Protein Folding Simulations:

Confinement, External Fields and Sequence Design.

Inaugural-Dissertation

zur Erlangung der

Doktorwürde der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt beim Fachbereich Naturwissenschaften

der Universität Kassel

von

Pedro Armando Ojeda May

aus Mérida, Yucatán, México

March, 2010

Supervisor, Prof. Dr. Martin E. Garcia

Institut für Physik

Universität Kassel,

Heinrich-Plett-Str. 40

34132 Kassel.

Day of disputation is 22-nd of March 2010

Abstract

The present Thesis looks at the problem of protein folding using Monte Carlo and Langevin simulations, three topics in protein folding have been studied: 1) the effect of confining potential barriers, 2) the effect of a static external field and 3) the design of amino acid sequences which fold in a short time and which have a stable native state (global minimum).

Regarding the first topic, we studied the confinement of a small protein of 16 amino acids known as 1NJ0 (PDB code) which has a beta-sheet structure as a native state. The confinement of proteins occurs frequently in the cell environment. Some molecules called Chaperones, present in the cytoplasm, capture the unfolded proteins in their interior and avoid the formation of aggregates and misfolded proteins. This mechanism of confinement mediated by Chaperones is not yet well understood. In the present work we considered two kinds of potential barriers which try to mimic the confinement induced by a Chaperon molecule. The first kind of potential was a purely repulsive barrier whose only effect is to create a cavity where the protein folds up correctly. The second kind of potential was a barrier which includes both attractive and repulsive effects. We performed Wang-Landau simulations to calculate the thermodynamical properties of 1NJ0. From the free energy landscape plot we found that 1NJ0 has two intermediate states in the bulk (without confinement) which are clearly separated from the native and the unfolded states. For the case of the purely repulsive barrier we found that the intermediate states get closer to each other in the free energy landscape plot and eventually they collapse into a single intermediate state. The unfolded state is more compact, compared to that in the bulk, as the size of the barrier decreases. For an attractive barrier modifications of the states (native, unfolded and intermediates) are observed depending on the degree of attraction between the protein and the walls of the barrier. The strength of the attraction is

measured by the parameter ϵ. A purely repulsive barrier is obtained for $\epsilon = 0$ and a purely attractive barrier for $\epsilon = 1$. The states are changed slightly for magnitudes of the attraction up to $\epsilon = 0.4$. The disappearance of the intermediate states of 1NJ0 is already observed for $\epsilon = 0.6$. A very high attractive barrier ($\epsilon \sim 1.0$) produces a completely denatured state.

In the second topic of this Thesis we dealt with the interaction of a protein with an external electric field. We demonstrated by means of computer simulations, specifically by using the Wang-Landau algorithm, that the folded, unfolded, and intermediate states can be modified by means of a field. We have found that an external field can induce several modifications in the thermodynamics of these states: for relatively low magnitudes of the field ($< 2.06 \times 10^8$ V/m) no major changes in the states are observed. However, for higher magnitudes than (6.19×10^8 V/m) one observes the appearance of a new native state which exhibits a helix-like structure. In contrast, the original native state is a β-sheet structure. In the new native state all the dipoles in the backbone structure are aligned parallel to the field.

The design of amino acid sequences constitutes the third topic of the present work. We have tested the Rate of Convergence criterion proposed by D. Gridnev and M. Garcia (*work unpublished*). We applied it to the study of off-lattice models. The Rate of Convergence criterion is used to decide if a certain sequence will fold up correctly within a relatively short time. Before the present work, the common way to decide if a certain sequence was a good/bad folder was by performing the whole dynamics until the sequence got its native state (if it existed), or by studying the curvature of the potential energy surface. There are some difficulties in the last two approaches. In the first approach, performing the complete dynamics for hundreds of sequences is a rather challenging task because of the CPU time needed. In the second approach, calculating the curvature of the potential energy surface is possible only for very smooth surfaces. The Rate of Con-

vergence criterion seems to avoid the previous difficulties. With this criterion one does not need to perform the complete dynamics to find the good and bad sequences. Also, the criterion does not depend on the kind of force field used and therefore it can be used even for very rugged energy surfaces.

Table of Contents

Abstract . . . ii
Table of Contents . . . v
List of Tables . . . vi
List of Figures . . . vii
1 Introduction . . . 1
2 Theory . . . 8
2.1 THE STRUCTURE OF A PROTEIN . . . 8
2.2 THERMODYNAMICS OF THE FOLDING . . . 12
2.3 THEORETICAL PROTEIN FOLDING MODELS . . . 18
2.3.1 Lattice Models . . . 18
2.3.2 Off-lattice Models . . . 20
2.4 INTERMEDIATE STATES IN THE FEL OF PROTEINS . . . 20
2.5 OPEN QUESTIONS: . . . 24
2.5.1 Effect of Confinement on Protein Folding . . . 24
2.5.2 Influence of an External Electric Field on Protein Folding . . . 26
2.5.3 Selection and Sequence Design . . . 29
2.6 ORGANIZATION OF THIS THESIS . . . 31
3 Results . . . 32
3.1 MODELS OF PROTEINS . . . 32
3.1.1 Model I . . . 33
3.1.2 Model II . . . 41
3.1.3 Reaction Coordinates . . . 43
3.2 COMPUTATIONAL ALGORITHMS . . . 45
3.2.1 Wang-Landau Algorithm . . . 45
3.2.2 Langevin Dynamics Algorithm . . . 53
3.2.3 Distance between Configurations and Rate of Convergence . . . 55
4 SIMULATION RESULTS AND ANALYSIS . . . 58
4.1 EFFECT OF CONFINEMENT ON THE INTERMEDIATE STATES OF A PROTEIN . . . 58
4.2 PROTEIN-FIELD INTERACTION . . . 68
4.2.1 Electric Field produced by a Nano-electrode . . . 79
4.3 SELECTION AND SEQUENCE DESIGN . . . 80
5 SUMMARY AND OUTLOOK . . . 94
Bibliography . . . 97
Publications related to this thesis . . . 108
Acknowledgements . . . 109
Curriculum Vitae . . . 110
Erklärung . . . 111

List of Tables

4.1 Transition temperatures T_f for different values of the radius R_c of the potential $V_1(r)$ (see main text). Note that T_f decreases for increasing R_c. T_f is the temperature at which the specific heat is a maximum. 60

4.2 Transition temperatures T_f for the confining potential $V_2(r)$ (see main text) for different degrees of hydrophobicity, ϵ= 0.0, 0.2, 0.4, 0.6, 0.8, 1.0 and for the bulk case. Notice that in general T_f decreases as ϵ increases. For $\epsilon = 1.0$ it is not possible to define T_f because the specific heat is almost completely attenuated. 63

4.3 The five sequences studied in this paper and their corresponding models. The folding time of the sequences is $t_f > 1 \times 10^7$ time steps. All the sequences have $N = 30$ monomers. 88

List of Figures

2.1 Structure of an amino acid showing the main atoms involved C, O, N and H, as well as the residue R. The bond lengths and angles are taken from Ref. [SF00]. 9

2.2 A sequence of 3 amino acids in a protein. The residues are displayed explicitly. 9

2.3 Hemoglobin structure with 574 amino acids. This molecule has a two-fold symmetry as revealed by X-ray diffraction. 11

2.4 Organization levels of proteins: a) the *primary structure*, which is the lowest level corresponds simply to the amino acid sequence. b) the *secondary structure* are the features produced by the hydrogen bonding, mainly α-helices and β-sheets structures. c) the *tertiary structure* is the resultant 3D shape of the protein resulting from the interactions between the α-helices and β-sheets structures. d) the *quaternary structure* is the arrangement of several protein chains. 12

2.5 Two of the most frequent protein structures: a) an α-helix where the torsional angles $\phi \sim -57^o$ and $\psi \sim -47^o$ and b) a β-sheet where $\phi \sim -139^o$ and $\psi \sim +135^o$. 13

2.6 a) Backbone structure of a protein showing the two degrees of freedom handled in the model, better known in the literature as the Ramachandran angles ϕ_i and ψ_i. b) Ramachandran plot for the protein PCNA, a human DNA clamp protein that is composed of both α-helices and β-sheets (PDB code 1AXC). The Ramachandran angles are Φ and Ψ. 14

2.7 The folding proceeds by minimizing the free energy at each step ΔF. The final state called the *Native State* is very compact and also stable. The hydrophobic residues (in black) are localized in the core of the *Native State*, while the hydrophilic residues are exposed to the water environment. 15

2.8 Schematic representation of the Free Energy Landscape (FEL) or the Potential Energy Surface (PES) of a protein with a funnel form. The y-axis refers to the internal energy E. The broadness of the funnel is a measure for the entropy. As the protein comes closer to the native state (global minimum of the PES), the loss of entropy (ΔS) is compensated by the decrease of internal energy (ΔE) whereupon the free energy is negative ($\Delta F < 0$) making the spontaneous change possible. 17

2.9 Lattice model of the native state of a protein with 27 amino acids. *Adapted from [SSK94a].* 19

2.10 Multicanonical histograms $H_{muca}(E, Q)$ of energy E and angular overlap parameter Q and the free energy landscapes $F(Q)$ at different temperatures for three sequences (a) S1, (b) S2 and (c) S3. Pseudo-phases are symbolized by D (denature states), N (native folds), I (Intermediates), and M (metastable states). *Taken from [SBJ07].* 21

2.11 Schematic view of the free energy landscape of the human prion as a function of pressure. The molar free energy differences of the four main conformations N_1, N_2, I_1 and I_2 are depicted as function of the pressure P at constant temperature T of 293K. *Taken from [KKZK06].* 22
2.12 Structure of the GroEL-GroES complex. 24
2.13 A schematic sketch of the cycle in the GroEl-GroEs-mediated folding of proteins. In step 1 the substrate protein is captured into the GroEl cavity. The ATPs and GroEs are added in step 2, which results in doubling the volume, in which the substrate protein is confined. The hydrolysis of the ATP in the *cis*-ring occurs in the step 3. After binding ATP to the *trans*-ring, GroES and the substrate protein are released that completes the cycle (step 4). *Taken from [ME04].* 26
2.14 The dipoles of NH and OC in the amide plane give rise to a total dipole moment for each amino acid which has the value 1.1×10^{-29}Cm. 27
2.15 Alignment of the amide-plane-dipoles in a $\alpha-$helix structure. *Taken from Hol [Hol85].* 28
2.16 Root Mean Square Deviation (RMSD) from the structure at $t = 0$ for a simulation without an external field (broken lines) and for a simulation with a static, homogeneous field 2×10^9 V/m. $T_0 = 100$K (solid lines). (a) shows the RMSD for a simulation under the influence of an electric field of duration 1, 2, and 3 ps. (b) shows the RMSD for a simulation with a static field in the long-time behavior. *Taken from [XPS96].* 29

3.1 Off-lattice model for proteins: backbone units are represented by spheres with diameter 3.7842 Å. Each unit contains five atoms: C, O, N, H and C^α atoms. R represents the side chain which is attached to the C^α-atom in a rigid way. 32
3.2 Backbone structure of a protein showing the two degrees of freedom handled in the model, better known in the literature as the Ramachandran angles ϕ_i and ψ_i. For a chain of N amino acids one has $2(N-2)$ of such angles. 33
3.3 Dipole-dipole interaction between a NH and CO pair. 34
3.4 Hydrogen bond interaction between a CO and a NH pair. σ_{HB} is the distance between O and H'. The three angles θ_1, θ_2 and θ_3 are defined as $\widehat{\text{BOH}}$', angle between CO and N'H', and $\widehat{\text{AH}}$'O, respectively. Their average values are in the right of this figure. 35
3.5 The water molecules prevent that the residues reach the global minimum at r' creating a local minimum at r''. This effect is simulated by means of a potential LJ with minimum at r' and two Gaussians at $r' + 1.5$ and $r' + 3$. The size of a water molecule is ~ 3Å. *Taken from [CGO02].* . . 36
3.6 Plot of Eq. 3.15 without ϵ_{XY}. The $\sigma_{local} = R_{small} + R_{small} = 5.20$ Å, for example. $E(r)$ is -1.0 when $r < \sigma_{local}$ and $V(r)$ is zero when $r > \sigma_{local}+0.5$. X and Y are any two residues. 38

3.7 Plot of Eq. 3.16 without ϵ_{XY}. The arrows above are the vectors defined by C^{α} to the C^{β}-atom in the residues. X and Y are any two residues. . . 39
3.8 Curve of $V_1(r)$ (solid line) and $V_2(r)$ for different values of the parameter ϵ. $\epsilon = 0.0$ means a purely repulsive barrier and $\epsilon = 1.0$ a barrier highly attractive. The minimum of the potential $V_2(r)$ is localized near the surface of the barrier. The radius of the barrier is 15 Å. 42
3.9 The rugged energy landscape of the HMP a) compared to the smooth landscape of the DHTP b). Observe that the DHTP has a very deep global minimum which corresponds to the native state. Pictures are derived from the conformations obtained during numerous dynamical runs of slow cooling. The energy of each conformation is plotted as a function of its distance from two fixed "reference" conformations. *Taken from [CMB98].* 44
3.10 Scheme of the Harmonic Oscillator potential (black-dashed line) $V(X)$ and its exact DOS (red-solid line). 50
3.11 Discretization of the DOS in energy bins E_i. At each Monte Carlo Step (MCS) the DOS is updated as $g(\widetilde{E_i}) \rightarrow g(\widetilde{E_i}) + \ln f$. 51
3.12 Logarithm of the exact DOS (red-solid line) and the simulated DOS (black-dashed line) at different stages of the simulation. At the beginning of the simulation (a-b) the simulated DOS shortly differ from the exact one but after 1×10^9 Monte Carlo Steps (MCS) the simulated DOS converges to the exact DOS (c). At 2×10^9 MCS the simulated DOS has already converged (d). 52

4.1 Ground-state structure (β-sheet) of the peptide 1NJ0 ($E_g \sim -135$ Kcal/mol). 59
4.2 Besides the Native State N and the Unfolded U states in the Free Energy Landscape ($F(E,Q)$), there are other two states which are intermediates in the folding process, in the picture they are denoted as I_1 and I_2. $F(E,Q)$ is plotted in terms of the configurational energy E and the End-to-End distance Q. 61
4.3 Logarithm of the density of states (DOS) $g(E)$ of the protein inside the confining potential $V_1(r)$ and for different values of R_c (15 Å, 20 Å, 25 Å) as well as for the bulk case. One notices the remarkable decrease of the DOS for decreasing R_c. 64
4.4 Specific heat for the bulk case and for confining potentials with radii 15 Å, 20 Åand 25 Å. $T_f = 321$ K is the transition temperature in the bulk case. T_f increases as the radius R_c decreases. The confining potential in this case is purely repulsive. 65

4.5 Contour plots of the free energy landscape $F(E,Q)$ as a function of the configurational energy E and the end-to-end distance Q for a purely repulsive confining potential. Plots a-d correspond to the bulk case and cages of radius 15 Å, 20 Åand 25 Å respectively. The unfolded state are strongly affected when the size of the cage decreases. The native state and the intermediates are only slightly modified. The contour lines represent the free energy difference with respect to the native state and are given in Kcal/mol. 66
4.6 Logarithm of the DOS $g(E)$ for different degrees of hydrophobicity (ϵ =0.0, 0.2, 0.4, 0.6, 0.8, and 1.0) and for the bulk case. Notice the abrupt decay of $g(E)$ by ~ 13 orders of magnitude as ϵ goes from 0.0 to 1.0. For high values of ϵ, the protein tends to be in the unfolded state. 67
4.7 Specific heat of the protein for different values of $\epsilon = 0.0$, 0.2, 0.4, 0.6, 0.8, and 1.0, compared to the bulk case. $T_f = 321$ K is the transition temperature for the bulk. Notice how T_f and the peak of the specific heat decrease as ϵ goes from 0.0 (purely repulsive wall) to 1.0 (strongly attractive wall). 68
4.8 Contour plots of the free energy landscape $F(E,Q)$ for a cage with an attractive inner surface. Different degrees of hydrophobicity are displayed in plots a-d, corresponding to ϵ =0.0, 0.4, 0.6 and 0.8. The native and the intermediate states are slightly modified for $0.0 < \epsilon < 0.4$ but for larger values of ϵ the intermediate states dissappear and the native structure is deformed. As a consequence $F(E,Q)$ represents a two-states landscape. The contour lines represent the free energy difference with respect to the native state and are given in Kcal/mol. 69
4.9 The dipoles of NH and OC in the amide plane give rise to a total dipole moment for each amino acid which has the value 1.1×10^{-29}Cm. 70
4.10 Free energy surface of the V3-loop as a function of the configurational energy E and the end-to-end distance Q for different strengths of the external electric field: $\chi = 0.0, 0.4, 0.8$ and 1.2. Local minima labeled as I_1 and I_2 correspond to intermediates. N_1 refers to the native state in absence of field, which becomes metastable (I_3) for $\chi = 0.8$. Note the formation of a new global minimum N_2 for the field strength $\chi = 1.2$. U corresponds to the unfolded states. The temperature in all cases is $T = T_f = 321$ K. 71
4.11 For low field magnitudes one native (N_1) a) and two intermediate (I_1 and I_2) b)-c) states are displayed in the FEL of the peptide 1NJ0. For high field strengths the peptide presents a new intermediate (I_3) d) and native (N_3) e) states. (The intermediate states are schematic). The native state N_3 is aligned to the field orientation given by the black (red in color) line in e). 73

4.12 Coordinates in the (E, Q)-plane of the conformations yielding the maximal contribution to the partition function for $\chi = 0.0, 0.4, 0.8$ and 1.2 and at different temperatures. Note that for $\chi = 0.0$ the observable structures lie around the point $(E = -135, Q = 5)$ (β-sheet) while for $\chi = 1.2$ they are located near the point $(E = -150, Q = 30)$ (helix). Dark (blue) and light (yellow) diamonds refer to low and high temperatures, respectively (see temperature scale). 76
4.13 Ramachandran plot of the V3-loop for different strengths of the external electric field at $T = T_f = 321$ K. The regions corresponding to helices and β-sheets are indicated. 77
4.14 Electric field inside the chaperon as a function of the distance to one end of the cavity. The field decreases because of the screening of the electrolytes in the cytoplasm medium. 81
4.15 The place of the designed sequence S_0 resulting after ordering the sequences by the Rate of Convergence in descending order versus the time period t_0. *Taken from [GG].* 85
4.16 The normalized Rate of Convergence versus temperature for the designed sequence S_0 for the time period $t_0 = 300$. Dash-dot: the same for the sequence S_1. Dashed line: the normalized Rate of Convergence of a bad folder. The vertical dotted line corresponds to the folding temperature of S_0. The temperature is given in dimensionless Miyazawa-Jernigan units multiplied by 100. *Taken from [GG].* 87
4.17 Native states for the sequences S_1 (left) and S_2 (right). Dotted lines connect those monomers that are in contact. The energies in the native state are $E_N(S_1) = -16.88$ and $E_N(S_2) = -14.29$. The number of native contacts for S_1 and S_2 is 34 and 27 respectively. Taken from [GG]. 88
4.18 Specific heat vs. temperature for the five sequences shown in Table V. SEQ1 and SEQ2 show a very well localized peak which is a consequence of the funnel structure of their potential energy surfaces. These two sequences are known to be good folders. SEQ3, SEQ4 and SEQ5 have not a defined peak but the curve is spread in the whole interval of temperatures, they are known to be bad folders. The temperature axis is normalized respect to the transition temperature of SEQ2, T_{f2}=15.3 in units of k_BT. 89
4.19 Global minima of the five sequences studied in this Work called SEQ1-5. 90
4.20 Main frame: short time behavior of the Rate of Convergence for $T = 190$K. We observe that already after the time step 200 there is a clear separation of good (SEQ1 and 2) and bad folders (SEQ 3,4 and 5). Inset: for very long times one can distinguish between good and bad folders, the top of the sequences is reached by SEQ2 after the time step 1×10^5. 91
4.21 Rate of convergence for a wide range of temperatures. We observe that the distinction between a good (SEQ1) and a bad (SEQ5) folders is independent on the temperature. 92

Chapter 1

Introduction

Proteins are essential parts of organisms and participate in virtually every process within cells, let us quote for instance:

- are passive building blocks of many biological structures, such as the coats of viruses, the cellular cytoskeleton, the keratin in our skin or the collagen in our bones and cartilages;
- transport and store other species, from oxygen or electrons to macromolecules;
- act as hormones, transmit information and signals between cells and organs;
- act as antibodies, defend the organism against intruders;
- are the essential components of muscles, converting chemical energy into mechanical one, and allowing the animals to move and interact with the environment;
- control the passage of species through the membranes of cells and organelles, they are doorkeepers;
- control gene expression;
- are the essential agents in the transcription of the genetic information into more proteins;
- as chaperones, protect other proteins to help them to acquire their functional 3D structure via the folding process that we will discuss later.

Due to this participation in almost every task that is essential for life, protein science constitutes a support of increasing importance for the development of modern medicine. On one side, the lack or malfunction of particular proteins is behind many pathologies; e.g., in most types of cancer, mutations are found in the tumor suppressor p53 protein [AVR02]. Also, abnormal protein aggregation characterizes many neurodegenerative disorders, including Huntington, Alzheimer, Creutzfeld-Jakob ('mad cow'), or motor neuron diseases [EAF$^+$06, Kel98, LM00]. Finally, to attack the vital proteins of pathogens (HIV, SARS, hepatitis, etc.) [BNO08], or to block the synthesis of proteins at the bacterial ribosome [BPZ$^+$07], are common strategies to battle infections in the frenetic field of rational drug design.

Apart from medicine, the rest of human technology may also benefit from the solutions that Nature, after billions (10^9) of years of "research", has found to the typical practical problems. And that solutions are often proteins: new materials of extraordinary mechanical properties could be designed from the basis of the spider silk, elastin or collagen proteins. Also, some attempts are being made to integrate these new biomaterials with living organic tissues and make them respond to stimuli from the patient. Even further away on the road that goes from passive structural functions to active tasks, no engineer who has ever tried to solve a difficult chemical problem can avoid to experience a feeling of almost religious inferiority when faced to the speed, efficiency and specificity with which proteins cut, bend, repair, carry, link or modify other chemical species. Hence, it is normal that we play with the idea of learning to control that power and have, as a result, nanoengines, nanogenerators, nanoscissors, nanomachines in general.

In the late 1950s Christian B. Anfinsen and his colleagues at the National Institutes of Health made a remarkable discovery. They were exploring a long-standing puzzle in biology: what causes newly made proteins which resemble loosely coiled strings and are inactive to wind into specifically shaped balls able to perform crucial tasks in a living

cell? Such a process of getting a specific 3D form is called protein folding. Anfinsen found an interesting answer for this question during his experiments on ribonuclease folding [Anf73]. The ribonuclease-A is a relatively small protein, with four distinct disulfide bridges. The first step, he made, was to denature the protein with the chaotropic agent urea and the disulfide-reagent mercaptoethanol. Since it is not too complicated, it was a wonderful opportunity to follow the changes in enzymatic activity of this protein. Not surprisingly, after this treatment the enzymatic activity of native ribonuclease disappeared. After denaturation, he extracted the disulfide-reagent mercaptoethanol from the solution, and measured the enzymatic properties. It did not change, but if he extracted both the mercaptoethanol and urea from the system, the enzymatic activity reappeared. This experiment gave proof for the following statement: that there are some proteins, that are able to fold from an unfolded to a folded state within a relatively short time range, in *in vitro* circumstances, without any helper molecules, such as chaperones. So there is no need for a special coding mechanism for protein folding, the information for folding is fully encoded in the primary structure (sequence of amino acids). It seemed the amino acid sequence of a protein, a one-dimensional trait, was fully sufficient to specify the molecule's ultimate 3D shape and biological activity. (Proteins are built from a set of just 20 amino acids, which are assembled into a chain according to directions embedded in the genes.) Outside factors, such as enzymes that might catalyze folding, did not have to be invoked as mandatory participants.

The discovery, which has since been confirmed many times at least for relatively small proteins suggested that the forces most responsible for proper folding in the cell could, in theory, be derived from the basic principles of chemistry and physics. That is, if one knew the amino acid sequence of a protein, all that would have to be considered would be the properties of the individual amino acids and their behavior in aqueous solution (the interior of Most cells is 70 to 90 percent water). In actuality, predicting the conformation

of a protein on the basis of its amino acid sequence is far from simple. More than 30 years after Anfinsen made his breakthrough, hundreds of investigators are still at work on that challenge, which has come to be widely known as the protein folding problem. The solution is of more than academic interest. Many major products of the developing biotechnology industry are novel proteins. It is already possible to design genes to direct the synthesis of such proteins. Yet failure to fold properly or "misfolding" is a common production concern. Therefore, the researchers are interested on the possible internal and external factors which intervene in the folding and eventually on how to control them.

Not surprisingly when proteins do not fold correctly there can be serious effects, including many well known diseases, such as Alzheimer's, Mad Cow (BSE), and Parkinson's disease. That is one of the reasons of why the scientific community is interested on protein folding. One of the possible reasons for the misfolding of proteins is the existence of stable intermediate states [FFC06]. Those are states different from the native one in which the protein stays for a very long time because there is a energetic barrier difficult to overcome with thermal excitations. The intermediate states act as check-points where a protein gets trapped and after some time it can continue the folding or even it stays there for indefinite time. The problem arises when the intermediate states are very stable, it means that a protein which reaches such a state cannot go out from this state easily. The protein stays in this state and cannot reach the native state in an appropriate time for its correct functioning inside the living organism. Then, a series of unexpected chemical reactions occur inside the organism and the final result is the appearance of several diseases. This fact makes the understanding of intermediate states of vital significance. In the present thesis we will study the presence of intermediate states in a certain protein, the 1NJ0 peptide, which is a small segment of the HIV. We will examine how those intermediate states can be modified and controlled by external factors such as an electric field or the confinement by potential barriers.

The understanding of protein folding was obtained from computer models *in silico* or from experiments in the laboratory *in vitro* in which an individual protein was denatured to observe it folding back into its original form. But, the situation is considerably more complex in the living cell *in vivo*. Although the fundamental energy rules also apply here, folding (at least of large proteins) rarely takes place spontaneously, as the ribosomes do not synthesize only one protein at a time. Instead, cells contain a vast number of proteins and other biomolecules at the extraordinarily high concentration of 340 grams per liter. Ordered protein folding in this cramped chaos is only possible under the supervision of specialized molecules, called chaperones, which accompany proteins and make sure that those that are being formed at the ribosomes do not clump together prematurely. Chaperones do not merely oversee the folding of the protein, they also protect its tertiary structure (3D shape of the protein) in situations in which the cell is under stress; for example, elevated body temperature, so these chaperones have also been classified as heat-shock proteins (HSPs). The HSP70s, so called because they have a molecular weight of 70 kilodaltons, are the most important class of chaperones. A chaperone is a molecule shape like a double ring which fits round the protein chain like a cylinder so that it can fold undisturbed inside. By confining the developing protein the chaperone protect those parts of the protein that are particularly sensitive to premature reaction with the environment and therefore to malformation. Although the cylindrical folding cage opens every 10 seconds, the protein only leaves the chaperone when it has achieved its required native structure. Even though several studies have been performed regarding the chaperones[TKT03, TKL03, RKP05, NSC06, JBS04, FS06] many questions remain open, for instance, what is the influence of the confinement on the folding?, what is the influence of the degree of hydrophobicity inside the chaperon?. We will give new insights into these questions along the present Thesis.

Solving the folding problem has enormous implications: exact drugs can be designed theoretically on a computer without a great deal of experimentation. Genetic engineering experiments to improve the function of particular proteins will be possible. Simulating protein folding can allow us to go forward with the modeling of the cell. We now understand better than ever how protein folding both *in vitro* and *in vivo* takes place. And this, in turn, has given us a better understanding of the origin and course of diseases that are associated with defective protein folding. However several questions remain open, for instance, given a certain amino acid sequence, how to know if it will fold into a unique native state in a relatively short time (compare to random sequences). Computer simulations cannot yet solve the folding code that is hidden in the primary structure by simply calculating the molecular dynamics atom by atom, as to work through just 50 milliseconds of folding would take even the fastest computer around 30,000 years. Any realistic hope of cracking the folding code, such as to produce special designed proteins that evolution had not planned, is probably a very long way off. To perform the complete dynamics (until the protein is folded) by the classical methods of just one sequence would take a long time. If we want to perform the whole dynamics of hundreds or even thousands of sequences to have more statistics, the required time for the simulations would be unimaginable. Therefore, it would be very helpful if we had at hand a criterion to decide if a protein will fold correctly to a native stable structure without performing the complete dynamics, which could take a very long time. In the present Thesis, we propose a new algorithm called the Rate of Convergence to decide if a protein is a good or bad folder from the very beginning of the dynamics. With our algorithm we save a lot of CPU time when trying to decide which amino acid sequence will fold correctly within a short time.

It would be wonderful if researchers had an atomic-level microscope that could take a movie of individual protein molecules folding up from their extended, unstable state to

their final, or native, state, which is more stable. From a collection of movies, all aspects of the reaction pathways could be seen directly. Unfortunately, no such instrument exists; investigators must fallback on much less direct measurements and very careful reasoning. One can gather helpful clues to the rules of folding by examining the three dimensional structures of unfolded and fully folded proteins and by analyzing the properties of individual amino acids and small peptides (linear chains of amino acids). Fortunately, the architecture of hundreds of native proteins has been determined by such imaging techniques as X-ray crystallography and, more recently, nuclear magnetic resonance (NMR). Both techniques have advanced dramatically in the past decade, as has theoretical work attempting to predict folding mathematically by computer. In particular the present Thesis is oriented to the computer simulations in protein folding. We have made use of Wang-Landau and Langevin algorithms of several proteins to give new insights into the protein folding problem. We have dealt with topics of actuality such as the confinement of proteins, the protein-electric field interaction and the sequence design.

Along this Thesis we will learn about the common models used to simulate proteins, about two of the most important algorithms to solve the dynamics and thermodynamics of proteins (Langevin and Wang-Landau algorithms respectively) and the essential features of the folding process.

Chapter 2

BASIC CONCEPTS ABOUT PROTEIN FOLDING

In this Chapter we explain the basic concepts behind protein folding. Section 2.1 is devoted to the description of the protein structure. The Section 2.2 describes the thermodynamics of protein folding. The Section 2.3 is a brief overview of the protein models commonly used in computer simulations. The Section 2.4 introduces the concept of intermediate states and finally the Section 2.5 describes three of the open questions in protein folding which we addressed in the present Thesis.

2.1 THE STRUCTURE OF A PROTEIN

An amino acid is a molecule containing both the amine and carboxyl functional groups, they have the general formula $H_2NCHRCOOH$ where R is an organic substituent called "Residue", (see Fig. 2.1). Only 20 kind of amino acids exist in the nature and differ among themselves just by the organic group R. The amino acids can bind to each other by means of polymerization reactions and form chains, as displayed in Fig. 2.2. These chains are known as Proteins.

Proteins play an essential role in all forms of life. Some of the functions of proteins include control gene expression [RTG+07], intercellular signaling [Gre98], control of histocompatibility [Con99] and transport of other proteins [RSS08]. Proteins show a high degree of *specifity*: it means that the function of a certain protein is highly determined by the 3D structure and the sequence of its amino acids. In general, one protein cannot be replaced by another one without altering the activities of the living organism.

The size of a protein can run from less than 50 amino acids in the chain, up to more

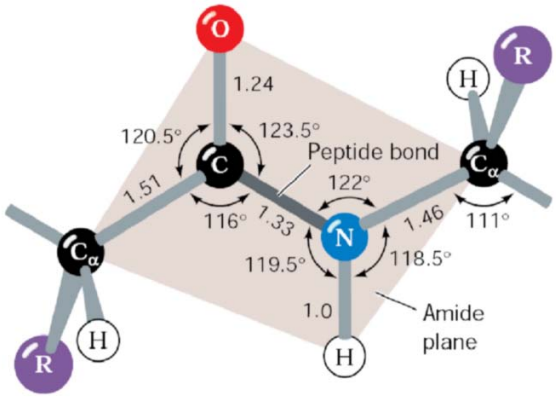


Figure 2.1: Structure of an amino acid showing the main atoms involved C, O, N and H, as well as the residue R. The bond lengths and angles are taken from Ref. [SF00].

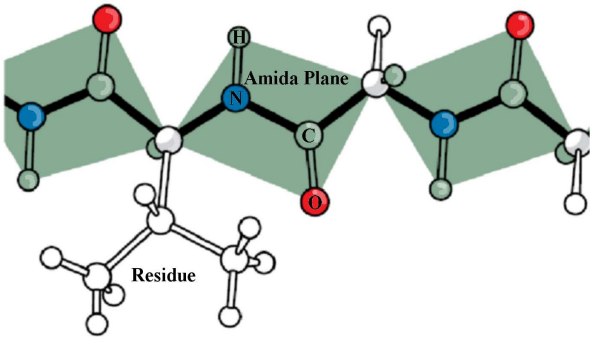


Figure 2.2: A sequence of 3 amino acids in a protein. The residues are displayed explicitly.

than 3000 amino acids. One of the largest amino acid chain is myosin, found in muscles, which consists of 1,750 amino acids. In Fig. 2.3 we show the structure of the protein Hemoglobin responsible for the transport of oxygen in the humans. Even for this middle-size protein we can already observe the high degree of complexity in the arrangement of the amino acids.

Much effort (about forty years worth) has been expended trying to understand how proteins fold up in nature. The goal is to fold up proteins from amino acid sequences which are easy to obtain (these days, the entire genome sequence for several organisms is available) into correct 3D structures (which are very few in number compared to the

number of amino acid sequences), theoretically (using a computer to do the actual folding steps). We are not very close in completing this goal, and so the Protein Folding problem remains one of the most basic unsolved problems in computational biology. With the advent of the computers, the people started the simulations of proteins using different kind of force fields between atoms. These simulations have the advantage that one can manipulate as many parameters as one wishes and observe how the system behaves. The goal of the computer simulations is to predict the real dynamics of the proteins and to see why the proteins behave as they do in our body. Several technological and pharmaceutical applications could be carry out using the results of protein studies. In spite of the fact that computer simulations provide a deep insight into the field of proteins they are limited by several factors. One of these factors is related to the CPU time needed because of the the long timescales required for folding processes. Another factor is related to the accuracy of the simulation: computer simulations make use of coarse-grained models, empirical potentials for the protein or solvent which in fact affect the accuracy of the simulation respect to the experimental results.

One can recognize different organization levels in Proteins. The lowest level corresponds to the amino acid sequence itself which is called the *primary structure*, see Fig. 2.4 a). The next level is the *secondary structure* which consists of the regularly repeating local structures stabilized by hydrogen bonds, the secondary structure is shown in Fig. 2.4 b). The most common examples are the α-helix and β-sheet, shown in Fig. 2.5. The protein can exhibit different secondary structures. Following the organization scheme, we distinguish the *tertiary structure*, which is the overall 3D shape of the protein, that is the spatial relationship of the secondary structures to one another, see Fig. 2.4 c). The tertiary structure is generally stabilized by nonlocal interactions, most commonly the formation of a hydrophobic core, but also through salt bridges, hydrogen bonds, and disulfide bonds. The term tertiary structure is often used as synonymous with the term

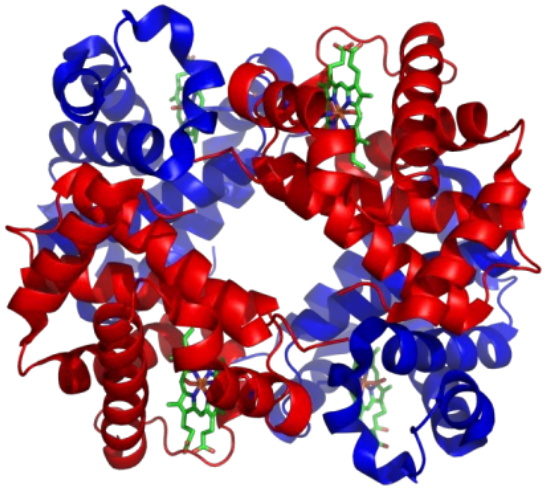


Figure 2.3: Hemoglobin structure with 574 amino acids. This molecule has a two-fold symmetry as revealed by X-ray diffraction.

"fold". The tertiary structure is what controls the basic function of the protein. The last level of organization is the *quaternary structure*, that is the structure formed by several protein molecules (polypeptide chains), usually called protein subunits, which function as a single protein complex. The quaternary structure is displayed in Fig. 2.4 d).

The structure of the protein is completely determined by the sequence of its torsional angles displayed in Fig. 2.6 a). Proteins can be characterized in general by their Ramachandran plots, this is a map which shows the possible torsional angles in the backbone structure. One can distinguish in the Ramachandran plot regions corresponding to α-helix and β-sheet structures among other conformations, as shown in Fig. 2.6 b). In this figure the white areas correspond to conformations where atoms in the polypeptide come closer than the sum of their Van der Waals radii. These regions are sterically disallowed for all amino acids except glycine which is unique in that it lacks a side chain. The black (blue in color) regions correspond to conformations where there are no steric clashes, i.e. these are the allowed regions namely the α-helix and β-sheet conformations.

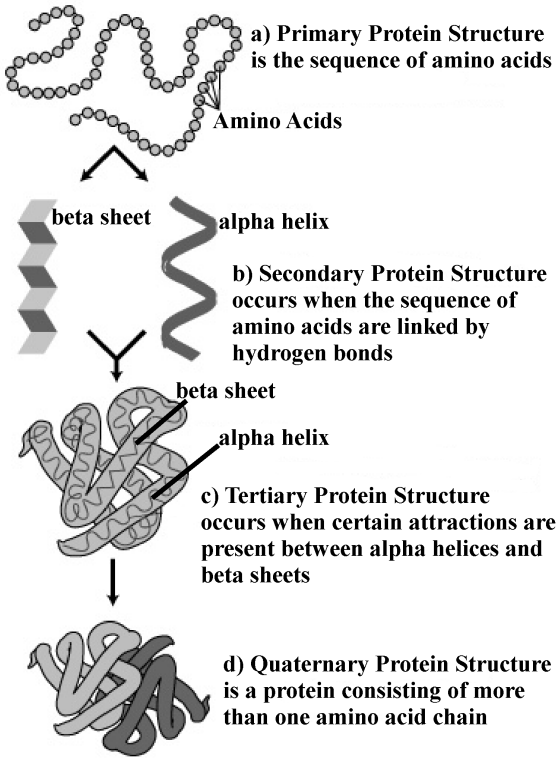


Figure 2.4: Organization levels of proteins: a) the *primary structure*, which is the lowest level corresponds simply to the amino acid sequence. b) the *secondary structure* are the features produced by the hydrogen bonding, mainly α-helices and β-sheets structures. c) the *tertiary structure* is the resultant 3D shape of the protein resulting from the interactions between the α-helices and β-sheets structures. d) the *quaternary structure* is the arrangement of several protein chains.

The grey (green in color) areas show the allowed regions if slightly shorter overlap between the residues occurs. This brings out an additional region which corresponds to the left-handed α-helix.

2.2 THERMODYNAMICS OF THE FOLDING

A long-standing problem in Biology has been the question of what makes proteins to fold, i.e what causes linear amino acid sequences to get the complex 3D stable structures which are vital for the function of a living organism. Proteins exhibit a particularly amazing

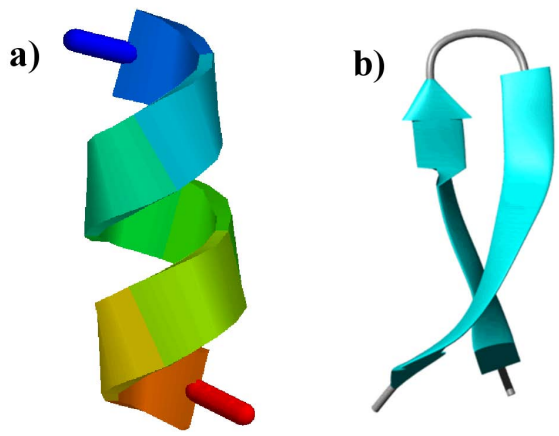


Figure 2.5: Two of the most frequent protein structures: a) an α-helix where the torsional angles $\phi \sim -57^o$ and $\psi \sim -47^o$ and b) a β-sheet where $\phi \sim -139^o$ and $\psi \sim +135^o$.

behavior when they are introduced into an aqueous environment. In this environment they tend to adopt a specific 3D form known as the *Native State*, this process is known as *Folding* and it is illustrated in Fig. 2.7. Under certain conditions of pH or temperature the native state can be unfolded and give place again to the random coil. Christian Anfinsen demonstrated that the process folding-unfolding is reversible for which he was awarded with the Nobel Prize in 1972 [Anf73].

The real problem with the folding of proteins is that, for a given sequence, we do not know *a priori* which 3D structure it will adopt. Therefore, it would be very useful to be able to predict the structure of a protein from its primary sequence for both scientific and industrial interests. For instance, we could design an artificial sequence which could acquire a determined native state and carry out a specific function. No less important is the fact that the misfolding of proteins is believed to cause diseases such as Creuztfeld-Jakobs and Alzheimer [EAF+06, Kel98, LM00]. Knowing the mechanisms of the folding one could, in principle, avoid such misfoldings or replace a given sequence by any other which could be less prone to misfolding.

The protein folding process can be compared to crystallization in the sense that a protein "condenses" in a unique stable structure. On the contrary, ordinary polymers

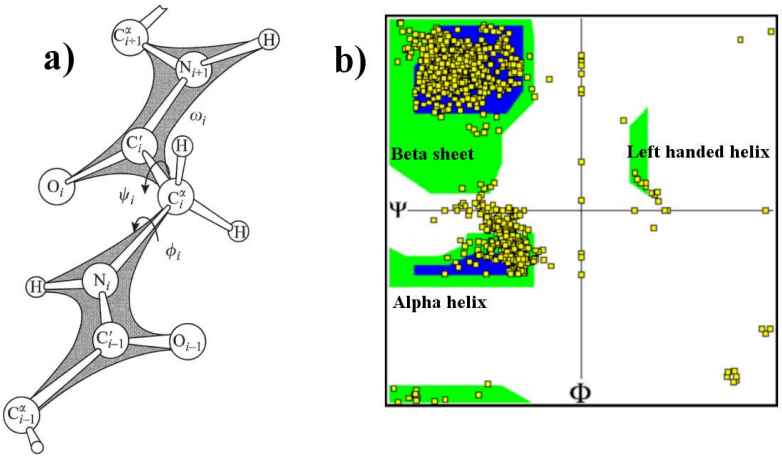


Figure 2.6: a) Backbone structure of a protein showing the two degrees of freedom handled in the model, better known in the literature as the Ramachandran angles ϕ_i and ψ_i. b) Ramachandran plot for the protein PCNA, a human DNA clamp protein that is composed of both α-helices and β-sheets (PDB code 1AXC). The Ramachandran angles are Φ and Ψ.

typically freeze to form amorphous globules, i.e. poly-peptides with random sequences which generally do not fold to unique structures.

A natural question which immediately arises is: what are the forces driving the folding of proteins? It has been established that the main forces involved are the electrostatic and hydrophobic interactions including the hydrogen bonds. There is consensus that the hydrophobic interaction is the major contributor to the stability of the native state of the protein. A way of understanding the hydrophobic effect is the example of a hydrophobic substance in water. Pure water molecules adopt a structure which maximizes entropy (S). A hydrophobic molecule will disrupt this structure and decrease entropy, and creates a "cavity" as it is unable to interact electrostatically with the water molecules. When more than one "cavity" is present, the surface area of disruptions is high, meaning that there are fewer free water molecules. To counter this, the water molecules push the hydrophobic molecules together and form a "cage" structure around them which will have a smaller surface area than the total surface area of the cavities. This maximizes the amount of free water and thus the entropy. Therefore the hydrophobic effect might

also be understood as the "the lipophobicity of water"

As a remark, the hydrophobic interaction between exposed non-polar amino acid residues on the surfaces of the protein molecule is, in general, attractive, short-range, and orientation dependent. By using these forces the amino acids are able to get the native state in a relative short time. The dynamics of this process was not clear till some years ago. The current picture is that the secondary structure forms at the very beginning of the folding. In an initial stage the protein collapse into a compact structure in whose center the hydrophobic amino acids are localized, leaving the hydrophilic amino acids exposed to the water. This condensed structure, called, *molten globule* in the literature, evolves through an even smaller ensemble of structures to a thermally jittered final tightly packed "single" structure. The thermodynamic guiding forces of protein folding will be most active in the early stages of folding because that is when the density of states is quite large while in the last stages of folding, when entropy has been reduced, the glass transition could well intervene.

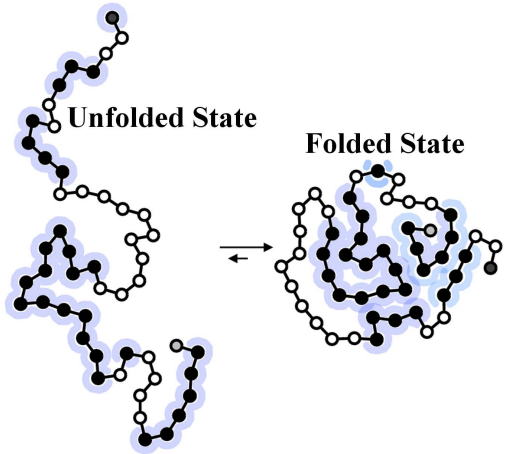


Figure 2.7: The folding proceeds by minimizing the free energy at each step ΔF. The final state called the *Native State* is very compact and also stable. The hydrophobic residues (in black) are localized in the core of the *Native State*, while the hydrophilic residues are exposed to the water environment.

Over a long time there was a dramatic discrepancy between the theoretical and exper-

imental folding times. On the one hand, it was found in the laboratory that the average folding time was between 10^{-3} and 1 sec. On the other hand, by using the random sampling hypothesis, one would conclude that the average folding time must be of almost 4 times the age of our universe. This disagreement between theory and experiment was called the *Levinthal Paradox* , after Cyrus Levinthal [Lev68].

In fact, such a paradox lacks of sense when we analyze carefully the details of the folding process. The Levinthal paradox would have validity in the assumption of every possible configuration sampled with uniform probability through the space of configurations. Explained in an illustrative way, this would be similar to leave a blind man in a landscape with many valleys and hills and wait until he finds the lowest place of the surface. It would take a very long time in average until the man reaches such a place.

What happens in reality is that the folding process is not random but it follows *routes* that minimize the Helmholtz free energy, $F = E - TS$. Along the folding path the changes in F are expressed as,

$$\Delta F = \Delta E - T\Delta S, \tag{2.1}$$

where ΔE and ΔS are changes in the internal energy and entropy respectively and T is the temperature of the environment. All this means that a compromise between energetic and entropic changes must exist so that a spontaneous transition from a configuration C_1 to another configuration C_2 can take place in such a way that $F(C_2) - F(C_1) \leq 0$. The presence of maxima and minima in the routes of folding allows us to introduce the concept of an energy surface where the thermodynamics of the folding evolves. Such a surface is called the Free Energy Landscape (FEL) in the literature [Wal03]. A very closed concept to the FEL is that of the Potential Energy Surface (PES) which can be observed as the FEL for a temperature $T = 0$ (without considering the electronic part of the molecule). The FEL (PES) is in general described in terms of some conformational

parameters or reaction coordinates which are supposed to captured the essential features of the folding. The FEL (PES) of the proteins has in most cases numerous roughnesses and entropic traps which make the global minimum (the native state) not attainable in a reasonable time or make it unstable.

The FEL (PES) roughnesses owes to the incapability to satisfy all the possible interactions in a single conformation or what is called *frustration*. In most of the proteins it is observed nevertheless that there exists a stable native state that can be reached in a relatively short time, which lead to the idea that the FEL (PES) of proteins should have a funnel form [LO92]. This is illustrated in Fig. 2.8.

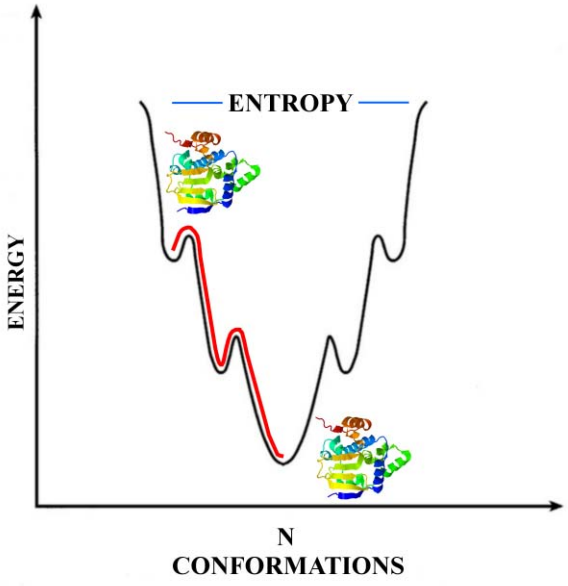


Figure 2.8: Schematic representation of the Free Energy Landscape (FEL) or the Potential Energy Surface (PES) of a protein with a funnel form. The y-axis refers to the internal energy E. The broadness of the funnel is a measure for the entropy. As the protein comes closer to the native state (global minimum of the PES), the loss of entropy (ΔS) is compensated by the decrease of internal energy (ΔE) whereupon the free energy is negative ($\Delta F < 0$) making the spontaneous change possible.

The form of funnel of the FEL (PES) tends to diminish the degree of frustration of a protein because the energetic traps present are small enough that they do not compete with the global energetic minimum that defines the native structural ensemble. This

has given place to the *Principle of Minimal Frustration* [BOSW95], which asserts that evolution has selected the amino acid sequences of natural proteins so that interactions between side chains largely favor the acquisition of the folded state. Interactions that do not favor folding are selected against, although some residual frustration is expected to exist. In general different kinds of funnels can exist depending on the amino acid sequence [Dil99].

2.3 THEORETICAL PROTEIN FOLDING MODELS

In the literature we find two general ways to describe the proteins depending on how they are allowed to move, that is, depending on whether they are confined or free to move in space. We will describe in the following paragraphs the essential ideas behind these two ways to describe the proteins in Computer Simulations.

2.3.1 Lattice Models

Lattice proteins are highly simplified computer models of proteins which were intensively used in the 90's to investigate protein folding. Actually, the first theoretical results in the field of protein folding came from lattice models [LD89, SSK94a]. Because proteins are such large molecules, containing hundreds or thousands of atoms, it is not possible with current technology to simulate more than a few microseconds of their behavior in all-atom detail. Hence real proteins cannot be folded on a computer. Lattice proteins, however, are simplified in two ways: the amino acids are modeled as single "beads" rather than modeling every atom, and the beads are restricted to a rigid (usually cubic) lattice. This simplification means that they can fold to their energy minima in a time quick enough to be simulated, see Fig. 2.9.

Lattice proteins are made to resemble real proteins by introducing an energy function, that is, a set of conditions which specify the energy of interaction between neighboring

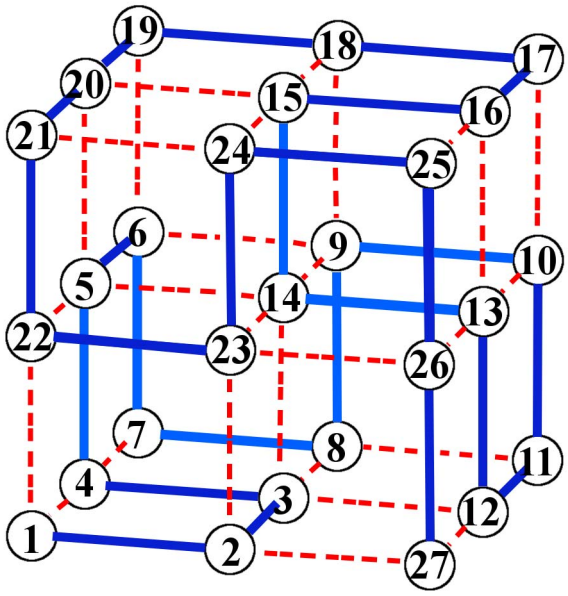


Figure 2.9: Lattice model of the native state of a protein with 27 amino acids. *Adapted from [SSK94a].*

beads, usually taken to be those occupying adjacent lattice sites. The energy function mimics the interactions between amino acids in real proteins, which include steric, hydrophobic and hydrogen bonding effects. The beads are divided into types, and the energy function specifies the interactions depending on the bead type, just as different types of amino acids interact differently. One of the most popular lattice models, the HP model, features just two bead types - hydrophobic (H) and polar (P) - and mimics the hydrophobic effect by specifying a negative (favorable) interaction between H beads [LD89, SBJ07, SSK94a]. The energy of a single chain $\mathcal{C}$ in the lattice models is given commonly as $E(\mathcal{C}) = \sum_{i,k=1}^{N} V_{ik}\Delta_{ik}(\mathcal{C})$, where the interaction matrix for monomers V_{ik} is determined by the Miyazawa-Jernigan matrix [MJ96]. This matrix, whose elements are statistically deduced pair-wise interaction potential energies among the twenty types of amino acids in proteins of known structure, has been widely applied to protein design and folding simulations [JB96, Sha94, PGT95]. $\Delta_{ik}(\mathcal{C})$ is the so called contact matrix, that is, $\Delta_{ik}(\mathcal{C}) = 1$ if the amino acids i and k are in a distance less than 8 Å and

$\Delta_{ik}(\mathcal{C}) = 0$ otherwise. We make use of this kind of energy in Chapter 4 where we study the folding of proteins from the lattice-models point of view.

2.3.2 Off-lattice Models

The problems regarding the oversimplification of the lattice models are solved by using the off-lattice models. These models are not restricted to a particular geometry of the grid, and all the atoms can move freely in space. The potentials used to simulate the proteins range from the Ab-initio [DWK98] to the so called Minimalist models [CSM06, SBJ07]. The Ab-initio models consider an all-atom potential and are very accurate. The problem with the Ab-initio models is that they are very time consuming. The Minimalist models consider an average of the forces over certain degrees of freedom in the protein structure and treat therefore the potential in a *mean field* approximation. The simulation time of these models is reduced considerably with respect to the Ab-initio approaches but these models are obviously less accurate. The degree of accuracy can vary depending on the type of approximation used for the forces. In the present work we employ only minimalist potentials (see description of Model I and Model II in Chapter 3).

2.4 INTERMEDIATE STATES IN THE FEL OF PROTEINS

The free energy landscape of a protein at a certain temperature could have several minima. Depending on the number of minima, we can have a two-state folding (with two minima), a folding through intermediates (more than two minima) and a glass-like folding into metastable conformations (more than two minima with almost the same free energy) [SBJ07]. The duration of the folding process varies dramatically depending on the protein of interest because of the presence of the intermediates. When studied outside the cell, the slowest folding proteins require many minutes or hours to fold primarily due to proline isomerization, and must pass through a number of intermediate states,

like checkpoints, before the process is complete [KB90]. Time scales of milliseconds are the norm and the very fastest known protein folding reactions are complete within a few microseconds [KHE04].

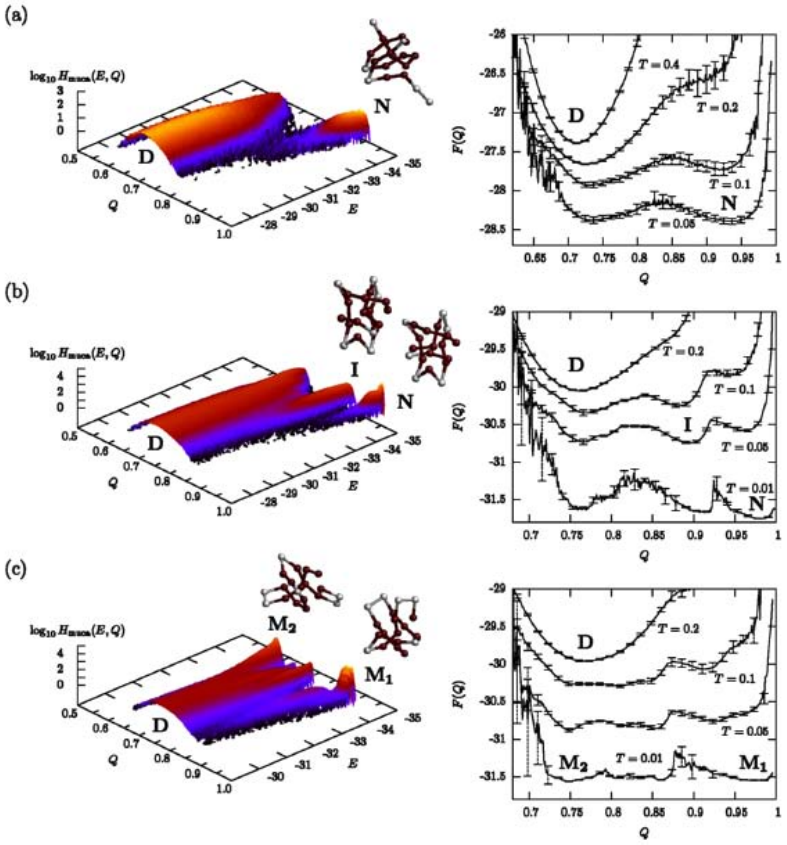


Figure 2.10: Multicanonical histograms $H_{muca}(E, Q)$ of energy E and angular overlap parameter Q and the free energy landscapes $F(Q)$ at different temperatures for three sequences (a) S1, (b) S2 and (c) S3. Pseudo-phases are symbolized by D (denature states), N (native folds), I (Intermediates), and M (metastable states). *Taken from [SBJ07].*

The intermediate states of proteins are important for the technological and medical applications. Suppose, for example, that one designs a protein in the laboratory which folds through intermediates. On the one hand it could be a problem the presence of intermediates because it would take a long time to reach the native state if the protein gets trapped in certain intermediates (local minima). In this case we would never reach the native state in an appropriate time. On the other hand it could be an advantage whether

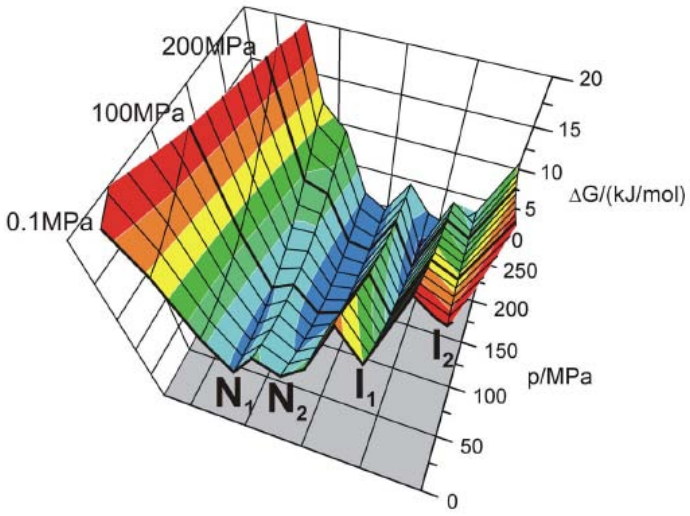


Figure 2.11: Schematic view of the free energy landscape of the human prion as a function of pressure. The molar free energy differences of the four main conformations N_1, N_2, I_1 and I_2 are depicted as function of the pressure P at constant temperature T of 293K. *Taken from [KKZK06].*

we want for a certain application that the protein stays in a determined intermediate state for a long time.

The concept of Intermediates comes from the fact that from time to time the protein conformations arrive to a local minimum, and they are not able to scape unless one gives them some external energy, by increasing the temperature for example. Because of the roughness of the FEL (PES) for proteins we expect to have several intermediates. However the folding mechanism was not always well understood in the early days of protein folding research [PK74]. A that time it was believed that the process follows a simple two-state transition. A two-state folding transition is explained simply as an equilibrium between a single folded conformation and an unfolded state as described above. This means that the transition involves only these two states with no accumulation of stable intermediates. The reaction coordinate of such a process will consist of two energy minima separated by a single energetic barrier. Nowadays we know that the free energy landscape can have in fact some intermediates besides the minima cor-

responding to the folded and unfolded states. Recent works confirm this hypothesis as in Refs. [SBJ07, OLYG09] and demonstrate that one can find different kinds of folding. For instance, Schnabel et. al, [SBJ07] could observe the three kinds of folding by using a minimalist model. The important point to remark is that even with a simple model of springs and Lennard-Jones potentials, as the one used by Schnabel, one can already observe many features in the free energy landscape, see Fig. 2.10.

Intermediates in protein folding, such as those observed for Lysozyme and Barnase, could result either from kinetic traps, which slow the folding process, or simply from additional free energy minima along the pathway, which could speed up the reaction [Kie95]. In either scenario folding is no longer a two-state, first-order-like transition. Evidence that both types of intermediates may occur, depending on the protein is provided by experimental results for Ubiquitin and Cytochrome$-c$ [WBCJ04]. Changing the balance between entropy and enthalpy can produce a change in behavior from rapid folding, without an intermediate, to mechanisms involving collapsed intermediate states or traps. In the laboratory such intermediates are achieved by creating mutants where the hydrophobic interaction is modified. These larger hydrophobic terms increase the chance that an intermediate free energy minimum exists, corresponding to a relatively compact state or molten globule.

Experimentally, the intermediate states are difficult to observe because there is no available technique which allows directly to monitor the folding on real time (order of microseconds). Techniques such as hydrogen exchange [KHLE04] and NMR [KKZK06] monitor the folding indirectly by measuring the number of hydrogen bonds present at a certain time. This give us an idea of how the 3D structure of the protein is, but, because not each atom is monitored, the resultant structure is not the real one but only a guess-average structure (since many structures satisfy the condition of having a certain number of hydrogen bonds). Intermediate states have been detected experimentally by Kachel

et. al.[KKZK06] using high pressure NMR spectroscopy, these intermediate states are displayed in Fig. 2.11.

2.5 OPEN QUESTIONS:

In spite of the numerous efforts that have been done to understand the folding process, many open questions remain. In the present thesis we addressed the following problems: the effect of confinement potentials, the influence of an external electrical field, and sequence design. In the following we describe the problems treated in our investigations.

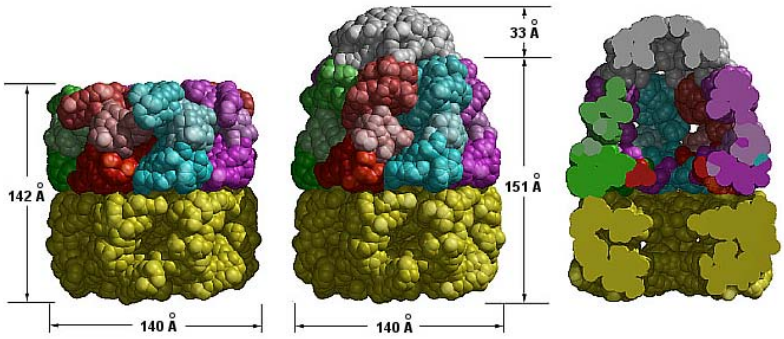


Figure 2.12: Structure of the GroEL-GroES complex.

2.5.1 Effect of Confinement on Protein Folding

The confinement effect is an important issue when the proteins are in the cellular environment surrounded by thousands of biomolecules. It has been found that crowding biomolecules make the folding process of a single protein sometimes almost impossible to be carried out. To overcome this problem certain structures called Chaperones play a major role. More than 50 families of Chaperones are known. The Groel-Groes found in bacteria is shown in Fig. 2.12. These chaperones are in fact hollow cylindrical proteins in whose interior smaller proteins can fold. The chaperones avoid the formation of undesirable aggregates of proteins and can sometimes assist misfolded protein to acquire

its correct native form [Ell06]. As a remark, aggregates of proteins are believed to be a cause of diseases like Alzheimer [EAF$^+$06, Kel98, LM00].

Considerable progress in understanding the mechanism of this nano-machine has become possible due to a combination of an extraordinary body of experimental work [FH97, XS99] and some contributions from theoretical studies [Thi94, GW94]. The function of the chaperone can be described as follows [PRF02], (illustrated in Fig. 2.13): the substrate (folding) protein is captured by the open cavity of the GroEl particle. To a first approximation, the mouth of the cavity can be thought of as a continuous hydrophobic surface. The interaction between the substrate protein and the GroEl is due to the attraction between the exposed hydrophobic residues of the substrate protein and the hydrophobic surface of the frontiers of the GroEl complex. Upon binding of ATP and GroEs significant chemical reactions occur in the GroEl particle.

The series of chemical reactions inside the Chaperone alter, in a fundamental way, the nature of interaction between GroEl and the substrate protein. Whereas in the process of capture the substrate protein-GroEl interaction is attractive, the interaction is either neutral or even repulsive after encapsulation. The surface remains hydrophilic until the restoration of GroEl to the initial state. This alternation between hydrophobic and hydrophilic surface enables this system to function as an annealing machine. The release of GroES and the protein occurs when the folding is finished.

The simplest form to model a Chaperon is by considering it as a potential barrier of impenetrable walls. Nevertheless since the walls have some unbalanced charges [Ell06], it is suitable to introduce an attractive potential which interacts with the residues of the proteins. Studies using molecular dynamics simulations [TKT03, LLW06] and considering confinement barriers whose size depend on the time, show that the size of the barrier is an essential factor in the stability of the protein and that the barrier not only does the folding more effective but also it collaborates in the correct folding of already misfolded

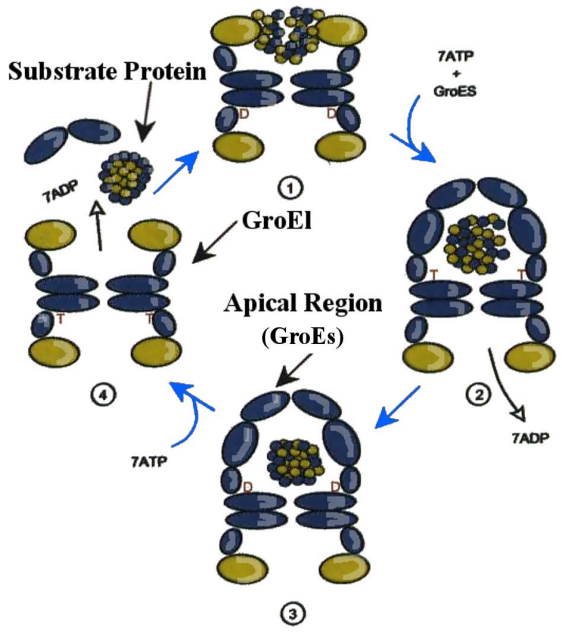


Figure 2.13: A schematic sketch of the cycle in the GroEl-GroEs-mediated folding of proteins. In step 1 the substrate protein is captured into the GroEl cavity. The ATPs and GroEs are added in step 2, which results in doubling the volume, in which the substrate protein is confined. The hydrolysis of the ATP in the *cis*-ring occurs in the step 3. After binding ATP to the *trans*-ring, GroES and the substrate protein are released that completes the cycle (step 4). *Taken from [ME04].*

proteins. In some way, the effect of a chaperon on the protein is the elimination of undesirable local minima, making the folding time shorter.

One of the goals for this Thesis is to study the behavior of small peptides under different kinds of confinement potentials. In Chapter 4 we will see that the thermodynamics of the folding is modified depending on the degree of confinement (induced by means of the barrier size) and also on electrostatic effects (caused by the attractive walls of the barrier).

2.5.2 Influence of an External Electric Field on Protein Folding

Due to the presence of electrical unbalanced charges in the structure of a protein (those belonging to the C, N, H, O atoms), permanent dipoles are present. In Fig. 2.14 one can

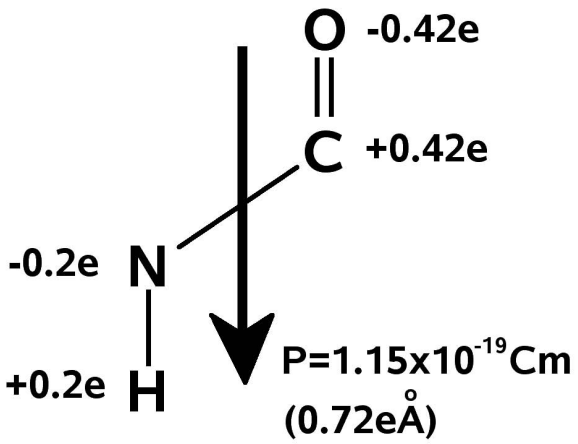


Figure 2.14: The dipoles of NH and OC in the amide plane give rise to a total dipole moment for each amino acid which has the value 1.1×10^{-29}Cm.

see the dipoles lying on the amide-plane of one amino acid. Note, that both dipoles are in the same direction and therefore there is a net dipole different from zero.

The dipoles in the plane of the amide can be lined up by means of an external field. Hol [Hol85] gave an experimental value of the dipolar moment on the amide plane of $1.1{\times}10^{-29}$Cm. As a remark the dipolar moment of a molecule of water is 6.1×10^{-30}Cm. In the same article, Hol mentioned that in structures such as α-helix, where all the dipolar moments are aligned, the total dipolar moment cannot be neglected. On the contrary, a structure such as the β-sheet where any two consecutive dipoles are antiparallel, the total dipolar moment is almost zero, see Fig. 2.15.

The interaction of an electric field (EF) with a protein has been used recently to align biomolecules in X-rays experiments [SSW$^+$05, RCF$^+$09]. The alignment is necessary in biomolecules particularly when these cannot form crystals. The crystals of the biomolecules are essential in order to conduct diffraction experiments to know the internal structure.

The alignment of dipoles by means of an EF finds an analogon in the Ising model of spins under an external magnetic field [Bin01]. Depending on the orientation of the

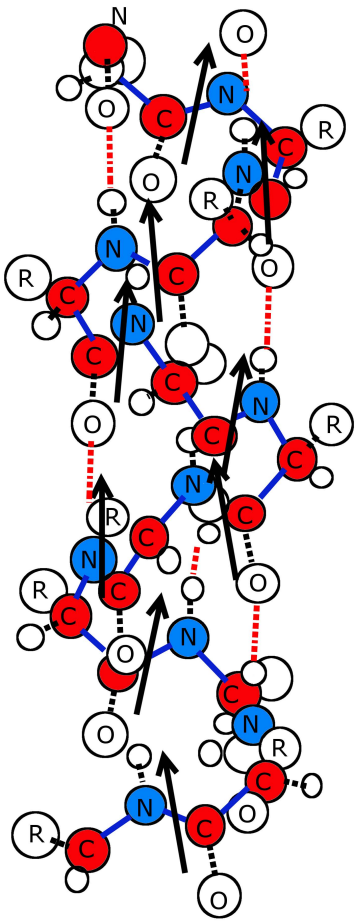


Figure 2.15: Alignment of the amide-plane-dipoles in a $\alpha-$helix structure. *Taken from Hol [Hol85].*

EF, the total energy could decrease or increase if the dipoles are parallel or antiparallel aligned to the EF.

Molecular Dynamics simulations of the folding in an external field have revealed that only high enough fields can interfere on the dynamics of the folding. Histeresis and relaxation effects have also been observed in big proteins [XPS96]. Fig. 2.16 displays one of the results of Schulten et. al [XPS96]. Here one can observe in the part a) the short time behavior of the Root Mean Square Deviation (RMSD) for a protein without any field (broken lines) and for homogeneous fields (2×10^9 V/m) of different pulse durations.

One observes that unless the duration of the pulse is grater than 1ps, the trajectory is not modified appreciably. As for the part b) of the figure, Schulten performed a simulation without a field (broken line) and with an static field (same magnitude as before) for long times. In this case, one observes that the trajectories end at completely different configurations even when they started with the same configuration and velocities. This means that a permanent field can in principle modify the dynamics of a protein and induce a different native state if the magnitude is high enough.

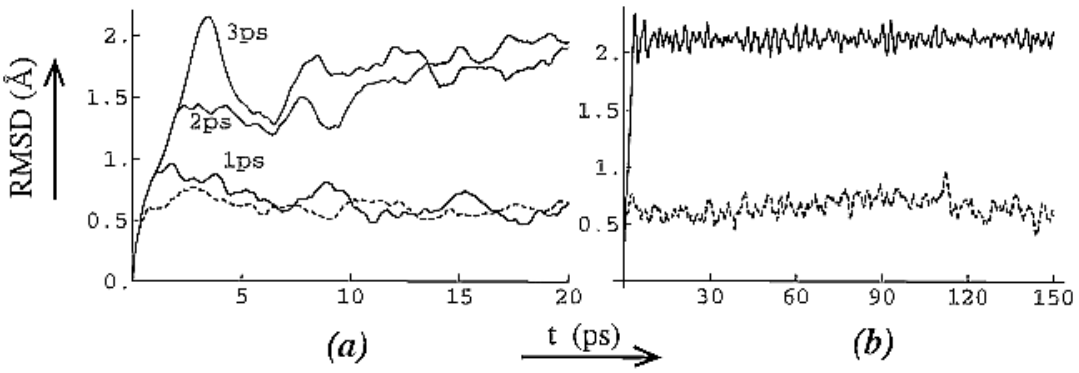


Figure 2.16: Root Mean Square Deviation (RMSD) from the structure at $t = 0$ for a simulation without an external field (broken lines) and for a simulation with a static, homogeneous field 2×10^9 V/m. $T_0 = 100$K (solid lines). (a) shows the RMSD for a simulation under the influence of an electric field of duration 1, 2, and 3 ps. (b) shows the RMSD for a simulation with a static field in the long-time behavior. *Taken from [XPS96].*

We will discuss in Chapter 4 how a protein interacts with an external field and also how the intermediate states of the protein are modified by the field. Our goal will be to study the possibility of inducing a native state in a protein different as the original one.

2.5.3 Selection and Sequence Design

It is of fundamental importance in the field of proteins to know how long a particular sequence would require to get its native structure and if this native structure would be stable at all. In what follows we mean by a good folder a sequence which possesses a very well defined native state and gets it in a relatively shorter time than the random

sequences. The problem of classifying good and bad sequences is of interest for the pharmaceutic industry where alternative sequences having a similar native structure and a shorter folding time are needed.

One possible way to know if a given sequence is a good or bad folder is by doing an exhaustive sampling of the PES using the Monte Carlo methods or Molecular Dynamics. However, doing the complete dynamics just to know if the protein would reach the native state in a short time could be a waste of time. It would be helpful if we had at hand a criterion to know which sequence will fold in a short time and which not without performing the whole dynamics.

Some work has been done in this direction, as an example we mentioned here the widely used criterion to characterize a good folder by looking at the energy gap between its global energy minimum and the minimum energy configurations, which are structurally dissimilar to the configuration of the global minimum [SSK94b, SG93, Sha94]. This energy gap ensures the "thermodynamic stability" and there is a strong correlation between the energy gap and the ability to fold into the global minimum in a reasonable time. Yet, without knowing the native state, there is still no good way to check whether a given amino acid sequence is a good folder other than letting it dynamically evolve from various initial conformations and checking if it does actually fold into a unique native state. Due to an unknown folding time it may take very long before one could identify some amino acid chain as a good/bad folder. A recent method [MC06b] for distinguishing proteins by their ability to fold suggests studying the curvature fluctuations of the energy surface along dynamical trajectories. However, the method is feasible for coarse-grained models with a smooth potential energy surface [MC06b]. Another impressive idea to distinguish folders comes from the analysis of the Microcanonical and the Canonical ensembles [HRL08], the main fact is that the features related to good and bad folders can be appreciated in the caloric curves obtained by means of the Microcanonical

ensemble. However, again to obtain the caloric curves one should perform before the complete dynamics.

In this topic we will investigate, in Chapter 4, to which extent the convergence of dynamical trajectories on the very initial stages could be a distinguishing feature for a good folder. The dynamics used for the description of an amino acid will be the Langevin dynamics, but other kinds of dynamics could be used for instance the Monte Carlo dynamics. We will propose in Chapter 4 a criterion to decide if a given sequence is a good or bad folder without doing the complete dynamics and even at the very beginning of the dynamics. This will be accomplished by defining a "distance" between structures in the configurational space.

2.6 ORGANIZATION OF THIS THESIS

This thesis is organized as follows: in the Chapter 3 we describe the protein model employed in our simulations, and we explain in detail the theoretical methods for both Langevin and Monte Carlo dynamics. In the Chapter 4 we present the results of our simulations and analyze them in detail and finally in Chapter 5 we give our conclusions and perspectives for the future work.

Chapter 3

THEORETICAL METHODS IN PROTEIN FOLDING

In this chapter we describe the two protein models employed in the present Thesis which simulate a Protein and the Computational Algorithms to describe the thermodynamics and the dynamics of the folding. The models are described in Section 3.1.

The Computational Algorithms are explained in Section 3.2. We used the Wang-Landau Algorithm [WL01] to calculate the Density of States of proteins. We also solved the Langevin equation to dynamical calculations.

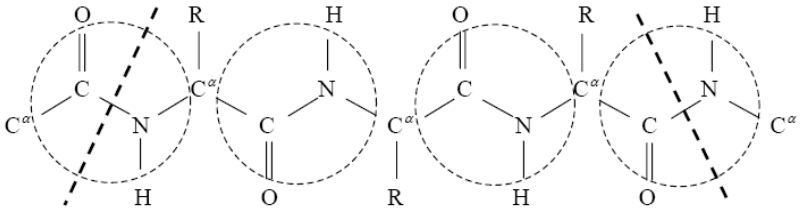


Figure 3.1: Off-lattice model for proteins: backbone units are represented by spheres with diameter 3.7842 Å. Each unit contains five atoms: C, O, N, H and C^{α} atoms. R represents the side chain which is attached to the C^{α}-atom in a rigid way.

3.1 MODELS OF PROTEINS

In this Section we explain the two force fields considered in our Work which we call Model I and II. Because the Model I includes more interactions than the Model II, it is expected that the dynamics of the folding is better described by Model I. However, Model II is much faster than Model I for computer simulations and allows to perform more statistics. At the end of this Section a few lines are dedicated to the Reaction Coordinates which allows us to describe the folding in terms of a few parameters.

3.1.1 Model I

The structure of the protein is simulated using the reduced off-lattice model developed in Ref. [CSM06]. The amino acids are represented by means of backbone units. Each backbone unit contains the atoms N, C_α, C', O and H. The residues are modeled as spherical beads, R, attached to the C_α's, see Fig. 3.1. The only remaining degrees of freedom are the Ramachandran angles ψ and ϕ, see Fig. 3.2. The values for the bond lengths and angles are given in Ref. [SF00].

The force field containing all relevant interactions in the protein is given by

$$E_{Protein} = E_{Steric} + E_{HB} + E_{DD} + E_{MJ} + E_{LocalHP}. \tag{3.1}$$

Here, E_{Steric} represents a hard-core interparticle-potential to avoid unphysical contacts and is given by,

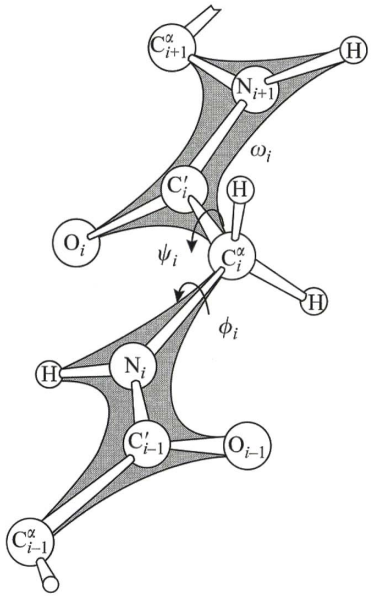


Figure 3.2: Backbone structure of a protein showing the two degrees of freedom handled in the model, better known in the literature as the Ramachandran angles ϕ_i and ψ_i. For a chain of N amino acids one has $2(N-2)$ of such angles.

$$E_{Steric} = \epsilon_{st} \sum_{ij} \left(\frac{\sigma_i + \sigma j}{r_{ij}} \right), \tag{3.2}$$

where σ_i and σ_j are the global radii of backbone or side units. r_{ij} is the distance between units.

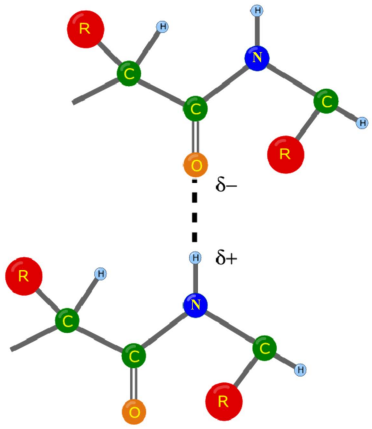


Figure 3.3: Dipole-dipole interaction between a NH and CO pair.

The second term of Eq. (3.1), E_{HB}, accounts for the hydrogen bonding. This interaction is the most important one in our model and is the major interaction responsible for stabilizing secondary structures [GT96, TSW99, ISW00, FIW02]. The hydrogen bonding is illustrated in Figs. 3.3 and 3.4. E_{HB} reads,

$$E_{HB} = \epsilon_{st} \sum_{ij} u(r_{ij}) v_1(\theta_{1,ij}, \theta_{1,ave}) v_2(\theta_{2,ij}, \theta_{2,ave}) v_3(\theta_{3,ij}, \theta_{3,ave}), \tag{3.3}$$

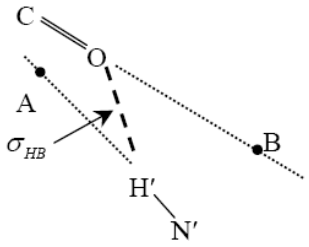
with

$$u(r_{ij}) = \epsilon_{HB} \left[5 \left(\frac{\sigma_{HB}}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma_{HB}}{r_{ij}} \right)^{10} \right], \tag{3.4}$$

and

$$v_1(\theta_{1,ij}, \theta_{1,ave}) = \begin{cases} \frac{1}{3} \left[4 \cos^2(\theta_{1,ij} - \theta_{1,ave}) \right] & \text{when } \theta_{1,ij} < (\theta_{1,ave} + \theta_r), \text{ l=1,2,3} \\ 0 & \text{otherwise} \end{cases}, \tag{3.5}$$

where i, j represent amino acids in which O_i and H_j atoms are belonging to, respectively. ϵ_{HB} is the strength of the interaction. r_{ij} is the distance between O_i...H_j. $\theta_{1,ij}$ are those angles defined above and $\theta_{1,ave}$ their average values.



the average of $\theta_1 \cong 26.77°$
the average of $\theta_2 \cong 11.60°$
the average of $\theta_3 \cong 17.98°$

the average of $\sigma_{HB} \cong 1.7834$ Å

Figure 3.4: Hydrogen bond interaction between a CO and a NH pair. σ_{HB} is the distance between O and H'. The three angles θ_1, θ_2 and θ_3 are defined as $\widehat{\text{BOH'}}$, angle between CO and N'H', and $\widehat{\text{AH'O}}$, respectively. Their average values are in the right of this figure.

In Eq. (3.1), E_{DD} describes the dipole-dipole interaction. This potential is one of the features that makes this model different from others. It is known that both the CO-NH groups on an amide-plane have partial charges; N and O atoms have an excess of negative charge and H and C atoms have an excess of positive charge. Because of this distribution, there is a dipole moment on the amide-plane as shown in Fig. 2.14 of Chapter 1. In CO-NH group this dipole moment is almost parallel to the CO and NH bonds and it has a value of $p = 1.15 \times 10^{-29} Cm$ [Hol85]. Moreover, when peptide chains are organized into a regular structure such as alpha helices or beta sheets, the total sum of small dipole moment may results in a net large moment. E_{DD} is given by,

$$E_{DD} = \epsilon_{DD} \sum_{ij} \left(\frac{p_i \cdot p_j}{r_{ij}^3} - \frac{3(p_i \cdot r_{ij})(p_i \cdot r_{ij})}{r_{ij}^5} \right), \tag{3.6}$$

where ϵ_{DD} is the global dipole-dipole interaction strength. p_i and p_j are any pair of CO (NH) dipoles and r_{ij} is the distance between them. The sum runs over all non-successive CO and NH dipoles. One simplification of E_{DD} occurs when we have two

dipoles located in consecutive amino acids. In this case the distance between dipoles is almost constant and E_{DD} is not more distance dependent. Therefore instead of using Eq. (3.6) to calculate the dipole-dipole interaction it is more convenient to take the following definition,

$$E_{DN} = \epsilon_{DN} \sum_{i,i+1} \left(\frac{p_i \cdot p_{i+1}}{|p_i||p_{i+1}|} - 1 \right), \tag{3.7}$$

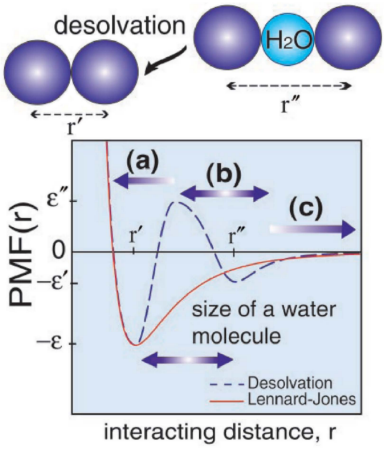


Figure 3.5: The water molecules prevent that the residues reach the global minimum at r' creating a local minimum at r''. This effect is simulated by means of a potential LJ with minimum at r' and two Gaussians at $r'+1.5$ and $r'+3$. The size of a water molecule is ~ 3Å. *Taken from [CGO02].*

As the reader probably has observed, in this model the dipoles are treated as localized spins on each amino acid. A similar situation occurs in an Ising model where spins are localized on a lattice [Bin01]. The main difference between the two models is that in our case the spins are free to move in space. We will make use of these localized dipoles in Chapter 4 when we discuss the interaction of a protein with an external electric field.

E_{MJ} is a distance-dependent version of the Miyazawa-Jernigan (MJ) matrix [MJ96], which describes the interactions between residues. In principle, two residues separated by a distance r_{ij} should be attracted until they are in contact. In our model we introduce this interaction by means of a Lennard-Jones potential. Note that this potential describes

the interaction in the absence of water. However this interaction is more complicated in the presence of water molecules. A single water molecule could enter the space between two residues and prevent them to reach the equilibrium distance, generating a potential barrier. To simulate explicitly the water molecules demands a very long CPU time because of the big number of water molecules required (~ 100000). In fact, most of the simulation time would be expended on the dynamics of these water molecules. Therefore in our model we consider the effect of the water molecules on the interaction between side chains using an effective potential that considers in addition to the Lennard-Jones potential, two gaussians that take into consideration the potential barrier and the local minimum caused by the presence of a water molecule, this kind of approximation was used previously, see for instance Refs. [HGG$^+$96, HGG$^+$98, CGO02]. Thus, the general expression for this interaction is,

$$E_{MJ} = \epsilon_{MJ} \sum_{i,j} \left[V_{LJ}(r_{ij}) + V_{Gaussian1}(r_{ij}) + V_{Gaussian2}(r_{ij}) \right], \tag{3.8}$$

$$V_{LJ}(r_{ij}) = \epsilon_{ij} \left[\left(\frac{r_1}{r_{ij}} \right)^{12} - 2 \left(\frac{r_1}{r_{ij}} \right)^{6} \right], \tag{3.9}$$

$$V_{Gaussian1}(r_{ij}) = \epsilon_b e^{-\sigma_w (r_{ij} - r_b)^2}, \tag{3.10}$$

$$V_{Gaussian2}(r_{ij}) = \epsilon_2 e^{-\sigma_w (r_{ij} - r_2)^2}, \tag{3.11}$$

$$\epsilon_b = |5\epsilon_{ij}/9| - \epsilon_{ij} \left[\left(\frac{r_1}{r_b} \right)^{12} - 2 \left(\frac{r_1}{r_b} \right)^{6} \right], \tag{3.12}$$

$$\epsilon_2 = -|\epsilon_{ij}/3| - \epsilon_{ij} \left[\left(\frac{r_1}{r_2} \right)^{12} - 2 \left(\frac{r_1}{r_2} \right)^{6} \right], \tag{3.13}$$

Here ϵ_{MJ} is the relative strength of global hydrophobic interaction. r_{ij} is the distance between two C^{β}-atoms. $r_1 = \sigma_i + \sigma_j$, $r_2 = \sigma_1 + 3$ and $r_b = \sigma_1 + 1.5$.

It is important to point out that E_{MJ} partially includes the effect of water polarization [WL00]. The values of the parameters of this potential are given in the original work by Chen et al. [CSM06].

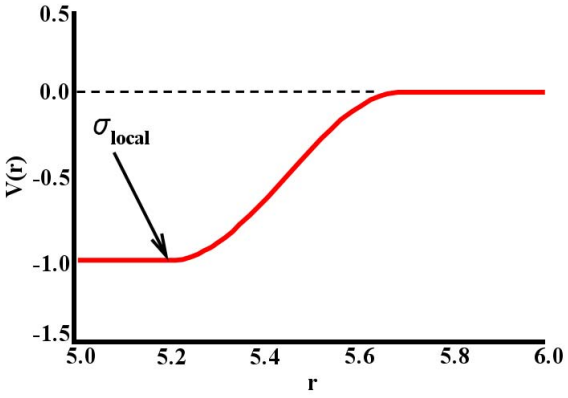


Figure 3.6: Plot of Eq. 3.15 without ϵ_{XY}. The $\sigma_{local} = R_{small} + R_{small} = 5.20$ Å, for example. $E(r)$ is -1.0 when $r < \sigma_{local}$ and $V(r)$ is zero when $r > \sigma_{local} + 0.5$. X and Y are any two residues.

A further term in Eq. (3.1), $E_{LocalHP}$ accounts for local hydrophobic effects. This interaction is also a sequence-dependent interaction. The idea of the local hydrophobic interaction is originated from the so-called HP model and the classification of amino acids. Amino acids are divided into two main classes as indicated in many textbooks. Some residues that try to to get away from water molecules are classified into the *hydrophobic* class. The interaction force between these residues has not a physical origin but a thermodynamic one. In fact the hydrophobic interaction is a consequence of the maximization of the total entropy for the system protein-water. Other kind of residues that interact with water favorably (are dissolved) are classified into the *polar* class.

The hydrophobic residues are located inside the 3D protein structure while the polar ones face the aqueous environment. From the microscopic point of view, if two successive

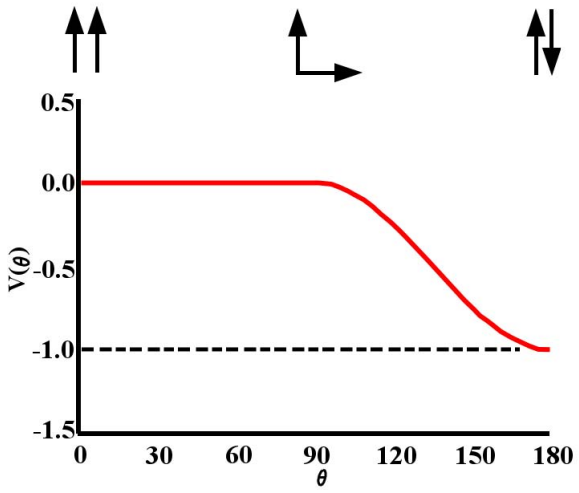


Figure 3.7: Plot of Eq. 3.16 without ϵ_{XY}. The arrows above are the vectors defined by C^{α} to the C^{β}-atom in the residues. X and Y are any two residues.

side chains belong to the same class, either they tend to bury themselves inside or expose themselves on the surface of the protein. This result in an attractive interaction between them. Hence, two residues in the same class tend to minimize the potential when getting together. On the other hand, if two residues are not in the same class there is an unfavorable contribution to the potential and a repulsive interaction between them arises. We write the the local hydrophobic interaction as,

$$E_{LocalHP} = \sum_{i,i+1} V_{XY}(r_{i,i+1}) \tag{3.14}$$

where the sum runs over all two successive side-chain units and $r_{i,i+1}$ is the distance between them. V_{XY} varies depending on the side-chain properties, the curve of V_{XY} is displayed in Fig. 3.6. If the interactions are between two hydrophobic, two polar or two different charged residues, the potential is written as,

$$E_{LocalHP} = \begin{cases} -1 & \text{when } r < \sigma_{local} \\ 0.5\left[-1 - \cos\left(\pi \frac{r-\sigma_{local}}{0.5}\right)\right] & \text{when } \sigma_{local} < r < \sigma_{local} + 0.5 \\ 0 & \text{when } r > \sigma_{local} \end{cases} \tag{3.15}$$

On the other hand if the interaction occurs between a hydrophobic and a polar residues or between residues with the same electrical properties, the potential is given by,

$$E_{LocalHP} = \begin{cases} 0.5\left[\cos\left(\pi\frac{\theta-90}{90}\right)\right] & \text{when } S_i \cdot S_{i+1} \leq 0 \\ 0 & \text{when } S_i \cdot S_{i+1} > 0 \end{cases} \tag{3.16}$$

where the S_i is the vector pointing from the C^{α} to the C^{β}-atom in the ith residue. V_{XY} is depicted in Fig. 3.7. As the reader has probably observed, the local hydrophobic interaction is given in two forms, Eqs. 19 and 20, depending on the kind of amino acids involved. We can see that Eq. 19 shows a distance-dependent behavior while Eq. 20 shows an orientation-dependent behavior, both terms for the local hydrophobic interactions were obtained by fitting the experimental data over hundreds of structures.

In order to simulate the confinement of a protein moduled by, for instance, a chaperone molecule, we add a further term to the potential. In the present work we use two different kinds of spherically symmetric potentials depending on a radius R_c, which is a measure of the size of the cage. In a first approach, we use an external potential $V_1(r)$ which allows the protein to fold freely for distances r smaller than R_c, but has a strongly repulsive part for larger distances, which simulates the presence of the walls of the cage. The potential $V_1(r)$ reads [RKP05]

$$V_1(r) = \frac{0.01}{R_c}\left[e^{r-R_c}(r-1) - \frac{r^2}{2}\right], \tag{3.17}$$

where $r = |\vec{R}|$ denotes the position of each residue.

Since $V_1(r)$ might represent a too simple description of the confining potential of a chaperon, we also investigated the effect of an improved external potential $V_2(r)$ simulating attractive walls [LLW06], which reads

$$V_2 \;=\; 4\epsilon_h\frac{\pi R_c}{r}\left(\frac{1}{5}\left[\left(\frac{\sigma}{r-R_c}\right)^{10} - \left(\frac{\sigma}{r+R_c}\right)^{10}\right]\right.$$

$$-\frac{\epsilon}{2}\left[\left(\frac{\sigma}{r-R_c}\right)^4-\left(\frac{\sigma}{r+R_c}\right)^4\right]\Bigg). \tag{3.18}$$

The physical meaning of the different parameters in Eq. (3.18) can be described as follows. A uniform distribution of beads spreads out on the surface of the cage with a number density $1/\sigma^2$. The parameter ϵ is used to simulate the degree of attraction of the inner surface of the cage. A wall with a purely attractive lining has a value of $\epsilon = 1$ whereas a purely repulsive lining has a value $\epsilon = 0$. In Eq. (3.18) we set $\epsilon_h = 1.25$ kcal/mol and $\sigma = 3.8$ Å. The external potential $V_1(r)$ has the only effect of confining the protein inside the cage whereas the external potential $V_2(r)$ not only confines but also interacts with it by slightly reducing its energy as ϵ increases. As a consequence, the residues tend to be far apart of each other in the region close to the walls of the cage. The curves of $V_1(r)$ and $V_2(r)$ are presented in Fig. 3.8, we can observe that while $V_1(r)$ is purely repulsive, $V_2(r)$ shows some attraction at the barrier surface (15 Å) and that this attraction can be controlled with the parameter ϵ.

3.1.2 Model II

The second model used to describe the dynamics of protein folding considers less interactions than the model I described in the previous section. Originally this model was described by Clementi et. al [CMB98], and it has been demonstrated to yield a good qualitative description of the folding. Several works aimed at describing the folding of proteins have used the Clementi potentials [Ark08, HRL08, CMB98, Cle08, SBJ07, MC06a].

Within the framework of the Clementi potential, a protein is modeled as a sequence of N beads in the 3D space. Each bead represents a C^α in a real protein. The interaction between amino acids is given by,

$$U_{ij} = \delta_{i,j+1}a(r_{ij}-r_0)^2 + (1-\delta_{i,j+1})4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], \tag{3.19}$$

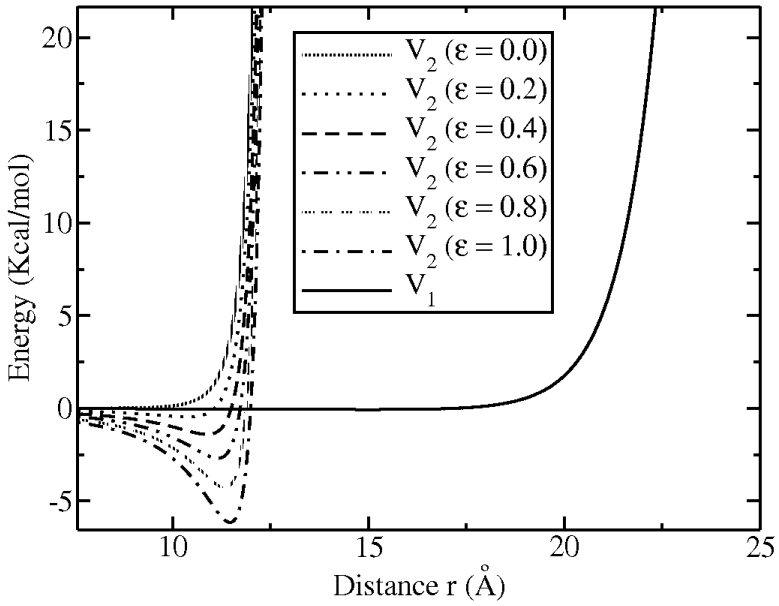


Figure 3.8: Curve of $V_1(r)$ (solid line) and $V_2(r)$ for different values of the parameter ϵ. $\epsilon = 0.0$ means a purely repulsive barrier and $\epsilon = 1.0$ a barrier highly attractive. The minimum of the potential $V_2(r)$ is localized near the surface of the barrier. The radius of the barrier is 15 Å.

where $a = 50$ Å^{-2} and equilibrium distance $r_0 = 3.8$ Å. In the present work we study chains with $N = 30$ monomers. Three different kinds of sequences were studied, which differed in the value of the parameters ϵ_{ij} and σ_{ij}. In the first sequence, known in the literature as Homopolymer (HMP) [HRL08], all the residues are equal, meaning that the parameters $\epsilon_{ij} = 10$ and $\sigma_{ij} = 6.5$ Å are chosen equal for all i and j. One expects for this case that the Potential Energy Surface has many local minima very similar to the global minimum and therefore that a protein with these characteristics should not have the stable native state. Instead, it would jump from one minimum to another with relative facility. This is the characteristic of a bad folder.

The second sequence studied was a Designed Heteropolymer (DHTP). This sequence was originally designed to exhibit a global minimum much deeper than any other local minimum. The DHTP should have a stable native state which is characteristic for a good folder. The model parameters for the DHTP have values in the range $0.25 < \epsilon_{ij} < 10$

and $5 < \sigma_{ij} < 17$ Å.

Both sequences, mentioned above have been studied previously [HRL08, CMB98]. It was found that the HMP possesses a very rough energy landscape with several local minima which are not distinguishable from the global minimum. This can be observed in the energy landscape as a function of the Root Mean Square Displacement in Fig. 3.9 a). In contrast to the energy surface for the HMP, the energy surface for the DHTP is very smooth and exhibits a global minimum which is deeper as any other local minima. Both factors make the global minimum stable and attainable in a relative short time. The energy landscape of the DHTP can be observed in Fig. 3.9 b).

Finally, the third sequence treated in this work is the so-called Random Heteropolymer (RHTP), in which $\sigma = 6.5$ Å is a constant value and the ϵ_{ij}'s have the same range of values as for the DHTP model. The values for ϵ_{ij} are distributed randomly. The energy landscape for the RHTP is very rough as in the case of the HMP and therefore these sequences behave as bad folders.

3.1.3 Reaction Coordinates

In order to be able to understand the folding process and make the study computational possible we need to find suitable parameters to reduce the description of the whole dynamics in terms of a few variables. Many parameters have been proposed for such purpose like, for instance the number of native contacts. However this parameter has the disadvantage of being a discrete quantity. A continuous parameter would be more suitable for our description in terms of the Density of States. One of the continuous parameters commonly used in the literature is the end-to-end distance, which is calculated between the initial and final C^{α} atoms in the protein structure,

$$Q = |r_{C^{\alpha}_{inicial}} - r_{C^{\alpha}_{final}}|. \tag{3.20}$$

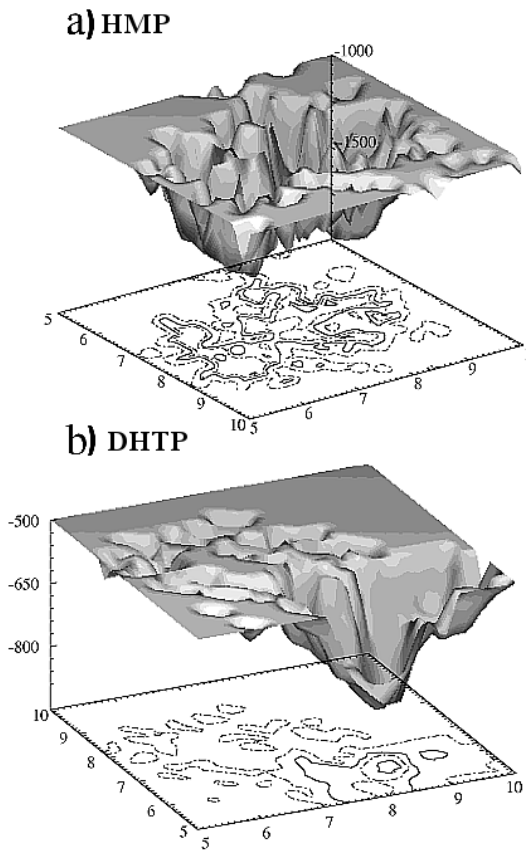


Figure 3.9: The rugged energy landscape of the HMP a) compared to the smooth landscape of the DHTP b). Observe that the DHTP has a very deep global minimum which corresponds to the native state. Pictures are derived from the conformations obtained during numerous dynamical runs of slow cooling. The energy of each conformation is plotted as a function of its distance from two fixed "reference" conformations. *Taken from [CMB98].*

Another parameter frequently used is the configurational energy E, which is the self-energy of a certain structure. The calculation of this parameter does not represent an additional CPU time because it is a by-product of the simulation.

Other reaction coordinates commonly used in the literature, but not treated in this Thesis are the path variables, the normal modes, the helicity of the backbone, among others. The reader is referred to Ref. [LG08] for more details about the different conformational parameters commonly used in Biomolecules.

3.2 COMPUTATIONAL ALGORITHMS

Because of the many-body character of proteins, the dynamics of the folding does not have an analytical solution. Therefore, one is forced to used computer simulations in order to study proteins and biomolecules in general. In the present Section we study two algorithms which allow us to solve the dynamics and the thermodynamics of the folding. Section 3.2.1 is dedicated to the Wang-Landau algorithm, which is a Monte Carlo method that calculates the Density of States "on the Fly" and it is suitable for complex PESs as in the case of Proteins. The Wang-Landau algorithm belongs to the kind of Stochastic Methods, because they are based on the sampling using random variables. The Markovian and Non-Markovian processes are the basis of such methods. Section 3.2.2 describes the Langevin Dynamics in the Overdamped limit, which is adequate for biomolecules because of the mass and the velocities of the atoms. In contrast to the Wang-Landau algorithm, the Langevin Dynamics is one of the so-called deterministic methods, because in this case there exists an equation which guides the evolution of the system given a set of initial conditions. Finally, the section 3.2.3 is devoted to the development of the Rate of Convergence concept. This concept will help us to study the problem of sequence design in Chapter 4.

3.2.1 Wang-Landau Algorithm

The problem that concerns us now is the determination of the thermodynamic properties of a protein that could be measured in the laboratory. The simplest method to investigate the Space of Configurations (SC) of a protein and then obtain the thermodynamic properties consists of using a canonical ensemble, where the configurations at temperature T are weighted by the Boltzmann factor,

$$P_B(E) = e^{-\beta E}, \beta = 1/k_B T. \tag{3.21}$$

where $k_B = 1.987 \times 10^{-3}$ Kcal/molK is the Boltzmann constant.

The resulting probability distribution is therefore,

$$\rho(E) \propto g(E)P_B(E), \tag{3.22}$$

where $g(E)$ is the Density of States (DOS). Because $g(E)$ is a rapidly increasing function of E and the Boltzmann factor $P_B(E)$ decreases exponentially $\rho(E)$ has in general a bell-like shape. At a finite temperature, the value of $\rho(E)$ for low E is smaller by many orders of magnitude than the maximum value of $\rho(E)$, this fact makes the sampling of the SC not uniform. Some regions are over-explored and other regions are not explored at all. In systems with a complex potential energy surface, as in the case of proteins, the weight factor $P_B(E)$ is practically incapable of sampling correctly the SC. Therefore, one cannot obtain a good estimation of the thermodynamic properties. In order to avoid these problems regarding the sampling of the SC, F. Wang and D. Landau [WL01] proposed to use a probability distribution defined in such a way that a configuration with any energy is accepted with a uniform probability:

$$\rho(E) \propto g(E)P_{WL}(E) = \text{const.} \tag{3.23}$$

Then, it follows that the Wang-Landau weight factor should have the form, $P_{WL}(E) \propto g^{-1}(E)$. Choosing in this manner the weight factor, all the conformations are the equally likely and therefore the SC is sampled in detail and the thermodynamic properties are more trustable than in the classical canonical ensemble. So far, there would not be any problem in using the exact DOS to perform the sampling because with it the examination of the SC would be optimal. Nevertheless the real problem is that the exact DOS is not known in general for many models of interest, let us quote the case of clusters, polymers, proteins, spin systems among others.

The novel contribution of Wang and Landau was the design of an algorithm that allows us to calculate the DOS during the simulation process without having any previous knowledge of it. The utility of this method has been observed in a variety of models which study properties of molecules [RKP05, OLYG09, TPB09], magnetic systems [DCJP04], quantum systems [CLST09] and numerical solutions of integrals [BMP08].

Other methods based on the philosophy of the Wang-Landau algorithm, have been proposed to compute the thermodynamical properties of finite systems. They include, for instance, multi-canonical simulations [BN91], simulated annealing [KGV83] and Parallel Tempering (Replica Exchange) [SW86]. In contrast to alternative sampling methods, the Wang-Landau almost seems to be a universal approach, because it does not rely on working out a good range of energy or distribution function to sample in advance [ZS08]. One of the main advantages of Wang-Landau simulations is that they allow to obtain directly the DOS of the system, which is, of course, independent of the simulation temperature. Once the DOS is known, one can obtain all the thermodynamical properties of the system at any temperature. As a remark, the convergence of the Wang-Landau algorithm is guaranteed upto a factor proportional to $\sqrt{f}$, f being the modification factor (see below).

Within the Wang-Landau framework, the transition probability between two conformations before and after a MC trial move, $\mathbf{X_1}$ and $\mathbf{X_2}$ respectively, is calculated as

$$P(\mathbf{X_1} \rightarrow \mathbf{X_2}) = \min\left[1, \frac{g(\mathbf{X_1})}{g(\mathbf{X_2})}\right], \tag{3.24}$$

where $g(\mathbf{X})$ is the DOS of the system and $\mathbf{X}$ is a generalized reaction coordinate, which in our case is represented by a vector with many possible entries $\mathbf{X} = (K_1, K_2, ..., K_N)$, and K_i being a characteristic parameter of the system. In our case we employed only two parameters: E, the configurational energy and Q, the end-to-end distance of the protein structure. Then $\mathbf{X} = (E, Q)$.

The original scheme developed by Wang and Landau can be briefly described as follows: one sets the initial function $g(\mathbf{X})$ together with an auxiliary histogram $H(\mathbf{X})$ to be equal to 1. Then, each time the bin $\mathbf{X}$ is visited, one updates the histogram $H(\mathbf{X})$ and modifies $g(\mathbf{X})$ as $g(\mathbf{X}) \rightarrow g(\mathbf{X}) \times f$, with $f = e = 2.718281...$. This procedure is continued until a "flat" histogram is obtained, that is when the fluctuations in the histogram are relatively small compared to the average of the histogram. The average of the histogram is defined below. At this step the histogram $H(\mathbf{X})$ is reset and the factor f is reduced. The usual way to perform this reduction is by taking $f_{i+1} = \sqrt{f_i}$. The convergence of the algorithm is achieved when a value for f_{i+1} close enough to 1 is obtained. The last step must be compatible with the desired accuracy, for example $f = \exp(10^{-7})$. As a remark, the Wang-Landau algorithm is a non-Markovian Monte Carlo method, because the transition probabilities are changing during the simulation. Therefore, we take as the DOS of the system, the function $g(\mathbf{X})$ at the final run.

We use the modified Wang-Landau approach proposed in Ref. [BP07], which has been shown to speed up simulations and to partially avoid the problem of saturation error. According to the new scheme, one does not need to wait until the histogram $H(\mathbf{X})$ is "flat", but it is enough to require that all the entries of $H(\mathbf{X})$ are visited. Then $H(\mathbf{X})$ is set to zero and $f_{i+1} = \sqrt{f_i}$ is updated.

Following Ref. [BP07] we employ a second histogram $H_2(\mathbf{X})$ which is never reset during the whole simulation and define the Monte Carlo time-step as $t = j/N$, N being the number of points in the energy axis and j the number of trial moves performed. If $f_{i+1} \leq t^{-1}$ then $f_{i+1} = f(t) = t^{-1}$ and from this point on $f(t)$ is updated at each Monte Carlo time step. $H(\mathbf{X})$ is not used during the rest of the calculation. The convergence is achieved when $f(t) < f_{final}$. In the present simulations we used $f_{final} = \exp(10^{-7})$.

Application to the 1D Harmonic Potential

Let us take a simple case to exemplify how the Wang-Landau algorithm works. We consider for example the potential of the 1D harmonic oscillator $V(x) = x^2$. The exact density of states in this case is calculated easily and it is given by,

$$g(E) = \int_{-\infty}^{\infty} \delta(V(x) - E)dx = \int_{-\infty}^{\infty} \delta(x^2 - E)dx = \int_{-\infty}^{\infty} \frac{\delta(y)dy}{,} 2(y+E)^{\frac{1}{2}} \tag{3.25}$$

where we have made the substitution $y = x^2 - E$ and $dx = dy/2(y+E)^{\frac{1}{2}}$. Finally we obtain,

$$g(E) = \frac{1}{2E^{\frac{1}{2}}}. \tag{3.26}$$

In the present 1D case we have obtained that the DOS is proportional to E to the power 1/2. The potential $V(X)$ and the exact DOS are shown in Fig. 3.10. In the most general case where we have a n-dimensional oscillator, the DOS changes with E according to,

$$g(E) \propto E^{\alpha}. \tag{3.27}$$

Now the problem that concerns us is the computational implementation of the Wang-Landau algorithm. Essentially, one discretizes the parameter under consideration, for instance the energy E, in bins up to a certain precision $\Delta E = (E_{max} - E_{min})/N$, where E_{max} and E_{min} are respectively the maximum and minimum of the energy range considered and N is the required number of bins. The discretization of the energy is illustrated in Fig. 3.11. Due to the rapid growth of $g(E)$, the update criterion $g(E) \rightarrow g(E) \times f$ looses sense, since $g(E)$ turn rapidly into not manageable numbers for the computer. This small difficulty is solved by considering the logarithm of g(E) ($\ln g(E)$) instead of $g(E)$. In this way the multiplications turn into sums. In what follows $\widetilde{g(E)}$ will refer to $\ln g(E)$.

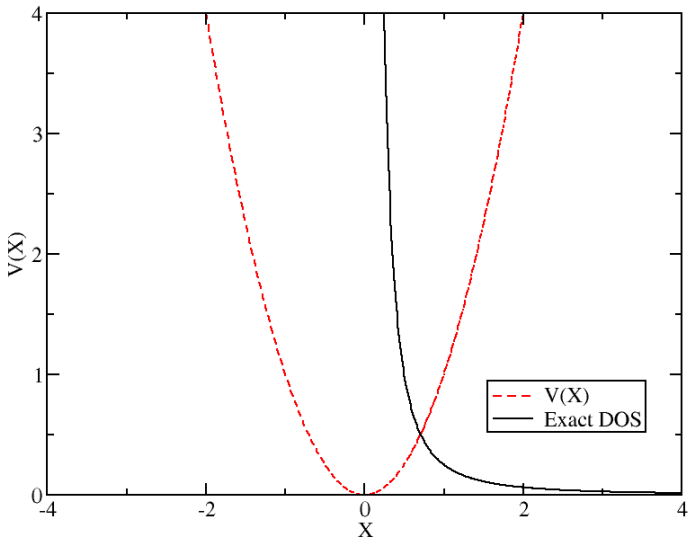


Figure 3.10: Scheme of the Harmonic Oscillator potential (black-dashed line) $V(X)$ and its exact DOS (red-solid line).

Then, during the whole simulation, one stores number of visits to each energy bin E_n in two arrangements $g(\widetilde{E_n})$ and $H(E_n)$. If we have planned to use the *1/t*-algorithm one more histogram $H_2(E)$ will be necessary. $g(\widetilde{E})$ is updated with the same parameter f until a certain criterion is reached. The most common one is to wait until de fluctuations of the histogram are less than a threshold value, for instance

$$\frac{< H^2(E) - < H(E) >^2 >}{< H(E) >} \leq 0.3. \tag{3.28}$$

here $< H(E) >$ represents the average of the histogram. It is given by $\sum_i H(E_i)/N$, where $H(E_i)$ is the number of visits to the energy bin E_i and N the total number of bins considered in the simulation. In a similar way one can define $< H^2(E) >$. After this criterion is satisfied, the modification factor is updated by $\ln f_{i+1} \rightarrow 0.5 \times \ln f_i$ and the histogram $H(E)$ is set to zero again. The whole simulation is stopped when $\ln f$ is sufficiently small, say $\ln f_i \sim 1 \times 10^{-8}$. The $1/t$ algorithm serves to reach a better convergence, it uses an extra histogram $H_2(E)$ which is never rever reseted during the simulation. Within this framework one waits until $\ln f$ is relatively small (say $\ln f \sim 10^{-5}$)

and then use the update criterion $\widetilde{g(E)} \rightarrow \widetilde{g(E)} + 1/t$, where $t =< H_2(E) > /N$.

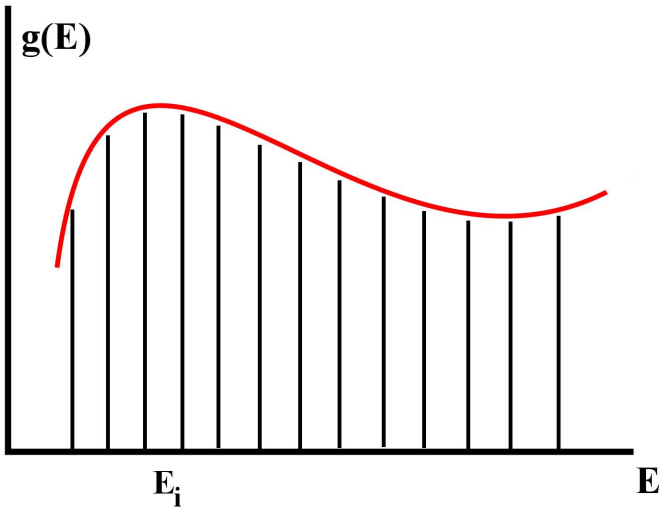


Figure 3.11: Discretization of the DOS in energy bins E_i. At each Monte Carlo Step (MCS) the DOS is updated as $\widetilde{g(E_i)} \rightarrow \widetilde{g(E_i)} + \ln f$.

By using the Wang-Landau scheme together with the *1/t*-algorithm we obtain the relation $g(E) = 1/2E^{1/2}$ as in the analytical solution in Eq. 30 except for a normalization constant. Both the analytical and simulated $log[g(E)]$ are shown in Fig. 3.12 at different stages of the simulation. We stopped the simulation when the modification factor was $\sim 10^{-6}$ and the average error $< \epsilon >= 10^{-5}$.

Once we have obtained the DOS of the system, we can calculate straightforwardly the thermodynamical properties such as the free energy $F(T)$, internal energy $U(T)$, entropy $S(T)$ and specific heat $C(T)$. For the specific case treated in this work, where $\mathbf{X} = (E, Q)$ the corresponding definitions read

$$F(T) = -k_B T \ln \left(\int dE \int dQ \, g(E,Q) \, e^{-\beta E} \right), \tag{3.29}$$

$$U(T) = \langle E \rangle_T = \frac{\int dE \int dQ \, E g(E,Q) \, e^{-\beta E}}{\int dE \int dQ \, g(E,Q) \, e^{-\beta E}}, \tag{3.30}$$

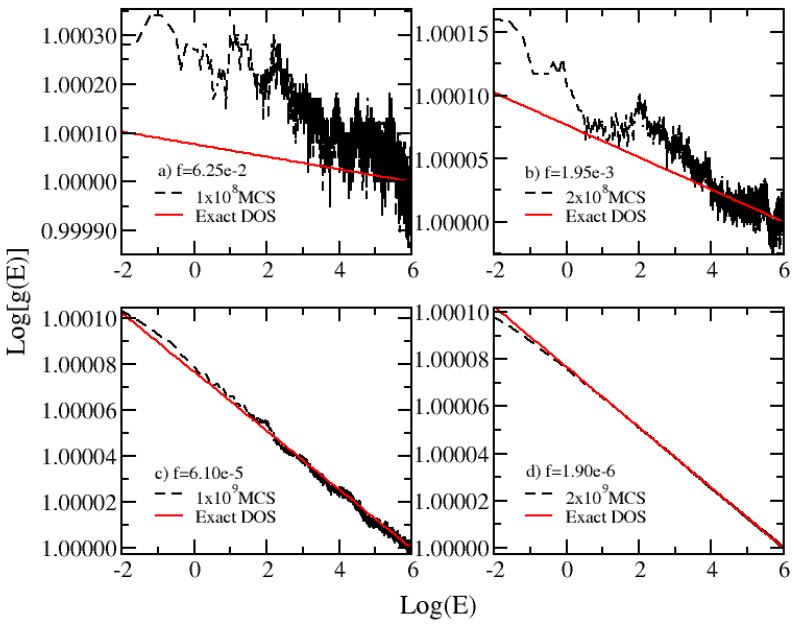


Figure 3.12: Logarithm of the exact DOS (red-solid line) and the simulated DOS (black--dashed line) at different stages of the simulation. At the beginning of the simulation (a-b) the simulated DOS shortly differ from the exact one but after 1×10^9 Monte Carlo Steps (MCS) the simulated DOS converges to the exact DOS (c). At 2×10^9 MCS the simulated DOS has already converged (d).

$$S(T) = \frac{U(T) - F(T)}{T}, \tag{3.31}$$

$$C(T) = \frac{\langle U^2 \rangle_T - \langle U \rangle_T^2}{k_B T^2}, \tag{3.32}$$

where $\beta = 1/k_B T$ and k_B is the Boltzmann constant. The free energy landscape as a function of E and Q can be computed as

$$F(E,Q) = -k_B T \ln \left(\frac{g(E,Q)\, e^{-\beta E}}{\int dE \int dQ\, g(E,Q)\, e^{-\beta E}} \right), \tag{3.33}$$

and

$$F(E) = -k_B T \ln \left(\frac{\int dQ g(E,Q)\, e^{-\beta E}}{\int dE \int dQ\, g(E,Q)\, e^{-\beta E}} \right). \tag{3.34}$$

It is important to notice that the free energy landscape described by $F(E,Q)$ gives us information not only about the native and the unfolded states but also about the presence of metastable or intermediate states. These are important regarding the folding because they can make the process faster or slower depending on the amino acid sequence and the external factors such as the temperature. We will make use of the free energy landscape in Chapter 4 when we discuss about the modifications in the intermediate states of proteins.

3.2.2 Langevin Dynamics Algorithm

The Langevin Dynamics is an approach to describe the evolution in time of molecular systems. It was originally developed by Paul Langevin. In this approach some degrees of freedom are neglected by making use of stochastic differential equations [Sch02, Str05]. We explain the main facts in the Langevin Dynamics approach below.

At the microscopic level, the Reynold's number plays a crucial role since the velocity of the atoms becomes much more important than the acceleration [Wai07]. At this scale of lengths the force acting on the atoms is proportional to the velocity and one is in a regime known as *overdamped*. In addition to fact, there are forces of impact coming from molecules in the environment. These forces exhibit a random behavior and they give place to the so-called Brownian Motion. Langevin dynamics takes into account the random character of the environment implicitly. In this way, one is able to simulate the dynamics of proteins in solvent without requiring a long CPU time. Langevin dynamics controls the temperature of the system like a thermostat, therefore it acts in the canonical ensemble. Following this scheme the protein follows a dynamics guided by the Langevin's

equation

$$m\ddot{\vec{r}_i} = \vec{f}_i - \gamma\dot{\vec{r}_i} + \vec{\eta}_i, \tag{3.35}$$

for each of the N residues. Here $\vec{r_i}$ is the position of the residue i and $\vec{f_i}$ is the deterministic force acting on it, γ is the effective friction and $\vec{\eta_i}$ is a random force which simulates the liquid environment and which has to fulfill two conditions,

$$< \vec{\eta_i} >= 0, \tag{3.36}$$

$$< \eta_{i,\alpha}(t)\eta_{j,\beta}(t') >= 2\gamma k_B T \delta_{i,j}\delta_{\alpha,\beta}\delta(t-t'), \tag{3.37}$$

with cartesian coordinates $\{\alpha, \beta\} = \{x, y, z\}$, t and t' being to given times. The diffusion equation is straightforwardly obtained as

$$< (\vec{r_i}(t) - \vec{r_i}(0))^2 >= \frac{6k_B T}{\gamma} t = 6Dt, \tag{3.38}$$

D being the diffusion coefficient.

For biomolecules in water solution the Langevin equation can be simplified. Because the Reynolds number is of the order of 10^{-3} we are in the strongly damped or overdamped regime. This value of the Reynolds number means that the viscous friction can stop a protein in a distance of around 10^{-3}nm, much shorter than the size of the atoms. The Langevin equation can then be reduced to,

$$\gamma\dot{\vec{r}_i} = \vec{f}_i + \vec{\eta}_i. \tag{3.39}$$

This equation is integrated by means of the Euler algorithm in the following form,

$$\vec{r_i}(t + \Delta t) = \vec{r_i}(t) + \frac{1}{\gamma}\left[\Delta_t \vec{f_i} + \sqrt{2k_B T \gamma \Delta_t}\hat{\eta}_i\right], \tag{3.40}$$

$k_B = 1.987 \times 10^{-3}$ Kcal/mol K. The units of the energy are given in Kcal/mol and the units of time in picoseconds.

3.2.3 Distance between Configurations and Rate of Convergence

In the Chapter 4 we will make use of the concept of "distance" between configurations, therefore we will explain in the following lines what this concept refers to and why it is important. We will also explain the concept of the Rate of Convergence, which will be useful in order to classify good and bad folders.

We are interested now in defining a parameter which allow us to monitor the folding process and also which could distinguish between the good and bad folders. To this aim we developed in our group the concept of distance between structures [GG]. The distance between two arbitrary structures C_1 and C_2, of a given amino acid sequence S_i, is defined as,

$$\mathcal{O}(\mathcal{C}_1, \mathcal{C}_2) = \sum_{i,k=1}^{N} \Delta_{ik}(\mathcal{C}_1)\Delta_{ik}(\mathcal{C}_2), \tag{3.41}$$

where the sum runs over all pairs of monomers i, k and Δ_{ik} is an intrinsic property of the protein. For lattice models Δ_{ik} denotes the contact map of configuration $\mathcal{C}$ [VND99], that is $\Delta_{ik} = 1$ if monomers i and k are in contact and $\Delta_{ik} = 0$ otherwise. Remember that two monomers are said to be in contact if the distance between them is less than a certain predefined value, in our calculations we take 8 Å. For off-lattice models where the distance between consecutive monomers is not the same, as in the Model II of this Chapter, the former definition of Δ_{ik} lack of sense. We found that for off-lattice models one possible way to define the Δ_{ik} is by taking the negative part of the Lennard-Jones potential between two non-consecutive monomers, that is,

$$\Delta_{ik}(\mathcal{C}) = -\min\left[0, 4V_{LJ}(r_i, r_k)\right], \tag{3.42}$$

with $V_{LJ}(r_i, r_k)$ being the LJ potential between the monomers i and k which have the coordinates r_i and r_k respectively. The factor 4 is just to normalize $V_{LJ}(r_i, r_k)$ to -1. In fact, this is a generalization of the contact map for lattice models because $\Delta_{ik}(\mathcal{C})$ is equal to 1 only when the monomers i and k are in the equilibrium distance of the LJ potential, otherwise when they are farther or closer we obtain in general $0 < \Delta_{ik}(\mathcal{C}) < 1$. Note that the more compact and structurally similar two configurations are the larger is the distance between them.

Having defined the concept of distance or overlap between configurations we introduce the concept of Rate of Convergence. First of all we fix a time scale t_0, which is longer than the typical time required for the dynamically evolving chain to overcome local minima on the energy surface. We then let a given amino chain start from two randomly chosen configurations $\mathcal{C}_1$ and $\mathcal{C}_2$ and dynamically propagate it from these positions over the time t_0. We denote the resulting configurations as $g^{t_0}\mathcal{C}_1$ and $g^{t_0}\mathcal{C}_2$ and the overlap between them as

$$R(t_0, T) = \mathcal{O}(g^{t_0}\mathcal{C}_1, g^{t_0}\mathcal{C}_2), \tag{3.43}$$

where T is the temperature (the dependence on T is hidden in the dynamical transformation). Note that $R(t_0, T)$ gets a positive contribution from the terms in (3.41) if and only if the monomers i and k find themselves in contact in both chains $g^{t_0}\mathcal{C}_1$ and $g^{t_0}\mathcal{C}_2$. Sampling over several randomly chosen initial configurations $\mathcal{C}_1$ and $\mathcal{C}_2$ we calculate the average $\langle R(t_0, T)\rangle$, which we call it the Rate of Convergence for the amino acid sequence S. The rate of convergence can be ascribed to any amino acid sequence and the larger $\langle R(t_0, T)\rangle$ the better are the chances for this sequence to be a good folder. If one needs to find the best candidates for being a good folder from a number of given amino acid sequences one can sort all sequences by their rate of convergence. The degree to which this sorting algorithm is effective depends on how t_0, which is sufficient for sorting, relates to the mean folding time. The algorithm consists in computing the rate of convergence

at a given temperature for various amino acid sequences and then ordering the sequences by this value.

Chapter 4

SIMULATION RESULTS AND ANALYSIS

In this Chapter we present the results of our simulations. In Section 4.1 we discuss about the confinement of proteins in potential barriers. Section 4.2 describes the effects of electric fields on proteins. Finally, Section 4.3 is devoted to the sequence design problem.

4.1 EFFECT OF CONFINEMENT ON THE INTERMEDIATE STATES OF A PROTEIN

In this Section we focus on the problem of protein folding assisted by Chaperones, which is one of the mechanisms present in nature to avoid aggregation and misfolding. Chaperones are molecules in the form of a cage inside which proteins fold correctly. Recently, some progress has been achieved in the understanding of the folding of proteins inside chaperones. These studies have shown that stability and folding kinetics are strongly correlated with the geometry and the degree of confinement inside the cage [TKT03, TKL03, RKP05, NSC06, JBS04, FS06]. However, many details of the folding under confinement still remain uncovered.

Here we consider the folding of the peptide V3-loop, Protein Data Bank ID 1NJ0, and analyze it under two kinds of time-independent confining potentials. The first potential simulates a cage being composed by rigid walls, while the second potential describes a cage with an attractive inner surface. The influence of both potentials is reflected in the thermodynamical properties, which we calculate using the Wang-Landau algorithm [WL01, BP07] described in Section 3.1.

As one of the main results of this work we obtain that the folding process of V3-loop

occurs through metastable intermediate states [SBJ07], and that the presence of those states can be controlled by the confining potential.

For the description of the protein we use the force field of Model I (see Section 3.2) which does not depend on the previous knowledge of the native structure and is also able to describe folding of proteins into both helices and β-sheets with the same set of parameters [CSM06].

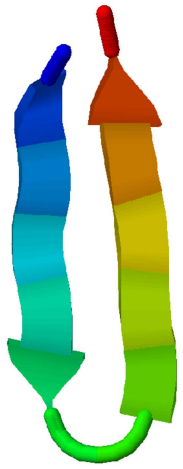


Figure 4.1: Ground-state structure (β-sheet) of the peptide 1NJ0 ($E_g \sim -135$ Kcal/mol).

We focused our attention on a peptide composed of 16 amino acids with PDB code 1NJ0 to study the folding mediated by confining potentials. This peptide conforms the V3-loop of the exterior membrane glycoprotein (GP120) of the Human Immunodeficiency Virus type 1 (HIV-1).

To explore the relevant part of the phase space of the protein we have chosen an energy window between -135.0 kcal/mol and -30 kcal/mol and the end-to-end distance Q ranging from 5 Å to 50 Å. This region is enough to cover both the highly ordered structures (present at $T \sim 0$) and the fully disordered random coils (stable for $T \sim \infty$). The MC search was generated by changing each pair of Ramachandran angles ψ_i and ϕ_i at each MC step using cutoffs with values $|\Delta\psi_c| \leq 40^o$ and $|\Delta\phi_c| \leq 40^o$. In order to

Temperature (K)	Radius (Å)
329.2	15
323.4	20
323.2	25
321.0	∞

Table 4.1: Transition temperatures T_f for different values of the radius R_c of the potential $V_1(r)$ (see main text). Note that T_f decreases for increasing R_c. T_f is the temperature at which the specific heat is a maximum.

reach $f_{final} = \exp(10^{-7})$ 8×10^9 trial moves were necessary.

We first analyze the properties of the peptide without confinement (bulk case). The obtained ground state structure of the V3-loop is depicted in Fig. 4.1. It consists of a $\beta-$sheet structure with energy ~ -135.0 Kcal/mol and an end-to-end distance of $\sim$5.5 Å.

A new feature described by our force field is the presence of intermediate structures between the native (N) and the unfolded (U) states as shown in Fig. 4.2. We obtain two intermediate states in the free energy profile $F(E, Q)$ at the transition temperature. The intermediates, denoted as I_1 and I_2 in the figure, appear as local minima of $F(E, Q)$. In order to analyze the nature of the intermediate states we have split the energetic and entropic parts of the free energy. Results indicate that the intermediates are mainly stabilized by their energy, but that there is a non negligible entropic contribution.

The next problem to be addressed is the influence of confinement on the free energy landscapes and folding behavior of the V3-loop. For this purpose, we calculated first the DOS and the specific heat of the protein assuming the rigid-wall confining potential $V_1(r)$ described above. We considered different cage-diameters (R_c = 15 Å, 20 Å, and 25 Å). In Fig. 4.3 we show influence of $V_1(r)$ on the behavior of the DOS. Note that, due to confinement, $\log[g(E)]$ considerably decreases at high energies compared to the bulk case ($R_c \to \infty$). For energies close to the ground state, $\log[g(E)]$ does not exhibit any noticeable change because the protein is almost folded. Since its gyration radius in the ground state is $R_g \sim 13$ Å, barriers of radii equal or larger than 15 Å do not

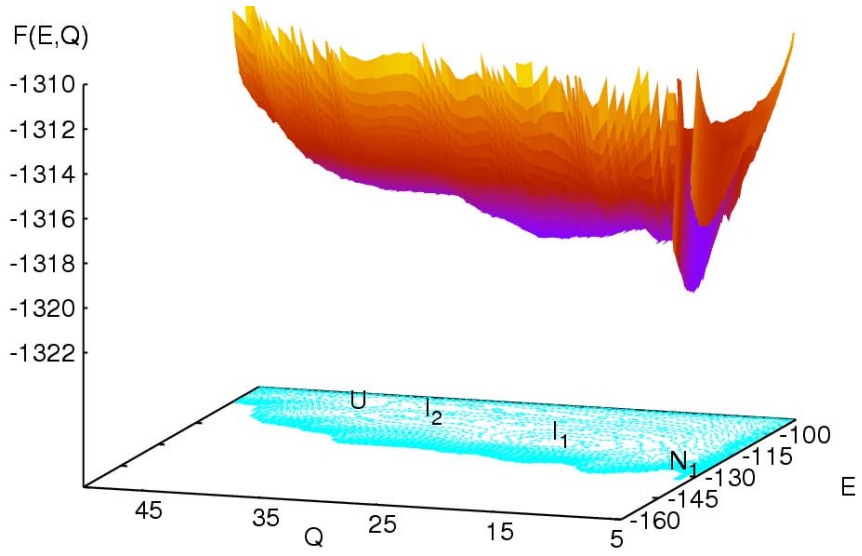


Figure 4.2: Besides the Native State N and the Unfolded U states in the Free Energy Landscape ($F(E,Q)$), there are other two states which are intermediates in the folding process, in the picture they are denoted as I_1 and I_2. $F(E,Q)$ is plotted in terms of the configurational energy E and the End-to-End distance Q.

affect folded structures. This result is consistent with the intuitive picture that cages, for instance chaperones, restrict the otherwise huge phase space for high energies, making the number of available structures, and consequently the entropy, considerably smaller than in absence of a cage.

The effect of confinement can be also observed in the specific heat of the V3-loop, which we show in Fig. 4.4. Here, we plot the specific heat for different values of the cage radius (15 Å, 20 Å, 25 Å) and for the bulk case ($R_c \to \infty$) as a function of T/T_f^0, where $T_f^0 = 321$ K is the transition (unfolding) temperature in absence of a cage. The transition temperature T_f^0 is the temperature at which the specific heat is a maximum. The effect of the rigid-wall potential $V_1(r)$ is to increase the transition temperature (see Table 4.1) and to make the curve of the specific heat broader as the radius of the cage decreases. A broader curve means that there are more structures with energies close to the native state than in the bulk-case where only the native state is the most important compact structure. For radii larger than 25 Å the transition temperatures are equal to T_f^0 within

the statistical error of our simulations. We conclude that the protein is more stable as the radius of the cage decreases. This results are in agreement with Ref. [RKP05], in which Monte Carlo simulations were used, and with Refs. [TKT03] and [LLW06], where Langevin simulations were performed. It is important to mention, however, that in those cited simulations a simplified Go-type force field was used. Thirumalai [TKL06] made a considerable improvement to the force field by introducing the effect of the non-native interactions. However, important interactions such as dipole-dipole and hydrogen bonds were not taken into account. The presence of intermediates was not reported either in those studies.

The main goal of this Section is the study of the influence of confinement on the potential landscape and, consequently, on the stability of the native and intermediate states. Note that the intermediates can be better characterized by analyzing the free energy F as a function of both the energy E and the order parameter Q. In Fig. 4.5 a) we show the contour plot of $F(E, Q)$ for the bulk case. Clearly, the two intermediate structures, which we denote as I_1 and I_2 can be identified as local minima of $F(E, Q)$. It is important to point out that the values of the end-to-end distance Q in the intermediates I_1 and I_2 are larger than in the native structure, but smaller than in the unfolded state. Fig. 4.3 shows the importance of choosing the adequate order parameters to plot the free energy.

In order to study the effect of a cage with purely repulsive walls (potential $V_1(r)$) we have determined $F(E, Q)$ for different values of the cage-radius R_c. In Figs. 4.5(b)-(d) we show the corresponding contour plots for $R_c = 25$ Å, $R_c = 20$ Å and $R_c = 15$ Å, respectively. The different minima of $F(E, Q)$ are shown together with representative snapshots of the corresponding structures. The main effect of the cage of $R_c = 25$ Å is to restrict the size of the unfolded states (U), which is reflected in a shift of the local "U"-minimum to a smaller value of Q (see Fig. 4.5 (b)). Further decrease of R_c leads to a

Temperature (K)	Radius (Å)
321.0	BULK
324.2	$\epsilon = 0.0$
324.1	$\epsilon = 0.2$
314.5	$\epsilon = 0.4$
292.1	$\epsilon = 0.6$
253.5	$\epsilon = 0.8$
–	$\epsilon = 1.0$

Table 4.2: Transition temperatures T_f for the confining potential $V_2(r)$ (see main text) for different degrees of hydrophobicity, ϵ= 0.0, 0.2, 0.4, 0.6, 0.8, 1.0 and for the bulk case. Notice that in general T_f decreases as ϵ increases. For $\epsilon = 1.0$ it is not possible to define T_f because the specific heat is almost completely attenuated.

stronger reduction of the size of the unfolded states. For example, for $R_c = 15$ Å the local "U"-minimum is situated at $Q \sim 20$ Å, i.e., which means that the end-to-end distance of the unfolded states has been halved in value with respect to the bulk case. This can also be observed by the change in the form of the unfolded structures shown in the figure. Interestingly, there is also a small shift of the "U"-minimum to lower energies for decreasing R_c. This is simply due to the fact that strong spatial confinement necessarily leads to the formation of contacts which were not present in the bulk. The shift of the "U"-minimum to lower energies also explains the reduction of the DOS upon confinement shown in Fig. 4.3.

In contrast to the unfolded states, the native structure is practically not affected by confinement, at least up to $R_c = 15$ Å, and the position of the "N"-minimum remains almost unchanged (see Fig. 4.5).

Different and interesting features are observed in the behavior of the intermediate states I_1 and I_2 upon repulsive confinement. In the bulk, the minima corresponding to the intermediates are well defined and separated by a potential barrier. Although both the position of the minima and the structure of the intermediates remain unchanged when the radius of the cage is reduced, the depth of the energy minima and the potential

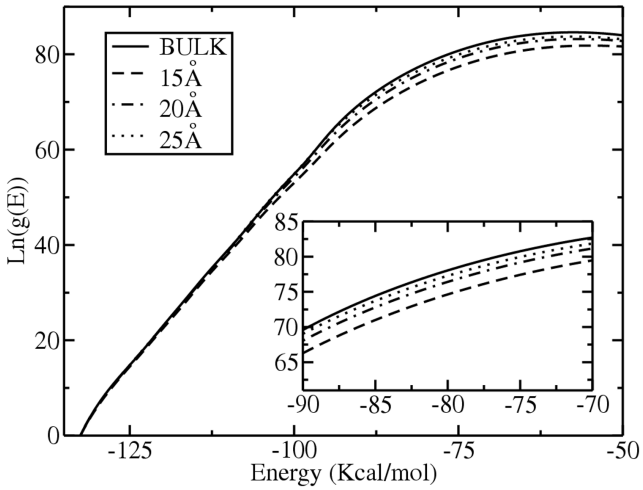


Figure 4.3: Logarithm of the density of states (DOS) $g(E)$ of the protein inside the confining potential $V_1(r)$ and for different values of R_c (15 Å, 20 Å, 25 Å) as well as for the bulk case. One notices the remarkable decrease of the DOS for decreasing R_c.

barrier between them decrease. Already for $R_c = 20$ Å, both minima start to merge and form an extended and shallow minimum. This effect is even stronger for $R_c = 15$ Å. Note that this happens when the radius of the confining potential becomes comparable to the end-to-end distance of the intermediate states in the bulk.

Now, we report on the influence of hydrophobic effects in the inner surface of the cage. Attractive cage-walls were considered by using the confining potential $V_2(r)$ (Eq. 3.18) with radius $R_c = 30$ Å. The degree of attraction is described by the coefficient ϵ. A completely attractive cage-wall is obtained when $\epsilon = 1.0$, whereas $\epsilon = 0.0$ corresponds to a completely repulsive or neutral inner surface of the cage. The effect of ϵ can be visualized in the following way: as ϵ increases from 0 to 1, the walls of the cage tend to attract the residues because of the relative minimum generated by the potential $V_2(r)$. The deepest minimum of $V_2(r, \epsilon)$ is reached when $\epsilon = 1.0$ and corresponds to $V_2^{min} \sim 5$ Kcal/mol. This energy is comparable to the energy required to break one hydrogen bond, $\Delta E_{HB} \sim 4.8$ Kcal/mol. Therefore, for $\epsilon \sim 1.0$ the potential is able to destroy the structure of the protein (denaturation).

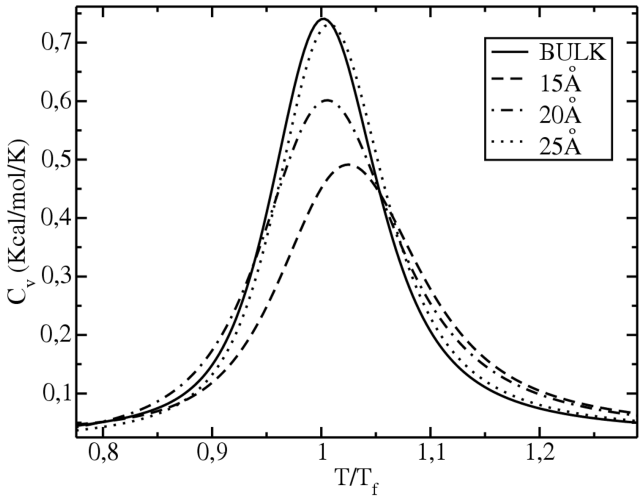


Figure 4.4: Specific heat for the bulk case and for confining potentials with radii 15 Å, 20 Åand 25 Å. $T_f = 321$ K is the transition temperature in the bulk case. T_f increases as the radius R_c decreases. The confining potential in this case is purely repulsive.

The influence of the confining potential $V_2(r)$ on the DOS of the protein is shown in Fig. 4.6, where different degrees of attraction and the bulk case are considered. Two clear features can be distinguished. First, the DOS at energies close to the native state increases for increasing ϵ. This means that the potential landscape is changed near the global minimum. Furthermore, for large energies one can clearly observe a dramatic reduction of $g(E)$ by up to ~ 13 orders of magnitude as ϵ goes from 0 to 1. However, this remarkable reduction of the phase space in this case does not help the protein to fold correctly but forces it to acquire a denatured conformation. This effect occurs because the peptide decreases its energy by placing some of the residues close to the border of the cage. Then, the number of accessible states at those energies decreases and residues are no longer allowed to be far apart from the border, since it would cost much energy. As a consequence, the peptide sticks to the wall of the cage.

The influence of the potential $V_2(r)$ on the specific heat $C(T)$ of the V3-loop is shown in Fig. 4.7. As ϵ increases, the curve $C(T)$ becomes broader. The transition temperatures for different values of ϵ and for the bulk case are presented in Table 4.2. Interestingly,

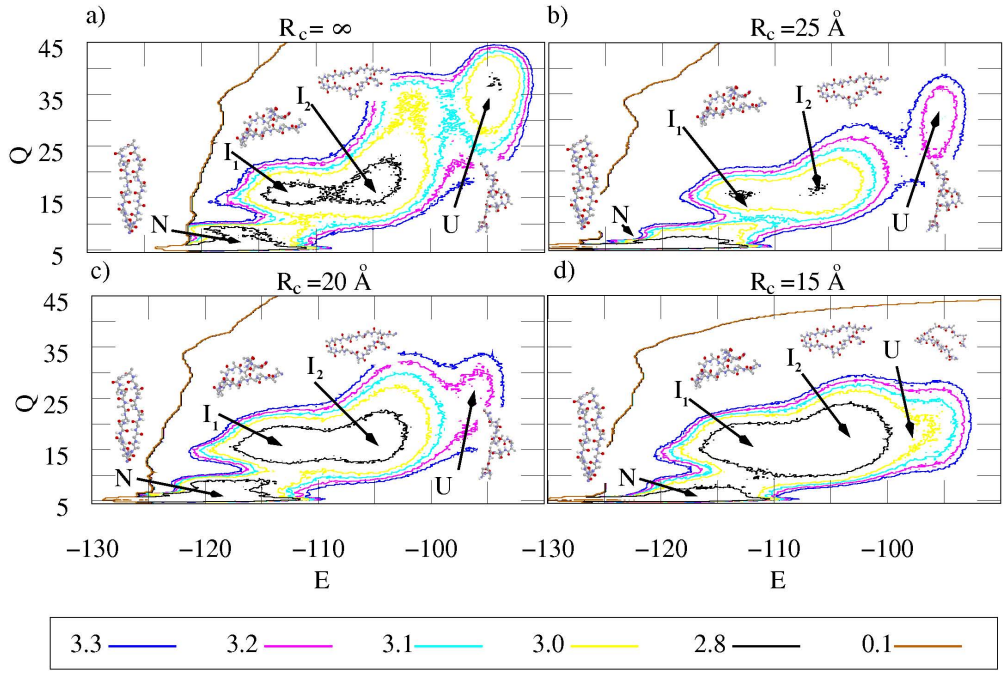


Figure 4.5: Contour plots of the free energy landscape $F(E, Q)$ as a function of the configurational energy E and the end-to-end distance Q for a purely repulsive confining potential. Plots a-d correspond to the bulk case and cages of radius 15 Å, 20 Åand 25 Å respectively. The unfolded state are strongly affected when the size of the cage decreases. The native state and the intermediates are only slightly modified. The contour lines represent the free energy difference with respect to the native state and are given in Kcal/mol.

for $\epsilon = 0 - 0.4$ we obtain an increase of the transition temperature compared to the bulk case. The range $0.3 \leq \epsilon \leq 0.4$ seems to be the optimal one regarding stability. For that range of ϵ the protein is more stable than in the absence of a cage. For higher values of ϵ the transition temperatures become lower. For $\epsilon = 1.0$ the curve of the specific heat is extremely broad and attenuated, reflecting the fact that the protein is almost denatured.

One of the main results of the present paper is illustrated in Fig. 4.8, where we show the contour plots the free energy $F(E, Q)$ for different values of ϵ. The radius of the cage is equal to 30 Åin all cases.

For $\epsilon = 0$ (Fig. 4.8 (a)) the presence of the native state (N), the intermediates (I_1 and I_2) and the unfolded states (U) can be clearly observed. The increase of ϵ leads to

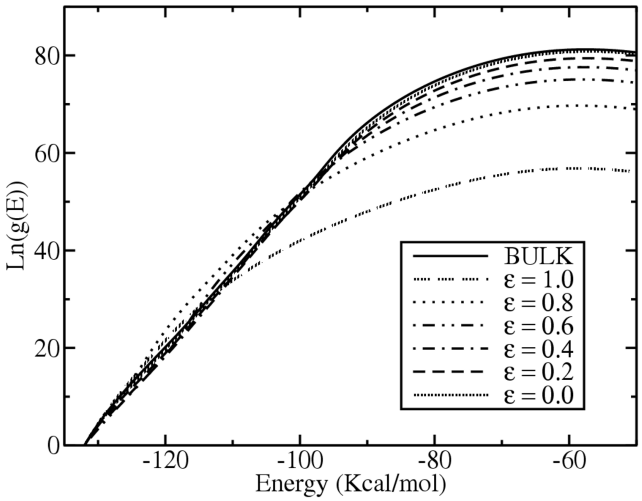


Figure 4.6: Logarithm of the DOS $g(E)$ for different degrees of hydrophobicity (ϵ =0.0, 0.2, 0.4, 0.6, 0.8, and 1.0) and for the bulk case. Notice the abrupt decay of $g(E)$ by ~ 13 orders of magnitude as ϵ goes from 0.0 to 1.0. For high values of ϵ, the protein tends to be in the unfolded state.

an effective increase of the confinement, and also to the presence of shallower minima for the intermediate states. For $\epsilon = 0.4$ (Fig. 4.8 (b)) the global minimum and the three local minima can still be distinguished. However, the intermediates become almost unstable and the energy landscape has practically only two well defined minima. A further increase of ϵ completely changes the energy landscape. For $\epsilon = 0.6$ (Fig. 4.8 (c)) the native state and the intermediates I_1 and I_2 are no longer present. Instead, a new intermediate state N' appears, which has more native contacts than I_1 and I_2, but less than N, and exhibits a much lower value of energy E. By looking at the structure corresponding to the minimum denoted by N' it is clear that is very similar to N, but not completely folded. We interprete that the N' state is a slight deformation of the native state produced by the presence of attractive wall. Note that also the "U"-minimum is shifted to much lower energies. This clearly indicates that the protein sticks to the wall of the cage. The net effect is that the potential landscape shows for this value of ϵ a two-state situation. Finally, for $\epsilon = 0.8$ (Fig. 4.8 (d)) only the unfolded states are present.

Shea and coworkers [JBS04] showed that for a particular protein one metastable state

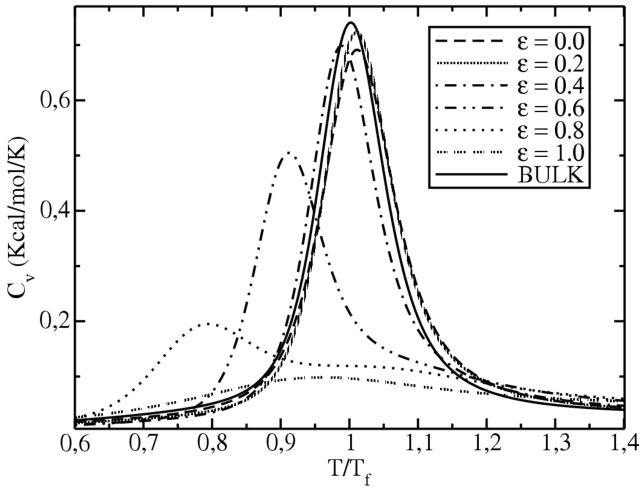


Figure 4.7: Specific heat of the protein for different values of $\epsilon = 0.0$, 0.2, 0.4, 0.6, 0.8, and 1.0, compared to the bulk case. $T_f = 321$ K is the transition temperature for the bulk. Notice how T_f and the peak of the specific heat decrease as ϵ goes from 0.0 (purely repulsive wall) to 1.0 (strongly attractive wall).

might exist in the presence a weakly hydrophobic barrier. In this work and for the peptide V3-loop we obtain a different result, namely, that the protein shows a folding behavior through intermediates in the bulk, but an attractive barrier with an optimum degree of hydrophobicity can lead to weaken the intermediate states and to induce a quasi two-state folding process.

Summarizing, we have studied the folding of the peptide 1NJ0 under different kinds of confining potentials. We used a force field which is independent on the native structure and includes relevant interactions such as the dipole-dipole and hydrogen bonds. We demonstrated the presence of intermediate states not reported before. These intermediates are strongly affected by the confining potentials.

4.2 PROTEIN-FIELD INTERACTION

Misfolding of a protein occurs when it becomes trapped in a local minimum of the potential energy surface (PES) where the conformation differs from the native-state struc-

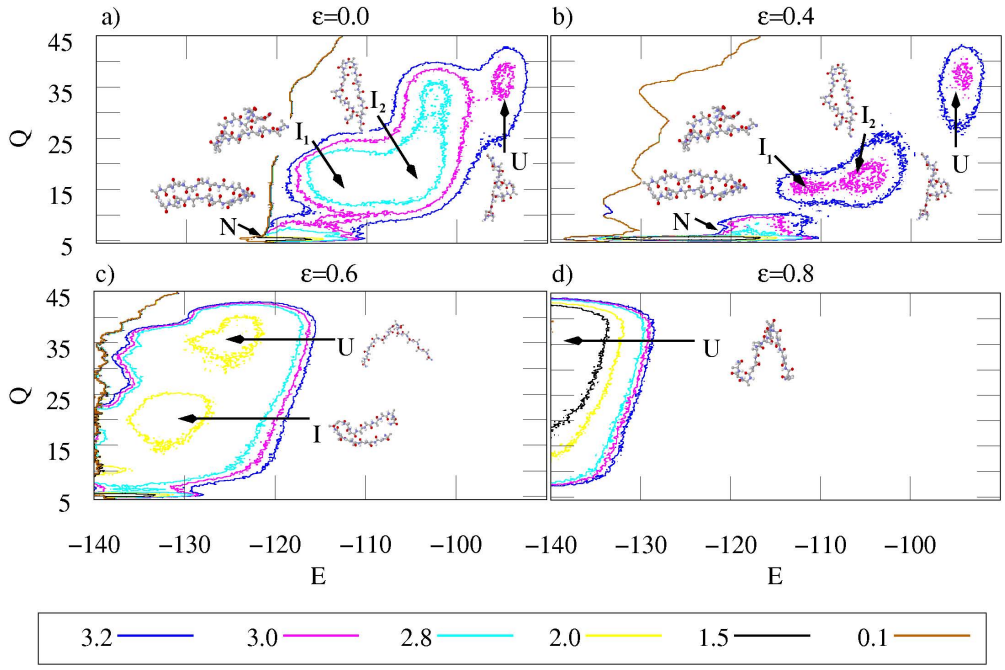


Figure 4.8: Contour plots of the free energy landscape $F(E,Q)$ for a cage with an attractive inner surface. Different degrees of hydrophobicity are displayed in plots a-d, corresponding to ϵ =0.0, 0.4, 0.6 and 0.8. The native and the intermediate states are slightly modified for $0.0 < \epsilon < 0.4$ but for larger values of ϵ the intermediate states dissappear and the native structure is deformed. As a consequence $F(E,Q)$ represents a two-states landscape. The contour lines represent the free energy difference with respect to the native state and are given in Kcal/mol.

ture. If the local minimum is stable enough, serious diseases may be caused, especially if the secondary structure of the misfolded conformation differs from the native-one [EAF+06, Kel98, LM00].

In this work we demonstrate that an external constant electric field can directly induce a dramatic conformational change in the secondary structure of proteins. Most importantly, we show that a transition from a β-sheet to an α-helix-like structure can be induced by field strengths which can be generated in a micro-electrode with an electrolyte inside [EML06].

The external electric field couples to the dipoles in the peptide units, which are parallel to the direction of the $\vec{\text{OC}}$ and $\vec{\text{NH}}$ bonds in the amide planes (see Fig. 4.9) and have a

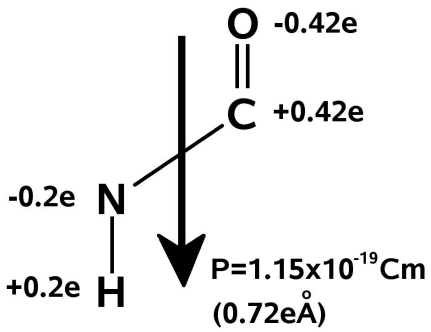


Figure 4.9: The dipoles of NH and OC in the amide plane give rise to a total dipole moment for each amino acid which has the value 1.1×10^{-29}Cm.

magnitude equal to 1.1×10^{-29}Cm [Wad76], i.e., 20 times larger than the dipole moment of a water molecule. This interaction can lead to structural changes. In the last years the alignment of hydrated proteins and small molecules in the gas phase was achieved experimentally using polarized light [SSW$^+$05, RCF$^+$09]. The alignment of the tertiary structure of large macromolecules under static and oscillating electric fields has also been described by molecular dynamics (MD) simulations [LBTY06, TLBZ09].

Here, we show that the folding dynamics in the presence of an electric field can be analyzed in a similar way as classical spin systems under magnetic fields. For that we generalize and implement a Monte Carlo approach used for spin glasses in order to be able to describe structural changes of macromolecules in real space.

Note that the superposition of the individual dipoles in a protein gives rise to a total dipole moment $\sum_i p_i \sim N$ ($N \equiv$ number of amino acids) when all the dipoles are aligned ferroelectrically as in case of the α-helix [Wad76, Hol85]. In contrast, the total dipole of a $\beta-$sheet should vanish because in this case the individual dipoles corresponding to the $i-th$ and $i+1-th$ amino acids in the sequence are oriented antiparallel to each other.

Now, it is known for polymers in general that a rotation of local dipoles can occur without any significant change in bond length and bonding angle [CWD$^+$07]. Therefore, a protein under an external electric field decreases its potential energy through dipole

alignment. On the other hand, the dipole-dipole interactions lead to an interplay between the conformation and the dipole arrangement in the peptide. Very high field strengths will force the protein-structure to be aligned. However, for field intensities corresponding to a coupling strength comparable to the energy of hydrogen bonds, we expect a more complex and physically more interesting potential energy surface.

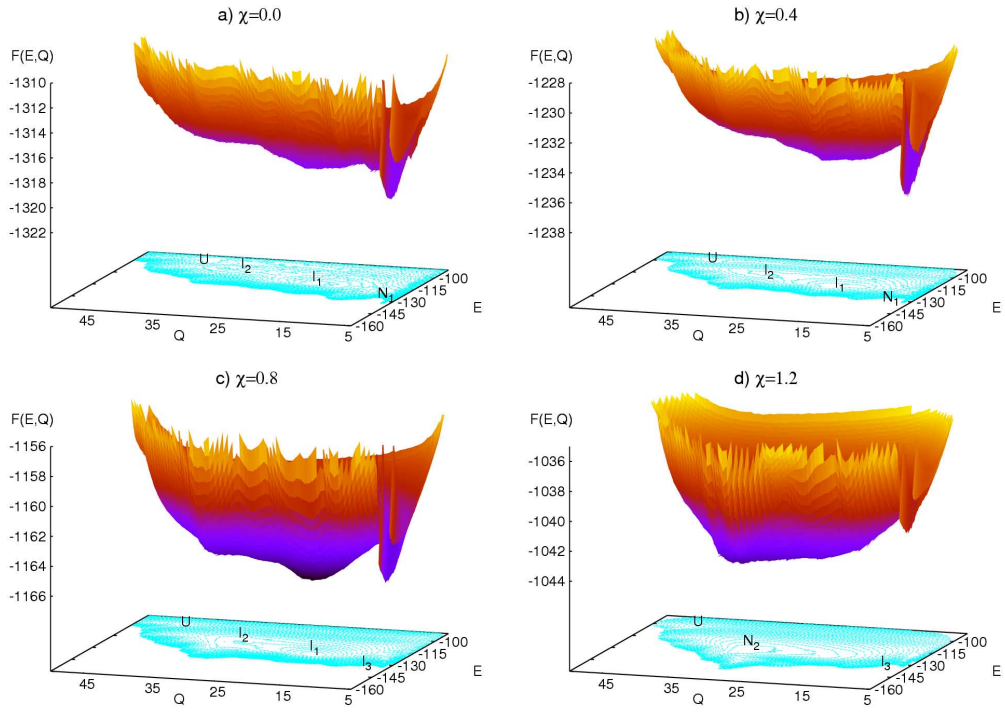


Figure 4.10: Free energy surface of the V3-loop as a function of the configurational energy E and the end-to-end distance Q for different strengths of the external electric field: $\chi = 0.0, 0.4, 0.8$ and 1.2. Local minima labeled as I_1 and I_2 correspond to intermediates. N_1 refers to the native state in absence of field, which becomes metastable (I_3) for $\chi = 0.8$. Note the formation of a new global minimum N_2 for the field strength $\chi = 1.2$. U corresponds to the unfolded states. The temperature in all cases is $T = T_f = 321$ K.

In order to analyze the direct influence of external fields on the secondary structure of proteins we study in this work the small peptide V3-loop, Protein Data Bank ID 1NJ0, which consists of a β-sheet structure in its native state. This peptide conforms the V3-loop of the exterior membrane glycoprotein (GP120) of the Human Immunodeficiency Virus type 1 (HIV-1). We describe the protein by using an unbiased off-lattice

model recently developed by Yow and coworkers [CSM06]. This coarse-grained model contains the most important ingredients needed to describe folding. In particular, local hydrophobic interactions between the residues i and $i+1$ and dipole-dipole interactions are treated on the same footing.

In the following we briefly describe the model (for more details see Ref. [CSM06]). Each amino acid is represented by a unit which contains the atoms N, C_α, C', O and H. The residues are modeled as spherical beads R attached to the C_α's. The only remaining degrees of freedom are the Ramachandran angles ψ and ϕ. Thus, the force field is given by

$$V_{Protein}(\psi,\phi) = V_{St} + V_{HB} + V_{DD} + V_{MJ} + V_{L-HP}, \tag{4.1}$$

where V_{St} represents hard-core potentials to avoid unphysical overlaps, V_{HB} accounts for the hydrogen bonding and V_{DD} for the dipole-dipole interaction. V_{MJ} is a distance dependent version of the Miyazawa-Jernigan (MJ) matrix [MJ96], which describes the interaction between residues. V_{L-HP} represents the local hydrophobic effect. The role of the presence of water molecules is taken into account both by the term V_{MJ} and V_{L-HP}. Notice that V_{MJ} partially includes the effect of water polarization [WL00]. The dipole-dipole interactions V_{DD} are divided into local and non-local terms. The latter account for the interactions between dipoles belonging to amino acids which are not nearest neighbors, and are described by the term

$$V_{DD}^{nl} = \chi_{DD}^{nl} \sum_{i,j\neq i\pm1,\mu,\nu} \left[\frac{\mathbf{p}_{i\mu}\cdot\mathbf{p}_{j\nu}}{r_{i\mu j\nu}^3} - 3\frac{(\mathbf{p}_{i\mu}\cdot\mathbf{r}_{i\mu j\nu})(\mathbf{p}_{j\nu}\cdot\mathbf{r}_{i\mu j\nu})}{r_{i\mu j\nu}^5} \right], \tag{4.2}$$

where μ and ν refer to OC or NH and $\mathbf{p}_{i\mu}$ is the corresponding OC or NH dipole moment in the i-th amino acid of the sequence. The dipole-dipole interactions between amino acids which are nearest neighbors (local terms) are approximated as $V_{DD}^{l} = \epsilon_{DD}^{l} \sum_i (\mathbf{P}_i \cdot \mathbf{P}_{i\pm1}/|\mathbf{P}_i||\mathbf{P}_{i\pm1}| - 1)$, where $\mathbf{P}_i = \mathbf{p}_{i\mathrm{CO}} + \mathbf{p}_{i\mathrm{NH}}$ refer to the total dipole moment of the i-th amino acid. ϵ_{DD}^{nl} and ϵ_{DD}^{l} are coupling constants. Notice that the

nonlocal interaction has a dependence on the distance between dipoles while the local one is roughly independent of it, because the dipoles are localized in the center of the amide plane and the distance between nearest neighbors remains unchanged upon conformational transformations.

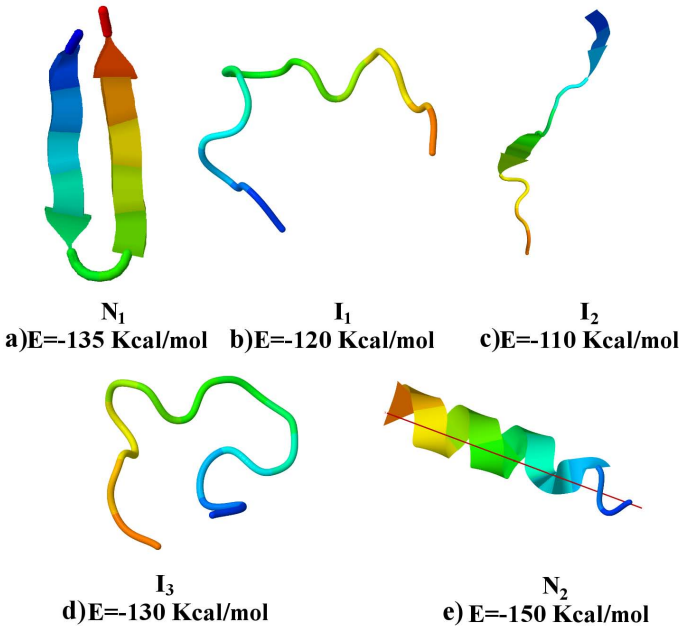


Figure 4.11: For low field magnitudes one native (N_1) a) and two intermediate (I_1 and I_2) b)-c) states are displayed in the FEL of the peptide 1NJ0. For high field strengths the peptide presents a new intermediate (I_3) d) and native (N_3) e) states. (The intermediate states are schematic). The native state N_3 is aligned to the field orientation given by the black (red in color) line in e).

The term describing the coupling of the protein to the external electric field $\mathbf{E}$ reads $V_{FD} = \chi \sum_i \mathbf{P}_i \cdot \mathbf{E}$, where the strength χ of the interaction term is dimensionless. To give an idea of the order of magnitude of the field strengths analyzed here, $\chi = 1$ corresponds to 5.16×10^8V/m. We use the parameter χ instead of $\mathbf{E}$ to avoid the repetition of the factor 10^8V/m in the text and plots.

We have determined the thermodynamic properties of the V3-loop under an external electric field using the Wang-Landau algorithm [WL01]. Since the joint density of states

g(E,Q) for biomolecules is difficult to obtain with the classical Wang-Landau algorithm, we implemented here the *1/t* convergence criterion, recently proposed by Belardinelli and coworkers [BP07] in order to avoid saturation errors, which is essential for the appropriate treatment of complex Potential Energy Surfaces (PES). This modified algorithm has not been reported before for proteins.

More details of the implementation of the Wang-Landau algorithm can be found in Refs. [WL01, BP07, OLYG09]. We calculated the joint density of states $g(E, Q)$ as a function of the configurational energy E (in Kcal/mol) and of the end-to-end distance of the protein Q (in Å). With the help of the Wang-Landau simulation we explored the volume $[-160 < E < -90] \times [4 < Q < 50]$ in the reduced phase space defined by E and Q. At each Monte Carlo step we changed the Ramachandran angles ϕ and ψ. After 1×10^{10} Monte Carlo steps we obtained the unnormalized density of states (using the convergence criterion $f_{final} \sim \exp(10^{-8})$).

From the so computed $g(E, Q)$ we obtained the free energy surface as a function of E and Q for different values of χ, running from 0 to 2 Kcal/mol, which is shown in Fig. 4.10. The temperature used for this calculation was $T = T_f$ =321 K, which is the optimal folding temperature of the V3-loop peptide. In the absence of an external field ($\chi = 0.0$) the free energy shows the typical funnel-like form around the native state N_1. In addition to the native- and the unfolded states, two characteristic local minima, I_1 and I_2 can be distinguished, which correspond to intermediates [OLYG09]. In the presence of an external field the whole free energy landscape is modified. For $\chi = 0.4$ (see Fig. 4.10 b)) still no considerable changes occur. However, already for $\chi = 0.8$ the native state N_1, which is strongly by the field, is no longer the global minimum of the free energy, but it becomes degenerate with the intermediates. The protein exhibits no global minimum. Therefore, we also consider N_1 as a metastable state and label it as I_3 (see Fig. 4.10c)).

For larger field strengths, dramatic qualitative changes in the potential landscape

occur, which give rise to a new phenomenon, namely, a *transition from a $\beta-$sheet to an α-helix-like secondary structure*. This effect can be directly observed in Fig. 4.10 (d): for $\chi = 1.2$ a conformation which does not correspond to any local minima in the absence of the field, becomes the new native state (N_2). The state N_2 is characterized by the coordinates (E = -145, Q = 30). Notice, for comparison, the native state in the absence of the field N_1 is located at the point (-135,5) in the plane (E, Q). The native states N_{1-2} and the intermediate states I_{1-3} are shown in Fig. 4.11.

This result suggests that by increasing the magnitude of the external field one should be able to control the conformation of the V3-loop and, in general, the secondary structure of proteins.

In order to visualize the field induced transition from the native state N_1 to the state N_2 we looked for the structure which yields the largest contribution to the partition function $\mathcal{Z}(T,\chi) = \sum_{E,Q} g(E,Q)e^{-E(\chi)/k_BT}$ for each value of the temperature T and the strength χ. Such conformation constitutes the observable structure, i.e., that having the highest probability to be present. In Fig. 4.12 we show the coordinates (E,Q) of the observable conformations for $\chi = 0, 0.4, 0.8$ and 1.2 and for temperatures running from 10K up to 600K. In the absence of the field and for $T < T_f$ the native state N_1 (E = -135, Q = 5) contributes most to $\mathcal{Z}(T,\chi)$. At very high temperatures only the unfolded state can be observed. The intermediates I_1 and I_2, although present as local minima in the PES (see Fig. 2), do not provide the dominant contribution at any temperatures. New features appear when the external field is switched on. For $\chi = 0.4$ the intermediate state I_2 ($E = -110, Q = 35$) becomes observable for temperatures above T_f. At very high temperatures again the unfolded states dominate. A peculiar situation occurs for $\chi = 0.8$, where both intermediates I_1 and I_2 can be observed and yield the dominant contribution on a wide range of temperatures. Interestingly, I_1 ($E = -120, Q = 20$) can be observed even below T_f. The state N_1 is only dominant at low temperatures.

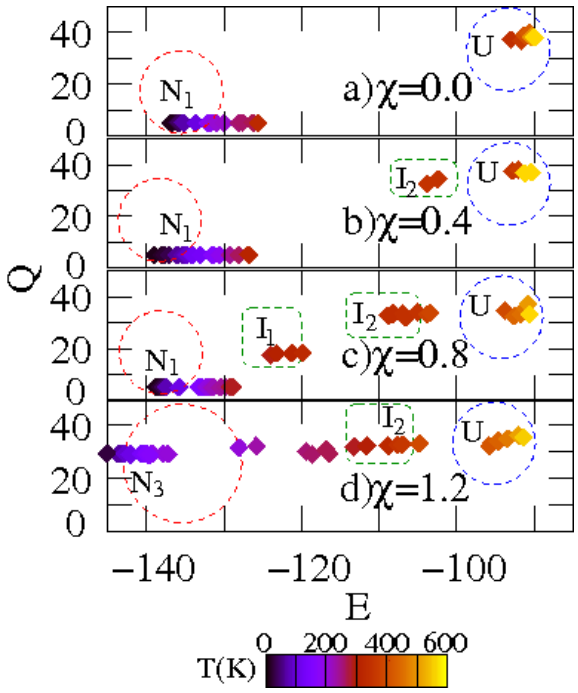


Figure 4.12: Coordinates in the (E, Q)-plane of the conformations yielding the maximal contribution to the partition function for $\chi = 0.0, 0.4, 0.8$ and 1.2 and at different temperatures. Note that for $\chi = 0.0$ the observable structures lie around the point $(E = -135, Q = 5)$ (β-sheet) while for $\chi = 1.2$ they are located near the point $(E = -150, Q = 30)$ (helix). Dark (blue) and light (yellow) diamonds refer to low and high temperatures, respectively (see temperature scale).

Finally, the conformation which yields the largest contribution to $\mathcal{Z}$ for $\chi = 1.2$ and $T < T_f$ is the helix-like structure with coordinates $(E = -150, Q = 30)$, corresponding to the new native state N_2. As the temperature is further increased, the intermediate I_2 starts to be the most probable structure. As in the case of smaller field strengths, at high temperatures ($T > 400$ K) the unfolded state U yields the largest contribution to the partition function.

A more graphical description of the transition and particularly of the structure of the field-induced new native state N_2 can be gained from the Ramachandran plots. In Fig. 4.13 we show the Ramachandran plots for $\chi = 0.0, 0.4, 0.8$ and 1.2 and at the folding temperature $T = T_f$. In the absence of the external field most of the bond-angles of the peptide lie in the upper left part of the Ramachandran plot, inside the region which

characterizes the β-sheet of the state N_1 (see the crosses in Fig. 4.13). There are, of course, angles which lie outside this region. This dispersion is due to the bonds at the ends and in the turn of the β-sheet.

As the magnitude of the field is increased, the local dipole moments of the amino acids start to change their orientation. This leads to rearrangements in the structure and, in particular, in the bond angles. As a consequence, for $\chi = 0.4$ and 0.8 the dispersion of the points in the Ramachandran plot increases. For $\chi = 0.8$ a considerable fraction of the angles lies not only outside the β-sheet region, but they are distributed over the four panels of the plot. Now, if $\chi = 1.2$ again a qualitative change can be observed. A considerable fraction of the angles is concentrated inside and around the region which characterizes α- and a 3_{10}-helices. This happens because for large fields all dipoles in the peptide tend to be aligned in the direction opposite to the applied field in order to minimize the interaction energy (note that $\chi > 0$). The resulting helix-like structure has a nonzero total dipole moment. This structure is shown in Fig. 4.11 e). One can observe that the helix is aligned to the field orientation given by the black (red in color) line.

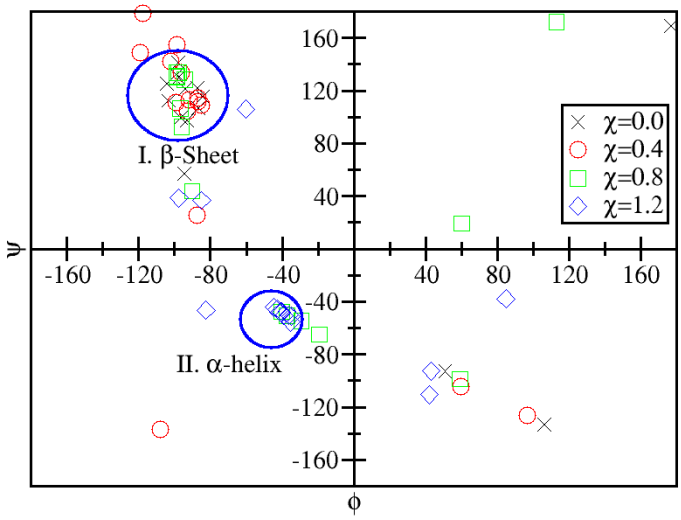


Figure 4.13: Ramachandran plot of the V3-loop for different strengths of the external electric field at $T = T_f = 321$ K. The regions corresponding to helices and β-sheets are indicated.

The energy of the native state N_1 undergoes slight changes with increasing χ with respect to the the case without field ($E(N_1)|_{\chi=0} = -137.5$Kcal/mol). In particular, $\Delta E/|E| \sim -0.009$ for $\chi = 0.4$ and $\Delta E/|E| \sim -0.015$ for $\chi = 0.8$. The energy of the state native state N_2 for $\chi = 1.2$ is equal to -150.0 Kcal/mol. The energy difference of ~ 12.5 Kcal/mol between the native states $\chi = 0$ and $\chi = 1.2$ results from the balance between the field-peptide interaction energy, i.e., the energy gained from the alignment of the dipole moments, and the different number of hydrogen bonds in the configurations N_1 and N_2.

From our calculations it is clear that an external electric field forces the peptide to undergo a conformational change to a structure in which the dipoles are aligned parallel to each other and antiparallel to the field. For $\chi = 1.2$ this structure is a helix-like one. If the field strength is further increased the native state displays more features related to helices, and for a sufficiently high value of χ a perfect α-helix is formed. Note that the situation is analogous to that of a classical spin system interacting with a magnetic field. The paramagnetic state would represent the unfolded protein structure, in which dipoles are oriented randomly in the absence of a field, the antiferromagnetic state would be the analog of a $\beta-$sheet, where neighboring dipoles are antiparallel to each other, and the ferromagnetic state would correspond to the α-helix. However, and in addition to the case of spin systems, the protein is free to move in space and the Hamiltonian consists of different competing many-body interaction terms.

At this stage it is important to mention that recent molecular dynamics simulations performed to study the effect of static and oscillating electric fields on insulin [LBTY06] and an β-Amyloid peptide [TLBZ09], yield a destabilization of α-helices and a change to a β-sheet or to a random-coil structure. Note, however, that in both cases the studied proteins are rather long and exhibit a tertiary structure consisting of subunits forming different secondary structures, and that the most important effect of the field was the

alignment of the tertiary structure. Therefore, the destabilization of a helix is only part of this global alignment process and does not contradict our results, because the field does not interact directly with the secondary structure, as is the case in our study.

Moreover, from the results of our calculations and the above discussion we can propose the following alternative picture for externally assisted folding: a transition from a β-sheet to a α-helix within a large protein can be induced by an electric field which directly interacts with the secondary structure. This is only possible if the tertiary structure of the protein cannot be aligned. A way to achieve a "frozen" tertiary structure is just through confinement. Therefore, assisted folding could result as the interplay between an electric field and confinement. Nano-capacitors provide both requirements at the same time.

Finally, it is important to point out that a much more efficient field induced transition from a β-sheet to an α-helix should occur if both conformations are already present as minima in the free energy in absence of the field. Research in this direction is in progress.

4.2.1 Electric Field produced by a Nano-electrode

We now show that the effect described throughout the letter might be induced in *in vivo* experiments to repair misfolding or favor correct folding. We considered the cytoplasm as a dielectric medium in which free ions are present (ionic solution) and assume that a capacitor consisting of square (4 nm $\times$ 4nm) nano-electrodes is introduced into the cell. The interior of the nano-capacitor has therefore a relative dielectric constant $\varepsilon_{cyto} \sim 60$, which corresponds to the dielectric constant of the cytoplasm. Assuming that the separation between nano-electrodes is 5 nm, and in order to create an electric field of the magnitude $|E_{ext}| = 6.2 \times 10^8$N/C, an external voltage of $|V| = 3.79$ V has to be applied. This value is perfectly accessible in such experiments [Yua07]. The charge density of electrolytes in the cytoplasm close to the surface of a nano-electrode is $\sigma =$

$4.59 \times 10^{17} e/m^2$ [LBK$^+$05]. Therefore, the cytoplasm inside the nano-capacitor produces the usual response of an electrolyte and generates an induced electric field $|E_{ind}| = 1.38 \times 10^8$N/C (calculated using $E = \sigma/A\epsilon_{cyto}$). The magnitude of the electric field inside the capacitor is given in Ref. [EML06] (with parameters adapted for our purposes) by,

$$E(z) = (E_{ext} - E_{ind}) \coth(\sqrt{A}z + \phi) \tag{4.3}$$

here $\sqrt{A} = e(E_{ext} - E_{ind})/(2K_BT) = 1.1/\text{Å}$, $\coth(\phi) = E_{ext}/(E_{ext} - E_{ind}) = 1.22$ ($\phi = 1.15$) and z (in Å) denotes the distance from the capacitor plate. Substituting these values in the expression of the electric field we obtain, $E(z) = 6.2 \times 10^8 \coth(1.1z + 1.15) V/m$. The dependence of $E(z)$ on z is shown in Fig. 4.14. Note that at the plate ($z = 0$) the electric field is maximal (vacuum value). For distances $z >> 0$ the electric field decreases due to the screening of the ions in the cytoplasm and reaches the asymptotic magnitude $E(z \to \infty) = 6.2 \times 10^8 V/m$. This value is corresponds to $\chi = 1.2$ in Fig. 4.10, which we have shown to be enough to strongly affect the folding of a protein.

4.3 SELECTION AND SEQUENCE DESIGN

In this Section we will make use of the Model II described in Chapter2. The dynamics employed will be the Langevin Dynamics but the main idea does not depend on the method one uses to describe the dynamics. As we mentioned in Chapter 2, it would be of crucial importance to have a simple test to know if a given sequence of amino acids will fold in a short time in comparison to random sequences. We will investigate to which extent the convergence of dynamical trajectories on the very initial stages could be a distinguishing feature for a good folder.

It is well-known, that most proteins fold rapidly and reliably to a unique native

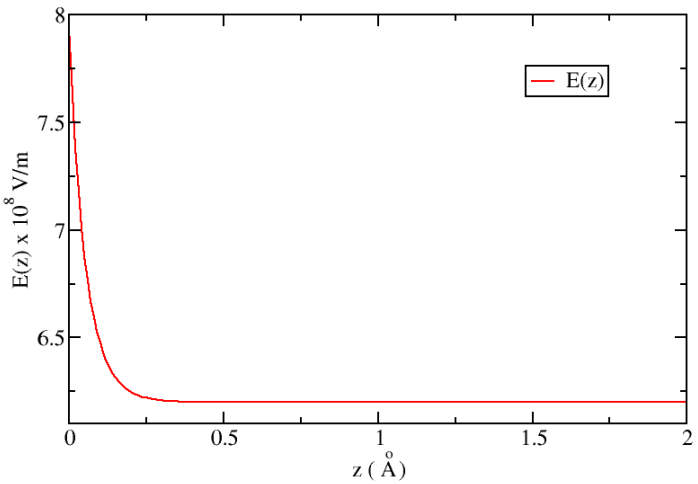


Figure 4.14: Electric field inside the chaperon as a function of the distance to one end of the cavity. The field decreases because of the screening of the electrolytes in the cytoplasm medium.

state from any one of a vast number of initial unfolded conformations [Cre92, FP02]. Nowadays, the consented answer to Levinthal's paradox is found in the specially designed energy landscape of a foldable protein, which resembles a many-dimensional funnel, where moving along the free-energy gradient narrows the accessible configuration space and guides to the unique native structure lying at the bottom of the funnel [GLSW92, SSK94c, BOSW95]. The funnel is also rough, giving rise to local minima, which can act as traps during folding. Most random amino acid chains have numerous funnels to different low-energy states but an evolutionary designed protein sequence has usually one. Typically a random amino acid chain will not fold to its lowest free-energy minimum in times less than that needed to explore completely the configuration space, thus making these times astronomically large [GLSW92].

In the present work we shall call good folders those amino acid sequences, which exhibit a protein behavior, *i.e.* those that fold into the unique native state within a reasonable time. Characterizing good folders is of vital importance for scientific and techno-

logical applications. One of the most accepted criteria to characterize a good folder is by measuring the energy gap between its global energy minimum and the minimum energy of configurations in the disordered ensemble. The disordered ensemble consist of configurations which are structurally dissimilar to the configuration of the global minimum (the native state) [SSK94c, SG93, Sha94]. The thermodynamic stability of the native state is strongly correlated to the magnitude of the gap. The energy gap also indicates the ability of the sequence to fold into the global minimum in a reasonable time. The main problem of calculating the energy gap is that one needs to know *a priori* the native state. Therefore, one should perform the complete folding dynamics at first. Due to an unknown folding time it may take very long before one could identify some amino acid chain as a bad folder. The situation is more complicated when we try to compute the energy gap for hundreds or thousand of sequences. Recently, Cassetti et. al [MC06b] proposed a method for distinguishing proteins by their ability to fold. This study suggests monitoring the curvature fluctuations of the energy surface along dynamical trajectories. However, the Cassetti's method is feasible only for coarse-grained models with a smooth potential energy surface [MC06b]. In the present Thesis we investigate a new method to monitor the convergence of dynamical trajectories on the very initial stages. We will show that the convergence of the trajectories is a distinguishing feature for a good folder.

Different kinds of dynamics can be used for the description of an amino acid chain, for instance Langevin dynamics for atomistic models [Gil93] or Monte Carlo dynamics for lattice models [HD75, Sha94]. Commonly, the time development of the configuration can be written as $\mathcal{C}(t) = g^t\mathcal{C}(0)$, where $\mathcal{C}(0)$ is the initial configuration and g^t denotes the dynamical transformation, which strongly depends on temperature and has probabilistic nature if it simulates how water molecules governed by chaotic motion affect the amino acid chain.

The dynamical transformation approach can be used to explain the funnel form of

the energy surface for proteins: if the dynamical transformation acts on two arbitrary points in the configuration space then the "distance" between them becomes enlarged $d(\mathcal{C}_1(t), \mathcal{C}_2(t)) > d(\mathcal{C}_1(0), \mathcal{C}_2(0))$, where d stands for "distance" between configurations. The time t should be not less than a time required for overcoming typical local traps in the folding funnel. The enlargement of the distance means that after a certain time t there should emerge structural similarities between two propagated yet initially unrelated chains. A short convergence time t of the distance is a characteristic for a good folder.

Now imagine the following problem being posed: out of N amino acid sequences one has to sort out the best candidates for folding in some reasonable time. One of the most radical solutions to this problem would be to perform the dynamics of each sequence starting from various randomly chosen initial positions and to check whether the dynamical trajectories reach the same native conformation. This would be, however, extremely time consuming (especially in the case of all-atom molecular dynamics simulations with solvent molecules). Besides this fact, it is *a priori* unclear how long the dynamical simulation must be run because the value of the folding time is initially unknown.

D. Gridnev and M. Garcia proposed an alternative solution to this problem based on comparing the amino acid sequence by a properly defined Rate of Convergence, discussed in Chapter 3. They applied it successfully to the lattice model [GG].

In this Thesis we extend the method to off-lattice models. We just summarize the results of [GG] on the lattice model. We tested our approach on both a standard lattice and an off-lattice models of proteins [SSK94c, SG93, BTR$^+$99]. Although geometrically poor, the lattice model is protein-like in the sense that lattice proteins fold to a unique native structure from an astronomically large number of possible initial conformations and do so rapidly and reproducibly. The structure of the amino acid chain is represented as a self-avoiding walk on the cubic lattice. A random configuration is then a

self avoiding random walk. The sequences are composed of amino acids of 20 types and the chain contains 36 monomers. Two monomers are considered "in contact" if they occupy neighboring positions on the lattice but are not sequence neighbors. The energy of two monomers in contact is calculated through the 20×20 Miyazawa-Jernigan matrix (Table VI of Ref. [MJ96]).

For example, the energy of the chain $\mathcal{C}$ having N monomers is expressed through the contact map as $E(\mathcal{C}) = \sum_{i,k=1}^{N} V_{ik}\Delta_{ik}(\mathcal{C})$, where the interaction matrix for monomers V_{ik} is determined by the Miyazawa-Jernigan matrix. The similarity parameter Q, which expresses on the scale from 0 to 1 the "closeness" of some conformation $\mathcal{C}$ to the native state configuration $\mathcal{C}_N$, becomes in our notations $Q = \mathcal{O}(\mathcal{C}, \mathcal{C}_N)/\mathcal{O}(\mathcal{C}_N, \mathcal{C}_N)$. The dynamic transformation g^t is implemented through the Monte Carlo dynamics [BTR$^+$99] with move set including end moves, corner flips, and crankshaft moves.

To demonstrate the efficiency of the Rate of Convergence approach Gridnev and Garcia [GG] have chosen a designed sequence [AGS94] of 36 monomers S_0 = SQKWLERGATRIADGDLPVNGTYFSCKIMENVHPLA. The native state of S_0 has the energy $E_N = -16.5$ in dimensionless $k_B T_{room}$ units, where T_{room} stands for the room temperature [MJ96]. At the folding temperature $T_f = 0.25$ (in Miyazawa-Jernigan dimensionless units) the chain S_0 always reaches its native state starting from any conformation and the mean folding time (obtained by sampling 10^3 self-avoiding random walks in initial configurations) is $t_f = 1.5 \times 10^6$ steps. As a remark, the folding time was calculated as the average time (over many trajectories) required for the protein to get its native state. The folding temperature was determined as the optimal temperature at which all the structures get the native state.

Gridnev and Garcia generated 800 sequences with a random amino acid decomposition and the designed sequence S_0 was hidden among random sequences as "a needle in a haystack". For each amino acid sequence the Rate of Convergence was calculated and

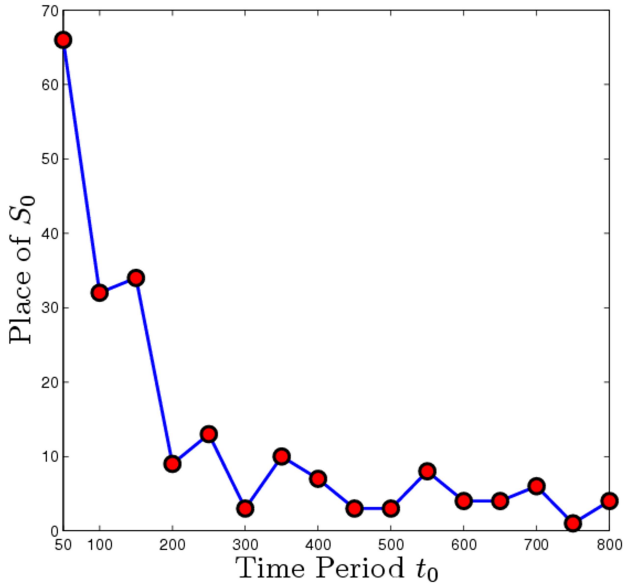


Figure 4.15: The place of the designed sequence S_0 resulting after ordering the sequences by the Rate of Convergence in descending order versus the time period t_0. *Taken from [GG].*

all sequences by the Rate of Convergence in descending order were sorted. The Rate of Convergence, $\langle R(t_0, T_f)\rangle$, was computed over 500 randomly chosen pairs of positions, $T_f = 0.25$ is the folding temperature of S_0. The Rate of Convergence was calculated each 50 time steps starting with $t_0 = 50$.

It is important to notice that for each new time period the 800 random sequences were generated anew. Fig. 4.15 shows the place of S_0 after sorting the sequences by the Rate of Convergence in descending order for each time period t_0. For $t_0 = 200$ the designed sequence S_0 took the 9-th place and for $t_0 = 700$ even the first place. One can see that for $t_0 \leq 150$ the designed sequence gets lost among other random sequences. The reason for that is that the time $t_0 \leq 150$ is insufficient for overcoming local minima through potential barriers. For $t_0 \geq 200$ the sequence S_0 gets into to the top ten, which allows to conclude that $t_0 \geq 200$ is sufficient for distinguishing the sequences by their ability to fold.

The values of $\langle R(t_0 = 300, T = T_f)\rangle$ are distributed among the first 15 places. The

place number three is occupied by the designed sequence S_0. Though there are overall 800 sequences one can see that the gap in $\langle R(t_0 = 300, T = T_f)\rangle$ values between the first placed sequences is large. (The mean value of $\langle R(t_0 = 300, T = T_f)\rangle$ for 800 sequences is 0.94).

One can check how the Rate of Convergence $\langle R(t_0, T)\rangle$ depend on the number of generated pairs. One founds that the square root of variance of $R(t_0 = 300, T = T_f)$ values increases in the range from 1.2 to 1.4 as one increases t_0 from 50 to 800. For the number of initial pairs $n \geq 100$ the distribution of $\langle R(t_0 = 300, T = T_f)\rangle_n$, where the average is calculated over n pairs, is almost Gaussian (as it should be by the central limit theorem). The calculations show that for 500 pairs the precision in determination of $\langle R(t_0 = 300, T = T_f)\rangle$ is sufficient for locating the first placed sequences, which is also due to a large gap between $\langle R(t_0 = 300, T = T_f)\rangle$ values of these sequences.

The Rate of Convergence at $T = 300$, $\langle R_{random}(t_0 = 300, T)\rangle$, was computed by generating 1000 random sequences. Fig. 4.16 shows the normalized Rate of Convergence for good folders S_0 and S_1 (see the definition of S_1 below) and for a bad folder given by the amino acid sequence S_{bad} = QACYDGVHWPMEANGYVTKWTVFRLSLWSFKLTKSW. It is remarkable that the normalized rate of S_0 measuring the folding ability peaks exactly at the folding temperature. On the contrary, the same dependence for a bad folder does not show any pronounced peak, but rather a monotonous, almost temperature independent behavior (see Fig. 4.16).

In order to show that the Rate of Convergence approach is also able to perform sequence design Gridnev and Garcia applied the algorithm to 5000 randomly generated amino acid sequences having 36 monomers. The top 5 sequences turned out to be good folders. The Rate of Convergence was calculated at $t_0 = 200$ and the sampling was done over 300 pairs of initial positions. The temperature was set to the folding temperature of the designed sequence S_0, namely $T = T_f$. By its Rate of Convergence the

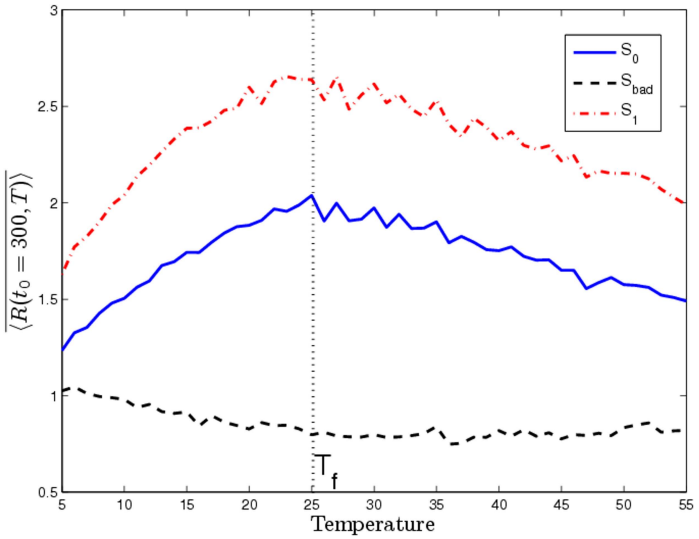


Figure 4.16: The normalized Rate of Convergence versus temperature for the designed sequence S_0 for the time period $t_0 = 300$. Dash-dot: the same for the sequence S_1. Dashed line: the normalized Rate of Convergence of a bad folder. The vertical dotted line corresponds to the folding temperature of S_0. The temperature is given in dimensionless Miyazawa-Jernigan units multiplied by 100. *Taken from [GG].*

designed sequence S_0 occupied the position 33. The results for the top two sequences S_1 = KWEEHEWGKDNLSDLHMHENEERFAQEQHNRDPQTD and S_2 = NALCD-DCSTEWCIPSMCCMCFEFIDFYKKKQQWRQM are analyzed in the following lines. The native states of both sequences are shown in Fig. 4.17. The energies of the native states are $E_N(S_1) = -16.88$ and $E_N(S_1) = -14.29$ respectively. The energy of S_1 in its native state is even lower than that of the designed sequence S_0, despite the fact that S_1 has the number of native contacts by 6 less than S_0 (note that the structure of S_0 was specifically designed to maximize the number of native contacts and 40 native contacts is the maximal reachable number for the sequence length of 36 monomers).

Both sequences $S_{1,2}$ have the folding temperature equal to T_f and their folding time is approximately 50 times longer than the folding time of S_0. This is the fact which deserves a discussion: in spite of $S_{1,2}$ having at all temperatures a better normalized Rate of Convergence compared to S_0, their folding time is substantially longer. This

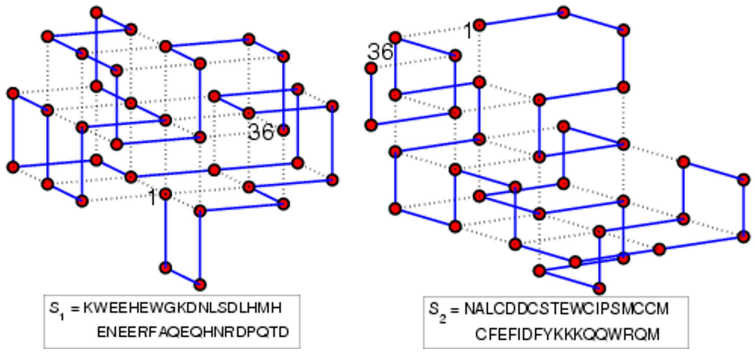


Figure 4.17: Native states for the sequences S_1 (left) and S_2 (right). Dotted lines connect those monomers that are in contact. The energies in the native state are $E_N(S_1) = -16.88$ and $E_N(S_2) = -14.29$. The number of native contacts for S_1 and S_2 is 34 and 27 respectively. Taken from [GG].

Name	Sequence	Model
SEQ1	311114442344312212224434333334	DHTP
SEQ2	443234423233421321132243424311	DHTP
SEQ3	321224314333113213344411112243	RHTP
SEQ4	414124323443321423324242141441	RHTP
SEQ5	444444444444444444444444444444	HMP

Table 4.3: The five sequences studied in this paper and their corresponding models. The folding time of the sequences is $t_f > 1 \times 10^7$ time steps. All the sequences have $N = 30$ monomers.

supports the idea that for a good folder there are so-called "hot contacts", which are formed in the first place, and then the chain undergoes the process of fine tuning. The discrepancy in the Rate of Convergence might be explained by the fact that though the "hot contacts" are formed quicker in $S_{1,2}$ compared to S_0, the process of fine tuning for $S_{1,2}$ takes a longer time.

One can also calculate the energy gap for both sequences S_1 and S_2. This was done by considering the bulk of dissimilar configurations satisfying $Q < 0.3$. For the designed sequence S_0 one obtains the energy gap $\delta = -2.0$. For the sequences $S_{1,2}$ the gap is -0.88 and -0.8 respectively. It is interesting to note how in this case the longer folding time is correlated with the lower energy gap.

In order to demonstrate that the method of Rate of Convergence also works for more

complicated protein models we tested it on a more sophisticated off-lattice model of proteins proposed by Clementi et. al [CMB98], for sequences of $N = 30$ monomers. An Overdamped Langevin Dynamics (OLD), studied in Chapter 3, described the time evolution g^t of the monomers. As a remark, OLD is suitable for Biomolecules because of the number of atoms and the velocities. Also, the time required for the simulation of the solvent is speed up because it is treated as an implicit random fluid

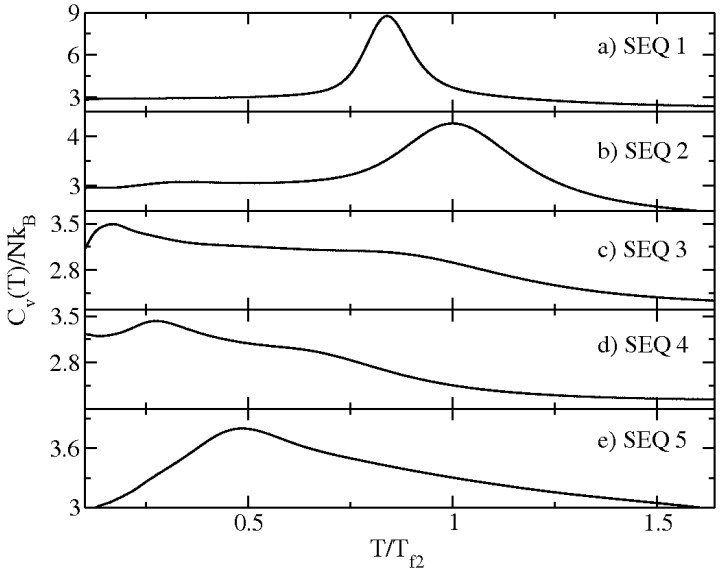


Figure 4.18: Specific heat vs. temperature for the five sequences shown in Table V. SEQ1 and SEQ2 show a very well localized peak which is a consequence of the funnel structure of their potential energy surfaces. These two sequences are known to be good folders. SEQ3, SEQ4 and SEQ5 have not a defined peak but the curve is spread in the whole interval of temperatures, they are known to be bad folders. The temperature axis is normalized respect to the transition temperature of SEQ2, T_{f2}=15.3 in units of k_BT.

As we did for the lattice model, the Rate of Convergence for the off-lattice model is given at the time t_0 and temperature T as an average of the Eq. 3.43 over several independent trajectories, in our case 200.

Rojas et. al [HRL08] have already distinguished three possible models in the Clementi potentials depending on how the parameters are fixed. The first model considered in this Work is the designed heteropolymer (DHTP) which intends to represent a natural

protein, i.e., a good folder. The second and third models corresponds to the random heteropolymer (RHTP) and the homopolymer respectively, both are expected to be bad folders because of their very rugged energy landscape.

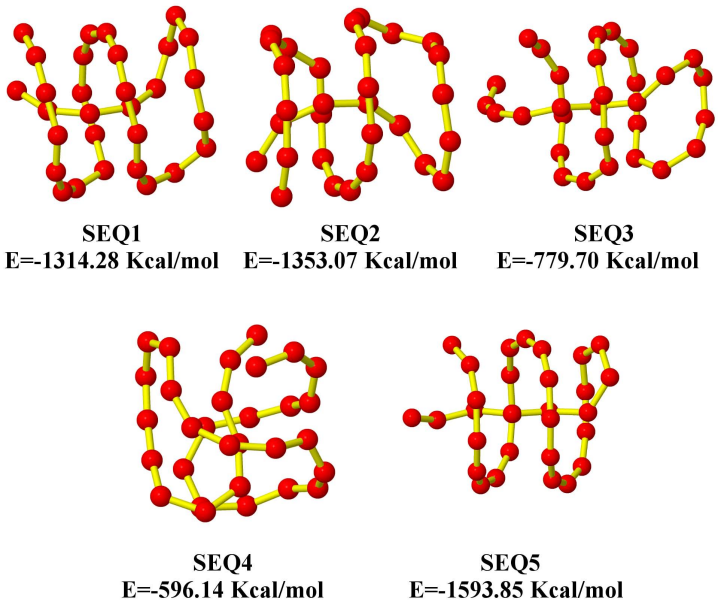


Figure 4.19: Global minima of the five sequences studied in this Work called SEQ1-5.

Wang-Landau simulations were performed over five different sequences called SEQ1-5 (see Table 4.3) to test the folding ability reflected in the specific heat curve. A picture of the global minima of these five sequences is displayed in Fig. 4.19. For the sequences SEQ1 and SEQ2, which belong to the DHTP model, we observe a very sharp peak in the heat capacity, see Fig. 4.18 (a) and (b). The sharp peak demonstrates that the transition from the unfolded state to the native state is very abrupt (first order transition) because there is a clear separation of the unfolded states respect to these states close to the native one. This is also a consequence of the big energy gap between folded and unfolded states of this designed model. In contrast to the SEQ1 and SEQ2, the SEQ3 and SEQ4, which belong to the RHTP model, they do not exhibit a defined peak (see Fig. 4.18 (c) and (d)) from which we can conclude that no real transition from an unfolded to a folded

states exists at all. The last sequence SEQ5, or HMP, shown in Fig. 4.18 (e), displays a peak which is not sharp but extended over a long range of temperatures. This means that the first derivative of the free energy (which is proportional to the specific heat) is a continuous function of the temperature. However, the second derivative of the free energy is a discontinuous function because there is an inflection point at the temperature where the specific heat has a maximum. Therefore, SEQ5 exhibits a second order transition. One expects in this case a glass-like transition where local minima are at the same level of the global minimum. The protein can be easily trapped in any of those local minima and never reach the native state. This fact is a characteristic of a bad folder.

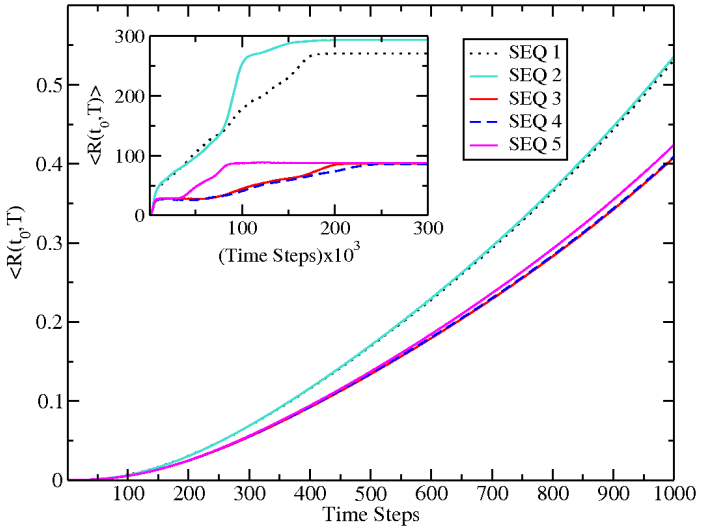


Figure 4.20: Main frame: short time behavior of the Rate of Convergence for $T = 190$K. We observe that already after the time step 200 there is a clear separation of good (SEQ1 and 2) and bad folders (SEQ 3,4 and 5). Inset: for very long times one can distinguish between good and bad folders, the top of the sequences is reached by SEQ2 after the time step 1×10^5.

The Rate of Convergence criterion for the sequences SEQ1-5 is shown in Fig. 4.20. In the outset we present the short time behavior of the Rate of Convergence at $T = 190K$, $< R(t, T = 190K) >$, for the five sequences displayed in Table 4.3. The temperature, $190K$, was chosen arbitrarily. One can observe that already at the 200 time step, the curves corresponding to the good folders SEQ1 an SEQ2 are separated from the rest of

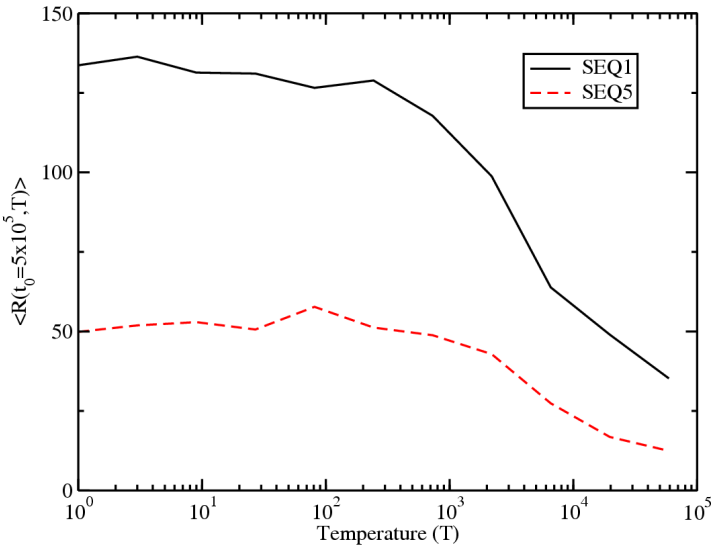


Figure 4.21: Rate of convergence for a wide range of temperatures. We observe that the distinction between a good (SEQ1) and a bad (SEQ5) folders is independent on the temperature.

the sequences which are known to be bad folders. In the inset of Fig. 4.20 we display the long time behavior of the Rate of Convergence. One can see that even for very long times, there exists a clear separation between good and bad folders. The rates of SEQ1 and SEQ2 are close to each other up to the time step 10^5. For longer times the rate of SEQ2 reaches the top. The folding time of the sequences studied in this work was $> 10^7$ time steps. This folding time is given as the time when the Rate of Convergence reaches a constant value. It is important to notice that by using the Rate of Convergence we reduced in five orders of magnitude ($10^2/10^7$) the simulation time required to know whether a sequence is a good folder. However, to know which is the best in a collection of random sequences we have only reduced in two orders of magnitude ($10^5/10^7$) the simulation time.

The Rate of Convergence at different temperatures for a good sequence (SEQ1) and a bad one (SEQ5) is shown in Fig. 4.21 with $t = 5 \times 10^5$ time steps (time chosen arbitrarily). One observes that at low temperatures the Rate of Convergence is high for both sequences but it tends to decrease as the temperature increases. That is because

the structures are more dissimilar to each other at high temperatures than at low temperatures. The important point to notice from Fig. 4.21 is that the distinction between good and bad folders is independent of the temperature at which one calculates the Rate of Convergence.

In spite of the fact that we have distinguished the good and bad folders in both the off-lattice and the lattice models by using the Rate of Convergence, we were not able to compute the folding temperature for the off-lattice model as Gridnev and Garcia did for the lattice model. The reason is probably the much smaller number of random sequences considered to make the average of the Rate of Convergence in the off-lattice simulations. In the off-lattice model only five sequences were considered while in the lattice model 800 sequences. For a future work we will make the simulations with a bigger number of sequences.

Summarizing this Section, we extended the Rate of Convergence method proposed by Gridnev and Garcia [GG] for sorting amino acid sequences by their ability to fold. The idea is applicable in all model frameworks, including accurate atomistic descriptions. For both the lattice and the off-lattice models the method showed the ability to select good folders at the very beginning of the simulations. We need to point out that by using the Rate of Convergence criterion we did not need to perform the complete dynamics but it was clear at the initial stages of the simulations which were the good and which the bad folders.

Chapter 5

SUMMARY AND OUTLOOK

In the present work we have studied three important aspects concerning the properties of proteins: 1) the confinement of a single chain inside a potential barrier which can be repulsive or attractive, 2) the interaction of a protein with a static external field and 3) the design of amino acid sequences which have a stable native state reachable in a relatively short time. In the following paragraphs we will summarize the most important results derived from the present Thesis.

The first main result of the present work is the identification of intermediate states in the folding process. These intermediates are of fundamental importance because they can even speed up or slow down the folding. The presence of intermediate states has been observed previously in simplified models of proteins [SBJ07]. However, a detailed study of intermediate states in more sophisticated protein models and the modification of these intermediates under external factors had not been done before. We have studied the peptide 1NJ0, which is a part of the HIV virus. This peptide has a β-sheet structure as the native state. We obtained the free energy landscape (FEL) of this peptide by means of the Wang-Landau algorithm. The free energy landscape shows the presence of two intermediates besides the native and the unfolded states.

The second important topic of this Thesis is regarding the influence of confinement potentials on the protein folding behavior. The confinement potentials mimic the effects of Chaperones on proteins. We found that the native, unfolded and intermediate states are modified by the presence of a confinement barrier. We studied two kinds of potential barriers, the first one which is purely repulsive and the second one which besides the repulsive part has an attractive effect on the protein. In the case of the repulsive barrier,

the intermediate states get closer to each other in the plot of the FEL as the radius of the barrier decreases. Eventually, they tend to collapse to a single minimum when the radius of the barrier is very small. Additionally, the unfolded state gets more compact in the presence of confinement. In the case of the attractive barrier, the situation is different. We observed the extinction of the intermediate and the native states when the degree of attraction is comparable to the energy necessary to brake the hydrogen bonds. For a sufficiently high magnitude of the field only the unfolded state is observed.

The third important result of the present work is related to the effects of an external field on the protein folding. We demonstrated in this Thesis that the intermediates can be modified by an external electric field. In fact, the magnitude of the electric fields used in the present work can be reached in the laboratory. From our simulations one can conclude that the presence of an external field can modify the intermediates and that for a sufficiently high electric field we can induce a new native state. In the present simulations we observed that the native state in the presence of a high field exhibits an α-helix structure. In contrast, the native state in the absence of an external field is a β-sheet structure. In our simulations the ionic contribution of the medium was neglected. This contribution could be important because of the high magnitudes used for the electric field. Some chemical reactions could occur which would modify the features in the native and intermediate states.

The fourth result concerns to the problem of sequence design. We have tested the Rate of Convergence criterion of Gridnev and Garcia [GG] on off-lattice models. The criterion allows to identify the good and bad sequences without performing the complete folding dynamics. In this way we save a lot of CPU time. We can analyze hundreds of sequences in a reasonable time because the criterion is not very time consuming. The Rate of Convergence criterion seems to be universal in the sense that it does not depend on the kind of potential energy surface. For lattice models one is able to compute the

folding temperature by means of the Rate of Convergence. This was not the case for the off-lattice model in the present Thesis because of the relatively small number of sequences considered. In a future work we will extend this number.

The present work is a contribution to the field of proteins. We have tried to make the simulations as realistic as possible but as in any computer simulation, there are several points where our work can be improve for future studies. Therefore, we consider necessary to discuss briefly in the following lines how the present work could be extended.

So far, we have discussed the influence that a static electrical field would have on the protein folding. An even more interesting case would be that of a time-dependent electric field. For such an intention we would need to develop a code of molecular dynamics to take into consideration the temporary dependency of the field. The implications derived from this study would be very varied, for quoting an example, one could observe the effects that oscillating fields proceeding from common electronic devices might generate in proteins in our body (i.e. *cellphones*).

Another possible extension of our work could be the introduction of the water molecules explicitly and the presence of ions in the aqueous environment to study the effects of chemical reactions on the thermodynamics of protein folding.

It could be also interesting to explore the problem of the influence of an external field with a protein which has a β-sheet structure as the native state and an α-helix as an intermediate state. We could observe in this case the transitions between helix and sheet using a suitable frequency for the field.

Bibliography

[AGS94] V. Abkevich, A. Gutin, and E. Shakhnovich. Specific nucleus as the transition state for protein folding: Evidence from the lattice model. *Biochemistry*, 33:10026–10036, 1994.

[Anf73] C. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181, 1973.

[Ark08] H. Arkin. Determination of the structure of the energy landscape for coarse-grained off-lattice models of folding heteropolymers. *Phys. Rev. E*, 78:041914, 2008.

[AVR02] D. Alarcon-Vargas and Z. Ronai. p53-mdm2 the affair never ends. *Carcinogenesis*, 23:541, 2002.

[Bin01] K. Binder. *Ising Model*. Kluwer Academic Publishers, 1th. edition, 2001.

[BMP08] R. E. Belardinelli, S. Manzi, and V. D. Pereyra. Analysis of the convergence of the 1/t and wang-landau algorithms in the calculation of multidimensional integrals. *Phys. Rev. E*, 78:067701, 2008.

[BN91] B. Berg and T. Neuhaus. Multicanonical algorithms for first order phase transitions. *Phys. Lett. B*, 267:249–253, 1991.

[BNO08] C. Blish, M. Nguyen, and J. Overbaugh. Enhancing exposure of hiv-1 neutralization epitopes through mutations in gp41. *PLoS Medicine*, 5:e9, 2008.

[BOSW95] J. Bryngelson, N.J. Onuchic, D. Socci, and P.G. Wolynes. Funnels, pathways, and the energy landscape of protein folding- a synthesis. *Prot. Struc. Func. Gen.*, 21:167–195, 1995.

[BP07] R. E. Belardinelli and V. D. Pereyra. Fast algorithm to calculate density of states. *Phys. Rev. E*, 75:046701, 2007.

[BPZ+07] M. Borovinskaya, R. Pai, W. Zhang, B. Schuwirth, J. Holton, G. Hirokawa, H. Kaji, A. Kaji, and J. Cate. Structural basis for aminoglycoside inhibition of bacterial ribosome recycling. *Nat. Struct. Mol. Biol.*, 14:727, 2007.

[BTR+99] R. Brolia, G. Tiana, H. Roman, E. Vigezzi, and E. Shakhnovich. Stability of designed proteins against mutations. *Phys. Rev. Lett.*, 82:4727, 1999.

[CGO02] M. Cheung, A. Garcia, and J. Onuchic. Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Nat. Aca. Sci. USA*, 99:685, 2002.

[Cle08] C. Clementi. Coarse-grained models of protein folding: Toy models or predictive tools? *Curr. Opin. Struc. Biol.*, 18:10, 2008.

[CLST09] Y. Crespo, A. Laio, G. Santoro, and E. Tosatti. Calculating thermodynamics properties of quantum systems by a non-markovian monte carlo procedure. *Phys. Rev. E*, 80:015702, 2009.

[CMB98] C. Clementi, A. Maritan, and J. Banavar. Folding, design, and determination of interaction potentials using off-lattice dynamics of model heteropolymers. *Phys. Rev. Lett.*, 81(5):3287–3290, 1998.

[Con99] MHC Sequencing Consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature*, 401:921, 1999.

[Cre92] T. Creighton. *Protein Structure and Molecular Properties*. Freeman, New York, 1th. edition, 1992.

[CSM06] N.-Y. Chen, Z.-Y. Su, and C.-Y. Mou. Effective potentials for folding proteins. *Phys. Rev. Lett.*, 96:078103–078107, 2006.

[CWD+07] L. Cai, X. Wang, Y. Darici, J. Zhang, and P. Dowben. Energetics of the dipole flip-flop motion in a ferroelectric polymer chain. *Jour. Chem. Phys.*, 126:124908, 2007.

[DCJP04] N. Douarche, F. Calvo, P. Jensen, and G. Pastor. Model simulations of ground-state and finite-temperature properties of disordered magnetic nanostructures. *Eur. Phys. Jour. D*, 24:77, 2004.

[Dil99] K. Dill. Polymer principles and protein folding. *Prot. Science*, 8:1166, 1999.

[DWK98] Y. Duan, L. Wang, and P. Kollman. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. *Proc. Nat. Aca. Sci. USA*, 95(17):9897–9902, 1998.

[EAF+06] M. Etienne, J. Aucoin, Y. Fu, R. McCarley, and R. Hammer. Stoichiometric inhibition of amyloid -protein aggregation with peptides containing alternating a,a,-disubstituted amino acids. *Jour. Am. Chem. Soc.*, 128(11):3522–3523, 2006.

[Ell06] R. Ellis. Inside the cage. *Nature*, 442:360–362, 2006.

[EML06] A. Esztermann, R. Messina, and H. Löwen. Localisation-delocalisation transition of electrolytes between micro-electrodes. *Europhys. Lett.*, 73:864, 2006.

[FFC06] S. Ferreira, F. Felice, and A. Chapeaurouge. Metaestable, partially folded states in the productive folding and in the misfolding and amyloid aggregation of proteins. *Cell Biochem. and Biophys.*, 44:539, 2006.

[FH97] W. Fenton and A. Horwich. Groel-mediated protein folding. *Prot. Sci.*, 6(2):743–760, 1997.

[FIW02] F. Favrin, A. Irbäck, and S. Wallin. Folding of a small helical protein using hydrogen bonds and hydrophobicity forces. *Prot. Struc. Funct. Gen.*, 47:99, 2002.

[FP02] A. Finkelstein and O. Ptitsyn. *Protein Physics: A Course of Lectures*. Academic Press, New York, 1th. edition, 2002.

[FS06] F. Fang and L. Szleifer. Controlled release of proteins from polymer-modified surfaces. *Proc. Nat. Aca. Sci. USA*, 103(15):5769–5774, 2006.

[GG] D. Gridnev and M. E. Garcia. Method for selection and design of proteins *unpublished*.

[Gil93] M. Gilson. Multiple site titration and molecular modeling 2 rapid methods for computing energies and forces for ionizable groups in proteins. *Prot. Struc. Func. Gen.*, 15:266, 1993.

[GLSW92] R. Goldstein, Z. Luthey-Schulten, and P. Wolynes. Optimal protein folding codes from spin-glass theory. *Proc. Nat. Aca. Sci. USA*, 89:4918, 1992.

[Gre98] I. Greenwald. Lin-12/notch signaling: lessons from worms and flies. *Genes Dev.*, 12:1751, 1998.

[GT96] Z. Guo and D. Thirumalai. Kinetics and thermodynamics of folding of a it de Novo designed four-helix bundle protein. *Jour. Mol. Biol.*, 263:323, 1996.

[GW94] K. Gulukota and P. Wolynes. Statistical mechanics of kinetics proof reading in protein folding *in vivo*. *Proc. Nat. Aca. Sci. USA*, 91:9292–9296, 1994.

[HD75] H. Hilhorst and J. Deutch. Analysis of monte carlo results on the kinetics of lattice polymer chains with excluded volume. *Jour. Chem. Phys.*, 63:5153, 1975.

[HGG+96] G. Humer, S. Garde, A. Garcia, A. Phorille, and L. Pratt. An information theory model of hydrophobic interactions. *Proc. Nat. Aca. Sci. USA*, 93:8951, 1996.

[HGG+98] G. Humer, S. Garde, A. Garcia, M. Paulaitis, and L. Pratt. The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins. *Proc. Nat. Aca. Sci. USA*, 95:1552, 1998.

[Hol85] W. Hol. The role of the alpha helix dipole in the protein function and structure. *Progress in Biophys. and Mol. Biol.*, 45:149, 1985.

[HRL08] J. Hernandez-Rojas and J. Llorente. Microcanonical versus canonical analysis of protein folding. *Phys. Rev. Lett.*, 100:258104, 2008.

[ISW00] A. Irbäck, F. Sjunnesson, and S. Wallin. Three-helix-bundle protein in a ramachandran model. *Proc. Nat. Aca. Sci. USA*, 97:13614, 2000.

[JB96] R. Jernigan and I. Bahar. Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.*, 6:195, 1996.

[JBS04] A. Jewett, A. Baumketner, and J. Shea. Accelerated folding in the weak hydrophobic environment of a chaperonin cavity: Creation of an alternate fast folding pathway. *Proc. Nat. Aca. Sci. USA*, 101(36):13192–13197, 2004.

[KB90] P. Kim and R. Baldwin. Intermediates in the folding reactions of small proteins. *Annu. Rev. Biochem.*, 59:631–660, 1990.

[Kel98] J. Kelly. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.*, 8(1):101–106, 1998.

[KGV83] S. Kirkpatrick, C. Gelat, and M. Vecci. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[KHE04] J. Kubelka, J. Hofrichter, and W. Eaton. The protein folding speed limit. *Curr. Opin. Struct. Biol.*, 14:76–88, 2004.

[KHLE04] M. Krishna, L. Hoang, Y. Lin, and S. Englander. Hydrogen exchange methods to study protein folding. *Methods*, 34:51, 2004.

[Kie95] T. Kiefhaber. Kinetic traps in lysozyme folding. *Proc. Nat. Aca. Sci. USA*, 92:9029, 1995.

[KKZK06] N. Kachel, W. Kremer, R. Zahn, and H. Kalbitzer. Observation of intermediate states of the human prion protein by high pressure nmr spectroscopy. *BMC Struc. Biol.*, 6:1, 2006.

[LBK+05] H. Lodish, A. Berk, C. Kaiser, M. Krieger, M. Scott, and A. Bretscher. *Molecular Cell Biology*. Macmillan, 6th. edition, 2005.

[LBTY06] F. Legge, A. Budi, H. Treutlein, and I. Yarovsky. Protein flexibility: Multiple molecular dynamics simulations of insulin chain b. *Biophys. Chem.*, 119:146, 2006.

[LD89] F. Lau and K. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22:3986, 1989.

[Lev68] C. Levinthal. Are there pathways for protein folding? *Jour. Chem. Phys. et Phys.-Chem. Biol.*, 65, 1968.

[LG08] A. Laio and F. Gervasio. Metadynamics: A method to simulate rare events and reconstruct the free energy in biophysics, chemistry and materials science. *Rep. Prog. Phys.*, 71:126601, 2008.

[LLW06] D. Lu, Z. Lu, and J. Wu. Structural transitions of confined model proteins: Molecular dynamics simulation and experimental validation. *Biophys. Jour.*, 90:3224–3238, 2006.

[LM00] D. Lynn and S. Meredith. Review: Model peptides and the physicochemical approach to beta-amyloids. *Jour. Struc. Biol.*, 130(2-3):153–173, 2000.

[LO92] P.E. Leopold and N.J. Onuchic. Protein folding funnels - a kinetic approach to the sequence relationship. *Proc. Nat. Aca. Sci. USA*, 89:8721–8725, 1992.

[MC06a] S. Matysiak and C. Clementi. Minimalist protein model as a diagnostic tool for misfolding and aggregation. *Jour. Mol. Biol.*, 363:297, 2006.

[MC06b] L. Mazzoni and L. Casetti. Curvature of the energy landscape and folding of model proteins. *Phys. Rev. Lett.*, 97:218104, 2006.

[ME04] J. Meller and R. Elber. Protein recognition by sequence-to-structure fitness. In R. Friesner, editor, *Computational Methods for Protein Folding*, page 114. Willey, 2004.

[MJ96] S. Miyazawa and R. Jerningan. Residue residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Jour. Mol. Biol.*, 256(3):623–644, 1996.

[NSC06] A. Netto, C. Silva, and A. Caparica. Wang-landau sampling in three-dimensional polymers. *Braz. Jour. Phys.*, 36(3A):619–622, 2006.

[OLYG09] P. Ojeda, A. Londono, N. Yow, and M. Garcia. Monte carlo simulations of proteins in cages: Influence of confinement on the stability of intermediate states. *Biophys. Jour.*, 96:1076, 2009.

[PGT95] V. Pande, A. Grosberg, and T. Tanaka. Freezing transition of random heteropolymers consisting of an arbitrary set of monomers. *Phys. Rev. E*, 51:3381, 1995.

[PK74] P. Privalov and N. Khechinashvili. A thermodynamic approach to the problem of stabilization of globular protein structure: A calorimetric study. *Jour. Mol. Biol.*, 86:665, 1974.

[PRF02] I. Prigogine, S. Rice, and R. Friesner. *Advances in Chemical Physics, Computational Methods for Protein Folding*. Wiley, 1th. edition, 2002.

[RCF+09] P. Reckenthaeler, M. Centurion, W. Fuss, S. Trushin, F. Krausz, and E. Fill. Time-resolved electron diffraction from selectively aligned molecules. *Phys. Rev. Lett.*, 102:213001, 2009.

[RKP05] N. Rathore, T. Knotts, and J. Pablo. Confinement effects on the thermodynamics of protein folding: Monte carlo simulations. *Biophys. Jour.*, 90:1767–1773, 2005.

[RSS08] L. Randau, L. Schroeder, and D. Soll. Life without rnase p. *Nature*, 453:120, 2008.

[RTG+07] S. Riccardo, G. Tortoriello, E. Giordano, M. Turano, and M. Furia. The coding/non-coding overlapping architecture of the gene encoding of the drosophila pseudourine synthase. *BMC Mol. Biol.*, 8:15, 2007.

[SBJ07] S. Schnabel, M. Bachmann, and W. Janke. Two-state folding, folding through intermediates, and metastability in a minimalistic hydrophobic-polar model for proteins. *Phys. Rev. Lett.*, 98:048103, 2007.

[Sch02] T. Schlick. *Molecular Modeling and Simulation*. Springer, 1th. edition, 2002.

[SF00] G. Solomons and C. Fryhle. *Organic Chemistry*. John Wiley & Sons, Oxford, 7th. edition, 2000.

[SG93] E. Shakhnovich and A. Gutin. Engineering of stable and fast-folding sequences of model proteins. *Proc. Nat. Aca. Sci. USA*, 90:7195, 1993.

[Sha94] E. Shakhnovich. Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett.*, 72:3907, 1994.

[SSK94a] A. Sali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369:248, 1994.

[SSK94b] A. Sali, E. Shakhnovich, and M. Karplus. Kinetics of protein folding - a lattice model study of the requirements for folding to the native-state. *Jour. Mol. Biol.*, 235:1614, 1994.

[SSK94c] A. Sali, E. Shakhnovich, and M. Karplus. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. *Jour. Mol. Biol.*, 235:1614, 1994.

[SSW+05] J. Spence, K. Schmidt, J. Wu, G. Hembree, U. Weierstall, B. Doak, and P. Fromme. Diffraction and imaging from a beam of laser-aligned proteins: resolution limits. *Acta Cryst. A*, 61:237, 2005.

[Str05] M. Streek. *Brownian Dynamics Simulatio of Migration of DNA in structured Microchannels*. PhD. Thesis, 1th. edition, 2005.

[SW86] R. Swendsen and J. Wang. Replica monte carlo simulation of spin glasses. *Phys. Rev. Lett.*, 57:2607, 1986.

[Thi94] D. Thirumalai. Theoretical perspectives on *in vitro* and *in vivo* folding. *J. Struct. Biol.*, pages 115–134, 1994.

[TKL03] D. Thirumalai, D. Klimov, and G. Lorimer. Caging helps proteins fold. *Proc. Nat. Aca. Sci. USA*, 100(20):11195–11197, 2003.

[TKL06] D. Thirumalai, D. Klimov, and G. Lorimer. Nanopore-protein interactions dramatically alter stability and yield of the native state in restricted spaces. *Jour. Mol. Biol.*, 357(20):632–643, 2006.

[TKT03] F. Takagi, N. Koga, and S. Takada. From the cover: How protein thermodynamics and folding mechanisms are altered by the chaperonin cage: Molecular simulations. *Proc. Nat. Aca. Sci. USA*, 100(20):11367–11372, 2003.

[TLBZ09] F. Toschi, F. Lugli, F. Biscarini, and F. Zerbetto. Effects of electric field stress on a beta-amyloid peptide. *Jour. Phys. Chem.*, 113:369, 2009.

[TPB09] M. Taylor, W. Paul, and K. Binder. Phase transitions of a single polymer chain: A wang-landau simulation study. *Jour. Chem. Phys.*, 131:114907, 2009.

[TSW99] S. Takada, Z. Schulten, and P. Wolynes. Folding dynamics with nonadditive forces: A simulation study of a designed helical protein and a random heteropolymer. *Jour. Chem. Phys.*, 110:11616, 1999.

[VND99] M. Vendruscolo, R. Najmanovich, and E. Domany. Protein folding in contact map space. *Phys. Rev. Lett.*, 82:656, 1999.

[Wad76] A. Wada. The alpha-helix as an electric macro-dipole. *Adv Biophys.*, 9:1, 1976.

[Wai07] T. Waigh. *Applied Biophyics: A Molecular Approach for Physical Scientists.* Wiley-Interscience, 1th. edition, 2007.

[Wal03] D. Wales. *Energy Landscapes with Applications to Clusters, Biomolecules and Glasses.* Cambridge University Press, 1th. edition, 2003.

[WBCJ04] H. Went, C. Bentez-Cardoza, and S. Jackson. Is an intermediate state populated on the folding pathway of ubiquitin? *FEBS Letters*, 567:333, 2004.

[WL00] Z. Wang and H. Lee. Origin of the native driving force for protein folding. *Phys. Rev. Lett.*, 84(3):574 – 577, 2000.

[WL01] F. Wang and D. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86:2050–2053, 2001.

[XPS96] D. Xu, J. Phillips, and K. Schulten. Protein response to external electric fields: Relaxation, hysteresis, and echo. *Jour. Phys. Chem.*, 100:12108, 1996.

[XS99] Z. Xu and P. Sigler. Groel/groes: Structure and function of a two-stroke folding machine. *J. Struct. Biol.*, 124(2):129–141, 1999.

[Yua07] T. Yuan. Electroporation: an arsenal of application. *Cytotech.*, 54:71, 2007.

[ZS08] C. Zhou and J. Su. Optimal modification factor and convergence of the wang-landau algorithm. *Phys. Rev. E*, 78:046705, 2008.

Publications related to this thesis

P. Ojeda, A. Londono, N.-Y. Chen and M. Garcia,

Monte Carlo Simulations of Proteins in Cages: Influence of Confinement on the Stability of Intermediate States.

Biophysical Journal, **96**, 1076 (2009).

P. Ojeda and M. Garcia,

Influence of an External Field on the Secondary Structure of Proteins (Sent to Biophysical Journal).

D. Gridnev, P. Ojeda and M. Garcia,

Method for Selection and Sequence Design for Proteins (In preparation).

Acknowledgements

My acknowledgments will go first to my thesis supervisor Prof. Dr. Martin E. Garcia. I am deeply indebted to him, whose patience and help were very important during this work. This thesis would not be achieved without his stimulating suggestions and encouragements. I am also grateful to Drs. Nan-Yow Chen and Dmitry Gridnev, for their help and nice collaboration. The thesis would not be possible without their assistance. I express also many thanks to MSc. Aurora Londoño for her help and nice suggestions on the part regarding the simulations of chaperones.

I thank Prof. Dr. Burkhard Fricke, Prof. Dr. Friedrich W. Herberg and Prof. Dr. Arno Ehresmann for their interest about my work.

I want to thank the university of Kassel in general for giving me the opportunity to do this thesis, and particularly the personnel of the physics department.

This work is dedicated to my wife Nadia my mother, my syster, for the long and sincere support. I am particularly very indebted to my wife Nadia Teresa for his care and love. I cannot ask for more from my mother, Rosa, as she is simply wonderful. I have no suitable word that can fully describe her love and support to me.

This work would not be possible without the generous support of the Deutsche Akademischer Austausch Dienst (DAAD) which allowed me to realize my ambition to do my thesis in Germany.

Curriculum Vitae

Name Pedro Armando Ojeda May

Date of Birth June 29, 1980

Place of Birth Merida, Yucatan, MEXICO

Education

2006-2010: Institut für Physik,
Universität Kassel / Germany,
Ph.D Student

2003-2005: Instituto Potosino de Investigación Científica y Tecnológica,
San Luis Potosi / MEXICO,
Master of Research : Nanoscience and Nanotechnology.

2002-2003: Universidad Autónoma de Yucatán,
Yucatán / MEXICO,
Bachelor of sciences: Physics.

1999-2001: Preparatoria 2 de la Universidad Autónoma de Yucatán,
Yucatán / MEXICO,
High School Diploma

Erklärung

Hiermit versichere ich, daß ich die vorliegende Dissertation selbständig und ohne unerlaubte Hilfe angefertigt und andere als die in der Dissertation angegebenen Hilfsmittel nicht benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen sind, habe ich als solche kenntlich gemacht. Kein Teil dieser Arbeit ist in einem anderen Promotions- oder Habilitationsverfahren verwendet worden.

March 29, 2010, Kassel

Pedro Ojeda May