ISAS - INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

Folding, Stability and Design of proteins.

Thesis submitted for the degree of
"Doctor Philosophiæ"

CANDIDATE

Michele Vendruscolo

SUPERVISOR

Prof. Amos Maritan

October 1996

SISSA - SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI
TRIESTE
Strada Costiera 11

TRIESTE

Table of Contents

1 Introduction

1.1 Foreword

In this work we will address the problem of protein folding. Proteins are long chain molecules capable to fold in a well defined spatial structure, the *native state*, in which they are biologically functional. The native state has several intriguing features. The principal one is that, unlike non-biological polymers, it is stable against varying environmental conditions. Under normal physiological circumstances a protein shows only small scale fluctuations and retains its overall shape. This stability is marginal and stronger perturbations can promote unfolding, or denaturation, of the protein. This fact is possibly a product of evolution, since biological functions should be responsive to the environment and to the presence of intervening regulatory molecules. The delicate balance of forces responsible for folding is likely to be encoded within the one dimensional sequence of amino acids, that are the elementary entities forming the protein. Understanding the physical process underlying protein folding would be a major step in medicine, allowing for the design novel proteins with desired functionality, and for the comprehension of the structures and functions of newly identified sequences. In particular, an impressive effort is being devoted to sequentiate to complete genome of man (human genome project). Such knowledge would be hardly beneficial unless we will not be able to predict protein function protein from the knowledge of amino acid sequences. The recent interest in the physics of protein folding is partly inspired by developments in the statistical mechanics of disordered systems, in particular polymers and spin glasses. The expectation is that emergent features of a complex systems can be recognized by the study of model systems. Our attention will be focused on simple models of proteins with the aim to unravel organizing principles rather than on the detailed discussion of the chemistry of the amino acid sequences.

The correspondence between sequence and structure constitutes the so called protein folding problem and it is discussed in the second chapter of this work. We will deal with two complementary views that are currently debated. In the *foldability* approach, it is proposes that existent proteins are the outcome of evolution, by selection of those sequences that

manifest a propensity to fold rapidly. The opposite view, termed *designability*, it is asserted that the selection acts primarily on structures, by choosing only those conformations that can accommodate a stable state of a protein.

Since the biological activity of proteins is mainly controlled by their structure, a method to design structures by suitably tailoring of sequences would permit the engineering of artificial proteins with predetermined functionality. This issue, known as the *inverse* folding problem, is discussed in chapter three, where we are able to present its solution on general grounds.

In chapter four we further discuss how natural selection could have possibly produced existent proteins. We show that the protein design procedure by optimization of thermodynamic stability, introduced in the third chapter, produces also robustness against mutations for the designed sequences. In the model we introduced, mutations are of very general character and can represent evolutionary changes in the composition of the sequence or perturbations in the properties of the solvent in which protein are plunged. We introduce an alternative, evolution-oriented, protein design scheme based on the optimization of the stability threshold against mutations and we show that yields thermodynamic stability as a consequence. We suggest that such design, starting from existing proteins, can produce artificial homologous sequences with a better functionality. Moreover, we discuss how these findings can provide a possible explanation to the observed occurrence of families of protein folds from the analysis of protein structure databases.

In the first four chapter the study has been dwelt within lattice models of proteins. Although there are sensible reasons to believe that such models are reasonably appropriate for describing some features of proteins, it would be interesting to extend the analysis to off-lattice situations. In chapter five an efficient Monte Carlo method for the simulation of off-lattice polymers is presented and applied to simple models of proteins.

In the final chapter we address three issues currently under investigations. The first one is the determination of a suitable effective interaction potential between protein constituents. The approximative knowledge of such potential has been hampering progresses in the design of new proteins and in the understanding of the principles of protein folding. The second problem discussed is the determination of a numerical optimization technique to find the ground state of a given protein. Finally, we discuss how correlations in protein sequences could be the signature of the unknown folding code, and can reveal a part of its nature.

1.2 Overview of Protein Folding

1.2.1 Protein Structures

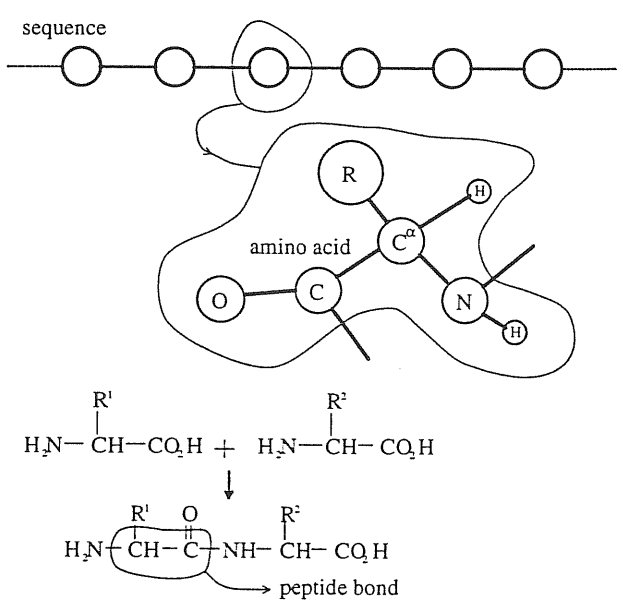

Figure 1.1: Schematic representation of a protein as a sequence of amino acids.

A protein is a long chain molecule whose building blocks are the 20 naturally occurring amino acids. The chemical structures of the amino acids are similar and are depicted schematically in Fig. 1.1. A central carbon atom C, traditionally labeled by α, is bonded to a side chain R. The specific chemical structure of the side chain R characterizes a particular species of amino acid. The synthesis of proteins is realized by sequential formation of peptide bonds between amino acids. The peptide bond is illustrated in Fig. 1.1. This polypeptide chain is termed *primary* structure of the protein. Under physiological conditions enzymatic proteins fold into a unique three dimensional close-packed globular conformation, known as *native state*, with few molecules of solvent in their interior. In such native state a protein is biologically active. Typically, an enzyme has a specific interaction with a particular biological molecule, or ligand. Its function is to catalyse a chemical reaction with changes of chemical bonds in the ligand. The structure plays a decisive role in the recognition of the target ligand. X-rays crystallography and NMR spectroscopy have made it possible to experimentally determine the structure of hundreds of proteins, demonstrating that there is a definite spatial organization in the native state. The atoms fluctuate weakly around rather localized positions in space [1]. An example of a protein molecule is shown in a pictorial way in Fig. 1.2. The all-atom picture appears daunting. However, it is instructive to consider a schematic representation of the molecule in which only the backbone of amino acids appears, without the side chains (see Fig. 1.2). A much better defined structure emerges. Small portions of the chain, consisting typically in a dozen of amino acids, are organized in local substructures, called *secondary* structures. These units have usually a helical shape (known as α helices, depicted as ribbons in Fig. 1.2), or are formed by parallel strands (called β sheets, represented as large arrows). Such secondary structure organize themselves into an overall three dimensional structure, known as *tertiary structure*, which often reveals a high degree

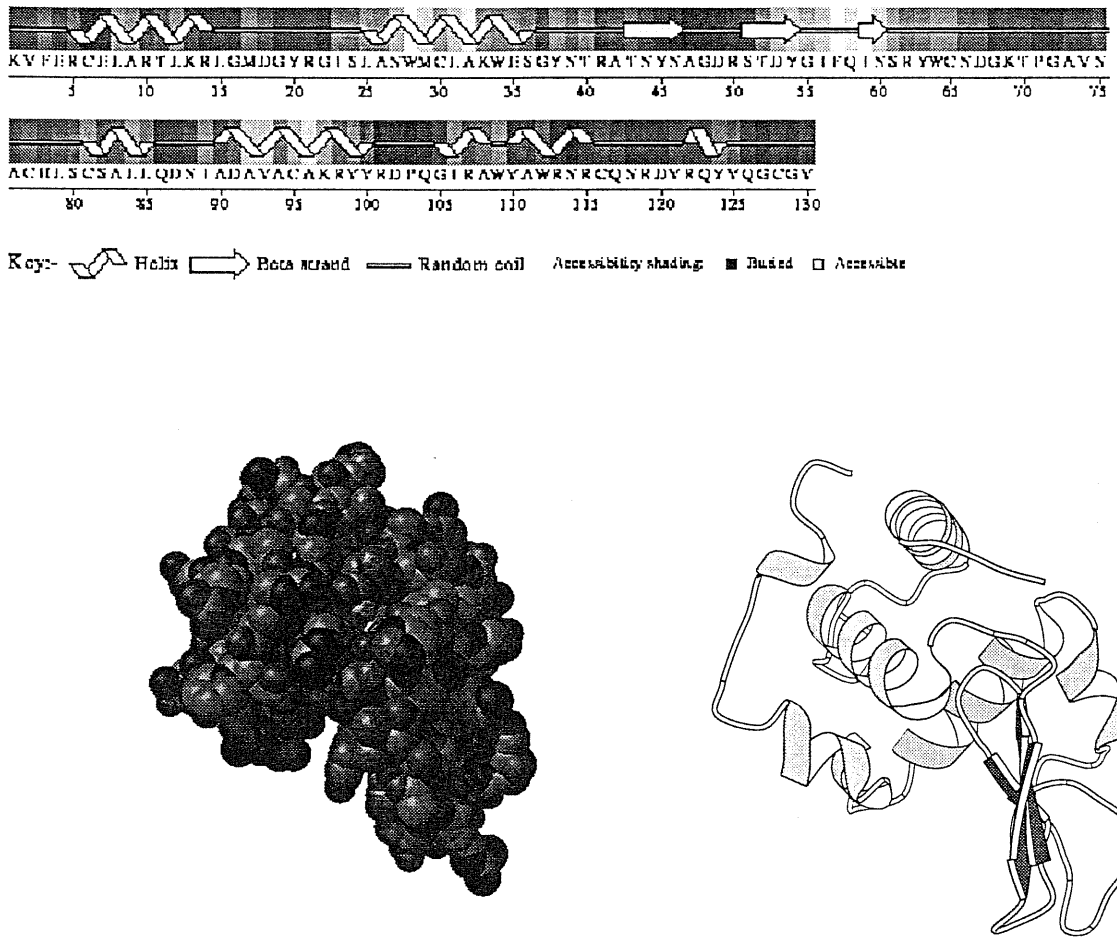

Figure 1.2: (Above) Primary structure of lysozyme, a well characterized enzyme. Secondary structures are also shown as they are located along the chain. Ribbons represent α helices and arrows β sheets. (Bottom left) All atom picture. (Bottom right) Schematic plot, showing the underlying structural organization in secondary structures.

of symmetry, showing bundles of helices or sandwiches of β sheets.

The native state is commonly believed to be the global minimum of the free energy [2]. The overall tertiary structure and the secondary structures are stabilized by non covalent interactions. Among the 20 amino acids, some have a net charge and all can form hydrogen bonds. About half of them are non polar to varying degrees and they can be classified as hydrophobic, if they have an unfavorable interaction with water, or hydrophilic in the opposite case. To avoid contact with water, hydrophobic amino acids tend to be buried inside the core of the folded protein and this force is the main responsible for the stabilization of the tertiary structure. Secondary structures, as α helices and β sheets, are formed by hydrogen bonding.

1.2.2 Thermodynamics

The present theoretical understanding of equilibrium aspects of protein folding is grounded on the concept of heteropolymer freezing [3]. When a random heteropolymer is cooled, it undergoes a freezing transition. In the high temperature phase the number of conformations that dominate equilibrium is exponentially large ($O(e^N)$) in the number N of monomers. Instead, below the freezing point only very few states ($O(1)$) are thermodynamically relevant. When the freezing transition was first discovered [3] it was given a large credit and many believed that a heteropolymer description of proteins would have captured the essential features of the folding process [4, 5]. Complex systems have been successfully modeled by using random interactions [6], so it appeared natural to apply the same approach on proteins. Many ideas are borrowed from the statistical mechanics of disordered systems, in particular from the Random Energy Model (REM) [7].

The REM model represent a family of models with disorder in the limit of negligible correlations between energy levels. It is defined as a system with 2^N energy levels E_i. These levels are assumed to be *independent* random variables extracted from a gaussian distribution

$$P(E) = \frac{1}{\sqrt{N\pi J}} \exp(-\frac{E^2}{NJ^2}) \,. \tag{1.1}$$

The density of states $n(E)$ fluctuates for different realizations, however the average $\langle n(E) \rangle$ over the distribution $P(E)$ is easy to write as

$$\langle n(E) \rangle = 2^N P(E) \propto \exp[N(\log 2 - (\frac{E}{NJ})^2)] \,. \tag{1.2}$$

The model is characterized by a critical energy $E_c = NJ\sqrt{\log 2}$. For $|E| < E_c$ the average number of levels in the interval $(E, E + dE)$ is much larger than 1, whereas for $|E| > E_c$ such number is much smaller than 1, meaning that at a given E for most of the realizations $n(E) = 0$. Below E_c, for a certain realization, levels are discrete, and the difference between them scales as $\sqrt{N}$. In the thermodynamic limit this difference becomes negligible, and a spin glass scenario is realized. Many states are almost degenerate and the system can be kinetically trapped in any of them. Thus, below a critical "glass" temperature $T_c = E_c/(2\log 2)$ the system is frozen in its ground state and the specific heat vanishes in the whole low temperature phase.

Two parallel approaches have been proposed, which despite the apparent diversity, share the same underlying philosophy, both leading to a REM scenario. Bryngelson and Wolynes [8] were the first to propose to replace a complex hamiltonian for protein folding with a stochastic one with the same statistical characteristics. Their starting hamiltonian was

$$\mathcal{H} = -\sum_i \varepsilon_i(\alpha_i) - \sum_i J_{i,i+1}(\alpha_i, \alpha_{i+1}) - \sum_{i,j} K_{i,j}(\alpha_i, \alpha_j, r_i, r_j) \,, \tag{1.3}$$

where α_i is the state of the i-th amino acid, and r_i its position. The first term represent the energy of a single amino acid in the system, the second term is a nearest neighbor interaction giving rise to secondary structures, and the third term is a long range interaction such as the hydrophobic force responsible for the collapse of the chain and for stabilizing the overall tertiary structure. They replaced all these terms with random variables with gaussian distributions. In addition, by adding terms favoring native conformation, they built in the principle of "minimal frustration", which states that interactions giving rise to secondary and to tertiary structures, otherwise conflicting, should be maximally compatible. They estimated the glass transition temperature for such stochastic hamiltonian, establishing one of the first theoretical results about thermodynamics of model proteins.

Shakhnovich and Gutin [9, 3] considered a path integral formulation for the partition function

$$Z = \int_{x(0)=0} \exp[-\mathcal{H}\{x(\tau)\}] Dx(\tau)\,, \tag{1.4}$$

where the hamiltonian is

$$\begin{aligned} \mathcal{H} &= \frac{1}{2}a^{-2}\int_0^N \left(\frac{\partial x(\tau)}{\partial \tau}\right)^2 d\tau + \frac{1}{2}\int_0^N B(\tau,\tau')\delta(x(\tau)-x(\tau'))d\tau d\tau' \\ &\quad + \frac{C}{6}\int_0^N \delta(x(\tau)-x(\tau'))\,\delta(x(\tau)-x(\tau''))d\tau d\tau' d\tau''. \end{aligned} \tag{1.5}$$

Disorder enters in the two terms interaction coefficient $B(\tau,\tau')$ which is assumed to be a gaussian variable with width B. They used the replica trick to average over disorder, finding again a REM-like glass transition for B sufficiently large. In the low temperature phase few states dominate, supporting the idea that the native state of a protein is the global free energy minimum, which in turn corresponds to the global energy minimum due to a negligible contribution of entropy. The validity of these results have been tested numerically [10] and in particular the small structural similarity between low energy states has been confirmed. Moreover they found that quantities characterizing the low energy states are non self-averaging. As a general observation, average over disorder is meaningful when one expects that physical quantities would depend weakly on the realization of disorder [6]. This situation is often realized, and indeed in most cases it is impossible to select a specific realization of disorder. Proteins are different in that nature has provided a replication mechanism able to reproduce, with virtually no mistake, macroscopic amounts of copies of the same realization of disorder.

1.2.3 Dynamics

Protein dynamics is characterized by frustration, arising from antagonistic interactions between amino acids, and by the chain topology constraint, which produces a complex connectivity pattern between low energy states. As a result, the energy landscape of a polypeptide

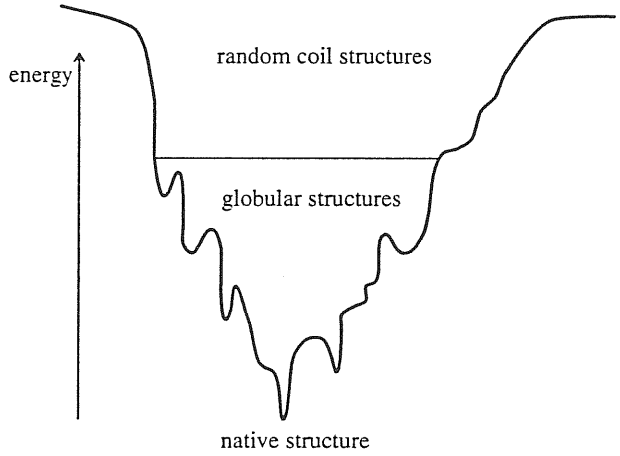

Figure 1.3: Schematic picture of a folding funnel.

sequence is rugged, with energy barriers of any height [11]. Shakhnovich and colleagues [5, 12] have shown by Monte Carlo simulations, that the ground state could possibly be found only for special protein sequences. They observed that a large energy gap above the ground state (a non self-averaging property) is the hallmark of sequences that can be driven to their ground state. It has also been shown that folding a random polypeptide sequence (within the reasonable model discussed in Ref. [13]) is a NP hard problem [13]. Unfolded polypeptide chains inside a cell are broken down in amino acids by specific proteases [14]. Evolution should have then selected sequences thermodynamically stable in their native state and capable to reach it rapidly, by overcoming energy barriers at a temperature at which the native state is stable. The kinetic accessibility of the native state to biological polypeptides is believed to occur through a funnel-shaped free energy landscape [15, 11], shown in Fig. 1.3, which prevents trapping in long living metastable states and biases the folding reaction towards the native state. These remarks were foreseen in the formulation of the minimal frustration principle [8].

Since the work of Anfinsen [2] it is generally believed that the native state is the global free energy minimum under physiological conditions. Experimental evidence of this fact is that upon changing back and forth temperature or the solvent conditions, the protein folds and unfolds with the same rate constants [2]. It may well be supposed that the native state is only in a local minimum of the free energy, thermodynamically metastable. In this hypothesis, the folding process would not be ruled by free energy alone but also by some kinetic mechanism, generally known as *folding pathways* [16], preceding the folding transition, such the formation of local secondary structures [17], or of a folding nucleus [18]. These mechanisms reduce the accessible conformational space, arising the possibility of the presence of a stable local minimum. Protein folding is then thought to occur with a sequential mechanism. Starting from a random coil, the chain self assemble by going through a succession of intermediate states which are more and more close to the folded

conformation. It has been conjectured that folding is a process that statistically select one pathway among a multitude of them [11]. It has been further proposed that a kinetic partitioning takes place [19]. A fraction ϕ of pathways are characterized by a specific collapse to a folding nucleus from where the native state is reached. The remaining fraction $1-\phi$ follows a different kind of route. Along these pathways, there is a non specific collapse to a compact globular shape with a large entropy. Then there is a diffusive search among these compact conformations for native like intermediate conformations. When one of these folding intermediates is reached there is an activated reaction (reminiscent of a possible first order phase transition at the onset of thermodynamic limit) to the folded final structure. Denaturation/renaturation experiment are actually not in contradiction with the conjecture of metastability of the native state. Once native conditions have been restored an "entry point" for a folding pathways can be readily found driving the protein towards its native structure. The marginal stability of a protein, realized through a metastable native state, would enforce the response to environmental changes, which are likely to characterize a biological system.

2 Protein Folding

The number of possible sequences that can be constructed from the 20 species of naturally occurring amino acids is gigantic. Assuming a typical protein length of 250 amino acids, or residues, there could be 20^{250} potential sequences [1]. Existing proteins have been selected by evolution through stochastic mutations to perform specific biological functions. Moreover, if every amino acid is assumed to have 3 possible conformational states, a sequence of 250 residues can be found in 3^{250} spatial structures. Remarkably, only 10^3 of these structures are estimated to be the native state of some sequence [20, 21].

In this chapter we address the question of how these structures and sequences have been selected by evolution. A deeper understanding of this issue will shed light on the physical mechanisms underlying the folding process in proteins

A protein can be biologically functional only if it is able to assume its native state rapidly and reliably. The key question is then to uncover the *intrinsic* features of sequences which encode the folding process. The most commonly accepted idea is that sequences with a ground state with a large gap above it are natural candidates to have good foldability properties [5]. Evidence of this fact comes mostly from Monte Carlo and from exact enumeration studies of simple models. More recently a new idea has been proposed focusing on properties of the structures [22]. Special structures emerge from the ensemble of all possible structures and they are characterized by their designability. The designability of a structure is measured by the number of sequences that possess such structure as their ground state. This idea is supported by a technically impressive exact enumeration study. However, this study relies on the assumption of compactness of the ground state [22]. We show that generally this assumption is not justified and that a more detailed study lead to different conclusion. In chapter four we will further develop evolutionary implications of these findings.

2.1 Foldability

A possible explanation for the existence of champion sequences with folding properties well above the average of random sequences has been termed *foldability*, or propensity to fold [5, 23]. Only those sequences that have a unique ground state that is both thermodynamically stable and kinetically accessible are biologically relevant. This conjecture relies mainly on numerical studies of minimal models which are thought to capture some of the more relevant protein-like features [24, 5, 12, 25, 26]. A protein is usually represented as a self-avoiding walk (SAW) on a lattice. Monomers are placed on lattice points and are a coarse-grained representation of amino acids. In the simplest model, the HP model [27] there are only two species of amino acids, mimicking hydrophobic (H) or polar tendency (P).

Hydrophobic interaction is believed to play a central role in protein folding as well as in other self-assembly processes, in micelle formation and in biological membrane structure stabilization [28]. It has an entropic origin. Water molecules in liquid form have a strong tendency to form hydrogen bonds. The hydrogen bond is an electrostatic interaction with a strong orientational character between an H atom, which indeed remains close to its parent O atom, and another O atom of a neighboring water molecule. Each water molecule participate on average to 3-3.5 hydrogen bonds. When a non polar molecule is present in water solution, it cannot form hydrogen bonds, so water molecules rearrange around it to optimize their number of hydrogen bonds. The loss of entropy so induced is responsible to the high insolubility of non polar substances, such as hydrocarbons, in water. For example the free energy of transfer of methane from bulk liquid to water is 14.5 kJ mol^{-1} at 25 °C. The same rearrangement effect induces an effective "hydrophobic" interaction in water solution between two non polar molecules. Such interaction is believed to have a range between 0 and 10 nm, with an exponential decay length of 1 nm. The tendency of hydrophobic amino acids to avoid water drives the collapse of the protein chain with the formation of a hydrophobic core [29]. It has also been shown experimentally that certain proteins can be designed by binary patterning of polar and non polar amino acids [30]. The strategy consists in specifying explicitly the sequence locations of hydrophobic and hydrophilic amino acids with no constraint on the precise identity of the side chains. Most of the designed sequences fold into the desired four helix bundle conformation, clearly showing that an opportune arrangement of polar and non polar residues can drive polypeptide chains to collapse into globular folds.

The HP hamiltonian for a sequence S in a conformation Γ is

$$\mathcal{H}_S(\Gamma) = \sum_{ij} B(s_i, s_j)\Delta_\Gamma(\mathbf{r}_i - \mathbf{r}_j). \tag{2.1}$$

The i-th amino acid s_i is located on the lattice site at position $\mathbf{r}_i$. The contact matrix $\Delta_\Gamma(\mathbf{r}_i - \mathbf{r}_j)$ is 1 if $\mathbf{r}_i$ and $\mathbf{r}_j$ are nearest neighbor sites that are not occupied by consecutive

amino acids along the chain, and zero otherwise. The amino acid s_i can be either H or P and $B(H,H) = -\epsilon$ (attractive interaction) whereas $B(H,P) = B(P,P) = 0$.

It has been shown [9, 3, 10] that the thermodynamics of random heteropolymers can be adequately described by the random energy model (REM) [7] (see Chapter 1). The most commonly used random hamiltonian giving rise to a REM type spectrum is the so called random interaction model ($B_{i,j}$ model) [31, 5, 12]

$$\mathcal{H}_S(\Gamma) = \sum_{ij} B_{i,j}\Delta_\Gamma(\mathbf{r}_i - \mathbf{r}_j). \tag{2.2}$$

where $i,j = 1,...,N$ are the amino acid labels. The N monomers are assumed to be distinct. The $B_{i,j}$ matrix is symmetric and has $N(N+1)/2$ elements. In order to obtain a random heteropolymer, these elements are drawn from a Gaussian distribution with mean value $B_\circ$, which is an overall attractive term favoring collapsed states, and variance σ_B which controls the degree of heterogeneity. Effectively, the matrix B represents a certain sequence.

A more realistic model is defined by Eq. (2.1) where s_i labels one of the 20 different species of amino acids. In this case B is a 20×20 matrix which is usually taken from suitable parameterizations of the contact energies given by Miyazawa and Jernigan (MJ) [32, 33] or Kolinski, Godzik and Skolnik [34]. In a recent work [35], the MJ matrix has been studied and it has been found that its matrix elements $B(s_i,s_j)$ can be simply expressed as $B(s_i,s_j) = q_i + q_j + \beta q_i q_j$, where β is a constant, and q a real variable associated with each of the 20 amino acids. The term $q_i + q_j$ constitutes the main contribution to the contact energy $B(s_i,s_j)$, and the q values correlate well with the hydrophobicities of the amino acids. This results explicitly support the view that the hydrophobic interaction is the dominant driving force for protein folding.

Random realizations of the interaction matrix in the $B_{i,j}$ or in the MJ case correspond to polypeptide sequences randomly synthesized. Biological properties of random polypeptides differ dramatically from those of enzymatic proteins. Under alteration of the environment (e.g. changes in the temperature, pH or pressure) the former change gradually their physical properties whereas the latter do not up to a critical strength of the perturbation. Below this threshold they are biologically active and above they suddenly loose this ability. This phenomenon is called denaturation [1]. Enzymatic proteins are special in that they are in approximate correspondence to *atypical* realizations of disorder in the two simple models presented. The native state of such special sequences becomes populated at a "folding" temperature T_f higher than the freezing (or glass) transition temperature T_g. Random heteropolymer have $T_g > T_f$ as a common feature [36, 37]. For generic sequences the freezing transition takes place first, preventing the possibility to reach the ground state. The system remains trapped in long living metastable states and the native state becomes kinetically inaccessible [5, 15]. Dynamical properties of folding sequences have been related to their thermodynamical properties, in particular to their energy spectrum. The relation

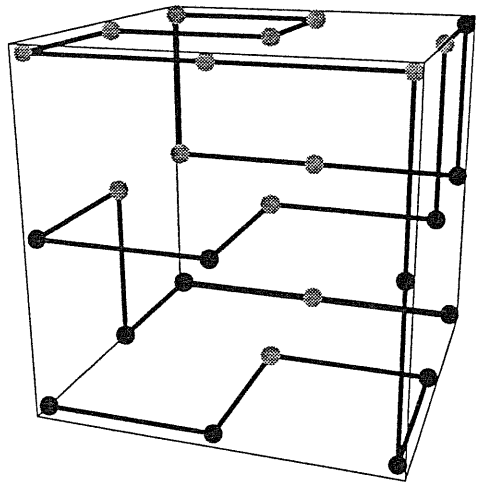

Figure 2.1: A typical compact conformation of a HP sequence of 27 monomers on a $3 \times 3 \times 3$ cube.

is justified by the observation that sequences should have been selected by simultaneously satisfying the requirements to be quickly foldable and to be stable in their functional state. Thermodynamical properties are then expect to dictate to some extent the overall kinetic behavior. The folding ability is lacking unless the spectrum does not present special features favoring the stability of the native state (see e.g. Fig. 4.1). The exact nature of these features has been a matter of debate in the last few years [24, 5, 38, 25]. Shakhnovich *et al.* [5, 12] have studied a 27 monomers $B_{i,j}$ chain in 3D. The ground state is known by exact enumeration of compact conformations on a $3 \times 3 \times 3$ cube. A typical compact conformation, or hamiltonian walk, is shown in Fig. 2.1 for the HP model. They have related the energy gap, defined as the minimum energy required to change the ground state structure to a different compact structure, to the Monte Carlo folding time which is taken as the mean first passage time (MFPT) to the ground state. A large energy gap is a necessary and sufficient condition for foldicity. Foldicity is defined from the dynamical behavior of a sequence as the fraction of Monte Carlo runs that starting with a random conformation finish in the native state under different initial conditions. It is a necessary condition since it guarantees that the native state is stable, and it is a sufficient condition because a stochastic search among globular structures will rapidly find its way to the ground state without getting trapped in deep metastable minima. More recently it has been pointed out that more general characteristics of the energy spectrum should be invoked to explain foldicity. Klimov and Thirumalai [25, 26] observed that the foldicity of a protein increases exponentially with the sequence intrinsic quantity

$$\sigma = \frac{T_\theta - T_f}{T_\theta} , \tag{2.3}$$

where T_θ is the collapse transition temperature from random coil to random globular con-

formations [39]. A small σ implies $T_f \sim T_\theta$. As a result, all possible globular structures are explored well above the ensuing glass transition. The sequences is then enrouted to its native state when it finds a "transition" state which has a significant structural similarity to the native state [5]. Wolynes *et al.* [15, 11, 40] have discussed a scenario in which the statistical properties of the entire energy landscape determine those features of the folding processes common to all sequences and to distinguish them to specific processes peculiar to individual proteins, The energy landscape has an overall funnel shape biasing the sequence to its ground state (see Fig. 1.3. Three thermodynamical parameters describe the properties of the funnel. The first is the ruggedness of the landscape, which is a measure of the heights of the energy barrier between conformations. Rugged landscape are usually found when there is competition between microscopic interactions, a phenomenon known as frustration [6]. The second is the gradient towards the folded state which is measured by the difference in energy between the native state and the average energy of globular states. The third parameter is the search problem size, which is given by the configurational entropy.

2.2 Designability

In recent work, Li *et al.* [22, 23] propose an alternative fascinating mechanism to the sequence selection hypothesis presented in the preceding section. In order to explain the high degree of regularity found in natural protein structures which are organized in secondary structures with tertiary symmetries (see Fig. [?] and Fig. [?]), they introduce the idea that natural selection acted on structure rather than on sequences. Structures that are found in nature are those that are characterized by a high *designability*, measured as the number N_S of sequences that have their ground state on them. Conformations differ markedly in terms of their designability. Those that are highly designable emerge from the vast ensemble of all possible conformations and exhibit protein-like secondary and tertiary structures. Sequences that fold on them are characterized by high thermodynamic stability and moreover they are stable against mutations. Are these facts mere coincidences or does indeed nature select highly designable structures to accommodate proteins?

Within the framework of the HP model on a 3D lattice they performed exact enumeration on 27 monomers chains. By enumerating all the possible compact conformations on a $3 \times 3 \times 3$ cube (see one example in Fig. 2.1) for each of the 2^{27} possible sequences, they showed that 4.75% of the sequences have a unique ground state. They have chosen the HP parameters in order to ensure that 1) compact shapes have lower energy than any non compact shapes; 2) H monomers are buried as much as possible, which is realized by choosing $B(H,H) < B(H,P) < B(P,P)$; 3) different types of monomers tend to segregate, which is expressed by $2B(H,P) > B(H,H) + B(P,P)$. The choice B(H,H)=-2.3, B(H,P)=-1, and B(P,P)=0 enforces conditions 2) and 3). However the requirement of condition 1) is

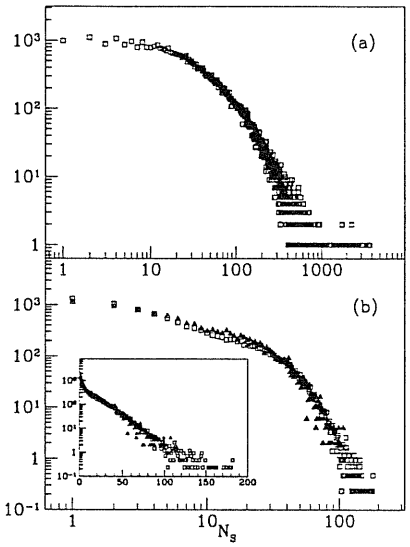

Figure 2.2: (a): Histogram of number of structures with a given number N_S of associated sequences for 3D 3 × 3 × 3 case, in a log-log plot. (b): Histogram of number of structures with a given N_S for 2D 6 × 5 (filled triangle) and 6 × 6 (open square) case, in a log-log plot. Insert: same data in a semi-log plot.

questionable, as we will discuss in detail below.

Fig. 2.2 shows the distribution of the number N_Γ of conformations having a given N_S. The top structure can be designed by $N_S = 3794$ sequences, and there are 4256 structures that are not designable by any sequence. The distribution has a long tail which clearly indicates that the conformations are not equivalent. If all the structures were statistically equivalent sequences with a unique ground state would distribute uniformly on them, and on average $\overline{N_S} \simeq 123$. In this case, the probability of finding a conformation Γ with $N_S > 200$ can be obtained from the Poisson distribution

$$P(N_S > 200) = \sum_{k>200} \frac{\overline{N_S}^k e^{-\overline{N_S}}}{k!} \tag{2.4}$$

and would results of the order of 10^{-10}. It is possible to speculate about the resemblance of the highly designable structures to real proteins. Li *et al.* find that these structures have symmetries and subunits that are absent in random compact structures. On a coarse grained level it is possible to relate these subunits to tertiary symmetries and secondary structures found in real proteins. The appeal of this remark is that the request of designability implies also a possible answer to the question of "why proteins look like proteins".

A study of a possible size dependence of these results is performed on a 2D square lattice for systems of size 4×4, 5×5, 6×5 and 6×6. A random sampling in sequence space is carried out and for each of the extracted sequences full enumeration of compact conformations is

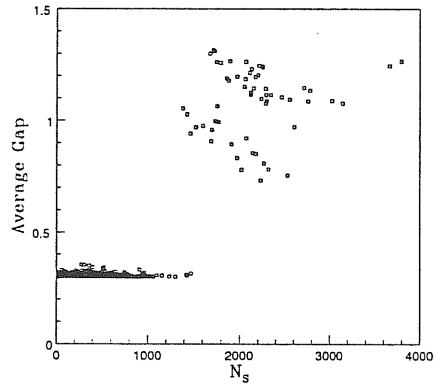

Figure 2.3: Average gap of 3D $3 \times 3 \times 3$ structures plotted against N_S of the structures.

performed. The behavior of $N_\Gamma(N_S)$ is qualitatively the same as in 3D, however in this case the tail is markedly exponential. The parameters for the HP models are kept unchanged, and it is known that the surface-volume ratio of 2D models approaches that of real proteins. Also in this case the evidence is that highly designable structures have bundles of pleats and long strands that are longer than expected if the conformations were statistically equivalent, and that are reminiscent of α helices and β sheets.

The analysis of the thermodynamic stability of the highly designable structures reveals that the structures in the tail of the distribution are indeed special. Let's introduce the average energy gap $\overline{\delta_\Gamma}$ for the sequences that fold on structure Γ. A large gap is assumed as a measure of the stability of the ground state of the sequence. Highly designable structures are characterized by a large average energy gap which in turn becomes a property of the structure. The behavior in $\overline{\delta_\Gamma}$ with respect to N_S shows a dramatic threshold around $N_S \sim 1400$ (see Fig. 2.3). Below this threshold the average gap is small, whereas above it is rather large. According to the criterion of high designability, only 0.12% of the compact conformations are candidates for being selected by evolution.

Sequences that have evolved from a common ancestor are said to be homologous [1]. They often originated from mutations conserving the initial ground state structure. Given a highly designable structure the sequences that have it as their ground state are remarkably different in their amino acid composition, yet their energy gap is large on the average. From this viewpoint, a large gap has then an evolutionary implication. If a sequence has a ground state with a large gap it is fairly probable that a mutation will not produce such a large shift in the energy to surmount the gap to the competing structures. A large gap implies stability also against mutations. The picture is rather self-consistent: highly designable

structures have a larger probability to have been chosen through random biosysthesis of sequences in the primordial age and moreover they are also more stable against mutations.

2.3 Which Came First, Protein Sequence or Structure?

Li *et al.* propose that structure selection is a relevant factor in the evolution of proteins since structures that are highly designable can host the ground state of many sequences and are characterized by a large average gap which implies thermodynamic stability and robustness against mutations. To what extent the assumption of compactness of the ground state affects these results? In this section we show that the jump in the average gap shown in Fig. 2.3 is indeed an artifact of having retained only compact conformations. Foldability, or sequence selection hypothesis, still deserves its credit.

Within the same HP model, we first consider a N=16 chain on a 2D square lattice which is amenable to exact enumeration of *all* possible conformations, either compact or not. After discarding walks that are related by rotation and reflection symmetries, there are 802075 possible SAWs and, among them, 69 are compact.

In order to enumerate the walks we used the backtracking algorithm [41]. It is a well know algorithm used to generate all the possible walks of a given length on a lattice. At the beginning the first walk is drawn on the lattice by placing steps on lattice edges as shown in Fig. 2.1 for the 3D case. Systematic attempts are made to place the last step on the lattice. If all possible new routes are found blocked, the algorithm retreats one step, the next to last, and move it to a new edge, if possible, and then advances forward again to the next step. Full enumeration is completed when the algorithm retreats back to the first step. Replacing steps with monomer species, we used the same idea to generate all possible HP sequences.

Following Li *et al.* we study the behavior of the number N_S of sequences S that have their unique ground state on a given structure Γ. By fully enumerating all conformations and all sequences, we found that, among 2^{16} possibilities, 9494 sequences have a unique ground state and that there are 1275 conformations that can be designed by at least one of these special sequences. We present in Fig. 2.4 the number N_Γ of structures having a given N_S. There is a comfortable agreement with their 6×6 2D case. Fig. 2.2 of Li *et al.*.

However, a major difference becomes manifest when we consider the average gap $\overline{\delta_\Gamma}$. The average gap is significantly constant within small fluctuations in a range [0.8,1.2], and no clear threshold is visible. Indeed, if we consider the same calculations restricted to only compact conformations $\overline{\delta_\Gamma}$ has a remarkably similar behavior respect to the result of Li *et al.*, as shown in Fig. 2.5. A different regime in the fluctuations of the average gap appears around $N_S \sim 400$, signalling a differentiation between highly designable compact

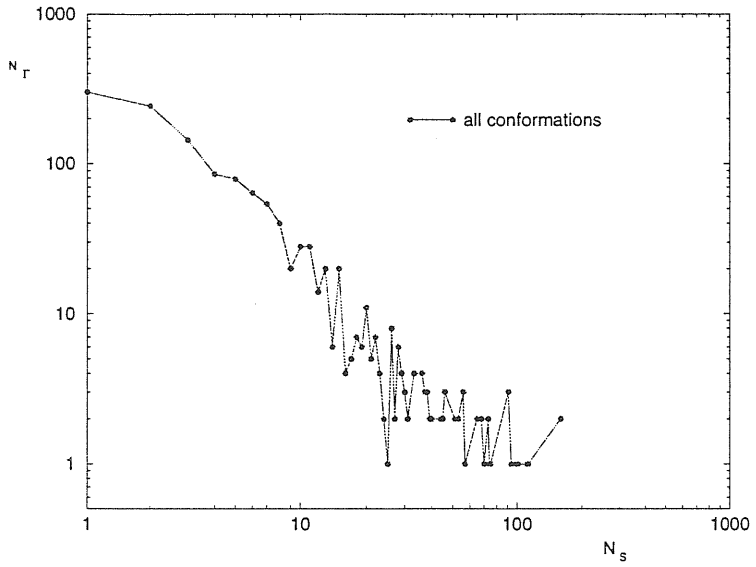

Figure 2.4: Histogram of number of structures with a given N_S for 2D 4×4 in a log-log plot in the full enumeration case.

conformation from the ordinary compact ones.

A possible explanation for these discrepancies comes from the observation that the gap to the first excited state could not be a good indicator of the thermodynamic stability [25, 26, 38, 40]. Moreover it has been observed that there is a difference in considering a gap defined only between compact conformations and between all conformations [42]. In the latter case there is the possibility to find conformations close in energy since they differ only by a few contacts. These conformations should not be considered as distinct within the assumption that lattice models are coarse grained representation of real 3D protein structures. Since all the collapsed conformations have a non negligible probability to be assumed by the sequence, a better measure of the thermodynamic stability is given by the Z score [43] which is the energy gap to the average energy $\langle \mathcal{H} \rangle$ of the collapsed conformations, scaled with its dispersion $\sigma = \sqrt{\langle (\mathcal{H} - \langle \mathcal{H} \rangle)^2 \rangle}$.

$$Z = -\frac{\mathcal{H} - \langle \mathcal{H} \rangle}{\sigma}. \tag{2.5}$$

$\langle \mathcal{H} \rangle$ and σ were calculated as averages over all conformations with seven or more contacts, which are those competing to be the native state. Compact conformations have nine contacts. A mean field approximation [44] for $\langle \mathcal{H} \rangle$ is obtained from the estimation of the mean single contact energy

$$\mu = \frac{\sum_{j>i}^{N} P_{ij} \epsilon_{ij}}{\sum_{j>i}^{N} P_{ij}} \tag{2.6}$$

The sum is over the N amino acids. P_{ij} is the probability that amino acids i and j are in contact, and $\epsilon_{ij} = B(s_i, s_j)\Delta_\Gamma(\mathbf{r}_i - \mathbf{r}_j)$ is their energy (see Eq. 2.1). In the mean field

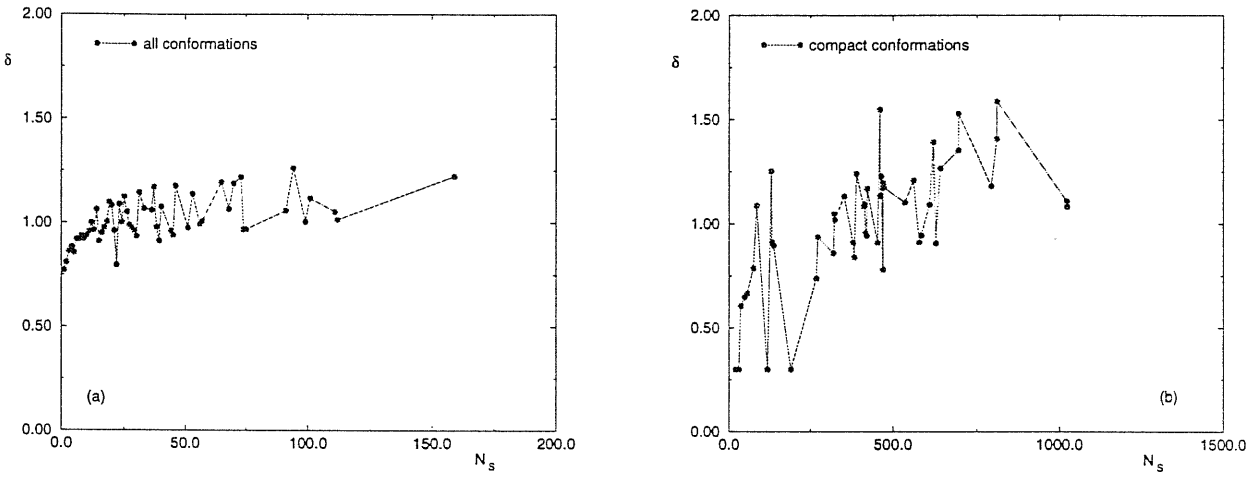

Figure 2.5: Average gap δ of 2D 4×4 structures plotted against N_S of the structures. (a) Enumeration of all the conformations. (b) Enumeration of only compact conformations.

approximation $P_{ij} = 1$ if i and j are allowed to be in contact by the chain connectivity and $P_{ij} = 0$ otherwise. The average energy is then given by $\langle \mathcal{H} \rangle = N_c \mu$ where N_c is the number of contacts in the compact case. The second moment μ_2

$$\mu_2 = \frac{\sum_{j>i}^{N} P_{ij} \epsilon_{ij}^2}{\sum_{j>i}^{N} P_{ij}} \tag{2.7}$$

of μ is also calculated to obtain the deviation $\sigma = \sqrt{\mu_2 - \mu^2}$. We have verified that results are not significantly affected using the former or the latter approximation for Z. The three curves shown in Fig. 2.6 correspond to the highest, the mean and the lowest Z score. We explore all the conformations and only the compact ones separately. In both cases there is no evidence of a jump in the thermodynamic stability beyond a certain value of N_s, in contrast to the suggestion by Li *et al.*. Indeed, for a given structure, there are variations in the stability on tuning the sequences showing that, at least, in the two dimensional model, the selection process primarily involves the sequences and not the structure.

A more careful comparative inspection of Fig. 2.2 and of Fig. 2.4 raises the suspicion that the 2D and the 3D cases could differ to some extent. In 3D, highly designable structures are found in the extreme tail of the distribution, for $N_S > 1400$, as shown in Fig. 2.2. Such a tail is much less pronounced, if at all present, both in the distribution for the 6×6 2D case, shown in Fig. 2.2 as well as in the 4×4 2D case of our study, shown in Fig. 2.4. Can be this difference be invoked to justify the different behavior of the average gap?

Li *et at.* conclude their paper by addressing the important questions of what is the kinetic accessibility of highly designable structures, and if there are any other selection

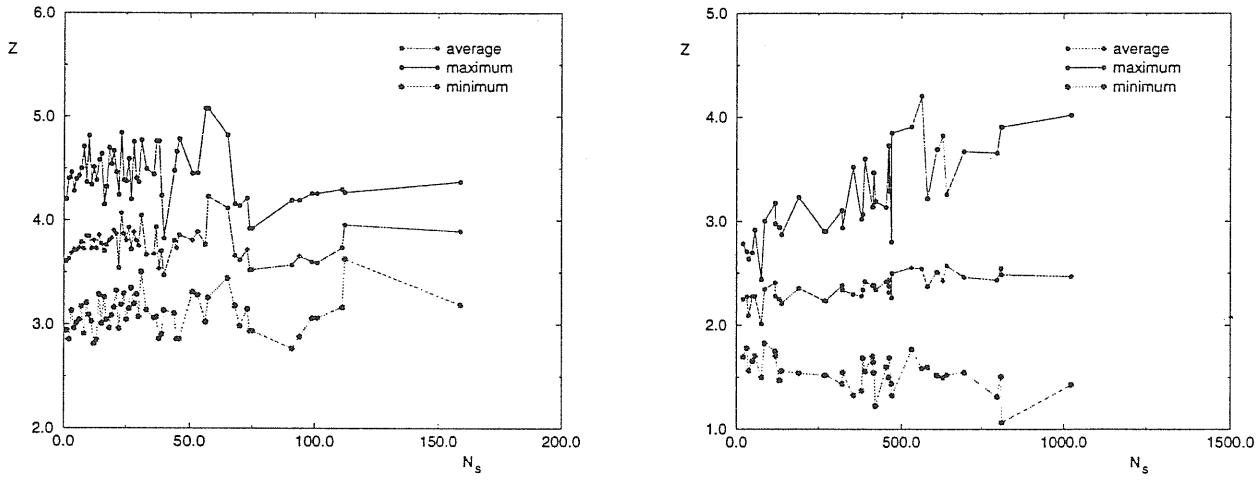

Figure 2.6: The Z score plotted against N_S for the 2D version of the 3D model considered by Li *et al.* [22]. (a) The Z score is the difference between the average energy $\langle \mathcal{H} \rangle$ of the compact and semi-compact conformations and the ground state energy, $\mathcal{H}$, scaled by the dispersion σ. (b) Z score obtained by considering only compact conformations.

principles imposed by the kinetics. We have undertaken this study by first searching the putative ground state of a set of sequences by enumeration of the 103346 distinct compact conformations on a $3 \times 3 \times 3$ cube. Sequences with a unique lowest energy conformation, or *putative* ground state, were then subjected to 3D Monte Carlo simulations. We have found, in 81% of the 242 cases we have looked at, that the native state is not compact. An example is given in Fig. 2.7. A possible explanation for such a large rate of failures can be formulated by considering the 2D $N = 16$ case, which is amenable to exact enumeration of all conformations, compact or not. We introduce the parameter $\nu = N_c^{\max} - N_c$, which gives the difference between the maximum number $N_c^{\max}$ of contacts that a conformation can have ($N_c^{\max} = 9$ in the case considered) and the number N_c of contacts of the ground state conformation of a given sequence. In Fig. 2.8 we show the behavior of the average $overline\nu$ of ν over conformations with a given N_S against N_S. Highly designable conformations have a large N_S and are characterized by $\overline{\nu} = 0$. This is the fingerprint of their compact nature. Below $N_S \sim 60$, on the average, designable conformations are markedly non compact, and ν is greater than zero. Since from Fig. 2.4 we know that for the large majority of sequences the ground state conformation is not highly designable, by picking up randomly a sequence, as we have done in the 3D case, it is most probable that the ground state would not be found among compact conformations only.

Technically, to generate only compact conformations we used a modified version of the backtracking algorithm. In general, to constrain the walk inside a given domain of a D

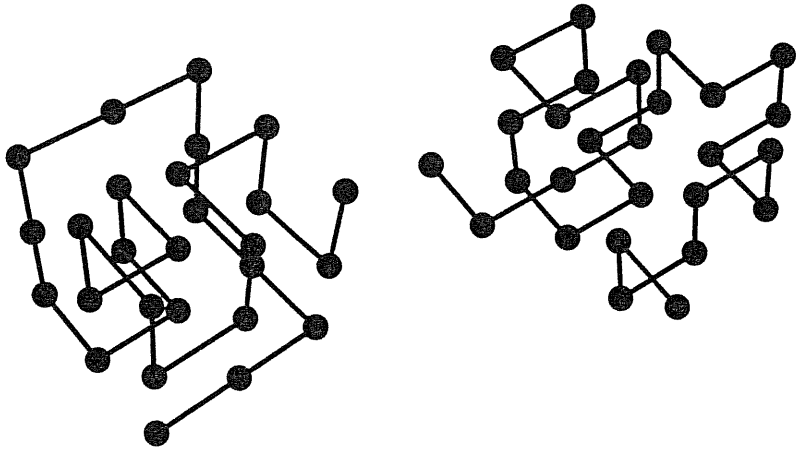

Figure 2.7: (Left) Unique lowest energy conformation among the compact structures for the sequence PPHHPHPPHHPHHHPHHPPHHPHHPHP. The energy is $\mathcal{H} = -40.5$ with a gap to the first excited state of 0.3. (Right) A Monte Carlo-generated non compact conformation with an energy $\mathcal{H} = -41.5$.

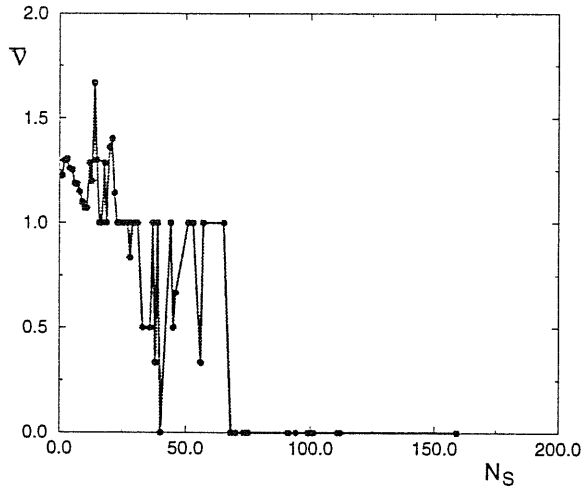

Figure 2.8: Relative number of contacts ν.

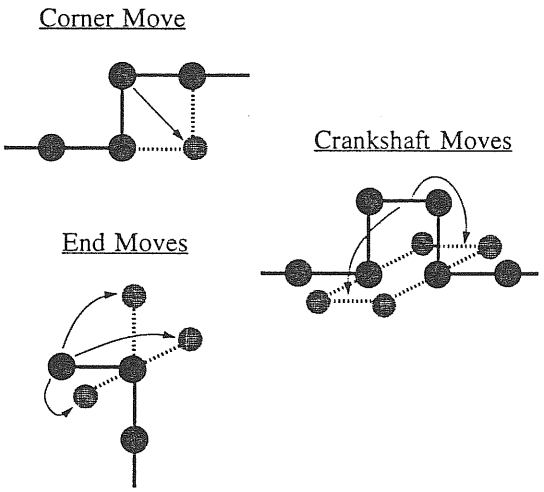

Figure 2.9: The three types of possible Monte Carlo moves used in simulations of protein dynamics.

dimensional lattice one introduces a matrix of contacts C. $C(i,j) = 1$ if amino acids i and j form a non bonded contact, that is they are non consecutive nearest neighbors, otherwise $C(i,j)$=0. In the backtracking algorithm when the walker is on site i, the matrix $C(i,i+1)$ is used to give the possible successive step. In this way one can generate all the hamiltonian walks. Care must be taken to discard symmetry related conformations, the usual trick being to place the first few monomers by hand. Given the definition of the energy Eq. (2.2), it is important to know the number K of non bonded contacts that can be present in a compact conformation. It's easy to see that K for an hamiltonian walk on a 3D parallelepiped of sides L, M and N is given by

$$K = M[(N-1)L + N(L-1)] + (M-1)NL - (NLM-1). \tag{2.8}$$

In the case discussed here we have a $3 \times 3 \times 3$ cube ($N = L = M = 3$) and one has $K = 28$ [45].

Monte Carlo simulations for polymer chains will be thoroughly discussed in Cap. 5. Here we give only a brief description of the algorithm we used, which has been introduced by Verdier and Stockmayer [46], and is most commonly used in lattice simulations of folding of simple models [12, 47]. Monte Carlo elementary moves involve local rearrangements of the SAW. Typical moves can be a end move, a corner flip or a crankshaft, as shown in Fig. 2.9. The simulation start from a random initial configuration. Successive elementary moves are attempted and if they are allowed by the SAW condition they are accepted according to the Metropolis criterion [48]. The change ΔE in energy for the proposed move is evaluated. If the the energy is lowered the move is accepted, otherwise $w = \exp[-\Delta E/T]$ is confronted with a random number $x \in [0,1]$ and the move is accepted $w \geq x$. T is a parameter in the

simulation used to tune the acceptance ratio of the attempted move to around 50% and that plays the role of a temperature.

3 Protein Design

Globular proteins, such as enzymes and antibodies, are characterized by the ability to recognize target molecules – their biological activity is mainly controlled by their spatial conformation. A successful method for protein structure design would have widespread implications in medical sciences, giving way to the design of new therapeutic drugs, such as inhibitors that can suppress the activity of harmful proteins, or small peptides performing the action of large proteins but easier to product and to handle. The advent of powerful recombinant DNA techniques allows for the routine determination and modification of the amino acid sequences. An optimal tailoring of the structure, by altering the amino acid sequence, will enable the creation of proteins with desired functionality. Thus a fundamental and vital issue in this field is the "inverse" folding problem: how does one design a sequence of amino acids that has a desired structure Γ as its ground state?

In this chapter, we formulate the solution of the problem of protein design on general principles. The method is based on an analysis in sequence space of the Boltzmann weight for a given target structure Γ^*. An efficient Monte Carlo code that explores sequence space (and for each sequence, the space of compact and non-compact conformations) is implemented. We tested the method on simple models in two and three dimensions. In two dimensions these models are amenable to complete enumeration and thence to an exact check of the numerical technique. Although very reliable, the proposed Monte Carlo strategy is numerically intensive. In order to improve the efficiency of our design procedure we introduce and discuss two approximation schemes. In three dimension a dynamic Monte Carlo is used to fold designed sequences to their ground state conformation, which would correspond to the target structure in case of success. The relation with earlier methods [45, 49, 50] is discussed and we show that a dramatic increase in the rate of successfully designed sequences is obtained.

We stress that the method described here is general and is not restricted to the toy lattice models that we have studied here. It can be readily implemented for more realistic three dimensional off–lattice models of real proteins, provided accurate potentials are known.

3.1 Advances in Protein Design

In this section we discuss previously proposed methods to solve the inverse folding problem within the framework of the simple models previously introduced. Four requirements are generally considered to mark success in designing proteins. First, the engineered sequence should have the target conformation in its ground state. Second, ground state degeneracy is unwanted. Other conformations degenerate or competing in energy with the target fold should be "designed out" [51]. Third, the ground state should be thermodynamically stable. Fourth, the target structure should be dynamically accessible to the designed sequence [45].

The most widely studied strategy to carry out sequence design is due to Shakhnovich and Gutin (SG) [45]. Their method consists in minimizing the energy $\mathcal{H}_S(\Gamma^*)$ in sequence space S for a given target conformation Γ^*. Even assuming that the minimization of $\mathcal{H}_S(\Gamma^*)$ is carried out correctly to determine the true minimum S^* of $\mathcal{H}_S(\Gamma^*)$ for fixed Γ^*

$$S^* = \min_S \mathcal{H}_S(\Gamma^*) \tag{3.1}$$

an intrinsic difficulty with the SG method is that design scheme does not guaranteed that the designed sequence has its ground state on the target conformation, that is

$$\Gamma_{S^*} = \min_\Gamma \mathcal{H}_{S^*}(\Gamma) \tag{3.2}$$

with $\Gamma_{S^*} \neq \Gamma^*$, as depicted in Fig. 3.1. As an example we show in Fig. 3.2 that the SG scheme is feeble in designing out alternative folds, at least within the 2D HP model for chains of length N=16. In the left part of the figure we show a target conformation with a sequence $S_1 \equiv$ PHPPPPHPPHHHPPHP having its unique ground state there. Black monomers are of the H type and there are 5 non consecutive nearest neighbors contacts then $\mathcal{H}_{S_1}(\Gamma^*) = -5$. As far as the design problem is addressed, S_1 solves it. The sequence $S_2 \equiv$ PHHPPHHPHHHHPPHH has a lower energy $\mathcal{H}_{S_2}(\Gamma^*) = -7$. However S_2 has a ground state degenerate with the conformation in the right part of figure. To avoid proliferation of H monomers in the designed sequence, more recently Abkevich *et al.* [52] have introduced the Z score [43] (see Eq. (2.5)) as the quantity to be maximized. Since the Z score is a measure of the distance between the ground state and the average energy of the collapsed states it automatically takes in account negative design.

The rank-ordered assignment of contact energies recently discussed by Shrivastava *et. al.* [53] is similar in spirit to the SG scheme and is guaranteed to work only in simple cases in which Γ^* is compact and there are no constraints on the types of monomers making up the heteropolymer chain.

Pande *et al.* [50] introduced a thermodynamic procedure to synthesize heteropolymers that is related to SG method. They envisage a concentrated solution of the constituents monomers. Each monomer species i has a probability p_i to appear in the solution related

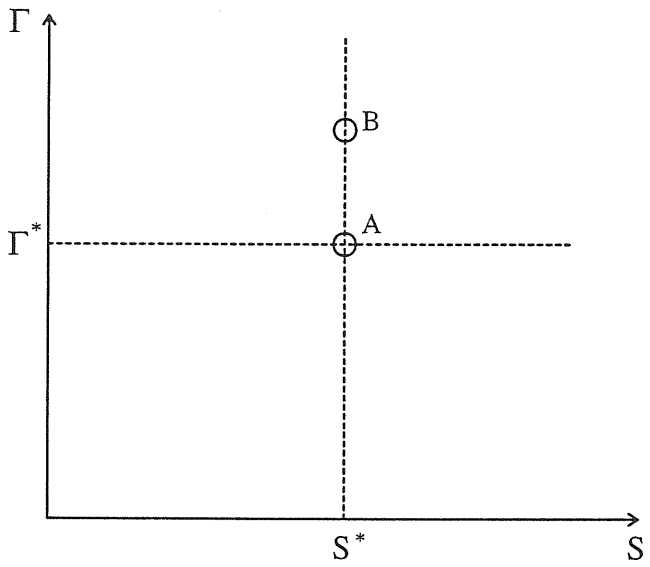

Figure 3.1: Schematic view of the protein folding as an optimization problem in the sequence – conformation ($S - \Gamma$) plane. To solve the direct problem one minimizes the energy $\mathcal{H}_{S^*}(\Gamma)$ with respect to Γ at fixed S^*, moving in the vertical direction. The inverse problem can be solved moving in the horizontal direction fixing a target conformation Γ^*. In the optimization scheme proposed by Shakhnovich and Gutin [45], the sequence S^* with minimal energy on Γ^* is sought (point A). However the search in Γ space for the ground state of S^* can give a conformation different from Γ^* (point B).

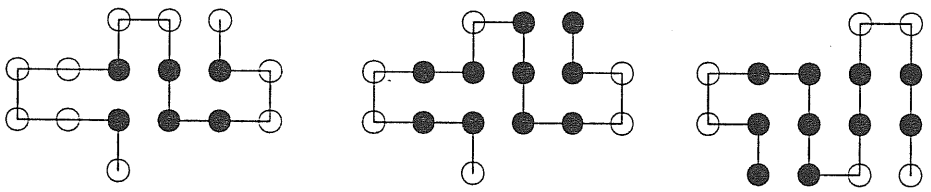

Figure 3.2: A typical case of failure of the SG scheme [45] in designing non degenerate sequences. (Left) Exact solution for the design of conformation Γ^*. The sequence S_1 has its ground state on Γ^*. (Center) The sequence S_2, designed by the SG procedure has its ground state on Γ^*, but (Right) there is a degeneracy with the another conformation.

to its natural abundance. The probability of a sequence S of N amino acids is then

$$P_S^{(0)} = \prod_{i=1}^{N} p_i. \tag{3.3}$$

The energy of such sequence on the target conformation Γ^* is $\mathcal{H}_S(\Gamma^*)$. The probability P_S in the "sequence space soup" is

$$P_S = \frac{P_S^{(0)} \exp[-\mathcal{H}_S(\Gamma^*)]/T_{des}]}{\sum_{S'} P_{S'}^{(0)} \exp[-\mathcal{H}_{S'}(\Gamma^*)/T_{des}]}, \tag{3.4}$$

where T_{des} is the design temperature. For lower T_{des}, the SG method is recovered, and sequences with the lowest possible energy in their ground state are selected. For higher T_{des}, selection is inactive. At intermediate T_{des} it is possible to select sequences, such as S_1, that solve the design problem without being the minimum of the energy in S space. Ramanathan and Shakhnovich [54], showed that the SG method is in close analogy with the method of Pande *et al.* if implemented with a Monte Carlo search in sequence space with a simulation temperature T_{sel}.

What we learn from the weaknesses of the SG and related methods? Having in mind the solution of the direct folding problem one would be lead to the naive approach to choose the energy $\mathcal{H}_S(\Gamma)$ as the optimization function. Unfortunately this guess turned out to be wrong, at least in the general case. In a recent work, Deutsch and Kurosky (DK) [55, 49] observed that the function to be minimized in sequence space is the difference ΔF between the free energy of a sequence pushed in the target conformation Γ^* by a clamping potential $V(\Gamma)$ and the free energy of the unconstrained sequence. The clamping potential $V(\Gamma)$ is assumed to have a deep minimum on Γ^* and scaffolds the sequence to the target conformation Γ^*. The sequence S having its ground state in Γ^* does not pay any price to $V(\Gamma)$ and the difference ΔF is zero (having assumed, without loss of generality that $V(\Gamma^*) = 0$). Specializing the discussion to the HP model and to $V(\Gamma) = \delta(\Gamma - \Gamma^*)$ in order to pick out one specific structure, they approximated the free energy to be temperature independent and given by the lowest order cumulant

$$F_S = -T\log(Z_S) \sim \sum_{ij} B(s_i, s_j)\langle\Delta(\mathbf{r}_i - \mathbf{r}_j)\rangle_\Gamma = \frac{1}{N_\Gamma}\langle\mathcal{H}_S\rangle_\Gamma, \tag{3.5}$$

where the average $\langle\cdot\rangle_\Gamma$ is performed over all the N_Γ conformations having 7 or more contacts (compact conformations have 9 contacts for the $N = 16$ case on the square lattice). We note that the DK approach is an approximate high temperature cumulant expansion of F_S and leads to a 50%-70% success rate for the HP model.

3.2 Optimal Design Procedure

In this section we give the solution to the inverse folding problem on general grounds and we present a method to implement it.

The statistical weight of an arbitrary sequence S in an arbitrary conformation Γ at temperature T is

$$P_S(\Gamma) = \frac{\exp[-\mathcal{H}_S(\Gamma)/T]}{Z_S} \tag{3.6}$$

where $\mathcal{H}_S(\Gamma)$ is the energy of the sequence S in the conformation Γ and

$$Z_S = \sum_{\Gamma} \exp[-\mathcal{H}_S(\Gamma)/T], \tag{3.7}$$

is the partition function. Let Γ_G be one of the ground state conformations of sequence S. As $T \to 0$, $P_S(\Gamma_G) = 1/g$, where g is the ground state degeneracy. For the $g = 1$ case, the folding transition temperature T_f is defined as the temperature at which

$$P_S(\Gamma_G)\,|_{T=T_f} \equiv 1/2. \tag{3.8}$$

Given a target structure Γ^*, the desired goal is to find a sequence S^* such that $P_{S^*}(\Gamma^*) \to 1$ in the $T \to 0$ limit. A brute-force way of achieving this is to calculate $P_S(\Gamma^*)$ at sufficiently low temperatures for all sequences and identify S^* as the sequence that maximizes $P_S(\Gamma^*)$. We will show that it is possible in many cases to find the correct sequence S^* by working at high temperatures, which is generally simpler. Situations in which the chosen Γ^* is not the unique ground state of any sequence S are also correctly identified. In other cases, there could be many sequences that have Γ^* as their unique ground state. From the temperature dependence of $P_{S^*}(\Gamma^*)$ for each of these sequences, one may select a sequence with a desired T_f. Even though T_f is obtained from thermodynamics and not kinetics, earlier studies have shown an excellent correlation between a propensity for rapid folding and thermodynamic stability [56, 15, 11, 53, 25].

A simple and general way of implementing these observations is by means of an importance sampling dual MC procedure, both in sequence and in conformation space. The first element of the dual MC scheme is the computation of $P_S(\Gamma)$ from Eq. (3.6) for a given sequence S, a given conformation Γ and at a fixed temperature T. The partition function Z_S can be trivially recast as [57, 58].

$$Z_S = C_{tot}\frac{\sum_\Gamma \exp[-\mathcal{H}_S(\Gamma)/T]}{\sum_\Gamma 1} \tag{3.9}$$

where $C_{tot} = \sum_\Gamma 1$ is the total number of conformations of a sequence S, and the sum over Γ is extended over all these conformations. Generally, it is not practical to generate all conformations of a given sequence in order to evaluate the associated Z_S. Noting that

all conformations do not contribute equally to the partition function, a simple importance-sampling scheme entails the dynamical growth of M independent conformations in a step-by-step manner so that each conformation is generated with a probability

$$\frac{\exp[-\mathcal{H}_S(\Gamma)/T]}{W(\Gamma)}, \tag{3.10}$$

where $W(\Gamma)$ is its Rosenbluth weight [59] which will be detailed in the following. Eq. (3.9) can be reexpressed as

$$Z_S = C_{tot}\frac{\sum_\Gamma W(\Gamma)\left[\frac{\exp[-\mathcal{H}_S(\Gamma)/T]}{W(\Gamma)}\right]}{\sum_\Gamma W(\Gamma)\left[\frac{\exp[-\mathcal{H}_S(\Gamma)/T]}{W(\Gamma)}\right]\exp[\mathcal{H}_S(\Gamma)/T]} \tag{3.11}$$

Since conformations are generated with the probability given by Eq. (3.10), undoing the bias in the sampling we obtain as an estimate,

$$Z_S \simeq C_{tot}\frac{\sum_{\Gamma=1}^{M} W(\Gamma))}{\sum_{\Gamma=1}^{M} W(\Gamma)\exp[\mathcal{H}_S(\Gamma)/T]}. \tag{3.12}$$

This procedure can be extended in a straightforward manner to discretized off-lattice models.

The Rosenbluth weight $W(\Gamma)$ is obtained by the following considerations. A walk is grown step by step. The attempt to add step $i+1$ is shown in Fig. 3.3. Without violating the self-avoidance condition there are $j = 1, .., k$ possible routes, ($k = 2$ in the case shown). Each new position j is characterized by a set of variables ψ_j describing its environment due to previous steps. Let $\epsilon_i(\psi_j)$ be the incremental energy associated with the $(i+1)$-th step into the site j. The total energy at the end of a walk of N steps is then

$$\mathcal{H}_S(\Gamma) = \sum_{i=1}^{N-1} \epsilon_i(\psi_{j(i+1)}), \tag{3.13}$$

where $j(i)$ denotes the site visited at the i-th step. The $(i+1)$-th step is chosen to terminate in site j with a probability

$$e^{-\epsilon_i(\psi_j)/T} / \sum_k e^{-\epsilon_i(\psi_k)/T}. \tag{3.14}$$

For a chain Γ the Rosenbluth weight $W(\Gamma) \equiv W_N$ is obtained using the recursion relation

$$W_{i+1} = W_i \sum_k e^{-\epsilon_i(\psi_k)/T}. \tag{3.15}$$

The second element of the dual MC code is the search for the sequence S^* that maximizes $P_S(\Gamma^*)$ for the given target conformation Γ^*. We start with an initial sequence S and estimate $P_S(\Gamma^*)$ using the procedure described in the previous paragraph. We then make a trial change in the sequence and determine the new value of $P_S(\Gamma^*)$. If $P_S(\Gamma^*)$ increases, we

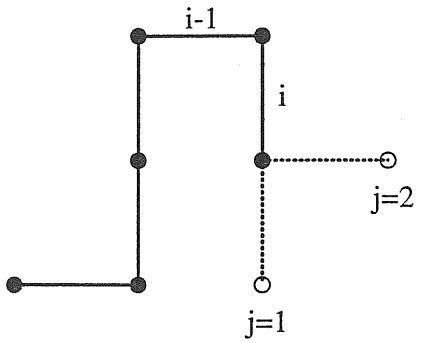

Figure 3.3: Step by step growth of the walk. The step $i+1$ can be chosen in two ways.

accept the change. If $P_S(\Gamma^*)$ decreases, we accept the change with a probability controlled by a fictitious temperature θ (not related to T_f) using the standard Metropolis algorithm [48]. We start with a high value of the fictitious temperature θ at which most sequence changes are accepted, and lower it slowly (and approximately linearly with time) until no further changes are accepted typically after about 1000 sequences have been run through. We have written such a dual Monte-Carlo code for carrying out the maximization procedure at any arbitrary temperature T.

Consider a desired ground state conformation Γ^*. In general, we can classify the sequences S into three classes that we will denote as "good", "medium" and "bad" for the particular conformation. Good sequences have Γ^* as a unique ground state, medium sequences have, in general, g degenerate ground states ($g > 1$) with one of them being Γ^*, whereas bad sequences have ground state conformation(s) that do not include Γ^*. With this classification, both good and medium sequences have the desired ground state conformation Γ^*. In more realistic models, one might expect that degeneracy would play less of a role and fewer medium sequences would be be present.

Let us consider the temperature dependence of $P_S(\Gamma^*)$. Good sequences have $P_S(\Gamma^*) = 1$ at zero temperature and this value decreases monotonically to $1/N_\Gamma$ at very high temperatures (N_Γ is the total number of conformations), as shown in Fig. 3.4. Medium sequences are similar in their behavior except that at low temperatures they asymptote to $1/g$. Bad sequences have P equal to zero at $T = 0$ whereas P approaches the $1/N_\Gamma$ value at very high temperatures. There are two simple scenarios – a monotonic increase of P as the temperature increases or a relative maximum at some intermediate temperature. In either case, we find that, at high temperatures, the sequence with the largest value of P is either a good or a medium sequence. We have verified this explicitly through an exhaustive enumeration of one of the lattice models described below and in all the cases studied for the second lattice model.

Armed with this insight, we suggest that one may use the MC scheme in the high temperature limit to narrow down one's search to good and medium sequences. In this limit,

simple high T expansions of the denominator of Eq. (3.6) suffice. An even simpler way to weed out bad sequences is to evaluate the derivative of P with respect to T and to discard sequences with a positive derivative. The analysis may then be extended to lower temperatures to sort out the medium from the good sequences, to deduce the folding transition temperature and to confirm that one has obtained the right answer.

A well know method to extract more information from a MC run is the histogram technique [60, 61]. The basic idea is to reconstruct the density of states $n(E)$ from the knowledge of the calculated distribution function $h_\beta(E)$ of the energy E at some temperature $T = 1/\beta$. By definition

$$h_\beta(E) = \frac{n(E)\exp(-\beta E)}{Z_\beta}. \tag{3.16}$$

Inverting this equation we get an estimate of $n(E)$

$$n(E) = h_\beta(E)\exp(\beta E)Z_\beta, \tag{3.17}$$

up to a multiplicative constant, Z_β. The average of an observable A is then given by

$$\langle A\rangle_\beta = \frac{\sum_E A(E)n(E)\exp(-\beta E)}{\sum_E n(E)\exp(-\beta E)}, \tag{3.18}$$

where the constant Z_β cancels out.

In the case of our MC, the density of state requires a further reweighting. By definition

$$n(E) = \sum_\Gamma \delta(E - E_\Gamma) \tag{3.19}$$

Again we can recast as

$$n(E) = C_{tot}\frac{\sum_\Gamma \delta(E-E_\Gamma)}{\sum_\Gamma 1} = C_{tot}\frac{\sum_\Gamma \delta(E-E_\Gamma)W(\Gamma)\left[\frac{\exp[-\mathcal{H}_S(\Gamma)/T]}{W(\Gamma)}\right]\exp[\mathcal{H}_S(\Gamma)/T]}{\sum_\Gamma W(\Gamma)\left[\frac{\exp[-\mathcal{H}_S(\Gamma)/T]}{W(\Gamma)}\right]\exp[\mathcal{H}_S(\Gamma)/T].} \tag{3.20}$$

Since our configurations are generated with weight of Eq. 3.10 we get the following estimate for the density of states

$$n(E) \simeq C_{tot}\frac{\sum_\Gamma \delta(E-E_\Gamma)W(\Gamma)\exp[\mathcal{H}_S(\Gamma)/T]}{\sum_\Gamma W(\Gamma)\exp[\mathcal{H}_S(\Gamma)/T].} \tag{3.21}$$

It is worth noting that in a standard MC the latter equation can be used setting $W(\Gamma) = 1$. From $n(E)$ it is easy to obtain the partition function $Z_{\beta'}$ at any temperature

$$Z_{\beta'} = \sum_E n(E)\exp(-\beta' E), \tag{3.22}$$

together with all the other averages of interest. This procedure is very useful, since we are interested in following the behavior in temperature of $P_S(\Gamma)$, after having estimated it at some high enough T.

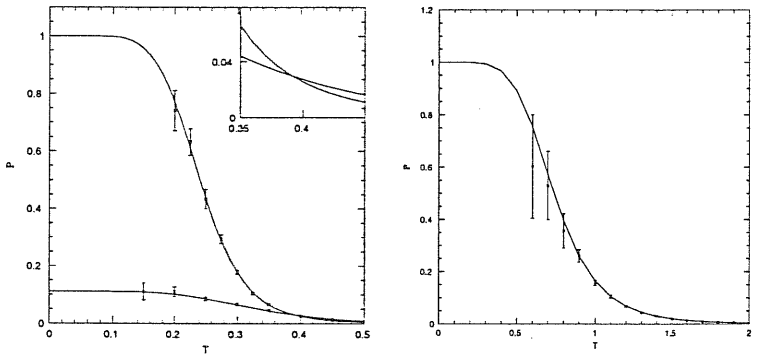

Figure 3.4: Dependence of P on T obtained from a MC calculation in conformation space. The optimal sequence for the given target conformation was first determined using the dual MC code. The points shown were deduced using 30 runs of 10^5 dynamically generated conformations each. The error bars are standard deviations over the independent runs. As a check of the overall precision, we also show the exact curve obtained by a complete enumeration of all the conformations. (Left) The good sequence for the HP model (upper curve) and the medium sequence as determined by the DK method (lower curve). In both cases, the target conformation is the one shown in Fig. 3.2. As shown in the inset, the DK method, being a high T expansion, fails because there is a crossing around $T \sim 0.4$ of the good and the medium curves. (Right) A good sequence for the 16 amino acids model.

3.3 Protein Design on 2D Simple Models

We now specialize our discussion to the two dimensional HP model. In order to test our procedure we have considered relatively short chains (N=16 and N=12 on a square and triangular lattice respectively) for which exact enumeration is possible. On a square lattice for $N = 16$ monmers, there are 802075 different (compact and non-compact) conformations Γ unrelated by simple symmetries. There is a set $\{\Gamma^*\}$ of 456 good conformations and a set $\{S^*\}$ of 1539 sequences with a unique ground state in one of these good conformations (different sequences can have the same good conformation as their native state). Likewise, for the triangular lattice (N=12), there are 16 good conformations (out of 1472412) and a corresponding set of 16 good sequences. Our tests were carried out on all of the good conformations using the above described dual MC algorithm and resulted in a success rate of 100% in the determination of good or medium sequences. Fig. 3.2 shows an example of a conformation for which the methods proposed in references [45] and [49] fail. Our method can also be used to design optimal sequences not only for good but also for generic target conformations, although in the latter case we find only medium sequences.

In order to reduce the number of medium sequences and make the model slightly more re-

alistic, we extended the HP model on a square lattice and considered chains of 16 monomers made up of one each of 16 kinds of amino acids. The Hamiltonian of this model is again given by Eq. (2.2), where B is a 16×16 matrix whose elements are drawn from a Gaussian distribution with mean value -2 and variance 1. The model is similar in spirit to that of Sali *et al.* [56]. Such random contact energies are in approximate correspondence to a more realistic parameterization of the contact energies given by Miyazawa and Jernigan [32, 33] or Kolinski, Godzik and Skolnik [34]. For a given sequence S, we are still able to enumerate exactly all possible conformations so that we can have an independent check on what the ground state is and whether it is unique. However, the total number of sequences is $16! \sim 10^{13}$. Whereas an exact enumeration of the 2^{16} sequences of the HP model was feasible, now the MC annealing procedure in sequence space is essential. We have randomly selected 100 conformations Γ on a square lattice with the number of contacts between 7 and 9. For each of them we have applied the dual MC scheme in order to maximize Eq. (3.6) at a temperature T=0.5. Typically 10^3 sequences S where sampled during the annealing procedure and for each of them 10^5 conformations were dynamically generated to estimate Z_S in Eq. (3.6). Again we have 100% success, i.e. we always succeeded in finding a sequence S^* which has Γ^* as its ground state so that $P_{s^*}(\Gamma^*) \to 1$ in the $T \to 0$ limit. We first identified the desired sequences for both models using our dual MC technique. As a check, we then determined the temperature dependence of P for these sequences using a MC scheme in conformation space. Typical results, along with an exact calculation, are shown in Fig. 3.4 . For a given target structure Γ^*, if there are many good sequences, it is also possible to choose among them the one with the folding transition temperature closer to a desired target design temperature as shown in Fig. 3.5.

As we have seen, a relation between energy gap and folding temperature is expected [5, 12]. This conjecture relates a property of the spectrum of the protein to its dynamical behavior and it is very interesting to verify its validity. In Fig. 3.6 we show the correlation between the folding temperature T_f of the sequences we have designed and the energy gap for the $B_{i,j}$ model. In the HP model the energy gap between the ground state Γ of a given sequence and its first excited state Γ' is typically 1, since the energy defined in Eq. (2.1) essentially counts the number of HH contacts in a given conformation. We ought to resort to another property of the lower part of the energy spectrum. We consider the density of states $n(E)$ at energy E. We plot $n(E_{fe})$, where n_{fe} is the energy of the first excited energy level, as a function of T_f in Fig. 3.6. For the set $\{S^*\}$ of 1539 good sequences we have identified, the stability of the ground state at low temperatures correlates to $n(E_{fe})$.

We have seen that, within the simple models investigated in this chapter, a small folding temperature T_f is related to the presence of a large number of decoys, or conformation competing in energy with the ground state, which make the design procedure hard. This is possibly the reason why the SG and the DK methods are weak in designing sequences

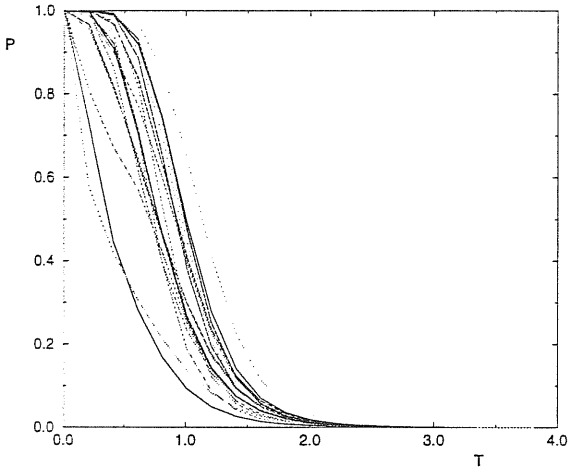

Figure 3.5: Probability as a function of temperature for many $B_{i,j}$ sequences having their ground state on the same conformation. The most stable designed sequence is that with the highest T_f.

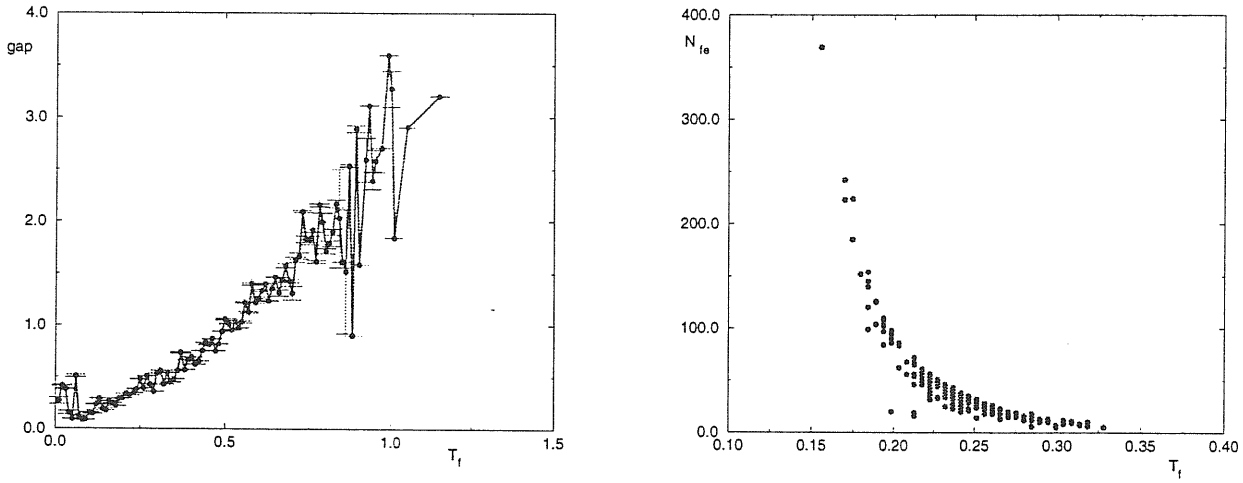

Figure 3.6: (Left) Energy gap to the first excited state versus the folding temperature T_f for the $B_{i,j}$ model. Designed sequences having a unique ground state are considered. (Right) First excited state population $n(E_{fe})$ versus T_f for the HP model.

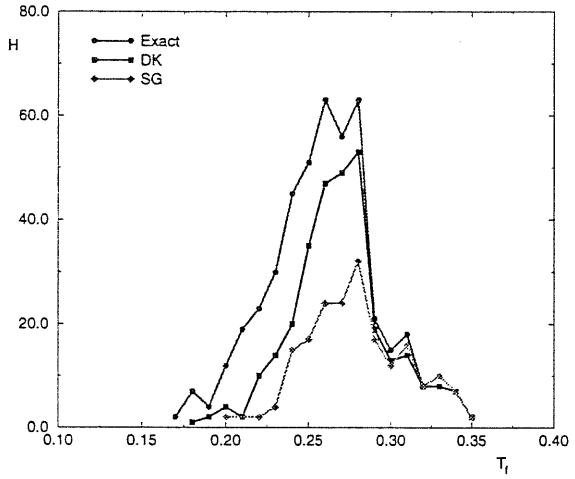

Figure 3.7: Histogram of the folding temperatures of the correctly designed sequences for the three methods discussed.

with a low T_f, as can be seen from the histogram in Fig. 3.7. The upper curve is the number $H_{\text{exact}}(T_f)$ of sequences in the set $\{S^*\}$ having a given folding temperature T_f, as determined by our Monte Carlo optimization scheme, which coincides with the exact results, known by full enumeration. The other two curves are the number H_{DK} and H_{SG} of sequences in $\{S^*\}$ correctly identified by the two other methods. The overall number of successes is 308 for the DK method and 192 for the SG method.

3.3.1 Landau-Ginzburg Expansion

The design procedure consists in a biased random walk in sequence space and, for each new sequence S, requires the evaluation of the Boltzmann probability $P_S(\Gamma^*) = \exp[-\mathcal{H}_S(\Gamma^*)/T]/Z_S$. We have shown how to compute Z_S by means of a Monte Carlo in configuration space. However this procedure is rather time consuming and it would be appreciable to develop an approximation scheme to avoid a new calculation of Z_S for every new sequence S. In this section we present a Landau–Ginzburg approach [62] for the evaluation of the free energy $F_S = -T\log(Z_S)$ for the HP model. In general, since F_S is a result from an averaging over configurations one may expect that it is a simple function of a continuous variable like a local magnetization resulting from grouping s_i variables along the chain

$$M_x = \sum_i^{N_x} s_i, \tag{3.23}$$

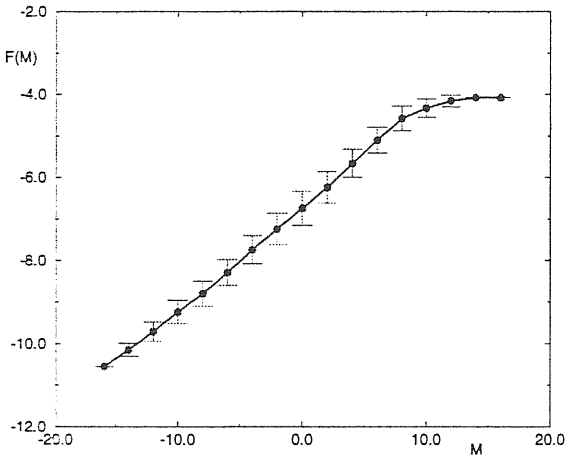

Figure 3.8: $F(M)$ in the Landau approximation of Eq. (3.24) at T=0.3, obtained by exact enumeration for the HP model in 2D for $N = 16$.

having identified H=+1 and P=-1. Here, N_x is the number of monomers in the coarse graining procedure at the position x. The free energy will be approximated by

$$F_S \sim F(\{M_x\}) = a_0 + \sum_x (a_1 M_x + a_2 M_x^2 + (\nabla M_x)^2 + \cdots), \tag{3.24}$$

where ∇M_x is a gradient term. As a first step we considered the magnetization M of the entire chain. The average and the standard deviation of $F(M)$ are evaluated by exact enumeration over all sequences of given M for the$N = 16$ case. Fig. 3.8 shows a plot of $F(M)$ versus M for $T = 0.3$. The fact itself the standard deviation is rather small indicates that $F(M)$ is a rather good approximation for F_S. However, it is not correct to assume, as in SG approximation, that F_S is a constant. Even fixing the composition M (as they do [45]), the variance of $F(M)$ is large enough to make this approximation wade (see Fig. 3.8). In this zero-th order approach, Eq. (3.24) becomes

$$F_S \sim F(M) \sim a_0 + a_1 M + a_2 M^2 + C_1 \,, \tag{3.25}$$

where as the gradient term C_1 we used the average number of contact that a given sequence S can have. (This is exactly $\langle \mathcal{H}_S(\Gamma) \rangle / N_\Gamma$, i.e. the DK approximation.) In general, in the practical implementation of this method, since the standard deviation of $F(M)$ is small, one can give an estimate of $F(M)$ at the begin inning by averaging F_S (obtained by a Monte Carlo in configuration space) over a small number of sequences. As a check of the method, we used the exact form of $F(M)$ at T=0.3, shown in Fig. 3.8, to give a "best fit" of the constants a_0, a_1 and a_2. Then using such values we performed a Monte Carlo optimization in sequences space and we found 314 good configurations out of the 456 for

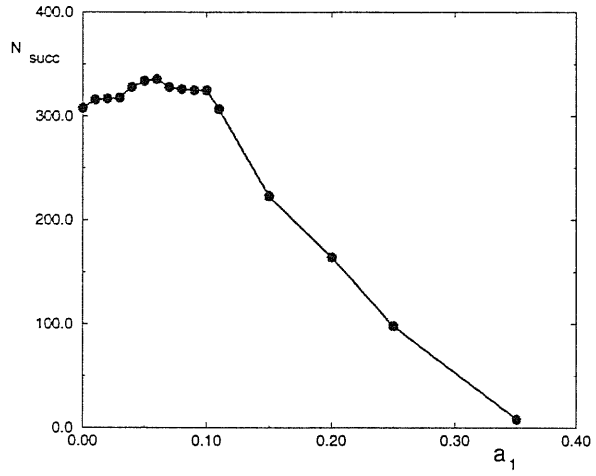

Figure 3.9: Number of successfully designed sequences considering only the term a_1 in Eq. (3.24).

the square lattice case. It is interesting to note that there is a good resemblance between the exact $F(M)$ at T=0.3 and the one obtained using the DK approximation, which is temperature–independent. This coincidence explains the good rate of successes (308) of the DK approximation, which is able to reproduce $F(M)$ at a temperature close to the typical folding temperature (see Fig. 3.7) where the optimization in sequence space is expected to be most effective.

There is a systematic increase in the number of successes including more terms in the expansion. As a first step we neglect the gradient and the quadratic terms. From a linear fitting of the exact form of $F(M)$ at $T = 0.2$ one gets the best fit estimate a_1=0.25. We used the approximation in Eq. (3.25) for F_S with a_1 ranging from 0 to 0.5 to design the good sequences over the 456 good configurations. Quite consistently, if we plot the number of successes versus a_1 we found a sharp maximum in the number of successes around a_1=0.28 where there are 345 successes (see Fig. 3.9). We note that at the "best fit" value a_1=0.25 we got 281 successes. To improve this result we did a quadratic fit which gave a_1=0.245 and a_2=-0.003. The sequence design gave 303 successes. Finally we included the gradient term C_1 and fitted a_0, a_1 and a_2 from $F(M)(T = 0.2) - C_1 = a_0 + a_1 M + a_2 M^2$, finding 314 successes. We see that there is an increase in the number of successes including more terms in the expansion Eq. (3.25) and using the "best fit" values for the constants, (although these values can be adjusted, in a non–systematic way, to get better results). To summarize, being able to give an estimate of $F(M)$ at a temperature where most of the sequences have folded, either with MC or otherwise, then an approximation of the type of Eq. (3.25) works yielding a rather high rate of successes.

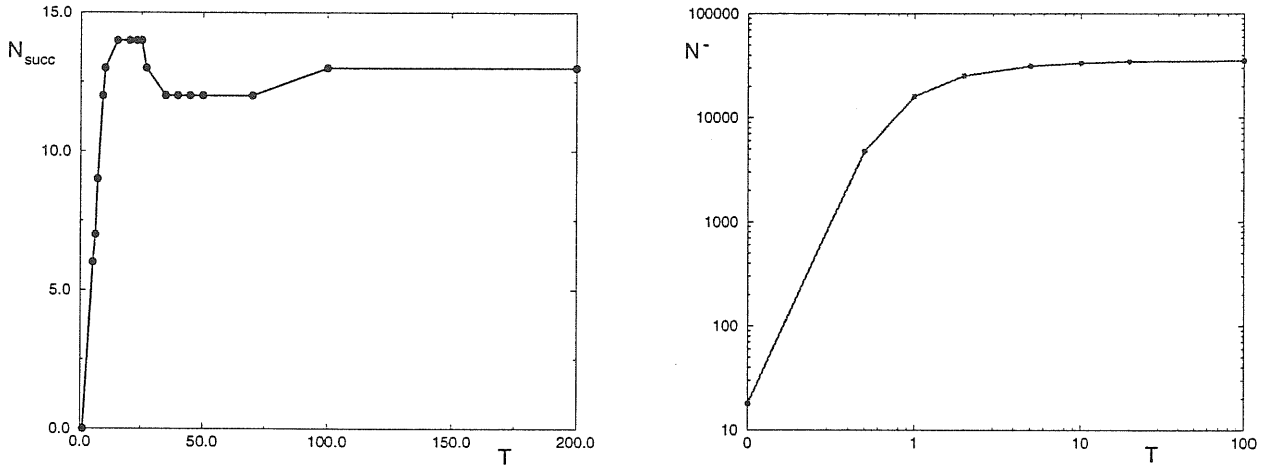

Figure 3.10: (Left) Number of successes with the second cumulant in the triangular lattice case. (Right) Number $N^-(T)$ of sequences with $P_S(\Gamma^*)$ with a negative derivative as a function of temperature.

3.3.2 High Temperature Expansion

In order to avoid the overload of computing the average in conformation space needed to evaluate the Boltzmann probability of Eq. (3.6), we pursue father the idea of DK to do a high T expansion. The DK method is a high–T cumulant expansion truncated to the first term and it is temperature independent ($T = \infty$ case). Adding the second term to this expansion, an improvement is expected. The free energy is approximated by

$$F_S \simeq -\frac{1}{\beta}\log N_\Gamma + \frac{\langle \mathcal{H}\rangle_\Gamma}{N_\Gamma} - \frac{\beta}{2}\left(\frac{\langle \mathcal{H}^2\rangle_\Gamma}{N_\Gamma} - \frac{\langle \mathcal{H}\rangle_\Gamma^2}{N_\Gamma^2}\right) \tag{3.26}$$

Fig. 3.10 shows the result of this approximation at various temperatures as compared with the $T = \infty$ case for HP sequences of length N=12 on a triangular lattice, where DK method melds 13 successes in the design of the 16 good conformations. At intermediate T it may happen that the number of successes decreases with respect to the $T = \infty$ case - the DK case. This indicates that some of the successes at $T = \infty$ are accidental. The trend is confirmed by the results for the square lattice. At T=10 we find 318 successes, and at T=1 we find 323 successes, at T=0.5 we find 336 successes, and at T=0.2 the number of successes decreases. These values are to be compared to 308 obtained with the first cumulant [49]. Another possible way of using the information offered from the second cumulant is to look at the derivative of $P_S(\Gamma^*)$ with respect to T in a range of T where the approximation is expected to be accurate. Since good sequences must have a negative derivative we can discard sequences with a positive derivative. At high T the number N^- of sequences with a negative derivative turns out to be very high, of the order of 50%, as for example in the

N=16 HP case shown in Fig. 3.10. Lowering the temperature this number decreases, and this method gains validity.

It is possible to further improve the expansion by adding more terms. Morrissey and Shakhnovich [44] have recently proposed a mean field approximation to compute the average in conformation space needed to estimate higher order cumulants. The same result can be obtained by means of our Monte Carlo method. The following formula can be derived to obtain numerically the first cumulants

$$\frac{d^k \log Z}{dx^k} = \frac{1}{Z}\frac{d^k Z}{dx^k} - \sum_{\{s_1,s_2,\ldots\}\{n_1,n_2,\ldots\}} \left[\frac{d^{n_1} \log Z}{dx^{n_1}}\right]^{s_1} \left[\frac{d^{n_2} \log Z}{dx^{n_2}}\right]^{s_2} \cdots \frac{k!}{\prod s_i!(n_i!)^{s_i}} \tag{3.27}$$

where the sum is restricted to couples (s_i, n_i) such that $\sum_i s_i n_i = k$, $n_i < k$, and all the n_i different.

3.4 Protein Design on 3D Simple Models

In this section we apply the Monte Carlo optimization to 3D lattice models. It is very instructive to start with a description of a failure, from which we can learn many difficulties hampering progress in protein design [63]. Two well known research teams in protein science, the Harvard and the San Francisco groups, undertook the following challenge. The Harvard group chose 10 target structures of length 48 on a 3D cubic lattice and designed them within the HP model by using the SG procedure [45]. As their outcome, they found 10 putative sequences, each one expected to have their native state on one of the target structures. The sequences were sent, without the correspondent target structures, to the San Francisco team, who performed a blind test, folding them to find the global minimum. They used two numerical methods, hydrophobic zippers (HZ) [64] and a constrained hydrophobic core construction (CHCC) [65]. Only in one case a structure with a lower energy has not been found. Moreover, for each sequence the global minimum was found to have a large degeneracy, with 10^3 to 10^6 energetically equivalent conformations. There could be a number of reasons behind this failure. A binary letter code, embodying only hydrophobic and hydrophilic features of amino acids, can be too unspecific to ensure designability of structures. This observation raises the intriguing question of which is the minimal number of species of amino acids that is required to provide the necessary diversity underlying uniqueness and stability of the native state.

We point out that although in principle with our method we are able to solve the inverse folding problem, we should be able to have the control on the direct problem. To state it in a suggestive way we can say that unless we are able to fold a protein we cannot design it. The model studied by Socci and Onuchic (SO) [47] is used as a benchmark. Sequences are designed with out dual MC, and as target conformations we used the ground states of

sequences 002 and 006 of Ref. [47]. The folding transition in this model is first order-like [47], and below T_f the probability distribution function of the energy

$$P_S(E,\beta) = n(E)\frac{\exp[-\mathcal{H}_S(\Gamma)/T]}{Z_S(\beta)} \tag{3.28}$$

is double peaked [66]. To compute $Z_S(\beta)$ we growth conformations in the usual way. However, the lower part of the energy spectrum escape an efficient sampling, and this fact hampers the design at a temperature close to T_f. As the analysis in 2D has shown, it is still possible to carry out the design procedure at a temperature T above T_f. In this range of temperature $P_S(E,\beta)$ is single peaked and centered around a value of the energy that is easily accessible to our sampling scheme. At this temperature our estimate of Z_S is reliable and our dual MC is efficient. A run time test is given by estimating $P_S(\beta')$ below the temperature T of simulation, using the reweighting procedure discussed above. If the selected sequence has its ground state on the target conformation the $P_S(\beta')$ increases in decreasing T whereas it goes very quickly to zero if the ground state is on some other conformation. The designed sequence is subjected to a Monte Carlo folding simulation to verify the native state. (In the cases of sequences 002 and 006 of SO we have been able to reproduce their results in full details with our folding Monte Carlo.) Then we took the ground state structures Γ_2 of sequence 002 and Γ_6 of sequence 006 and we designed them. There are several sequences with their ground state on the same conformation As the optimization procedure in sequence space goes on, it selects more and more stable sequences. This can be verified by looking at $P_S(\Gamma)$ as a function of T. The more stable sequences are those with higher T_f. In order to investigate further our design scheme we repeated the same analysis on Γ_2 and Γ_6 on the HP and the MJ models. As a first step a series of simulations of folding has been done on randomly selected sequences in order to have an estimate of a typical T_f. Then the usual design scheme has been carried out at T slightly above T_f. And finally the designed sequences were refolded in order to check if a configuration with a lower energy can be found. The same conformations Γ_2 and Γ_6 were used, and we verified that the design scheme is always successful.

In perspective, it would be interesting to apply our design scheme to the problem of molecular recognition. Enzymatic proteins typically are functional when they chemically bind some specific ligand. Such chemical bond is often realized by a few specific amino acids in a strategic position in the folded structure. It is believed that the overall 3D structure of the protein is a product of evolution to ensure the stability of the functional section. Examples are the process where a ligand (O_2 or CO) binds the internal heme site in hemoglobin or myoglobin [67]. and antibody-antigene complex [1]. We are interested in the thermodynamics of such interactions, namely in predicting a sequence of amino acids that has a native structure capable of binding a ligand. In our method this requirements enters as a constraint in the search in sequence space.

4 How Does Natural Selection Work on Proteins?

Because the number of possible random amino acid sequences and the number of possible conformations is huge, a key issue is understanding the selection principles that apply to protein sequences and native state structures. In this chapter we discuss the mechanism of evolution through natural selection in proteins and we discuss its implication in the protein design problem. We address this issue by probing the stability of the native state against perturbations in the effective interaction potential between amino acids. Our calculations, within the framework of simple 2D models, suggest that random heteropolymers are not stable against mutations, whereas "evolved" sequences are characterized by a non-zero stability threshold. Protein design strategies relying on thermodynamic stability optimization have been proposed [45, 50, 49, 68]. Here, we explicitly show that a large energy gap implies also robustness against mutations. However, we suggest that natural selection, taking advantage from diversity supplied by the 20 species of amino acids, acts in the reverse way by promoting sequences that are not easily mutated away. We introduce an evolution-like protein design scheme that works by maximizing the stability threshold and we show that mutation stability implies also thermodynamic stability, which would emerge as an indirect consequence of the evolution mechanism. The requirement of stability against mutations suggest an explanation to the emergence of families of folds gathering ensembles of homologous sequences ("twilight zone") [37]. There are special structures that can host a large number of sequences providing them high stability for varying physiological conditions. As Li *et al.* [22] recently proposed, structure selection is complementary to sequence selection. However, as we have shown in Chapter 2, they do not support convincing evidence for this fact.

4.1 Protein Design and Stability against Mutations

In real proteins, mutations are realized by specific mechanisms which affect the unknown interatomic potential in a complex way [1, 69, 70]. We introduce a general type of pertur-

bations of which naturally occurring mutations are a particular subset.

Using the previously introduced $B_{i,j}$ and MJ models in 2D, we study the sensitivity of the native state to mutations modeled as perturbations in the interaction potential between amino acids. For a given sequence with $N \leq 25$, we enumerate the energies of all possible conformations. We are therefore able to determine the native state conformation exactly. In the $B_{i,j}$ model, in order to model evolved sequences with a large stability gap, we follow the rank-ordered procedure outlined by Shrivastava *et al.* [53] of shuffling the $B_{i,j}$ entries to assign the most favorable attractive interactions to the native contacts of a desired compact ground state. In the MJ model, each monomer is chosen to represent one of the twenty amino acids with the interactions determined by Miyazawa and Jernigan [32, 33]. A random sequence would correspond to a random choice of the amino acids. In order to mimic evolved sequences, it is no longer possible to follow the rank ordering procedure because the $B_{i,j}$ entries cannot be shuffled at will. Instead, one is allowed to move in sequence space by changing one amino acid into another. To obtain evolved sequences with a desired native state conformation and significant thermodynamic stability (or equivalently a large folding transition temperature T_f at which the probability of occupancy of the native state is 1/2), we used our recently introduced protein design procedure [68]. The designed sequences are then subjected to random perturbations.

Our calculations begin with the selection of two statistically similar but distinct interaction matrices which we shall call B and C. We shall consider 4 choices: the random and the evolved $B_{i,j}$and MJ models. The ground states of the B and C sequences are generally distinct. We now consider mutations of the sequence along a trajectory parameterized by a mixing coefficient $a \in [0, 1]$ that changes the interaction matrix from B to C:

$$B_a = (1 - a)B + aC. \tag{4.1}$$

The coefficient a is a measure of the distance in sequence space between B and B_a. The structural similarity of the ground state conformations of these two sequences is given by the normalized distance $\Delta = d(B_a, B)/d(C, B)$, where the distance $d(X, Y)$ is defined by

$$d(X, Y) = \sqrt{\sum_{i,j=1}^{N} (r_{i,j} - r'_{i,j})^2} \tag{4.2}$$

where $r_{i,j}$ and $r'_{i,j}$ are the Euclidean distances between amino acids i and j in the the two native states of sequence X and Y respectively, Note that Δ has been normalized so that it is 1 when $a = 1$, as long as the ground states of B and C are distinct. Our primary probe of the stability to mutations is via a study of the dependence of Δ on a. Qualitatively similar trends are found for both the models – the signature of the selection in sequence space is in the quite distinct behavior of random and evolved sequences.

As the mixing coefficient a is increased starting from zero the ground state of B_a can be changed respect to that of B. In Fig. 4.1 we show how the energy levels of the 69 compact

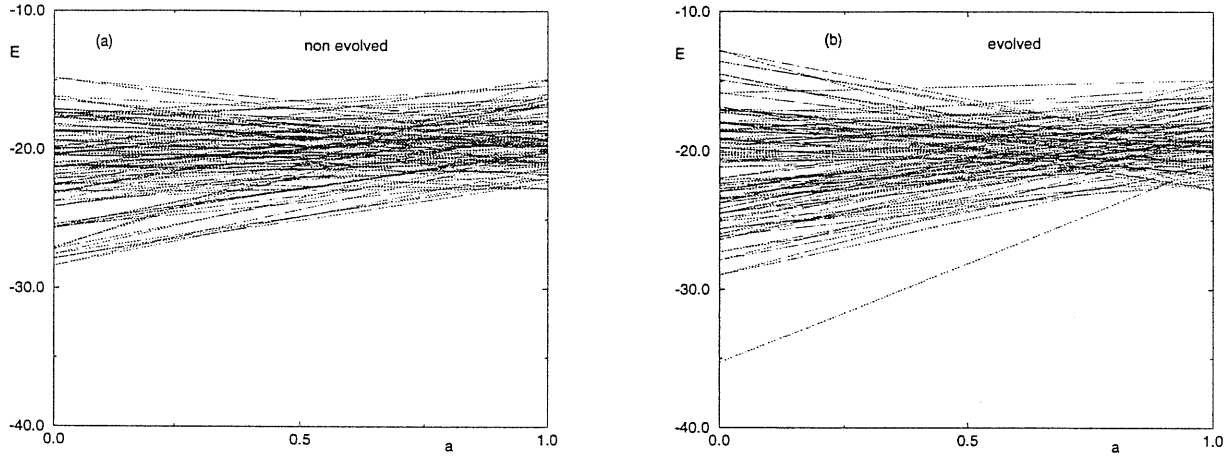

Figure 4.1: Level crossing as a function of the mixing coefficient a. a) The non evolved case. b) The evolved case. For $a = 0$ it is clearly seen the large gap to the first excited state for the rank ordered sequence.

conformations cross as a is varied from 0 to 1. From increasing a, when the lower curve intersects another curve coming from above, the ground state changes. In the non evolved case, the unperturbed sequence B has a very small gap, so its ground state is destabilized for small a. Instead in the evolved case, the gap of B is rather large and the perturbation should become strong to change its ground state. However the detailed behavior of the crossings depends both on the spectrum of B and on the spectrum of C. A large gap would imply stability only in a statistical sense for short chains.

A summary of our results for the behavior of the average Δ as a function of a for $N = 16$ is shown in Fig. 4.2. The curves have been obtained as an average over 1000 realizations of independently chosen B and for each of them over 1000 realizations of C for the $B_{i,j}$ model and over 10 realizations of B and for each of them over 1000 realizations of C for the MJ model. The average stability threshold is zero for random heteropolymers and is distinctly non-zero for the evolved cases.

We define an individual stability threshold $a_t(B, C)$ in the strength a of the perturbation above which Δ becomes non-zero for the first time – the native structure of sequence B is destabilized. Normalized probability distribution $P(a_t)$ of the individual stability thresholds for the random and evolved $B_{i,j}$ models are shown in Fig. 4.3. They underscore the different behaviors in the two cases. Furthermore, the stability threshold goes up with the overall thermodynamic stability as measured by the folding transition temperature. The threshold is somewhat reduced but is clearly non-zero when one considers rank ordered $B_{i,j}$ sequences that have native states in conformations that are not compact. In the evolved

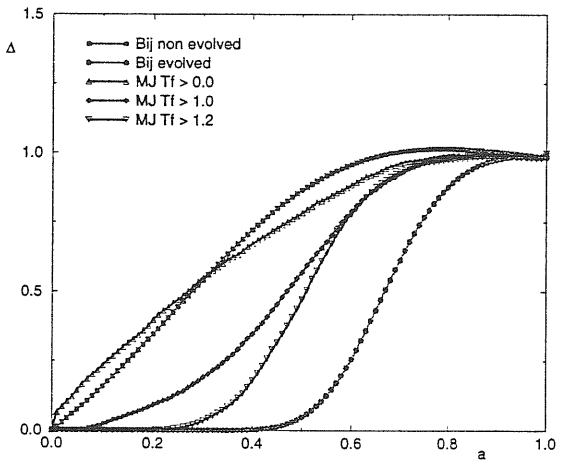

Figure 4.2: Average structural similarity Δ as a function of the perturbation a. The non evolved and evolved cases for both the models used are shown. For the $B_{i,j}$ model the averages have been taken over 1000 realizations of B and for each of them over 1000 realizations of C. For the MJ model we averaged over 10 realizations of B and for each of them over 1000 realizations of C. The design parameter T_f means that in our design procedure we selected only sequences with a folding temperature T_f greater than the indicated value. $T_f = 0$ indicates no selection.

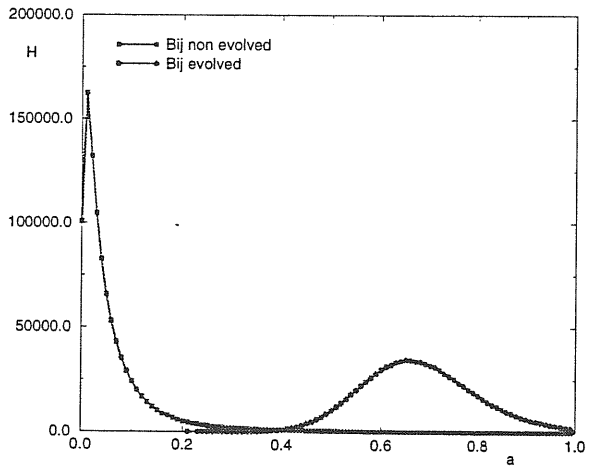

Figure 4.3: Normalized probability distribution $P(a_t)$ of the individual stability thresholds a_t for the non evolved and evolved $B_{i,j}$ case.

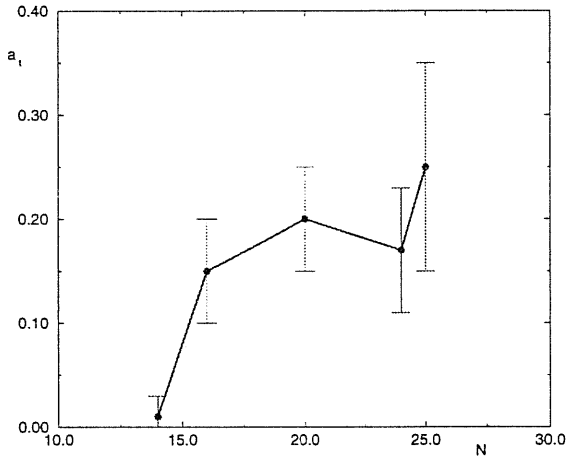

Figure 4.4: Scaling behavior of the threshold for increasing chain length N.

case the distribution of the threshold is well described by a gaussian. We define the overall threshold a_t as the lowest realized $a_t(B, C)$.

For evolved sequences, the stable phase along the a-axis increases in size as shown in Fig. 4.4, along with a sharpening of the $\Delta - a$ curve. This is suggestive of a sharp phase transition at the onset of the thermodynamic limit. We are then tempted to represent the behavior of Δ by a power law as can be seen from Fig. 4.2

$$\langle \Delta(a) \rangle = K(a - a_t)^{\alpha}. \tag{4.3}$$

The actual value of the exponent α can be extracted using a least square fit or the Padè approximants. However we observe that the $B_{i,j}$ model is ill-defined as far as our problem is concerned. There is no lower bound in the interactions between monomers. There is always a non zero probability to extract a matrix C which destabilizes the ground state of sequence B for any non zero strength a. In particular if we approximate $\Delta(a) = \vartheta(a - a_t)$ and $H(a_t) = e^{-(a_o - a_t)^2/\sigma^2}$ (a gaussian), we find that there is no threshold in the average $\langle \Delta(a) \rangle_C$, which starts smoothly from 0 as $\langle \Delta(a) \rangle_C \simeq 1 - \mathrm{erf}((a_o - a)/\sigma)$. To circumvent this problem, one can reasonably assume, as in the case of the MJ model, that the actual distribution of interactions between monomers is limited from below resulting in a distribution of a_t which is limited as well.

Our results are related to a recent study of Bryngelson [71] who used a mean field theory to estimate the probability of predicting the correct structure of a sequence of monomers if the interatomic potential is known only to an accuracy of η. His principal result is

$$P = 1 - k\frac{\sqrt{N}\eta}{B} \tag{4.4}$$

where P is the probability to predict the correct structure for a sequence of N monomers, if the "true" potential is know up to an accuracy η. B is the energy scale of the interactions, and k is a constant of order 1. For $\eta = 0$ the right potential is recovered and the sequence fold with probability 1 in its true native state. A non-zero η could arise

- from variations in the solvent properties
- from imperfect parameterization or determination of the potential between amino acids
- from mutations in the sequences, as in our case.

This result is consistent with a REM approach. In a REM-like model the typical energy difference between low energy states scales as $\sqrt{N}$, and since they are structurally different, a perturbation can lead to a dramatic change in the ground state. This results strongly suggest that real proteins, which have remarkably stable native states, are not well described as random heteropolymers.

In a recent work [72] it has been observed through a mapping of a related model for random heteropolymer folding over the random field Ising model that in 2D even a small perturbation destabilizes the ground state, whereas in 3D a non zero threshold is expected. Indeed this result is consistent with work in that for non evolved sequences we found zero threshold. Our non evolved sequences are in fact random heteropolymers. Only when we perform a sequence design a non zero threshold appears. It would be interesting extend our study to 3D where even for the random heteropolymer case a non zero threshold is expected, to see if such threshold would be enhanced.

We recall as related results recently obtained by Gutin *et al.* [73]. Their design scheme is aimed at optimizing the folding time of in sequence space. A random initial sequence is chosen and its mean first passage time (MFPT) through the native state is calculated. Then a mutation is performed and if the new sequence has a smaller MFPT the mutation is accepted. Sequences so optimized have a Z score significantly greater than random sequences, meaning that evolutionary selection promotes also thermodynamic stability.

4.2 Protein Design by Threshold Optimization

In this section we develop the idea of using threshold optimization to mimic evolutionary activity in sequence as it is found in nature. Sequence are designed to be resistant against perturbations. It is well known that inside the cell imperfectly folded proteins are demolished by proteolytic enzymes [1]. Unfolded proteins can be present either as a product of a destabilizing mutation or due to a variation of the solvent properties. The folded structure

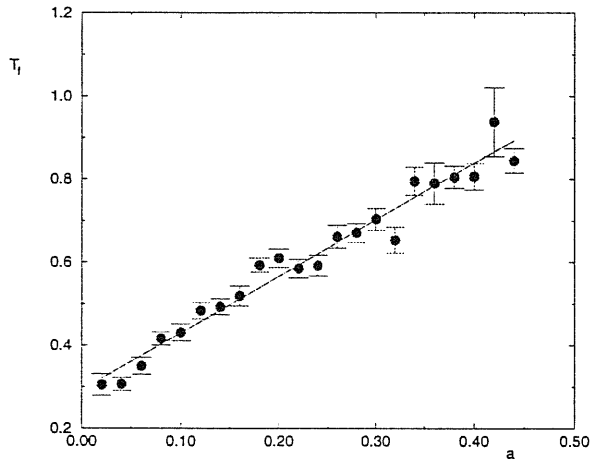

Figure 4.5: Folding temperature T_f as a function of the stability threshold a in the MJ model. Sequences are designed through threshold optimization.

must be robust against perturbations to survive in the cell. Sequence selection primarily produces robust sequences and as a consequence these sequences are rapidly folding on a target conformation with a specific function. Beyond the simple model under study, we propose to use our design scheme by starting from existing functional sequences and producing artificial homologous. In this way we hope to find new sequences with better functionality.

The design scheme works by first selecting an initial random sequence. We compute its MJ matrix B and its stability threshold by extracting a set of 100realizations of the perturbation C. The sequence is then subjected to Monte Carlo optimization. Monomers are swapped and the new sequence is accepted if its threshold is increased. After 1000 Monte Carlo steps the folding temperature T_f of the sequence is computed. In Fig. 4.5 we show the T_f averaged over sequences with the same threshold with its dispersion. A good correlation is found. Our calculation rely on exact enumeration. To generalize the method we can use a Monte Carlo dynamical simulation to evaluate the stability threshold, as for example in Ref. [72]. The same approach will give us also T_f for that sequence, using the histogram technique as in Ref. [47]. We added the constraint that sequences have a compact ground state, since it is expected that the stability gap is more pronounced. However we can relax this request without changing the results.

4.3 The Twilight Zone

Our work provides a characterization of the "twilight zone" and enables the elucidation of the basic mechanism underlying the observation that sequence homology implies similarity of the native structures [1]. Homologous sequences are those deriving from a common ancestor. The concept of the twilight zone has been introduced in the context of sequence similarity detection [74, 75, 76]. The number of amino acid matches obtained by pairing two random sequences of the same length is given by the binomial distribution with $p = 1/20$ if one assumes that the 20 amino acids have the same probability to occur in natural polypeptides. For sequences of length N there will be an average of pN identical amino acids, with a variance of $Np(1-p)$. For example, for random chains of 200 amino acids, 95% of comparisons will yield a similarity between 0% and 9%. Mutual correlations, if they are present, are then undetectable if their amount is inside the above mentioned range – a similarity as high as 9% can arise in the case of no correlation for $N = 200$. The *fidelity* of a comparison, or alignment, is the extent to which mutual correlations are detected.

It is well known that proteins form families according to the spatial conformation of their native states [21] (see Fig. 4.6). Inside a family, a high degree of sequence homology is found. It is then possible to predict the structure of a protein by standard alignment techniques, that is by making comparisons with proteins whose sequences are homologous to the one under study. The exact degree of homology necessary to make the inference is in the range of 25-40% sequence identity. Below this limit, a twilight zone emerges in which sequence homology does not imply structural similarity [74, 37]. Structural similarities probably reflect the evolution of the current array of protein structures from a small number of primordial folds. It is tempting to assume that indeed only a finite number of folds exist. Once this set is eventually determined, the deduction of the native state of a new sequence would reduce, in the simplest situation, to the identification of the fold maximally compatible with this sequence among the limited repertoire of existing configurations. [43] Screening methods such as threading are based on such a scenario [77, 37, 78].

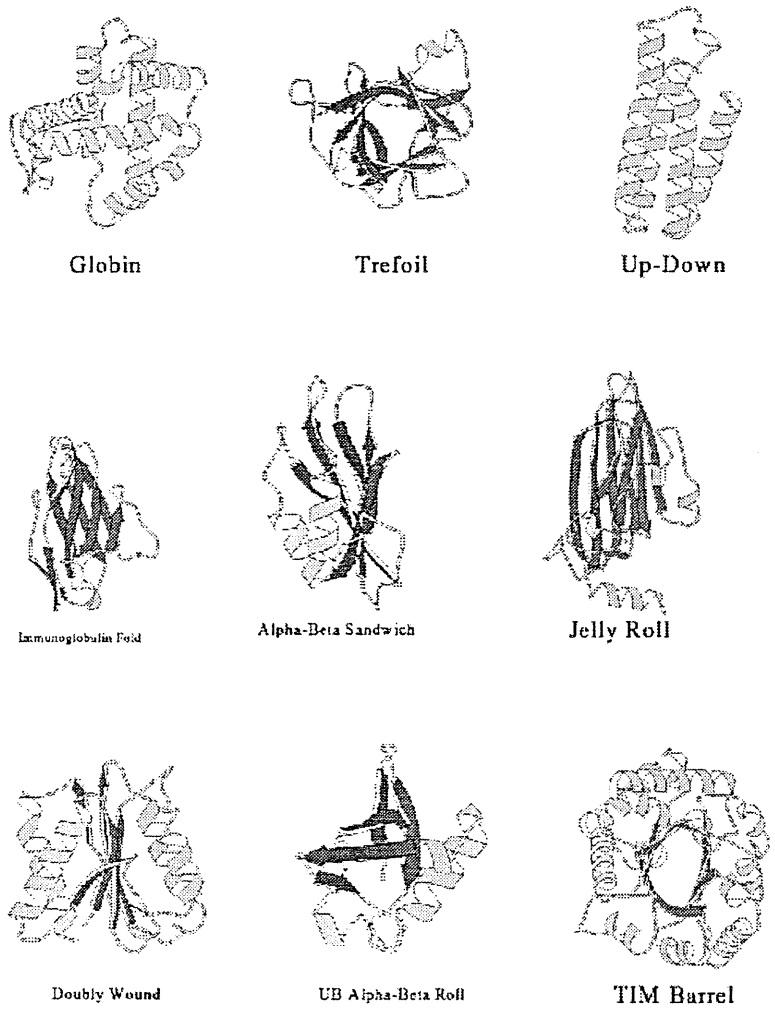

Figure 4.6: The nine superfolds which dominate the structure database [21]

5 A Monte Carlo Method for the Simulation of Realistic Off-Lattice Proteins

The virtue of coarse-grained lattice models, of the type discussed in previous chapters, is that their ground state can often be known exactly, many of their properties are well understood and are believed to resemble those of real proteins [17]. It would be of great importance to extend our predictive ability to off-lattice situations. In this chapter we present a Monte Carlo method for the efficient simulation of off–lattice polymers and more in general of polymer systems. It is an extension of the Configurational Bias Monte Carlo method [79, 80]. Elementary moves consist in regrowing internal segments of a polymer chain. We show that the method satisfies the detailed balance condition. We apply it to three well known simple models used in protein studies, namely homopolymers, random heteropolymers and random copolymers, showing that it is a highly competitive Monte Carlo (MC) algorithm.

5.1 Monte Carlo Simulation of Polymer Systems

We will focus our discussion on the Monte Carlo (MC) simulation of proteins although our method can be applied to a broader class of polymer problems, including polymer mixtures, polymer melts, cross-linked networks, ring polymers and branched chains. Much attention has been recently devoted to the development MC schemes for the simulation of polymer systems [79, 80, 81, 82, 83, 84, 85, 86]. From the computational point of view, the main issue is to obtain an efficient sampling of the conformational space of chain molecules. In this, MC is certainly more flexible than Molecular Dynamics (MD): it does not require to follow the correct time evolution of the system, and it allows non physical moves. How rapidly one can obtain statistical averages depends on the kind of trial moves. In usual MC approaches, trial moves consist of an attempted displacement of a single particle. Since they involve local changes in the conformation, the sampling of the phase space goes on

quite slowly.

A major advance in this field has been the introduction of the Configurational Bias MC (CB) [79, 80]. This method is based on the Rosenbluth sampling scheme [79, 59], and its important feature is that it allows to perform large scale conformational changes in a single step. The polymer conformation is built up step by step by choosing each new position among a set of trials. Hard core overlaps are avoided by selecting trials according to their Boltzmann probability. The well known shortcoming of the Rosenbluth technique, namely the fact that configurations are not obtained with the correct Boltzmann weight, is prevented by introducing a Rosenbluth weight to correct this bias. With the CB method it is possible to regrow the entire chain or a end part of it. The efficiency of the CB method decreases as the length of the chains increases, since the concentration of end parts becomes small. In this work we present a new algorithm, the Modified Configurational Bias MC method (MCB), which allows to regrow internal segments of the chain. Following Ref. [81], where the same modification is presented in the lattice case, the regrowth is guided by a bias towards the fixed end of the segment to be reconstructed.

Other collective moves involving inner segments of the chains have been recently presented [82, 83, 84]. Our approach is expected to be an improvement of the method of Ref. [82], because introduce a bias towards the end. Other collective modes involve concerted rotations of the bonds comprising the segments to be regrown [83, 84], and become quickly very complex as the number of monomer involved increases. In particular, we compare the MCB method to other MC methods currently in use. Along with the CB method [79, 80], we discuss the reptation method (RMC) [87], the pivot method (PMC) [88, 89] and the traditional Metropolis method (MMC) [48]. We show that MCB is an improvement as compared to these methods, especially when it becomes important to move sections of polymer chains that do not contain free chain-ends. It turns out that, in the simple cases studied, the CB method is the only competitive method among those considered. Both the reptation move and the Metropolis move involve only one monomer per step so they do not provide relevant changes in the polymer conformation. On the other hand the pivot method, which is extremely well suited for the calculation of some high temperature scaling properties of single non interacting chain [89], fails when interactions are considered and for dense systems. A pivot move consists in a rotation around single bonds. As a consequence, even for small rotations, distant opposite parts of the chains undergo a large displacement and this results almost unavoidably in a hard core overlap with some monomers not involved in the move. As a further observation, the CB method is not expected to perform well in glassy phases which are characterized by a freezing of the overall mobility of the polymer chains. More in general, in high molecular weight polymer systems, one assumes that configurational rearrangements typically would involve small motions of localized sections of the chains.

We present the MCB algorithm in the case of a single chain with fixed bond length l.

The extension to a multi-chain system in solution is straightforward. A remark is in order about the choice of fixed bond length. Polypeptide chains are characterized by essentially two energy scales. The first scale, of the order of ~ 5 eV, is that of the covalent bonds. The second scale is much smaller, of about ~ 0.1 eV, and is that of all the other interactions, such as the interaction of the amino acids with the solvent molecules, the off–neighbor interactions of amino acids either of the same chain or of other chains, and so on. As a result, at room temperature, the covalent bonds are essentially stable, and cannot be broken neither through thermal fluctuations nor through other interactions. In turn, the subtle interplay of the other interactions within the three dimensional folded structure of the chain and the environment determines the biological functionality of the polypeptide.

5.2 Outline of the MCB algorithm

We address the problem of calculating averages in a given sample space according to some distribution function. In a MC study, a sequence of correlated samples is generated from a Markov process whose unique equilibrium distribution corresponds to the desired one. A move is defined as the transition from a sample to the successive one. Depending upon the system under study, several different kinds of moves are possible, and it is desirable to find those moves that are most effective in reducing the correlations between successive samples. The autocorrelation time τ of an observable is a measure of the number of moves needed to obtain two substantially different samples. In addition, a crucial point is the minimization of the mean computer time t_m needed to perform a move. This time t_m can be regarded as the computational complexity of an algorithm. The overall efficiency of a MC algorithm can be roughly estimated by looking at the quantity τt_m [89].

In the MCB method a move consists in the replacement of an internal portion of the polymer chain. The total number of beads remains fixed during the simulation. The endpoints of the chain are also kept fixed. This last restriction can be removed by using the MCB method in combination with other methods.

A polymer is represented by a chain of N beads. A configuration b of the chain is defined by the positions $\mathbf{r}_1, \ldots, \mathbf{r}_N$ of the beads. The Boltzmann weight of configuration b is

$$\pi_b = \frac{1}{Z} e^{-\beta U_b}, \tag{5.1}$$

where Z is the chain partition function and U_b is the chain energy. A simple choice for the inter-particle interaction is the Lennard–Jones potential [86, 4]

$$V_{ij} = 4\varepsilon[(\frac{\sigma}{r_{ij}})^{12} - (\frac{\sigma}{r_{ij}})^{6}], \tag{5.2}$$

where the $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ are the inter-particle distances. We define $u_i = \sum_{j \neq i} V_{ij}$ as the interaction energy of monomer i with all the other monomers. More complex pair-wise

interactions can be also used. A typical form of the potential, currently used in MD or MC simulations of polypetides has the form [90]

$$\begin{aligned}\mathcal{H} &= \sum_{\text{bonds}} K_b(b-b_0)^2 + \sum_{\text{angles}} k_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} k_\phi(1+\cos(n\phi-\delta)) \\ &+ \sum_{\text{impropers}} k_\nu(\nu-\nu_0)^2 + \sum_{i<j}{}' 4\varepsilon_{ij}\left\{\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right\} + \sum_{i<j}{}' \frac{332}{\varepsilon_r}\frac{q_i q_j}{r_{ij}}, \quad (5.3)\end{aligned}$$

where distances are in Angströms, angles in radians and energy in kcal/mol. The first four term represent bonded terms, respectively referred to valence bonds, angles, dihedrals and improper dihedrals. The last two term represent respectively the Van der Waals and the Coulomb energy (with partial charges, and dielectric constant ε_r), and the apex over summation means sum over non bonded terms. The parameters entering in Eq. (5.3) can be found in literature [91, 92]. They are obtained semi-empirically, and adjusted to fit the structure organic molecules and their vibrational and rotational frequencies. As their ad hoc origin, there is no guarantee that the native state of the protein is related to minima (global or local) of the energy in Eq. (5.3).

The correct sampling of the distribution π_b is assured if the detailed balance condition

$$\pi_a P_{ab} = \pi_b P_{ba} \quad (5.4)$$

is fulfilled. P_{ab} is the transition probability from configuration a to configuration b. In the Metropolis, prescription P_{ab} is split into the product of two terms

$$P_{ab} = A_{ab} T_{ab} \quad (5.5)$$

where T_{ab} is a proposal transition matrix, and A_{ab} is an acceptance matrix to be chosen so as to satisfy detailed balance. A well known choice for A_{ab} is

$$A_{ab} = \min(1, \frac{T_{ba}\pi_b}{T_{ab}\pi_a}). \quad (5.6)$$

Our choice for the proposal transition matrix T_{ab} is based on the extension to the off–lattice case of the regrowth procedure of Dijkstra *et al.* [81].

We give a description of this procedure:

1. Select at random two internal sites i_0 and i_1 of the chain (with $i_1 > i_0$) and remove the $n = i_1 - i_0 - 1$ monomers between them.

2. Regrow this subsection of n beads one monomer at a time. The position of the first new monomer (i=1) is chosen with a probability p_j to be specified below ($j = 1, ..., k$) among k trials randomly generated on a sphere of radius l centered on site i_0. After the first monomer has been put into place, the second monomer (i=2) is chosen

among k trial positions centered on the first monomer. This scheme is repeated up to monomer $i = n - 1$. The last monomer is chosen among k trials generated by crank–shaft moves [93].

3. The probability p_j of each trial j ($j = 1, ..., k$) of monomer i is given by the product of two factors: the Boltzmann weight of the inserted monomer $e^{-\beta u_j}$, and a factor that biases the overall walk to reach its fixed target, that is the site i_1. This bias factor is given by the probability $P(\mathbf{R}_j)$ of a random walk connecting the jth trial position of monomer i, $\mathbf{r}_i^{(j)}$, to the final point $\mathbf{r}_{i_1}$ (with $\mathbf{r}_{i_1} - \mathbf{r}_i^{(j)} = \mathbf{R}_j$) in $n - i + 1$ steps of fixed length l. This $P(\mathbf{R}_j)$ can be calculated analytically, (see below). Then $p_j = P(\mathbf{R}_j)e^{-\beta u_j}/Z_{\{\mathbf{R}\}_i}$ where $Z_{\{\mathbf{R}\}_i} = \sum_{j=1}^{k} P(\mathbf{R}_j)e^{-\beta u_j}$, The notation $\{\mathbf{R}\}_i$ represents the set of k trials $\mathbf{R}_j$ generated to select the i-th monomer.

4. Quite generally, if a change of coordinate system is performed, then a Jacobian factor must be included [82] in evaluating the probabilities appearing in Eq. (5.6), Such caution can be avoided since each trial position j for the new monomer i is generated with uniform distribution on a sphere of radius l centered on monomer $i - 1$. In other words our Jacobian is 1. A special case is the last monomer which is positioned by a crank–shaft move. This requires the introduction of a suitable Jacobian factor, $J = 1/(l^2 d)$, where d is the distance between monomer $n - 1$ and site i_1, as discussed in Ref. [82].

Here an important remark on detailed balance is in order. In the detailed balance condition, Eq. (5.4), P_{ab} is the transition probability from configuration a to configuration b. This requirement is not suited to the off–lattice case since it is possible to generate configurations out of only a particular finite ensemble of trials[80]. Following Ref. [80], we use the notation $\{b\}$ to represent the ensemble of all the sets $\{\mathbf{R}\}_i$ ($i = 1, n$) employed to give rise to configuration b. To recover the correct P_{ab} we would have to sum over all the possible sets $\{b\}$ giving rise to configuration b. In the continuum case this sum is clearly impossible to carry over in a finite time. We introduce the super-detailed balance condition [80]

$$\pi_a P_{ab}\{a\}\{b\} = \pi_b P_{ba}\{a\}\{b\}, \tag{5.7}$$

where $P_{ab}\{a\}\{b\}$ is the probability of generating configuration b out of the trial positions $\{b\}$ and configuration a, out of the trial positions $\{a\}$. Writing down the sums over all the admitted sets $\{a\}$ and $\{b\}$ it is easily seen that the super-detailed balance condition implies the detailed balance one.

In order to apply the super-detailed balance condition, we introduce the proposal transition probability

$$T_{ab}\{a\}\{b\} = \prod_{i=1}^{n} P_{\{\mathbf{R}\}_{i_b}} P_{\{\mathbf{R}\}_{i_{a'}}} \frac{P(\mathbf{R}_{i_b})e^{-\beta u_{i_b}}}{Z_{\{\mathbf{R}\}_{i_b}}}, \tag{5.8}$$

which is the probability of generating configuration b out of the set $\{b\}$ and of generating a similar set $\{a'\}$ of $n(k-1)$ positions, in addition to the already existing positions, relative to the configuration a. Here $P_{\{\mathbf{R}\}_{i_b}}$ is the probability of generating the set $\{\mathbf{R}\}_{i_b}$ of trial bonds for the configuration b. We impose then Eq. (5.7), by choosing $P_{ab}\{a\}\{b\} = A_{ab}T_{ab}\{a\}\{b\}$. Introducing

$$G_b = \prod_{i=1}^{n} P(\mathbf{R}_{i_b}), \tag{5.9}$$

and

$$W_b = \prod_{i=1}^{n} \frac{Z_{\{\mathbf{R}\}_i}}{k}, \tag{5.10}$$

we finally get the acceptance matrix[1]

$$A_{ab} = \min(1, \frac{J_b W_b G_b^{-1}}{J_a W_a G_a^{-1}}). \tag{5.11}$$

The crucial point in our method is that $P(\mathbf{R})$ can be given analytically. Consider the polymer as a random walk in the continuum 3D space. Let $p(\mathbf{r}_1)$ be the probability of finding a walker in $\mathbf{r}_1$. Let $p^{(1)}(\mathbf{r}_2, \mathbf{r}_1)$ be the conditional probability of finding the walker in $\mathbf{r}_2$ at the successive step. In our case $p^{(1)}(\mathbf{r}_2, \mathbf{r}_1) = \delta(|\mathbf{r}_2 - \mathbf{r}_1| - l)/4\pi l^2$. Using the Chapman-Kolmogorov identity [94], we obtain the probability of going from $\mathbf{r}_1$ to $\mathbf{r}_3$ in two steps, as

$$p^{(2)}(\mathbf{r}_3, \mathbf{r}_1) = \int d^3 r_2 p^{(1)}(\mathbf{r}_3, \mathbf{r}_2) p^{(1)}(\mathbf{r}_2, \mathbf{r}_1). \tag{5.12}$$

Similarly, the probability of going from r_1 to r_{n+1} in n steps is

$$p^{(n)}(\mathbf{r}_{n+1}, \mathbf{r}_1) = \int \prod_{i=2,n} d^3 r_i \, p^{(1)}(\mathbf{r}_{n+1}, \mathbf{r}_n) \cdots p^{(1)}(\mathbf{r}_2, \mathbf{r}_1). \tag{5.13}$$

Taking $\mathbf{r}_{n+1} - \mathbf{r}_1 = \mathbf{R}$, we have

$$P(\mathbf{R}) = \int d^3 r_{n+1} \delta(\mathbf{r}_{n+1} - \mathbf{r}_1 - \mathbf{R}) p^{(n)}(\mathbf{r}_{n+1}, \mathbf{r}_1), \tag{5.14}$$

which is independent of $\mathbf{r}_1$. In order to calculate the integral in Eq. (5.14) we change variables to $\mathbf{y}_i = \mathbf{r}_{i+1} - \mathbf{r}_i$,

$$P(\mathbf{R}) = \frac{1}{(4\pi l^2)^n} \int \prod_{i=1,n} d^3 y_i \, \delta(\sum_{i=1,n} \mathbf{y}_i - \mathbf{R}) \prod_{i=1,n} \delta(|\mathbf{y}_i| - l). \tag{5.15}$$

Using the integral representation of the delta function, we finally obtain

$$\begin{aligned} P(\mathbf{R}) &= \frac{1}{(4\pi l^2)^n} \frac{1}{(2\pi)^3} \int d^3 k e^{-i\mathbf{k}\cdot\mathbf{R}} [\int d^3 y e^{i\mathbf{k}\cdot\mathbf{y}} \delta(|\mathbf{y}| - l)]^n \\ &= \frac{1}{2\pi^2 l^n R} \int_0^\infty dk \sin(kR) \frac{\sin^n(kl)}{k^{n-1}} \end{aligned} \tag{5.16}$$

[1] As opposed to usual crank–shaft moves, where the use of the Jacobian can be avoided, since $J_a = J_b$ always, here, due to rule (4) of the construction procedure, it must be included.

that can be easily evaluated through Fourier sine transforms.

The last monomer is special, since it requires no calculations due to the rigid bond length constraint. In this case we can trivially put $P(\mathbf{R}_j) = 1$ for every trial.

5.3 Applications to Simple Models

In this section we show that the MCB algorithm can be used efficiently in the study of simple off-lattice polymer system. We focus our attention to three well known models adopted in protein studies, a homopolymer chain with hamiltonian given by Eq. (5.17), a random heteropolymer chain (Eq. (5.24)), and a copolymer chain (Eq. (5.25)). The extension to more complex polymer systems is straightforward.

As a first test, we present the case of a single homopolymer chain with fixed bond length between successive beads [85, 86]. The model is defined by the hamiltonian

$$\mathcal{H}_o = \sum_{i=1}^{N} \sum_{j>i} V_{ij} \tag{5.17}$$

where V_{ij} is the Lennard–Jones potential similar to that of Eq. (5.2),

$$V_{ij} = \varepsilon[\frac{R}{r_{ij}^{12}} - \frac{A}{r_{ij}^{6}}], \tag{5.18}$$

with R and A as two adjustable parameters. For $R = 1$ and $A = 2$ the model is well characterized [85, 86]. The Lennard–Jones potential is a very common choice for the interaction between non–charged groups of atoms and molecules, with essentially Van der Waals type interactions [95]. At high temperatures, entropy dominates and the chain is in a swollen coil state. As a typical linear dimension, the average end-to-end distance R_{ee} is usually considered. R_{ee} scales with N as

$$R_{ee} \propto N^{\nu}. \tag{5.19}$$

Equivalently, if one takes the average gyration radius R_g one has

$$R_g = \sqrt{\sum_{i<j} \left(\frac{\mathbf{r}_i - \mathbf{r}_j}{N}\right)^2} \propto N^{\nu}. \tag{5.20}$$

In both cases $\nu \simeq 0.59$ in 3 dimensions [86]. Lowering the temperature the attractive energy gains more and more importance and at the θ point ($T \simeq 3.71\varepsilon$ [86]) a transition takes place to a compact globular state. Below the θ point the scaling regime holds with $\nu = 1/3$.

As discussed in the preceding section, it is instructive from the technical point of view, to consider the computational cost $C = \tau t_m$, where t_m is the CPU time per move, and τ is

the integrated autocorrelation time at equilibrium. If $\rho(t)$ is the normalized autocorrelation function of a quantity $A(t)$, measured at time-step t during the simulation,

$$\rho(t) = \frac{\langle A(s)A(s+t)\rangle - \langle A\rangle^2}{\langle A^2\rangle - \langle A\rangle^2}, \tag{5.21}$$

then

$$\tau = \frac{1}{2}\sum_{t=-\infty}^{\infty} \rho(t). \tag{5.22}$$

At large separations in time, $\rho(t)$ is dominated by statistical noise. As shown in Ref. [96], Eq. (5.22) should be truncated to a certain window $[1, t_o]$. One looks then for the convergence of

$$\tau(t_o) = \frac{1}{2} + \sum_{t=1}^{t_o} \rho(t) + r(t_o), \tag{5.23}$$

with respect to the truncation time t_o, with the remainder $r(t_o)$ evaluated from the exponential decay of $\rho(t)$. In the following we will take the gyration radius R_g as our observable $A(t)$.

In the homopolymer chain case we compare the computational efficiency of the MCB algorithm with other commonly used MC methods, namely the reptation method [87], the Metropolis method [48], the pivot method [88, 89] and the CB method [79, 80]. The plot of R_g as a function of the number of MC steps, showing the convergence of the results for the various methods is shown in Fig. 5.1. In Fig. 5.2 we show the computational efficiency C (referred to R_g), as a function of the degree of polymerization N, for two different temperatures: (a) At $T = 10\varepsilon$, where the polymer is well above the θ point and in a swollen coil state, and (b) at $T = 3\varepsilon$, where the polymer is in a compact globular state. The following observations apply to both cases (a) and (b). The reptation method and the Metropolis method become computationally very costly when N grows over 30. As for the pivot method we note that, as expected, this scheme becomes very inefficient in the globular phase (case (b)) for long chains ($N > 50$). We observe that the MCB method is defined at fixed chain extrema. In order to make a comparison with the CB method, in which the chain ends are moved, we allowed for a fraction of CB moves (typically from 10% to 50%) in the MCB simulations. At both temperatures it is possible to see that the MCB method improves by an appreciable factor over the CB method.

The second system studied is the random heteropolymer chain of Ref. [4]. The hamiltonian is

$$\mathcal{H} = \mathcal{H}_o + \sum_{i=1}^{N}\sum_{j>i} \frac{\eta_{ij}}{r_{ij}^6}, \tag{5.24}$$

where $\mathcal{H}_o$ is given by Eq. (5.17). Here η is a random variable with zero mean $\langle\eta_{ij}\rangle$=0 and a correlation of the form $\langle\eta_{ij}\eta_{kl}\rangle = \eta_o\delta_{(ij)(kl)}$. The random interaction term is introduced to represent many interplaying different factors, typically the complex interactions between

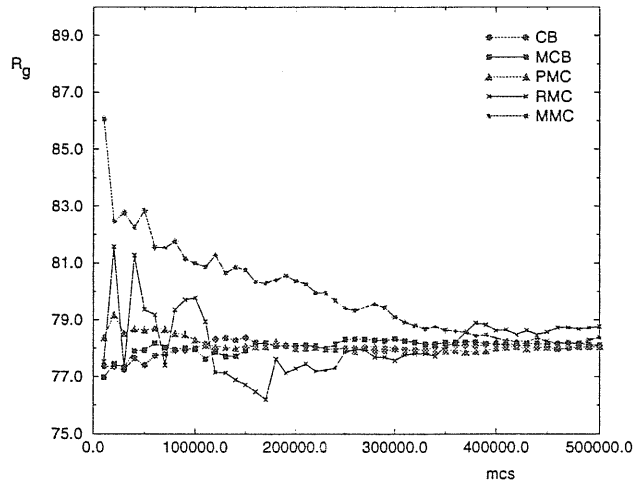

Figure 5.1: Convergence for the results on the average gyration radius R_g for the methods considered, as a function of the MC steps (mcs).

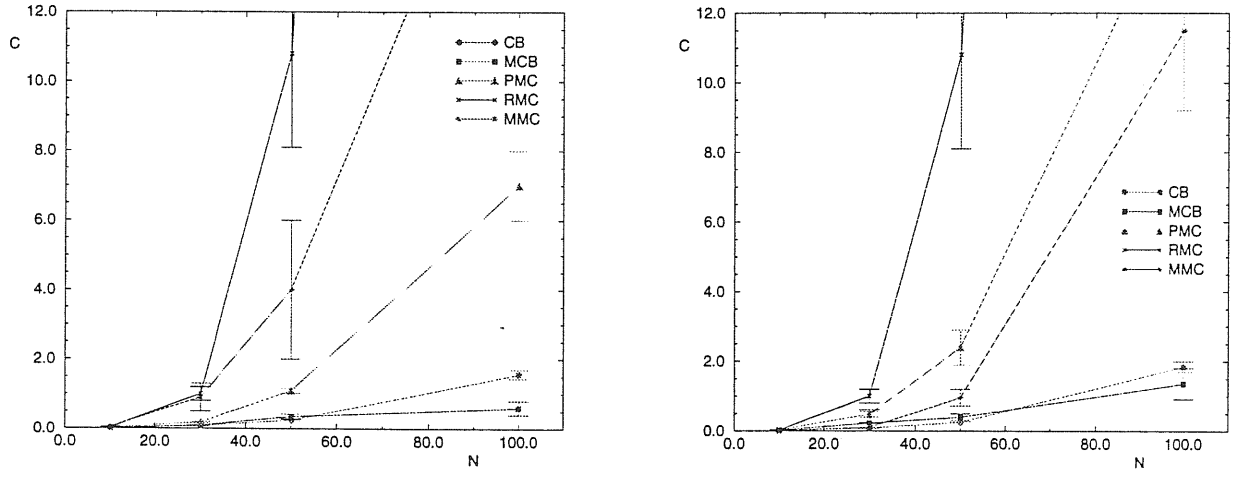

Figure 5.2: Computational efficiency $C = \tau t_m$ of the MC methods considered in the homopolymer case. τ is the gyration radius autocorrelation time, and t_m is the mean cpu time per move. Two cases are shown: (a) $T = 10\varepsilon$, where the polymer is in a swollen coil state, and (b) $T = 3\varepsilon$, where the polymer is in a globular state. The θ point is at $T \simeq 3.71\varepsilon$.

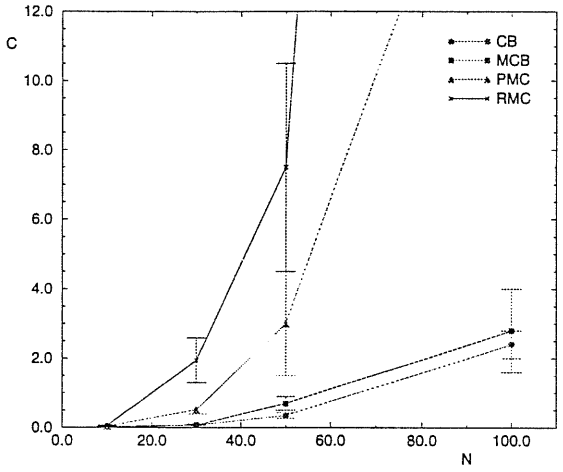

Figure 5.3: Computational efficiency of the MC methods considered in the case of a heteropolymer. The temperature is $T = 5\varepsilon$, slightly below the θ point.

different groups of amino acids and the effect of the solvent. Since we deal with fixed bond lengths we have dropped a term of the type $\sum_i K(\mathbf{r}_i - \mathbf{r}_{i+1})^2$ considered in Ref. [4]. Such an harmonic potential is meant to represent the strong peptide bond between successive amino acids along the chain. (A more realistic interaction term would not be invariant for rotation around the bonding axis).

We fixed the strength of the random potential to $\eta_o = 3$ and the temperature of the system at $T = 5\varepsilon$, which is slightly below our rough estimate of the collapse temperature θ obtained through the analysis of the scaling exponent ν in Eq. (5.19). In Fig. 5.3 we show the behavior of the computational cost C as a function of N. As discussed in the introduction, MC simulations in the collapsed phase are hindered by the glassy behavior of the polymer. Our study indicates that local MC schemes, such as the reptation or the Metropolis method fail to a large extent to sample the phase space, unless prohibitively long runs are performed. CB and MCB methods give better performances, comparable within the statistical errors, since they are able to redraw large segments of the chain, thus allowing appropriate conformational rearrangements.

The third model considered is a random copolymer chain. According to their affinity to water the 20 species of amino acids can be grouped as hydrophobic and polar [97]. Within this classification scheme only two kinds of amino acids are considered, namely hydrophobic (H) and polar (P) amino acids. The potential $V_{i,j}$ in Eq. (5.17) is

$$V_{ij} = \varepsilon[\frac{R(\sigma_i, \sigma_j)}{r_{ij}^{12}} - \frac{A(\sigma_i, \sigma_j)}{r_{ij}^{6}}], \tag{5.25}$$

where R and A are now 2×2 matrices, whose energies are referred to the solvated states

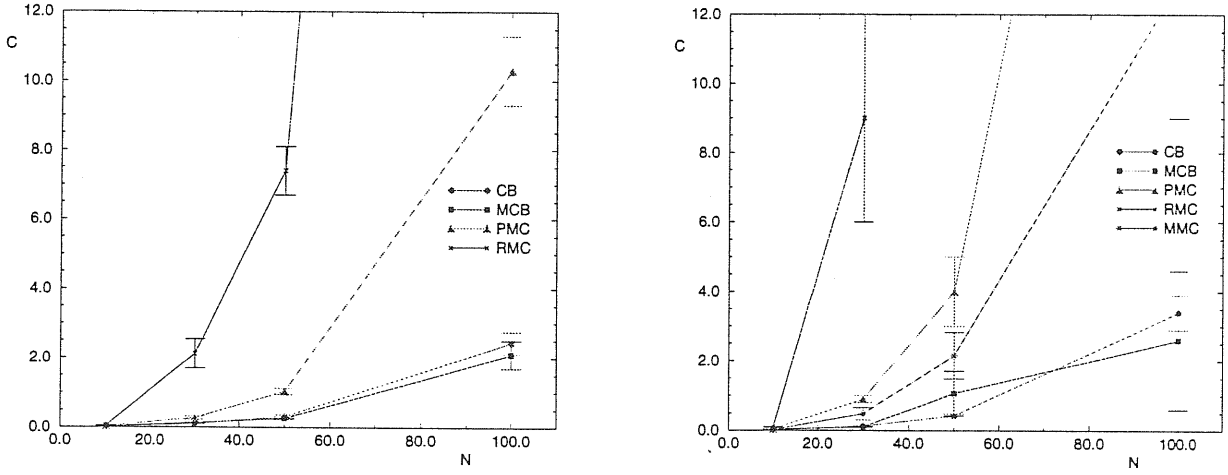

Figure 5.4: Computational efficiency of the MC methods considered in the case of a copolymer. Two cases are shown: a) $T = 10\varepsilon$. b) $T = 1.3\varepsilon$ which the θ point estimated from the scaling behavior of R_{ee}.

of the monomers. The label σ of monomer i can be either H or P and a fraction f of monomers is assigned with an H label. In the minimal parameterization, $A(H,H)$ is set to some positive value and all the other elements of A are set to 0. This assignment captures the bare fact that, in proteins, collapsed states with H monomers buried inside are preferred, and it is believed that hydrophobicity is the major driving force for protein collapse [17]. Non polar amino acids experience an effective attraction in water solution, so that the polypeptide chain has the tendency to form a hydrophobic core.

Other choices are possible for R and A. In microphase separation studies, one requires segregation of different monomers. In the Lennard–Jones potential of Eq. (5.25), the interaction between monomers of different character is then taken to be purely repulsive, i.e. $A(H,P) = 0$, whereas the interaction between monomers of the same type is taken to be attractive. The phase diagram of a copolymer model has been predicted in Ref. [98] using the replica technique.

As in the heteropolymer case, we give a rough estimate of the θ temperature by a scaling analysis giving the exponent ν, finding $\theta \sim 1.4\varepsilon$. We analyze the performance of the MCB method both above and below the θ point. In Fig. 5.4 we show the computational cost C vs N at a) $T = 10\varepsilon$ and at b) $T = 1.3\varepsilon$. Considerations entirely similar to those made for the previous two models apply to the present case.

5.4 Conclusions

The main purpose of this chapter has been to present a new algorithm, the MCB method, for the simulation of realistic off-lattice polymer systems.

The MCB method is based on the CB method [80] and extends it allowing to regrow inner chain segments. In the original CB method, end segments of the chain are redrawn by imposing a bias towards energetically favorable conformations. With respect to another extension to the CB method, proposed by Escobedo and de Pablo [82], we introduced an additional probabilistic bias towards the fixed end of the segment to be regrown. In this respect, our scheme is much similar to the one presented by Dijkstra, Frenkel and Hansen [81] for polymers on a lattice. The concerted–rotation method, recently proposed by Dodd, Boone and Theodorou [83], is also aimed at improving the efficiency of off-lattice MC simulations of polymer systems by introducing substantial rearrangements of internal segments of the chain, lacking however in both the energetic and probabilistic biases. Elementary moves consist in coordinated rotations of adjacent torsion angles along the chain that leave all bond angles and bond lengths unaffected. Another possible extension of the CB scheme, joining it to the bisection method of Ceperley *et al.* [99] is currently under study [?].

We showed that the MCB method satisfies the detailed balance condition and that its performance is very good when compared to traditional MC methods.

As an application, we considered the simulation of three models commonly used in protein studies. The first case considered has been the homopolymer model [85, 86], with a purely deterministic Lennard–Jones potential between different monomers. We presented results also on two other simple models more directly related to proteins, namely the random copolymer model [97, 17], and the random heteropolymer model, introduced by Iori, Marinari and Parisi [4].

We gave evidence that, as it could be expected from previous studies on the CB method, the MCB method is more suited for the study of complex polymer systems than more traditional MC methods as the reptation method [87], the pivot method [88, 89] or the traditional Metropolis method [48]. In all the cases considered we showed that the MCB method perform as well as, or better, than the CB method.

The discussion presented in this chapter has been referred to a single chain problem, however, the extension to more complex systems is straightforward. Cross linked polymer structures, branched chains and polymers in constrained environments appear to be amenable to simulations with the MCB method. Moreover, from a more biologically motivated point of view, it would be interesting to consider more accurate potentials, such as Eq. (5.3), for protein folding within our scheme.

6 Perspectives

In this final chapter we address a list of issues that are currently under investigation. Although very few results are presented, we believe that the topics discussed are relevant to an understanding of the major problems in the study of protein folding.

Ideally, the study of a physical system would start from the knowledge of the ground state. Are we able to determine such state in proteins? Although X-rays crystallography and NMR spectroscopy have offered the structures of hundreds of proteins, from the theoretical point of view the answer is generally negative.

Two complementary approaches are undertaken. In the first approach, dealt with in the first section of this chapter, the aim is the determination of suitable interaction potential between amino acids. Realistic numerical calculation would then yield the ground state. This point is of overwhelming importance since, even in the case that the general principles of protein folding would be completely clarified, no effective prediction could be made unless realistic force field would be known. In the second approach, presented in the second section, we address the problem of finding a reliable algorithm to fold a protein. Despite strenuous efforts in the last thirty years, this procedure has escaped a formulation. Monte Carlo algorithms are appealing, since their dynamics can be completely unphysical and the only requirement is to sample efficiently compact conformations. In the third section we present a method that could in principle provide a way to navigate in the conformation space of a disordered system as a protein and select the ground state.

Do we have any understanding of the physical principles determining the folding of proteins into their native state? Also in this case the answer is negative. From simple models, we have seen that hydrophobic interaction would possibly play an important role. We have also learn that random heteropolymers are a good approximation to real proteins, if a selection procedure of sequences is performed. This lead us to the formulations of the two complementary concepts of foldability and designability. A different point of view, discussed in the fourth section, is to investigate how long range correlation can be the signature of the folding code buried in the protein sequence and encoding the three dimensional structure.

6.1 Effective Interaction Potential between Amino Acids

A major problem in protein science is that although structures are experimentally known relatively accurately (typically with a 2Å resolution in the position of H atoms in the residues [1]) the interaction stabilizing a particular conformation have escaped quantitative enumeration. The gap between the knowledge of the structure and the correct energetics has hampered the solution of the protein folding problem and the design of new enzymes. In the past thirty years two approaches have been pursued. The more rigorous approach would to derive from quantum mechanical calculations, or from spectroscopic experimental data, the forces between amino acids [100, 101]. Typical time-scale for all-atoms calculations is far too short to be used for the folding of a real protein, although some insight can be obtained as far as secondary structure formation is concerned. Much more amenable are coarse grained models of proteins where amino acids are represented in a simplified way, typically as interacting centers which may or may not have internal degrees of freedom. The interaction between such entities has been traditionally derived from statistical analysis of protein structure databases, as for example in the seminal work of Miyazawa and Jernigan [32, 33]. Invoking a Boltzmann distribution, the energy e_{ij} of a contact between amino acids i and j is assumed to be proportional to the logarithm of the relative frequency f_{ij} of its appearance in the database

$$e_{ij} \propto \log f_{ij}. \tag{6.1}$$

In a recent important work, Thomas and Dill (TD) [102, 103] have employed exact lattice models to rigorously test the assumptions and approximations of these traditional approaches and have identified their weaknesses. In particular, they noted that the the usual approach neglect the effect of excluded volume, chain connectivity and amino acid sequence in the calculation of the relative frequencies.

More recently a new approach has formulated the determination of effective potentials as an optimization problem [78, 104]. Since it is believed that the folded state is the global energy minimum or at least a stable one, the problem is to find a set of parameters that define the potential in such a way that the folded state is recovered among a large ensemble of alternative conformations. Such an approach has been pioneered by Maiorov and Crippen [78] and more recently developed by Mirny and Shakhnovich [104]. The scheme is the following

- choose a given parameterization of the potential. usually a pairwise additive potential is chosen (210 parameters if the 20 species of amino acids are considered).
- select from a database of M sequences S_i $(i = 1, \ldots, M)$ and the correspondent crystal structures Γ_i^* to be used as alternatives.
- define a cost function that attains its minimum in parameters space when all the

sequences S_i are simultaneously as close as possible to their true folded state Γ_i^*.

- adopt the corresponding set of parameters as the results of the procedure.

For a given parameterization, the energy of a sequence S_i on a conformation Γ is

$$\mathcal{H} = \mathcal{H}(S_i, \Gamma, \{\alpha\}), \tag{6.2}$$

where $\{\alpha\}$ is the set of parameters characterizing the potential. The Z_i score [43] (see chapter 2) is a measure of the relative thermodynamic stability of the native state Γ_i^* of the sequence S_i. The more negative the Z_i score is, the more stable is the native state of the sequence S_i. Ideally, for the set $\{\alpha\}_{\text{true}}$ of parameters corresponding to the true potential, all the sequences S_i have their ground states respectively on Γ_i^*. A change in the set $\{\alpha\}_{\text{true}}$ would almost inevitably result in a destabilization of the ground state with an increase of the Z_i score. As cost function to be minimized, Mirny and Shakhnovich introduced the harmonic mean of the Z score over the M sequences S_i

$$\langle Z \rangle = \frac{M}{\sum_i^M 1/Z_i} \tag{6.3}$$

The best set $\{\alpha\}$ is the one that stabilizes simultaneously all the sequences S_i in their ground state Γ_i^*. Technically however, to compute $\langle Z \rangle$, they employed an ideal gas approximation that is probably the reason of their low rate of success.

In the same spirit, the work of Ref. [105] present the problem of determining effective interaction potentials as an optimization problem. Let us consider M sequences denoted by $\{S\} = \{S_1, S_2, \ldots, S_M\}$, each made up of N amino acids. Each sequence S_i is postulated to have a unique native state in a conformation Γ_i^* that is known experimentally or otherwise. The corresponding set of native conformations is denoted by $\{\Gamma\} = \{\Gamma_1^*, \Gamma_2^*, \ldots, \Gamma_M^*\}$. The questions addressed are: what are the potential energies of interaction between the amino acids that are consistent with the above data? Can one predict the native structures of other sequences using these derived interaction energies?

For a given set $\{\alpha\}$ of parameters consider, for each sequence $S_i \in \{S\}$, the energies $\mathcal{H}_{il}$ in each of the M conformations $\Gamma_l \in \{\Gamma\}$. For the correct set $\{\alpha\}_{\text{true}}$ of parameters,

$$\mathcal{H}_{ii} = \min_{l=1,m} \{\mathcal{H}_{il}\}, \tag{6.4}$$

because Γ_i^* is the ground state of the sequence S_i. Two versions of a cost function are defined which, when minimized, ensure this requirement. The cost function is not unique and can be tailored to satisfy the physical constraints. Let $F_i = \mathcal{H}_{ii} - \min_{l=1,M}\{\mathcal{H}_{il}\}$; $\mathcal{H}_{if} = \min_{l=1,M;\ l\neq i}\{\mathcal{H}_{il}\}$;, $g_i = \mathcal{H}_{ii} - \mathcal{H}_{if}$ and $\bar{g} = \sum_{i=1}^M g_i$. Define $C_i = F_i$, if F_i is positive. Otherwise $C_i = g_i$. The first of the two cost functions is

$$C_{\text{tot}} = \sum_{i=1}^{M} C_i \,. \tag{6.5}$$

The cost function ensures that the native state of each of the M sequences is in the correct conformation and that the gap between the native state energy among the other conformations is high. Such a gap ought to be large for good folders [5, 12].

As a test case, they study the 2D HP model, where the true interaction energies are independently known. The HP model [22] captures the dominance of the hydrophobic and polar amino acids in determining the native state of proteins in a solvent. The contact potential is characterized by three parameters $B(H,H), B(H,P)$ and $B(P,P)$ denoting the interactions energies for HH, PP and HP contacts. Two monomers are said to be in contact if they are nearest neighbors but not next to each other along the chain. For short chains, containing N monomers, one can carry out a complete enumeration of all 2^N sequences and all conformations to determine all the native states. Our test begins with a choice of the "true" interaction parameters $B(H,H)$, $B(H,P)$ and $B(P,P)$. First, the complete set of $\widetilde{M}$ sequences which have unique ground states are found. Typically, the total number of distinct ground state conformations is a number M which is smaller than $\widetilde{M}$ because several sequences may have the same ground state conformation. As a data bank, select M of the $\widetilde{M}$ sequences are selected, so that each has a distinct ground state. This constitutes the set $\{S\}$ and the M conformations the set $\{\Gamma\}$. Using the two sets as input, they determined the values of the interaction energies using the optimization schemes. These interaction energies are then used to predict the ground states of all the $\widetilde{M}$ sequences within the space of all possible conformations. An alternative cost function is defined to be $C'_{\rm tot} = C_{\rm tot}$ if at least one of the $F'_i s$ is positive. Otherwise

$$C'_{\rm tot} = \sqrt{\frac{\sum_{i=1}^{M}(g_i - \bar{g})^2}{M}} \tag{6.6}$$

Results for several cases are summarized in Table I along with the TD results of tests of the Miyazawa-Jernigan [32, 33] scheme. Since the energy scale is arbitrary the value of $B(H,H)$ is set and the other two parameters are determined. While the ground states of all M sequences comprising the set $\{S\}$ are correctly predicted by both optimization schemes, only the second scheme yields a 100 % success rate for all $\widetilde{M}$ sequences. This suggests that the gap is indeed roughly comparable for all sequences. In order to assess the robustness of the approach, next nearest neighbor HP and PP interactions are allowed for the case with $B(H,H) = -5$, $B(H,P) = -4$ and $B(P,P) = -1$. Our results for the nearest neighbor interactions are unchanged and the derived values of the next nearest neighbor interactions are found to be less than 0.001 in magnitude. In all cases, a 100% success rate is found for ground state recognition independent of the specific values of the derived potentials.

Our aim is the application of this scheme to real proteins using data from Protein Data Bank (PDB), and to deduce a new set $\{\alpha\}$ of parameters for a pairwise additive short range potential. The computation scheme does not differ much from the one used in the test case, except that here we don't know the *true* set of parameters. In practice we

retrieve protein crystal structures from PDB. The correspondent sequences are of different lengths and threading is possible to enlarge the set of alternative conformations. Solving the optimization problem we would obtain the 210 parameters that minimize the cost function for the given set of structures and sequences. From chapter 4 we know that homologous sequences (those sequences descending from a common ancestor [1]) often have a similar 3D structure. It is convenient to consider a set of homologous sequences since they differ only for a set of contacts. In this case many parameters can be kept fixed in order to reduce the search size of the optimization problem.

Table I

True				TD test [102]			
$B(H,H)$	$B(H,P)$	$B(P.P)$		$B(H,H)$	$B(H,P)$	$B(P,P)$	PS
-5	-4	-1		-5	-3.0	$+0.8$	74
-5	-1	-2		-5	-1.1	-2.1	100
-5	-5	-1		-5	-3.7	$+1.4$	84
-5	-3	$+1$		-5	-2.6	$+2.5$	96
-5	-3	-1		-5	-2.4	0.0	64

1st Method				2nd Method			
$B(H,H)$	$B(H,P)$	$B(P.P)$	PS	$B(H,H)$	$B(H,P)$	$B(P,P)$	PS
-5	-3.75	0^-	100	-5	-4.03	-1.31	100
-5	0^-	-1.66	98	-5	-1.35	-2.31	100
-5	-7.5^+	0^-	98	-5	-4.61	-0.89	100
-5	-2.5^-	$+2.5^-$	100	-5	-2.97	1.21	100
-5	-2.50	0^-	100	-5	-2.87	-0.77	100

Table I. Summary of our results. PS represents the percentage of success in the prediction of native states of the $\widetilde{M}$ sequences having a unique native conformation within the space of all possible connected self-avoiding conformations . N is the chain length.

6.2 Topological Annealing Monte Carlo

Monte Carlo methods are known for their ability to perform large scale moves that ought not to be necessarily related to the true physical dynamics of the system, as for example in cluster algorithms for critical spin systems [106, 96, 107], or in the "mountain to valleys" algorithm for the study of the roughening transition in SOS models of surfaces [108]. In such models coherent large scale fluctuations are relevant in determining the dynamical behavior. Traditional simulation techniques, such as the Metropolis algorithm [48], perform incoherent small scale moves that are extremely ineffective in modifying correlated large scale structures. The underlying large scale physics has to be understood in order to devise

efficient updating schemes. The requirement is to "break" large scale fluctuations that are responsible for the slowing down of the dynamics. As in most frustrated systems, the nature of excitations in heteropolymers is extremely complicated and the resulting motion is embedded in a rugged energy landscape, with many almost degenerate ground states. and local metastable minima. Real proteins are complex systems containing thousands of atoms, however they are able to find their ground state on a time-scale of seconds. Although this time could appear to be fantastically short, it is extremely large as compared to microscopic motions which range from 10^{-15} for bond vibrations to 10^{-9} for hinge motions between secondary structures [29]. It can be argued then that folding is a *slow* process. The Monte Carlo technique most commonly adopted in the simulation of proteins [12, 25](see also chapter 2) relies on a local updating scheme, which is believed to reproduce the true dynamics of protein chains [12]. As such, it is probably very inefficient since microscopic motions *in machina* require much more time than for real systems. The key question is then if it is possible to envisage the nature of the large scale excitations and to devise a cluster technique able to accelerate the dynamics by breaking such mechanisms.

We address this problem by formulating a Monte Carlo optimization strategy to find the ground state conformation Γ^* of a given sequence S of amino acids. To find the ground state is a zero temperature problem and detailed balance has not to be required [109]. Technically, the ground state is found by performing a "random walk" in conformation space. The native state of a protein in the cell is known to be rather compact, with very few molecules of solvent inside the core of the folded structure [1, 17]. It would be important to restrict the search problem to the subset of maximally or nearly maximally compact conformations. It has been proposed that the native state Γ^* should be characterized by the maximal compatibility in the interactions giving rise to secondary and tertiary structures, (the "minimal frustration" principle [8]). Can we develop a scheme that embodies this insight? We start from a gas of N unconnected amino acids within the simple B_{ij} model in 2D (see Eq. (2.2)). The idea is to switch gradually on the constraint of chain connectivity. To each amino acid a label i (i=1,...,N) is attached. We introduce a set of $N-1$ "chain" potentials $V(r_{i,i+1})$, where $r_{i,i+1} = |\mathbf{r}_i - \mathbf{r}_{i+1}|$ is the relative distance between two successive amino acids and $\mathbf{r}_i$ is the position of amino acid i. The potential $V(r_{i,i+1})$ is typically of a Lennard-Jones type (see Eq. (5.2) and Fig 6.1) and acts only between amino acids i and $i+1$. The two amino acid are forced to migrate one towards the other and to form a bond. The self avoidance condition is built in, by letting $V(r_{i,i+1})$ be positive in $r = 0$ and it is gradually switched on as well. An annealing scheme [110] is set up to gradually switching on both the potential B_{ij} and the potential $V(r_{i,i+1})$. As a preliminary result, amino acid sequences are reconstructed with all the monomers correctly enchained.

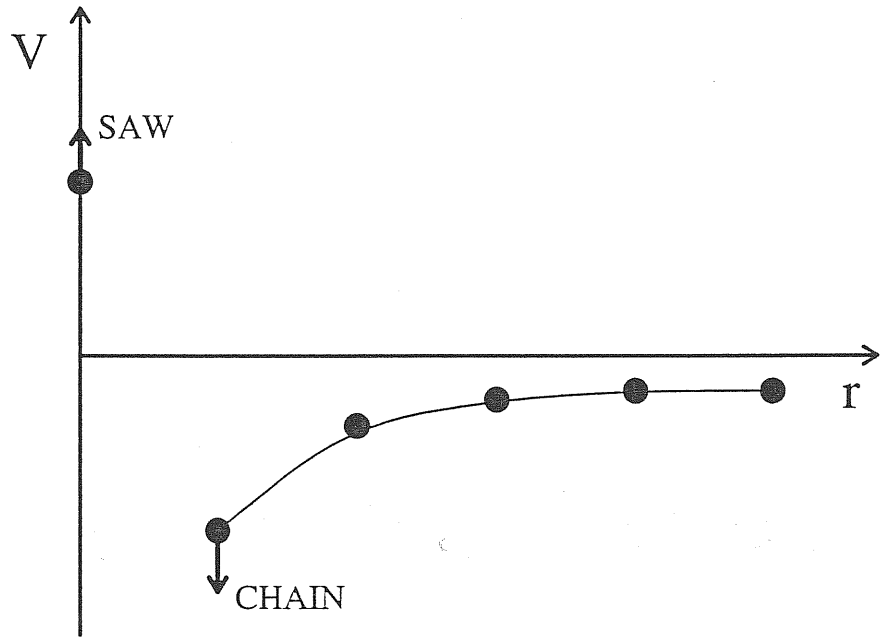

Figure 6.1: Chain potential $V(r)$.

6.3 Random Walk with Memory

Proteins are able to find their optimal folded configuration, because of a very specific organization of phase space, with a funnel-like free energy minimum [15, 11]. A protein can be seen as a disordered system such that the sequence of amino acids is quenched to form the polypeptide chain. Apart from spin glasses, which can be described as systems with a randomly "rugged free energy landscape", there are other systems in which the disorder has a peculiar organization. For example in the Hopfield model [111] the dynamics is such that stored information can be retrieved under particular conditions: the free energy minima are localized and well defined attraction domains exist. In this section we present a method that, starting from a process in quenched disorder, prescribes how to construct an annealed dynamics which yields the statistical properties of the original process. In principle we are able to construct a dynamics that retrieves a random realization of the disorder with any preassigned distribution. To describe the origin of specific structure of disorder that characterizes the phase space of a protein the minimal frustration principle [8] has been formulated. We think that this result can be relevant on protein folding in that we are able in principle to give an alternative description of a possible dynamics which can achieve some memory retrieval or "folding".

The dynamics of disordered systems is a very active subject of research of statistical physics. In non equilibrium systems, such as driven interface growth [112] and charge density waves [113], disorder leads to very interesting effects as depinning transitions, creep phe-

nomena and self organization. In out of equilibrium systems, like spin glasses, aging effects arise which, at least at a mean field level, has been related to the lack of time translational invariance and the failure of fluctuation dissipation relations [114]. The main complication brought by the presence of disorder is that, in order to compute a physical quantity, apart from the "dynamic" average over different stochastic time evolutions, quenched dynamics requires a second average over the realizations of disorder. This, operationally, implies that one has to evolve the system in several disorder configurations and at the end average the result over the realizations of disorder. On one hand, the dynamics explicitly depends on the particular realization of the disorder (typically through transition rates). On the other, in most systems, one expects the physical quantities to be self averaging and therefore to depend weakly on the disorder configuration. This situation is rather unsatisfactory, in our opinion, because only after this second average over disorder it is possible to appreciate the general features of the dynamics. It has recently been pointed out [115, 116] that this problem can be overcome in non equilibrium models based on extreme dynamics, by appealing to an annealed dynamics (we shall use this term as opposed to quenched dynamics) which does not make reference to a particular realization of disorder. The advantage of this point of view is that only the average over different stochastic time evolutions need to be taken: the effective dynamics is indeed such that the averages over disorder are taken "run time", i.e. at each time step, by the process itself. Moreover this approach provides also the statistical weight of the history of the process, which is hardly available in dynamics with disorder. The key point, in the derivation of such annealed dynamics, is that the future evolution has to be statistically consistent with the past history. The mathematical translation of this principle relies on the concept of conditional probability. The process thus acquire time dependences which naturally explain the emergence of memory effects in quenched dynamics. It has also been shown that, from this point of view, the relation between extremal dynamics and self organization are a simple consequence of a more general relation between dynamical processes with memory and self organization [117].

In this section we apply the same considerations to an equilibrium system. We shall deal with the simplest such system, i.e. a one dimensional random walk in random environment. For this we will derive the exact corresponding annealed dynamics. This dynamics, by definition, does not depend on any particular realization of the disorder. However, as we shall see, the process has the same statistical properties. Asymptotically, for large times, the process singles out a particular realization of the disorder, which is the only one which is consistent with the past history of the process. A simple generalization of the dynamics with memory we find, shows that, interestingly enough, the disordered dynamics lies on the border line between random dynamics and deterministic dynamics. The random walker, in the latter case will sooner or later localize on some site. Finally we shall generalize our arguments to the problem of a random walk with traps and draw some conclusions.

The random random walk (RRW) on a line is defined by assigning at each site $i = 0, \pm 1, \pm 2, \ldots$ a random variable $p_i \in [0,1]$ drawn from a distribution $P\{p \leq p_i < p + dp\} = \phi(p)dp$. The evolution of the position x_t of the RRW is defined by $x_{t+1} = x_t + 1$ with probability p_{x_t} and $x_{t+1} = x_t - 1$ otherwise. In spite of its simplicity this model has been studied by many authors as a toy model for localization [118], depinning transitions [119] and aging effects [120]. The most striking feature is that the diffusion is extremely slow: The typical size visited by the walker after a time t is $\delta x \sim (\ln t)^2$. Comparing this result, originally derived rigorously by Sinai [121], with the diffusion of a random walk without disorder, $\delta x \sim \sqrt{t}$, suggests that disorder has really dramatic effects on the dynamics.

In order to introduce our model, let us consider the case of a uniform distribution $\phi(p)$=1. Imagine to observe the walker in its motion, without knowing the realization $\{p_i\}$ of the disorder. The only information available is what one sees, namely the number $n_{i,t}$ of times that the random walker has visited site i and the number $k_{i,t}$ of times in which it has moved from site i to site $i+1$. As we shall now show, it is possible, using this information, to describe a RRW even if the values of p_i are not known. This is accomplished by observing that the probability that the number of right jumps $i \to i+1$ is k, given that site i has been visited n times and the transition probability is $p_i = p$, is simply given by the binomial distribution

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{6.7}$$

where the notation $P(A|B)$ stands for the probability of the event A, conditional to the occurrence of B. Regarding k as the "effect" of the "cause" p, we can invert this statistical relation to obtain the probability $dP(p|n,k)$ that $p \leq p_i < p + dp$ given k and n. Using Bayes rule of causes (see [94] p. 124), it is easy to find that $dP(p|n,k) = (n+1)P(k|n,p)dp$. From this we can obtain an "effective" transition probability

$$p^a_{n,k} = \int dP(p|n,k)p = \frac{k+1}{n+2} \tag{6.8}$$

where the last equality holds for $\phi(p) = 1$ (see later). The content of Eq. (6.8) is that, among all the processes and all the realizations of the disorder, the probability that the random walker will jump from site i to site $i+1$, given that it has made the same jump k times after the n previous visits, is $p^a_{n,k}$. This is the transition probability which is consistent, in a conditional way, to the past history of the process. The history of the process is in general encoded in the effective distribution of the variable p_i at time t, which was named run time statistics in [116]. In our case the distribution of p_i is parameterized by only two numbers n_i and k_i, and therefore a direct expression of the effective dynamics in terms of k_i and n_i only is possible. The structure of the memory can be described by placing a Polya urn on each site [94].

The model defined by Eq. (6.8) will be hereafter called a random walk with memory (RWM). Its evolution is defined as follows: define on each site i of the lattice two integer

"dynamical" variables $n_{i,t}$ and $k_{i,t}$ which count the number of visits on site i and the number of jumps $i \to i+1$. At time $t = 0$, $n_{i,0} = k_{i,0} = 0$ and the walker is at site $i = 0$. At time t, if the random walker is at site i, then with probability $p^a_{n_{i,t},k_{i,t}}$ it will move to site $i+1$ and $k_{i,t+1} = k_{i,t} + 1$. Otherwise the walker moves to site $i-1$ and $k_{i,t+1} = k_{i,t}$. In either case $n_{i,t+1} = n_{i,t} + 1$ increases by one. This process, by construction, is expected to reproduce the same results of the RRW with a random realization of $\{p_i\}$. In the RWM, the transition probabilities depend on the dynamical variables $\{k_{i,t}, n_{i,t}\}$ and therefore evolve in time. On the contrary, in the RRW, the transition probabilities p_i are fixed before the process starts. The equivalence of the dynamics of the two walkers results from the fact that each realization of the RWM asymptotically singles out a realization of the disorder, in the sense that $p^a_{n_{i,t},k_{i,t}} \to p_i$ as $t \to \infty$, where p_i is a uniform random number in $[0,1]$. This has been explicitly checked in numerical simulations, but it can also be argued from the distribution $dP(p_i|n,k)/dp$ of p_i. This is indeed sharply peaked around the mean value $p^a_{n,k}$, with a width of order $1/\sqrt{n}$. The statistics of the asymptotic value of $p^a_{n,k}$ as $n \to \infty$ can be explicitly shown to be that of uniform random variables by analyzing the moments of the effective transition probability $p_i(n_i) = p^a_{n_i,k_i}$. Dropping the i index for the moment, one observes that at the $n-1^{\text{st}}$ visit $p(n-1)^q$, with probability $p(n-1)$ increases to $\left[\frac{(n+1)p(n-1)+1}{n+2}\right]^q$ while with probability $1-p(n-1)$ it becomes $\left[\frac{(n+1)p(n-1)}{n+2}\right]^q$. Taking the average over realizations, leads to a recursion relation for the moments of $p(n)$ which, with a little algebra, can be solved to find

$$M_q(n) = \langle p(n)^q \rangle = \frac{1}{n+1} \sum_{k=1}^{n+1} \left(\frac{k}{n+2}\right)^q . \tag{6.9}$$

Note that $M_1(n) = 1/2$ for all n. Moreover all central moments $\langle [p(n) - \langle p(n) \rangle]^q \rangle$ with q odd vanish identically. For $n \gg 1$, one easily finds $M_q(n) = (1+q)^{-1} + O(n^{-1})$, i.e. the moments of $p(n)$ tend indeed to those of a uniform distribution in $[0,1]$. Therefore, the distribution of the transition probabilities, for a RWM in a box of size L with periodic boundary conditions, will asymptotically tend to a delta function around a random value p_i whose statistics is uniform in $[0,1]$. However, strictly speaking, even with periodic boundary conditions, the random walk will never reach a stationary state. This is reminiscent of systems out of equilibrium.

Another interesting observation is that one can easily calculate the probability of a realization of the process, i.e. of a given history $\{x(\tau) : \tau = 1, t\}$. This is indeed given simply by $P\{n_{i,t}\} = \prod_i [n_{i,t}+1]^{-1}$ citenote. Note that to obtain such a quantity in the RRW, one needs to evaluate it for a given realization of the disorder and then average over all realizations.

The diffusion law $\delta x \sim (\ln t)^2$ can be understood, in the context of the RWM, with the following argument. First we note that the values of k_i and n_i on different sites are not independent. For example it is easy to check that $t = \sum_i n_i$ and $x_t = \sum_i (2k_i - n_i)$. In

general $n_i = k_{i-1} + n_{i+1} - k_{i+1}$. In this relation the k's are distributed uniformly between 0 and the n's. Then, approximately, this relation has the form $n_{i+1} \simeq C_i n_i$ with C_i a random variable. In other words the variable $\ln n_i$ will have the shape of a random walk over i, which means that typically the maximum value of n_i for $i \in [0, L(t)]$ will be $n_{\max} \sim \exp \sqrt{L(t)}$. Since this value will also dominate the sum $\sum_i n_i = t$, we can conclude that $L(t) \sim (\ln t)^2$.

One striking feature of the RRW is the lack of time translational invariance. It was pointed out [120] that two times correlation functions are not functions of the difference of the times, as is normally the case, but also depend on the "waiting" time (i.e. the smallest time). This was related in Ref. [120] to the aging phenomena observed in spin glasses and glasses. The calculation of $\langle A_t A_{t+\tau} \rangle$, where A_t is any observable, depends only on processes between times t and $t + \tau$. If the transition probabilities involved in these process are constant in time, time translation invariance follows naturally. The lack of time translational invariance is no surprise in the RWM, because the transition probabilities explicitly depend on the "waiting" time t. This point can be hardly appreciated in the framework of the RRW, where the transition probabilities are fixed from the beginning. The absence of quenched disorder in the RWM evidences the fact that aging effects result from local memory effects. These effects, as shown by the equivalence of the RRW and RWM, are also present in disordered dynamical systems.

One might wonder what happens if instead of a uniform distribution one considers a general distribution $\phi(p)$. It is not difficult to show that all the above considerations hold the same, apart from the specific form of the moments and of the distribution of $p_i(n)$. Indeed Eq. (6.7) still holds. However when one inverts it to find the distribution $dP(p|n,k)$ one has to account for the fact that the probability that $p \leq p_i < p + dp$ is $\phi(p)dp$ with $\phi(p) \neq 1$ in general. In practice Eq. (6.8) is slightly modified, but only up to factors of order n^{-1}. For example, if $\phi(p) = \Gamma(\alpha+\beta)x^{\alpha-1}(1-x)^{\beta-1}/[\Gamma(\alpha)\Gamma(\beta)]$, one finds $p^a_{n,k} = (k+\beta)/(n+\alpha+\beta)$. Our numerical check of the diffusion as a function of α for $\beta = 1$ confirm the depinning transition for $\alpha > 2$ found by Derrida [122].

To address the problem of localization we note that on each site the RWM can create a barrier. If the walker has failed to pass a site after n visits, its probability to overcome it at the next visit is $p^a_{n,0} = 1/(n+2)$. Even though this probability decreases, it decreases so slowly that any barrier will sooner or later be overcame. This results from a straightforward application of the Borel-Cantelli lemma [94]. It is worth to observe that this behavior is the probabilistic counterpart of the "marginal" localization properties of the RRW [118]. Indeed it is easy to show, by the same argument, that if $np^a_{n,0} \to 0$, as $n \to \infty$ the RWM would surely localize, sooner or later on some site. This marginality seems to be even stronger as suggested by the following argument. For any regular distribution $\phi(p)$, we found $np^a_{n,0} \to 1$

as $n \to \infty$. Let us therefore generalize our model by taking

$$p^a_{n,k} = \frac{k+1}{n+2} + a \sin\left(2\pi \frac{k+1}{n+2}\right). \tag{6.10}$$

This describes a generalized symmetric ($p^a_{n,k} + p^a_{n,n-k} = 1$) random walk with memory. Note that $np^a_{n,0} \to 1 + 2\pi a$. We expect that, for $a < 0$ the walker localizes, whereas for $a > 0$, for large times, the dynamics becomes that of a random walker without disorder (i.e. $p_i = 1/2$). This expectation is based on the fact that the function $f(x) \equiv p^a_{n,xn}$ seen as a map [i.e. $x_{n+1} = f(x_n)$] has two stable fixed points (0 and 1) and one unstable fixed point (in $x = 1/2$) in the first case ($a < 0$) while in the second case the stability is reversed (0,1 are unstable and 1/2 is stable). Our problem is not a map, but it is similar (it has also randomness). However, numerical investigation shows that our expectation is correct. For $a < 0$ the walker localizes, whereas for $a > 0$ all the transition probabilities $p_i \to 1/2$ as $t \to \infty$. In other words, as shown in Fig. 6.2, the dynamics recovers different distributions of the disorder in the three cases:

$$\begin{aligned} &\phi(p) = \tfrac{1}{2}\delta(p) + \tfrac{1}{2}\delta(p-1) && \text{for } a < 0 \\ &\phi(p) = 1 && \text{for } a = 0 \\ &\phi(p) = \delta\left(p - \tfrac{1}{2}\right) && \text{for } a > 0 \end{aligned} \tag{6.11}$$

From this point of view the case $a = 0$ is very peculiar. It is the only case for which the distribution which is recovered by the dynamics is continuous. The case $a < 0$ bears some resemblance with systems, such as the Hopfield model [111] or folding proteins [15], where the phase space has a peculiar organization and the dynamics "localizes" on a particular low energy state.

The above model can be generalized straightforwardly to higher dimensions d. This only requires the introduction of d dynamical variables $k_i^{(j)}$, $j = 1, \ldots, d$, one for each direction on each site. An even simpler generalization is the case of a d dimensional random walker with random traps: Assign a uniform variable $p_i \in [0, 1]$ to each site of the lattice. If the walker is on site i at time t, with probability p_i it remains on the same site at $t + 1$, and with probability $1 - p_i$ it diffuses to one of the neighbor sites. Still we can use $p^a_{n_i,k_i}$ for the probability of jumping out of site i, conditional to n_i visits and k_i previous jumps out of the trap. It is easy to see how the diffusion law is modified in this case. Indeed, apart from the fact that the walker can spend a time $n_i > 1$ over a given site before jumping to the next one, the diffusion is the same. This means that $\delta x^2 \sim N$ where N is the number of sites visited (i.e. the number of jumps). This is related to the time t by summing all the times spent on different sites: $t = \sum_{i=1}^{N} n_i$. This sum is dominated by the large n_i values. The probability that the walker has been trapped for n_i steps on site i is $(n_i + 1)^{-1}$. The probability that it will jump out of the trap is $p^a_{n_i,0} = 1/(n_i + 2)$. Therefore the distribution of n_i is $D(n) = [(n + 1)(n + 2)]^{-1}$. This means that, for $N \gg 1$, $t = \sum_{i=1}^{N} n_i \sim N \ln N$, which yields the diffusion law $t \sim \delta x^2 \ln \delta x^2$. We checked the logarithmic corrections to the

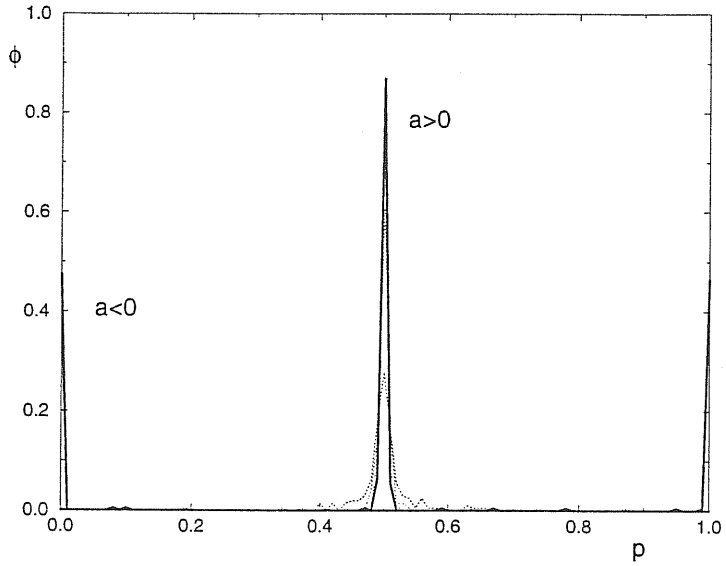

Figure 6.2: Probability density $\phi(p)$. The solid curve centered in $p = 0.5$ is obtained for $a = 0.1$ and is reminiscent of a random walk. The dotted lines are previous stages of simulation. The solid curve with two peaks in $p = 0$ and $p = 1$ refers to the $a = -0.1$ case where localization takes place.

diffusion numerically. In this case, using the generalized model of Eq. (6.10), it is easy to find that $D(n) \sim n^{-2-2\pi a}$. Therefore for $a > 0$, the above argument yields the standard diffusion $\delta x^2 \sim t$, whereas for $a < 0$ one finds anomalous diffusion $\delta x^2 \sim t^{1+2\pi a}$. Also in this case, therefore, disorder dynamics appears to be a borderline case.

In conclusion we have derived and discussed some simple models of random walks which reproduce the behavior of diffusion in disordered media *without* specifying the disorder. We have seen that the dynamics itself retrieves a realization of the disorder with the proper statistical properties. Our results may well be used to generate dynamically a random realization of the disorder in any model with quenched variables. It is tempting to conjecture that such an algorithm could provide an alternative to the simulated annealing [110] procedure used to find optimal configurations in disordered systems. The annealing procedure has indeed the drawback that, once the disorder realization is fixed, the starting configuration of the dynamical variables may be "far" from a reasonably good optimal state. Using the above results would instead produce dynamically a realization of the disorder which is "consistent" with the configuration of the dynamical variables.

6.4 Long-range Correlations in Protein Sequences

We have seen that protein are sequences built up with the 20 naturally occurring amino acids. Pictorially, the amino acids can be viewed as the *alphabet* in which protein sequences

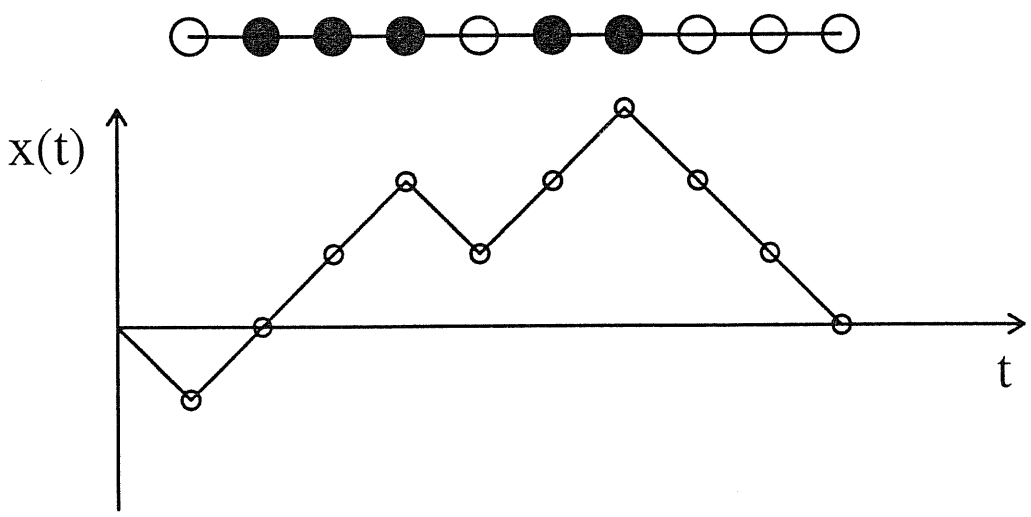

Figure 6.3: Mapping of an HP sequence on a random walk $x(t)$ in one dimension. H amino acids (full circles) correspond to a step up, P amino acids to a step down.

are written. Since it is assumed that the functional state of a protein is in its ground state conformation [2] the *grammar* used has to be that of the interactions between amino acids [123]. Random heteropolypeptides are not able to assume a unique stable and functional spatial structure (see chapter 4), and natural evolution should have selected special realizations of randomness. The selection criteria is that of functionality, which in turn depends on the structure. It is then interesting to ask if correlations in sequences would show up as a signature of non randomness. We use the idea of mapping the sequence on a random walk, originally introduced by Peng *et al.* [124] to study DNA sequences, and extended to amino acid sequences by Pande *et al.* [123]. Since the encrypted language is dictated by energetics the breaking code should depend crucially on the identification of the driving interactions. It is known that hydrophobic interaction is determinant in promoting the collapse of a protein into a globular state [29, 22]. Hydrophilic character of amino acids correspond to a step down and hydrophobic character to a step up of a one dimensional random walker. Each sequence then is mapped on a particular realization of a random walk $x(t)$, as shown in Fig. 6.3. The random walk is initially in the origin ($x(t=0)=0$) and at time $t+1$ it is in $x(t+1)=x(t)+1$ if the t-th amino acid is H, otherwise $x(t+1)=x(t)-1$. The problem of characterizing the correlations in a protein sequence is then reduced to the determination of the properties of such stochastic process. In addition to HP interactions Pande *et al.*, also considered the capability of forming hydrogen bonding, which is deemed responsible for the stabilization of secondary structures, and Coulomb interactions. They found that sequences extracted from protein databases show correlations remarkably higher that those expected for a pure random walk.

The relevance of the result of Pande *et al.* consist in having shown that the strategy underlying the encoding of sequences is of energetic nature. In this sense, correlations are the fingerprints that the selection of sequences is done by requesting stability in the ground state.

Is the correlation connected to the designability of the structure? It is known that there are structures that can host several sequences. Is a certain correlation needed to encode a given structure?

We fix a conformation Γ^*, the probability to find a given sequence S on Γ^* is given by its Boltzmann weight

$$P_S(\Gamma^*) = \frac{e^{-\beta \mathcal{H}_S(\Gamma^*)}}{\sum_\Gamma e^{-\beta \mathcal{H}_S(\Gamma)}}, \tag{6.12}$$

where $P_S(\Gamma^*)$ is normalized to 1 in conformation space. An important statistical quantity characterizing the random walk is the root mean square fluctuation $\delta x(t)$ of the displacement $x(t)$

$$\delta x(t) = \overline{x(t+\tau)^2} - \overline{x(t+\tau)}^2 , \tag{6.13}$$

where the average is taken over all times τ. For the pure random case $\delta x(t) \sim t^\alpha$, with $\alpha = 1/2$. Deviation from $\alpha = 1/2$ would signal presence of correlations. We assume that the average fluctuation $\langle \delta x_{\Gamma^*}(t) \rangle_S$ that is present when a sequence encode a given Γ^* is measured by an average in sequence space

$$\langle \delta x_{\Gamma^*}(t) \rangle_S = \sum_S \delta x_{\Gamma^*}(t) P_{\Gamma^*}(S) , \tag{6.14}$$

where $\delta x_{\Gamma^*}(t)$ is measured for a given S. The probability $P_{\Gamma^*}(S)$ is normalized as

$$P_{\Gamma^*}(S) = \frac{P_S(\Gamma^*)}{\sum_{S'} P_{S'}(\Gamma^*)}. \tag{6.15}$$

The meaning of Eq. (6.14) is that a sequence S contributes to the average correlation with a weight $P_{\Gamma^*}(S)$ proportional to the Boltzmann probability for that sequences to be on Γ^*. If a sequence has its ground state on Γ^* then, at low temperature, it will give a large contribution to the sum in Eq. (6.14). In the opposite case, if a sequence has a small statistical weight on Γ^*, then its contribution to the sum will be negligible. In other words only sequences that have a unique ground state on Γ^* are relevant in the calculation of correlation $\langle \delta x_{\Gamma^*}(t) \rangle_S$. Note that this automatically excludes the bias given by sequences comprised by all hydrophobic or all hydrophilic residues, a typical shortcoming of design strategies such that of Ref. [45]. In particular, at $T = 0$ the average is simply

$$\langle \delta x_\Gamma(t) \rangle_S = \frac{1}{N_S} \sum_S \delta x_\Gamma(t), \tag{6.16}$$

where N_S is the total number of sequences considerated that have their ground state on Γ^*. The Monte Carlo method presented in chapter 3 [68] can be used to obtain $P_S(\Gamma^*)$ for given

S and Γ^*. To avoid dependence of results on specific features of a given Γ^*, we consider an overall average over several Γ^* of $\langle \delta x_\Gamma(t) \rangle_S$.

Acknowledgements

Among the people I would like to express my deepest gratitude to, Amos is in the first place. I am sincerely convinced that his good influence on me has been really decisive. My attitude towards research has changed drastically under his advice. The presence of Erio has been equally fundamental. With his wise patience, he has been strongly stimulating me to overcome my weaknesses throughout my entire permanence in the school. I am especially grateful to Giuseppe, who has always been present to clarify my uncertainties and to show me, through his example, how to be much more careful in many circumstances. A particular thanks is also for Massimo and Stefano, who have offered me a very good chance, that regretfully I have not been able to fully exploit. I would like to say special words of gratitude to Matteo, whose strange way of thinking has been a continuous source of wonder. The unconventional standpoint of Josh has been a real surprise, and with him, physics has revealed some of its astonishing sides. Jayanth was the person who probably showed me in the strongest way the importance and the pleasure of being determined in looking for clear and firm concepts. Although I has few occasions to speak with him, Eytan has demonstrated me the great impact that authoritative ideas can have. I am also grateful to Lazslo. In the days he has been here, I had the opportunity to learn much about his very active approach to physics. Many times I felt lost in dazing technical problems, and in many times Flavio rescued me. Along the years, Nicola's attitude towards difficulties has been always a precious example for me. I am grateful to Alessandro and Giovanni, who helped me feel at home in Paris, and to Tanya, Hemant and Carsten with whom a had a very good time in Santa Cruz. I have spent particularly pleasant time with Francesca and Sandro. My room mates Paolo and Guido deserve special gratitude, the former for having been trying his best to bear me and the latter for having been, maybe unwantedly, a hard challenge for me. Besides working with me, Cecilia was the first person to really believe in my capacity to grow plants. A thought is for Francesco, whose strange words have been traveling with me almost all the time. The final words are for Laura, for having been always close to me, even from the distance.

Bibliography

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

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

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

[4] G. Iori, E. Marinari, and G. Parisi, J. Phys. A: Math. Gen. **24**, 5349 (1991).

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

[6] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond*, World Scientific, (1987).

[7] B. Derrida, Phys. Rev. Lett. **45**, 79 (1980).

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

[9] E. I. Shakhnovich and A. M. Gutin, Europhys. Lett. **8**, 327 (1989).

[10] E. I. Shakhnovich and A. M. Gutin, J. Chem. Phys. **93**, 5967 (1990).

[11] J. Bryngelson, J. N. Onuchic, J. N. Socci, and P. G. Wolynes, Proteins: Structure, Function, and Genetics. **21**, 167 (1995).

[12] A. Sali, E. I. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).

[13] R. Unger and J. Moult, Bull. Math. Biol. **55**, 1183 (1993).

[14] A. L. Goldberg and K. L. Rock, Nature **357**, 375 (1992).

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

[16] P. S. Kim and R. L. Baldwin, Ann. Rev. Biochem. **51**, 459 (1982)).

[17] K. A. Dill *et al.*, Protein Science **4**, 561 (1995).

[18] E. I. Shakhnovich, V. Abkevich, and O. Ptitsyn, Nature **379**, 96 (1996).

[19] D. Thirumalai, J. Phys. I France **5**, 1457 (1995).

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

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

[22] H. Li *et al.*, Science **273**, 666 (1996).

[23] M. Kardar, Science **273**, 610 (1996).

[24] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. USA. **90**, 6369 (1993).

[25] D. K. Klimov and D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996).

[26] D. K. Klimov and D. Thirumalai, (1996).

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

[28] C. Tanford, *The Hydrophobic Effect: Formation of Micelles and Biological Membranes* (Wiley & Sons, New York, (1980)).

[29] H. S. Chan and K. A. Dill, Physics Today **46**, 24 (1993).

[30] S. Kamtekar *et al.*, Science **262**, 1680 (1993).

[31] E. I. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).

[32] S. Miyazawa and R. Jernigan, Macromolecules **18**, 534 (1985).

[33] S. Miyazawa and R. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[34] A. Kolinski, A. Godzik, and J. Skolnik, J. Chem. Phys. **98**, 7420 (1993).

[35] H. Li, C. Tang, and N. Wingreen, (1995), cond-mat/9512111.

[36] E. I. Shakhnovich and A. M. Gutin, Nature **346**, 773 (1990).

[37] R. A. Goldstein, Z. A. Luthey-Shulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA. **89**, 4918 (1992).

[38] H. S. Chan, Nature **373**, 664 (1995).

[39] P. G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca and London, (1979)).

[40] J. N. Onuchic, P. G. Wolynes, Z. A. Luthey-Shulten, and N. D. Socci, Proc. Natl. Acad. Sci. USA. **92**, 3626 (1995).

[41] D. C. Rapaport, Comp. Phys. Rep. **5**, 265 (1987).

[42] M. Karplus and E. I. Shakhnovich, (1996), cond-mat/9606037.

[43] J. U. Bowie, R. Lüthy, and D. Eisenberg, Science **253**, 164 (1991).

[44] M. P. Morrissey and E. I. Shakhnovich, (1996), cond-mat/9601120.

[45] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. USA. **90**, 7195 (1993).

[46] P. H. Verdier and W. H. Stockmayer, J. Chem. Phys. **71**, 2662 (1962).

[47] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **103**, 4732 (1995).

[48] N. Metropolis *et al.*, J. Chem. Phys. **21**, 1087 (1953).

[49] J. M. Deutsch and T. Kurosky, Phys. Rev. Lett. **76**, 323 (1996).

[50] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Proc. Natl. Acad. Sci. USA. **91**, 12976 (1994).

[51] K. Yue and K. A. Dill, Proc. Natl. Acad. Sci. USA. **89**, 4163 (1992).

[52] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, J. Mol. Biol. **252**, 460 (1995).

[53] I. Shrivastava *et al.*, Proc. Natl. Acad. Sci. USA. **92**, 9206 (1995).

[54] S. Ramanathan and E. I. Shakhnovich, Phys. Rev. E **50**, 1303 (1994).

[55] T. Kurosky and J. M. Deutsch, J. Phys. A: Math. Gen. **28**, 1387 (1995).

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

[57] F. Seno and A. L. Stella, J. Phys. France **49**, 739 (1988).

[58] F. Seno and A. L. Stella, Europhys. Lett. **7**, 605 (1988).

[59] M. N. Rosenluth and A. W. Rosenbluth, J. Chem. Phys. **23**, 356 (1955).

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

[61] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).

[62] P. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics*, Cambridge, (1995).

[63] K. Yue *et al.*, Proc. Natl. Acad. Sci. USA. **92**, 325 (1995).

[64] H. S. Chan and K. A. Dill, J. Chem. Phys. **94**, 3475 (1993).

[65] K. Yue and K. A. Dill, Phys. Rev. E **48**, 2267 (1993).

[66] K. Binder, Rep. Prog. Phys. **50**, 783 (1987).

[67] W. Nadler and D. L. Stein, Proc. Natl. Acad. Sci. USA. **88**, 6750 (1991).

[68] F. Seno, M. Vendruscolo, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **77**, 1901 (1996).

[69] T. Alber *et al.*, Nature **330**, 41 (1987).

[70] J. F. Reidhaar-Olson and R. T. Sauer, Science **241**, 53 (1988).

[71] J. D. Bryngelson, J. Chem. Phys. **100**, 6038 (1994).

[72] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, (1996), cond-mat/9606136.

[73] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. USA. **92**, 1282 (1995).

[74] G. Vogt, T. Etzold, and P. Argos, J. Mol. Biol. **249**, 816 (1995).

[75] R. F. Doolittle, Science **214**, 149 (1981).

[76] T. Hwa and M. Lässig, Phys. Rev. Lett. **76**, 2591 (1996).

[77] M. S. Friedrichs and P. G. Wolynes, Science **246**, 371 (1989).

[78] V. N. Maiorov and G. M. Crippen, J. Mol. Biol. **227**, 876 (1992).

[79] J. I. Siepmann and D. Frenkel, Mol. Phys. **75**, 59 (1992).

[80] D. Frenkel, G. C. A. Mooij, and J. I. Siepmann, J. Phys.: Cond. Matt. **4**, 3053 (1992).

[81] M. Dijkstra, D. Frenkel, and J. P. Hansen, J. Chem. Phys. **101**, 3179 (1994).

[82] F. A. Escobedo and J. J. de Pablo, J. Chem. Phys. **102**, 2636 (1994).

[83] L. R. Dodd, T. Boone, and T. N. Theodorou, Mol. Phys. **78**, 961 (1993).

[84] E. Leontidis and U. W. Suter, Mol. Phys. **83**, 489 (1994).

[85] D. Ceperley, M. H. Kalos, and J. L. Lebowitz, Phys. Rev. Lett. **41**, 313 (1978).

[86] A. Baumgärtner, J. Chem. Phys. **72**, 871 (1980).

[87] F. T. Wall and F. Mandel, J. Chem. Phys. **63**, 4529 (1975).

[88] N. Madras and A. D. Sokal, .

[89] A. D. Sokal, Monte Carlo Methods in Statistical Mechanics: Foundations and new Algorithms, Cours de Troisième Cycle de la Physique en Suisse Romande, (1989).

[90] M. Karplus and G. A. Petsko, Nature **347**, 631 (1990).

[91] B. R. Brooks *et al.*, J. Comp. Chem. 4, 187 (1983).

[92] J. C. Smith and M. Karplus, J. Am. Chem. Soc. .

[93] K. Kremer and K. Binder, Comp. Phys. Rep. **7**, 259 (1988).

[94] W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley & Sons, New York, (1971)).

[95] D. Chandler, J. D. Weeks, and H. C. Andersen, Science **220**, 787 (1983).

[96] U. Wolff, Nucl. Phys. B **322**, 759 (1989).

[97] K. A. Dill, Biochemistry **24**, 1501 (1985).

[98] C. D. Sfatos, A. M. Gutin, and E. I. Shakhnovich, Phys. Rev. E **48**, 465 (1993).

[99] D. M. Ceperley and E. L. Pollock, *in Monte Carlo Methods in Theoretical Physics* (edited by S. Caracciolo and A. Fabrocini, ETS Editrice, Pisa, Italy, Pisa, Italy, (1990)).

[100] V. Dagget and M. Levitt, Annu. Rev. Biophys. Biomol. Struct. **22**, 353 (1993).

[101] E. M. Boczko and C. L. Brooks, Science **269**, 393 (1995).

[102] P. D. Thomas and K. A. Dill, J. Mol. Biol. **257**, 457 (1996).

[103] P. D. Thomas and K. A. Dill, (1996), preprint.

[104] L. A. Mirny and E. I. Shakhnovich, (1996), cond-mat/9607024.

[105] F. Seno, A. Maritan, and J. R. Banavar, (1996), preprint.

[106] R. H. Swendsen and J. S. Wang, Phys. Rev. Lett. **58**, 86 (1987).

[107] D. Kandel and E. Domany, Phys. Rev. B **43**, 8539 (1991).

[108] H. G. Evertz *et al.*, Nucl. Phys. B (Proc. Suppl.) **20**, 80 (1991).

[109] D. Kandel, R. Ben-Av, and E. Domany, Phys. Rev. B **45**, 4700 (1992).

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

[111] J. J. Hopfield, Proc. Natl. Acad. Sci. USA. **79**, 2554 (1982).

[112] S. V. Buldyrev *et al.*, Phys. Rev. A **45**, R8313 (1992).

[113] G. Gruner, Rev. Mod. Phys. **60**, 1129 (1988).

[114] L. Cugliandolo and J. Kurchan, Phys. Rev. Lett. **71**, 1 (1993).

[115] L. Pietronero and W. R. Schneider, Physica A **119**, 249 (1989).

[116] M. Marsili, J. Stat. Phys. **77**, 773 (1994).

[117] M. Marsili, G. Caldarelli, and M. Vendruscolo, Phys. Rev. E **53**, R13 (1996).

[118] E. Tosatti, M. Zannetti, and L. Pietronero, Z. Phys. B **73**, 161 (1988).

[119] J. P. Bouchaud, A. Comtet, A. Georges, and P. L. Doussal, Ann. Phys. **201**, 285 (1990).

[120] E. Marinari and G. Parisi, J. Phys. A: Math. Gen. **26**, l1149 (1993).

[121] Y. G. Sinai, Theory of Prob. and its Appl. **27**, 256 (1982).

[122] B. Derrida, J. Stat. Phys. **31**, 443 (1983).

[123] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Proc. Natl. Acad. Sci. USA. **91**, 12972 (1994).

[124] C. K. Peng *et al.*, Nature **356**, 168 (1992).