [114].) One may show that helices are excluded from being the ground states, when the tube thickness exceeds $R_0^{max,hel} \sim \left(\sqrt{1+R_1^2}\right)/2 \sim 0.943$ which is obtained when two parallel straight lines (successive turns of the helix treated as circles with infinite radius) are parallel and face each other at the bond length R_1.

For $R_0 > 0.98$, the ground state structures become more and more planar, first locally then globally. For large N, the winning planar structures entail the combinations of strands into a sheet structure. (We find that sheet structures persist for N as large as at least 33, whereas the persistence length for helices is expected to be somewhat smaller.) For two zig-zag antiparallel strands facing each other, one can easily show that the maximum thickness is obtained (leaving aside the edge effect of how the strands are connected together in a hairpin) when the global radius of curvature is constant and furthermore when the local and the smallest non-local radii of curvature have the same value. Indeed, this condition leads to the following relationship between the tube thickness R_0 and the bond length R_1

$$R_1^2 + 2 + \frac{R_1}{R_0} - 4R_0^2 = 0, \tag{5.1}$$

which yields a value of $R_0 \sim 1.2124$, when $R_1 = 1.6$. In order to get Eq. 5.1, it is enough to use elementary geometry considerations, and then to maximize the function $\min_{i<j<k} R_{ijk}$ with respect to θ, the angle between two successive links in the zigzags, and x, the distance between the two strands.

The swollen phase, which occurs for even larger values of tube thickness has two energy plateaus. The first of these plateaus has just one contact and comprises all swollen conformations whose two ends make a contact – the thickest tube which is able to make 1 contact has a unique ground state of a closed polygon with N edges of unit length and 1 of length R_1. Likewise, the plateau of 0 contacts has the limiting thickness situation of a straight, infinitely fat tube. Indeed, starting from the zig-zag conformation, the unique conformations corresponding to the largest possible thickness compatible with a given energy (or number of contacts) all share the intriguing property that the local and the smallest non-local radius are exactly equal (It is interesting to note that the optimal helix of Fig. 5.3 (A1) has a ratio of the smallest

non-local to the local radius of around 0.97 which is very close to the corresponding value for α helices occurring in proteins [90]).

In Fig. 5.2 we also plot (solid line) the energy obtained from numerical minimization over a subset of configurations, over which the optimization can be handled with very high accuracy (indeed the result is 'exact' for any practical purpose). The subset of configurations used for the solid line of Fig. 5.2 is that of regular and generalized discrete helix. A regular discrete helix or a discrete helix with constant pitch, whose axis we can take with no generality loss to be the z axis, and in which two successive beads stay at unit distance, is characterized uniquely by the rise per bead, v, and by the azimuthal (around z) rotation angle per bead, π/m. From here it can be seen that m, which is not necessarily integer, is the number of beads per turn in the helix. The position of the $i-$th bead, $\mathbf{r}(i)$, for a helix, ($i = 1, \ldots, N$), is thus parametrized as follows:

$$\mathbf{r}(i) = \begin{pmatrix} R\cos(2\pi i/m) \\ R\sin(2\pi i/m) \\ iv \end{pmatrix}, \tag{5.2}$$

where the radius R is fixed by the condition that two successive beads have a relative distance 1, i.e.:

$$2R^2(1 - \cos(2\pi/m)) + v^2 = 1. \tag{5.3}$$

Generalized helices, on the other hand, are characterized by a number of beads per turn, a real number, and a variable rise per bead, which is now no longer constant, but depends on the bead index, and is a periodic function of this variable.

We now come to the discussion of our simulation outputs. Helices and sheets are of course the well-known building blocks of protein structures [36, 37] (see Fig. 5.3 (A1) and (D1) for two examples). In addition to the prediction of these motifs in our calculations, it is interesting to note that some of the other marginally compact conformations bear a qualitative resemblance to secondary folds in biopolymers. Helices analogous to Figure 5.3 (A3) with an irregular contact map occur, e.g., in the HMG protein NHP6a [115] with pdb code 1CG7. Fig. 5.3 (C1) shows the "kissing hairpins" [116] of RNA (pdb code 1KIS), each of which is a distorted and twisted hairpin structure while Fig. 5.3 (C2) is a saddle conformation, which is a hairpin distorted into a three-dimensional structure. Figure 5.3 (B1) shows a helix of strands found experimentally in Zinc metalloprotease [117] (pdb code: 1KAP), whereas Figure 5.3 (B2) is the corresponding marginally compact conformation obtained in our calculations.

We point out that the results above, specific for $N = 14$, apply in general for 'short' thick polymers, where 'short' here means, though it is not too easy to quantify,

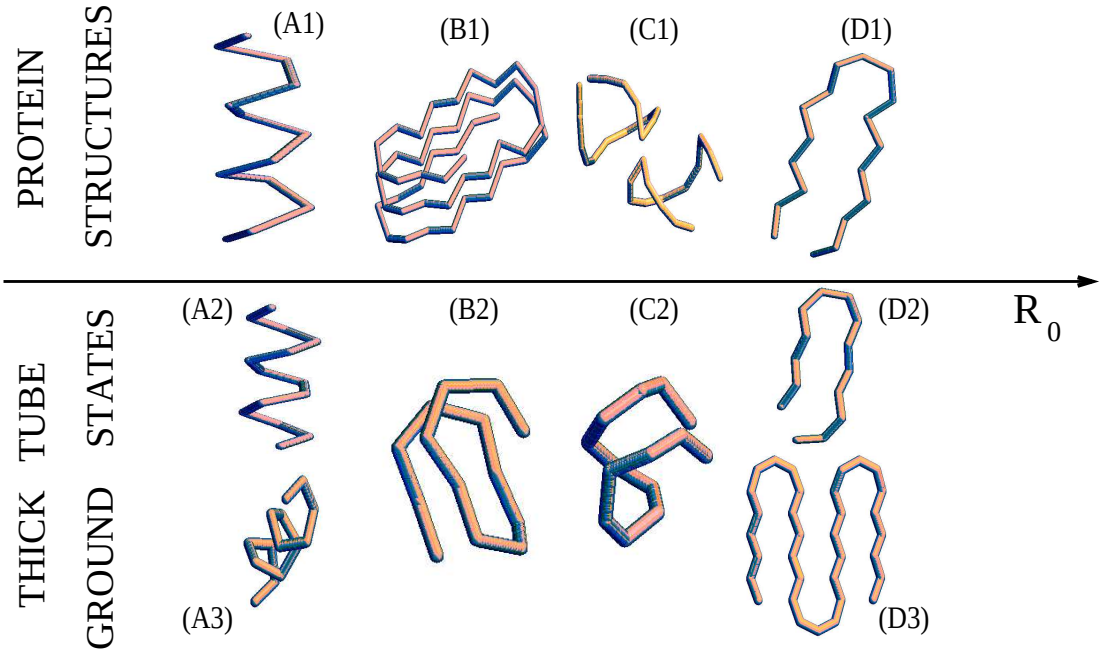


Figure 5.3: Building blocks of biomolecules and ground state structures associated with the marginally compact phase of a short tube. The axis in the middle indicates the direction along which the tube thickness R_0 increases. The top row shows some of the building blocks of biomolecules, while the bottom row depicts the corresponding structures obtained as the ground state conformations of a short tube. (A1) is an α-helix of a naturally occurring protein, while (A2) and (A3) are the helices obtained in our calculations – (A2) has a regular contact map and is obtained when $R_0 = 0.80$ whereas (A3) ($R_0 = 0.83$) is a distorted helix in which the distance between successive atoms along the helical axis is not constant but has period 2. (B1) is a helix of strands in the alkaline protease of pseudomonas aeruginosa, whereas (B2) shows the corresponding structure ($R_0 = 0.88$) obtained in our computer simulations. (C1) shows the "kissing" hairpins of RNA and (C2) the corresponding conformation obtained in our simulations with $R_0 = 0.95$. Finally (D1) and (D2) are two instances of quasi-planar hairpins. The first structure is from the same protein as before (the alkaline protease of pseudomonas aeruginosa) while the second is a typical conformation found in our simulations when $R_0 > 0.98$. All the cases shown correspond to tubes of 14 C_α atoms, except for the sheet-like structure (D3), which employed 33 C_α atoms.

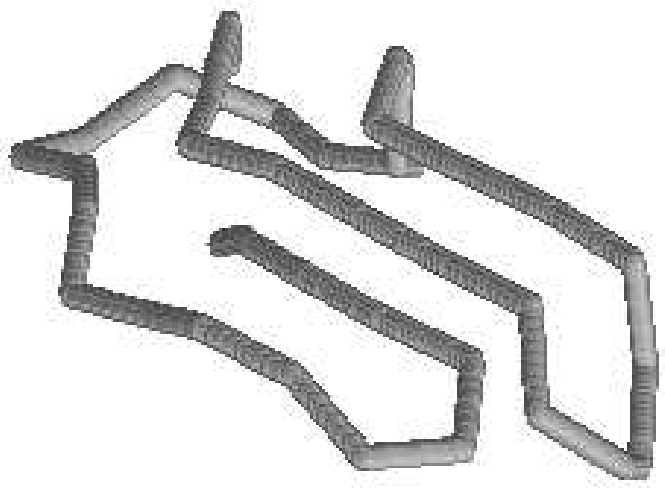


Figure 5.4: This is the ground state, or at least one of the low energy states for a tube of thickness 0.9 with 41 beads.

less than approximately $20 - 30$ monomers depending on the value of R_0. If we consider longer tubes, a preferred ground state at least for polymers with thickness greater then $R_0 = 0.8$ is one in which the different portion of the tube are parallel (see Fig.5.4). This is expected to be, in the continuum, the ground state of a thick tube in the thermodynamic limit. The fact that section of the tubes in contact must position themselves parallel with respect with each other can be rationalized rather simply also in the continuum model. If we computed the interaction energy of two cylinders whose axis points are subject to a two-body potential such as that in Eq. 4.10, we would see (as is apparent from intuition) that for thin tubes ($R_0 \ll R_1$) the attraction roughly is not dependent on the angle between the two axis, whereas for thicker tubes ($R_1 \sim R_0$) there is a marked maximum for parallel axis [118] (because otherwise the tubes would 'stay close' for a smaller portion of their length leaving the remaining portion unable to avail of the attractive potential).

In experiments and real situations, a situation for some aspects analogous to what is obtained in our simulations often occurs. First, in real proteins it is known that when the protein concentration in the cell is higher than a certain critical threshold, then the proteins misfold and tend to aggregate, by forming amyloid fibers. In the case

of prion, for example, this form of protein aggregation causes the arising of the bovine spongiform encephalopathy (BSE) disease. The shape of the aggregate can be visualized in a coarse grained way by thinking of different parts of the aggregate as portion of tubes which are aligned with respect to each other. Another famous example is that of DNA subject to condensation due to the presence of polyvalent counterions in solution. The globule structure formed by DNA upon condensation is often chosen between a circular donut or spool with a hole in its interior, when DNA is short or concentration is low, and a 'nematic-like' state with different portions of the molecule aligned along a local director, when the concentration is high. Our model can on one hand explain the qualitative shape of the minima and also that the crossover between the two different kinds of behaviour should be observed by changing the tube length, which is equivalent to increasing concentration in real experiments. In the case of DNA, we should mention however in the first place that one gets a donut or spool, i.e. a kind of not space filling helix, because one has to take into account the stiffness of the double-stranded DNA molecule which makes it energetically unfavourable to bend the tube. Second, the crossover value of N is much higher than the one proposed here. However, the values of R_0 and R_1 are very different from the true ones in DNA and this can explain in part this discrepancy. The agreement thus is completely qualitative but remarkable all the same.

5.2 A new phase for polymers ?

At the beginning of this Chapter, and in Fig. 5.1 we spoke of a marginally compact 'phase'. Though we used the term in a non rigorous way there, at this point this definition deserves a more thorough discussion. Self interacting polymers (with zero thickness) are usually said to exist in one of three phases: compact, swollen or Θ polymers according to the temperature (or equivalently to the nature of the solvent interactions with the polymer). The compact or collapsed phase can be however of different kinds: either isotropic or anisotropic, i.e. with nematic order, as occurs for semiflexible chains and for our long thick polymers. Neither of these compact phases, however, has the property of being protein-like, in the sense of having a secondary structure which consists of helices and sheets. In this sense, we are tempted to say that our marginal compact structures obtained for short tubes and for short globules under a stretching force do constitute a novel phase for polymers, which is on the edge of compactness and yet not swollen. However, we should remark that strictly speaking for us to speak of a phase we should need those characteristics to persist

in the thermodynamic limit, i.e. for long chains. In the case treated in Chapter 3, marginally compact structures simply disappear after a model dependent crossover length. The situation for tubes is more interesting. For infinite length, the marginally compact regime observed for shorter chains maps into a regime of nematic order plus most probably a small region or a single point in which the structures are planar or quasi planar. The planar point is characterized by different exponents than the compact phase and in this case does represent a different phase, namely one in which the walk is compact in two dimensions. Further work could be done in order to find whether the peculiar behaviour found in $d = 3$, that compact structures upon inflation choose to arrange themselves in an ordered way on a $2d$- plane before disrupting, is common also to cases in $d > 3$.

Chapter 6

Phase diagram for a thick polymer

In this Chapter we want to find the phase diagram, i.e. to study the thermodynamics, at a temperature T of a polymer with non-zero thickness R_0 whose non-consecutive beads are subject to an attractive pair-wise potential as in the previous Chapter. This model should be equivalent to a chain of coins, whose individual circle radius is R_0, as regards universal properties (this is useful because for a chain of coins we can exploit results coming from liquid crystal theory and make analytical progress in the treatment, see below). In the next Section we write down mean field approximations for these two models and solve them explicitly. This is used as a guide towards the numerical solution. Then, in the following Section, we report the results coming from a series of Monte-Carlo computer simulations aiming at characterizing the thermodynamics of a self-attracting tube of varying thickness R_0. Technical details concerning the mean field calculation are deferred to the Appendixes.

6.1 Mean field treatments

The chain of coins

Let us consider a chain of N coins, whose centers are identified by N vectors $\{\vec{r}_i\}_{i=1,...,N}$, each of radius R_0. The coins cannot compenetrate and there is a 2-body purely attractive potential $V_{2b}(\vec{r}_i - \vec{r}_j)$ whose argument is the distance between pairs of coin centers. The centers of the coin also cannot approach closer than twice a given hard core radius R_{hc}. This is necessary for our mean field treatment to give a Θ collapse, otherwise the free energy minimum would always have a non-zero density and would thus never be the swollen phase. The partition function of this system at

an inverse temperature $\beta \equiv \frac{1}{T}$ is:

$$\mathcal{Z} = \int \prod_{i=1}^{N} d\vec{r}_i \prod_{i=1}^{N-1} \delta\left(|\vec{r}_{i+1} - \vec{r}_i| - 1\right) e^{\left(\sum_{i<j;i,j=1}^{N} V_{2b}(\vec{r}_i - \vec{r}_j)\right)} \prod_{i<j;i,j=1}^{N} (1 + f_{ij}^{(1)}) \prod_{i<j;i,j=1}^{N} (1 + f_{ij}^{(2)}), \tag{6.1}$$

where $f_{ij}^{(1)}$ ($f_{ij}^{(2)}$) is -1 if the coins (spheres) centered in $\vec{r}_i$ and in $\vec{r}_j$ compenetrate, and is 0 otherwise. We have taken the length of a link between two successive coins equal to 1. We can also take $\vec{r}_1 = \vec{0}$ without loss of generality.

We now introduce a density field $\psi(\vec{r}, \vec{t})$, where $\vec{t}_i \equiv \vec{r}_i - \vec{r}_{i-1}$, defined as follows:

$$\psi(\vec{r}, \vec{t}) = \sum_{i=2}^{N} \delta\left(\vec{r} - \vec{r}_i\right) \delta\left(\vec{t} - \vec{t}_i\right). \tag{6.2}$$

We can expand the effective Hamiltonian in Eq. 6.1 in powers of $f_{ij}^{(1),(2)}$ as in a standard virial expansion. By enforcing the definition in Eq. 6.2 through the introduction of a conjugate density field $\hat{\psi}(\vec{r}, \vec{t})$, and by restricting the virial expansion in the free energy to two particle clusters in $f_{ij}^{(1)}$ and up to three particle clusters in $f_{ij}^{(2)}$ (see the Appendix for more details on these steps), we can rewrite Eq. 6.1 as:

$$\mathcal{Z} = \int \mathrm{D}\psi \, \mathrm{D}\hat{\psi} \exp\left(\mathcal{H}(\psi, \hat{\psi})\right) \tag{6.3}$$

where the explicit form of $\mathcal{H}$ is given in the Appendix. From Eq. 6.3 one obtains the saddle point equations $\delta\mathcal{H}/\delta\psi = \delta\mathcal{H}/\delta\hat{\psi} = 0$. Solving them is equivalent to finding the optimal density configurations of the system. However an explicit solution is not possible without further approximations. A very common one, which we adopt in the following, is the *mean field* approximation: in our present case it consists in looking for solutions of the saddle point equations which have the form $\psi(\vec{r}, \vec{t}) \equiv \rho\phi(\vec{t})$ (and consequently one has also $\hat{\psi}(\vec{r}, \vec{t}) \equiv \hat{\rho}\hat{\phi}(\vec{t})$). If we now use the self-consistent saddle point equations and insert them back into Eq. 6.3, we find that the mean field free energy functional per particle to be minimized is:

$$f_{mf}(\rho, \psi) = \int d\vec{t}\phi(\vec{t}) \log\left(\phi(\vec{t})\right) - \frac{1}{2}\beta \int d\vec{r} V_{2b}(\vec{r})\rho + \frac{2}{3}\rho\pi R_{hc}^3 + B\rho^2 + 2\rho\pi R_0^3 \int d\vec{t} \int d\vec{t}' \left(1 - \left(\vec{t} \cdot \vec{t}'\right)^2\right)^{\frac{1}{2}} \phi(\vec{t})\phi(\vec{t}'); \tag{6.4}$$

where the first term is the entropy contribution coming from the chain constraint, the next three terms arise due to the two-body potential (the first is the attractive interaction and the other two are the repulsive terms and are necessary in order to describe a Θ transition in our theory), whereas the last term is the coin excluded volume mean field interaction which brings forth the anisotropy in the model[119], and where $B \equiv \frac{8\pi^2\rho^2}{3}\int_0^{R_{hc}} dr_1 \int_0^{R_{hc}} dr_2 \int_{-1}^{1} d\cos\theta r_1^2 r_2^2 H(R_{hc}^2 - r_1^2 - r_2^2 + 2r_1 r_2 \cos(\theta))$ in the approximated treatment given in the Appendix. In the last formula $H(x)$ is 0 if $x < 0$ and 1 otherwise. It is to be noted, however, that the qualitative form of the resulting phase diagram (see below and Fig. 6.1), does not depend at all on the numerical value of B, but just on its presence and on its being greater than 0 in order to render the density of the globule finite when the system is in the compact phase.

We now have to minimize Eq. 6.4 with respect to ρ and ϕ. We call $\int d\vec{r} V_{2b}(\vec{r}) \equiv V_0$, the only feature of the two body attractive potential which affects the mean field solution.

If T is high (β is small), the minimum of f_{mf} occurs for $\rho = \psi(\vec{t}) = 0$: this is the swollen isotropic phase. At a critical value of the temperature, namely for $\beta_\Theta = T_\Theta^{-1} = \frac{\frac{\pi}{2}R_0^3 + \frac{4}{3}\pi R_{hc}^3}{V_0}$, there is a phase transition into a phase with $\rho \neq 0$ and $\phi = 0$, which is physically an isotropic globular phase. This transition is second order within the mean field approximation ($\rho \propto (\beta - \beta_\Theta)$ as $\beta \to \beta_\Theta^+$). If T is further lowered, we find a second phase transition to an anisotropic globule (with non-zero orientational order parameter ϕ) at a temperature:

$$T = T_F = \frac{T_\Theta}{1 + \frac{BT_\Theta}{R_0^3 V_0}} \tag{6.5}$$

This is a first order transition, as can be also seen by realizing that our system is equivalent to the liquid crystal theory treated in Ref. [120] and then following the calculations reported there. Another important point is that the ratio between the two critical temperatures (see Eq. 6.5) approaches 1 for large R_0 as $1 - \frac{a}{R_0^3}$, with $a > 0$ a constant. The precise form of the phase diagram is shown in Figure 6.1.

From that it is clear that we can distinguish two regimes: the first is one in which the tube is 'thin', which occurs when $R_0 < 1.5$ roughly, and in which there are two clearly distinct transition as T decreases, the first (the usual Θ transition) from coil to globule, and the other one, which we call 'folding transition', between an isotropic globule and a compact configuration in which there is orientational order. After this second transition, the chain of coins ground state resembles the aggregate that we found in the preceding Section as a ground state for a long thick polymer, hence the name 'folding temperature'.

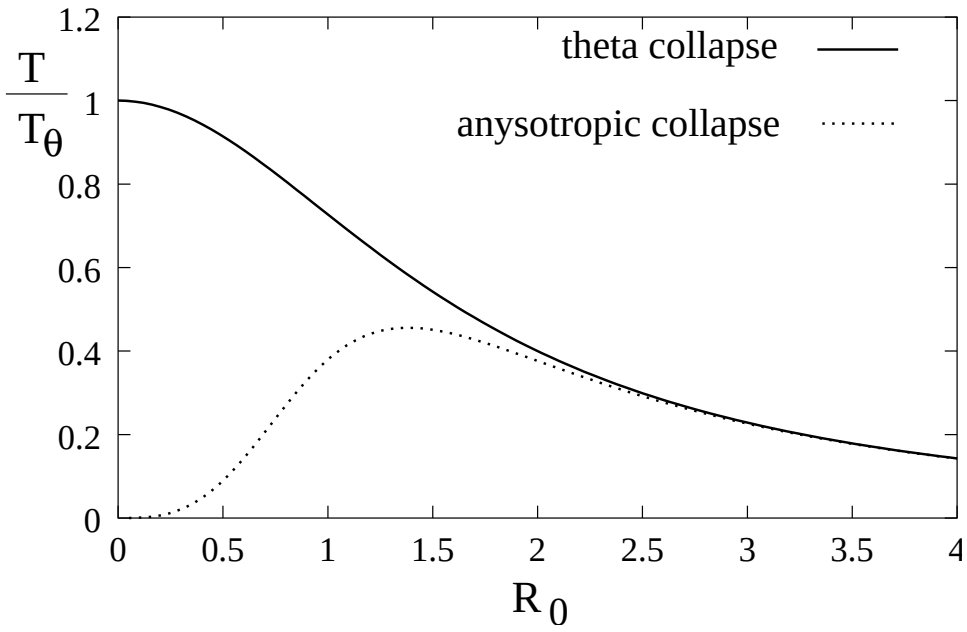


Figure 6.1: Phase diagram for a chain of coins in the (T, R_0) plane obtained in the mean field approximation.

A different regime is instead met when the thickness R_0 is above roughly 1.5: in this case we speak of 'thick tubes'. In this second regime there is practically only one transition, i.e. the system is almost never in the isotropic globule state. We also ascribe the fact that in our theory there is always a small isotropic globular phase to a weakness of our mean field approximation. A similar situation was observed in [121, 122], in which the authors find two transitions in the mean field approximation for stiff polymers whereas simulations indicate that the true solution should display a single first order transition. It is to be noted that from experiments and recent theoretical results on the thermodynamics of proteins [123, 124] it is rather well established that real proteins may fall in either of these two situations.

The 'tube', i.e. a polymer with a constraint on the triplet radii

We briefly give the results of the same mean field treatment applied to a tube-like polymer with a three body constraint (see preceding section). Here once more the suitable order parameters are ρ, the density, and $\phi(\vec{t})$, the link orientational order parameter. With respect to the chain of coins, there is no region in which there are two transitions. The transition temperature is $T_\Theta = \frac{V_0}{2B(R_0)+\frac{8}{3}\pi R_{hc}^3}$, where $B(R_0) \equiv$

$\frac{1}{V}\int\int\int d\vec{r}_1 d\vec{r}_2 d\vec{r}_3 f_3(\vec{r}_1,\vec{r}_2,\vec{r}_3)\delta(|\vec{r}_3-\vec{r}_2|-1)$, which goes to 0 as $R_0 \to \infty$. In this formula f_3 is 0 if the radius of the circle constructed with the triplet $\vec{r}_{1,2,3}$ is greater than R_0 and 0 otherwise: this term acts as an effective two-body hard core (but thickness dependent), whereas V is the volume occupied by the system in the canonical ensemble. We have also considered in general the presence of a standard two-body hard core which prevents the critical temperature from being infinitive in the $R_0 = 0$ case. The Θ collapse is from a swollen phase to an isotropic globule if $R_0 < \frac{1}{2}$, and to an anisotropic globule otherwise. This time the transition is second order in any case. We point out that the wormlike chain model (see e.g. [5] for the definition) with compacting potential would give a qualitatively similar phase diagram in this mean field calculation. This is so because indeed at first order in the cluster expansion the tube nature is important only in disallowing strong bends, as happens also for polymers with stiffness. However, higher order in the virial expansions would consider configurations with non-local thickness effects where different portions of the tube come into contact with one another. Below we present the phase diagram found with Monte-Carlo simulations, which does take every feature into account.

6.2 Monte-Carlo evaluation of the phase diagram

We now consider a tube of a given thickness, schematized as explained above with a three body effective constraint. We want here to find the thermodynamics of the system without relying on a mean field approximation.

The methods employed are Monte-Carlo simulations, with the parallel tempering or multiple Markov chain technique (see [34] and Chapter 1). The number of replicas equilibrated simultaneously ranges from 11 to 16 in our simulations. When needed, to estimate canonical averages at other values than those simulated, we used the multi-hystogram method as in Ref. [125].

In all formulas below $\langle\cdot\rangle$ denotes ensemble averaging. For each thickness considered, we calculate the specific heat C as:

$$C(T,R_0) = \beta^2\left(\langle E^2\rangle - \langle E\rangle^2\right), \tag{6.6}$$

where E is the internal energy (contact number) and $\beta \equiv \frac{1}{T}$ is the inverse temperature; and the radius of gyration R_g defined as:

$$R_g(T,R_0) = \langle\frac{\sum_{i=1}^{N}\left((x_i-x_{cm})^2+(y_i-y_{cm})^2+(z_i-z_{cm})^2\right)}{N}\rangle^{1/2}, \tag{6.7}$$

where $\{x_i, y_i, z_i\}_{i=1,\ldots,N}$ represent the coordinates of the N beads and $\{x_{cm}, y_{cm}, z_{cm}\}$ the coordinates of the center of mass of the polymer.

We also record three simulated probability distributions: the distribution of the chirality[1] χ, which is a number between -1 and 1 defined for every consecutive quadruplet of beads along the chain, whose position vectors are $\vec{r}_{i,i+1,i+2,i+3}$, as:

$$\chi \equiv \vec{r}_{i,i+1} \cdot \left(\vec{r}_{i+1,i+2} \times \vec{r}_{i+2,i+3}\right), \tag{6.8}$$

the distribution of $\cos(\theta)$, where θ is the angle between two consecutive links along the chain, and that of $\cos(\xi)$ where ξ is the angle between two links which are in contact. Note that we define two links ($\vec{r}_{i+1} - \vec{r}_i$ and $\vec{r}_{j+1} - \vec{r}_j$) to be in contact when $|\vec{r}_i - \vec{r}_j| < R_1$. We label these probability distributions $P(\chi)$, $P\left(\cos\left(\theta\right)\right)$ and $P\left(\cos\left(\xi\right)\right)$ respectively. For comparison it is useful to remember the shapes of these distributions for non-interacting random walk, which are immediately found to be:

$$\begin{aligned} P(\chi) &= \frac{2\sqrt{1-\chi^2}}{\pi} \\ P\left(\cos\left(\theta\right)\right) \equiv P\left(\cos\left(\xi\right)\right) &\equiv \text{const.} \end{aligned} \tag{6.9}$$

As with all Monte Carlo algorithms, finding canonical averages and particularly the specific heat below the transition is difficult. The multiple Markov chain (or parallel tempering) algorithm is a way to take care of this because the mobility of the chain at low T is enhanced. To further improve results at low T, we have first performed one or more parallel tempering runs collecting both data for canonical averages and specific heat and the low T configurations, starting from an open initial conditions (all chains more or less straight are chosen for all the replicas). Then another run is performed with the replicas starting from a 'folded' conformation, the one with the minimum energy found with the preceding runs. This is a way to sample more correctly the low T configurations, giving a bias towards the 'correct' ground state for very low T. On can check that the high T behaviour is the same for all runs.

From our data, we found that the phase diagram we obtain (see Figure 6.10 in which the phase transition critical lines are plotted in the (T, R_0) plane) suggests the presence of two quite distinct regimes as for the mean field approximation results: the first holds for 'thin polymers' (for thickness up to roughly 0.8), the second for 'thick tubes' ($R_0 > 0.8$). This is similar to what we found in the mean field analysis of the chain of coins.

[1]The chirality is mathematically the mixed product of three consecutive links (see Eq.6.8). It is useful for us because it gives information on the structure of the chain. A helix has a chirality which is constant throughout the chain, a planar structure on the other hand has $\chi = 0$ for any quadruplets.

In the graphs in Figures 6.5,6.9 we show the behaviour of the specific heat and of the gyration radius for two different thickness values, typical respectively of thin ($R_0 = 0.7$) and thick ($R_0 = 0.95$) tubes. In the other Figures 6.2-6.8 on the other hand we show typical histograms for the previously defined $P(\chi), P(\cos(\theta)), P(\cos(\xi))$. We considered polymers with $N = 20, 41$ and 60. We have performed Monte-Carlo runs scanning the thickness R_0 every 0.05 units.

Regime 1: thin polymers: $R_0 < 0.8$

For small thickness, the thermodynamic behaviour of the polymer is for many aspects found to be similar to the one reported in Ref. [126], which holds for conventional polymers whose thickness is actually 0. We consider the case $R_0 = 0.7$ as the representative one. By looking at the gyration radius (see Fig. 6.5), one would place the 'Θ-like' collapse transition point at the T for which the three curves intersect, since on the transition line for the usual Θ collapse one expects that $R_g \sim N^{1/2}$ (as it should be $\nu = \nu_\theta = 1/2$). From Fig. 6.5, we see that even for a polymer with a (small) non-zero thickness, the polymer critical exponent ν at the transition line is $1/2$. This means that roughly $T_\Theta = 1.5$. The specific heat however displays a peak only for T approximately equal to 1 (see Fig. 6.5), a value significantly smaller than the previous one. Moreover, our data suggest that there could be another transition for even smaller values of T. The specific heat points are however very noisy in this region (see Fig.6.5) and hamper a clear identification of this second transition. If we look at $P(\chi)$, we see that after we pass through the Θ-like point where the gyration radius drops, the distributions changes from one peaked around $\chi = 0$ (more strongly than in the random walk case), to a more or less flat distribution. This remains true after the first specific heat peak is gone through, whereas as the polymer meets the second peak of the specific heat the distribution develops many peaks for non-zero values of χ. The analysis of the distribution of the angle θ between successive links suggests that at high T the distribution is flat but is such that negative values are disallowed because of the thickness constraint. After the first transition, values of $\cos(\theta)$ correspondent to tubes not tightly bent locally are penalized; after the specific heat peak this is more pronounced (and two peaks develop in the small angle region). Note that after the 'Θ-like' transition, the values of $\cos(\theta)$ which are not penalized correspond to a local thickness roughly between 0.7 (the constraint), and slightly more than 0.8, in a way that one can summarize the situation by saying that 'locally thin' tubes only dominate the thick polymer distribution probability after the θ-like collapse. Last, analysis of $P(\cos(\xi))$ suggests that the second peak in the specific heat, corresponding to the hypothetical third transition temperature, is connected to an increased probability of contacts to occur between parallel links. This suggests that the

second specific heat might mean a transition between an isotropic to an anisotropic crystalline state for the polymer. This effect is rather mild though, because locally the tubes are not straight: this might be due to finite size effects and in this view only in the thermodynamic limit the mean field picture of a nematic-like phase would be recovered.

In Ref. [126] it was suggested that there is a first Θ transition from coil to globule, equivalent to the usual gas-to-liquid transition. This is signalled by the drop of the gyration radius as well as by a shoulder in the specific heat. Subsequently, the globule crystallizes (as in a liquid-to-solid transition for clusters of hard spheres) giving raise to the first peak in the specific heat, which should correspond to a first order transition in contrast with the second-order transition of standard Θ collapse. Lastly, there is a second transition, which the authors refer to a solid-to-solid first order transition.

It is interesting to notice that even with exact enumerations on interacting self-avoiding walks on the two-dimensional lattice, though for the rather small values manageable for these problems (which means roughly N of order 25), one gets a similar behaviour[127], in that the gyration radius drops at a higher T than the one at which the specific heat displays a peak, and at still lower T one finds a second specific heat peak.

In our case, we believe that the thermodynamics of a thin tube is equivalent to the one proposed in Ref. [126] up to the first specific heat peak. However, what happens below is most probably different and peculiar to the cylindrical shape of the polymer. In particular, we have only a partial numerical evidence that the second specific heat peak corresponds to a physical transition. Also, if there is such a transition, we are lead to believe that in our calculation the nature of this transition should be different than the second solid-to-solid transition observed with conventional polymers: it is likely that the second specific heat peak corresponds to the transition between an isotropic compact state into an anisotropic one, as in the mean field result.

Regime 2: thick polymers

For higher thicknesses (roughly from $R_0 = 0.8$ on), the thermodynamic behaviour is different. In this case, the point in which the scaled curves for R_g intersect and the location of the peak in the specific heat (see Fig. 6.9) are not very different as in the case of thin polymers. This suggests that in this case there is a single transition. Also the data for the specific heat per monomer, within the Monte-Carlo errors, suggest that this quantity diverges at the transition. We thus predict that this is a first order direct transition from a swollen phase to a 'nematic-like' globular phase. This prediction is in agreement with the mean field analysis of the preceding Subsection, and is also confirmed by looking at the other data from the simulation (see below). That

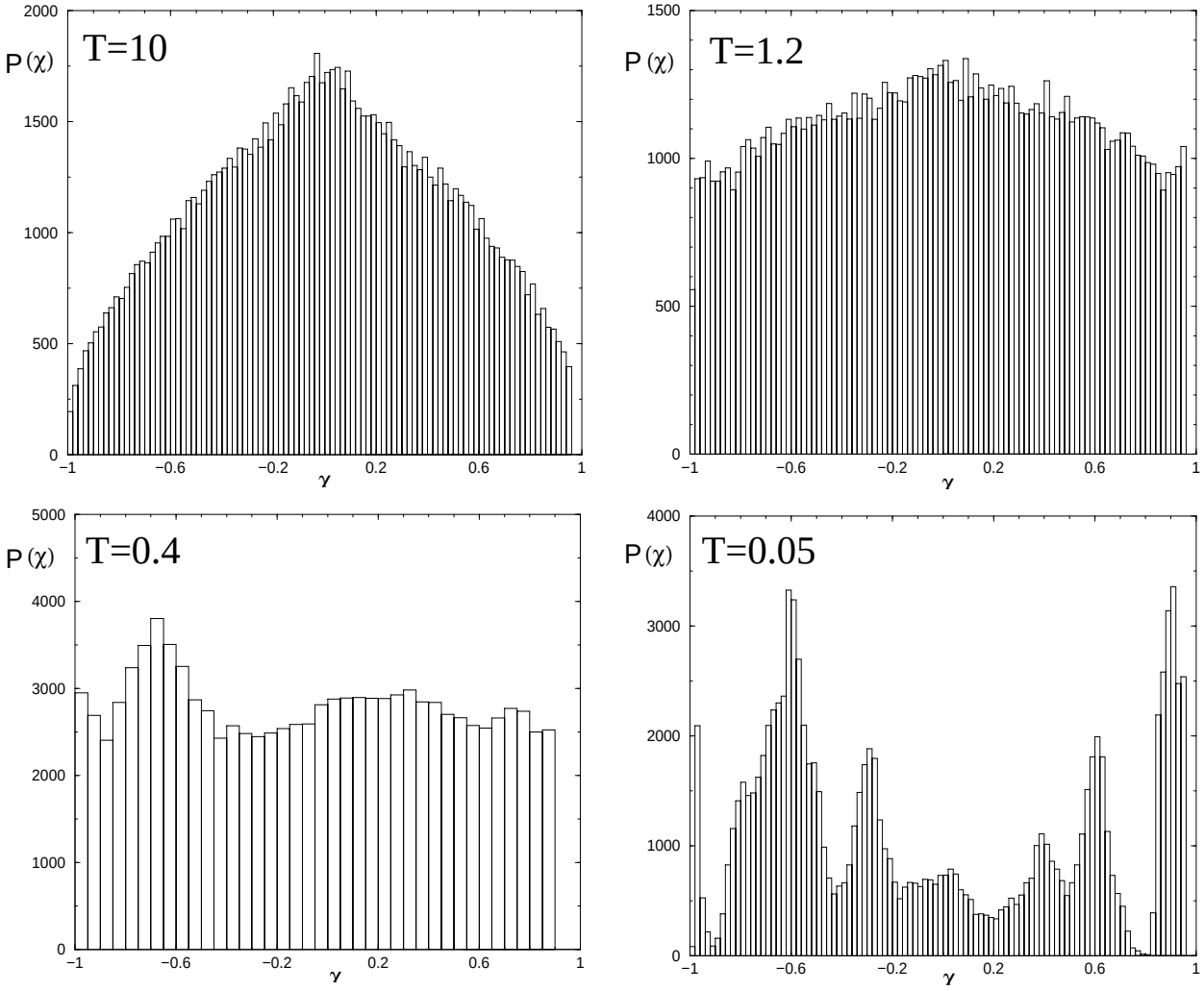


Figure 6.2: Chirality probability distribution from Monte-Carlo simulation for a thick polymer with $R_0 = 0.7$.

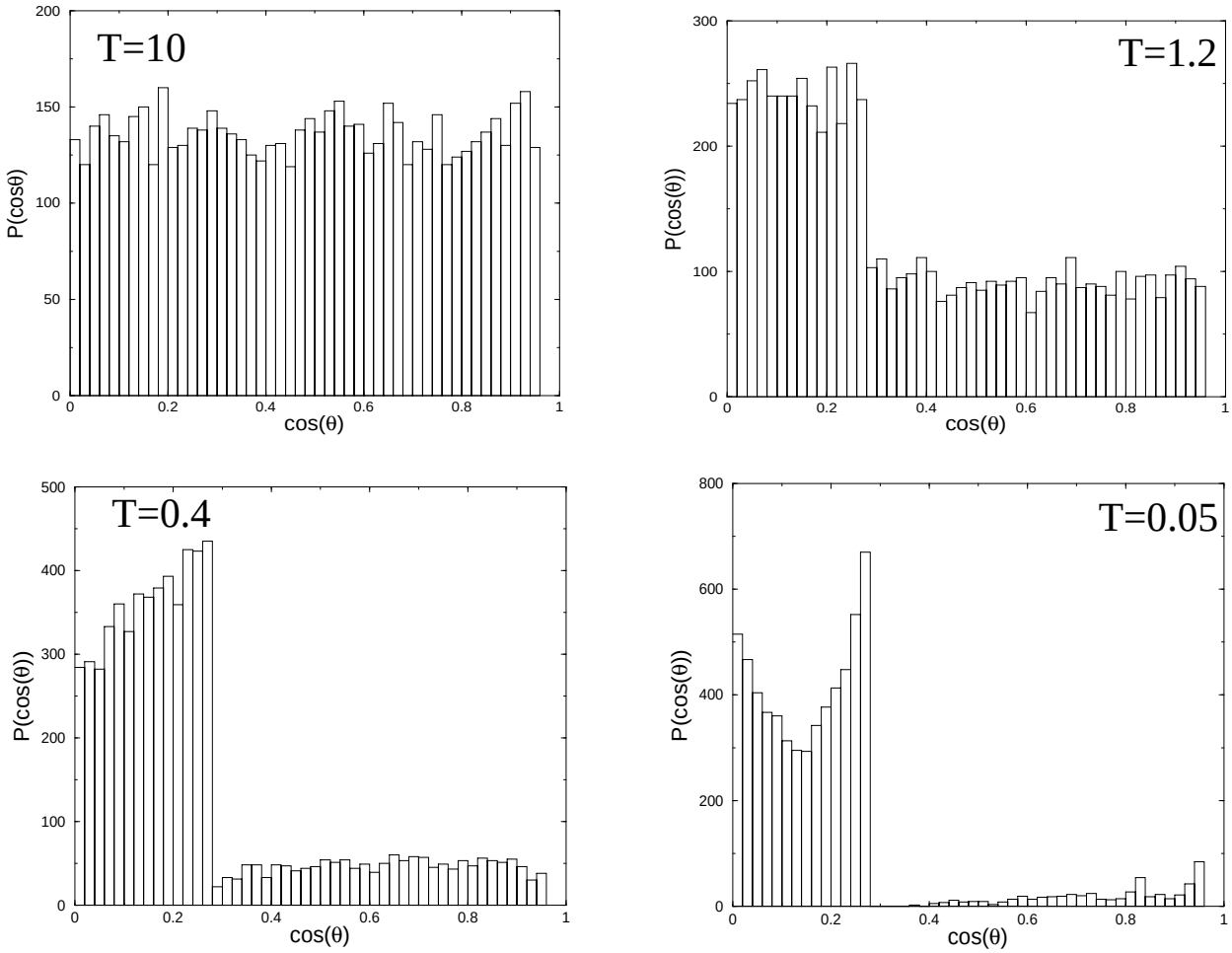


Figure 6.3: Probability distribution of $\cos(\theta)$ for $R_0 = 0.7$.

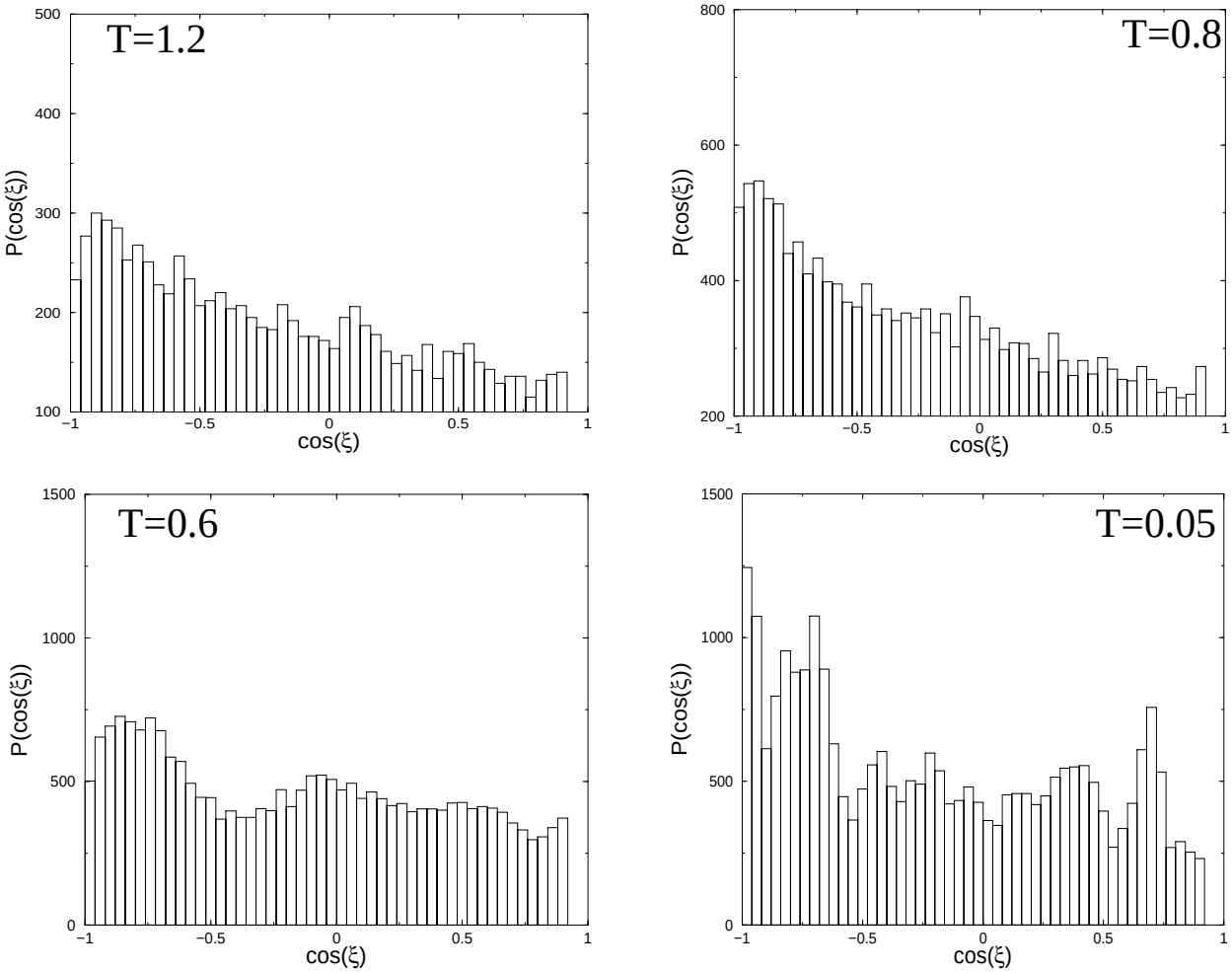


Figure 6.4: Probability distribution of $\cos(\xi)$ for $R_0 = 0.7$.

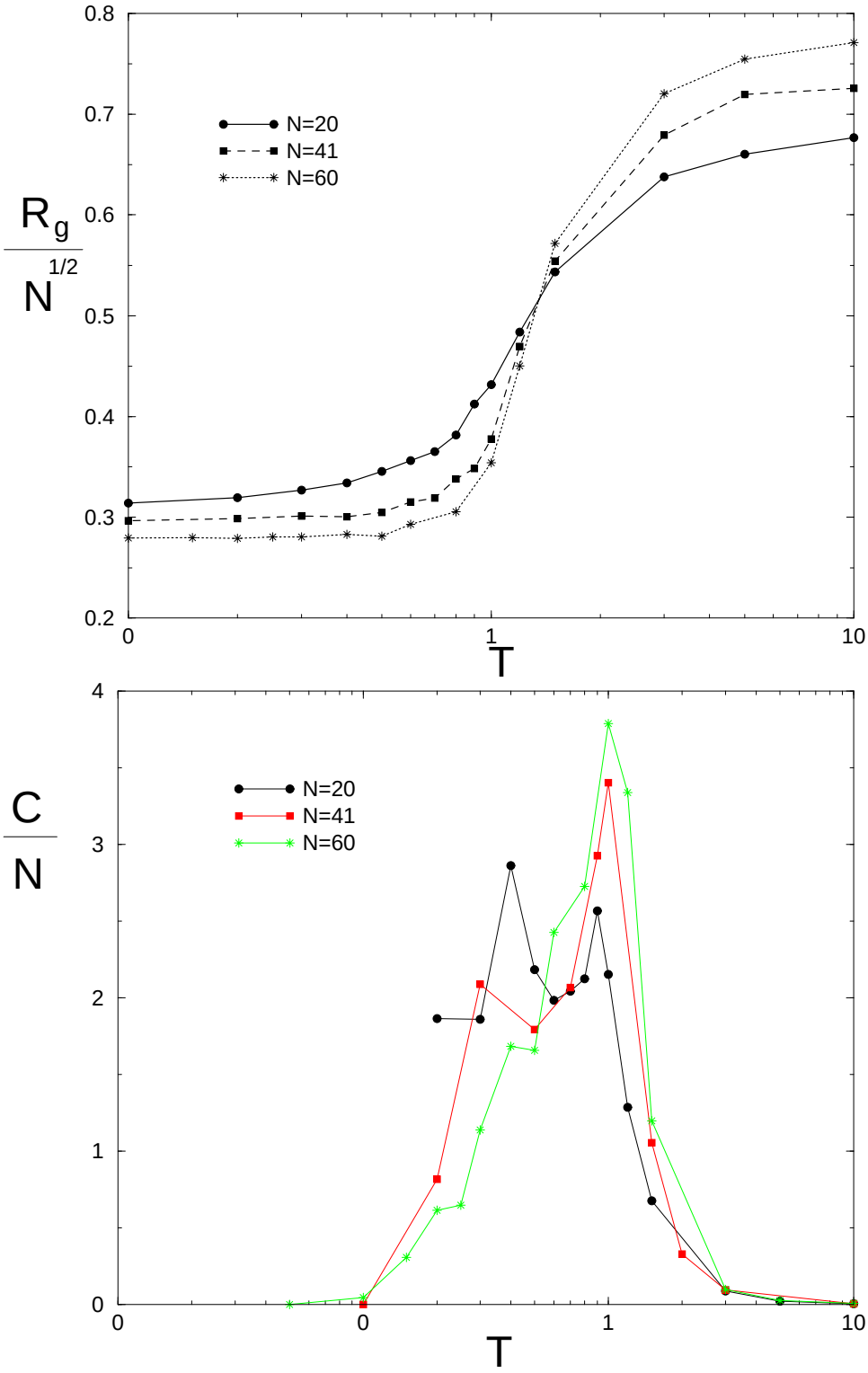


Figure 6.5: Gyration radius (top) and specific heat (bottom) for a thick polymer with $R_0 = 0.7$. Error bars are maximum near the specific heat peak, where they are at most $\sim 20\%$, otherwise they are comparable with point size.

the low temperature phase is indeed characterized by portion of the tubes trying to position themselves parallel with respect to one another is rather clear by looking at the distribution $P(\cos\xi)$, which at low T has sharp peaks in -1 and in 1. By analogy with the continuum ground state for long tubes, we also conjecture that for long tubes the ground state for very low T will happen with tube portions oriented parallel to one another (Abrikosov lattice state, as in Fig.5.4). It is to be noted however that the portion of the tubes which face each other are more or less straight only for not too thick tubes (around $R_0 = 0.9$ in our simulations, see Fig. 5.4), otherwise the thickness constraint makes so that, for $R_0 = 0.95$ for example, we do not get a peak corresponding to 1 in the distribution $P(\cos(\theta))$. The analysis of $P(\chi)$ (Fig.6.6) reveals that at high T the chirality distribution is wedge-shaped, so the peak in $\chi = 0$ is more enhanced with respect to the case of low thickness. This shape persists almost up to the transition point, and there are no pre-transition effects as in the thin polymer case, consistently with the transition being first order, i.e. sharp. Below the transition, three peaks develop, one for $\chi = 0$, not present for $R_0 = 0.7$, and the other two roughly symmetric at the edge of the χ spectrum visited by the system. The central one might be due to the richness of planar or quasi-planar structures in the sampled configurations (as a planar structure has $\chi = 0$).

The critical exponent ν for this case of thick polymers deserves some further discussion. Note that, once again, the gyration radius data (Fig. 6.9) intersect when scaled by $N^{1/2}$, which would suggest that the polymer critical exponent along the transition line is $\nu = \frac{1}{2}$ as in the standard Θ point transition. Though, a rather good scaling is also obtained with $\nu = 0.588\ldots$ (the exponent of polymers in the swollen phase). Indeed, if there is coexistence at the transition point (as is usually the case in a first-order transition), one would get $\nu = 0.588\ldots$. Note also, however, that if we move in the phase diagram at $T = 0$, we expect that the ground state obtained for $R_0 = 1.2124\ldots$ (the value in Eq. 5.1) is a compact configuration in $d = 2$, thus corresponding again to $\nu = \frac{1}{2}$. So if at the transition point the configuration is planar or quasi-planar, the exponent would be $1/2$ along all the transition line. Both these possibilities are equally likely from our data.

The phase diagram in Fig. 6.10 was computed by marking as transitions only the specific heat peaks, even if for thin polymers there may be one more transition corresponding to the gyration radius drop as discussed above. It is intriguing that the 'triple point' in the phase diagram, which discriminates between 'thin' polymers and 'thick' polymers (that we have identified as being roughly 0.8), in which anisotropy of the cylindrical chain starts to play a major role, seems to coincide with the thickness corresponding with the onset of the marginally compact region in the minima analysis

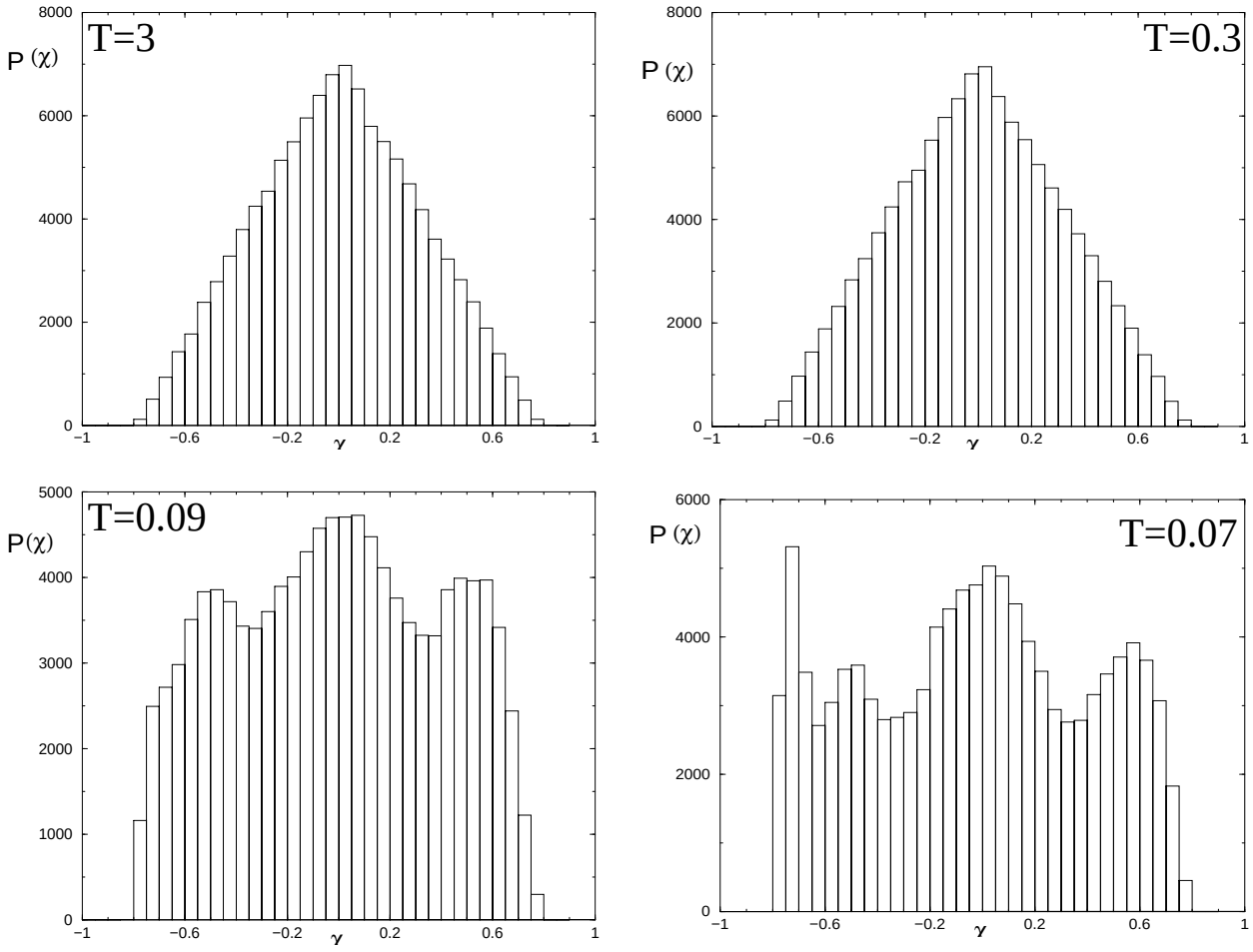


Figure 6.6: Chirality probability distribution from Monte-Carlo simulation for a thick polymer with $R_0 = 0.95$.

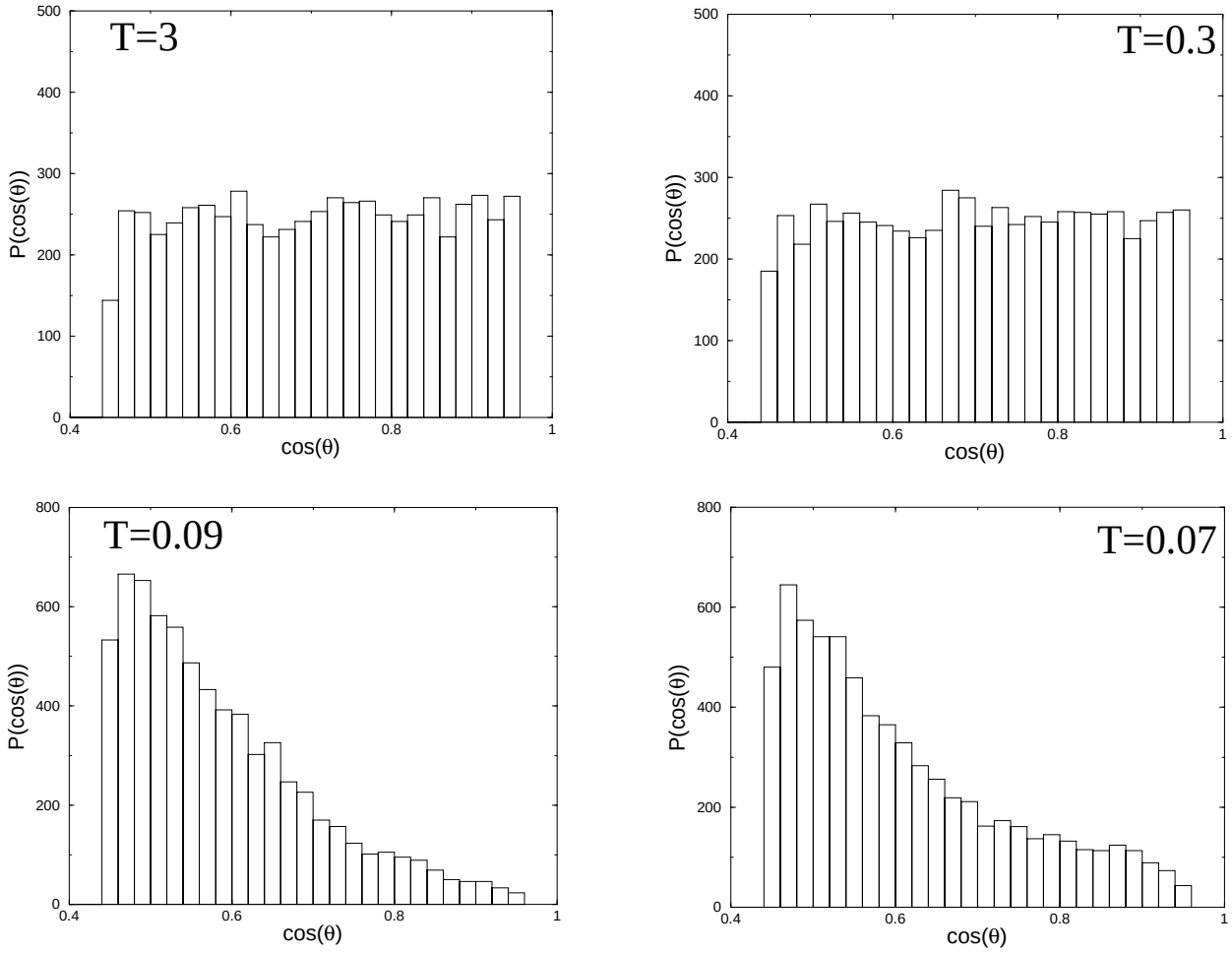


Figure 6.7: Probability distribution of $\cos(\theta)$ for $R_0 = 0.95$.

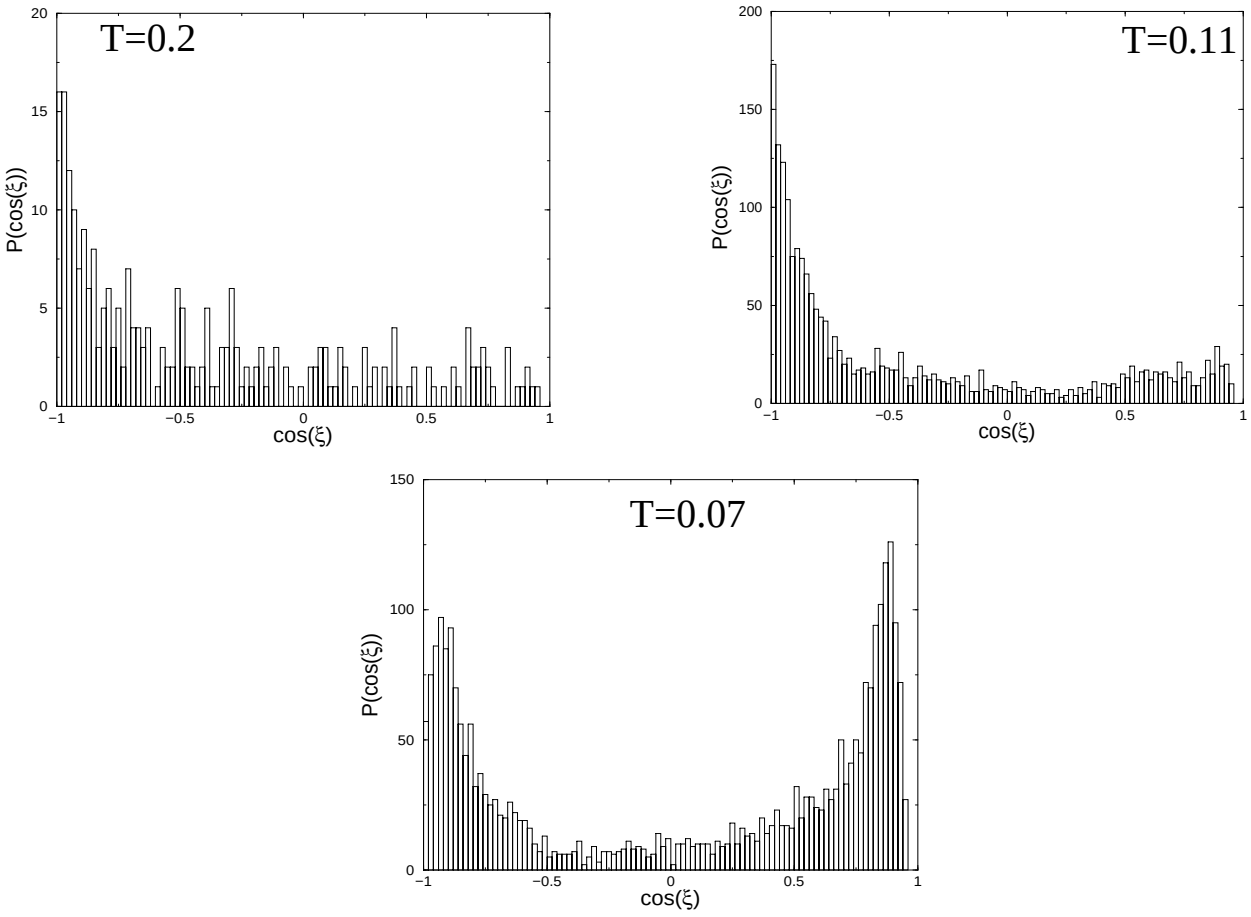


Figure 6.8: Probability distribution of $\cos(\xi)$ for $R_0 = 0.95$.

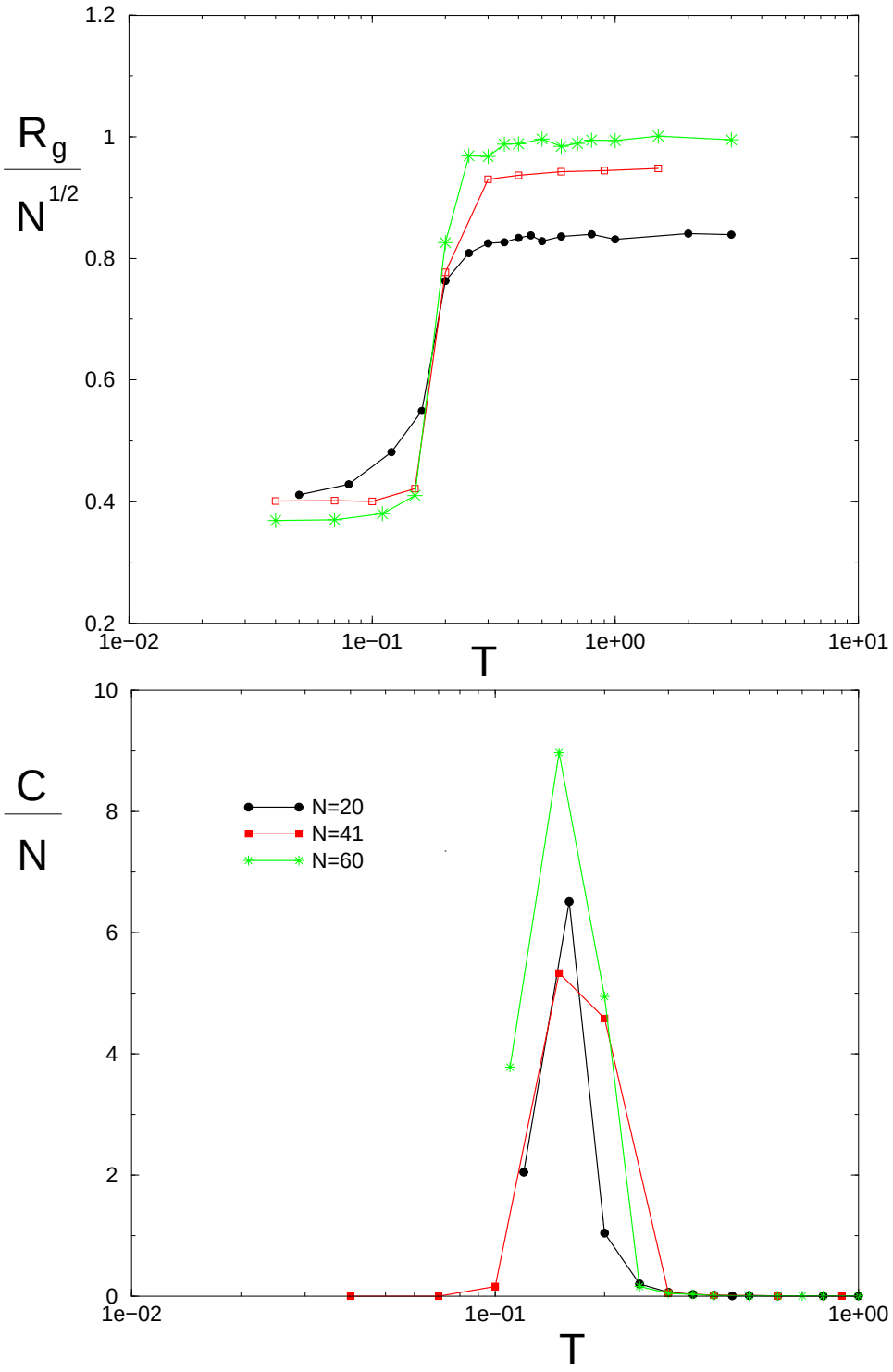


Figure 6.9: Gyration radius (top) and specific heat (bottom) for a thick polymer with $R_0 = 0.95$. Error bars are maximum near the specific heat peak, where they are at most 20%, otherwise they are comparable with point size.

for short polymers detailed in Section III. The phase diagram in Fig. 6.10 should be compared with that of a stiff polymer with no thickness (see e.g. Ref. [96, 121, 122]). Also stiff polymers are known to undergo either two transitions, one of second and the other of first order, or only one transition of first order, according to whether the stiffness is small or high. The thickness of the tube indeed makes it impossible for it to have sharp bends and this is reflected in the tube having a bigger persistence length than a polymer with no thickness: in other words, the thickness acts *locally* as a stiffness. However, the thickness has an important role also non-locally both in providing a simple mechanism through which distant portions of the polymers must orient themselves selectively in order to take advantage of attractive interactions, and, more important, provide a natural way in which naturally occurring secondary motifs arise with no need of heterogeneity in the model. To our knowledge no secondary structures appear in the resulting ground states when dealing with only stiff polymers, whereas with thick polymers we obtain both α helices and β sheets. Also, thick discrete polymers display a planar phase at high values of the thickness which cannot be obtained with stiff polymers.

We also performed an analysis of the 'unfolded state', i.e. of the *ensemble* of configurations typical for the polymer just above the transition point (the first peak of the specific heat when there is ambiguity). Again the situation is different for thin tubes and thick tubes. In the former case, the unfolded state is one in which some contacts are already formed, so it is distinct from a swollen phase, though it has no regularity typical of the crystalline phase. In the latter case, on the other hand, consistently with the fact that the transition is found to be first-order-like, we observe that just above the transition point the chain is still indistinguishable from a typical configuration in the swollen phase. One can get an idea of this by looking at the distribution of $P(\cos\xi)$ shown before. In the case of a thick tube, one can see there is almost no statistics for this quantity until at the transition temperature (i.e. for higher T there are practically never sampled configurations with one or more contacts, which is why we can only start at $T = 0.2$ with the measure of this probability distributions in Fig. 6.8), suggesting thus that the unfolded state in this case is swollen and that the transition is sharp, of all-or-none kind. In the case of a thin tube, on the other hand, even for $T = 1.2$ and higher (not shown) we sampled a reasonable number of configurations with self-contacts, which are better characterized as partially folded than as completely unfolded.

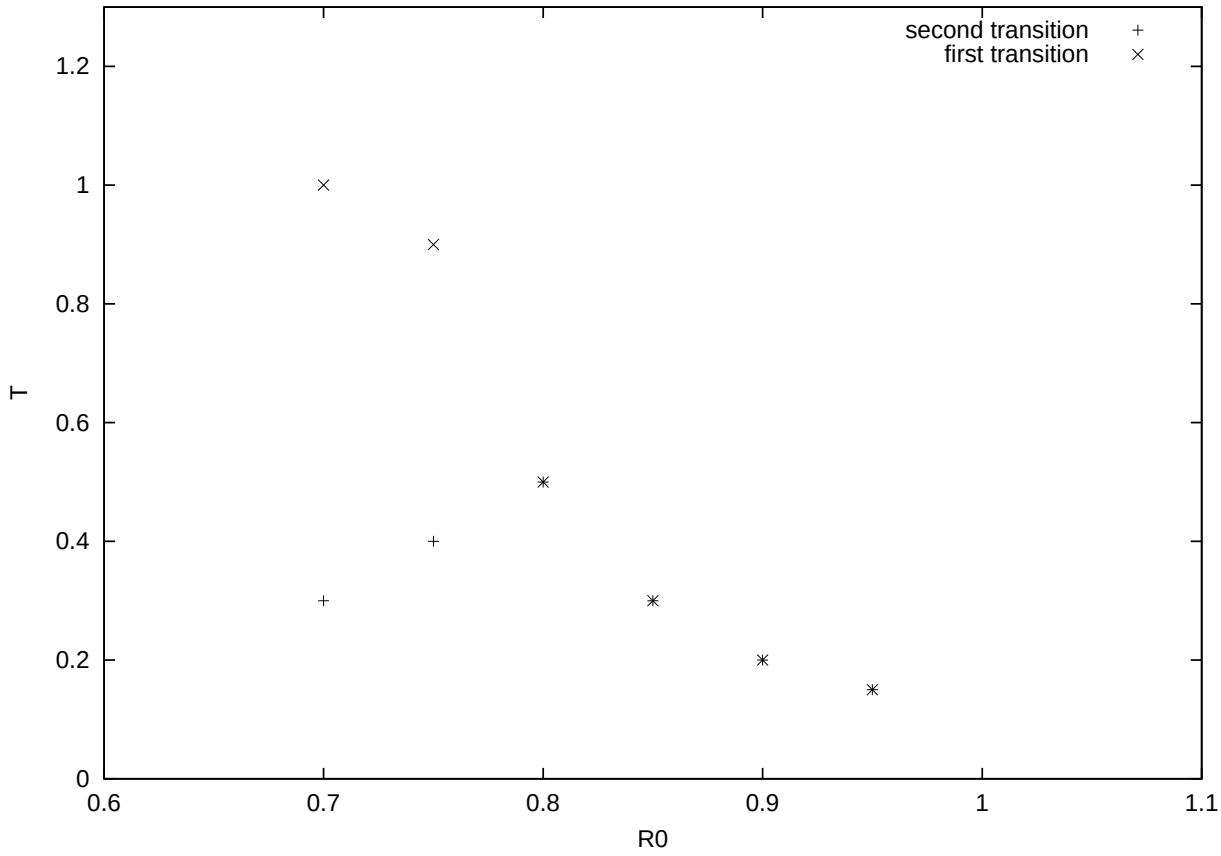


Figure 6.10: Phase diagram for a thick polymer in the (T, R_0) plane obtained with Monte-Carlo simulations. Points represent peaks in the specific heat curve for $N = 41$. Errors on the critical temperature determination can be taken to be ± 0.1.

Conclusions and Perspectives

This thesis consists of two parts. The first part dealt with original physical questions motivated by single molecule experiments in DNA unzipping and in stretching of collapsed polymer.

In the first Chapter, within a lattice model of self-avoiding walks, we proved that the force versus temperature phase diagram in DNA unzipping displays a regime in which it becomes harder to open the molecule as T increases. This corresponds to a reentrant transition line and has been explained as a rather simple energy-entropy effect. Even more realistic models, in which heterogeneity and dsDNA stiffness are fully taken into account would show the same behaviour, rendering this reentrance a robust phenomenon. However, estimates coming from our model suggest that in real experiments this effect should be relegated to values of the temperature which are lower than those which can nowadays be reached. A complete characterization of the dynamics in terms of the dynamical critical exponents has been given for the homogeneous system, and a partial analysis has been performed for random DNA, limited to the case when denaturation bubbles are neglected (which is a valid approximation as long as T is consistently below the melting temperature). The values of critical exponents have also been explained by heuristic arguments. To our judgment, the most relevant result is that the mechanism of unzipping below the denaturation transition is super-universal, in that it does not depend on the dimension and on the specific details of the model (complete self-avoidance, stiffness, etc.). On the other hand, the zero-force melting dynamical mechanism appears to be much more model dependent, thus giving in principle a 'dynamical criterion' according to which during replication a fluctuation induced bubble can be promptly distinguished from a proper force-induced unwinding of the strands. At this point, the model will have to be made more realistic in order for it to represent faithfully the first step of DNA replication. The most important missing ingredient is the intrinsic helicity of DNA in its double-stranded form. Careful further work is needed in order to ascertain its role on the var-

ious aspect of the phase transition. Nevertheless, the predictions highlighted above are expected to be robust with respect to the introduction of these realistic details.

In the second Chapter, we discussed the seemingly related topic of the stretching of a polymer in a poor solvent. We studied a model of a self-avoiding chain subject to a compacting potential and to a stretching force. Usually data obtained from these kind of experiments are compared with mean field theoretical calculations. We showed in this Chapter that mean field is often very unreliable in making predictions. In particular, we showed that for finite chains the transition is not an all-or-none transition, rather it proceeds in a multi-step fashion. It is to be noted that this multi-step character has nothing to do with the expectations that dynamically a globule might get trapped in metastable configurations during its opening: our results show that (for sufficiently short chains) at intermediate values of the force the *optimal* configuration (global minimum of the internal energy) is neither a compact globule nor an open state, but a state which is both elongated and has a large number of contacts: we might call the ensemble of such states a 'marginally compact phase' to underscore the strict analogy with what found in the second Part of the thesis. So above a crossover value for the chain length – for very long chains – in three dimensions, mean field theory is recovered and the transition is thermodynamically first order, but below it marginally compact states show up. Remarkably, among the selected ground states we find a regular helix. This fact does not depend on the potential details provided that the attraction range is of the order of the distance between successive beads. This is an intriguing fact because it suggests that helices are naturally chosen whenever there is the requirement that a polymer must withstand a significant applied force and still have portion of the chains which are able to form non covalent or hydrophobic interactions. This would concur in explaining why helices recur so often in nature. We also found it pleasing that our multi-step non-mean field theory is in qualitative, although non quantitative, agreement with experimental data from Ref. [13], which were not interpreted before. In this experiment the authors observed that for very long collapsed DNA molecules the force versus elongation characteristic curves have a single plateau and are in good agreement with mean field theory, whereas this is no longer true for data corresponding to somewhat shorter DNA's (roughly by a factor of 10). In this last case a series of peaks was observed. These oscillations mean that the elongation versus the stretching force evolves in a multi-step fashion as shown in the thesis. Though there is hysteresis and hence dynamical effects will have affected the data, the agreement remains qualitative but significant. In two dimensions, on the other hand, we find that marginally compact states persist up to infinitely long chains. This drives the thermodynamic transition second order. A real space renormalization

group argument can explain the nature of the thermodynamic unfolding in different dimensions satisfactorily. We hope that these results on polymer stretching can stimulate other experiments to verify whether helices are possible structures before the polymer is completely stretched. Theoretically, we have not treated dynamical effects: in the future we plan to consider them in our models in order to find whether they can spoil possible experimental verification of our predictions.

The second part of this thesis has been devoted to the development and characterization of a model to describe thick polymers. This was meant as a coarse-grained model for protein folding which takes into account only the common characteristics of all proteins. Real proteins indeed at first approximation are schematized by chains, whose beads stand for the C_α atoms in the polypeptide backbone. At every site (C_α atom) there is a hard core sphere (which acts in usual models on the other non-consecutive beads only) so that the chain is self-avoiding. Moreover, in conventional 'protein folding' coarse grained models all C_α interact via a (generally pair dependent) compacting potential. However, the presence of side chains and all other atoms imposes additional steric constraints on the allowed configuration of a chain which mimics a polypeptide. A way to encompass this is to replace a self-avoiding chain by a self-avoiding tube with a non-zero effective thickness.

First, in Chapter 4, we have shown that in order to have a model of self-avoiding interacting curves with a non-zero thickness and with a well-defined continuum limit (without singularities), it is necessary to introduce three body constraints in the potential. These three body terms have nothing to do with those which arise e.g. in Edwards's model and which come from the virial expansion: here they have a physical meaning as their typical scale gives the thickness of the polymer.

In Chapter 5 we have discussed the optimal configurations which arise when we want to compactify *discrete* thick polymers. Our main result is that if we tune the tube thickness to be of the order of the attraction range, the different portions of the polymer need to position themselves parallel to each other, and we obtain 'marginally compact structures' which are still able to shield water from their core but have also internal wiggle room for their side chains. When we go from compact to swollen, we thus encounter again a marginally compact phase. From thinner to thicker, the marginally compact structures come from helices to sheets, passing through a range of interpolating motives, namely helices of strands and saddles. All of them are found as more common (helices and sheets) or less common (helices of strands and saddles) naturally occurring secondary structures of biopolymers. We stress that no coarse grained model, which neglects chemical details, had been found before to have a

protein-like ground state (apart from the ad-hoc constructed Go-model). In addition, in our model no sequence specificity has been used in order to find the optimal states. For longer tubes, unless we are near the onset of planarity, we observe a crossover to a nematic-like state, rather similar to the structures met in amyloid fibers or in protein aggregates, which are usually linked to the inability to correctly fold of a protein consequent to a disease. It is interesting that the crossover to these states is controlled by polymer length, as it is similar to what occurs in real situations, where the amyloid or aggregate formation is enhanced by increasing the concentration of polymers in solution.

In Chapter 6, we have finally studied the thermodynamics of thick polymers. By mean field theory and Monte-Carlo simulations, we have identified two regimes. The first is that of 'thin' tubes: decreasing T, the tube first undergoes a standard Θ collapse (second order), then a second transition at a lower T from an isotropic globule into a 'nematic' phase (first order) where isotropy is broken. In the second regime the tube is 'thick' and there is a direct first order transition to the nematic phase (which for small N gives protein-like structures).

Coming back to the questions raised in the Introduction, we see now that the proposed model of a thick polymer with hydrophobic interactions may shed light on some of them. First, it is no longer true that a thick homopolymer when it is collapsed has a highly degenerate ground state: if its thickness is tuned, the optimal state is practically unique. Second, the free energy landscape is highly affected by the three body constraints: in the marginally compact phase there are only a few minima among which to choose, and the dynamics will thus be more efficient and less likely to get stuck. Third, our results occur for *homopolymers* and thus suggest that, regardless of the aminoacidic sequence, many different proteins may share the same secondary structural elements provided their effective thickness is similar. Last, the observed protein versatility might be connected to the fact that, as we have seen, the 'twilight zone' in which marginally compact structures exist is relatively small and in order to switch from an α helix to a β sheet the thickness must be increased only slightly. Even if these observations are to some extent speculative, we believe that our results are encouraging and convincing enough to prove that thickness is an important feature to consider or incorporate in coarse-grained models for protein folding subsequent to ours. Of course, further developments will be needed. In particular, though we obtain faithful analogues of protein secondary structures, upon compaction of bigger chains, as noted above, we obtain nematic structures, we never find tertiary folds within the present scheme. It would be both important and interesting to see how our model can be minimally expanded in order to include them. It is likely that a

dishomogeneity in the local thickness of the tube and/or in the compacting potential, due to aminoacid diversity, is needed. Preliminary results in this direction are rather promising in particular for the stabilization of bundles of helices, whereas mixed α/β folds are more elusive. In any case, the size of the polymers needed is larger and simulations are harder and more time consuming. We hope that our results on marginally compact structures and on thick polymers will spur interest and further more advanced work theoretically and, also, experimentally.

Acknowledgments

I would like to warmly thank Amos both for proposing to me very interesting and stimulating problems and for his help and his friendly and encouraging attitude that he had regarding me and all the work I have done in my PhD. This was extremely important when things became more difficult. I would also like very much if possible to continue collaborating with him as I believe that I would never end learning things from him and enjoying his ideas.

My deepest thank goes to my wife Ilenia for her invaluable company, and also for bearing me when ultimating this thesis. In this last task she was surely helped by our cat Ortica to whom as well I owe many thanks.

Also my former family (my mother, father, brother and dog) deserves heartly thanks for (though skeptically in the beginning!) supporting my choices in life.

I also like to specially thank Flavio and Somen, who taught me an enormous amount of things on physics, particularly on numerical algorithm and the renormalization group, and who have always behaved to me as true friends as well.

During my PhD I have had the pleasure and luck to collaborate also with Angelo, Antonio, Enzo and Somen for the topics in the first part of the thesis, and with Alessandro (who is also my newest roommate), Antonio, and (of course!) Jayanth for the topics on thick polymers. I am now collaborating with Cristian and Hoang from Amos's group. Each of them taught me something important. I also have enjoyed illuminating discussions with Fabio, who was in Amos's group when I came to Trieste.

During my stay here in Trieste I have known many great guys whom I hope to stay in touch with. Among others are my roommates in the ICTP office: Alessandro, Angelo and Hoang who have already been listed above, and my former roommates Gianluca and Gianni, with whom I shared many discussions within and without physics; as well as my Padova friends Fabrizio and Guido who have also been in Trieste these years.

And, of course, my stay in Trieste would not have been so nice without the glori-

ous Sissa football team (headed by Marco and then Fabio), of which I was goalkeeper for three years, and without the very warm company of my CM-SBP friends Alexei, Davide, Evgheny, Giuliano, Luca, Mahdi, Matteo, Michele, Oleg, Paola, Paul, together with the old stars Andemar and Lorenzo.

I acknowledge moral help by my old Cittadella friends Andrea and Cristiano (my witness, who also made me have some fun losing to him at chess!), and by my university friends Carlo, Lele and Massimo, as well as the e-mails and phones by my old schoolmates Alessandro, Eddy, Stefano, who have done so much to keep in touch with me despite me.

Bibliography

[1] J. D. Watson, F. H. C. Crick, *Nature* **171**, 737 (1953).

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

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

[4] C. Bustamante, J. C. Macosko, G. J. L. Wuite, *Nature Reviews* **1**, 130 (2000).

[5] E. D. Siggia, J. F., Marko, *Macromolecules* **28**, 8759 (1995).

[6] M. Reif, H. Clausen-Schaumann, H. E. Gaub, *Nature Struc. Biol.* **6**, 346 (1999).

[7] U. Bockelmann, P. Thomen, B. Essevaz-Roulet, V. Viasnoff, F. Heslot, *Biophys. J.* **82**, 1537 (2002).

[8] A. Kishino, T. Yanagida, *Nature* **34**, 74 (1988).

[9] S. B. Smith, Y. J. Cui, C. Bustamante, *Science* **271**, 795 (1996).

[10] J. Liphardt, B. Onoa, S. B. Smith, I. Tinoco, C. Bustamante, *Science* bf 292, 733 (2001).

[11] M. Rief *et al.*, *Science* **276**, 1109 (1997), M. S. Z. Kellermayer *et al.*, *Adv. Exp. Med. Biol.* **481**, 111 (2000).

[12] A. F. Oberhauser, P. K. Hansma, M. Carrion-Vasquez, J. M. Fernandez, *Proc. Natl. Acad. Sci. USA* **98**, 468 (2001).

[13] C. G. Baumann *et al.*, *Biophys. J.* **78**, 1965 (2000).

[14] B. J. Haupt, T. J. Senden, E. M. Sevick, *Langmuir* **18**, 2174 (2002).

[15] A. Kornberg, T. Baker, *DNA Replication* (3rd ed., W. H. Freeman) (1992).

[16] S. M. Bhattacharjee *J. Phys. A* **33**, L423, and 9003(E) (2000); S. M. Bhattacharjee, *Ind. J. of Phys. A.* **76A**, 69 (2002).

[17] D. K. Lubensky, D. R. Nelson, *Phys. Rev. Lett.* **85**, 1572 (2000).

[18] E. Boye, T. Stokke, N. Kleckner, K. Skarstad, *Proc. Nat. Acad. Sci. USA* **93**, 12206 (1996); W. M. Xia, W. Dowhan, *Proc. Nat. Acad. Sci. USA* **92**, 783 (1995) .

[19] J. Marx *Science* **270**, 1585 (1995) ; ibid. 1667, 1671,1674.

[20] B. Essevaz-Roulet, U. Bockelmann, F. Heslot, *Proc. Nat. Acad. Sci. USA* **94**, 11935 (1997); B. Essevaz-Roulet, U. Bockelmann, F. Heslot, Phys. Rev. E **58** (1998) 2386.

[21] R. E. Thompson, E. D. Siggia, *Europhys. Lett.* **31**, 335 (1995).

[22] S. M. Bhattacharjee, D. Marenduzzo, *J. Phys. A* **35**, L349 (2002).

[23] M. Peyrard, *Europhys. Lett.* **44**, 271 (1998).

[24] C. Branden, J. Tooze, *Introduction to Protein Structure*, (2nd edition, Garland Publishing) (1999).

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

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

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

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

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

[30] J. Miller, C. Zeng, N. S. Wingreen, C. Tang, *Proteins* **47**, 506 (2002).

[31] C. Micheletti, F. Seno, A. Maritan, J. R. Banavar, *Proteins* **32**, 80 (1998).

[32] N. Go, *Macromolecules* **9**, 535 (1976); N. Go, H. Abe, *Biopolymers* **20**, 991 (1981).

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

[34] M. C. Tesi, E. J. J. van Rensburg, E. Orlandini, S. G. Whittington, *J. Stat. Phys.* **82**, 155.

[35] R. M. Wartell, A. S. Benight, *Phys. Reports* **126**, 67 (1985).

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

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

[38] G. N. Ramachandran, C. Ramakrishnan *et al. J. Mol. Biol.* **7**, 95 (1963).

[39] P. J. Flory, *Principles of Polymer Chemistry*, Cornell University Press, Ithaca, (1971).

[40] P. G. de Gennes, *Scaling concepts in polymer physics*, Cornell University Press, Ithaca (1979).

[41] M. Doi, S. F. Edwards, *The theory of polymer dynamics*, Clarendon Press, New York (1993).

[42] K. Huang, *Statistical Mechanics*, John Wiley, New York (1987).

[43] A. D. Sokal, *Nucl. Phys. B*, Suppl. **47**, 172 (1996).

[44] B. Duplantier, H. Saleur, *Phys. Rev. Lett.* **59**, 539 (1987).

[45] C. Vanderzande, A. L. Stella, F. Seno. *Phys. Rev. Lett.* **67**, 2757 (1991).

[46] H. Meirovitch, H. A. Lim, *Phys. Rev. Lett.* **62**, 2640 (1989).

[47] B. D. Hughes, *Random Walks and Random Environments*, Vol.2 (Clarendon Press, Oxford) (1995).

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

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

[50] M. Reif, H. Clausen-Schaumann, H. E. Gaub, *Nature Struc. Biol.* **6**, 346 (1999).

[51] D. Anselmetti, J. Fritz, B. Smith, X. Fernandez-Busquets, *Single Mol.* **1**, 53 (2000) .

[52] T. Strunz, K. Oroszlan, R. Schfer, H. H. Güntherodt, *Proc. Nat. Acad. Sci. USA* **96**, 11277 (1999).

[53] B. Maier, D. Bensimon, V. Croquette, *Proc. Nat. Acad. Sci. USA* **97**, 12002 (2000);

[54] G. V. Shivashankar, M. Feingold, O. Krichevsky, A. Libchaber, *Proc. Nat. Acad. Sci. USA* **96**,7916 (1999).

[55] G. J. L. Wuite, S. B. Smith, M. Young, D. Keller, C. Bustamante, *Nature* **404**, 103 (2000) .

[56] D. Poland, H. Scheraga, *J. Chem. Phys.* **45**, 1464 (1966).

[57] D. Marenduzzo, A. Trovato, A. Maritan, *Phys. Rev. E* **64**, 031901 (2001).

[58] E. Orlandini, S. M. Bhattacharjee, D. Marenduzzo, A. Maritan, F. Seno, *J. Phys. A* **34**, L751 (2001).

[59] D. Marenduzzo, S. M. Bhattacharjee, A. Maritan, E. Orlandini, F. Seno, *Phys. Rev. Lett.* **88**, 028102 (2002).

[60] K. L. Sebastian, *Phys. Rev. E* **62**, 1128 (2000) .

[61] H. Zhou, e-print cond-mat/0007015.

[62] J. Z. Y. Chen, e-print cond-mat/0111462.

[63] R. M. Neumann, e-print cond-mat/0201539.

[64] E. A. Mukamel, E. I. Shakhnovich, e-print cond-mat/0108447.

[65] D. K. Lubensky, D. R. Nelson, *Phys. Rev. E* **65**, 031917 (2002).

[66] S. Cocco, R. Monasson, *Phys. Rev.Lett.* **83**, 5178 (1999).

[67] S. Cocco, R. Monasson, J. F. Marko, *Proc. Natl. Acad. Sci. USA* **98**, 8608 (2001).

[68] S. Cocco, R. Monasson, J. F. Marko, *Phys. Rev. E* **65**, 041907 (2002).

[69] M. S. Causo, B. Coluzzi, P. Grassberger, *Phys. Rev. E* **62**, 3958 (2000)

[70] Y. Kafri, D. Mukamel, L. Peliti, *Phys. Rev. Lett.*, **85**, 1572, (2000).

[71] Y. Kafri, D. Mukamel, L. Peliti, *Eur. Phys. J. B* **27**, 135 (2002).

[72] Y. Kafri, D. Mukamel, L. Peliti, *PhysicaA* **306**, 39 (2002).

[73] E. Carlon, E. Orlandini, A. L. Stella *Phys. Rev. Lett.* **88**, 198101 (2002).

[74] M. Baiesi, E. Carlon, A. L. Stella, *Phys. Rev. E* **65**, 021804 (2002).

[75] N. Madras, A. Sokal, *J. Stat. Phys.* **47**, 573 (1987).

[76] E. Orlandini, F. Seno, A.L. Stella, *Phys. Rev. Lett.* **84**, 294 (2000).

[77] R.J. Rubin, *J. Chem. Phys.* **43**, 2392 (1965).

[78] S. Lifson, *J. Chem. Phys.* **40**, 3705 (1964).

[79] See, *e.g.*, A. L. Barabasi, H. E. Stanley, *Fractal Concepts in Surface Growth* (Cambridge University Press) (1995).

[80] S. M. Bhattacharjee, F. Seno, *J. Phys. A* **34**, 6375 (2001).

[81] B. Derrida, Y. Pomeau, *Phys. Rev. Lett.* **48**, 627 (1982). See also B. Derrida, *J. Stat. Phys.* **31**, 433 (1983) and F. Solomon *Annals of Probability* **3**, 1 (1975).

[82] A. Halperin, E. B. Zhulina, *Europhys. Lett.* **15**, 417 (1991).

[83] T. A. Vilgis, A. Johner, J. F. Joanny, *Eur. Phys. J. E* **2**, 289 (2000).

[84] P. L. Geissler, E. I. Shakhnovich, *Phys. Rev. E* **65** 056110 (2002); cond-mat/0111317.

[85] D. K. Klimov, D. Thirumalai, *Proc. Natl. Acad. Sci. USA* **96**, 6166 (1999).

[86] M. Cieplak, T. X. Hoang, M. O. Robbins, cond-mat/0112200, cond-mat/0112201.

[87] C. Vanderzande, *Lattice models of polymers*, Cambridge University Press (1998).

[88] P. Grassberger, H. Hsu, *Phys. Rev. E* **65**, 031807 (2002).

[89] D. Marenduzzo, A. Maritan, F. Seno, *J. Phys. A* **35**, L233 (2002).

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

[91] B. Derrida, *J. Phys. A* **14**, L5 (1981).

[92] A. Trovato, F. Seno, *Phys. Rev. E* **56**, 131 (1997).

[93] H. Saleur, *J. Stat. Phys.* **45**, 419 (1986).

[94] H. Zhou, cond-mat/0112090.

[95] M. Henkel, G. Schutz, *J. Phys. A* **21**, 2617 (1988)

[96] S. Lise, A. Maritan and A. Pelizzola, *Phys. Rev. E* **58**, R5241 (1998).

[97] D. Dhar, J. Vannimenus, *J. Phys. A* **20**, 199 (1987).

[98] A. Maritan, F. Seno, A. L. Stella, *Physica A* **156**, 679 (1989).

[99] Yamakawa, H., Modern theory of polymer solutions, Harper and Row, New York, 1971.

[100] G. des Cloiseaux, J. F. Jannink, *Polymers in solution: their modeling and structure*, Clarendon Press, Oxford, 1990.

[101] D. R. Nelson, T. Piran, S. Weinberg, Editors, *Statistical mechanics of membranes and surfaces*, Volume 5 of Jerusalem Winter School for Theoretical Physics, World Scientific, Singapore, 1989.

[102] K. J. Wiese, *Polymerized membranes, a review*, in Phase transitions and critical phenomena, Vol. 19, edited by Domb, C. and Lebowitz, J. L., Academic Press, New York, 2000.

[103] P. G. de Gennes, *Phys. Lett. A* **38**, 339 (1972).

[104] J. Zinn-Justine, *Quantum field theory and critical phenomena*, (3rd edition, Oxford, University Press) (1996).

[105] O. Gonzalez, J. H. Maddocks, *Proc. Natl. Acad. Sci. USA* **96**, 4769-4773 (1999).

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

[107] J. R. Banavar, O. Gonzalez, J. H. Maddocks, and A. Maritan, submitted to *J. Stat. Phys.*.

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

[109] C. Cerf, A. Stasiak, *Proc. Nat. Acad. Sci. USA* **97**, 3795 (2000).

[110] P. Pieranski, S. Przybyl, A. Stasiak, *Eur. Phys. J. E* **6**, 123 (2001).

[111] J. Yan, M. O. Magnasco, J. F. Marko, *Nature* **401**, 932 (1999).

[112] The protein data bank is available on line at http://www.rcsb.org/pdb.

[113] F. G. Wang, D. P., Landau, *Phys. Rev. Lett.* **86**, 2050 (2001).

[114] J. R. Banavar, A. Maritan, C. Micheletti, A. Trovato, *Proteins* **47**, 315 (2002).

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

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

[117] U. Baumann, S. Wu, K. M. Flaherty, D. B. and Mckay, *Embo J.* **12**, 3357 (1993).

[118] J. R. Banavar, A. Flammini, D. Marenduzzo, A. Maritan, A. Trovato, accepted for publication in *ComPlexUs* (2002).

[119] L. Onsager, *Ann. NY Acad. Sci.* **51**, 627 (1949).

[120] G. J. Vroege, H. N. W. Lekkerkerker, *Rep. Prog. Phys.* **55**, 1241 (1992).

[121] E. Pitard, T. Garel, H. Orland, *J. Phys. I* **7**, 1201 (1997).

[122] S. Doniach, T. Garel, H. Orland, *J. Chem. Phys.* **105**, 1601 (1996).

[123] J. Chahine, H. Nymeyer, V. B. P. Leite, N. D. Socci, J. N. Onuchic, *Phys. Rev. Lett.* **88**, 169101 (2002).

[124] K. W. Plaxco *et al.*, *Nat. Struct. Biol.* **6**, 554 (1999); O. B. Ptitsyn, *Adv. Protein Chem.* **47**, 83 (1995).

[125] A. M. Ferrenberg, R. H. Swendsen, *Phys. Rev. Lett.* **61**, 2635 (1988).

[126] Y. Q. Zhou, M. Karplus, J. M. Wichert, C. K. Hall, *Journal of Chemical Physics* **107**, 10691 (1997).

[127] F. Seno, private communication.

[128] A. J. Barrett, C. Domb, *J. Stat. Phys.* **77**, 491 (1994).

Appendix A

Monte-Carlo moves

The configuration space we want to sample is the set, $\mathcal{R}_n$, of $N-$beads discrete chains, $\{\vec{R}_i\}_{i=0,\ldots,N-1}$, having a fixed constant distance between consecutive beads, which we take to be 1. In all cases considered in this thesis there is a hard core repulsion at a distance $2R_{hc}$ between non-consecutive beads, while the detailed form of the potential, V, depends on the specific problem we want to tackle, and is defined in the body of this thesis every time a Monte-Carlo run is performed. Another possible way to define the configuration of a chain is via the Ramachandran angles introduced in Chapter 1, but here we have not used this angle formulation mainly for simplicity of the computer code.

We have employed four different types of moves in the Monte-Carlo dynamics. All moves do not violate the constraint of fixed constant difference between consecutive beads along the chain, but we have to check whether the propose updated configuration satisfies other constraints that we want to enforce, which in our case sums up to checking self-avoidance. When the trial configuration violates self-avoidance, it is rejected before performing the Metropolis test on it (see Chapter 1). We now list the moves we used and discuss their local/global character, with respect to the number of beads involved in the move, as is customary in polymer physics [43, 33].

Cranckshaft move: Select randomly two beads i, j ($i < j$) such that $j - i \leq n_c + 2$, with $n_c \ll N - 1$ (most frequently we have used $n_c = 6$). Then rotate beads $i + 1, \ldots, j$ of an angle $\Delta\phi_c$ around the axis $\vec{R}_j - \vec{R}_i$. The angle $\Delta\phi_c$ is chosen randomly with a uniform distribution in the interval $[-\Delta\phi_m/2, \Delta\phi_m/2]$. This is a *local* move, since only n_c beads are involved. So it may well correspond to a *physical* move of the polymer in a real dynamics (as the one used for DNA unzipping in Chapter 2 on the lattice).

Reptation move: This kind of move is also known as the *slithering-snake* move. It consists in deleting n_r ($n_r = 1, \ldots, N_r$, with N_r most frequently chosen as 5) beads from one end of the chain and appending them, after a rotation around a random axis of an angle $\Delta\phi_r$ chosen uniformly from 0 to π. Reptation is a *bilocal* move, since it alters two disjoint small groups of consecutive beads of the chain. Reptation is used as a physical dynamical moves in models in which the polymer migrates in a fixed environment or is subject to some displacing field and is subject to no constraints.

Pivot move: Select randomly one bead i ($0 < i < N - 1$) 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. The rotation is around a random axis chosen with uniform probability among all the versors on the unit sphere and the pivot rotation angle is uniformly chosen between $-\Delta\phi'_m$ and $\Delta\phi'_m$ (which in general is different from $\Delta\phi_m$). Pivot moves are *global* ones, since they involve a rearrangement of a macroscopic portion of the chain. A pivot move is unlikely as a physical dynamical move.

Cut-and-rotate move: This move consists in cutting all the links of the chain, i.e. in breaking the chain $\{\vec{R}_i\}_{i=0,\ldots,N-1}$, in rotating randomly (as above, by an angle from 0 to a maximum value $\Delta\phi''_m$) a fraction ρ_{c-a-r} of the links, and then in reassembling them in the same order as before. This can be seen as equivalent in spirit to making many pivot moves simultaneously. Cut-and-rotate moves are expected to rearrange drastically the whole shape of the chain, thus ensuring a more efficient search in configurational space, though they have an acceptance rate which is lower that that of the above moves. For optimization it proved to be very helpful, while in equilibration calculation it was less efficient and we did not use it.

To conclude, we note that the efficiency of Monte Carlo dynamics may depend crucially on tuning the different control parameters which we have introduced, namely n_c, N_r, ρ_{c-a-r}, $\Delta\phi_m$, $\Delta\phi'_m$, $\Delta_p hi''_m$, possibly considering them as non-constant functions of the temperature. Efficiency also depends on the relative frequency of the different moves that we use. In general of course, the best efficiency for different purposes (i.e. optimization or equilibration) is attained for different values of the control parameters.

Appendix B

Ground state of clusters of interacting hard spheres

We consider spheres with a hard core R_{hc} subject to a compacting potential V between all pairs. The potential is an asymmetrical square well, identical to that of Eq.4.10. We consider the two and three-dimensional cases in detail below. By keeping the hard core radius fixed to the value $R_{hc} = 0.5$, we want to find the ground state (gs) of the potential for different values of R_1, the range of the attracting potential.

Some features of the solution to this problem do not depend on d and on other details and are worth being mentioned. First, the ground state is unique only for particular values of the contact potential range R_1. In general there is a degeneracy. There are windows or steps of values of R_1: the degeneracy is one only at the left edge of each plateau and then it increases (compare Fig. B.1 valid for this problem with. Fig. 5.2 which is the analogous one for the thick tube problem). Second, the recurring ground state is different for different steps unless in the trivial one-dimensional case. Third, note that when R_1 tends to $2R_{hc}$ one reduces to the well studied task of finding the structure giving the maximally compact packing of hard spheres: in the thermodynamic limit, in $d = 2$ this structure is the triangular lattice, in $d = 3$ it is the face centered cubic (fcc) lattice. Last, in contrast to the case discussed in Section III of the thick polymer, here there is in general a greater degeneracy: this suggests that the anisotropy present there is a key to gs selection and degeneracy reduction.

Two-dimensional case

In $d = 2$ it is well known that there is a unique way of packing hard disks so that the compaction is maximal. The optimal packing is performed by placing the centers

of the disks in a triangular lattice, for which every site has six nearest neighbours. This is consequently the ground state of our problem when $R_1 = 2R_{hc}$. For generic values of the attraction range R_1, we study what the best possible way to compact hard disks at that particular value of R_1 is. So we look for the ground state in the thermodynamic limit.

We expect that the ground state for big N is again a lattice because the thermodynamic limit must be translationally invariant (contrarily to the case of a finite size sample). We show in Fig. B.1 below the best possible lattice as a function of R_1. The first plateau corresponds to a triangular lattice. As R_1 increases this ground state is more and more degenerate. A measure of the degeneracy inside the class of two-dimensional lattice only is given by the height of the vertical segment in Fig. B.1. Of course this measure being limited to the sub-class of lattices, it gives an underestimate of the real degeneracy. When the attraction range becomes $\sqrt{2}$, it is more advantageous for the points to arrange themselves on a square lattice, because in this way also the second neighbours have a distance which falls within the potential well. Intriguingly, the system continues oscillating between the triangular lattices and other kinds of Bravais lattices also for bigger values of R_1 as is shown in Fig. B.1 (in the first few steps the competing lattices comprise the square lattice, but in general this is not true). For big enough (size-dependent) R_1 the degeneracy becomes so huge that the gs is again a gas as in the case $R_1 < 2R_{h.c}$. We stress once more that Fig. B.1 shows the best possible *lattice* for the compaction at a given attraction range R_1. In order to support our expectation that there does not exist a structure which for a given R_1 (in the thermodynamic limit) has a bigger energy than one of the lattices in Fig. B.1, we have performed Monte-Carlo simulated annealing on a system of a finite number of disks in a finite box ($N = 25$ disks in the structures of Fig. B.2).

Though we only deal with finite size samples, we can define effectively a 'bulk coordination' or 'bulk contact number' of the structures coming out from the simulations. One first definition might be the maximum number of contacts performed by any bead given that the surface spheres can make fewer contacts than the bulk particles. The 'bulk contact numbers' defined in this way of the structure found with the simulated annealing are below those of the $2d$ lattices except in two cases (fig. B.3). These two counterexamples both occur near the end of one step where degeneracy is high and are due to defects in which there can be more contacts locally. If we define the 'bulk contact number' as the maximum number of contacts made by a bead with the constraint that this maximum must be attained at least a given number k of times this quantity is always below the coordination of the optimal lattice at the corresponding R_1 (provided wee choose at least $k = 3$ in our example with $N = 25$).

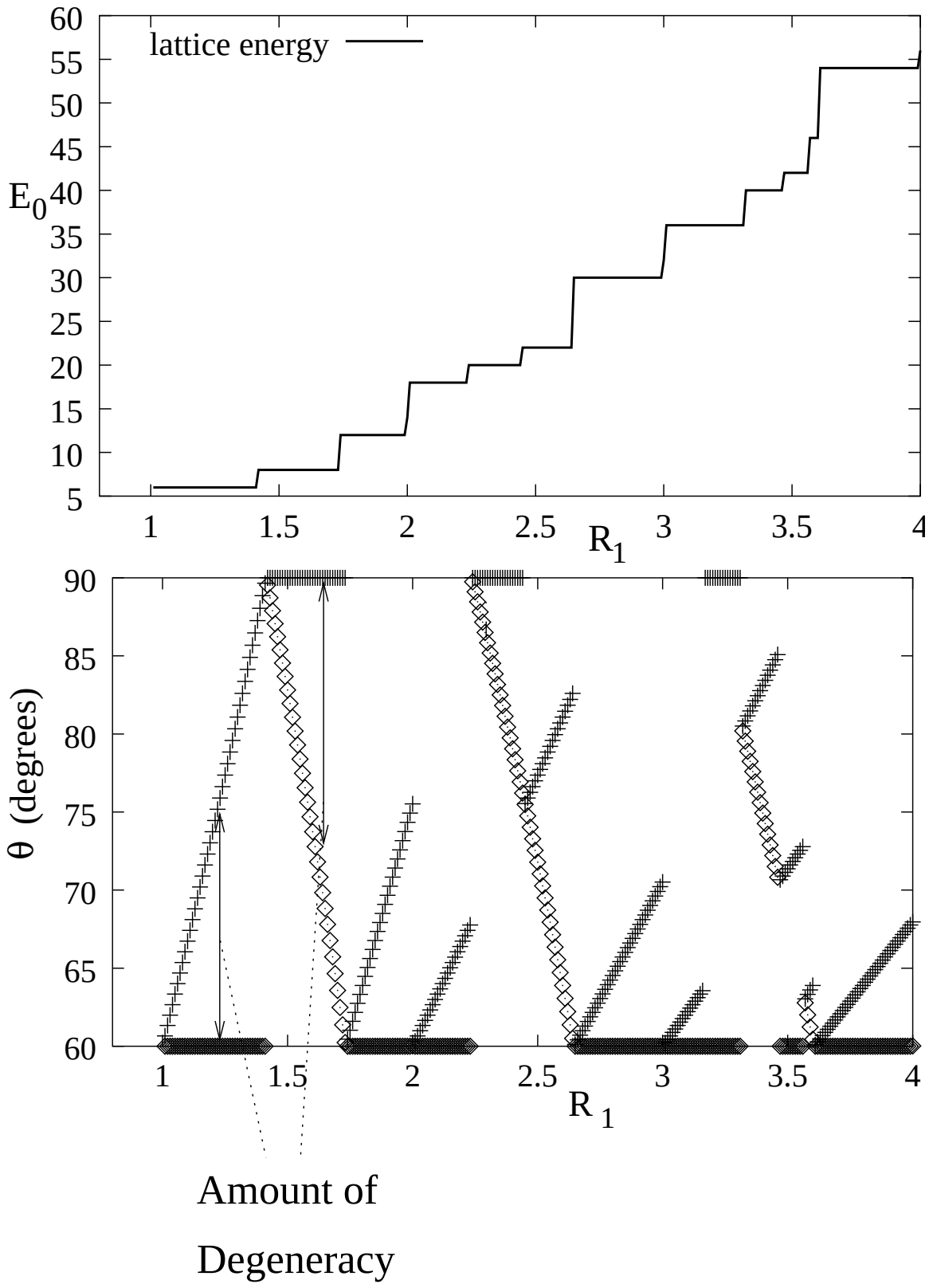


Figure B.1: Top figure: lattice energy vs. attraction range for two dimensional Bravais lattices. Bottom figure: this gives a measure of the degeneracy of the gs lattice.

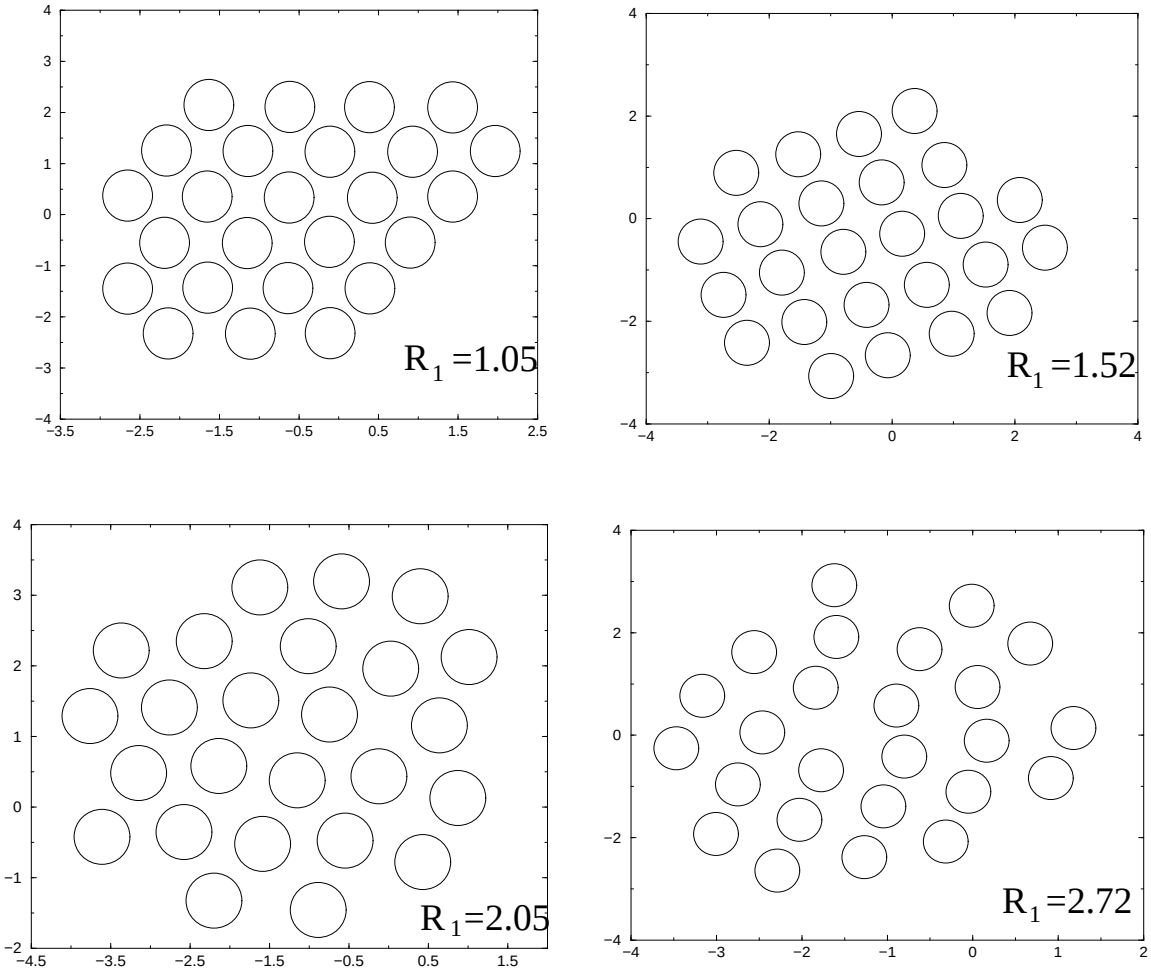


Figure B.2: Structures obtained as ground states of a box with 25 attracting hard disks, for various values of the attraction range R_1 (shown in the figure).

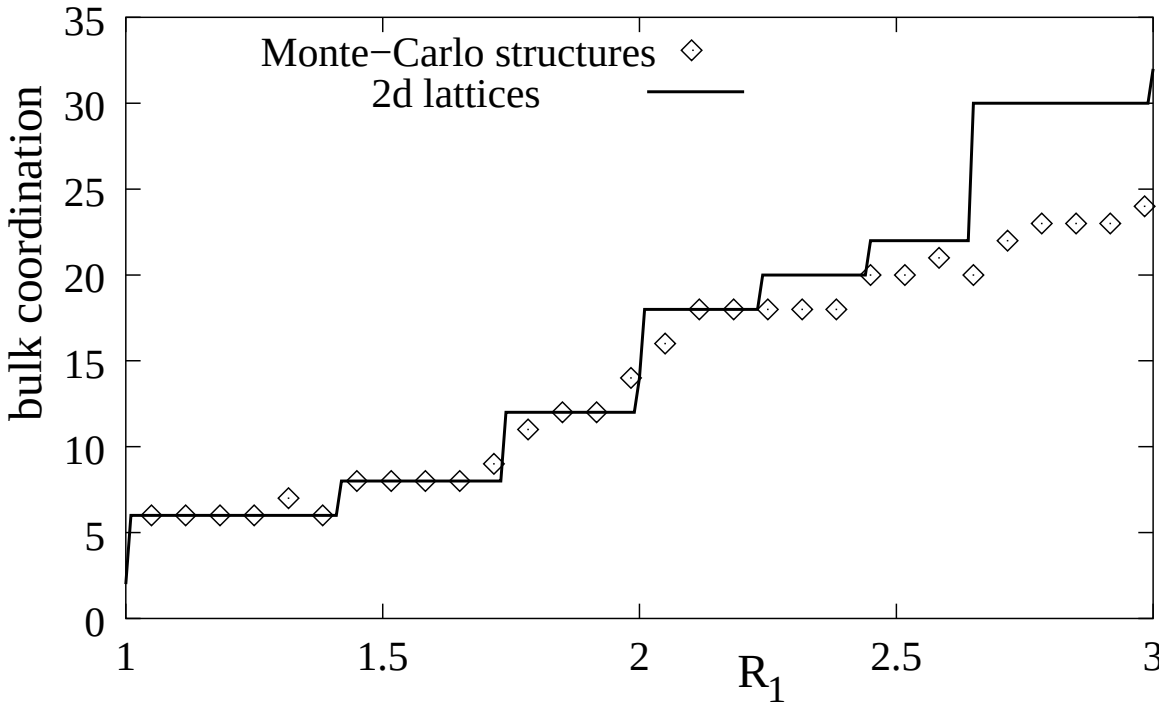


Figure B.3: Bulk coordination of lattices and Monte-Carlo ground-states.

Three-dimensional case

Though it has been proved rigorously only a few years ago, it has been conjectured since Kepler's time that the optimal packing in $d = 3$ for hard spheres is given by the face centered cubic crystal (fcc crystal) arrangement. In this Bravais lattice all the sites have 12 nearest neighbours so that the energy in the thermodynamic limit is $6N$.

In the three-dimensional case, however, the fcc lattice is not the unique ground state when $R_1 = 2R_{h.c}$. There are in fact an infinite series of degenerate ground states. These maximally compact states are found by superimposing one onto the other two-dimensional triangular lattices. If the configuration corresponding to the center of the spheres in the first layer is labelled 'A', then the second triangular lattice might position its centers either in the depression available between the spheres in the first layer (labelled 'B') or directly on top of the first layer ('A' again). Once these two layers are positioned, the position of the third triangular lattice might choose three different configurations: 'A' if it is placed on top of the first layer, 'B' if the spheres are put on top of the second layer, or 'C' if the centers are put into the de-

pressions unused by the second layer of spheres. The configuration repeated such that the following layers have successive configurations A and B, 'ABABAB' etc., is the hexagonal close packed lattice, or hcp; whereas the structure 'ABCABCABC...' is the fcc lattice. It is to be noted that the fcc lattice is the only Bravais lattice between those of this set of structures which maximally fill the space.

So our strategy for finding a guess for the ground states at varying attraction ranges is the following. First we look among all the 3d Bravais lattices (labelled by three angles: ϕ, between the first two Bravais vectors; θ, ψ the polar and azimuthal angle respectively for the third vector in the system generated by the first two vectors and the perpendicular to their plane) for the one which maximizes the potential V. The resulting plot of the energy versus R_1 is shown in Fig. B.4. Second, we try to position triangular or square lattice on top of each other and look for the optimal structures of this kind (not necessarily a Bravais lattice) obtained in this way. Third, as we cannot be sure that the ground state is found among either three-dimensional Bravais lattices or two-dimensional Bravais lattices laying one onto the other, we have performed Monte-Carlo simulations of hard spheres in a box.

By considering the bulk of the spheres cluster in practice we can simulate the thermodynamic limit of the systems. As in $d = 2$ with these simulations one is not able to improve over the energies found by enumerating the $3d-$ optimal Bravais lattices found before, meaning that the bulk number of pairwise contacts of the $3d$ Bravais lattices is greater than the corresponding Monte-Carlo generated structures. We can therefore consider the minima in Fig. B.4 with a sufficient degree of certainty to be the true ground states of our potential V.

In Fig. B.4 we show the gs energy found by following the strategy outlined above and also mention the corresponding structures for the first minima with increasing R_1 (the structure is known once we give the angles, as we are dealing with Bravais lattices). With respect to the two-dimensional case, there are many more states which come up with increasing R_1. In particular note that the third step corresponds to the simple hexagonal lattice; this situation corresponds to that studied in[126]. The authors there find a competition between the spheres either connected by a chain or free, in doing triangular and square lattices: this is precisely what happens in a hexagonal simple lattice, in which the unit cell is a prism in which two faces are triangles and the other are squares.

In Fig. B.5 we show some examples of ground state configuration found by packing $N = 15$ ($N = 64$) hard spheres. These finite size ground states do not have the symmetry of lattices, the number of constituent particles being so low that boundary effect become predominant and the chosen structures are those that lose less energy

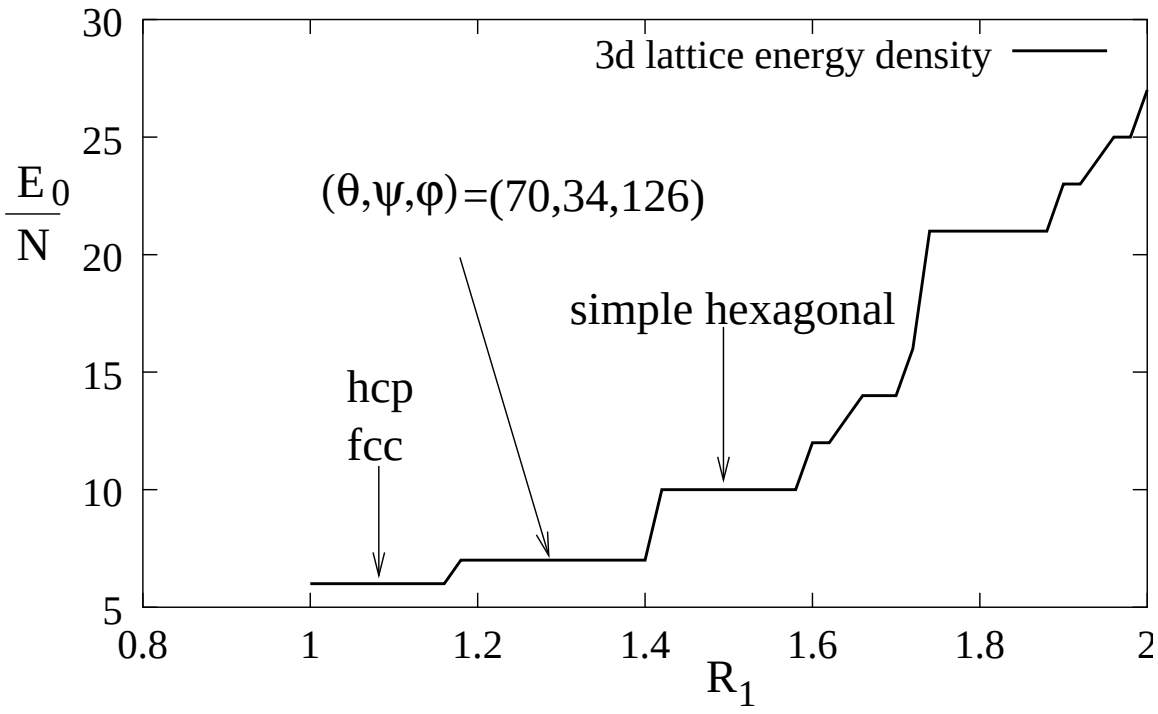


Figure B.4: Top: 3d Bravais lattice energy density for small values of R_1 with identification of ground state classes.

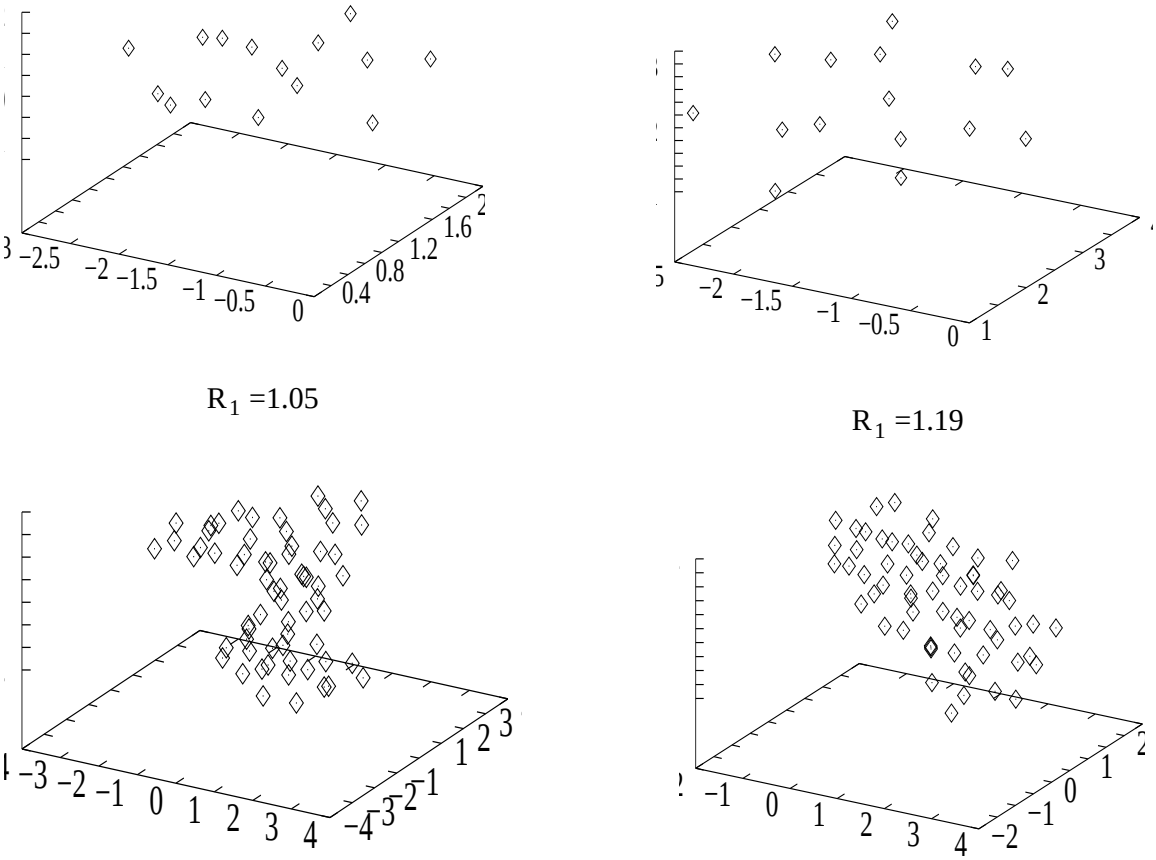


Figure B.5: We show in this figure typical minima resulting from simulated annealing.

due to surface exposure rather than those which make more contacts in the bulk. In order to find the ground states in Fig. B.5 we used simulated annealing with Monte-Carlo dynamics. It is to be noted that the Monte-Carlo dynamics is not trivial and one has to combine local one sphere moves with some global cluster moves that prevent the systems from dividing itself in several subcluster not interacting with one another.

Appendix C

Mean field calculations

In this Appendix we derive the formulas for the mean field free energy Eqs. 6.3, 6.4. We consider only the case of the chain of coins, the case of the polymer with triplet constraint is similar. We start from the partition function in Eq. 6.1:

$$\mathcal{Z} = \int D\psi \int \prod_{i=1}^{N} d\vec{r}_i \int \prod_{i=1}^{N-1} d\vec{t}_i \delta\left(|\vec{r}_{i+1} - \vec{r}_i| - 1\right) \tag{C.1}$$
$$\delta\left(\vec{t}_i - \vec{r}_{i+1} + \vec{r}_i\right) \delta\left(\psi - \sum_{i=1}^{N} \delta(\vec{r} - \vec{r}_i)\delta(\vec{t} - \vec{t}_i)\right)$$
$$e^{\left(\sum_{i<j;i,j=1}^{N} V_{2b}(\vec{r}_i - \vec{r}_j)\right)} \prod_{i<j;i,j=1}^{N} (1 + f_{ij}^{(1)}) \prod_{i<j;i,j=1}^{N} (1 + f_{ij}^{(2)}).$$

We can rewrite it as:

$$\mathcal{Z} = \int D\psi \int \prod_{i=1}^{N} d\vec{r}_i \int \prod_{i=1}^{N-1} d\vec{t}_i \delta\left(\vec{t}_i - \vec{r}_{i+1} + \vec{r}_i\right) \tag{C.2}$$
$$\delta\left(|\vec{r}_{i+1} - \vec{r}_i| - 1\right) \delta\left(\psi - \sum_{i=1}^{N} \delta(\vec{r} - \vec{r}_i)\delta(\vec{t} - \vec{t}_i)\right) \langle \mathcal{Z}_{int} \rangle(\psi).$$

We have called:

$$\langle \mathcal{Z}_{int} \rangle(\psi) = \frac{\int_\psi \prod_i d\vec{r}_i d\vec{t}_i \prod_{i<j;i,j=1}^{N} (1 + f_{ij}^{(1)}) \prod_{i<j;i,j=1}^{N} (1 + f_{ij}^{(2)})}{\int_\psi \prod_i d\vec{r}_i d\vec{t}_i}, \tag{C.3}$$

where with the symbol $\int_\psi$ we mean integration with respect to the density $\psi(\vec{r},\vec{t})$:

$$\int_\psi \prod_i d\vec{r}_i d\vec{t}_i \equiv \int \prod_i d\vec{r}_i d\vec{t}_i \delta\left(|\vec{r}_{i+1}-\vec{r}_i|-1\right)\delta\left(\vec{t}_i-\vec{r}_{i+1}+\vec{r}_i\right) \qquad \text{(C.4)}$$
$$\delta\left(\psi-\sum_{i=1}^{N}\delta(\vec{r}-\vec{r}_i)\delta(\vec{t}-\vec{t}_i)\right).$$

Through a standard virial or cluster expansion[120] truncated at second order in $f^{(1)}$ and at first in $f^{(2)}$[1], we get:

$$\log\left[\langle\mathcal{Z}_{int}\rangle(\psi)\right] \sim -\frac{1}{2V}\int d\vec{r}_1\int d\vec{r}_2\psi(\vec{r}_1)\psi(\vec{r}_2)f^{(1)}(\vec{r}_1,\vec{r}_2) \qquad \text{(C.5)}$$
$$-\frac{1}{3V}\int d\vec{r}_1\int d\vec{r}_2\int d\vec{r}_3\psi(\vec{r}_1)\psi(\vec{r}_2)\psi(\vec{r}_3)f^{(1)}(\vec{r}_1,\vec{r}_2)f^{(1)}(\vec{r}_2,\vec{r}_3)f^{(1)}(\vec{r}_1,\vec{r}_3)$$
$$-\frac{1}{2V}\int d\vec{r}_1\int d\vec{r}_2\psi(\vec{r}_1)\psi(\vec{r}_2)f^{(2)}(\vec{r}_1,\vec{r}_2),$$

where by $\psi(\vec{r})$ we mean $\int d\vec{t}\psi(\vec{r},\vec{t})$. Now in Eq. C.2 we use the well-know identity:

$$\delta\left(\psi-\sum_{i=1}^{N}\delta(\vec{r}-\vec{r}_i)\delta(\vec{t}-\vec{t}_i)\right) = \qquad \text{(C.6)}$$
$$\int D\hat{\psi}e^{i\int d\vec{r}\int d\vec{t}\hat{\psi}(\vec{r},\vec{t})\psi(\vec{r},\vec{t})-\sum_{i=1}^{N}\hat{\psi}(\vec{r}_i,\vec{t}_i)}.$$

In this way we can rewrite eq. C.2 as in Eq. 6.3 with:

$$\mathcal{H}\left(\psi,\hat{\psi}\right) = -\frac{1}{2}\int d\vec{r}_1 d\vec{r}_2\psi(\vec{r}_1,\vec{t}_1)V_{2b}(\vec{r}_1-\vec{r}_2)\psi(\vec{r},\vec{t})+i\int d\vec{r}d\vec{t}\hat{\psi}(\vec{r},\vec{t})\psi(\vec{r},\vec{t}) \qquad \text{(C.7)}$$
$$+\log\left(\zeta(\hat{\psi})\right)-\frac{1}{2V}\int d\vec{r}_1\int d\vec{r}_2\psi(\vec{r}_1)\psi(\vec{r}_2)f^{(1)}(\vec{r}_1,\vec{r}_2)$$
$$-\frac{1}{3V}\int d\vec{r}_1\int d\vec{r}_2\int d\vec{r}_3\psi(\vec{r}_1)\psi(\vec{r}_2)\psi(\vec{r}_3)f^{(1)}(\vec{r}_1,\vec{r}_2)f^{(1)}(\vec{r}_2,\vec{r}_3)f^{(1)}(\vec{r}_1,\vec{r}_3)$$
$$-\frac{1}{2V}\int d\vec{r}_1\int d\vec{r}_2\psi(\vec{r}_1,\vec{t}_1)\psi(\vec{r}_2,\vec{t}_2)f^{(2)}(\vec{r}_1,\vec{r}_2)$$

[1]The derivation of Eq. C.5 implies neglecting the chain constraint in the interaction part. It is possible but more cumbersome to take into account the chain constraint properly also here[128]. To our purpose however however it suffi ces to show that one obtains a term of second order in ρ in the free energy.

where we have called $\zeta(\hat{\psi})$ the polymeric partition function, which reads:

$$\zeta(\hat{\psi}) \equiv \left(\int \prod_i d\vec{r}_i \delta\left(|\vec{r}_{i+1} - \vec{r}_i| - 1\right) e^{-i \sum_{i=1}^{N} \hat{\psi}(\vec{r}_i, \vec{t}_i)} \right)^{\frac{1}{N}} . \tag{C.8}$$

The saddle point equation read $\frac{\delta \mathcal{H}}{\delta \psi} = \frac{\delta \mathcal{H}}{\delta \hat{\psi}} = 0$. In this form they are not solvable explicitly. As in the text, we here make the 'mean field approximation' $\psi(\vec{r}, \vec{t}) \equiv \rho\phi(\vec{t}) \equiv \psi(\vec{t})$ (and consequently from the saddle point equations one finds $\hat{\psi}(\vec{r}, \vec{t}) \equiv \hat{\rho}\hat{\phi}(\vec{t}) \equiv \hat{\psi}(\vec{t})$). After this one gets $\zeta(\hat{\psi}) = \int d\vec{t} \delta\left(|\vec{t}| - 1\right) e^{-i\hat{\psi}(\vec{t})}$ and the equation $\frac{\delta \mathcal{H}}{\delta \hat{\psi}} = 0$ takes on the form of a self-consistent equation for ψ and $\hat{\psi}$:

$$\psi(\vec{r}, \vec{t}) = \rho \frac{e^{-i\hat{\psi}(\vec{t})}}{\int d\vec{t} \delta\left(|\vec{t}| - 1\right) e^{-i\hat{\psi}(\vec{t})}} \tag{C.9}$$

or equivalently:

$$i\hat{\psi}(\vec{t}) = \log\left(\frac{\langle e^{-i\hat{\psi}(\vec{t})} \rangle \psi(\vec{r}, \vec{t})}{\rho} \right) \tag{C.10}$$

where with $\langle \cdot \rangle$ we denote ensemble averaging with respect to the measure $\int d\vec{t} \delta\left(|\vec{t}| - 1\right) \cdot$. In this way we recover the free energy in Eqs. 6.4, by recalling also the formula for the mean excluded volume of two coins, with axis $\vec{t}_1$ and $\vec{t}_2$, and radius R_0 which is due to Onsager[119] and reads:

$$\frac{1}{2V} \int d\vec{r}_1 \int d\vec{r}_2 f^{(2)}(\vec{r}_1, \vec{r}_2; \vec{t}_1, \vec{t}_2) = 2\pi R_0^3 \left(1 - \left(\vec{t}_1 \cdot \vec{t}_2\right)^2\right)^{\frac{1}{2}} . \tag{C.11}$$