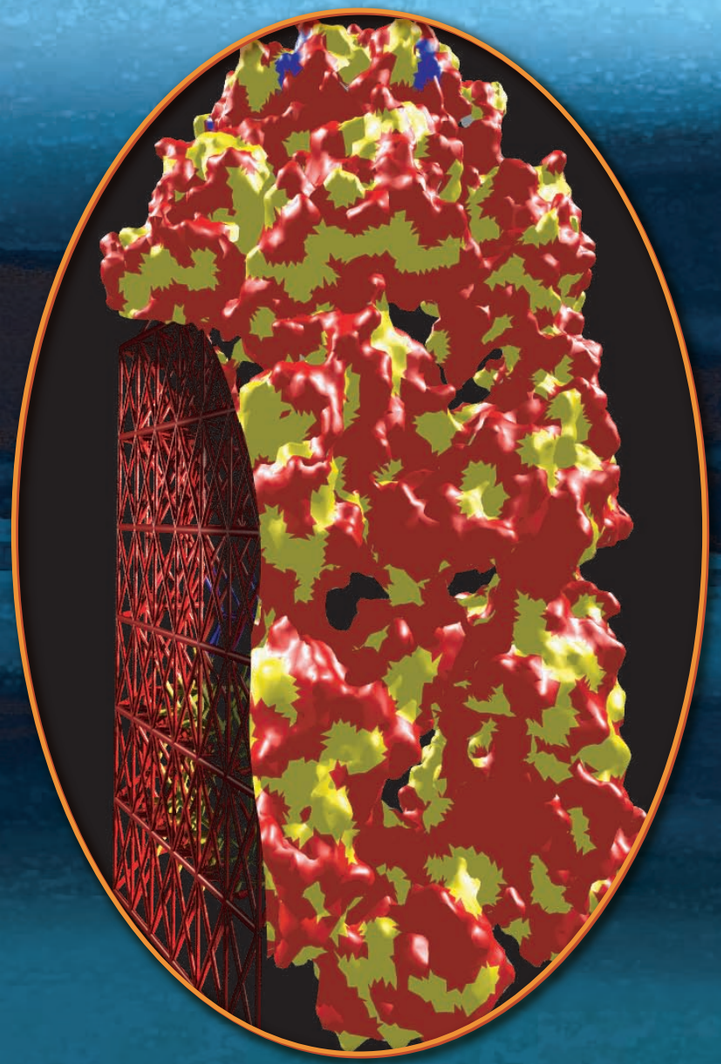
Proteins are the ultimate nano-machines. Moreover, they provide the building blocks for complex structures. Many protein-based machines and materials have evolved during the past 4 billion years to perform specific tasks with high selectivity. The aim of the present thesis is to gain a better understanding of how proteins function, by studying the behavior of simple model proteins. We show how by artificially evolving their amino acids sequence, it is possible to control their behavior.



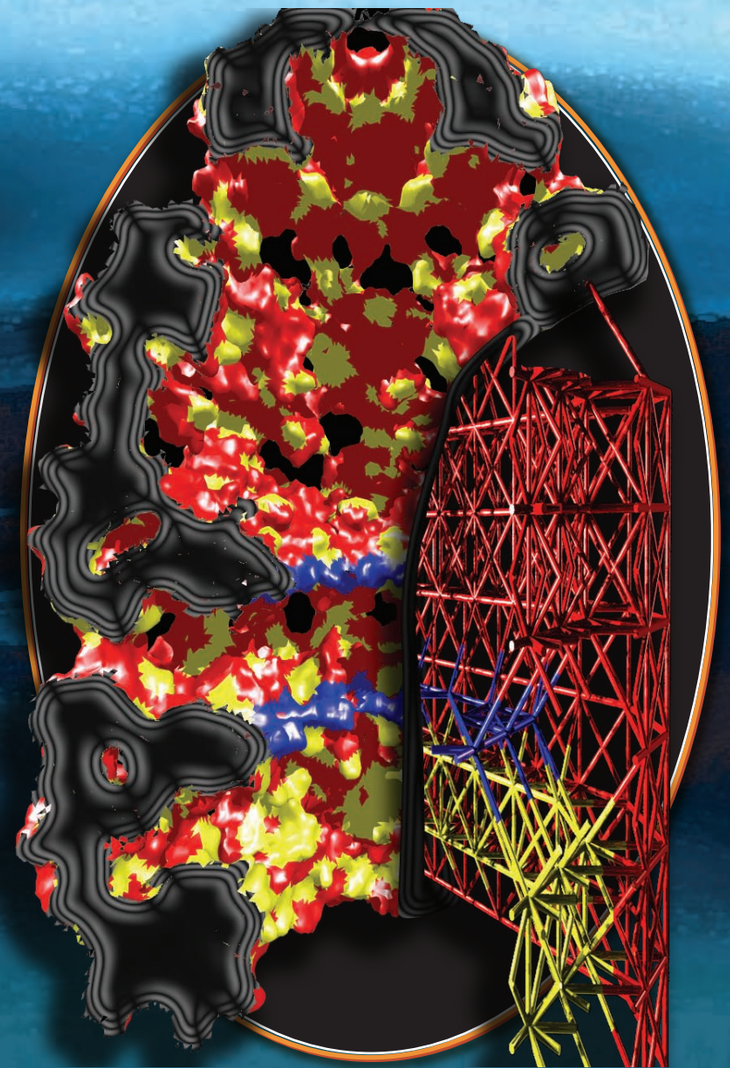
Cover : Artistic superimposition of the X-Ray structure of the GroEL/GroES complex with the cage structure used to model such a system in computer simulations

WRITING WITH AMINO ACIDS

IVAN COLUZZA

WRITING WITH AMINO ACIDS

DESIGNING THE FOLDING AND BINDING OF MODEL PROTEINS



IVAN COLUZZA

Writing with amino acids: designing the folding and binding of model proteins

Writing with amino acids: designing the folding and binding of model proteins

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit van Amsterdam op gezag van de Rector Magnificus prof. mr. P. F. van der Heijden ten overstaan van een door het college voor promoties ingestelde commissie, in het openbaar te verdedigen in de Aula der Universiteit op donderdag 23 juni 2005, te 11:00 uur.

door

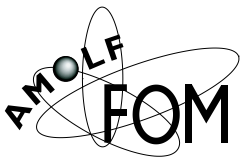
Ivan Coluzza

geboren te Rome, Italië

Promotor: prof. dr. D. Frenkel

Co-promotoren: prof. dr. H.G. Muller

Faculteit: Natuurwetenschappen, Wiskunde en Informatica



The work described in this thesis was performed at the FOM Institute for Atomic and Molecular Physics, Kruislaan 407, 1098 SJ, Amsterdam, The Netherlands. The work is part of the research program of the Stichting voor Fundamenteel Onderzoek der Materie (FOM) and was made possible by financial support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

A Bea
Alla mia famiglia

The work in this thesis covers the following pubblications

Chapter 3:
Virtual-move Parallel Tempering
(Accepted for publication in ChemPhysChem)
Coluzza I, Frenkel D

Chapter 4:
Designing refoldable model molecules,
(Phys. Rev. E 68 (4): Art. No. 046703 Part 2 OCT 2003)
Coluzza I, Muller HG, Frenkel D

Chapter 5:
Designing specificity of protein-substrate interactions
(Phys. Rev. E 70 (50): Art. No. 051917 Nov 2004)
Coluzza I, Frenkel D

Chapter 6:
Refoldable proteins and substrate interaction
*(*In preparation*)*
Coluzza I, Frenkel D

Chapter 7:
Translocation boosts efficiency of double-barreled chaperonins
*(*In preparation*)*
Coluzza I, Frenkel D

Contents

1 Introduction

Proteins are the ultimate nano-machines. Moreover, they provide the building blocks for complex structures. Many protein-based machines and materials have evolved during the past 4 billion years. For instance proteins can assemble into complex structures that can reach macroscopic sizes. Examples of such structures are actin filaments (Fig. 1.1) and microtubules. These protein-based structural units are responsible for the structure and elastic properties of many cells. Proteins perform specific tasks with high selectivity. The aim of the present thesis is to gain a better understanding of how proteins function, by studying the behavior of simple model proteins.

Proteins are heteropolymers composed of 20 different types of amino acids. Different proteins have different chain lengths. Depending on the amino-acid sequence, some proteins can collapse to form a well-defined "native" conformation whilst others can not. This process of forming a compact, native structure is called folding. Usually, proteins are only biologically active in their native state. The tasks that proteins perform are very diverse; they usually involve the interaction with other proteins or other biomolecules, such as DNA. These interactions are controlled by the same elements that encode for the native structure of the protein itself. Structure and function are, as consequence, strongly correlated, and are directly dependent on the sequence of amino acids along the protein chain. A better understanding of the relation between sequence, structure and function, is crucial for the design of biomolecular materials and molecular machines.

Another way to look at functional proteins is as information carriers. Proteins are an essential part of gene-regulatory networks that control the response of the cell to the environment. A simple way to understand the function of regulatory networks is to picture the stimuli coming from the world outside the cell membrane (e.g signaling proteins, changes in food concentrations...) as data input in a computer program that will generate a different output, depending on which conditions are satisfied. The code for the program *and* the design of the computer hardware are written in the DNA. The role of the network is to transmit the signal and translate it from a the chemical language to the expression of proteins that are coded for by the genetic material. Once the signal reaches the DNA, the output is provided in the form of promotion or repression of the expression of specific proteins that will then start a new cascade of reactions to translate back the answer from the genetic code to necessary chemical reactions. This signal-processing property can be achieved only if each element is designed to interact selectively with its molecular partners, otherwise the signal would generate "crosstalk" with potentially deleterious consequences. It is important to remember that such selectivity was designed through the evolution of the amino acids sequence of the protein. An example of the gene-regulatory network of the E-Coli bacterium is shown in schematically in Fig. 1.2.

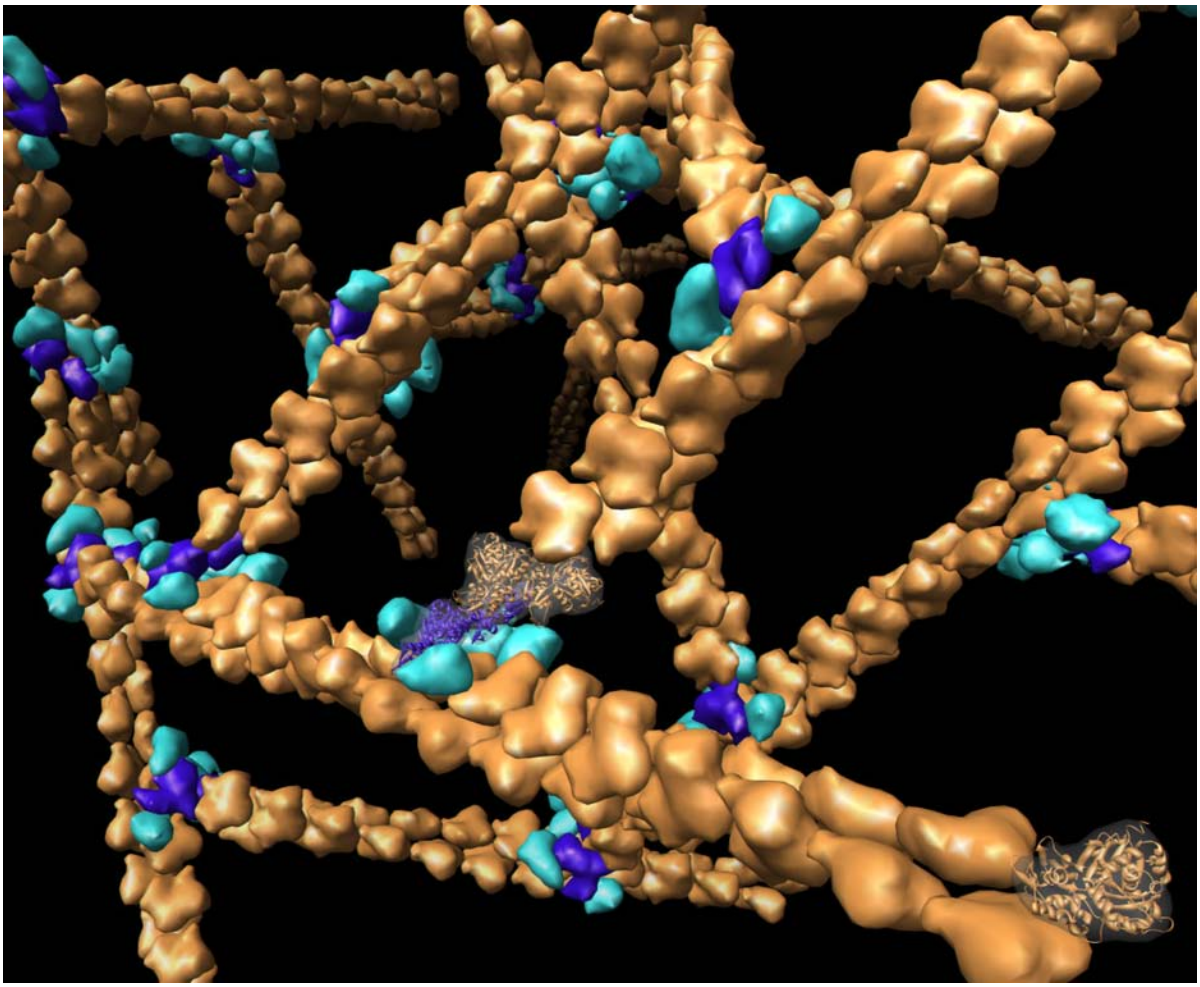


Figure 1.1: One of the most famous example of self assembly in nature, the actin filaments are long chain of the same repeating unit protein. This filaments are the building blocks of the cytoskeleton of many cells, and together with myosin is the most important components of muscular fibers in all animals [1] (http://www.cgl.ucsf.edu/chimera/ImageGallery/entries/actin/actin.html).

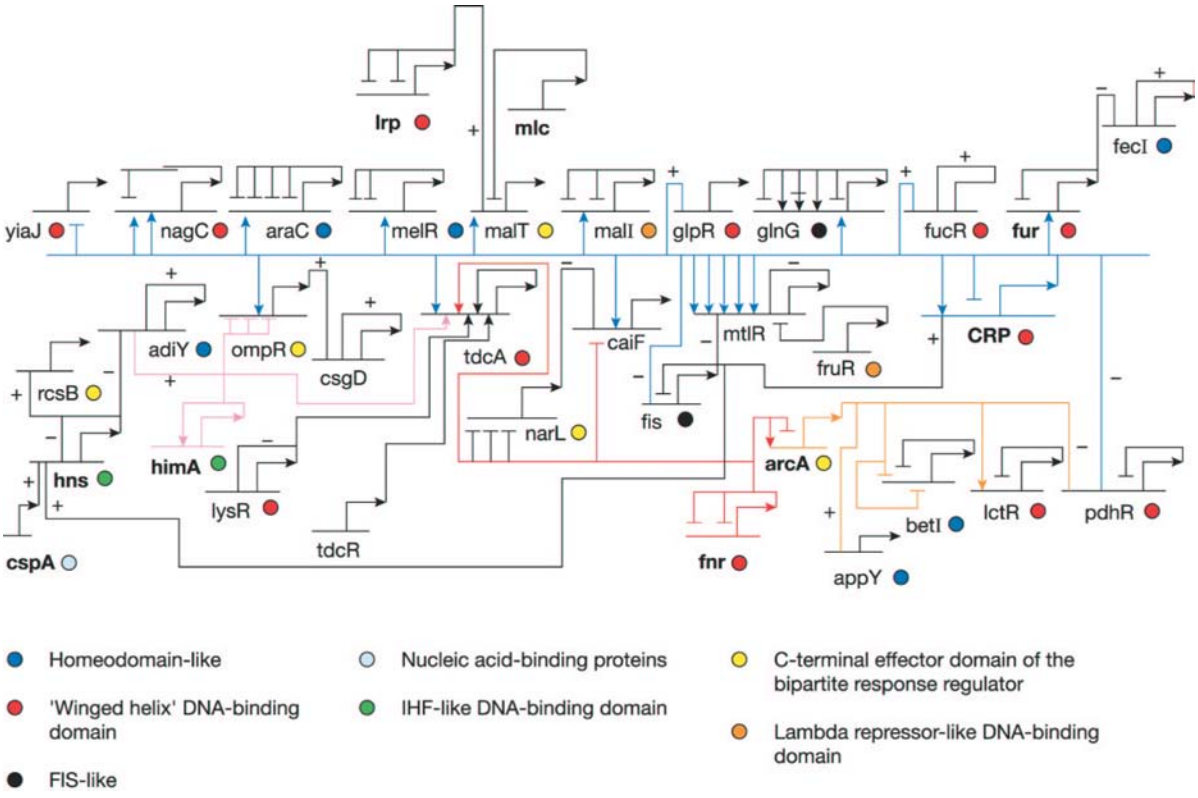


Figure 1.2: Gene Regulatory Network (GRN) of TFs in E-coli. This figure illustrates the complexity of the core of the GRN in E-coli, and gives a glimpse of the number of interaction that are involved in a subset of physiological activities of one of the simplest living organisms [2].

There are two ways to approach the modeling of the relation between sequences and function. One is to focus on the properties of specific proteins. Such an approach requires an accurate, atomistic model of the protein and its surrounding medium. The other approach focuses on the more general question how the heterogeneity of a heteropolymer can lead to both folding and specific interactions. In such studies, one can use simpler (cheaper) models to reproduce the specificity of proteins. The reason why we use a simplified model is that the numerical study of conformational changes in proteins tends to be computationally very demanding. The power of computers has only recently reached the level where it becomes feasible to simulate the folding of a single, relatively short, protein. However, for a systematic study of the relation between sequence and conformational change, it is necessary to explore the properties of a great number of molecules with different amino-acid sequences. For such studies, fully atomistic models are not an option. It should be stressed that a protein is not only optimized to fold into a specific structure, or to undergo a specific conformational change (as in a motor protein). Its is also designed to interact strongly with specific substrates but only weakly, if at all, with all other molecules that it encounters as it diffuses through the cell. Clearly these conditions increase the complexity of protein design.

In what follows, we will use lattice heteropolymers to represent the peptide chain. This model is one of the simplest representation of a protein and has been extensively used for the study of folding properties. In Chapter 2 we describe the methods that we developed to artificially evolve the model protein to perform specific, elementary tasks. Once a protein has been "designed" the next step is to study the conformational space of the protein to test if it performs as designed. In Chapter 3 we describe the algorithm that we used to sample the free energy landscape of the heteropolymer even in regions of high free energy. In Chapter 4, we describe the application of our model to the study of conformational changes in proteins induced by a chemical agent. We pay special attention to important role of thermal fluctuations that allow the system to reach state at which the cost of the transition to one structure to the other one is considerably reduced. In Chapter 5 we extend our design technique to include specific binding properties between a protein and a substrate. We study the influence of such substrates on the transition between two conformations and on the folding properties of random domains of proteins (Chapter 6). The final step is the study of a particular class of protein called chaperonins that act as folding catalyzers. We introduce a model for chaperonin action that suggest that protein translocation is a key step by which chaperonins refold misfolded proteins into their native state (Chapter 7).

2 Artificial Evolution of a Lattice Heteropolymer

The typical size of single-domain protein is of the order of 10^3atoms (see Fig. 2.1). This number of atoms does not included the surrounding solvent that is, however, an essential element of the protein activity. Upon inclusion of the first shell of water the number of atoms increases by a factor ten .

Using the fastest computers available,it is now becoming possible to perform fully atomistic simulations of the folding of small proteins. However, as explained in the introduction, our aim is different: we wish to explore the constraints placed upon protein design by the requirements of foldability and function. For such studies it is advisable to use a highly simplified protein model. In what follows, we describe proteins as heteropolymers "living" on a 3D cubic lattice.

2.1 General properties of heteropolymers on a lattice

This lattice-polymer model for proteins is highly simplified. First of all, the side chain of the single amino acids are not taken into account, leaving the protein as necklace of beads with isotropic interactions. The major effect of this approximation is to ignore the entropic contribution of the side chain and to reduce the effect of steric hindrance. The second approximation consists in constraining the residues of the chain on a cubic lattice of unit side length. Constraining a polymer to a lattice does not change the topology of the chain, but all internal degrees of the monomeric units, as well as the vibrations of the bonds between them, are ignored. The model assumes that there are only nearest-neighbor interactions between the amino acids. The total configurational energy of a particular sequence in a given structure is given by

$$E = \sum C_{ij} S_{ij} \,, \tag{2.1}$$

where i and j are particle indices, C is the contact map defined as

$$C = \begin{cases} 1 & \text{if i neighbor of j} \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.2}$$

and S is the interaction matrix. For S we use the 20 by 20 matrix fitted by Miyazawa and Jernigan [3]. From the PDB database these authors extracted the frequency of contacts for each pair of amino acids in a wide range of different proteins. The formation of a bond in

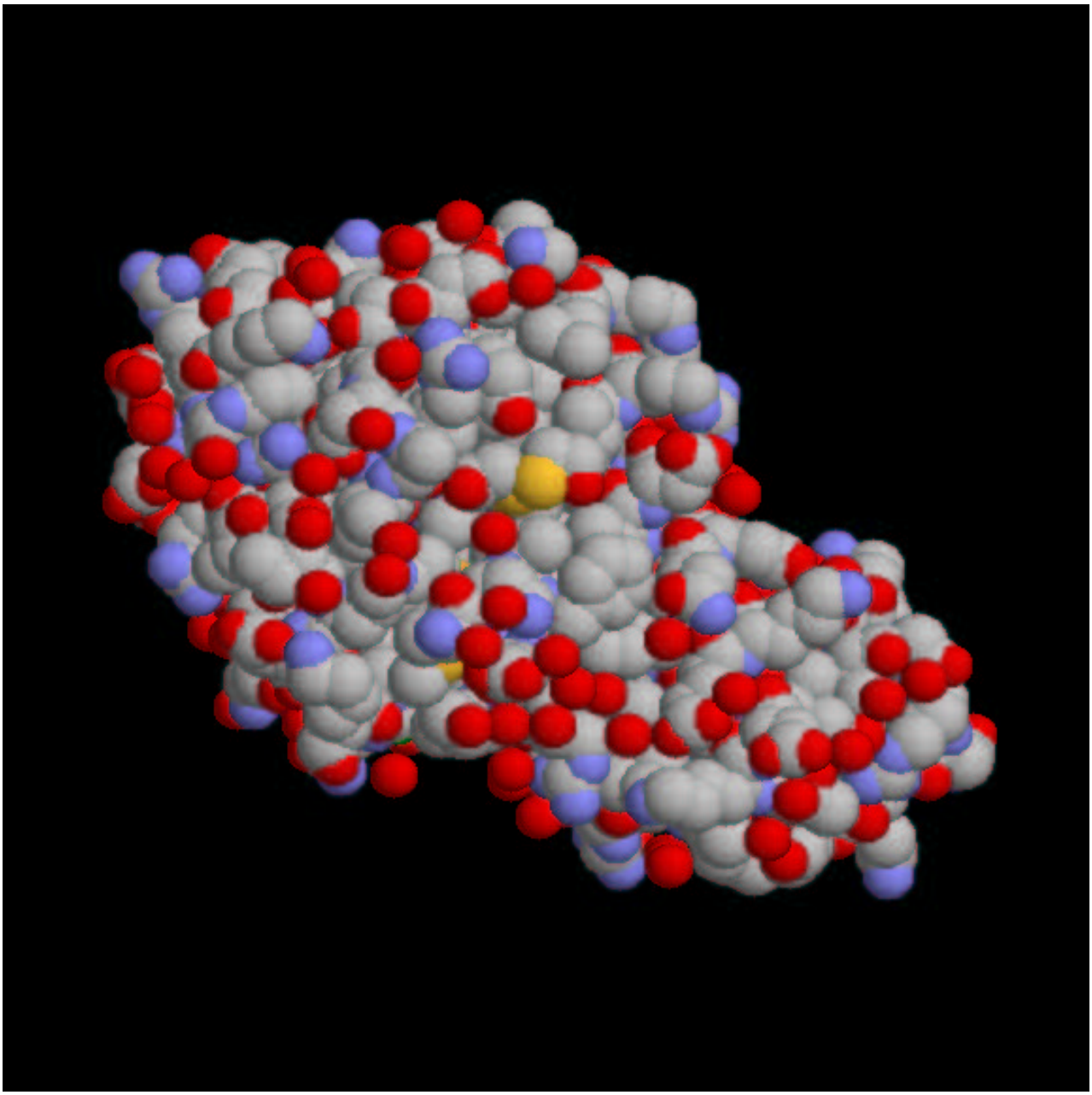


Figure 2.1: Hen egg white lysozyme is a single chain of 129 residues (~1000 atoms). It has an alpha+beta fold, consisting of five to seven alpha helices and a three-stranded antiparallel beta sheet. The enzyme is approximately ellipsoidal in shape, with a large cleft in one side forming the active site. Lysozyme on of the better known single domain proteins

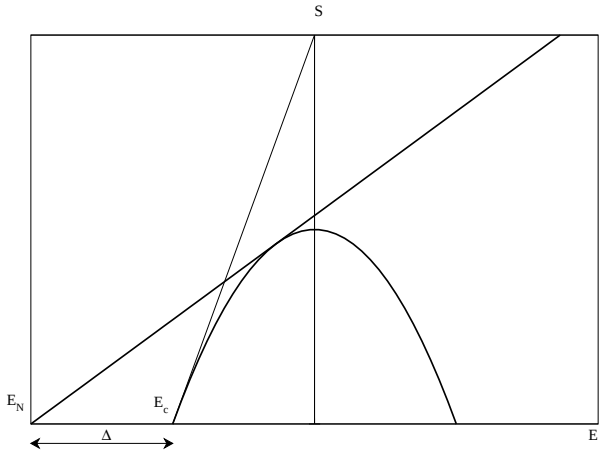


Figure 2.2: Energy spectrum of a heteropolymer on a lattice given by Eq. 2.3. On the horizontal axis is the configurational energy (Eq. 3.9), while on the vertical axis is the entropy. E_N is the energy of the native state, while E_c is the crossing point of the parabola with the abscissa. Δ the region of discrete states. The slope of the tangent passing through E_N defines the folding temperature, while the tangent in E_c gives the glass temperature.

the collapsed polymer can be schematically represented by the following chemical reaction, $(\sigma - *) + (\pi - *) \rightleftharpoons (\sigma - \pi) + (* - *)$ where $(\sigma - *)$is the average binding energy between an amino acids of type σ and all the other types of amino acids. Using the quasi-chemical approximation Miyazawa and Jernigan estimated the free energy associated with the formation of a bond between all possible pairs of amino acids. It should be stressed that it is not at all obvious that the frequency of contacts between amino-acids is a direct reflection of their binding strength. For one thing, it is obvious that the "training" set from which the interaction energies were deduced is a subset (and probably not a random subset) of all existing protein structures. Note that here and in the next chapters the energy is expressed in units of kT relative to the energies in the interaction matrix.

A given lattice polymer can form a large number of compact conformations. Obviously, every conformation is characterized by a different contact map. Hence, the energy of the polymer depends on its conformation. In a mean field approximation the energy spectrum of the compact structures of a random chain on a lattice has the shape shown in Fig. 2.2.

The mean field expression for the entropy is [4, 5, 6]

$$\mathcal{S}(E) = \begin{cases} N \ln \gamma - \frac{E^2}{2N\sigma_{\mathrm{B}}^2} & \text{if } E > E_c \\ 0 & \text{if } E \leq E_c \end{cases} \tag{2.3}$$

[1] where N is the number of elements in the chain, σ_B is the standard deviation of the interaction matrix, and γ is the coordination number for fully compact structures on the lattice. E_c is the

[1] In the definition of the entropy the contribution of the quantity $\sqrt{\pi\sigma_B^2}$ is ignored, as explained by Derrida [6]

(lower) crossing point of the parabola with the abscissa, $E_c = -N\sigma_B (2\ln\gamma)^{1/2}$. The finite width of this energy spectrum reflects the fact that the system is frustrated. The "native state" corresponds to the least frustrated structure. If the native state is non-degenerate, this lowest-energy conformation has zero entropy. The degree of frustration of a heteropolymer is linked to the number of different monomers that it contains. This is particularly obvious in the case of a homopolymer. For such molecules, all compact states are unfrustrated and have the same energy. This picture is confirmed in our simulation, where indeed we observe a non-degenerate native state for a well designed sequence. In the following we will refer to the lowest energy state as the native state of the heteropolymer.

In 1993 Shaknovich and Gutin [7, 8, 9] showed that it is possible to "design" a lattice protein in such a way that it will fold into a specific conformation. They achieved this by optimizing the sequence of amino acids, using a Monte Carlo algorithm that randomly exchanges amino acids within the chain molecule. The acceptance of such trial swaps depends on the energy change associated with the move

$$\Delta E = \sum \left(S'_{ij} - S_{ij}\right) C_{ij} \tag{2.4}$$

where S' (S) denotes the interaction matrix of the new (old) sequence of amino acids. During a Monte Carlo run of several million cycles, a large number of distinct sequences are generated. The sequence S^* with the lowest energy is assumed to be the best candidate to fold into the native state.

$$E_{\text{Native}} = \sum C_{ij} S^*_{ij} \,. \tag{2.5}$$

A closely related, but different method, for the design of heteropolymers that fold into a specific structure, is the so called "Painted globule" model (see, e.g. [13, 14]). The central idea behind this approach is to look at the target structure, and then distinguish between surface and core residues (hydrophilic and hydrophobic). The design consists of a sequence of folding and re-painting steps.

In our work we have chosen to not use this last method, because it does not allow for a complete control on the target configuration. The annealing process needs equilibration steps of the configuration of the chain. In practice this means that the minimization procedure is done on both the C and the S terms of the energy. To study designed configurational changes in protein we found it crucial to have the target structure equal to the final native state of the protein.

Pande et al. [10, 11, 12] have provided a theoretical analysis of the general design of a foldable protein sequence. In particular, these authors showed that, in the context of the Random Energy Model, the phase behavior of designed protein sequences can be predicted analytically. One of the main findings of Pande et al. was that the energy gap separating the target "native" state from the set of non-native compact states, is inversely proportional to the design temperature (the fictitious temperature at which we perform Boltzmann sampling of different sequences for a given target conformation). This a crucial result because it gives a theoretical basis to the feasibility of heteropolymer design, and also because it fixes a thermodynamic relation between the process of folding of a chain and the the design of its sequence of monomers. In the next section we will give the derivation of such a fundamental result.

2.2 Folding and Freezing Transition

In this section we derive the relation between the freezing transition for heteropolymers with random sequences and the folding transition of designed proteins. Although, the derivation it is valid only in a mean-field approximation, the final result will give a clear and simple physical explanation of what it means to design a protein.

In their 1997 review Pande et al. [11], presented a new approach to describe the statistical mechanics of protein folding. The approach that we follow in subsequent chapters has been inspired by the methods described in ref. [11].

Let us start by considering the total free energy of a random heteropolymer

$$\mathcal{F}(T) = \left\langle \mathcal{F}_{seq}(T) \right\rangle = -T \left\langle \ln \mathcal{Z}_{seq}(T) \right\rangle \tag{2.6}$$

where $\mathcal{F}_{seq}$ ($\mathcal{Z}_{seq}$) is the free energy (partition function) for a possible random sequence and the average is done on all possible realizations. As we discussed above, the entropy in REM (Eq. 2.3) vanishes for states with energy below E_c, while above it the average density of states $\langle n(E) \rangle$ is equal to $\mathcal{M}P(E)dE$ where $\mathcal{M} = \gamma^N$ is the number of states, dE in the thermodynamic limit can be approximated by $dE \simeq \sqrt{N}$ [2], and P is the Gaussian density probability

$$P(E) = \frac{1}{\left(2\pi N \sigma_B^2\right)^{1/2}} \exp\left[-\frac{E^2}{2N\sigma_B^2}\right] \tag{2.7}$$

where σ_B^2 is the variance of the interaction matrix . We can then rewrite the partition function in a simpler form by taking advantage of the exponential weight of states in the saddle point E^* of the density of state

$$\mathcal{Z}_{seq}(T) = \int n(E)\, e^{-E/T} dE = \int \mathcal{M}P(E) e^{-E/T} dE \simeq \mathcal{M}P(E^*) e^{-E^*/T}$$

where $E^* = -N\sigma_B^2/T$ as the value of the energy at which the argument inside the integral is maximum for a Gaussian distribution [3]. This representation is valid for temperatures for which E^* is located in the continuous part of the spectrum. In this regime, the partition function is independent of the particular sequence and in Eq. 2.6 we can swap the ordering of taking the average and the logarithm. The main assumption of the REM is that each interaction between the monomers is independent, and follows a Gaussian distribution. Using this approximation we can compute the average of $\mathcal{Z}$ over the all possible sequences, as the product of the contribution of each pair interaction:

$$\langle \mathcal{Z} \rangle = \mathcal{M} \left[\int P(E_{pair})\, e^{-E_{pair}/T} dE_{pair} \right]^{\mathcal{L}} = \mathcal{M} e^{\mathcal{L}\left[\overline{E} - \frac{\sigma_B^2}{2T}\right]} \tag{2.8}$$

$$\mathcal{F} = -T \ln \langle \mathcal{Z} \rangle = \mathcal{L}\left[\overline{E} - \frac{\sigma_B^2}{2T}\right] - TN\ln\gamma = \mathcal{L}\left[\overline{E} - \frac{\sigma_B^2}{2T}\left(1 + \frac{T^2}{T_g^2}\right)\right], \tag{2.9}$$

[2] The approximation for dE is an arbitrary power α of N, with $\alpha < 1$ Ref. [6]

[3] Note that by taking the thermodynamic limit of the log of the density of states $\ln \gamma^N P(E)\, dE \simeq \gamma^N P(E)\sqrt{N} \simeq N\left[\ln\gamma - \frac{E^2}{2N^2\sigma_B^2}\right]$ we recover the expression of the entropy in Eq. 2.3

where $\mathcal{L}$ is the (independent of the conformation) number of contacts between monomer in any particular compact conformation, and $\overline{E}$ is the average of the interaction matrix. We expressed the right-hand side as a function of the transition temperature T_g at which the entropy $S(T) = -\frac{d\mathcal{F}}{dT}$ vanishes

$$T_g = \sigma_B \sqrt{\frac{\mathcal{L}}{2N \ln \gamma}} \tag{2.10}$$

This temperature is called *glass temperature* because below it the system is trapped in one of the conformations that belong to the discrete region of the density of states. We now have the statistical tools to describe the phase behavior of a quenched random heteropolymer. The reason why we sketch the derivation of Eq. 2.9 explicitly is that it illustrates the peculiar temperature dependence of the system. Above the glass temperature T_g, the random-energy heteropolymer explores many states practically independent of the particular sequence of amino acids. However, as the temperature is lowered, the equilibrium is dominated by few discrete states of low energy, which are highly dependent on the particular sequence. The transition at $T = T_g$ is called the freezing transition [15, 16]. Initially it was suggested that the random-energy model might provide a useful model for protein folding, as it yields a unique ground state with a probability independent of the system size. However the energy differences between structurally distinct states in the discrete region of the energy spectrum are only of the order of $\sqrt{N}$, which does not allow for a robust equilibrium state. The question is then if it possible to design particular sequences that freeze into a robust ground state. In order for such an approach to work, the energy of the target state must be well separated from the boundaries of the continuous distribution of states, where the glassy states accumulate (at typical distances of order $\sqrt{N}$). Using mean-field arguments similar to the ones used above, we can derive an expression for the average energy of the designed state E_d as function of the temperature of the canonical ensemble of sequences T_d. We start by choosing a target conformation C_d as our tentative native state. This conformation is characterized by an energy $E_d = \mathcal{H}(S_d, C_d)$ that obviously depends on the sequence S_d. The partition function obtained by summing over all possible sequences is denoted by W and it defines a free energy F_W through

$$\begin{aligned} F_W \equiv -T_d \ln W(T_d) &= -T_d \ln \left[\left\langle \exp\left[-\mathcal{H}(S_d, C_d) \right] / T_d \right\rangle \right] \\ &\simeq \left\langle \mathcal{H} \right\rangle - 1/2T_d \left[\left\langle \mathcal{H}^2 \right\rangle - \left\langle \mathcal{H} \right\rangle^2 \right] \\ &= \mathcal{L} \left[\overline{E} - \frac{\sigma_B^2}{2T_d} \right]. \end{aligned}$$

where T_d denotes the design temperature. In terms of F_W we can write an approximate expression for the average energy of the designed sequence $\langle E_d \rangle = -\frac{\partial \ln W}{\partial (1/T)}\Big|_{T \to T_d}$, which does not depend on the target conformation, but instead show that the energy is inversely proportional to the design temperature

$$\langle E_d \rangle = \mathcal{L} \left[\overline{E} - \frac{\sigma_B^2}{T_d} \right], \tag{2.11}$$

This result implies that if the design procedure is carried out at a temperature lower than T_g, then the average energy of the designed state will be below the boundaries of the continuous

part of the density of states. The lower the design temperature, the larger the gap will be and, as a consequence, the designed state will be increasingly robust. Furthermore the freezing transition of sequences designed at low T_d will occur at a temperature higher than T_g because the energy of a designed sequence is lower that the one of the glassy states of a random heteropolymer. The folding temperature T_f is defined as the temperature at which there is equilibrium between the native (target) state and the random globule. T_f can be computed by comparing Eq. 2.11 with Eq. 2.9

$$\frac{1}{T_f^2} + \frac{1}{T_g^2} = \frac{2}{T_f T_d} \tag{2.12}$$

which is independent of the mean value of the interaction matrix, but it does depend on the variance σ_B (see Eq. 2.10). We expect that the rate at which a designed protein folds into the native state is faster than the rate at which a completely random heteropolymer reaches its lowest-energy state. The reason is that folding takes places at a temperature $T_f > T_g$ where the system has enough thermal energy to overcome local energy barriers. Using the relation in Eq. 2.12, we constructed a phase diagram that describes the general relation between design and folding in heteropolymers (see Fig. 2.3).

2.3 Design algorithm

In the previous section I have presented a mean field analysis of the design process of heteropolymers. The results proved the validity of a general design scheme, that would simply minimize the total energy of the polymer, with a fixed target compact configuration in a canonical ensemble of sequences. If this minimization process was carried out at a temperature T_d low enough, than the final heteropolymer would have shown a folding behavior similar to real proteins. I now will describe the technical details our implementation of the design algorithm. In order to design monomer sequences that yield a target conformation, we used a modified version of the Shakhnovich method. We perform point mutation of single amino acids and swap of identity between two randomly chosen residues. Unlike the latter method, our approach does not keep the amino acid composition of a chain fixed. Rather, we allow for random changes of amino acids. As a consequence we had to devise a criteria more sophisticated than the normal metropolis scheme to prevent that this compositional sampling results in the formation of homopolymers of the most attractive amino acid. We introduce a (purely fictitious) compositional "temperature". Increasing the compositional temperature increases the compositional entropy. To perform the sampling, we combine the following acceptance criterion with the normal acceptance Metropolis rule

$$P_{\text{acc}} = \min \left\{ 1, \left(\frac{N_P^{\text{new}}}{N_P^{\text{old}}} \right)^{T_p} \right\},$$

where T_p is the arbitrary parameter that plays the role of a temperature, and N_P is the number of permutations that are possible for a given set of amino acids. N_P is given by the multinomial

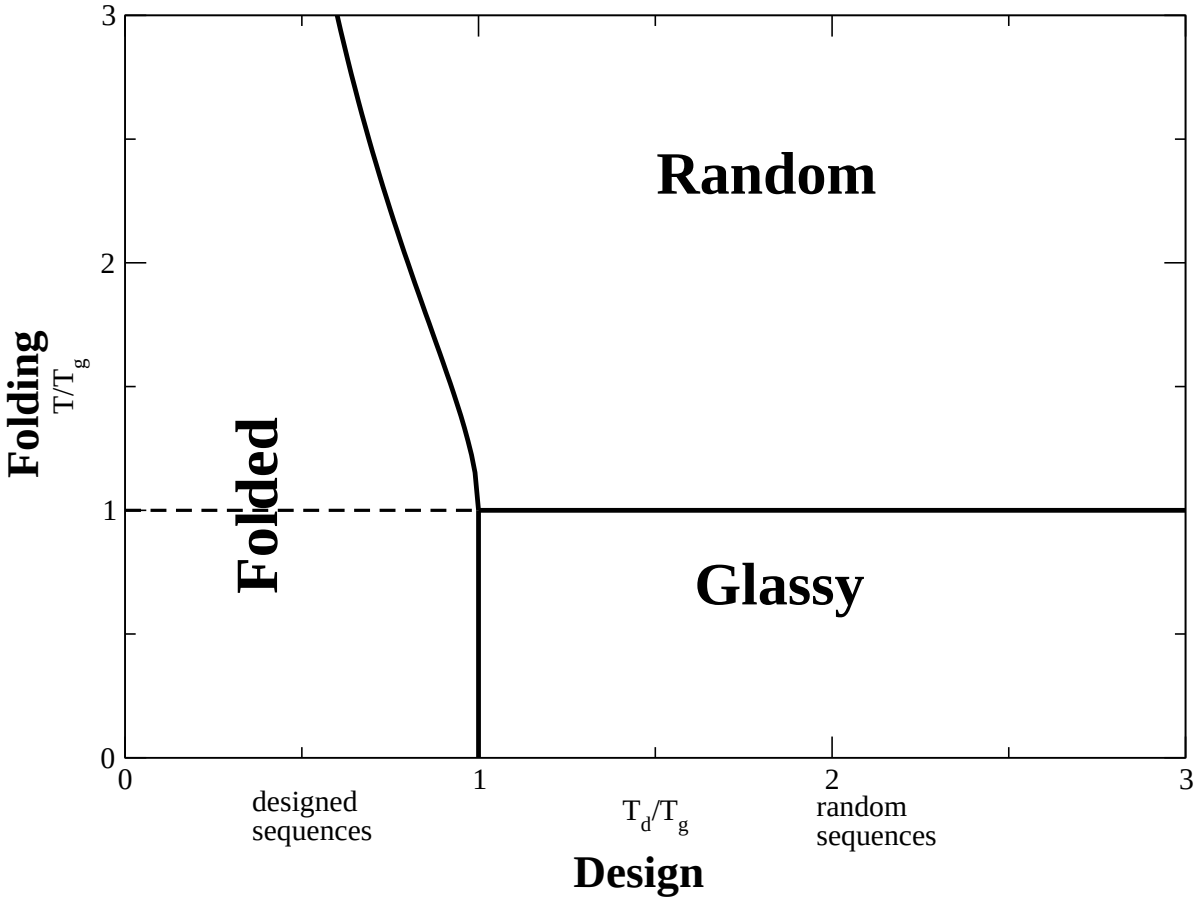


Figure 2.3: Phase diagram of the freezing transition in globular heteropolymers with a designed sequence. We have indicated with on the y axis we plot the temperature T of the system during the folding process, and on the x axis there is the temperature $T_d$of the design process. We can identify 3 phases: 1) a Glassy phase in the region $T_d/T_g > 1$ and $T/T_g < 1$, in which the protein is stuck in one of the low energy states in the discrete part of the energy spectrum of a random heteropolymer. 2) a Random phase for $T/T_g > 1$if $T_d/T_g > 1$and $T > T_f$, where $T_f$is the folding temperature and is calculated through Eq. 2.12 with the conditions $T_d/T_g < 1$and $T_f/T_g > 1$. This region is called Random because the system fluctuates between the different conformation of the continuous part of the density of states, and it correspond to the unfolded state of a design protein. 3) The last region is the Folded phase where the target conformation dominates the equilibrium. The thermodynamic stability of this phase is directly dependent on how much the design temperature T_d is lower than the glass one T_g

expression

$$N_p = \frac{N!}{n_1!n_2!n_3!\ldots} \tag{2.13}$$

where N is the total number of monomers and n_1, n_2 etc are the number of amino acids of type 1,2,.... With this condition we can explore a large set of sequences, yet avoid the formation of homopolymers. In the absence of any *a priori* criterion to fix T_p, we used trial and error. If T_p is too small, the chains will tend to become homopolymers (with a degenerate native state). In contrast, when T_p is too large, we noticed that the lowest energy sequences were no longer able to fold (the sequence becomes too random). We should therefore choose a value of T_p that yields a compromise between these two conflicting tendencies. To explore a range of values for T_p and at the same time limit the trapping in local minima of sequence space we introduced a parallel tempering algorithm for the sequence sampling at different pseudo-temperatures. We use the following criterion to decide whether the swapping of the sequences between two adjacent temperatures would be accepted:

$$P_{\text{acc}} = \min\left\{1, \left(\frac{N_P^{\text{new}}}{N_P^{\text{old}}}\right)^{\Delta T_p}\right\}. \tag{2.14}$$

In our simulations we found that good sequences could be designed if parallel tempering was performed with the set of values $T_p = \{1, 1/2, \ldots 1/14\}$. With this set of compositional temperatures we obtained native states that were both stable and non-degenerate. We stress that the fictitious temperature parameter T_p only plays a role in the generation of suitable sequence. It plays no role in the subsequent simulations of chain (re)folding. We used variations of the above design algorithm for each individual problem that we studied. In every chapter we describe the modifications introduced and we explain the reasons for introducing the approach used in that chapter. Of course, once we have generated a particular sequence (or two, as the case may be), we still need to test whether these sequences do indeed fold into the desired structures.

3 Free Energy Calculations

In the previous chapter we have introduced the model that we use to describe proteins. To study the folding of a particular model protein, we use a Monte Carlo algorithm with three basic moves: corner-flip, crankshaft, branch rotation. The corner-flip involves a rotation of 180 degrees of a given particle about the line joining its neighbors along the chain. The crankshaft move, is a rotation by 90 degrees of two consecutive particles. A branch rotation is a turn, around a randomly chosen pivot particle, of the whole section starting from the pivot particle and going to the end of the chain. With these moves we expect to have a good balance between cooperative moves and single-particle moves, as well as an efficient sampling of the compact configuration of the polymer, which are crucial for the study of the equilibrium properties of the native state.

Each move can be accepted or rejected according to the following acceptance rule, based on the configurational energy of the system,

$$P_{acc} = \frac{e^{-\beta(E_N - E_O)}}{1 + e^{-\beta(E_N - E_O)}},$$

where E_N and E_O represent the energy of the new state and of the old state respectively; $\beta = 1/T$ is the inverse temperature computed in reduced units. This acceptance rule satisfies detailed balance. If the system is ergodic (i.e. if every state can be reached from any other state in a finite number of Monte Carlo steps), the algorithm will ensure that every state is sampled with a frequency proportional to its Boltzmann weight. This choice of the acceptance rule is not unique; however we preferred it because it ensures that the total weight of the old and the new configuration is always one.

It is often convenient to group different protein conformations according to some "order parameter". For instance, one could classify states according to their radius of gyration or according to the number of nearest-neighbor contacts that they have in common with the lowest-energy ("native") state of the protein. A single value of such an order parameter may correspond to many different conformations. The probability to visit states with a given order parameter Q is determined by the free energy $F(O)$. In the case that all states with a given order parameter have the same energy, we can write $F(Q) = E(Q) - TS(Q)$, where $S(Q)$ is the entropy of the system with order parameter Q: $S(Q) = k \ln \Omega(Q)$, where $\Omega(Q)$ is the number of states with order parameter Q. In general, the free energy is simply defined by

$$F(Q) = -T \ln [P(Q)], \tag{3.1}$$

where $P(Q)$ denotes the probability that the system is in a state with order parameter Q at temperature T. Clearly, the most probable state of the system has the lowest free energy.

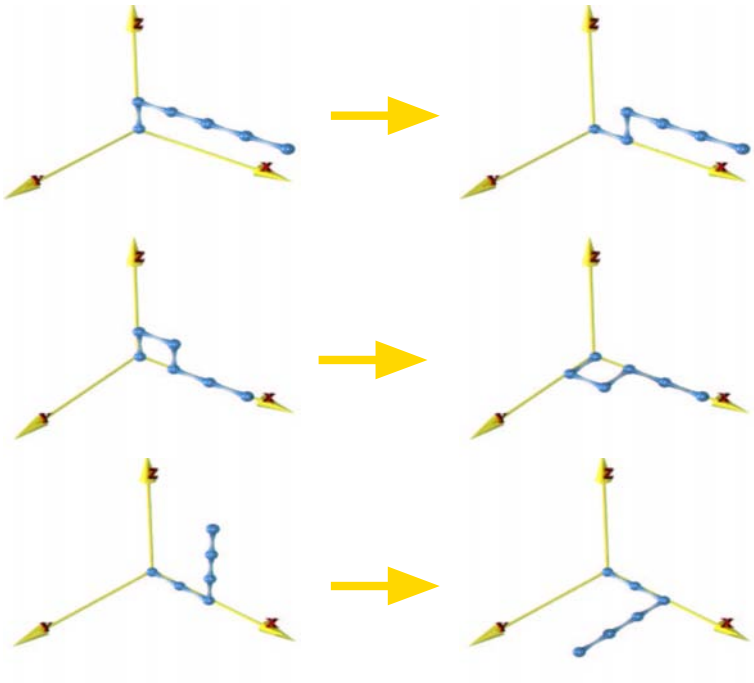


Figure 3.1: Monte Carlo moves used in the folding program. From top to bottom : corner flip, crankshaft, rotation

When studying the approach of the system to its lowest free-energy state, it is often useful to compute the free-energy "landscape" (i.e. the function $F(Q)$), as the topography of this landscape can suggest whether or not there are easy paths that bring the system from a non-native state to the native state.

During the simulation we measured F as function of several order parameters. One order parameter that we employed in all cases studied was the number of native contacts in a given conformation. This order parameter is a commonly used to measure the progress of the protein-folding process. In cases where we are considering a model with two target structures whose relative free energy can be changed by an external perturbation, we define an order parameter that measures the progress from one target state to the other: this order parameter is equal to the difference in the number of contacts that belong to two reference structures (e.g. 1 and 2) i.e.

$$Q(C) = \sum_{i<j}^{N} \left[C_{ij}^{(1)} C_{ij} - C_{ij}^{(2)} C_{ij} \right], \tag{3.2}$$

where $C_{ij}^{(1)}$ and $C_{ij}^{(2)}$ are the contact maps of the two target structures, whilst C_{ij} is the contact map of the instantaneous conformation To be more precise: as we consider two distinct target states (say 1 and 2), we give a value $+1$ to every contact that belongs to structure 1 and a value -1 to every native contact of structure 2. Contacts that appear in both 1 and 2 do not contribute to this order parameter. The reason why we assign negative values to native contacts of structure 2, is that we compute the free energy difference between the protein in

conformation 1 and 2. If we would have assigned 0 to the contacts of structure 2 then we would not have been able to distinguish it from unfolded states that do no have any native contacts at all.

In addition to the number of native contact we also introduced other order parameters to explore different properties of the system. For instance in chapter 4, we will investigate the design of specific interactions between a pair of proteins. For that calculations we found it useful to measure the number of contacts with a substrate regardless of whether they were native (i.e. involved in the selective binding) or not. The definition of the free energy does not substantially change and is easily extended to all the order parameters used in this thesis. In the methods subsection of each chapter, we give the problem-specific order parameters that are used.

Let us next consider the numerical determination of the free-energy landscape $F(Q)$. It is important to note that a direct (brute force) calculation of the histogram $P(Q)$ is not efficient because regions of high free energy are sampled infrequently, in particular at low temperatures. Different techniques are usually applied to increase the sampling in such situations. As the work described in this thesis progressed, we found it necessary to develop novel schemes to sample free-energy landscapes. Below, we briefly describe the methods used.

For the study of the free-energy landscape of relatively short chains, we made extensive use of the so called Adaptive Parallel Tempering (APT) scheme [17]. APT is an extension of the well-known Umbrella Sampling scheme. The latter method speeds up the sampling of a rugged free-energy landscape by flattening it. A simple way to flatten the landscape is to add a biasing potential to the normal Hamiltonian. Ideally, the biasing potential would be equal to $-F(Q)$, in which case the resulting free-energy landscape would be perfectly flat (the probability to observe the system in a state with order parameter Q is proportional to $\exp(-\beta(F(Q)+W(Q)))$. But, of course, we do not know $F(Q)$ a priori. This is where the "adaptive" part of APT comes in. APT uses an iterative method to estimate the biasing potential. During the simulation we sample the probability $P(Q)$ of sampling a conformation with order parameter Q (Eq. 3.2). After a specified number of steps we calculate the new biasing potential $W(Q)$ with the following recursive equation

$$W_i(Q,T) = W_{i-1}(Q,T) - \mathcal{U}\ln P(Q,T) \text{ with } W_0(Q,T) = 0\,, \tag{3.3}$$

where the index i indicates the iteration, and $\mathcal{U}$ is a constant which we give a value between 0.5 and 0.05 depending on how much the biasing potential varies from one iteration to the other. Once we have the new biasing potential we add it to the energy in the acceptance criterion of every move. The biasing potential $W(Q)$ depends on the conformation of the system through the order parameter Q, but it also depends on temperature. This temperature dependence is important when we combine umbrella sampling with multicanonical parallel tempering [17]. The basic idea behind parallel tempering is that we run two or more simulations in parallel. Each simulation runs at a different temperature (or, more in general, at a different value of an external control parameter). In addition to the normal Monte Carlo moves in the different systems, we occasionally attempt to swap the control parameter of two systems. In the present case, where the control parameter is temperature, the acceptance rule for a temperature swap

move is then

$$\begin{aligned} P_{\text{acc}} &= e^{\Delta\beta\Delta E+\Delta W} \\ \Delta W &= W(Q_i,T_j)-W(Q_j,T_j)+W(Q_j,T_i)-W(Q_i,T_i) \end{aligned} \tag{3.4}$$

where i and j are replica indices. The advantage of the APT algorithm is that it makes it possible to equilibrate systems with a rugged free-energy landscape. However, we found that the efficiency of the APT algorithm could be greatly increased by combining it with the so-called Waste Recycling Monte Carlo method [18, 19]. The resulting scheme, called Virtual-move Parallel Tempering (VMPT), turns out to be very efficient in sampling states over a wide range of free energies. This is an essential requirement for the study of conformational changes, because the states that are most stable in the absence of a perturbation will become less stable (high free energy) after the perturbation is applied. In the next section, we describe the VMPT scheme in more detail.

3.1 VMPT Virtual-move Parallel Tempering

The VMPT method boosts the efficiency of the accumulation of statistical averages by including information about all potential parallel-tempering trial moves, rather than just those trial moves that are accepted. As a test, we compute the free-energy landscape for conformational changes in simple model proteins. With the new technique, the sampled region of the conformational space in which the free-energy landscape could be reliably estimated, increases by a factor 20.

The exponential increase in the speed of computers during the past decades has made it possible to perform simulations that were utterly unfeasible one generation ago. But in many cases, the development of more efficient algorithms has been at least as important.

One of the most widely used schemes to simulate many-body systems is the Markov-chain Monte Carlo method (MCMC) that was introduced in 1953 by Metropolis et al. [20]. In this algorithm the average properties of a system are estimated by performing a random walk in the conformational space, where each state is sampled with a frequency proportional to its Boltzmann weight. In the Metropolis algorithm, this is achieved by attempting random moves from the current state of the system to a new state. Depending on the ratio of the Boltzmann weights of the new and the old states, these trial moves may be either accepted or rejected. Metropolis et al. showed that the acceptance probability of trial moves can be chosen such that Boltzmann sampling is achieved.

One important application of the MC method is the estimation of the Landau free energy F of the system as function of some order parameter

$$F(Q) = -T\left[\ln P(Q)\right].$$

There are many situations where the MCMC method does not yield an accurate estimate of F, because it fails to explore configuration space efficiently. This is, for instance, the case in “glassy” systems that tend to get trapped for long times in small pockets of configuration

space. In the early 1990's the so-called parallel-tempering (PT) technique was introduced to speed up the sampling in such systems [21, 22, 23, 24, 25, 26, 27].

In a parallel-tempering Monte Carlo simulation, n simulations of a particular model system are carried out in parallel at different temperatures (or other at different values of some other thermodynamic field, such as the chemical potential or a biasing potential). Each of these copies of the system is called replica. In addition to the regular MC trial moves, one occasionally attempts to swap the temperatures of a pair of these systems (say i and j). The swapping move between temperature i and j, is accepted or rejected according to a criterion that guarantees detailed balance, e.g.:

$$P_{acc}(ij) = \frac{e^{\Delta\beta_{ij}\Delta E_{ij}}}{1+e^{\Delta\beta_{ij}\Delta E_{ij}}} \tag{3.5}$$

where $\Delta\beta_{ij}$ is the difference of the inverse of swapping temperatures, and ΔE_{ij} is the energy difference of the two configurations. Although there are other valid acceptance rules, we used the one in Eq. 3.5 because it was easy to implement, and it ensures that the total weight of the old and the new configuration is always one.

To facilitate the sampling of high free-energy states,"difficult" regions, we use the Adaptive Umbrella Sampling [28, 29, 30, 31, 32]. In this (iterative) scheme, a biasing potential is constructed using the histogram of the states, sampled during an iteration as follows

$$W_I(Q,T) = W_{I-1}(Q,T) - a\ln(P_I(Q)), \tag{3.6}$$

where W is the biasing potential function of an order parameter Q, I is the iteration number, a is a constant that controls the rate of convergence of W (a typical value for a is 0.05), and T is the temperature. After iteration, W converges to the Landau free energy. As a consequence, $P(Q) \sim \exp(-\beta F(q))\exp(W(Q))$ becomes essentially flat and the biased sampling explores a larger fraction of the configuration space. During the MC sampling, we include the bias, and only at the end of the simulation we compute the free energy $F(Q)$ from

$$F(Q) = -T[\ln P(Q) + W(Q,T)],$$

where $P(Q)$ is the probability of observing a state characterized by the order parameter Q, and $W(Q,T)$ is the biasing potential of the last iteration computed at temperature T. Combined with Parallel Tempering, the acceptance rule for the temperature swapping move is then

$$acc_{ij} = \frac{e^{\Delta\beta_{ij}\Delta E_{ij}+\Delta W_{ij}}}{1+e^{\Delta\beta_{ij}\Delta E_{ij}+\Delta W_{ij}}} \tag{3.7}$$

$$\Delta W_{ij} = W_I(Q_i,T_j) - W_I(Q_j,T_j) + W_I(Q_j,T_i) - W_I(Q_i,T_i) \tag{3.8}$$

where i and j are replica indices, and I is the iteration number. We refer to this scheme as APT (Adaptive Parallel Tempering [33, 17].

In the conventional MCMC method all information about rejected trial moves is discarded. Recently Daan Frenkel has proposed a scheme that makes it possible to include the contributions of rejected configurations in the sampling of averages [18]. In the present section, we show how this approach can be used to increase the power of the parallel-tempering scheme.

In this scheme, we only retain information about PT moves that have been accepted. However, in the spirit of refs. [18], we can include the contribution of all PT trial moves, irrespective of whether they are accepted. The weight of the contribution of such a virtual move is directly related to its acceptance probability. For instance, if we use the symmetric acceptance rule for MC trial moves, then the weights of the original and new (trial) state in the sampling of virtual moves are given by

$$\begin{aligned} P_N &= \frac{e^{\Delta\beta\Delta E_{O\to N}+\Delta W_{O\to N}}}{1+e^{\Delta\beta\Delta E_{O\to N}+\Delta W_{O\to N}}} \\ P_O &= \frac{1}{1+e^{\Delta\beta\Delta E_{O\to N}+\Delta W_{O\to N}}}, \end{aligned}$$

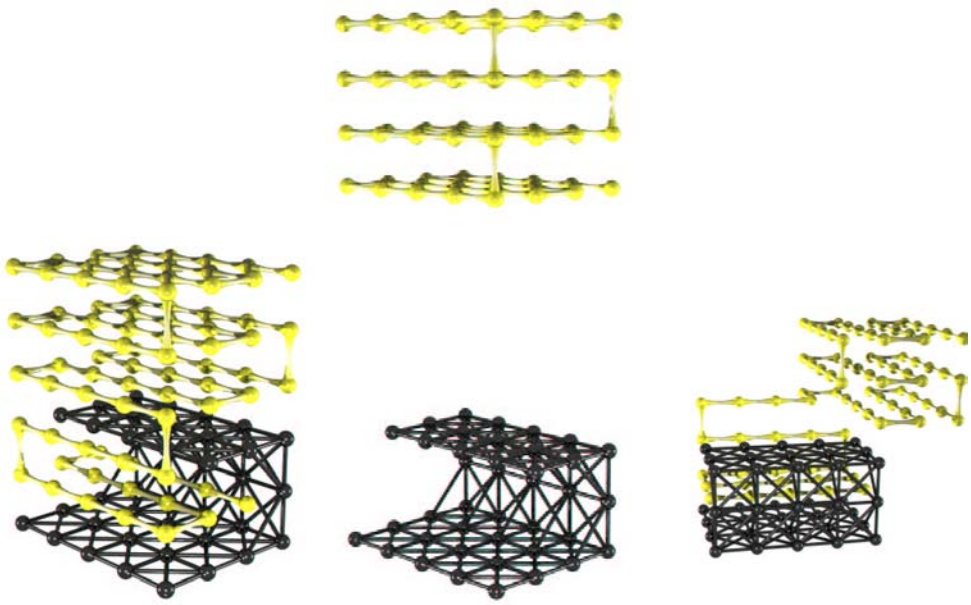
where $\Delta W_{O\to N}$ is defined in Eq. 3.8. We are not limited to a single trial swap of state i with a given state j. Rather, we can include all possible trial swaps between the temperature state i and all $N-1$ remaining temperatures. Our estimate for the contribution to the probability distribution P_i corresponding to temperature i is then given by the following sum

$$\begin{aligned} P_i(Q) = &\sum_{j=1}^{N-1}\left(\frac{1}{1+e^{\Delta\beta_{ij}\Delta E_{ij}+\Delta W_{ij}}}\right)\delta(Q_i-Q)+ \\ &\sum_{j=1}^{N-1}\left(\frac{e^{\Delta\beta_{ij}\Delta E_{ij}+\Delta W_{ij}}}{1+e^{\Delta\beta_{ij}\Delta E_{ij}+\Delta W_{ij}}}\right)\delta\left(Q_j-Q\right), \end{aligned}$$

where the delta functions select the configurations with order parameter Q. As we now combine the Parallel tempering algorithm with a set of parallel virtual moves, we refer to the present scheme as Virtual-move Parallel Tempering (VMPT).

In what follows we will extensively use this scheme to compute the free energy landscapes of different systems. However to show the efficiency of VMPT, we want to anticipate the calculation of the free energy landscape of a simple lattice-protein model. In this model, interaction with a substrate can induce a conformational change in the proteins.

Specifically, the model protein that we consider represents a heteropolymer containing 80 amino acids, while the substrate has a fixed space arrangement and contains 40 residues, see Fig. 3.2.



(a) (b) (c)

Figure 3.2: Spatial arrangement of the chain in the structures used to test the model (a , b) , and intermediate structure ($Q = 25$).

The configurational energy of the system is defined as

$$E_C = E_{\text{intra}} + E_{\text{inter}} = \sum_{i}^{N_C} \left[\sum_{j \neq i}^{N_C} C_{ij} S_{ij} + \sum_{j' \neq i}^{N_S} C_{ij'} S_{ij'} \right], \tag{3.9}$$

where the indices i and j run over the residues of the protein, while j' runs only over the elements of the substrate, C is the contact defined in Eq. 2.2 and S_{ij} is the interaction matrix. For S we use the 20 by 20 matrix fitted by Miyazawa and Jernigan [3] on the basis of the frequency of contacts between each pair of amino acids in nature.

We change the identity of the amino acids along the chain by "point mutations" which, in this context, means: changes of a single amino acid. In doing so we explore the sequence space of the protein and the substrate, and we minimize at the same time the configurational energy of the system in two distinct conformations, one bound (Fig. 3.2.a) and one unbound (Fig. 3.2.b). The design scheme is an extension of the basic one explained in section 2.2. In this scheme, trial mutations are accepted if the Monte Carlo acceptance criterion is satisfied for both conformations. Further details are given in Chapter 5 where will give an accurate description of the model system and the design techniques applied. For the purpose of this chapter let us assume that the design algorithm can generate a model protein that has the ability to change its conformation when bound to the substrate.

The sampling of the conformations is performed with three basic moves: corner-flip, crankshaft, branch rotation. The corner-flip involves a rotation of 180 degrees of a given

particle around the line joining its neighbors along the chain. The crankshaft move is a rotation by 90 degrees of two consecutive particles. A branch rotation is a turn, around a randomly chosen pivot particle, of the whole section starting from the pivot particle and going to the end of the chain. For all these moves we use a symmetric acceptance rule with the addition of the biasing potential calculated with the umbrella sampling scheme (Eq. 3.6)

$$acc_{O\to N} = \frac{e^{\beta\Delta E_{O\to N}+\Delta W_{O\to N}}}{1+e^{\beta\Delta E_{O\to N}+\Delta W_{O\to N}}}, \tag{3.10}$$

where $\Delta E_{O\to N}$ is the energy difference between the new and the old state (Eq. 3.9), and $\Delta W_{O\to N}$ is the difference in the bias potential from the same states (Eq. 3.6). We sample the free energy, as a function of two order parameters, of which the first is the conformational energy defined in Eq. 3.9, and the second is the difference of the number of contacts belonging to two reference structures (e.g. 1 and 2) $Q(C)$ from Eq. 3.2. For our specific case, $C_{ij}^{(1)}$ represents the structure in Fig. 3.2.a, while $C_{ij}^{(2)}$ corresponds to the one shown in Fig. 3.2.b, and Q has values between -15 and 30. Because the number of native contacts includes the contacts with the substrate of the reference state, it can be used to compute the free energy difference between the unbound state and the specifically bound one.

We performed 15 simulations, 5 of them with VMPT (using the parameters in Tab. 3.1.I) and the other 10 with APT (5 using the parameters in Tab. 3.1.I, and 5 with the parameters in Tab. 3.1.II).

Simulation	Temperatures	Number of Iterations	Sampling Steps	APT exec time (sec)	VMPT exec Time (sec)
I	0.1 0.125 0.143 0.167 0.2 0.222 0.23 0.25 0.270000 0.29 0.31 0.33 0.350000 0.37 0.4 0.444 0.5	400	$4\,10^8$	2600	3200
II	0.1 0.125 0.143 0.167 0.2 0.222 0.23 0.25 0.270000 0.29 0.31 0.33 0.350000 0.37 0.4 0.444 0.5	1000	$2\,10^{10}$	150000	

Table 3.1: Simulation parameters used for comparing the VMPT algorithm with the old scheme. In Simul. I we used the same parameters for both algorithms. The results in Fig. 3.3 show that VMPT was much more efficient in sampling the free energy. In Simul.II, we increased by two orders of magnitude the number of steps of the simulation with APT to obtain a sampling of $F(Q)$comparable to the one computed using the new VMPT scheme (Fig. 3.4). Execution times computed on a SGI Altix 3700 with Intel Itanium II, 1,3 GHz

In Fig. 3.3 we compare the average free energies at $T = 0.1$ (with error bars). In this figure, we only show those free energies that were sampled in all the 5 simulations of each group.

>From the figure it is clear that the VMPT approach leads to a much better sampling of the free-energy landscape. The advantage of the VMPT approach becomes even more obvious if we plot the free energy “landscape” as function of two order parameters (viz. the conformational energy (Eq. 3.9) and the number of native contacts). In this case the APT method is almost useless as only small fragments of the free-energy landscape can be reconstructed. The total number of points sampled with VMPT is 20 times larger than with APT, and the energy range that is probed, is one order of magnitude larger (see Fig. 3.5). To check the accuracy of the VMPT method, we compared the average free energy obtained by APT and VMPT at high temperatures where the APT scheme works reasonably well. As can be seen in Fig. 3.4 the two methods agree well in this regime (be it that a much longer APT simulation was needed). Even though the APT runs required 20 times more MC cycles, it still probes about 30% less of the free-energy landscape than the VMPT scheme.

As the implementation described above is not based on a particular feature of the system under study, the results obtained in this study suggest that the VMPT method may be useful for the study of any system that is normally simulated using Parallel Tempering. Examples of the application of Parallel Tempering in fully atomistic simulations of protein folding can be found in refs. [34, 35].

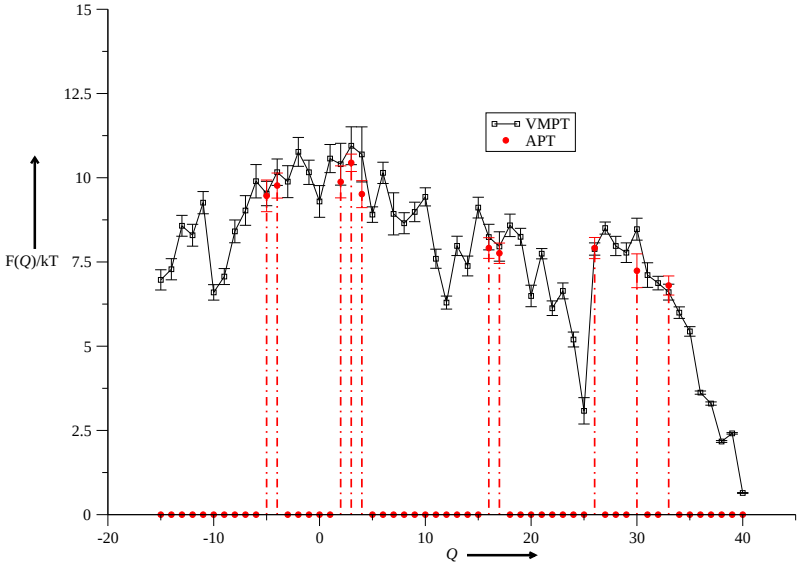


Figure 3.3: Average free energy computed with 5 run (10^8 MC steps Tab. 3.1.I) of the old scheme, compared with the result of 5 VMPT simulation (10^8 MC steps Tab. 3.1.I), at $T = 0.1$. The points with $F = 0$ correspond to values of Q that have not been sampled.

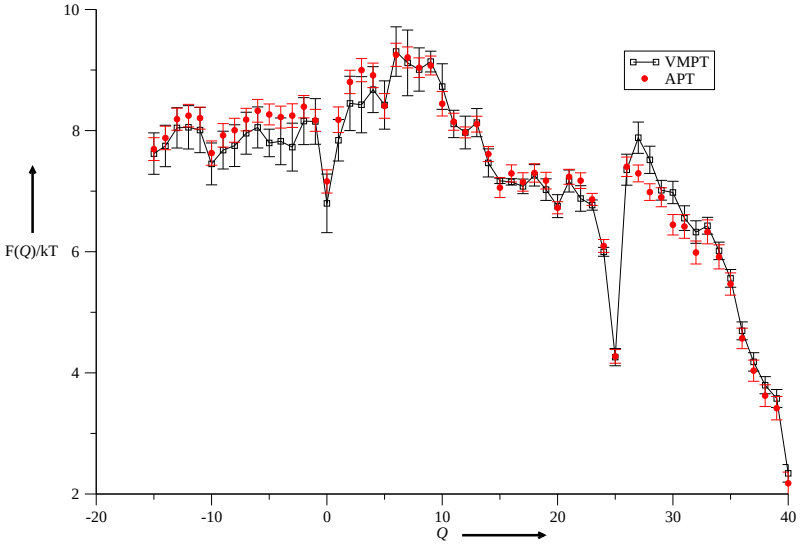
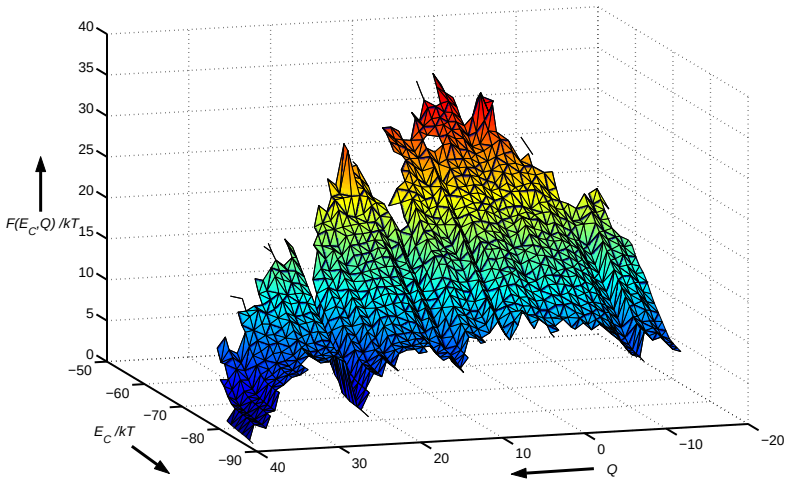
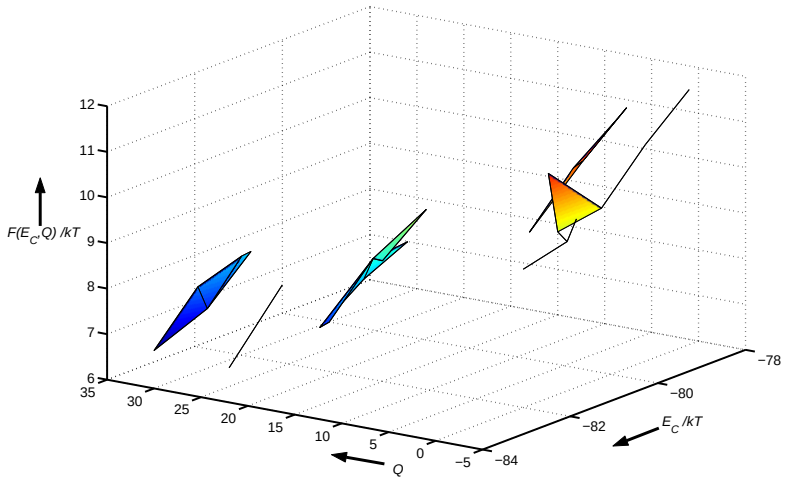


Figure 3.4: Average free energy computed with 5 long run (10^10 MC steps Tab. 3.1.II) of the old scheme, compared with the result of 5 shorter VMPT simulation (10^8 MC steps Tab. 3.1.I), at $T = 0.5$



(a)



(b)

Figure 3.5: Plot of the free energy landscapes computed with the algorithm VMPT (a) and the standard scheme APT (b). The free energies $F(E_C, Q)$ are function of the conformational energy E_C (Eq. 3.9) and of the number of native contacts Q (Eq. 3.2). It is important to notice the big difference in the sampling, in fact the number of points sampled with VMPT is 30 times bigger than the one obtained with APT.

4 Design Refoldable Molecules

We report a numerical study of the design of lattice heteropolymers that can refold when the properties of only a few monomers are changed. If we assume that the effect of an external agent on a heteropolymer is to alter the interactions between its constituent monomers, our simulations provide a description of a simple allosteric transition. We characterize the free energy surfaces of the initial and modified chain molecule. We find that there is a region of conformation space where molecules can be made to refold with minimal free energy cost. This region is accessible by thermal fluctuations. The efficiency of a motor based on such an allosteric transition would be enhanced by "borrowing" heat from the environment in the initial stages of the refolding, and "paying back" later. In fact, the power cycle of many real molecular motors does involve such a borrowing activation step.

4.1 Introduction

Molecular motors are molecules that can convert chemical energy into mechanical energy. The effect of the chemical reaction is to induce a conformational change in the molecule. As the final conformation has a lower free energy than the initial one, the molecule has the capacity to perform an amount of work that is, at most, equal to this free energy change. The amount of work that is delivered in practice depends on many factors, such as the speed of the transformation and the mechanical coupling of the relevant "reaction coordinate" to the outside world.

In molecular motors, as well as many others proteins, the transition between conformations is induced by a change in the environment (e.g. a pH change), the absorption of a photon, or the chemical transformation of a fuel molecule (e.g. ATP or lactose) attached to the protein. The effect of this external element is to change the structure or interactions of some "active" parts of the protein. These changes, in turn, lead to a rearrangement of the protein structure. The fact that molecular motors are microscopic, has important consequences for their mode of operation. In fact, the second law of thermodynamics makes it impossible for a macroscopic Carnot engine to "borrow" significant amounts of heat from the environment. In contrast, thermal fluctuations play an important role in the behavior of molecular motors.

Changes in conformation due to altered interactions between monomers are also of interest in a different context, namely in the design of mutations that significantly modify the native structure of a protein. In 1993, Rose and Creamer [36] formulated this problem as follows: given two distinct protein folds of similar length, what is the minimum number of amino acids that must be changed in order to transform one fold into the other? In fact, Rose and

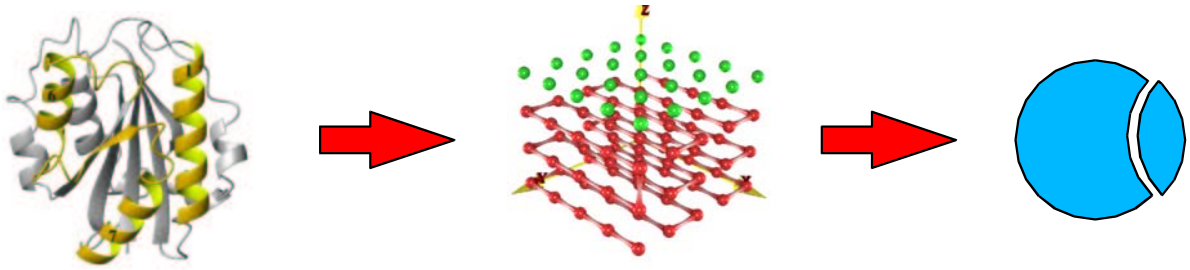


Figure 4.1: The figure represent the level of approximation that we intend to use for the description of the protein world. The lattice model can be viewed as a representation of proteins that is intermediate between a fully atomistic model (that would be computationally intractable except in very simple cases) and a representation where all internal degrees of freedom of the protein are ignored.

Craemer formulated the so-called “Paracelsus challenge”: the award of a prize to anyone who could convert one protein fold into another without changing more than 50% of the original protein’s sequence. A possible solution to this challenge was proposed by Dala et al. [37] who designed a protein sequence that could be converted from its native β -sheet conformation into an α -helix structure by changing fewer than half of the amino acids.

At the level of the relative stability of native structures, the present study of allosteric transitions is equivalent to the problem of conformational changes due to mutations. The difference appears when we consider the actual pathway by which the molecule refolds after the change has been introduced: this pathway has little physical meaning in the case of mutation, but is of considerable interest regarding allosteric transitions. In this chapter, we explore a simple model for allosteric transitions that is intermediate between a realistic, but prohibitively costly, atomistic model and a simple, but abstract, two-state model.

We model the chain as a linear, polypeptide-like heteropolymer, living on a lattice. In what follows, we shall refer to this molecule as a “protein” and, in fact, we shall use model parameters that apply to proteins. We stress, however, that the approach is not limited to protein-based conformation switches. Our central assumption concerns the effect of a chemical reaction on the chain molecule. We assume that the chemical reaction does not directly lead to a conformational change of the molecule. Rather, we assume that it leads to a change in the effective interaction between some of the monomers in the chain.

In our model, we account for this difference in the properties of individual monomers by changing their chemical nature. This modification could be thought of, for example, as a change of the ionization state of acidic or basic residues triggered by a pH change, or as resulting from binding to a metal ion. For simplicity, however, rather then introducing a new set of interactions between specific residues, we exchange them for other members of the standard set of twenty amino acids. In the language of proteins: we change the identity of one or more amino acids in the chain. Once an amino acid has been modified, the molecule may be able to lower its free energy by transforming into a different native state. In the present model, it is this thermodynamic incentive that drives the refolding to a new spatial arrangement.

WKCAVCEMNRCILCDTWKCFICEMERDGQKYPSRQK	Sequence A
WKCAVCEMNRCILCDTWKCFICEMERDGQKYPS IQM	Sequence B
WKCAVCEMNRCILCDDWKCFGCEMPRKNPMYTS EQH	Sequence C
WKCAVCEMNRCILCDDWKCFGCEMPRKNEHYT SIQP	Sequence D
HWKLHDMYVWRTKDMLPWREVDMYAQIPPITE NSKAFESCRGFQCLNG	Sequence E
HWKLHDMYVWRTKDMLPWREVDMYAQIPPIT EN-SKAFESCRGFQCNKG	Sequence F

Table 4.1: Sequences generated for the test structures (Fig. 4.2). The letters in bold are the amino acids chosen by the design program to induce the conformational change.

Below, we first describe the techniques used to simulate our system, we then present the simulations of the refolding process, and finally we discuss some of the implications of this work.

4.2 Results

To illustrate the mechanism by which allosteric transitions proceed in our model, we consider the refolding behavior of three different model molecules. In Fig. 4.2 we show the target structures between which the transitions occur: $1 \Leftrightarrow 2$, $3 \Leftrightarrow 4$, and $5 \Leftrightarrow 6$. Because the same procedure is applied in every case, we focus our explanation on the conformational change from structure 1 (Fig. 4.2a) to structure 2 (Fig. 4.2b).

4.2.1 Design of a switchable polymer

Following the procedure explained in section 2.2 we first designed a sequence that would fold into structure 1 (see Fig. 4.2a). We explore possible amino acid sequences by using both the conventional swapping move that does not change the composition and the switch move, that does. The acceptance criterion of the latter trial move depends on the parameter T_p that has to be chosen. A typical result after 10 million iterations is the sequence A (Table 4.1).

We applied the same technique to the other initial structures 3 and 5 (Fig. 4.2c,e), and the resulting sequences are respectively sequence C and sequence E (Table 4.1).

Sequences A, C and E, listed in Table 4.1, were used as the starting point to design the modified sequences that would refold into structures 2, 4 and 6, respectively (see Fig. 4.2b,d,f). We limited our search to those sequences that differed by a given number of residues. For the first and the last example we constrained the sequence that formed the initial conformation to differ by, at most, two amino acids from the sequence that formed the final conformation. For the transition $3 \Leftrightarrow 4$, we imposed a threshold of 4 residue differences. These “Paracelsus

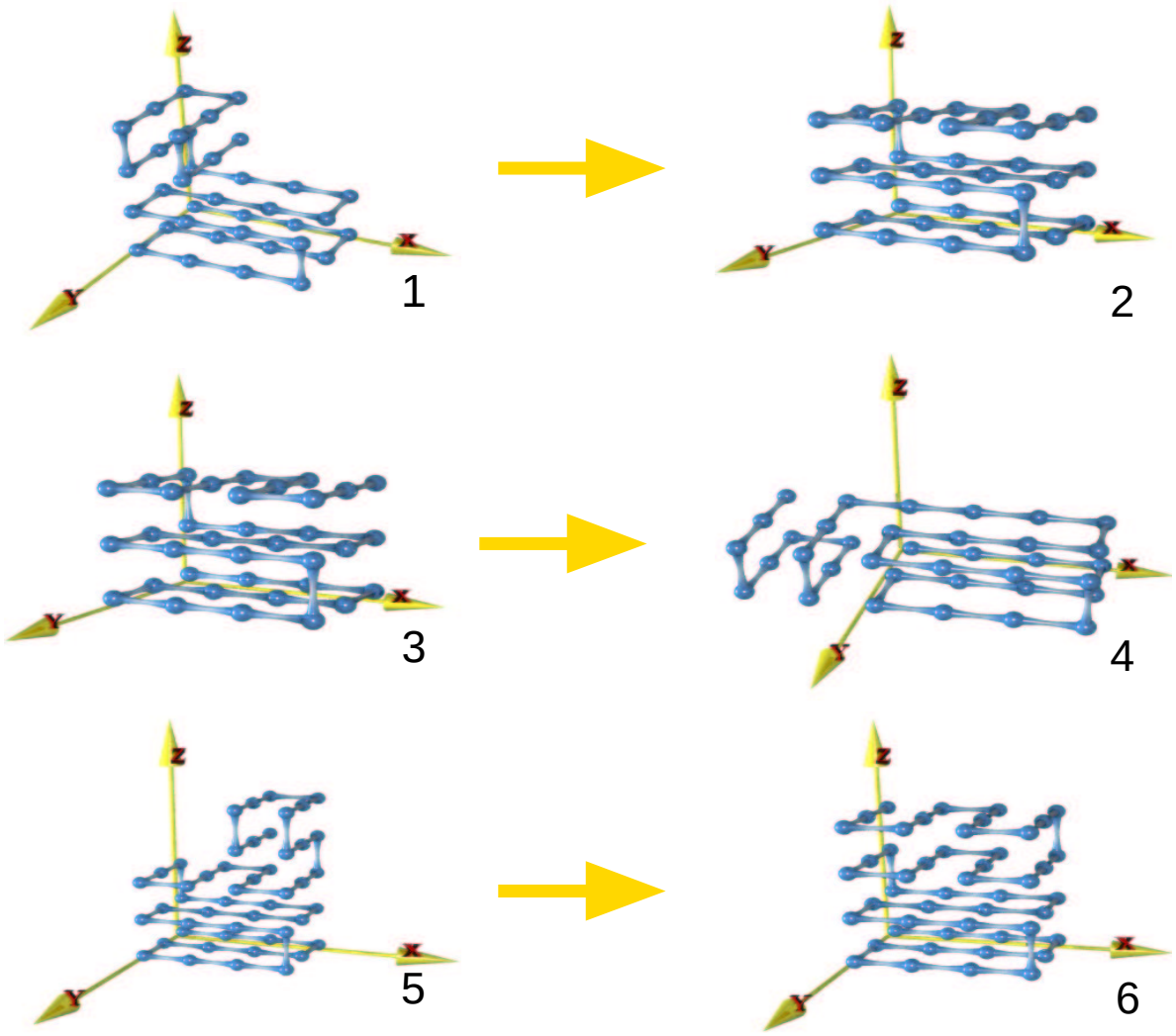


Figure 4.2: Spatial arrangement of the chain in the structures used to test the model

numbers" are purely empirical: they are the lowest threshold for which refolding to the desired structures could be obtained in each case.

The sequences that are listed in Table 4.1 are the ones used in the simulations described below. We stress that we did not impose the positions of the "mutations". But, not surprisingly, they appear to be concentrated in that part of the chain that is involved in the conformational change. Having constructed the two desired sequences, we performed a Monte Carlo (MC) simulation to study the equilibrium properties of the native state of each sequence.

4.2.2 Free energy Calculations

First, we checked if changing from sequence A to B (see Table 4.1) did, indeed, induce the desired conformational change. To this end, we started with a random coil of a molecule with sequence B. We used a standard MC simulation to let this structure fold. After the chain had reached its native structure (1), we changed the sequence from A to B and continued the simulation. After sequence B reached its native state (2), we switched back to sequence *A*, to verify that the refolding works both ways. In Fig. 4.3 we plot the conformational energy of the chain (Eq. 3.9) as a function of the number of MC steps, highlighting the time windows corresponding to each sequence. In each window, we see that immediately after the sequence switch, the system is in a state of very high potential energy, but then the chain quickly relaxes into its new native state. This shows that it is indeed possible to induce a conformational change with a relatively small modification of the chemical nature of some amino acids along the chain. The same procedure was also applied to the other sets of sequences. The results thus obtained were qualitatively similar to the one we obtained for the A-B pair.

Subsequently we studied the free energy profile using the parallel tempering described in section 2. The results of these simulations are shown in Fig. 4.4 and exhibit the characteristic behavior expected for a molecule that can undergo a folding transition. At low temperatures, the native state has a free energy that is lower than that of the molten globule (characterized by an order parameter close to zero). However, to study how one structure refolds into the other, we need to know the free energy landscape of, for example, sequence A in the vicinity of the native structure of B, and vice versa.

In order to improve the sampling of conformations that would hardly be sampled in a brute-force simulation, we proceed according to the method explained in section 3 and we use the APT method to construct an efficient biasing potential. The result of the simulation is the complete spectrum of the free energy for each of the two sequences. In Fig. 4.6 we plot the free energy of sequences A and B. As a consistency test, we compare in Fig. 4.5 the easily accessible part of the free energy of sequence B calculated with and without the umbrella-sampling scheme.

Now that we know the free energy curves for sequences A and B, we can study the effect of a sequence change. The crossing point of the curves (Fig,4.6) corresponds to a value of the order parameter for which we can change the sequence from A to B without changing the free energy.

While the order parameter that we have used thus far allows us to discriminate between states that are close to one native state or the other, it is less convenient to probe the intermediate region. The reason is that there are many different conformations with an order parameter

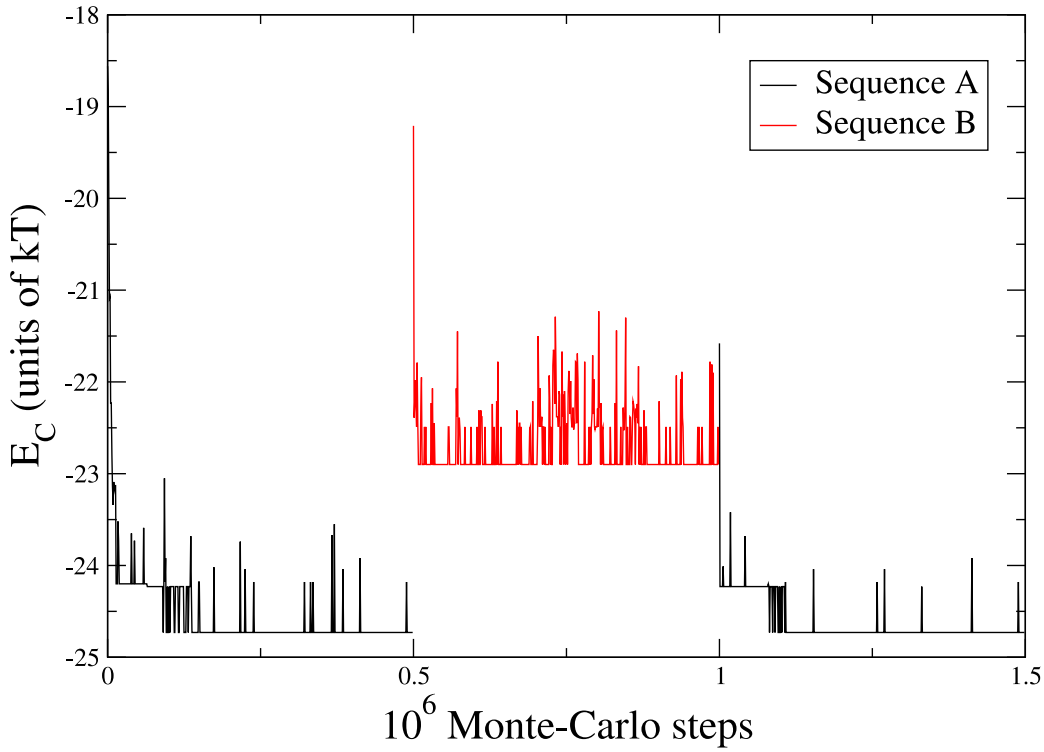


Figure 4.3: Sequence switching simulation. On the vertical axis we plot the conformational energy E_C (Eq. 3.9) while the abscissa is the number of Monte Carlo steps. During the simulation we switch from sequence A to sequence B, and look at the conformation corresponding to the lower energy. For the sequence A the native structure is 1 (Fig. 4.2.a), while for sequence B the structure is 2 (Fig. 4.2.b). The process is reversible; in fact, when, after 10 million steps we switch back to sequence A, the lowest energy conformation is structure 1.

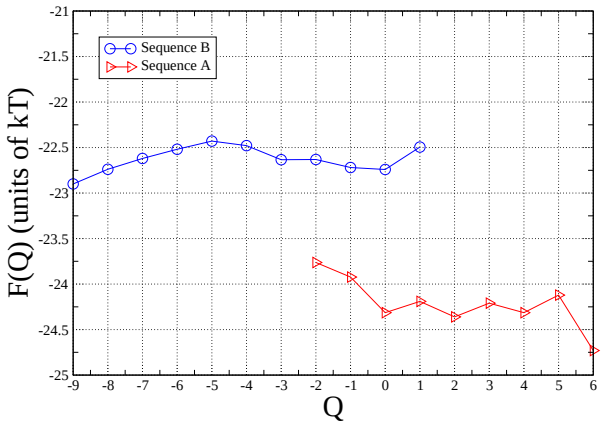


Figure 4.4: Result of the folding simulation for the sequences A and B, obtained using a parallel tempering algorithm. On the horizontal axis is the number of native contacts Q defined in Eq. 3.2. On the vertical axis is the free energy $F(Q)$. This plot shows the need for improved sampling.

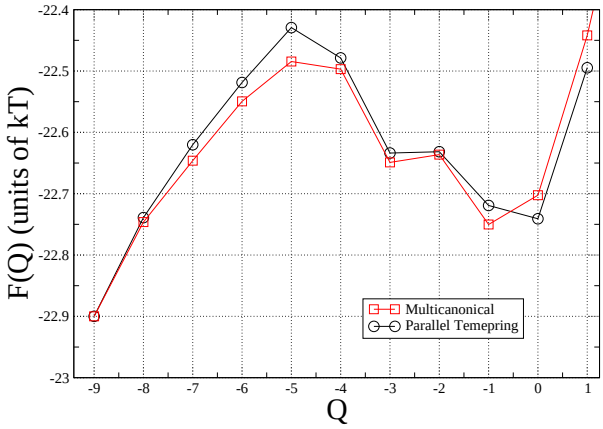


Figure 4.5: Comparison between the free energy $F(Q)$ calculated with an umbrella sampling simulation and a parallel tempering one, in a window of the order parameter Q (Eq. 3.2) were the two methods should give the same results. The agreement between the two methods provides confidence in the validity of our results.

close to zero. Not all of these conformations are equally important for the refolding process. We therefore need a second order parameter that allows us to get more detailed information about the free energy landscape in the intermediate state. We found that the conformational energy of the chain was suitable as a second order parameter.

Figure 4.7 shows the free energy landscapes for the folding of sequences A and B. Interestingly, the two surfaces show an overlap close to the crossing point of the curves in Fig. 4.6. This means that in the region of overlap, it is possible to change the amino acid sequence without changing the conformational energy of the chain. This suggests that in those conformations of the chain, the "mutated" amino acids are not in contact with the rest of the chain. The possibility of changing the sequence without affecting the potential energy of the chain facilitates the action of an external agent.

Similar behavior has been postulated for real protein motors that undergo a progressive change in the conformation. For instance, in the hand-over-hand model of kinesin by Schief and Howard [38], the external agent only acts if the protein is ready to accept it. Such behavior could easily be described by an extension of the present model where we only allow sequence switching when thermal fluctuations bring the chain into a favorable conformation. As can be seen from figure 4.4, both proteins can reach conformations with an order parameter between -2 and 1 by spontaneous thermal fluctuations of the order kT.

Clearly, the present model only deals with a single, albeit essential, aspect of a molecular motor, namely the property of a motor head to undergo an allosteric transition. For one thing, we do not consider explicitly the reaction between the external agent (e.g. ATP) and the chain molecule: we simply assume that the effect is to expose some other amino acids in the molecule. It should be possible to construct a model where these changes follow in a natural way from the chemical natures of both the chain molecule and the external agent. Then it would be interesting to see how the work that can be performed by the molecule (i.e. the difference in free energy between the initial and final states in the refolding process) depends on the free energy change associated with the chemical reaction with the external agent.

4.3 Discussion

Recently, Borovinskiy and Grosberg [39] reported a numerical study that focused on another aspect of refolding in lattice model proteins. In the latter work, the strategy was to design not just the initial and final states of the model protein, but also to design the sequence such that every single step of the conformational change would release approximately equal amounts of free energy. This imposed property was based on the idea that a protein stores free energy like a spring. Our model differs from this approach because we do not constrain the path by which the conformational changes proceed. In particular, by not imposing how the free energy is stored in the molecule, we find a barrier between the two states, the height of which depends strongly on the starting conformation of the chain. The refolding process is assisted by "uphill" thermal fluctuations that put the system in a favorable initial condition and effectively reduce the amount of chemical work that would be wasted in the refolding process. Evidence for the relevance of thermal fluctuations in initiating refolding comes from experimental studies on motor proteins [38] and is captured at a phenomenological level by

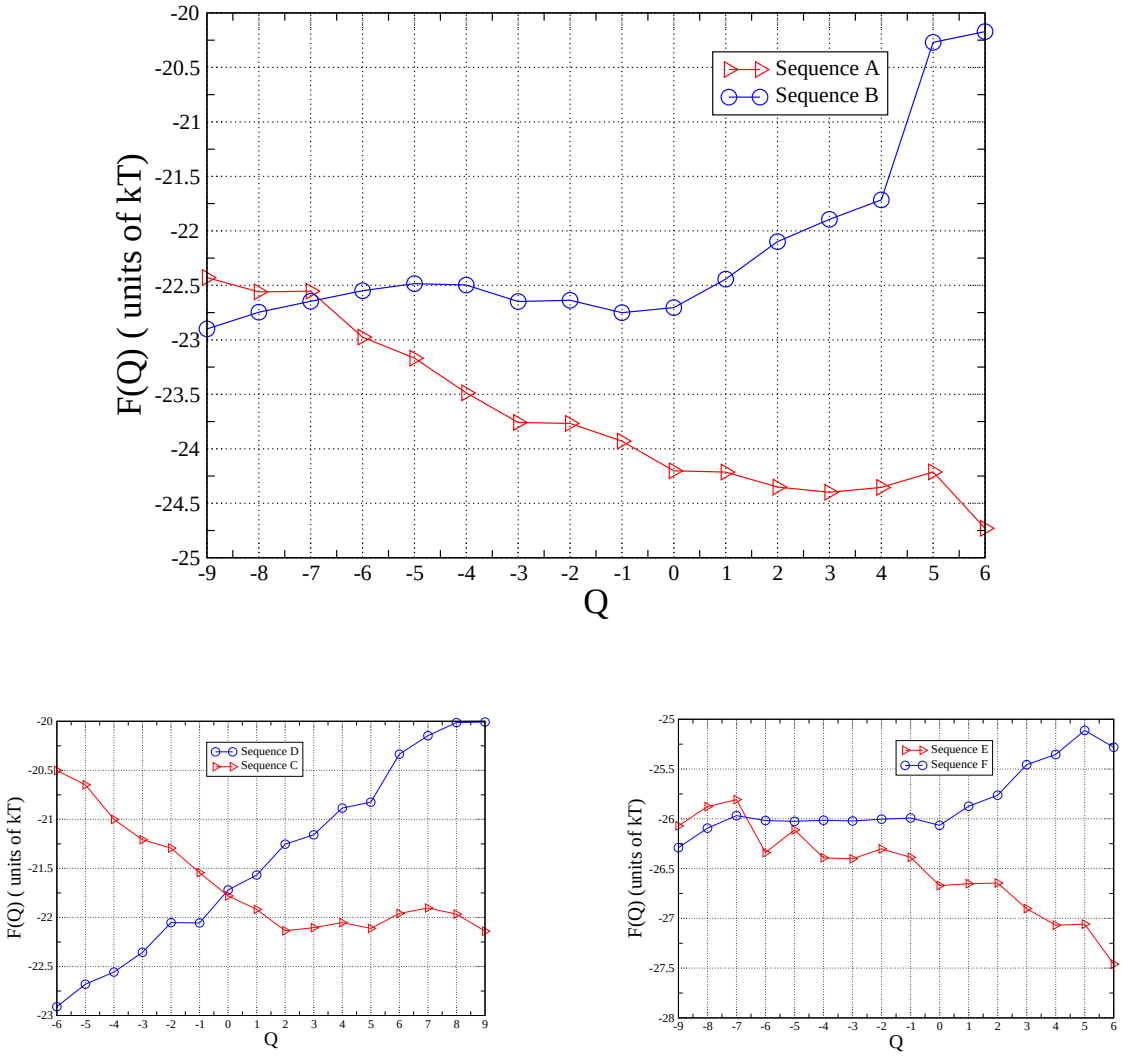


Figure 4.6: Plots of the free energy $F(Q)$ of all the sequence pairs (A⇔B,C⇔D, and E⇔F) as a function of the number of native contacts Q (Eq. 3.2). These data were obtained with a joint parallel tempering and umbrella sampling simulation. In this plot is visible the crossing point between the free energy curves where the energetic cost of the sequence switching is lower.

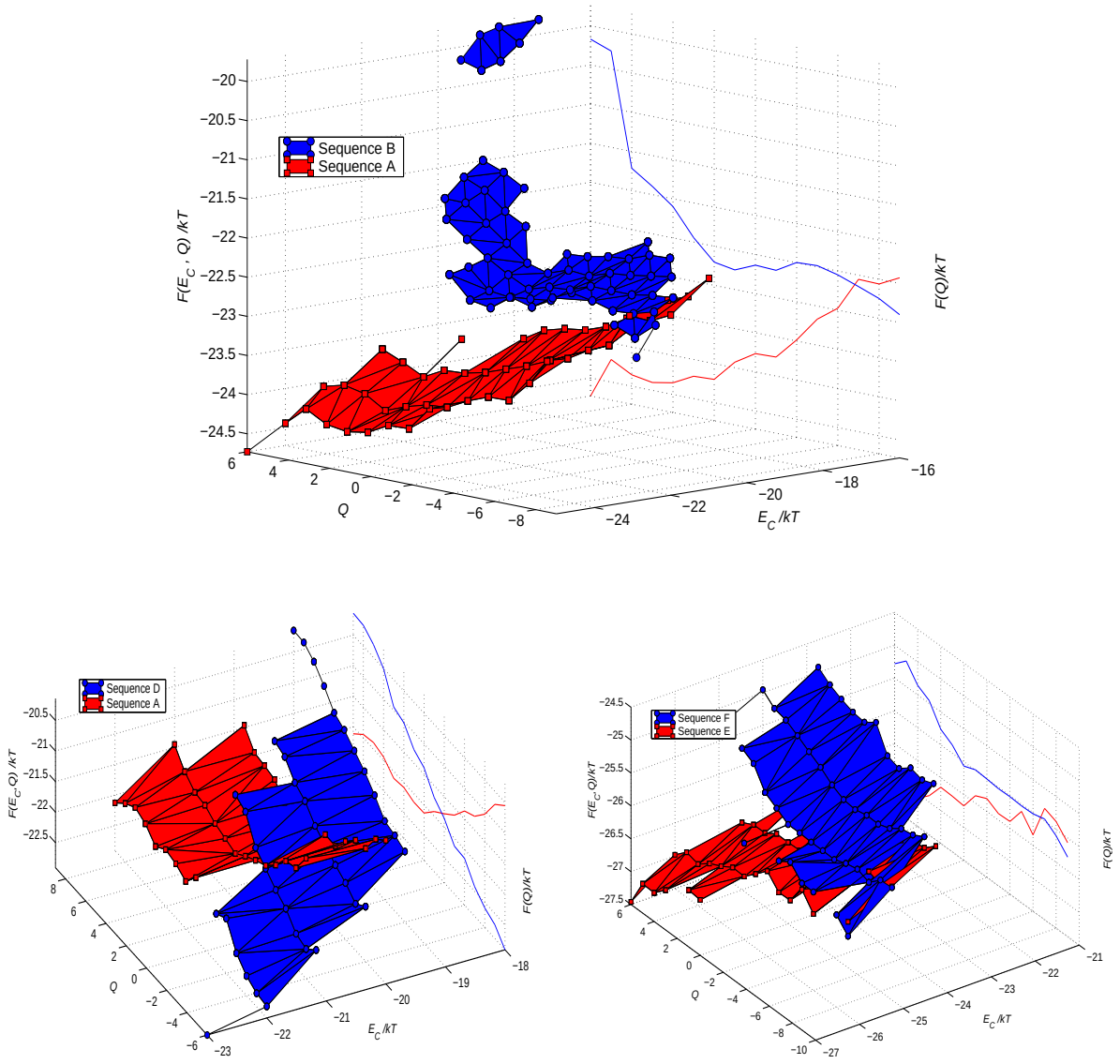


Figure 4.7: Plot of the free energy landscape of the folding of the sequences pairs A⇔B,C⇔D, and E⇔F. The free energies $F(E_C, Q)$ are function of the conformational energy E_C (Eq. 3.9) and of the number of native contacts Q (Eq. 3.2). The interesting feature of the plot is the overlapping area. The conformations in this area are common to both sequences. Comparing these plots with the corresponding ones in Fig. 4.6 (for convenience, we have replot it on the z-y plane) we see that close the states have the same conformational energy. From the parallel tempering simulation we know that those conformations can be reached just by thermal fluctuations.

thermal ratchet models [40]. Surprisingly, we find that, even without designing the pathway for refolding, our model spontaneously reproduces the "spring-like" gradual release of free energy during refolding.

To summarize, we have introduced a simple model to describe the behavior of a protein undergoing an allosteric transition. The protein is approximated by a linear heteropolymer on a lattice. The role of the external signal is played by an effective change of the amino acids along the chain. With this model we want to demonstrate that by destabilizing some essential elements of a conformation we can induce the chain to refold into a different structure. We can control this process by using a sequence design algorithm. With a folding program we characterized the equilibrium properties of the chain before and after the signal. Using the order parameters derived from the number of native contacts and the conformational energy, we compared the free energy landscape of the two sequences. The most important feature of the free energy plots is the overlapping region. The structures in this window are those where the contacts between the "mutated" residues and the rest of the amino acids are broken. In fact, they have the same conformational energy. In these particular states the energy cost for the transition is very low. We also emphasize that these states are accessible by thermal fluctuations. We believe that our model is able to reproduce the general behavior of allosteric transitions in proteins, where the external agent uses thermal fluctuations to lower the free energy cost of its action. This is also the basis of thermal ratchet models for molecular motors, where the thermal fluctuations are essential to drive the system.

5 Evolution of Protein Protein interaction

One of the key properties of biological molecules is that they can bind strongly to certain substrates yet interact only weakly with the very large number of other molecules that they encounter. Using a simple lattice model, we test several methods to design molecule-substrate binding specificity. We characterize the binding free energy and the binding energy as function of the size of the interacting units. Our simulations indicate that there exists a temperature window where specific binding is possible. Binding sites that have been designed to interact quite strongly with specific substrates are unlikely to bind non-specifically to other substrates. In other words, the conflict between specific interactions between small numbers of biomolecules and weak, non specific interaction with the rest, need not be a very serious design constraint.

5.1 Introduction

Biomolecules, such as proteins, tend to bind strongly to specific binding sites in target molecules. In addition, the binding needs to be selective: the molecules should bind strongly to one, or a few, partners and weakly, if at all, with all other biomolecules. The requirement that the binding should be strong and specific imposes constraints on the design of the binding sites. In particular, it suggests that binding sites should have a shape that is complementary to that of the substrate binding site and that its surface is patterned. Often, the total interaction (free) energy can be approximated as the sum of local intermolecular interactions that add coherently. In what follows, we focus on the role of the energetic patterning of binding sites.

It is important to recall that, even if the local intermolecular interactions are effectively random, binding is still possible. To see this, consider a non-specific interaction with an associated binding energy that is the sum of N terms. We assume that the individual contributions are Gaussian distributed with a zero mean and variance σ^2 [6]. The probability P of having a binding energy E is given by

$$P(E) = (2\pi N\sigma^2)^{-1/2} e^{-\left[E^2/2N\sigma^2\right]}, \tag{5.1}$$

where N is the size (the number of interaction sites) of the binding region. The probability to form a bond its determined by the Boltzmann factor $\exp(-\beta E)$ corresponding to the interaction energy E. Even if the average interaction energy is zero, two sufficiently large binding regions are still likely to bind, as the average Boltzmann factor is given by

$$\langle \exp(-\beta E) \rangle = \exp(N\sigma^2\beta^2/2)$$

This implies that for large N, a truly random binding site is not inert. The effect of non-specific ("random") interaction has been discussed in detail by Pande et al. in the context of a study of the freezing transition in heteropolymers [11]. Note that the effective interaction strength due to random interactions scales with N, just as is the case for the interaction strength of specific (designed) interactions. However, the average strength per monomer is larger for designed specific interactions and hence one might expect that for any N one can always find conditions where specific binding dominates. But this argument ignores the fact that the spread in the binding free energy for random sequences is proportional to $\sqrt{N}$. Hence, for small enough N there are, most likely, specific random sequences that bind at least as strongly as the "designed" sequence. As N increases, ($\sqrt{N}/N$ decreases) this becomes less of a problem.

The above discussion suggests that binding sites should contain a sufficiently large number of monomeric units in order to guarantee that a designed binding site binds significantly stronger to a given template than a random binding site. Yet the site should be sufficiently small that non-specific bonds can easily be disrupted by thermal fluctuations. One might think that this could be achieved by designing the individual site-site interactions to be small compared to the thermal energy k_BT. However, the same site-site interactions are responsible for the stability of the native state of the protein. Hence, weakening these interactions (or, equivalently, increasing the temperature), may result in denaturing of the protein, rather than in more specific binding.

There is a distinction between the specificity and selectivity of binding [41]. In order to quantify selectivity, it would be necessary to count the number of the substrate to which the protein can bind. In the present chapter, we do not attempt such an exhaustive search (as this would be prohibitively expensive for the model systems that we consider). However, Gutin et al. [42] showed, for a discrete version of the Random Energy Model (REM), that the probability of degeneracy of the lowest energy state decreases exponentially as its energy is lowered. This suggests that the specificity that we discuss below will, in most cases, also imply considerable selectivity.

In what follows, we consider under what conditions we can "design" a model substrate-binding site pair that binds significantly stronger than the corresponding "random-energy" pair, while maintaining the structural integrity of the native state of the protein in solution. Hence, binding and folding are both the consequence of the heterogeneous interactions between monomeric units. To this end, we explore the role of system size and temperature on the binding specificity in a model that mimics a general protein-substrate system. We consider two molecules, one of which (the "protein") is free to move, while the other is kept fixed and acts as the binding site of a substrate. We model the protein backbone as a linear, polypeptide-like heteropolymer living on a lattice. We then design ("evolve") the monomer sequence of the molecules according to three different scenarios, that we will refer to as OO,OR,and RR. First we consider the case of cooperative design, where the sequence of both the substrate and the ligand are evolved to increase the binding affinity. The second scenario is the model for a ligand that evolves to bind a substrate with a sequence that has been fixed a priori. The difference between model OO and OR lies in the role of the substrate. In scenario OO, the binding information is distributed over both protein and substrate: this approach should result in a protein-substrate pair that bind exclusively to each other. In the second approach, the protein

is designed to bind to a specific substrate which, in its turn, can have multiple binding partners (low selectivity). Case RR represents the case of a protein-substrate pair that does not bind. This is the reference state that allows us to define the specificity of the other two systems as a function of the substrate size and temperature. In the first section of this chapter we describe the simulation techniques that we used to design and study protein-substrate interaction. We then discuss the binding of the two different molecules on the same substrate. We conclude with a discussion of the potential implications of this work.

5.2 Methods

The system that we consider is a protein that is free to move in a box with hard walls in the presence of a substrate that is made of the same building blocks The box has a cubic shape and a lateral size of 3 time the length of the protein. The substrate is in the middle of the box. We model the chain as the linear, polypeptide-like heteropolymer, that we described in Chapter 2. The conformational energy of the system is given by the following expression

$$E = E_{\text{intra}} + E_{\text{inter}} = \sum_{i}^{N_C} \left[\sum_{j \neq i}^{N_C} C_{ij} S_{ij} + \sum_{j' \neq i}^{N_S} C_{ij'} S_{ij'} \right], \tag{5.2}$$

where the index i and j run over the residues of the protein, while j' runs over the elements of the substrate, C is the contact matrix, defined in Eq. 2.2.

We start by imposing the template configuration, which should give information on the structure of the protein and on the desired bound state (e.g. Fig. 5.1). From the mean field expression for the entropy in Eq. 2.3 we expect a wider distribution for the protein-substrate system, compared to the one of an isolated molecule. However, if the gap is still present, then the folded-bound state should be the equilibrium conformation. This condition does not exclude the case in which the interaction with the substrate is essential to keep the protein in the native state. Because we want to focus only on the binding properties regardless of the effect on the folding, we consider only a system with more intra-molecular than inter-molecular contacts – in other words we use only compact proteins with a large fraction of inter-molecular interactions.

In order to design the monomer sequence for the three different scenarios OO, OR, and RR we performed Monte Carlo sampling on a range of monomeric sequences. For each different scenario we applied the design process on a different subset of residues. In particular for case OO we include all the residues of both the protein and the substrate, while for the others the sampling is limited to the amino acids of the protein, while the structure of the substrate is fixed.

Again we are interested in the equilibrium properties of the system, and we calculate the free energy as function of the number of native contacts

$$Q = \sum_{i<j}^{N} \left[C^{*}_{ij} C_{ij} \right] \tag{5.3}$$

where C^* is the contact map of the native state. However, we will need also to measure the number of contacts with the substrate Q_s to measure the binding affinity of the protein with the substrate, and also to test whether the specific and random interactions can destabilize the native state of the protein inducing unwanted rearrangement of the peptide chain.

5.3 Results

5.3.1 Design the binding scenarios

To study the dependence of the binding specificity on system size and temperature, we consider a set of 4 different proteins with corresponding substrates. Each system was designed to reproduce the conditions of the three scenarios OO, OR, and RR. In order to design the first case we compute sequences of amino acids for the protein bound to the substrate, as shown in Fig. 5.1.a for a protein with 72 residues and a substrate with 24 amino acids. In this case the design program will optimize the sequence to minimize the energy of the contacts within the chain and between chain and substrate. For the case OR, we impose the same target configuration as before, but we limit the optimization to the amino acids of the protein, and we assign a random sequence to the substrate. The final scenario is for non-specific binding, this is achieved in two ways. First, we design a protein simply to fold into a given native structure, with no optimization of the substrate-binding energy. Second, we expose the protein form the OO and OR scenario to a random substrate without further design. It is important to stress that in the design of the OO and OR, the intra-molecular bonds are optimized together with the inter-molecular ones. In this way, we are able to construct model proteins that have the same internal structure both in the bound and unbound states. However, it is also possible to design structures that change upon binding. In Tab. 5.3, we list the amino-acid sequences that were the result of the design procedure described above.

5.3.2 Free energy calculations

As a first check, we verified that the generated sequences do indeed fold into the respective target structure. We show only the calculation of the binding free energy for a proteins consisting of 72 monomers (Fig. 5.1) as an example. In particular we consider the sequence OO (both protein and substrate optimized) and the sequence RR, where the protein sequence has been optimized to fold, but not to bind to a substrate, which has a random sequence. In Fig. 5.2 we plot the free energy of the sequences OO and RR, as function of the number of native contacts defined in Eq. 3.2. In each plot we distinguish between conformations that do and do not touch the substrate. As is to be expected (see Fig. 5.2), the binding free-energy is much larger in the case where both the binding site and the substrate have been optimized (OO),compared to the RR scenario. Moreover, in the case of the random interactions (RR), the free energy minimum is reached before all contacts with the substrate are satisfied. To characterize the system in this regime, we computed the free energy, $F(Q,Q_s)$, as function of both the number of native contacts and the number of non-specific contacts with the substrate. This should allow us to discriminate between conformations that are specifically and non-specifically bound

to the substrate. In Fig. 5.3.a and 5.3.b we plotted $F(Q,Q_s)$ for OO and RR, respectively. The "funnel" shape of the surface in Fig. 5.3.a demonstrates that the sequence OO does fold and sticks to the substrate in the designed way. In contrast, the free energy surface for the sequence RR is flat at the bottom of the slope. This indicates that, in this case, the folded protein does not have a unique bound state with significant binding free energy. So much so that the presumed target state is not even favorable from a free-energy point of view. For the other sequences that we studied we found that, in every case, the design process (OO,OR and RR) determined similar free energy landscape. The OR scenario (not shown in the figures) resulted in a free energy landscape similar to that obtained in the OO case, but the binding strength was less. It is important to notice that in all scenarios the free chain retains the native intra-molecular contacts, even in the unbound state. Ref. [43] discusses a different situation where the substrate is able induce conformational changes.

Next, we consider the dependence of the binding strength on the size of the binding site. In Fig. 5.4.a we plot the binding energy as a function of the size of the substrate for the three scenarios (OO, OR and RR). The error bars represent the spread of the random interactions given in Eq. 5.1 around the mean value (calculated at two sigma). From the interaction matrix that we used, we get a mean interaction energy of around zero [3]. The figure shows that there is a significant gap (more than 2σ) between the binding energy in the case of designed binding sites compared to that of the purely random case the designed energies and the boundaries of the distribution. The gap is large enough to guarantee that the designed binding is energetically favorable compared to the random case, even for the smallest substrate. As expected, the binding specificity increases with the substrate size.

As mentioned in the introduction, the presence of an energy gap between specific and non-specific binding is not a sufficient condition to guarantee specific binding at any given temperature. To ensure specific binding of a given protein, there should exist a range of temperatures that are low enough to ensure that the designed protein structure is stable, yet high enough to guarantee that random (non-specific) interactions are not strong enough to cause spurious bindings. As discussed in the introduction, it is not a priori obvious that such a temperature window always exists. However, in the present case, it appears possible to satisfy this condition. Fig. 5.4.b shows the free-energy difference between the bound and unbound states of the chain in the native conformation for the cases OO, OR and RR. As can be seen from the figure, the binding free energies behave more or less as the binding energies. In particular, a significant gap between specific and non-specific bonding is maintained. This holds both for the case where both protein and substrate have been optimized and even for the case where only the protein has been optimized. [1]

Clearly, the model used in the present study is highly simplified. Apart from the fact that we used a rather crude lattice model for the protein, we only considered the effect of binding energy on binding specificity. In reality, steric effects are at least as important and should be taken into account in any more realistic study. It would therefore be unwise to try to apply design calculations of the type described above to real protein systems. Nevertheless,

[1] Note that the definition of the binding free energy here is not yet system size independent. In order to correct for this we could perform the same simulation with periodic boundary conditions or compute the concentration of protein in the bulk.

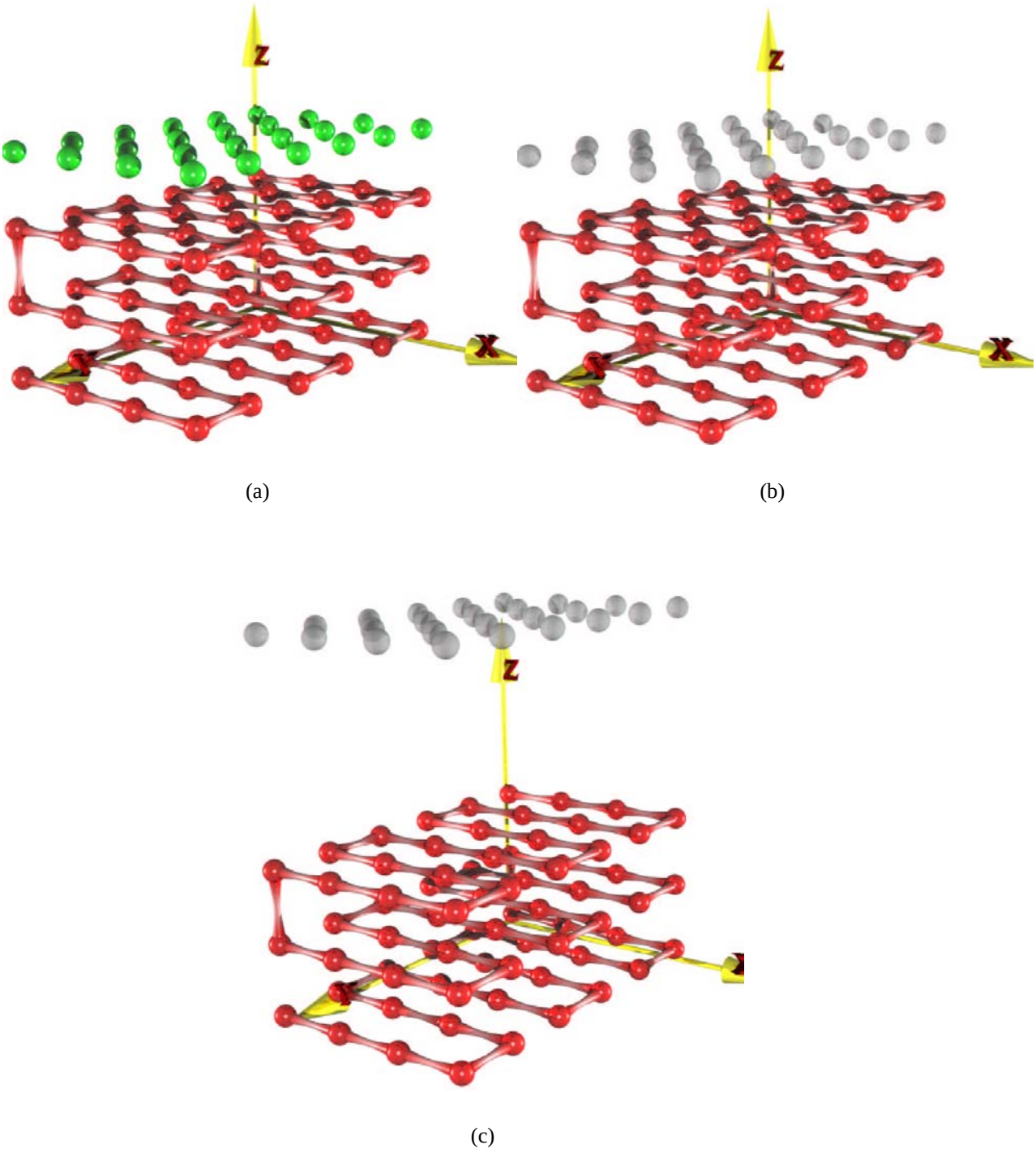
some of the conclusions that we reach are likely to survive the transition to a more realistic model. First of all, the existence of a temperature window where specific binding is possible, is also expected in models that take steric repulsion into account. Secondly (and interestingly), the present calculations suggest that binding sites that interact quite strongly with specific substrates are unlikely to bind non-specifically to other substrates. In other words, the conflict between specific interactions between small numbers of biomolecules and weak, non specific interaction with all the rest, need not be a serious design constraint. This latter statement should be qualified: as the number of distinct species increases, so does the probability that at least one pair of molecules will, by accident, have a strong, non-specific interaction. This will then result in an additional evolutionary pressure to keep non-specific protein-protein interactions weak.

We note that the design of specific binding sites also plays a role in experimental schemes to detect specific proteins [44]. In this case a clear differentiation of the binding affinity between a substrate and proteins in solution is essential to isolate a particular molecule. As before, this implies a temperature window in which non-specific bonds can be disrupted by thermal fluctuations, whilst the proteins themselves and the specific bonds that they form, are still stable.

Size	Scenario	Sequence	T_F
27	OO	YDCFRPIDGWRLQEMCKPNECWK NVEM **GSLYQFCTH**	0.2-0.5
27	OR	RQGCRDMDHIKWRELFKQSEVIK TMEL **YHYNGCNFP**	0.2-0.5
27	RR	MDSCRWLDCQKIMEFGKWMENQK WAER **HVPWYFKTP**	0.2-0.5
72	OO	NDCALCKNREFIDMKDPEWRVMRGY DWVQMKQREWRLFKDNECIACKNPE CTLCKYHEFIQMKDPEWPVMKH **GTFVTYHYSDWSLGHQNTGIACSS**	0.2-0.5
72	OR	CNQSLRECMKDIFREWWHQGARNPFND VGREMMKDGLREWCKQISPECAKQSLP ESMKQIGREWFKDTAHNF **YCTWTYHMPVPLFHDVYKVITYNC**	0.2-0.5
72	RR	GEQGDRKFLEQRNFKIIEMNSWHAIDMS NWKLLEMNDPKICEQRGPRFCDQADPK CLEMHQWKVIEMNSWRL **YCTWTYHMPVPLFHDVYKVITYNC**	0.2-0.5
75	OO	NDMRPCDWKNIEMRCIDFKLAEGRLFQ FKGIEMRLCDWKLNEMRCYQWKNSDM PPCQWKSIEMRCVQFKLGEFPV **VQGSTVTGSAHTWHAYDAHCYTWHY**	0.2-0.5
75	OR	NDGWSHMGRDREFWHCQFKDAELPCC QVKAREIPCYMLKQTEFWHSMFRGAD VWSYMLKAPEIWPCMLKQVEVPC **CYIYQHGGSNEMIKDKTTFTDRNNN**	0.2-0.5
75	RR	NMQESAKRWNIMDEACKRFLHGQDHCR PGYIFQECTKRWLNMDEASKRWNAMDE STKRWSIMQEGCKPFLHGQDC **WTYHMPVPLFHDVYKVITYNCVIFE**	0.2-0.5

Size	Scenario	Sequence	T_f
98	OO	YCMRDQFIRREWCHLCMKDDLGRKE WCINCMKEDGIRKEWFNIGMREDLVS KEWFLNFMKEDAGRKEWCNVCMKE DTIRREWCVYCMKDQLGPSQWCP **PCPYTPGLTTSVYYIAFQSHIGTYHP** **HANFPQHSALTQSMVNATFQHNV**	0.2-0.5
98	OR	IWSKICDQCLEDMLNWRHFCFPCFEEM NAWKKGDYVRGEDMTHWRHSPVAQS DDMYAWKKGDAPSGEEMANWKKFCQ HCLEEMNIWRKICYSCLEQMA **FNGTLTRRQYVTVIQYPFMCLRGYKV** **PCIFNQTTPHDTRSIRYHPWHWV**	0.2-0.5
98	RR	GMSIHQAYPELDWGNMKIKQHGREFEW NVMKCKDFASECEFLAMNCRSSASDCDW AVMKCKDAGRECEWVNMKCKQTYPELE WNGMRIKQHIPDLDWF **FNGTLTRRQYVTVIQYPFMCLRGYKV** **PCIFNQTTPHDTRSIRYHPWHWV**	0.2-0.5

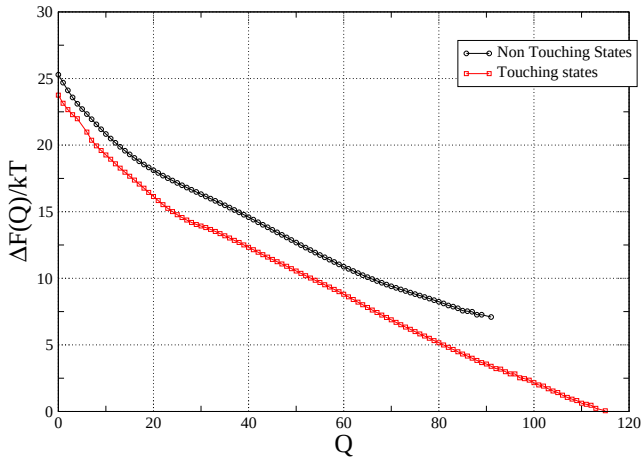
Table 5.3: Sequences designed in the three different evolutionary scenarios and for the different protein-substrate sizes. The parameters used where, the design temperature $\beta_D = 20$ and the permutation temperature $\beta_P = 24$ in the range. Each letter represents a different amino acid. The letters in bold are the amino acids of the substrate.



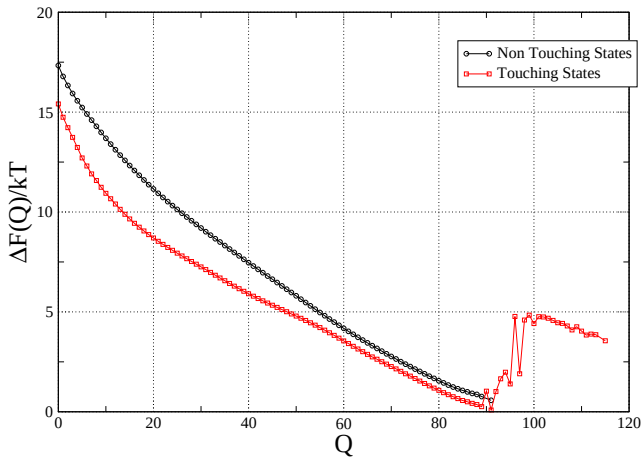
(a) (b)

(c)

Figure 5.1: Spatial arrangement of the 72 amino acid chain with its 24 residues substrate, for scenario OO (a), scenario OR (b), and scenario RR (c).

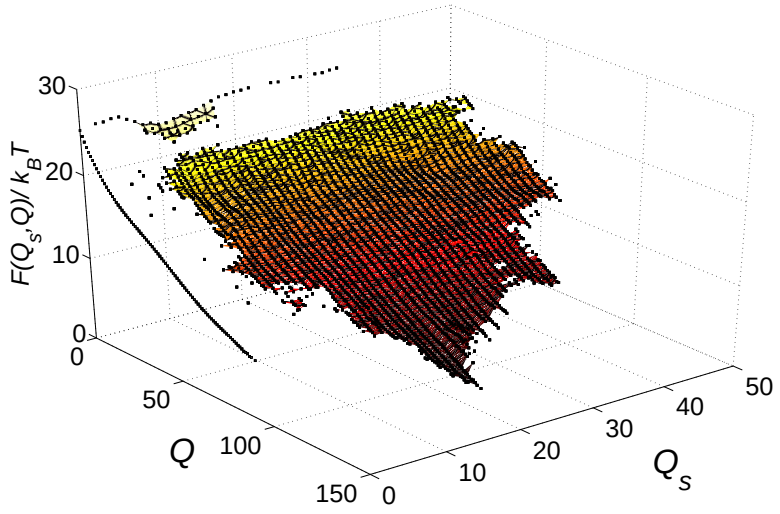


(a)

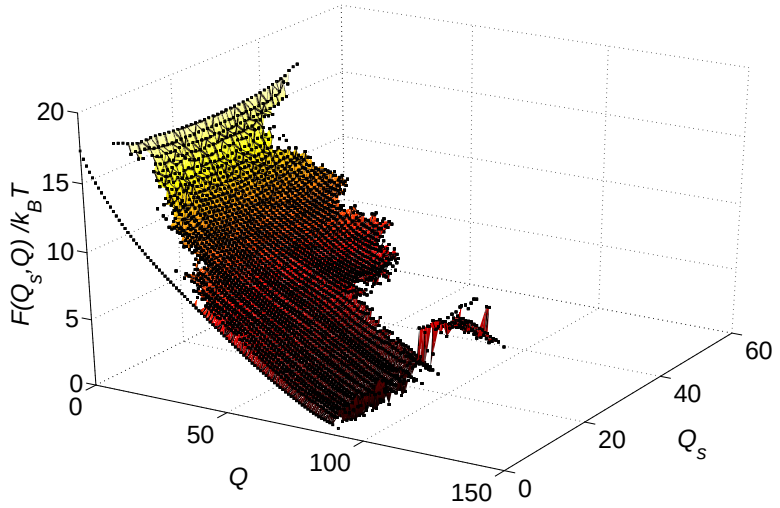


(b)

Figure 5.2: Plots of the free energy $F(Q)$ of the sequences OO (cooperative evolution) (a) and RR (independent evolution)(b), as a function of the number of native contacts Q (Eq. 3.2), at $T = 0.15$. States that touch the substrate (squares) have been plotted separately from those that do not (circles). The curve corresponding to the touching states is longer, because in the definition of the order parameter we take into account also the native contacts with the substrate. All data were obtained with a combined parallel tempering and umbrella sampling simulation.

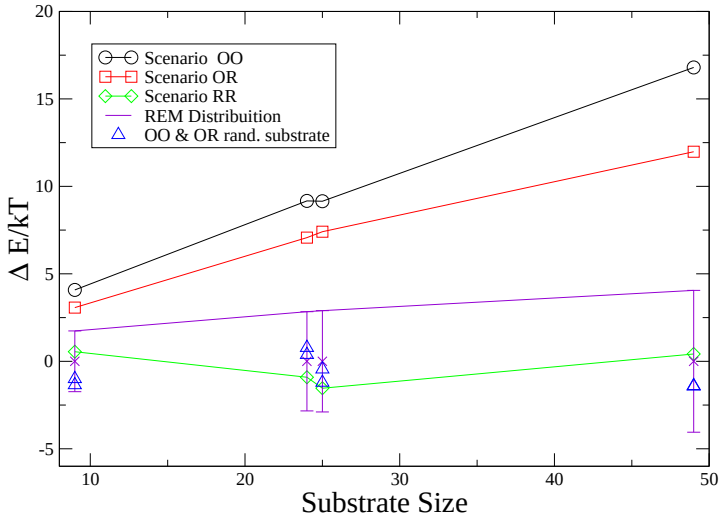


(a)

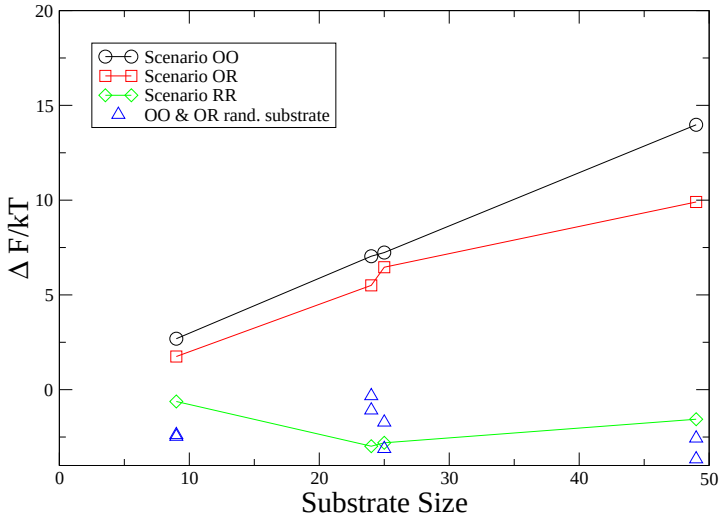


(b)

Figure 5.3: Plots of the free energy landscape $F(Q,Q_s)$ of the sequences OO (evolution for binding)(a) and RR (random interaction)(b), as a function of the number of native contacts Q (Eq. 3.2) and the number of contact with the substrate Q_s, at $T = 0.15$. The flat end of the slope in the second plot indicates that each bound state is equivalent in free energy to the unbound states. While in the first plot the funnel shape demonstrates that cooperative evolved sequence has a clear free energy advantage in the specific bind. The line separated form the surface represents the states that are not touching the substrate ($Q_s = 0$), and gap is caused by the poor sampling of the intermediate states.



(a)



(b)

Figure 5.4: Plot of the size dependence of the Binding Energy (a) and of the binding free energy (b). The error bars in the in (a) represents the 2σ width of the distribution of interaction given by the random energy model. The triangles represent the interaction of the proteins designed for the OO and OR scenario, with a random substrate [6].

6 Refoldable proteins and substrate interaction

Loosely speaking, folding is the process by which proteins explore their conformational space and find those conformations that have the lowest free energy. Many proteins can fold into more than one structure. The relative stability of these structures can often be changed by external agents such as the absorption of light, the binding to a substrate or the hydrolysis of a fuel molecule such as ATP.

In this chapter we explore the refolding of model proteins that is caused by binding to a substrate. To this end, we used the approach described in Chapter 2 to design protein-like lattice heteropolymers that have different stable conformations depending on whether they are bound to a substrate or free in solution. We considered three different systems that differ in the size of the interacting units. For small substrates, we observe that an increase in temperature induces the protein to unbind from the substrate, yet remain in its native state. For larger protein-substrate interaction sites, the bound conformation remains folded while the unbound conformation is denatured. We also considered the case where the unbound protein does not have a well-defined native structure. We found that, in that case, binding to a specific substrate could induce folding in the disordered protein.

6.1 Introduction

Proteins can change their conformation depending on their physical or chemical environment. A common example of such a structural change is unfolding, which happens when the protein is heated or exposed to denaturing substances. Denaturing is not a particularly subtle conformational change as the resulting state of the protein is disordered. However, in many cases proteins respond to an external stimulus by changing their spatial arrangement from one specific conformation to another. Driven conformational changes are thought to be at the root of many of the tasks that proteins perform. This is the reason for why proteins are often compared to nano-machines. Examples of proteins that perform a task by changing their conformation are motor proteins that generate forces or transport materials from one part of a cell to another.

The motor action is usually induced by the binding and subsequent hydrolysis of a “fuel” molecule, such as ATP. The energy that is released during the burning of this fuel is used to induce a large-scale rearrangement of the protein backbone. This, in its turn, causes the protein to move with respect to the substrate to which it is bound. Whilst this general picture of motor action is generally believed to be correct, the details of the motor action are, at present, largely

unknown.

Several models have been proposed to account for the mechanism of molecular-motor action. We will not review these models but rather describe some common features. The motor action is usually assumed to involve changes in the tertiary structure of the protein that is represented as a set of springs and links that connect larger structural units. When the APT is hydrolyzed, some of the physical links are cut and the energy stored in the springs force the protein to move to a new equilibrium conformation. In a number of cases, this picture could be supported by structural data derived from x-ray crystallography. In these experiments, the proteins are "frozen" during different stages of their power cycle. The resulting protein conformations are thereupon crystallized and studied by x-ray crystallography. In this way, conformational changes in the protein could be studied in detail. Clearly, this approach yields insight in the sequence of allosteric transitions that are involved in the motor action of proteins. However, necessarily, the technique neglects the contribution of thermal fluctuations and focuses on those parts of the molecule that are highly ordered. Much less is known about the evolution of disordered domains in the protein or about the role that fluctuations play during the power cycle of a motor protein. That fluctuations can be very important follows, for instance, from a recent study by Hawkins and McLeish [45] who proposed a coarse-grained model for *E-coli lac* and *trp* repressor. The binding of this repressor to its substrate can be strongly influenced by the binding of a second molecule to the repressor: even though this second binding hardly changes the conformation of the repressor, it does affect the thermal fluctuations in the molecule and this, in its turn, was shown to have a large effect on the binding of the repressor to its substrate.

In the present chapter, we use simulations of a simple lattice model to explore the effect of substrate binding on the structure and fluctuations of model proteins. Our approach is as follows: we consider pairs of proteins and substrates. The substrate are designed such that the binding will induce a conformational change in the protein. For simplicity, we assume that the interactions between the monomeric units of the protein and the substrate are similar to the intramolecular interactions between amino acids that belong to the same protein. In the context of our lattice model, this means that the amino-acid-substrate interactions are determined by the same interaction matrix as the intra-molecular interactions of the amino acids belonging to the same protein.

6.2 Methods

The model system that we consider is similar to the one discussed in the previous chapter, namely a protein confined in a cubic box in the presence of an immobile substrate. The conformational energy is defined by Eq. 5.2.

We start by designing simultaneously the sequence of amino acids to be compatible with both the initial (unbound) state (A) and the final bound state (B). Moreover, we impose that the most stable structure in the bound state is not the same as in the unbound state. As before, the design stage involves a Monte Carlo sampling over amino-acid sequences. The acceptance of a sequence-changing trial move is determined by three criteria: the first two are Metropolis-like rules that ensure that sequence changes that greatly increase the energy of either state

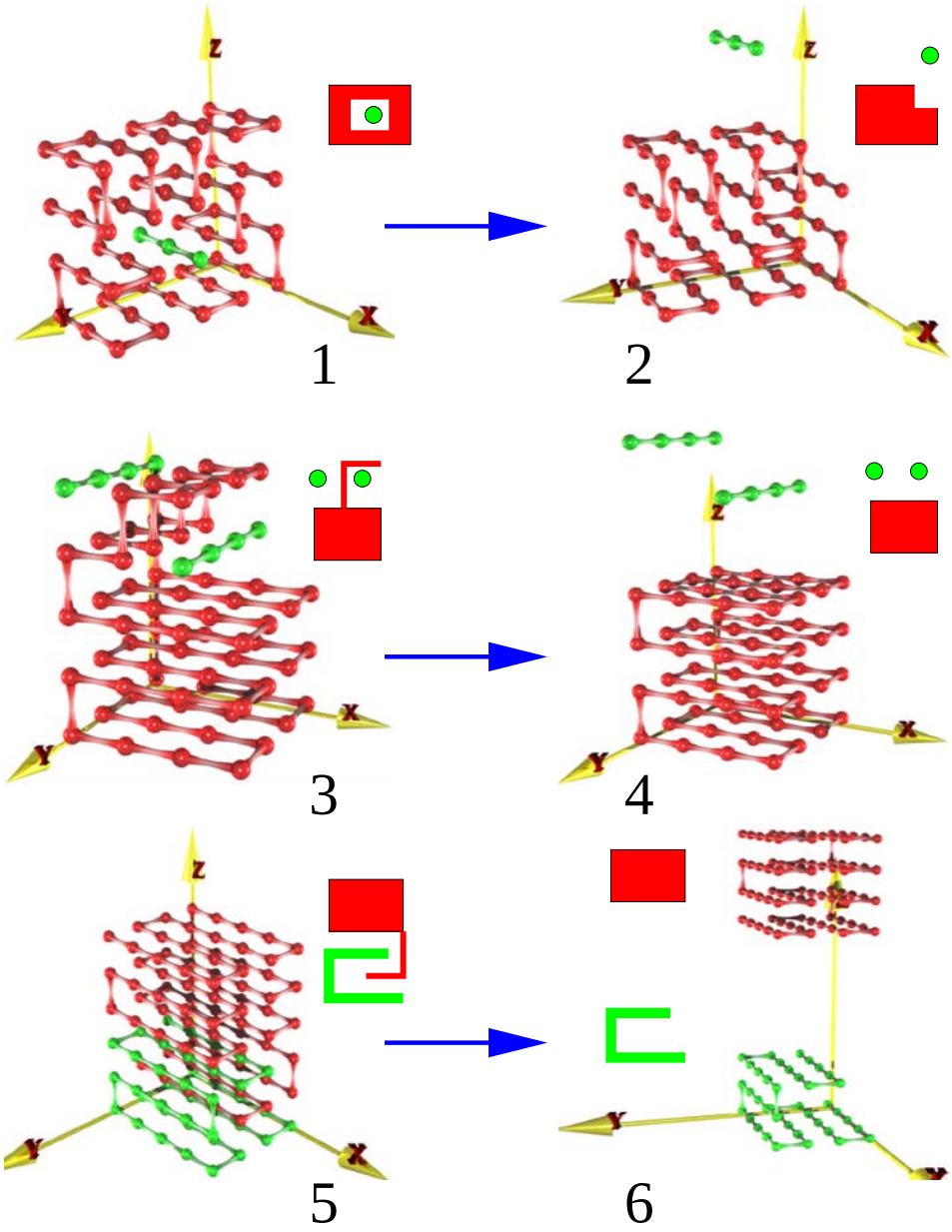


Figure 6.1: Spatial arrangement of the chain in the structures used to test the model

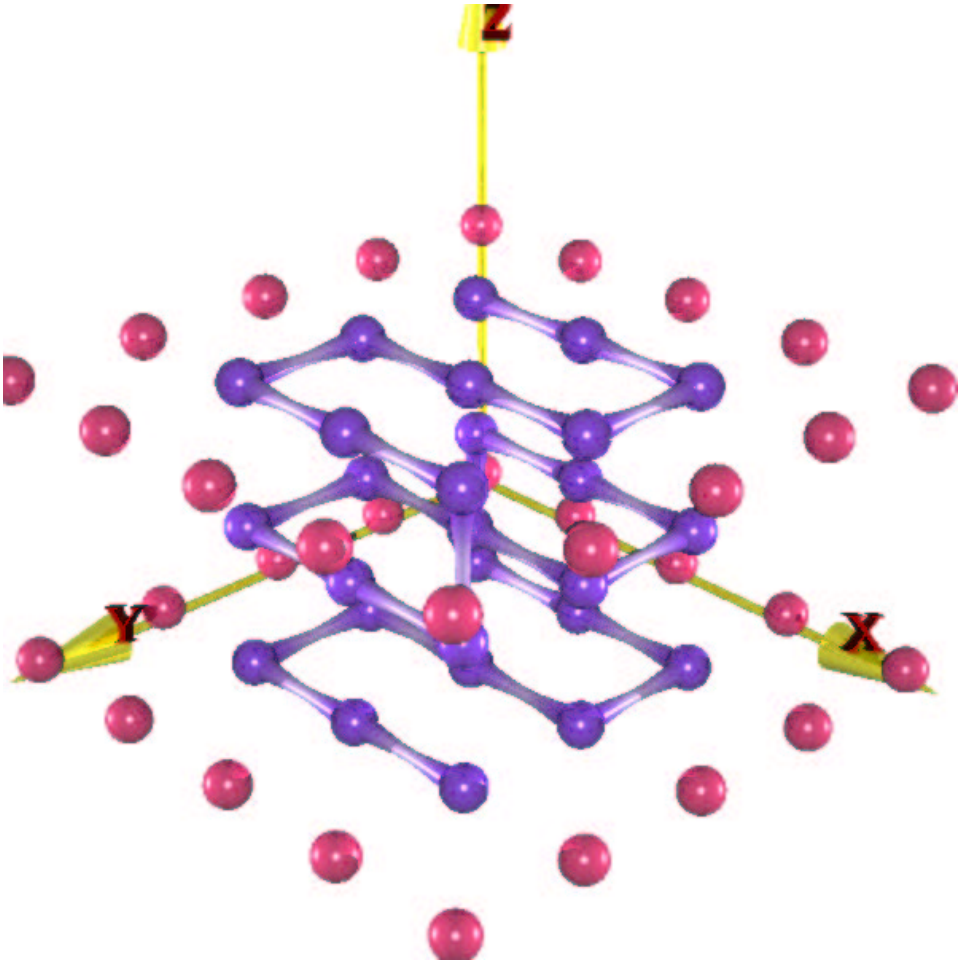


Figure 6.2: Spatial arrangements of the chain used to study the folding upon binding of a random domain

A or state B (or both) are rejected. The third acceptance criterion is related to the structural heterogeneity of the amino-acid sequences. This rule (Eq. 2.14) was described the preceding chapters.

Once a suitable sequence has been generated, we focus on the equilibrium properties of the system, and calculate the free energy as function of the number of native contacts (Eq. 3.2), and of the number of contacts with the substrate Q_s. In this way, we measure how the increase in the number of contact with the substrate correlates with the change in the native structure of the chain.

In the case where we studied protein folding *induced* by the substrate, we had to generate sequences that were in the coil state when unbound. We found that if we increased the configurational design "temperature" to a value that would generate random coil states in solution, then we would also find disordered bound states, yet when we chose a lower configurational temperature, the proteins in solution would typically fold into some compact state. To resolve this problem, we devised a scheme to control the randomness of the protein sequence. We randomly selected a certain percentage of amino acids that could evolve without constraints. In practice the design algorithm would randomly change their identity when selected for a mutation move, regardless of the Metropolis acceptance criteria. These amino acids would typically be irrelevant for folding. We found that, in this way, it was possible to design proteins that would be in the random coil state when unbound, yet in an ordered state when bound.

6.3 Results

To study the influence of a substrate on the equilibrium properties of our model protein we considered three different conformational changes induced by substrates of different sizes. In Fig. 6.1 we show the target structures between which the transitions occur: $1 \Leftrightarrow 2$, $3 \Leftrightarrow 4$, and $5 \Leftrightarrow 6$. In addition we studied the case of a protein that does not fold into a native state when it is in solution. However, upon binding the molecule assumes a designed target structure (see figure 6.2). This model could be relevant for the understanding of the role that random domains play in protein binding. The same design procedure was applied in every case. We therefore limit our explanation to design of the sequence that, upon binding, undergoes a conformational change from structure 1 (Fig. 6.1a) to structure 2 (Fig. 6.1b). Following the procedure explained in chapter 2 we optimize the conformational energy of the chain in both structure 1 (see Fig. 6.1.a) and 2 (see Fig. 6.1.b).

After eight independent simulations (typical length per run: 10^9 MC trial moves), we collect those sequences for which structure A is most stable in solution whilst structure B is most stable when bound. In all cases, we found at least one sequence that satisfied these constraints. In Tab. 6.2 we show the sequences that we selected to study the conformational changes shown in figure 6.1.

The design procedure is slightly different in the case where the molecule is in the disordered state in solution and folds upon binding (Fig. 6.2). Following the procedure explained in the Methods section above, we designed the protein in the bound state with different percentages of random amino acids ranging from 0% to 60%. The results are a group of sequences D0-D60 shown in Tab. 6.3.

RHFSYTRRGMDDRCWVCDACVMCT PHWLEYNKILENPKIMEQRKWGEDP KFAEQNKI **MSQ**	Sequence A
LEASPSKIREGYPGRTRDFYWCKDLEC MNCKILECNWCKIRECMHFRDPDF YWCKQVECM-NCKV **VATGQHQH**	Sequence B
PRDGLWGRDQPRDFMIFRDYMKDCLW CKEWNKECMICRENNKDCLWCK ENMKECMICK-EWFKDCLWCKEFNKECMI CREN **PRQFMIGHQHHHPGLVTSTYAVVAAVT SYYPSQAHVGSTQ**	Sequence C

Table 6.2: Sequences generated for the test structures (Fig. 6.1). Each letter represents a different amino acid [3]. The letters in bold are the amino acids of the substrate

MKCREWLKDREIMK DCEWNRFREPLKD **HQITVMFPWQCYCTSAYGDVVIYNSNQFAGTH**	D0
MKCREWLKDREIMK DCEWNRFREPLKD **HQITVMFPWQCYCTSAYGDVVIYNSNQFAGTH**	D6
MKCREWLKDREIMK DCEWNRFREPLKD **HQITVMFPWQCYCTSAYGDVVIYNSNQFAGTH**	D12
CRNPECFKQWEGCK MRECIKDWELGKM **PDVAFHHSCQNTNYSTAWQGVFILTLDRYHMP**	D18
KMIPWECMNDWCKM RLWERMIEWDFPR **NY-CFCKEADFVILYNSTQHGHGRQSTVAALKT**	D24
GECPRELWRWRFRE MCKDPEFVKQFNM **DMIYIK-TATHCACQDSVPNGSLNYHQKLYGIW**	D30
GECPRELWRWHIRE MCKDPEHVKQFNM **DMISIKFNTKCSCYDLSFNGAGPQAQHTVYVW**	D36
GGCPRELWRWRFKE MCKDPNHVNEFNM **DMISIK-TATTCACQDYIWNFLGKQSQHYHYVP**	D42
GGCPPELWDMQFRR WHDDPNEVEHFNI **KMISIK-TATGCWCYRHVLNFAMYQAQYKSKVL**	D48
GGCPPELWDMQFRR WHDDPNEVEHFNI **KMISIK-TATGCWCYRHVLNFAMYQAQYKSKVL**	D54
GGCPRELWDMHFRE WCKDPNWVKHFNI **RMI-AIKGSPTCYCQKVALDFSQEYAYNMTMVL**	D60

Table 6.3: Sequences generated for the study of binding random domain proteins (Fig.6.2). Each letter represents a different amino acid [3]. The letters in bold are the amino acids of the substrate

6.3.1 Free energy calculations

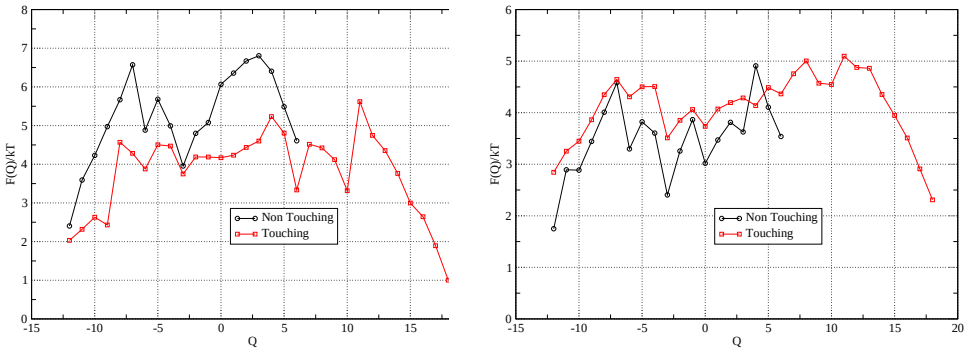
Having designed the sequences for the structures depicted in figures 6.1 and 6.2, we verified that the generated sequences do indeed fold into the respective target structures in both the bound and (where applicable) the unbound states. In Fig. 6.3.a-6.5.a, we plot the free energy of sequences A,B,C respectively, as a function of the number of native contacts Q (Eq. 3.2) at a temperature of $T = 0.1$. In each plot we distinguish between conformations that do and do not touch the substrate. A common feature of the three proteins is that they fold into different target structures, depending on their binding state. For example, in Fig. 6.3, the lowest free-energy conformation in the bound state corresponds to structure 1 ($Q = 18$), whilst in the unbound state structure 2 ($Q = -12$) has the lower free energy. The same applies for sequences B and C that were designed for the transitions $3 \Leftrightarrow 4$ and $5 \Leftrightarrow 6$ respectively. This result demonstrates it is feasible to design model proteins that undergo conformational changes upon binding to a substrate. In addition, in Fig. 6.3 the barrier to go from conformation 2 ($Q = -12$) to conformation 1 ($Q = 6$) is higher for the unbound conformation than for the bound conformations. This suggests that binding is likely to precede refolding. Additional evidence that this may be the case comes from the shape of the free energy surface $F(Q, Q_s)$ that is a function of the number of native contacts Q and of the number of contacts with the substrate Q_s (see Figs. 6.7.a-6.9.a). For instance, Fig. 6.7.a shows that when the protein has an order parameter close to that of the unbound native state (conformation 2), the bound states with high values of Q_s are not favorable. Rather, the free-energy landscape for refolding is fairly flat except in the vicinity of the target state (1). This suggests that the protein is first weakly absorbed on the substrate. From then on refolding and increased binding[1] to the substrate occur together.

Let us next compare the behavior of the different substrate sizes and consider the effect of temperature. To test the thermal stability of the different conformational changes, we increase the temperature until we reach a regime where the native state of the free protein is in equilibrium with the native bound conformation. For small substrates ($1 \Leftrightarrow 2$ and $3 \Leftrightarrow 4$) the temperature increase will favor the unbound conformation, without denaturing the protein. The situation is different for the transition $5 \Leftrightarrow 6$, where the size of the substrate is larger. In this case, there is a temperature region where the bound state still folds into the designed structure, whilst the unbound state is denatured. Hence in this case the substrate acts to increase the thermal stability of a particular protein conformation.

Note that an induced conformational change such as the one from $1 \Leftrightarrow 2$ could act as a form conformation-mediated signal transmission. The interaction of a protein with a small molecule or small binding site, induces a substantial rearrangement of the chain, which changes the nature of the exposed surface of the protein. It would interesting to study the nature of this signal transmission more extensively, such a study would fall outside the scope of this thesis.

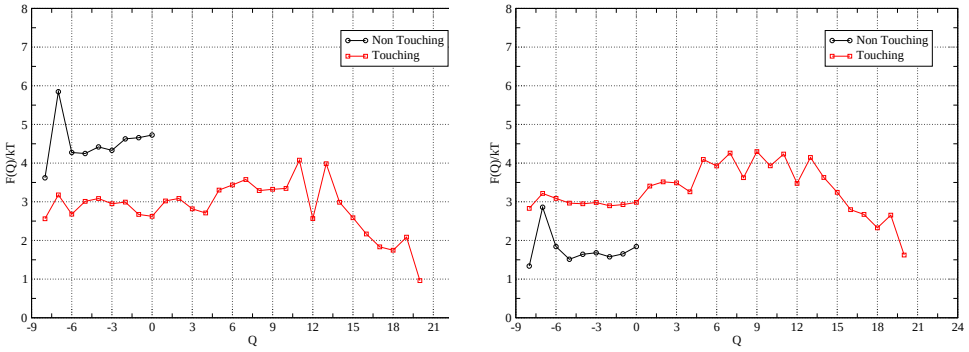
Let us finally consider the case of a protein coil that folds into a compact state when binds to a substrate. In Fig. 6.6.a-b we plot the free energy of, respectively, the free and bound states of sequences D, as function of the number of native contacts Q (Eq. 3.2). Not surprisingly,

[1] Note that the definition of the binding free energy here is not yet system size independent. In order to correct for this we could perform the same simulation with periodic boundary conditions or compute the concentration of protein in the bulk.



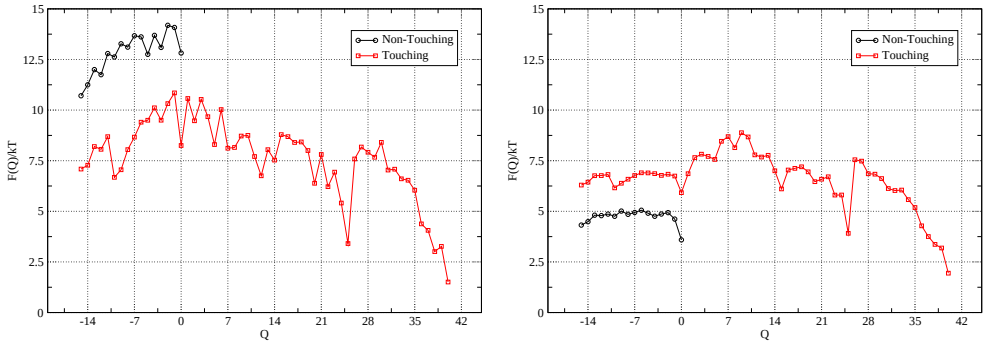
(a) (b)

Figure 6.3: Plots of the free energy $F(Q)$ of sequence A as a function of the number of native contacts Q (Eq. 3.2), at $T = 0.10$(a) and at temperature $T = 0.30$(b). States that touch the substrate (squares) have been plotted separately from those that do not (circles). The curve corresponding to the touching states is longer, because in the definition of the order parameter we take into account also the native contacts with the substrate. All data were obtained with a combined parallel tempering and umbrella sampling simulation.



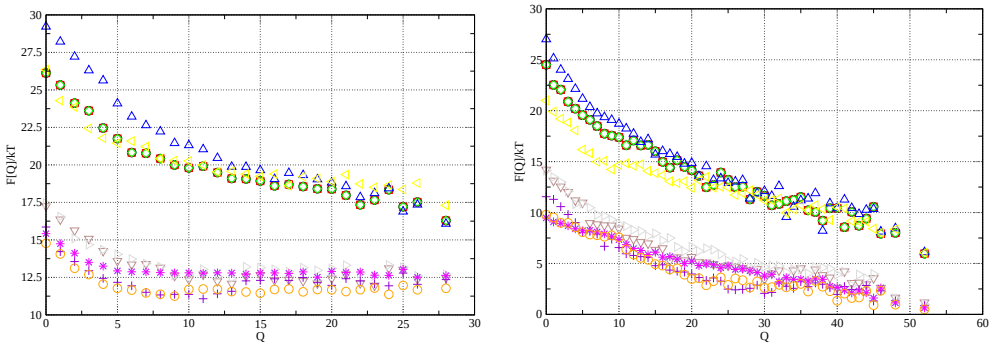
(a) (b)

Figure 6.4: Plots of the free energy $F(Q)$ of sequence B as a function of the number of native contacts Q (Eq. 3.2), at $T = 0.10$(a) and at temperature $T = 0.30$(b). States that touch the substrate (squares) have been plotted separately from those that do not (circles). The curve corresponding to the touching states is longer, because in the definition of the order parameter we take into account also the native contacts with the substrate. All data were obtained with a combined parallel tempering and umbrella sampling simulation.



(a) (b)

Figure 6.5: Plots of the free energy $F(Q)$ of sequence C as a function of the number of native contacts Q (Eq. 3.2), at $T = 0.10$(a) and at temperature $T = 0.50$(b). States that touch the substrate (squares) have been plotted separately from those that do not (circles). The curve corresponding to the touching states is longer, because in the definition of the order parameter we take into account also the native contacts with the substrate. All data were obtained with a combined parallel tempering and umbrella sampling simulation.

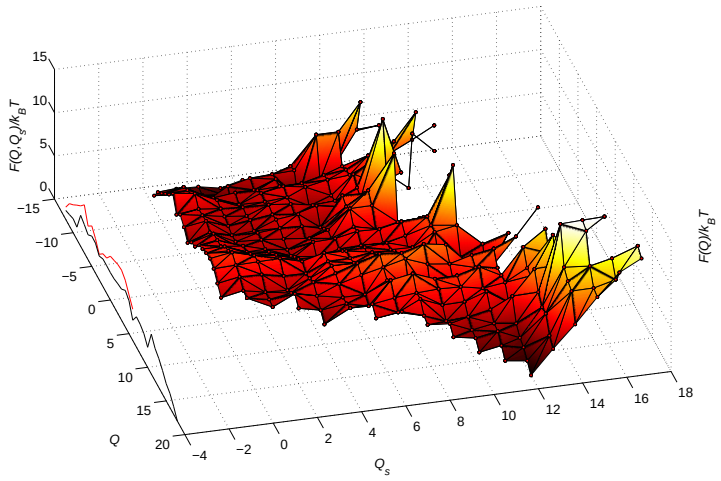


(a) (b)

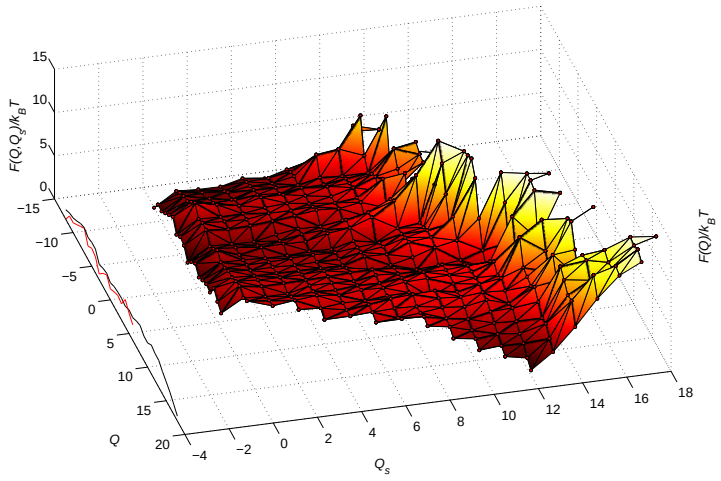
Figure 6.6: Plots of the free energy $F(Q)$ of sequences D0-D60 (0-60% of random amino acids) as a function of the number of native contacts Q (Eq. 3.2), at $T = 0.10$. States that touch the substrate are plotted separately (b) from those that do not (a). The gap between the curves does not have a physical meaning, and has been introduced only to separate the free energies of sequences that folds in the bulk from the one the do not. Note that all proteins fold upon binding.

above a certain threshold of “randomness” (30%) the unbound chain does not have a compact native state. However we found that even in the case where 60% of the amino acids were chosen at random, the substrate-bound state would still fold into the native state. Of course, these numbers should be treated with caution as the quantitative results of the simulations are expected to be model dependent. The main message of Fig. 6.6.a-b is that a disordered protein (or, for that matter, a disordered protein domain) can be involved in selective binding and, in the process, fold into an ordered conformation. Although we have not tested directly whether protein protein can be mediated by an apparently random domain, the present results strongly suggest that this is the case.

To summarize, we have used a lattice model to describe the conformational changes in protein systems. This particular phenomenon is often triggered by the interaction of the protein with an external agent, that we model as substrate fixed in the simulation box. The first result was the successful design of a system with two equilibrium conformations, one for the bound state and the other one for the free state. We then computed the free energy of the proteins at different temperatures – distinguishing between touching and non-touching conformations. The behavior at low temperature was mainly characterized by a strong preference for the bound states 1, 3 and 5. When the temperature of the system was increased, the free energy of the unbound states decreased because of the increased importance of translational entropy. However the response of the system, was not simply a gradual denaturation of the proteins. Rather, it depended on the number of interactions with the substrate. At higher temperatures the inter-molecular interactions of small substrates ($1 \Leftrightarrow 2$ and $3 \Leftrightarrow 4$) were not strong enough to compensate for the increase of the translational-entropy term, however the intra-molecular bonds were still stable and they could keep the protein in the native state 1 and 3 respectively. This picture starts to change when we consider a larger substrate. In this case the strength of the intra-molecular and inter-molecular interactions were comparable. As consequence, the translational-entropy term was never strong enough to favor the unbound states, before the protein was unfolded.

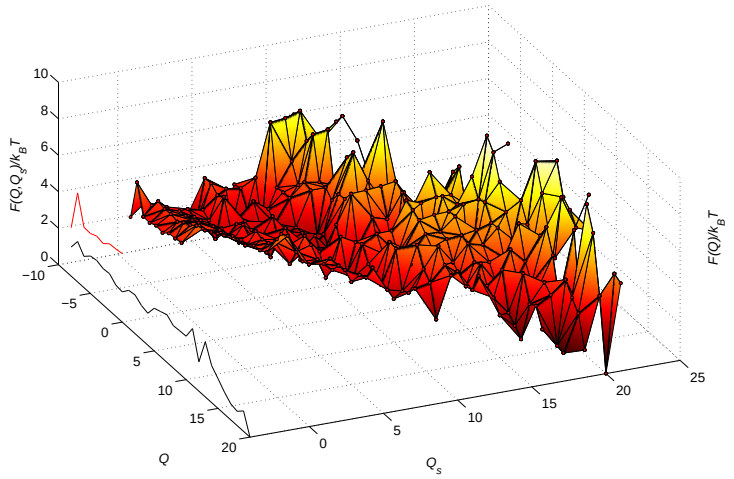


(a)

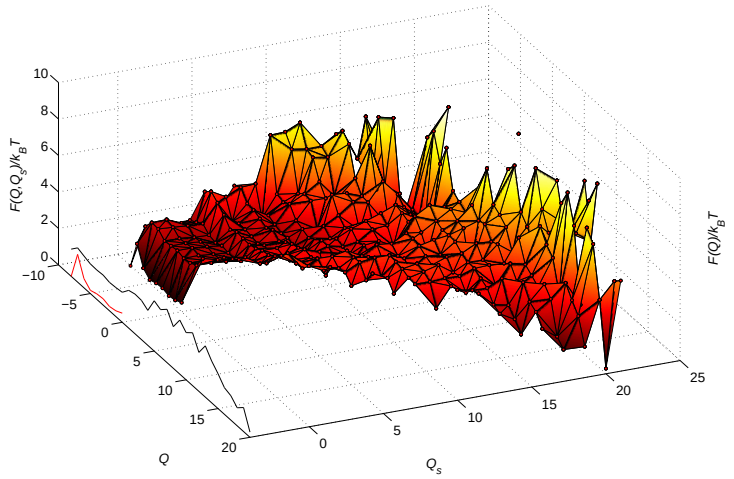


(b)

Figure 6.7: Plots of the free energy landscape $F(Q, Q_S)$ of sequence A as a function of the number of native contacts Q (Eq. 3.2) and of the number of contacts with the substrate Q_S, at $T = 0.10$(a) and at temperature $T = 0.30$(b). All data were obtained with a combined parallel tempering and umbrella sampling simulation.

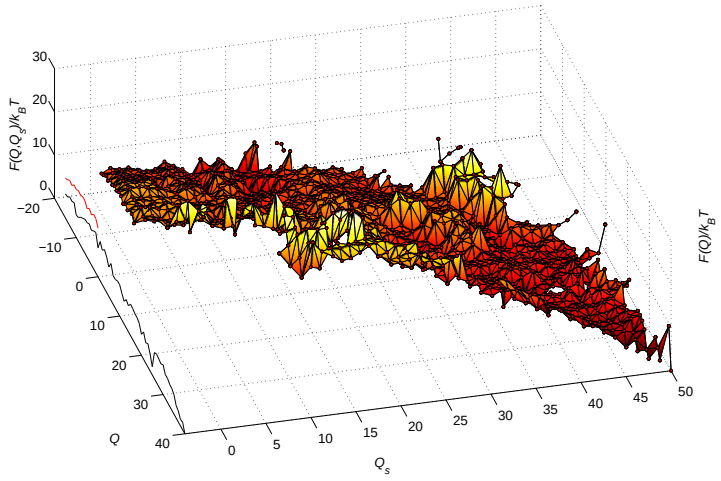


(a)

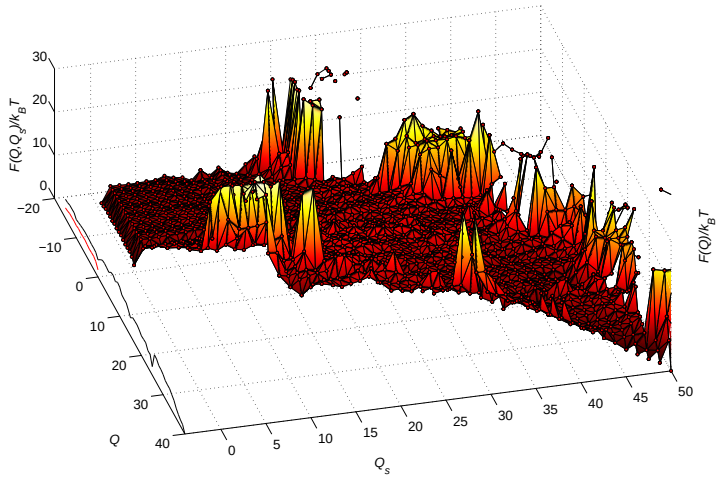


(b)

Figure 6.8: Color Online) Plots of the free energy landscape $F(Q, Q_s)$ of sequence B as a function of the number of native contacts Q (Eq. 3.2) and of the number of contacts with the substrate Q_S, at $T = 0.10$(a) and at temperature $T = 0.30$(b).



(a)



(b)

Figure 6.9: Color Online) Plots of the free energy landscape $F(Q, Q_s)$ of sequence C as a function of the number of native contacts Q (Eq. 3.2) and of the number of contacts with the substrate Q_S, at $T = 0.10$(a) and at temperature $T = 0.50$(b).

7 Simple model for chaperon action

7.1 Introduction

Misfolding of proteins in living cells may lead to serious malfunctioning of the cell machinery and even premature cell death. The problems are not only related to the fact that misfolding disrupts the normal activity of the protein itself, it can also induce uncontrolled protein aggregation as observed in diseases such as Alzheimer or BSE. To prevent such cellular disasters from happening, a healthy cell contains protein complexes ("chaperones") that assist many proteins in finding their correct native structure. Interestingly, a single chaperone complex can assist the folding of a variety of proteins, with quite different amino-acid sequence. This means that, although the chaperone specifically targets misfolded proteins, its action is not sequence-selective. An interesting example of such a specific but not selective chaperone is the GroEL/GroES complex. This complex has a cage structure in which it can capture misfolded proteins. The chaperone then makes such a misfolded protein go through a number of refolding cycles until it reaches its native state. This process is quite subtle because the chaperone must distinguish a misfolded state from a native state. As chaperones can refold proteins with different sequences, the mechanism cannot rely on a selective protein recognition process to identify the native state.

In their 1998 review, Sigler et al. [46] propose the following scenario for GroEL chaperonin action (Fig. 7.1), based on the available experimental evidence: initially the GroEL complex is in an "open-barrel" state, thus exposing a hydrophobic surface to which the target proteins can bind. The next step is an ATP-driven capping of the GroEL barrel by the GroES cap. The misfolded protein is now trapped inside the protein complex, the inner surface of which is mainly hydrophilic. The final step consists of the hydrolysis of the ATP and the release of the cap. At this stage the protein is presumably released in its native state. Interestingly, the GroEL complex has a symmetric double barrel structure and, after the first barrel releases its protein, the other barrel is ready to start refolding the next protein. The mechanism by which the chaperonin helps the misfolded protein to reach its native state is still matter of debate. A possibility is that the chaperonin simply captures misfolded protein and keeps them isolated from the rest of molecules in solution until they have had time to fold spontaneously into their native structure. But recent models suggest a more active role of the chaperonin. Jewett et al. [47] proposed a model for the chaperonin cavity, where the internal walls of the cage are designed to have a weak attractive interaction with the hydrophobic residues of the protein. In their simulations Jewett et al. found that, by a judicious choice of the strength of the attractive interactions, they could greatly enhance the refolding rate of model proteins.

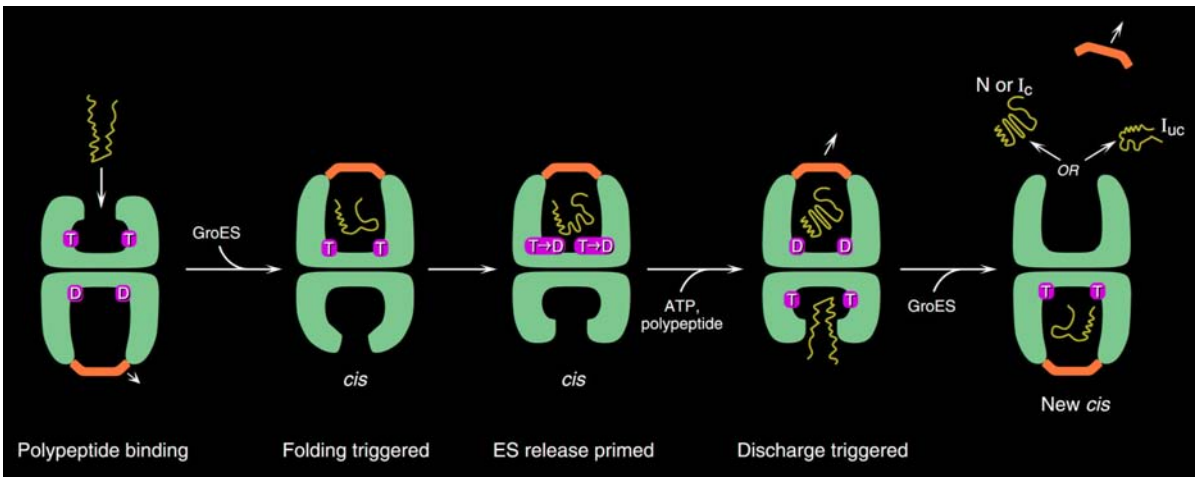


Figure 7.1: Schematic representation of the current interpretation of the refolding cycle of the GroEL/GroES chaperonin complex. The cycle consists of five steps. The overall duration of one cycle is approximately ~15s. During the first step the misfolded protein is trapped on the rim of the open cavity. During the subsequent two steps, the cage closes and the protein is assumed to refold. During these steps, some seven molecules of ATP are hydrolyzed to induce the conformational change, and another seven bind to the lower cavity to prepare it for the next encapsulation. The final step corresponds to the release of the target protein (folded or misfolded) and the encapsulation of the next target in the lower chamber.

This high refolding rate is presumably due to a partial adsorption of the protein on the surface of the cavity – the partial adsorption lowers the barrier to go from a misfolded state to the native state. In other words: the action of the chaperonin is mainly to assist the unfolding of non-native states, rather than to actively fold into the native state.

In the present chapter, we report simulations that lead us to propose an alternative unfolding mechanism that is effective, yet sequence-unspecific. Our model provides an interesting rationale for the double-barrelled structure of the GroEL complex.

In the scenario that we propose, the central step in the refolding process is a translocation of misfolded proteins from one barrel of the GroEL complex to the other. Such a translocation may be possible because there is a hole connecting the two barrels of the chaperonin complex. In fact, crystallographic studies suggest that the chaperonin complex has a well-defined structure, except for a fairly large ($\sim 30\,\text{Å}$) "hole" between the two barrels [46]. However, the hole in the X-ray structure does not mean that there are no amino acids in this region but only that they are disordered and highly mobile. Low resolution small-angle neutron scattering experiments [48] suggest that there is, in fact, a high density of disordered amino acids in the vicinity of the central hole. Yet, this does not preclude translocation through this region. In a different context [49] it is well known that disordered protein filaments near a translocation pore do not close such a pore but rather play an important role in adding selectivity to the translocation process. We assume that the something similar happens in the GroEL complex: there is a hole between the two barrels, but its effective diameter is less than 30 Å.

The efficiency that a translocation based chaperon can achieve in removing misfolded proteins from the solution is higher that in case of previous models. In fact if a first refolding cycle is not successful that the misfolded protein is released in solution and through diffusion should reach the other chamber of the chaperon which is in the trapping mode. We can give an upper limit to the probability that a just released protein will actually reach the other open chamber after a long time and this is given by [50]

$$P \simeq \left(\frac{r}{d}\right) \frac{k_r}{k_r + 4\pi D r},$$

where r is the radius of the trapping chamber (~ 45Å), d is the starting distance (~ 160Å), k_r is the association reaction rate, and D is the diffusion constant of the protein. If we consider the case of a diffusion limited reaction, then we can reduce the probability to the ratio between the r and d which in the case of the chaperon is ~ 0.3. This implies that in best scenario only 30 % of the misfolded proteins will rebind, while a translocating chaperon will have much higher probability of keeping the misfolded molecule away and try the necessary refolding cycles. We still need to demonstrate that such a small hole can play a crucial role in protein refolding. This is the central question that we address by simulation.

Our simulations show that the active confinement of a (misfolded) protein in a small hydrophilic cage is enough to cause it to translocate rapidly to the other (open) barrel. The translocation process will break any pre-existing compact structures. We find that, as the translocation proceeds, the protein refolds in the open barrel. In fact, this refolding provides an important contribution to the thermodynamic driving force for the translocation. There exist other examples of such a translocation-plus-refolding processes. One is the synthesis of a polypeptide in the Ribosome. Using a lattice model similar to the one we employ, Morrissey

et al. [51] showed that a protein that is extruded gradually from the ribosome folds much faster than from its fully denatured state. As the protein refolds in the open barrel of the chaperone complex, it need not end up in its native state. However, as the surface of the open barrel is predominantly hydrophobic, it has the tendency to trap misfolded proteins, but let correctly folded proteins escape.

Modelling of the GroEL complex is greatly facilitated by the fact that we know that its action is not sequence specific. This implies that the essence of the action of this chaperone can be represented by a model that can describe protein (mis)folding and the interaction of poly-peptides with heterogeneous protein surfaces. To this end, we use a lattice model to simulate the poly-peptide and the chaperone cage. The building blocks of both the poly-peptide and the chaperone cage are amino acids with interaction parameters as derived by refs. [52, 53].

Although this lattice model provides only a very crude representation of a protein in a chaperone cage, it retains important characteristics such as the amino-acid heterogeneity and the foldability of the protein chain. In our simulations, we design an artificial protein that folds into a given target structure. We then study the behavior of this model protein in the chaperon cavity and we compute the free-energy barrier that the protein must overcome in order to move from one chaperone barrel to the other.

7.2 Methods

We consider a lattice protein and a model chaperone in a cubic simulation box with hard walls. The lateral size of the simulation box is three times the contour length diameter of the protein. In the middle of the simulation box we placed a cubic "chaperone" cage (Fig. 7.2) with a lateral size of 5 lattice units. The volume of this cage is large enough to contain the protein in a compact (but not necessarily native) state.

We model the protein as a linear heteropolymer, living on a lattice. The amino-acids of the chain have nearest-neighbor interactions. The conformational energy E of the system is given by the following expression

$$E = E_{\text{intra}} + E_{\text{inter}} = \sum_{i}^{N_C} \left[\sum_{j \neq i}^{N_C} C_{ij} \varepsilon_{ij} + \sum_{j' \neq i}^{N_S} C_{ij'} \varepsilon_{ij'} \right], \tag{7.1}$$

where E_{intra} is the total interaction among the amino acids in the protein and E_{inter} is the binding energy between the protein and the walls of the cage. The indices i and j run over the residues of the protein, while j' runs over the elements of the cage, C is the contact matrix, defined in Eq. 2.2, and ε is the interaction matrix. For ε we use the 20×20 matrix derived with the method of Betancourt and Thirumalai [52] from the matrix determined by Miyazawa and Jernigan [53] on the basis of the observed frequency of contacts between each pair of amino acids. The matrix ε' has some inconsistency in reproducing the hydrophobic and hydrophilic nature of the amino acids because it is not straightforward to estimate the effective number of interaction between water molecules and the residues of a real protein in the native state. Betancourt and Thirumalai proposed to rescale all the values in the matrix with respect to the

interaction with the amino acid Thr in the following way

$$\varepsilon_{ij} = \varepsilon'_{ij} + \varepsilon'_{Thr-Thr} - \varepsilon'_{Thr-i} - \varepsilon'_{Thr-j}.$$

where ε' is the interaction matrix calculated by Miyazawa and Jernigan. The choice of Thr is justified because it gives the best correlation between experimental hydrophobicities and the self-interaction term $\varepsilon_{Thr-Thr}/2$. Although these interaction energies are strictly speaking neither energies nor free energies, they do provide a reasonable representation of the heterogeneity in the interactions between different amino acids.

We note that the chaperone cage is modelled as a rigid object and hence Eq. 7.1 does not include the interactions between the amino acids that form the cage.

7.2.1 Design of the folding and of the cavity coating

To design a lattice protein that will fold into a specific conformation, we use the approach described in Chapter 2. In this approach, we sample sequences for a given conformation, rather than conformations for a given sequence. The basic trial moves are single point mutations. As in the conventional Metropolis scheme, the acceptance of trial moves depends on the ratio of the Boltzmann weights of the new and old states. However, if this were the only criterion, there would be a tendency to generate homo-polymers that have a highly degenerate ground state, rather than a chains that fold selectively into a desired target structure. To ensure the necessary heterogeneity, we impose the following additional acceptance criterion

$$P_{\text{acc}} = \min\left\{1, \left(\frac{N_P^{\text{new}}}{N_P^{\text{old}}}\right)^{T_p}\right\},$$

where T_p is an arbitrary parameter that plays the role of a temperature, and N_P is the number of permutations that are possible for a given set of amino acids. N_P is given by the multinomial expression

$$N_p = \frac{N!}{n_1!n_2!n_3!\ldots} \tag{7.2}$$

where N is the total number of monomers and n_1, n_2 etc are the number of amino acids of type 1,2,.... While sampling the sequence space with a Monte Carlo scheme, we keep the temperature (T_{P}) associated with this quantity high. In doing so we generate an heterogeneous composition of amino acids. The importance of sequence heterogeneity for the design of specific structures is confirmed in our simulation, as it allows us to design hetero-polymer sequence that have a non-degenerate native state. There is another, subtler, meaning of the "temperature" associated with the structural heterogeneity: it also represents the "frustration" imposed on protein design by the fact that a protein lives in the presence of many other molecules to which it should not bind unspecifically. By increasing this "frustration" temperature, we make it less likely that the protein will form an undesired, specific bond to any of the other proteins in the system. During a Monte Carlo run of several million cycles, a large number of distinct sequences are generated. The sequence S^* with the lowest energy is assumed to be the best

candidate to fold into the native state. The energy of a given lattice polymer depends on its conformation.

$$E_{\text{Native}} = \sum C_{ij}\varepsilon^*_{ij}\,. \tag{7.3}$$

In this work we used this scheme to design a lattice heteropolymer to fold to a given target structure. During the design process we do not take in to account any interaction with the cage. In other words we consider only the intra molecular interaction term in Eq. 7.1.

In order to design the interior of a chaperone, we have to mimic the hydrophilic or hydrophobic nature of the cage, while excluding any sequence selectivity. To this end, we employ the approach used in refs. [47, 54] to make a totally structure-less cage wall. To represent a strongly hydrophobic surface, we select the amino acid that has the largest average attraction with all other amino acids. For a hydrophilic surface we select the amino acid with the largest average repulsion. That is, we selected from the interaction matrix the amino acids I_r with the strongest average repulsive interaction and I_a with the strongest attractive one:

$$\begin{aligned} I_r &= \max\left[\frac{1}{20}\sum_{j=1}^{20}\varepsilon_{1j},\ldots,\frac{1}{20}\sum_{j=1}^{20}\varepsilon_{20j}\right] \\ I_a &= \min\left[\frac{1}{20}\sum_{j=1}^{20}\varepsilon_{1j},\ldots,\frac{1}{20}\sum_{j=1}^{20}\varepsilon_{20j}\right], \end{aligned}$$

For the matrix that we used, the amino acid with largest average attractive interaction is Phe with a value of $I_a = -0.23$, while the most repulsive is Arg, $I_r = 0.38$. We also considered a cage with a milder attractive interaction, because, as will be discussed below, a cage made with Phe was so attractive that any protein is irreversibly absorbed on the inner surface of the chaperone. To model a more moderate hydrophobic surface, we used Tyr which has an average attraction strength of $I_a = -0.16$. We stress that, although the parameters of refs. [52, 53] are derived from experimental data, the interaction strengths in our model reproduce the interactions between the amino acids of real proteins only qualitatively. This is no problem for the generic model that we consider, but it would be inadequate for a quantitative description.

7.2.2 Folding

To explore the possible conformations of the lattice polymer, we use four basic Monte-Carlo moves: corner-flip, crankshaft, branch rotation, and translation. The corner-flip involves a rotation of 180 degrees of a given particle about the line joining its neighbors along the chain. The crankshaft move, is a rotation by 90 degrees of two consecutive particles. A branch rotation is a turn, around a randomly chosen pivot particle, of the whole section starting from the pivot particle and going to the end of the chain. The translation is simply a displacement of the center of mass of the protein of one lattice unit in a random direction.

We explore the equilibrium properties of the system by sampling the free energy as a function of two order parameters. The first is the number of native intra-molecular contacts of the protein in a given conformation

$$Q(C) = \sum_{i<j}^{N} C^{(1)}_{ij} C_{ij}\,, \tag{7.4}$$

where $C_{ij}^{(1)}$ is the contact map of the reference structure, and C_{ij} is the contact map of the instantaneous conformation. Only those contacts that belong to the reference structure contribute a value $+1$ to the order parameter. A second order parameter, Q_s, allows us to study the progress of the extrusion process. It is defined as the total number of residues that are still in the cavity.

The quantity that we aim to compute is the free energy F as a function the two order parameters. To compute $F(Q)$, we used the following relation:

$$F(Q) = -T \ln [P(Q)], \tag{7.5}$$

where $F(Q)$ is the free energy of the state with order parameter Q and $P(Q)$ is the histogram that measures the frequency of occurrence of conformations with order parameter Q. In practice, a direct (brute force) calculation of this histogram is not efficient, as the system tends to be trapped in local minima, especially at low temperatures. To solve this problem, we used the Virtual-move Parallel-tempering (VMPT) scheme that is described in Chapter 3.

7.3 Results

To study the competition between the extrusion and the folding process, we consider a protein (P-64) of 64 residues trapped inside a cubic cage of size 5x5x5 in lattice unit length with a hole in the shape of a 3 by 3 cross in one of the faces (Fig. 7.2). The ratio between the accessible volume and the volume of the protein for our model is around 1.95, which would correspond to ratio between the volume of globular protein with a radius of gyration of ~ 27Å, and the volume of the closed chamber in the GroEL/GroES complex (~ 170000Å^3).

Following the scheme in the methods section, we designed the protein to fold, in the absence of a confining cage, into the target structure shown in Fig. 7.2.a. In Fig. 7.3 we show the folding free energy of the unconfined P-64 (red curve) at $T = T_F/2$ where T_F is the temperature at which there is equilibrium between the unfolded and the native states. The free-energy profile illustrates that the protein has a strong tendency to fold into the target structure which is characterized by a value of the order parameter $Q = 81$.

Let us now consider the successive steps in the chaperonin-assisted protein refolded, as represented schematically in Fig. 7.6. In the initial conformation (1) the chaperonin is open and exposes a hydrophobic rim that should attract misfolded proteins. We model this conformation of the chaperonin with an open cubic box of size 2x5x5, approximately half of the closed one, surrounded by a repulsive outer layer 3x7x7 to avoid binding on the outside of the rim (lower section of Fig. 7.2.a). The attractive internal lateral surface was represented by the strongly attractive amino acid which is also a strongly hydrophobic Phe ($I_a = -0.23$) whilst a repulsive back surface was made of Arg ($I_r = 0.38$) which is strongly hydrophilic. In this way we mimic the rim structure of the real chaperonin.

Our simulations show that this barrel cannot trap the protein in its native state, but it can trap misfolded proteins. This can be seen in Fig. 7.3 where we plot the free energy surface as function of the number of native contacts Q (Eq. 7.4). The free energy curve shows that proteins in the native state ($Q = 81$) outside the cage have a lower free energy than inside,

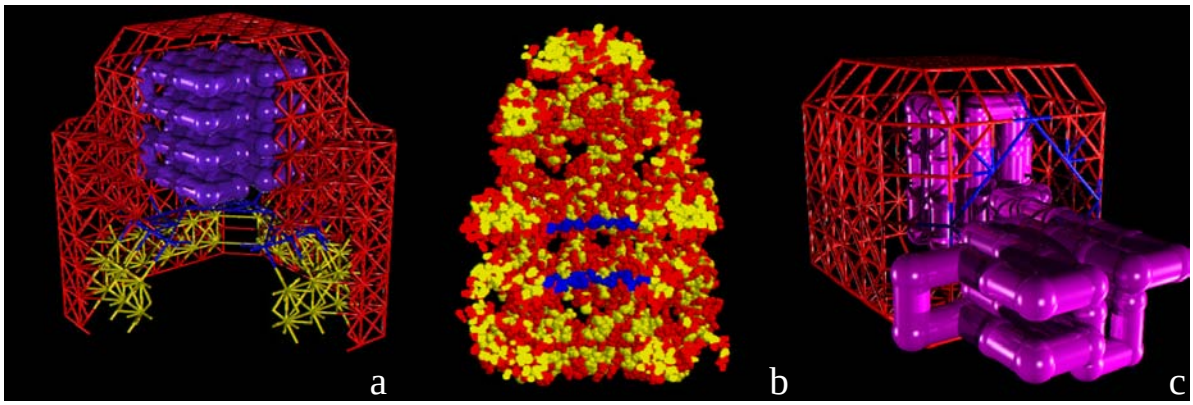


Figure 7.2: Lattice model for GroEL-GroES complex. (a) The closed GroEL-GroES compartment is modelled as a cage of 5x5x5 lattice units (red). It is connected by a hole (3x3 cross – blue) to the open barrel, modelled as a 2x5x5 box surrounded by a repulsive outer layer 3x7x7 to avoid binding on the outside of the rim. The attractive internal lateral surface was represented by the strongly attractive and hydrophobic amino acid Phe (average pair interaction API with the other amino acids $API = -0.23kT$ Supplementary methods Eq.6) whilst a repulsive and hydrophilic back surface was made of Arg ($API = 0.38kT$). The ratio between the volume of the cage and that of the 64-residue protein (purple) is around 1.95, which typical for the values found in experiments. (b) Space-filling representation of the X-Ray structure of the GroEL/GroES/ADP complex [46]. Colors represent the type of surface: all hydrophobic amino acids (A,V,L,I,M,F,P,Y) are in *yellow*, while the polar one (S,T,H,C,N,Q,K,R,D,E) are *red*. (c) Intermediate conformation during the extrusion process from the hydrophilic cage. If the inner surface of the closed GroEL-GroES compartment was made of the mildly hydrophobic Tyr ($API = -0.16kT$), extrusion did not take place.

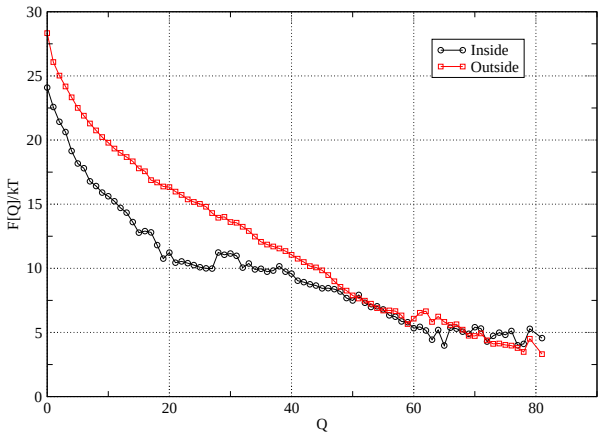


Figure 7.3: Plots of the free energy $F(Q)$ at $T = T_F/2$, as a function of the number of native contacts Q (Eq. 7.4) for the conformation that have at least one amino acid inside the open cage (black circles), and for the one free in solution (red squares). The open cage is covered with the most attractive amino acids Phe with $API = -0.23kT$ in the rim area, while for the hydrophilic back surface we used Arg. For chain conformation close to the native state ($Q > 45$) there is not free energy gain in the trapping indicating that those states can easily diffuse away. However for the misfolded states there is a strong preference (up to $5kT$) to bind to the rim.

whilst misfolded conformations (small values of Q) prefer to stay inside and bind to the hydrophobic rim (up to $5kT$ binding free energy). This is further demonstrated by the free energy function of the total number of contacts N_c and of number of amino acids inside the cavity Q_s for states with no native contacts Fig. 7.5.a, this shows a strong preference for trapped compact conformations. A similar behavior is observed also for states with a different number of contacts over native contacts ratio N_c/Q (Fig. 7.5.b).We stress that our results should not depend on the particular sequence of the target protein, because we use structureless cage walls.

After a misfolded protein is captured, the GroEL/GroES complex closes (i.e. the barrel gets capped) and we move to the actual translocation process (steps 2 and 3 in Fig. 7.6). The extrusion takes place through the hole in bottom of the hydrophilic cage (see Fig. 7.2.a). To model a hydrophilic cage, the walls of the cavity where coated with Arg, which in our interaction matrix is, on average, a strongly repulsive as well as strongly hydrophilic amino acid. In Fig. 7.7 we plot the free energy as function of the usual order parameters. The figure shows that the lowest free-energy state corresponds to values of the order parameters $Q_s = 0$ and $Q = 81$, which demonstrate that the most favorable conformations is the chain folded outside of the cage. Of course, the early stages of extrusion cost free energy, as the protein must unfold, at least locally, to initiate the extrusion. Interestingly, the free-energy barrier for extrusion is considerably larger for the native state of P-64 ($\sim 10K_BT$) than for a partially folded state ($\sim 4K_BT$). This implies that the hydrophilic cage will preferentially expel non-

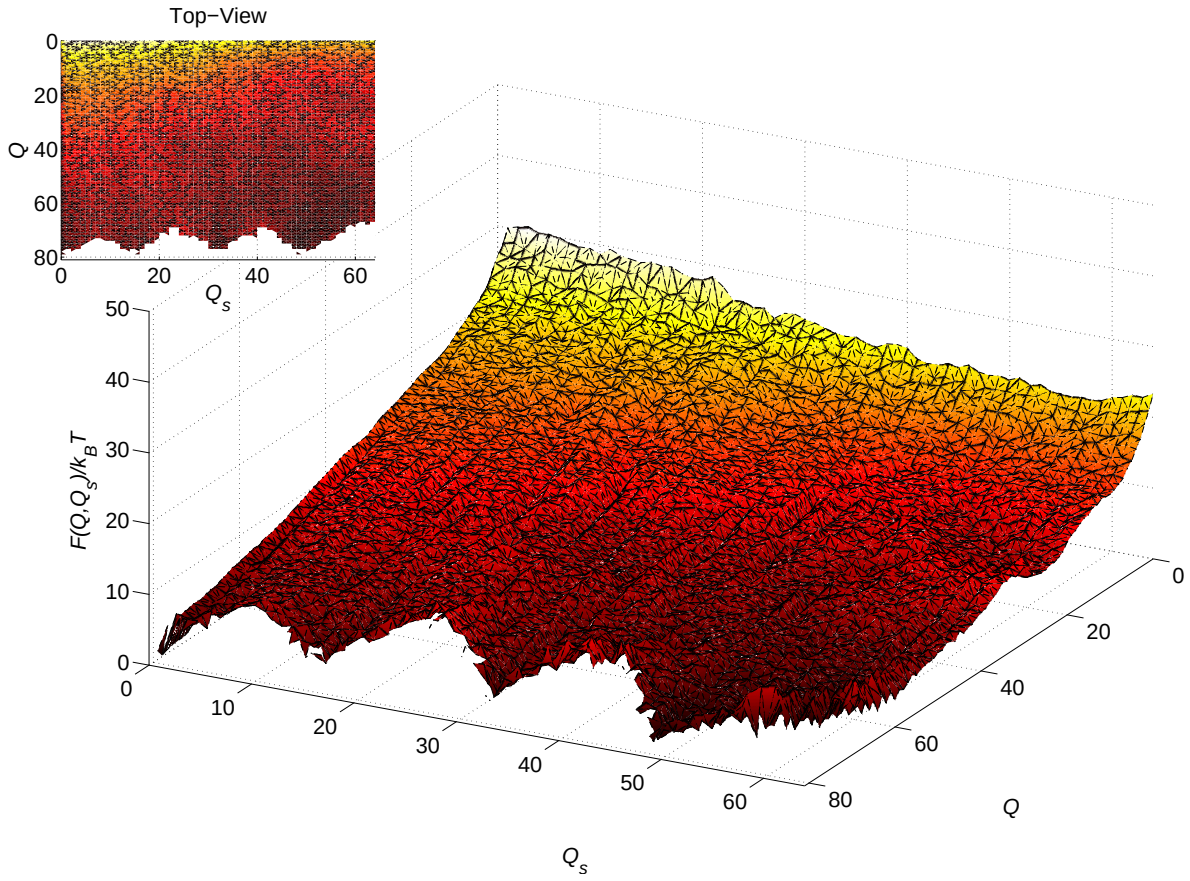
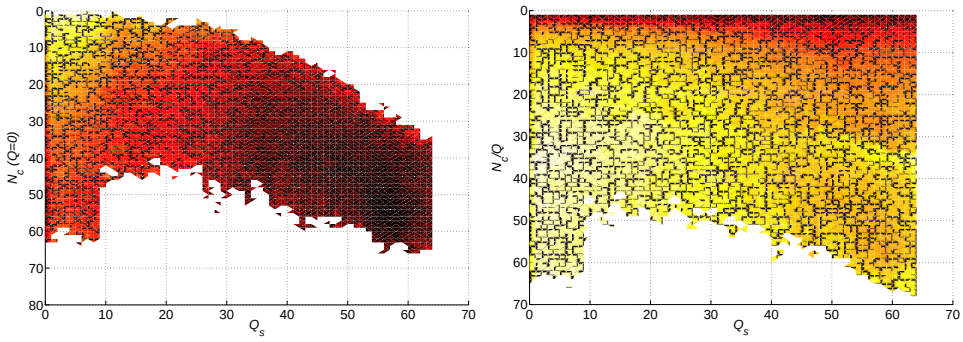


Figure 7.4: Plots of the free energy landscape $F(Q, Q_s)$ at $T = T_F/2$, as a function of the number of native contacts Q (Eq. 7.4) and of the number of residues inside the cage Q_s. The open cage is covered with the most attractive amino acids Phe with an average contact strength of -0.23 in the rim area, while for the hydrophilic back surface we used Arg. The states with the lowest free energy correspond to conformation of the chain mainly outside the cage (high values of the order parameter Q_s) and in the native state (high values of Q). However the lower states for non native conformation are inside the cage space. This indicate a binding selectivity for misfolded conformations.



(a) (b)

Figure 7.5: Plots of the free energy $F(N_c, Q_s)$ of the trapping in the open cage state as function of the number of contacts $N_c$and the number of residues inside the cage Q_s computed for states with no native contacts ($Q = 0$)(a). This free energy shows that open cage has a strong attraction for highly misfolded states, and this is because energetically they con open and bind strongly to the hydrophobic rim. (b) Free energy landscape $F(N_c/Q, Q_s)$ computed in the same conditions as before but now function of the ratio between $N_c$and the number of native contacts $Q$for $Q > 0$. The states shown here represents conformations that have common bonds with the native state ($N_c/Q = 1$). Even for this intermediate conformation the open cage is attractive.

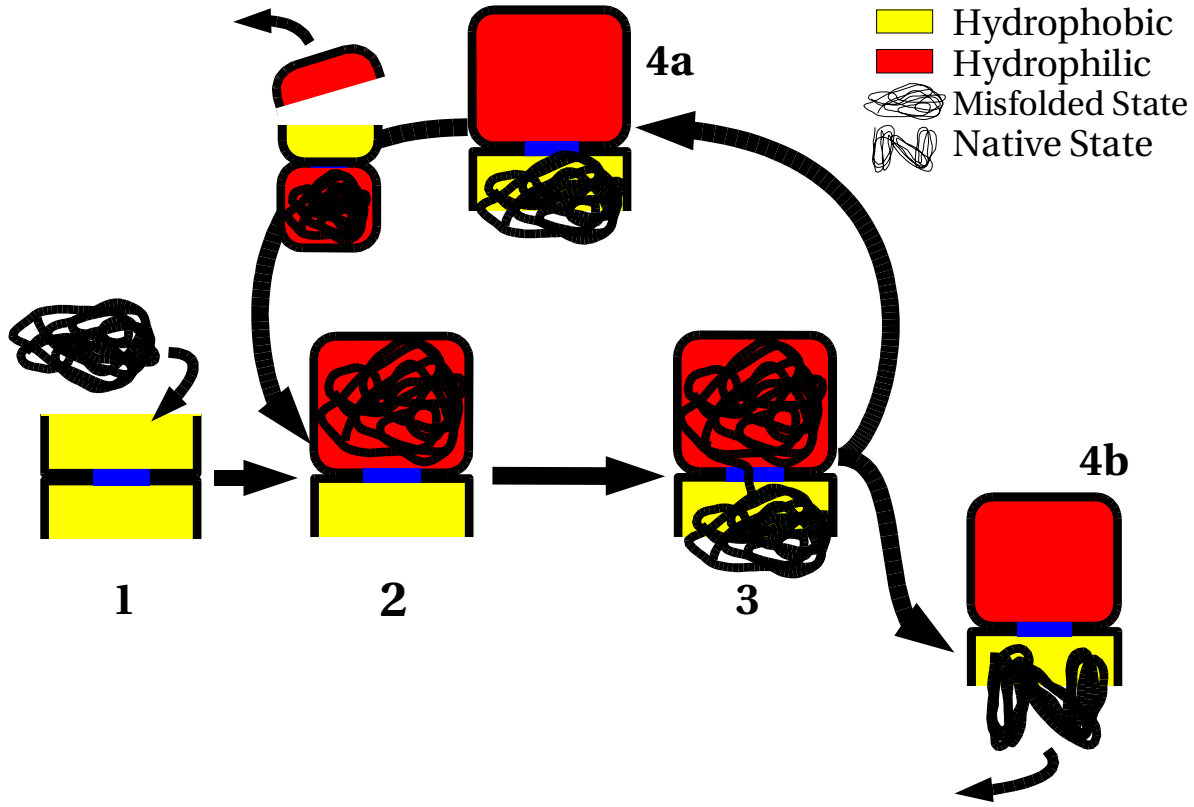


Figure 7.6: Refolding process of the double-cage chaperonin. The First step is the trapping of misfolded proteins in the open cage. The second step is encapsulation upon binding of the GroES cap, and change of the internal wall from hydrophobic (red) to hydrophilic (blue) (Fig. 7.2.a). At this point the extrusion process starts (7.2.b) with two possible outcomes:a) the protein does not fold in the native state, b) it does. If the case is a) then the protein goes through another round of extrusion.

native conformations. Analyzing the structure of the (partially) extruded protein we found that the chain is initially stretched across the hole but, as soon as a sufficient number of amino acids are outside the cage, they fold again into a compact structure. The free-energy gain due to refolding facilitates the extrusion process. During the extrusion process, the existing misfolded structure is completely broken up and a new compact conformation is formed outside.

We stress that in the in our simulations, translocation was always much faster than refolding of the proteins inside the chaperonin cage. Although folded state where observed inside the cage (Fig. 7.7), on the timescale of a translocation event, we never saw complete refolding inside the cage, whereas complete refolding was always observed as the end result of translocation. This is illustrated by Fig. 7.8 where we compare the rate of intra-cavity refolding with the rate of inter-cavity translocation. In a different context, it is well known that extrusion may speed up protein folding [51].

In order to test whether repulsion inside the chaperonin cage is important for the extrusion process, we repeated the previous calculations with a different (strongly hydrophobic) Phe coating of the internal walls of the cavity. The plot in Fig. 7.9 shows the free energy profile for extrusion from such a hydrophobic cage. As can be seen for the figure, the driving force for extrusion has now been reversed: rather than expelling the protein, the cage sucks it in. The attraction is strong enough to cause partial absorption and unfolding of the chain inside the cage. This is not the case for a moderately hydrophobic surface (modeled by Tyr – average interaction energy $I_a = -0.16$) where the native state is not disrupted by the absorption on the inner walls (see Fig. 7.10). However, the attraction is strong enough to inhibit the translocation process. Hence, for translocation to occur, the protein should initially be confined in a hydrophilic cavity. This offers a rationale for the strong hydrophilic nature of the closed cavity. When the protein refolds into the open barrel of the chaperonin complex (Fig. 7.6-4), it need not end up in its native state. However, the surface of the open barrel tends to trap misfolded proteins. In this way the refolding cycle can start again, with the capping of the second cavity and the opening of the first. This refolding scenario has one attractive feature: it offers a natural explanation of the double-barrel structure of the chaperonin, as it makes it plausible that misfolded proteins are shuttled forwards and backwards between the barrels until they reach the native state that can escape from the hydrophobic rim of the open barrel. Moreover, as the barrier for translocation is higher for native states than for misfolded states, native states that happen to be trapped in the GroEL/GroES complex stand a good chance of surviving until the barrel opens again. Of course, the scenario that our simulations suggest is, at present only a hypothesis. However, it should be testable. First of all, it should be possible to verify that proteins can move through the equatorial plane of the GroEL barrel. Moreover, we should expect that the translocation process will be very sensitive to the nature of the disordered protein segments near the hole in the equatorial plane. Any chemical modification that would block the hole should decrease the chaperonin activity of the GroEL/GroES complex.

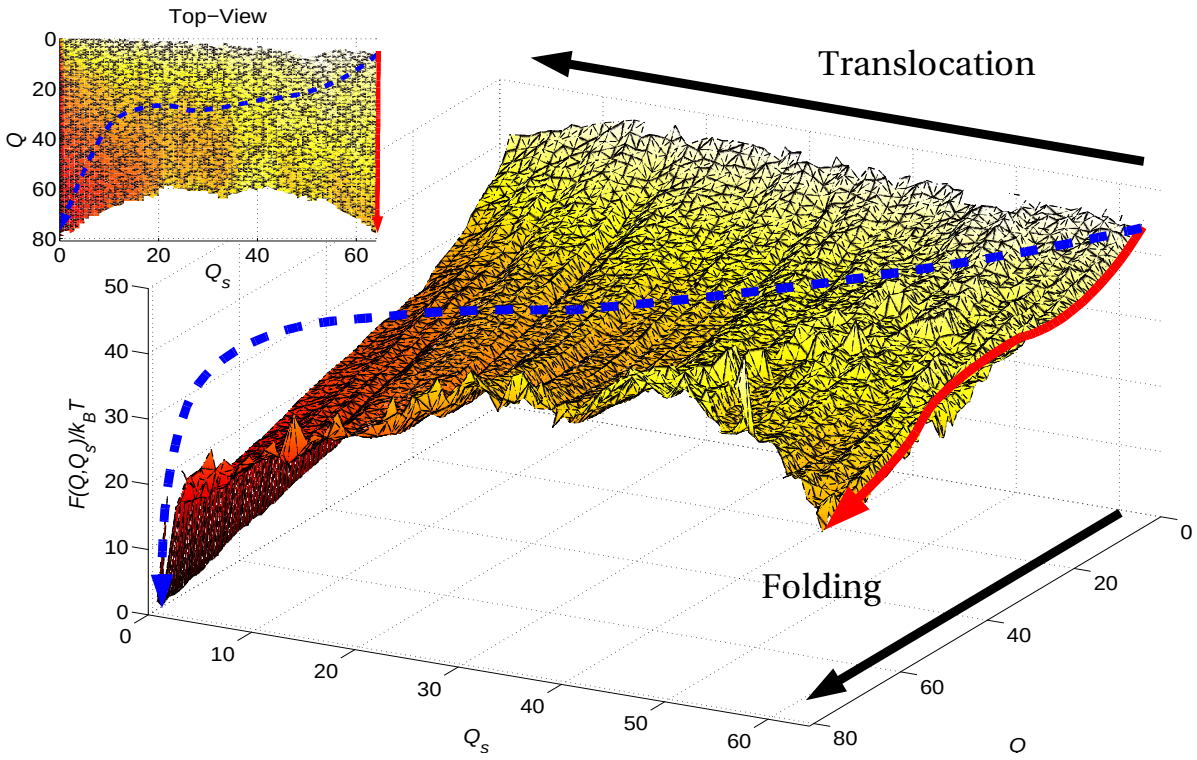


Figure 7.7: Plots of the free energy landscape $F(Q,Q_s)$ as a function of the number of native contacts Q and of the number of residues inside the cage Q_s, computed at $T = T_F/2$, where T_F is the temperature at which there is equilibrium between the unfolded and the native states. The states with the lowest free energy correspond to conformation of the chain folded outside the cage ($Q_s = 0$ and $Q = 81$), demonstrating a preference for the extrusion plus refolding process. The colored lines represents trajectories for the refolding plus translocation (*dashed blue*) and for the intra-cage refolding (*red continuous*). Although both possible, the first one is much faster than the second. The cage is covered with the most repulsive amino acids Arg with an average repulsive strength per contact of 0.38. The non-sampled region correspond to conformations of the chain too compact to exist across the hole.

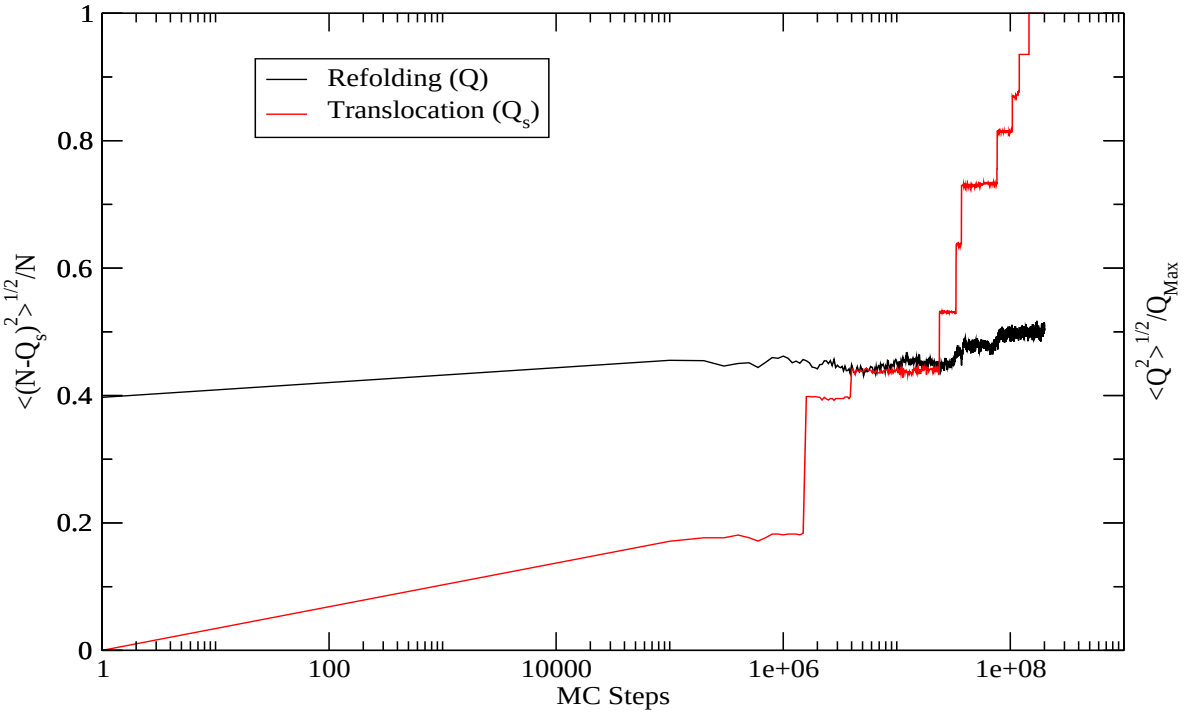


Figure 7.8: Comparison of the relative rate of intra-cavity refolding and inter-cavity translocation. The degree of translocation is measured by the fraction of the protein that has escaped form the original cage. In our simulations, we compute the root mean-square of this fraction ($\sqrt{\left\langle (N-Q_s)^2 \right\rangle}/N$) over many trajectories. Similarly, we quantify the degree of refolding by root mean-square $\sqrt{\langle Q^2 \rangle}/Q_{\mathrm{Max}}$. Here Q_s is the number of residues inside the cage, Q is number of native contacts, and Q_{Max} is the number of contacts in the native state. Time is measured in terms of the number of MC steps. We chose the reduced temperature of these simulations slightly higher than in the remainder of this work ($T = T_F/2$ rather than $T = T_F/1.5$). Because otherwise even translocation would be too slow to be observed on the timescale of the simulations. The figure shows that complete translocation occurs on a timescale where intra-cage refolding is still imperceptible.

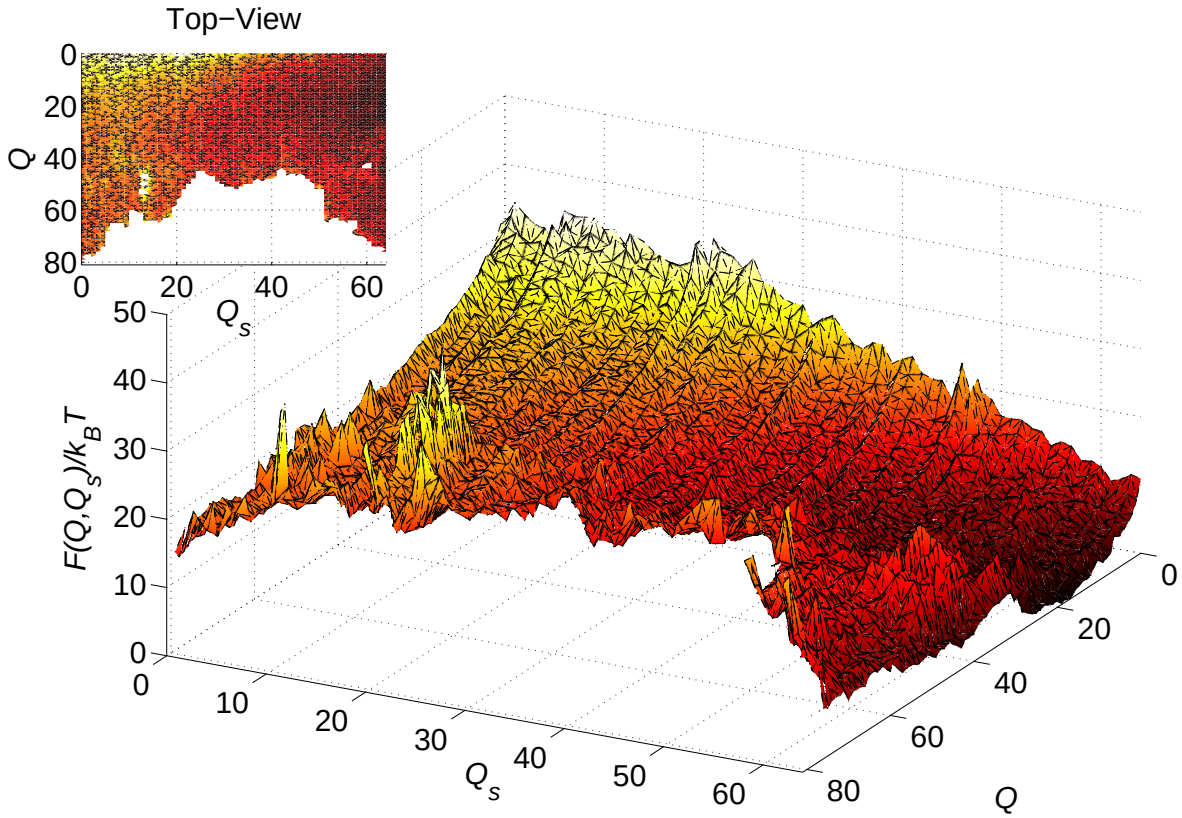


Figure 7.9: Plots of the free energy landscape $F(Q,Q_s)$ at $T = T_F/2$, as a function of the number of native contacts Q (Eq. 7.4) and of the number of residues inside the cage Q_s. The cage is covered with the most attractive amino acids Phe with an average repulsive strength per contact of -0.23. The states with the lowest free energy correspond to conformation of the chain mainly inside the cage (high values of the order parameter Q_s) and not in the native state (low values of Q). This indicates that the extrusion process has been reversed, and that the attraction is strong enough to absorb the protein on the walls of the cavity.

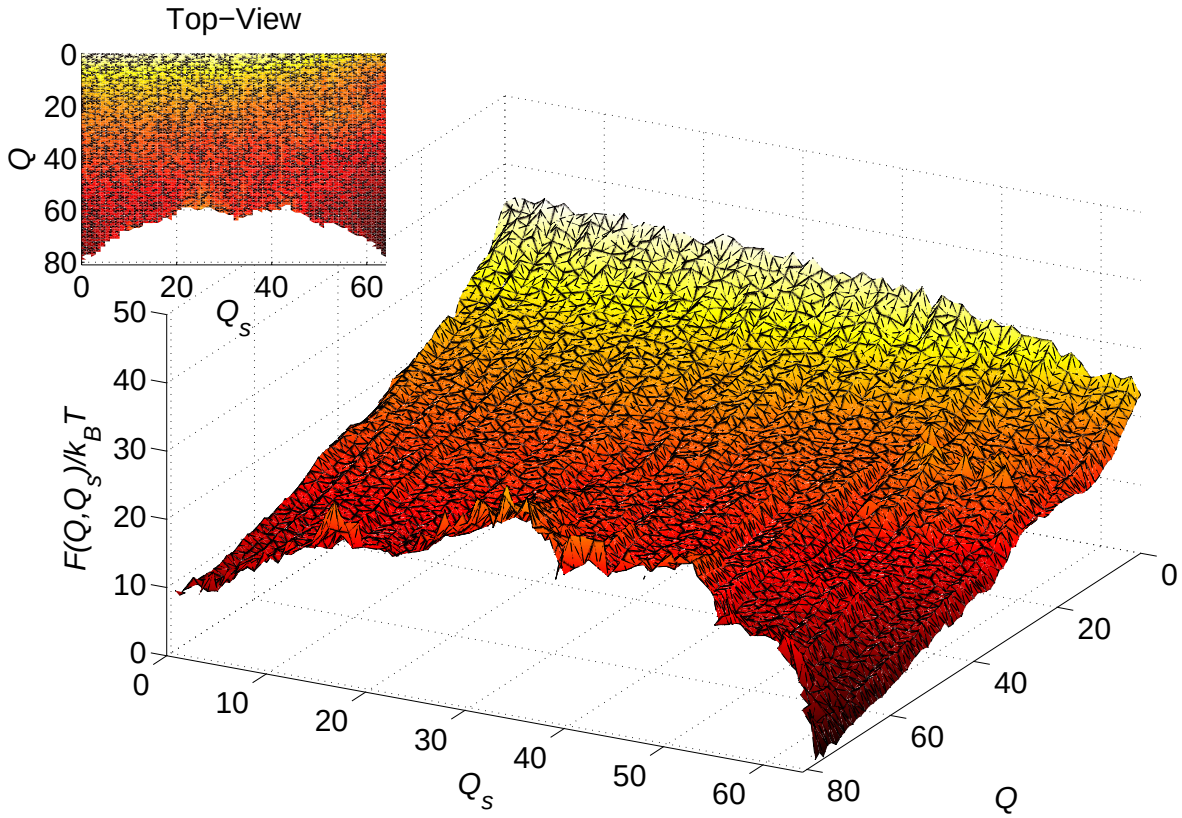


Figure 7.10: Plots of the free energy landscape $F(Q,Q_s)$ at $T = T_F/2$, as a function of the number of native contacts Q (Eq. 7.4) and of the number of residues inside the cage Q_s. The cage is covered with the most attractive amino acids Tyr with an average contact strength of -0.16. The states with the lowest free energy correspond to conformation of the chain mainly inside the cage (high values of the order parameter Q_s) and in the native state (high values of Q). This indicates that the extrusion process has been reversed, but the attraction is not strong enough to cause unfolding of the protein as in the case of the strong attractive surface.

8 Summary

> "....I have invariably found that our knowledge, imperfect though it be, of variation under domestication, afforded the best and safest clue. I may venture to express my conviction of the high value of such studies, although they have been very commonly neglected by naturalists."
> *ON THE ORIGIN OF SPECIES by Charles Darwin*

In this thesis we used numerical simulations to study the relation between different functionalities of a protein that are determined by the amino-acid sequence. In particular, we considered conformational changes, protein-protein binding, and chaperone-aided protein folding. The common link that connects these apparently different processes, is that they all involve an external perturbation of the protein. The order in which we presented the results of our research represents the evolution from a local perturbation to the global refolding of the protein.

In the introduction and in chapter 2 we motivated the use of a lattice-heteropolymer model to represent proteins. In particular, we argued that such a simple description provides a powerful theoretical tool to understand the physics of intra and intermolecular interactions of proteins and the effect they have on binding and folding. We reviewed the mean-field arguments that are at the basis of our numerical protein-design strategy. In particular, we discussed how foldable sequences can be selected by performing simulations at a design temperature that is below the typical glass transition of the corresponding random heteropolymer.

In chapter 3 we described the numerical techniques used to compute the free energy landscape of the lattice proteins that we designed. The free energy was computed as a function of several order parameters characteristic of the native state and of the process under study. In practice, the free energy landscape of such system if often highly corrugated. In order to sample states separated by high free energy barriers, we devised a new algorithm (Virtual-move Parallel Tempering) that greatly increased the sampling efficiency.

In chapter 4 we used our design strategy to study conformational changes in a single protein. The mechanism by which driven conformational transitions in proteins take place is still a matter of debate. The simplest picture stresses the analogy between molecular and macroscopic machines. In this picture, the biological molecules are supposed to consist of rigid sub-units that are connected to each other through springs and notches. The folded state is then a *stressed* state where some of the springs are compressed or extended. A conformational change starts when the external agent disrupts some attractive interactions that were essential for the stability of the original conformation. The energy stored in the springs is then released. One of the problems with this picture is that a sudden release of accumulated stress in a chain molecule will, in general, be irreversible. This irreversibility lowers the efficiency

of the molecular motor. In our model, instead, the transition is induced by an alteration of the chemical properties of few amino acids. By natural or artificial (design) evolution it is possible to control the free energy difference between the native state and a given secondary target conformation. In the presence of specific external perturbations the free energy of this secondary minimum can be lowered and become the most stable state. The system will then naturally diffuse toward this new equilibrium conformation. The role of thermal fluctuations becomes important because they allow the system to perform the transition from one structure to the other one at a considerably lower energetic cost. A similar process, repeated cyclically, could be responsible for generating the motion of motor proteins such as Myosin and Dynein.

The second step in our research was to characterize the role of a substrate on the refolding of a protein. When designing protein-substrate interactions, we must consider two challenges. The first concerns the design of a specific binding site on the surface of a protein that can attach to the substrate. The second challenge is to design the response (e.g. refolding) of the protein to the binding. Before going on to study how the binding could induce a conformational change, we focused on the problem of specific binding. The question can be briefly summarized as follows: "How can proteins bind strongly to a small number of substrates, yet weakly - if at all - to the large number of other biomolecules in the cell?". The cellular environment is highly crowded with thousands of different proteins species. The chances of an occasional encounter are therefore very high. In chapter 5 we addressed this question by looking at the influence of the binding on the folding free energy of different proteins and substrates. The results of the simulations showed that occasional non-selective interactions do indeed occur between the protein and the substrate, but the entropy cost was too high to create stable binding, even at a temperature equal to a third of the folding temperature. This is not the case for designed binding regions: these are able to bind strongly to their partner. Our explanation is that there is no fundamental difference between intra- and inter-molecular interactions, hence once the protein is in contact with the substrate they can be viewed as a single protein. This is the reason why he same algorithm that was used to select a sequence that would fold into a specific structure could be employed to design specific binding between protein and substrate.

In chapter 6 we addressed the question how a substrate could influence the conformation of a protein. To this end, we computed the free energy of the binding process, and we monitored the change in the protein conformation for different sizes of the binding region. In all cases, the structure of the free-energy landscape suggested that the first step of the binding process involved a local unfolding of the initial compact conformation to increase the contact with the substrate. Subsequently, the chain molecule would rearrange locally to reach the native structure corresponding to the bound state. The elements that we have collected so far demonstrate the ability of model proteins to partially refold in new conformation, to specifically bind to few partners, and refold upon binding. Next, we focused on the role of random domains on protein-protein interactions. Many proteins contain parts that are disordered and highly mobile. Experimentally, it is difficult to investigate these random domains as they are effectively invisible in X-ray crystallography. However, the fact that a domain has a random conformation does not necessarily mean that it has no function. To illustrate this statement, we designed a lattice-protein that would not fold in solution. We then introduced a substrate that would preferentially bind the protein in a specific conformation. The simulations showed

that the protein folds upon binding. This process is reversible, in the sense that the protein reverts to its disordered state upon unbinding. The next question is whether our model allows us to design a substrate that would help the protein to find the native state even after unbinding. In practice, such substrates act as folding *catalysts*. They provide a binding region that, after a successful folding event, would be freed to receive a new target. Such refolding machines exist in nature and are known as *chaperons*. There are different classes of such complexes. Generally, chaperons are believed to work as unfolding-refolding machines: in doing so, they lower the barrier that the protein must overcome in order to escape from a local free-energy minimum to reach the native state.

In chapter 7, where we focused on refolding properties of a fascinating subclass of chaperons, the double-barrelled chaperons or *chaperonins*. Such large protein complexes refold non-native proteins. Initially, the target proteins are confined inside one barrel of the chaperonin complex. What happens inside the cage is still a matter of debate. Previous models described the chaperonin as a cage with a special internal surface that would mildly bind a wide variety of protein independently of their conformation. Computer simulations showed that such a chaperonin could indeed assist the folding process, but only in a small range of the values of the attraction strength between the protein and the walls of the cage. Although, such a model provides a possible explanation for the enhanced folding rate, it does not explain the double-barrelled conformation of the chaperonin, nor its asynchronous activity. In our simulations, we considered an alternative mechanism that did not rely on the partial unfolding of the protein due to the interactions with the internal walls. Rather, in our simulations, we considered the possibility of translocation of the protein through an hole that connects the two barrels of the chaperonin complex. This mechanism turn out to be very efficient and completely independent of the particular sequence or initial structure of the misfolded state.

We stress that all models studied in this thesis are highly simplified. However, with more realistic models it would not (yet) be feasible to perform a systematic study of the phenomena that we describe. Of course, it is essential to link the simulations to experiments. In the case of the chaperonin complex (Chapter 7) this is fairly straightforward, as our model makes specific and testable predictions about the mode of operation of the chaperonin. But also for the topic discussed in the other chapters, it should be possible to make a link with experiments. This will, however, require the use of models that are designed to describe a particular process (e.g. a specific substrate-induced refolding process). We hope that this thesis will contribute to such studies.

Bibliography

[1] *http://ghr.nlm.nih.gov/ghr/picture/actin*.

[2] *http://www.biochemj.org/bj/381/0001/bj3810001f01.htm*.

[3] S. Miyazawa and R. Jernigan. *Macromolecules*, 18:534 table VI, 1985.

[4] G. Tiana. *Ph. D. Thesis (http://merlino.mi.infn.it/guido/lxmi/pubs.php)*.

[5] E.I. Shakhnovich and A.M. Gutin. *Proc. Nat. Acc. Sci.*, 90:7195, 1993.

[6] B.Derrida. *Phys. Rev. B*, 24:2613, 1981.

[7] E.I. Shakhnovich and A.M. Gutin. *Protein Engin.*, 6:793, 1993.

[8] E.I. Shakhnovich. *Phys. Rev. Lett.*, 72:3907, 1993.

[9] E.I. Shakhnovich and A.M. Gutin. *Nature*, 346:773, 1990.

[10] V. S. Pande el al. *J. Chem. Phys.*, 103:9482, 1995.

[11] V. S. Pande el al. *Biophysical Journal*, 73:3192, 1997.

[12] V. S. Pande el al. *Rev. Mod. Phys.*, 72:259, 2000.

[13] E. N. Govorun el al. *Phys. Rev. E.*, 64:040903, 2001.

[14] M-T. Kechadi et al. *Nuovo Cimento*, 20(12bis):2383, 1998.

[15] Bryngelson and Wolynes. *Proc. Nat. Acc. Sci.*, 84:7524, 1987.

[16] E. I. Shakhnovich and A. M. Gutin. *Biophys. Chem.*, 34:187, 1989.

[17] R. Faller el al. *J. Chem. Phys.*, 13:5419, 2002.

[18] D. Frenkel. *Proc. Nat. Acc. Sci.*, 101:17571, 2004.

[19] G. Boulougouris and D. Frenkel. *unpublished*.

[20] N. Metropolis et al. *J. Chem. Phys*, 21:1087, 1953.

[21] D. Frenkel and B. Smit. Understanding molecular simulations. *Accademic Press*, page 389, 2002.

[22] A.P. Lyubartsev el al. *J. Chem. Phys.*, 96:1776, 1992.

[23] E. Marinari and G. Parisi. *Europhys. Lett.*, 19:451, 1992.

[24] R. H. Swendsen and J.S. Wang. *Phys. Rev. Lett.*, 57:2607, 1986.

[25] C.J. Geyer. *Proc. 23rd Symp. on the Interface and Interface Foundation*, Fairfax Viginia p.156, 1991.

[26] K. Hukushima and K. Nemoto. *J. Phys. Soc. Jpn*, 65:1604, 1996.

[27] M.C. Tesi el al. *J. Stat. Phys.*, 82:155, 1996.

[28] G.M. Torrie and J.P. Valleau. *J. Comp. Phys.*, 23:187, 1977.

[29] B.A. Berg and T. Neuhaus. *Phys. Rev. Lett.*, 68:9, 1992.

[30] Y. Sugita and Y. Okamoto. *Chem. Phys. Lett.*, 329:261, 2000.

[31] J.P. Wilding. *Phys Rev E*, 52:602, 1995.

[32] N. Tsunekawa el al. *J. Chem. Phys.*, 116:6725, 2002.

[33] I. Coluzza el al. *Phys Rev E*, 68 (046703), 2003.

[34] C.Y. Lin el al. *Proteins-Str. Fun. Gen.*, 52:436, 2003.

[35] A. Schug and W. Wenzel. *Europhys. Lett.*, 67:307, 2004.

[36] G. D. Rose and T. P. Creamer. *Proteins: Struct, Func. Gen.*, 19:1, 1994.

[37] S. Dalal et al. *Nature Struct. Biol.*, 4:548, 1997.

[38] W. R. Schief and J. Howard. *Cur. Opin. Cell. Biol.*, 13:19, 2001.

[39] L.Borovinskiy and A.Yu.Grosberg. *J. Comp. Phys.*, 118:5201, 2003.

[40] F. Jülicher el al. *Rev. Mod. Phys.*, 69:1269, 1997.

[41] V.S. Pande el al. *Proc. Nat. Acc. Sci.*, 91:12976, 1994.

[42] A. M. Gutin and E. I. Shakhnovich. *J. Chem. Phys*, 98:8174, 1993.

[43] I.Coluzza and D. Frenkel. *in preparation*.

[44] J. C. Nam el al. *Science*, 301:1884, 2003.

[45] R. Hawkins and T. McLeish. *Phys. Rev. Lett.*, 93:098104, 2004.

[46] P. Sigler el al. *Ann. Rev. of Biochem.*, 67:581, 1998.

[47] A. Jewett el al. *Proc. Nat. Acc. Sci.*, 101:13192, 2004.

[48] P. Thiyagarajan el al. *Structure.*, 4:79, 1996.

[49] G. Rabut and J. Ellenberg. *Current Biology*, 11:R551, 2001.

[50] H. Kim and K. J.Shin. *Phys. Rev. Lett.*, 82:1578, 1998.

[51] M. Morrissey el al. *Polymer.*, 45:557, 2004.

[52] M. Betancourt and D. Thirumalai. *Protein Science*, 8:361, 1999.

[53] S. Miyazawa and R. Jernigan. *Macromolecules*, 18:534, 1985.

[54] F. Takagi et al. *Proc. Nat. Acc. Sci.*, 100:11367, 2003.

9 Samevatting

In dit proefschrift worden numerieke simulaties gebruikt voor het bestuderen van de relatie tussen verschillende functies van een eiwit die worden bepaald door de aminozuursequentie. Hierbij is in het bijzonder gekeken naar conformatieveranderingen, eiwit-eiwitbinding en chaperonne-afhankelijke eiwitvouwing. Wat deze ogenschijnlijk verschillende processen gemeen hebben is een externe verstoring van het eiwit. De volgorde waarin de resultaten van dit onderzoek gepresenteerd worden vertegenwoordigt de evolutie van een lokale verstoring tot de globale hervouwing van het eiwit.

In de introductie en in hoofdstuk 2 beargumenteerden wij het rooster-heteropolymeermodel voor de representatie van eiwitten. Het belangrijkste argument was dat een dergelijke simpele beschrijving een krachtig theoretisch hulpmiddel levert voor het begrip van de fysica achter intra- en intermoleculaire interacties van eiwitten en voor het begrip van het effect van deze interacties op vouwing en het vormen van verbindingen. Verder is een overzicht gegeven van de gemiddeld-veldargumenten die ten grondslag liggen aan onze numerieke eiwitontwerpstrategie. Er wordt beschreven hoe "vouwbare" sequenties kunnen worden geselecteerd door het uitvoeren van simulaties op een ontwerptemperatuur ónder de typische glasovergang van de overeenkomstige willekeurige heteropolymeer.

In hoofdstuk 3 zijn de numerieke technieken beschreven die zijn gebruikt om het vrije-energielandschap van het ontworpen roostereiwit te berekenen. De vrije energie is berekend als functie van verschillende ordeparameters, karakteristiek voor de gevouwen toestand en voor het bestudeerde proces. Het is gebleken dat het vrije-energielandschap van zo'n systeem vaak grote rimpels laat zien. Om toestanden te samplen die door hoge vrije-energiebarrières van elkaar gescheiden zijn, hebben wij een nieuw algoritme ontwikkeld (Virtual-move Parallel Tempering). Dit algoritme verbetert de efficiëntie van de sampling aanzienlijk.

In hoofdstuk 4 is onze ontwerpstrategie gebruikt om conformatieveranderingen in éen enkel eiwit te bestuderen. Hoe deze conformatieveranderingen plaatsvinden is nog steeds onderwerp van discussie. In het meest eenvoudige model wordt de analogie tussen moleculaire en macroscopische machines benadrukt. In dit model wordt verondersteld dat de biologische moleculen uit stijve subunits bestaan die met elkaar verbonden zijn door veren en scharnieren. De gevouwen toestand is dan een *gespannen* toestand waarin sommige veren ingedrukt of uitgerekt zijn. Een conformatieverandering begint als een externe factor interacties verstoort die essentieel zijn voor het behoud van de originele conformatie. De energie die is opgeslagen in de veren wordt dan vrijgelaten. Een van de problemen met dit model is dat een plotseling vrijlaten van opgebouwde spanning in een moleculaire keten over het algemeen ook onomkeerbaar zal zijn. Deze onomkeerbaarheid verlaagt de efficiëntie van de moleculaire motor. In ons model, daarentegen, is de overgang geïnduceerd door een verandering van de chemische

eigenschappen van enkele aminozuren. Door natuurlijke of kunstmatige (ontwerp-) evolutie is het mogelijk het vrije-energieverschil tussen de natieve (gevouwen) toestand en een gegeven tweede (doel-) toestand te bepalen. Bij bepaalde externe verstoringen kan de vrije energie van dit tweede minimum verlaagd worden en zodoende de meest stabiele toestand worden. Het systeem zal zich dan automatisch bewegen naar deze nieuwe evenwichtsconformatie. De rol van thermische fluctuaties wordt belangrijk omdat ze het systeem met een aanzienlijk lager energieverlies van de ene naar de andere structuur laten overgaan. Een dergelijk proces, herhaaldelijk uitgevoerd, zou verantwoordelijk kunnen zijn voor de beweging van motoreiwitten als myosine en dyneïne.

De tweede stap in ons onderzoek was het karakteriseren van de rol van een substraat op het hervouwen van een eiwit. In het ontwerpen van eiwit-substraatinteracties zijn twee uitdagingen. De eerste behelst het ontwerp van de specifieke bindingsplek op het oppervlak van een eiwit dat aan het substraat kan binden. De tweede uitdaging is het ontwerp van hoe het eiwit reageert op binding aan het substraat (bijvoorbeeld hervouwing). Alvorens te bestuderen hoe de binding een conformatieverandering kan veroorzaken, hebben we ons gericht op het probleem van specifieke binding. De vraag kan als volgt worden verwoord: “Hoe kunnen eiwitten sterk aan slechts een paar substraten binden, maar zwak - of helemaal niet - aan een groot aantal andere biomoleculen in een cel.“ De omgeving in een cel is vergeven van duizenden verschillende soorten eiwitten. De kans op een toevallige ontmoeting is daarom erg hoog. In hoofdstuk 5 hebben we deze vraag behandeld door te kijken naar de invloed van binding op de vrije-energie bij het vouwen van verschillende eiwitten en substraten. Het resultaat van de simulaties liet zien dat toevallige, niet-selectieve interacties inderdaad plaatsvinden tussen eiwit en substraat, maar het verlies van entropie was te hoog om stabiele binding te bewerkstelligen, zelfs bij temperaturen gelijk aan een derde van de vouwtemperatuur. Dit geldt niet voor ontworpen bindplekken op het eiwit: deze zijn wél in staat om sterk aan hun partner te binden. Onze verklaring is dat er geen fundamenteel verschil tussen intra- en intermoleculaire interacties is.Als het eiwit dus in contact is met het substraat kunnen beiden samen worden gezien als een enkel eiwit. Om deze redenen gebruiken we voor het ontwerpen van specifieke eiwit-substraatbinding hetzelfde algoritme als voor het vinden van een sequentie die in een specifieke structuur vouwt.

In hoofstuk 6 behandelden we de vraag hoe een substraat de conformatie van een eiwit kan beïnvloeden. Hiertoe berekenden we de vrije energie van het bindingsproces en keken we naar de verandering in eiwitconformatie voor verschillende formaten van de bindingsplek. In alle gevallen duidde de structuur van het vrije-energielandschap erop dat de eerste stap van het bindingsproces een lokale ontvouwing van de in beginsel compacte conformatie met zich meebracht om het contact met het substraat te vergroten. Vervolgens zou het ketenmolecuul zich lokaal herschikken om de natieve (gevouwen) structuur te bereiken die overeenkomt met de gebonden toestand. De elementen die we tot nu toe hebben verzameld, laten het vermogen zien van modeleiwitten om gedeeltelijk te hervouwen in een nieuwe conformatie, om te binden aan een klein aantal specifieke partners, en om te hervouwen als ze gebonden worden. Hierna richtten we onze aandacht op de invloed van willekeurige domeinen op eiwit-eiwitinteracties. Veel eiwitten bevatten delen die ongeordend en sterk beweeglijk zijn. Experimenteel is het moeilijk om deze willekeurige domeinen te onderzoeken, aangezien ze feitelijk onzichtbaar zijn in röntgenkristallografie. Uit het feit dat een domein een willekeurige conformatie heeft,

volgt echter niet noodzakelijkerwijs dat het geen functie heeft. Om deze uitspraak te illustreren, ontwierpen we een roostereiwit dat niet zou kunnen vouwen in oplossing. Hierop brachten we een substraat in dat het eiwit bij voorkeur in een bepaalde conformatie zou binden. De simulaties wezen uit dat het eiwit vouwt als het gebonden wordt. Dit proces is omkeerbaar, in de zin dat het eiwit terugkeert naar zijn ongeordende toestand als het loskomt. De volgende vraag is of ons model ons toestaat een substraat te ontwerpen dat het eiwit helpt de natieve toestand te vinden, zelfs nadat het losgekomen is. In de praktijk werken dergelijke substraten als *vouwkatalysatoren*. Ze voorzien in een bindingsplek die, na een succesvolle vouwgebeurtenis, vrij is om een nieuw doeleiwit te ontvangen. Dergelijke hervouwingsmachines komen voor in de natuur en zijn bekend als *chaperonnes*. Er zijn verschillende klassen van zulke complexen. In het algemeen worden chaperonnes veronderstelt te functioneren als ontvouwings-hervouwingsmachines: op deze manier verlagen ze de barrière die een eiwit moet overkomen teneinde uit een lokaal vrije-energieminimum te ontsnappen om de natieve toestand te bereiken.

In hoofdstuk 7 richtten we ons op de hervouwingseigenschappen van een fascinerende subklasse van chaperonnes, de *chaperonines* ofwel chaperonnes die bestaan uit twee ton-achtige structuren. Deze grote eiwitcomplexen hervouwen niet-natieve eiwitten. Aanvankelijk worden de doeleiwitten opgesloten in de holte van een van de twee ton-achtige structuren van het chaperoninecomplex. Wat er binnen de holte gebeurt is nog steeds een onderwerp van discussie. Eerdere modellen beschreven chaperonine als een holte met een speciaal inwendig oppervlak dat een grote verscheidenheid aan eiwitten licht zou kunnen binden, onafhankelijk van hun conformatie. Computersimulaties lieten zien dat een dergelijke chaperonine inderdaad het vouwproces kan helpen, maar slechts in een klein bereik van de waarde van de aantrekkingssterkte tussen het eiwit en de muren van de holte. Hoewel zo'n model een mogelijke verklaring biedt voor het verhoogde vouwtempo, verklaart het niet de dubbele-holteconformatie van de chaperonine, noch zijn asynchrone activeit. In onze simulaties beschouwden we een alternatief mechanisme dat niet afhing van de gedeeltelijke ontvouwing van het eiwit ten gevolge van interacties met de inwendige muren. In plaats daarvan beschouwden we in onze simulaties de mogelijkheid van translocatie van het eiwit door een gat dat de twee holtes van het chaperoninecomplex met elkaar verbindt. Dit mechanisme bleek zeer efficiënt te zijn en volledig onafhankelijk van de specifieke volgorde of beginstructuur van de verkeerd gevouwen toestand.

We benadrukken dat alle bestudeerde modellen in dit proefschrift sterk zijn gesimplificeerd. Met meer realistische modellen is het echter (nog) niet haalbaar om een systematische studie uit te voeren naar de verschijnselen die we beschrijven. Uiteraard is het noodzakelijk om de simulaties aan experimenten te koppelen. In het geval van het chaperoninecomplex (hoofdstuk 7), is dit redelijk simpel, aangezien ons model specifieke en testbare voorspellingen doet over de werkingswijze van de chaperonine. Maar ook voor de onderwerpen die in de andere hoofdstukken besproken zijn, zou het mogelijk moeten zijn een verband te leggen met experimenten. Hiervoor zullen echter modellen nodig zijn die ontworpen zijn om een bepaald proces te beschrijven (bijvoorbeeld een specifiek substraatgeïnduceerd hervouwingsproces). We hopen dat dit proefschrift zal bijdragen aan zulke studies.

Acknowledgments

> "Not everything that can be counted, counts and not everything that counts can be counted."
> *Albert Einstein*

It is now time to thank all the people that had a strong and positive influence on my life and on my work during the years of my PhD. Let me start by expressing my gratitude to Daan for being an extraordinary group leader. His scientific knowledge was a constant source of inspiration during my research. He was always able to indicate me the right path to get out of dead ends. However, following the path was sometime hard and I could not have done it without the constant support of my friends that where always there to demolish my pessimism. I profited very much for their moral and scientific support. Their methods where particular convincing and often involved huge quantities of food. I will always remember the Maltese tuna fish of Mark , the illusional chicken of Chinmay and the chocolate cake of Marcia and Paulo. I have to thank Mark also for all the times he read some of my writings giving me crucial suggestions to produce something that could be called an English written text. Since the beginning of my PhD, I could rely on the life expertise of Fabrizio and Marco, that often gave me crucial information for the survival in the Amsterdam life, and for that I will always thank them. I also want to thank Angelo, Josep, MarcoJ, and Jorgos for teaching me new entries in my *Important things in life* list. Angelo also for taking the burden of being the English native speaker corrector when Mark left. Also a special thank for Chantal, who took care that nobody was ever left isolated from the group. I am very gratefull to Ruud and Michael that translated the Summary into the Samevatting, and to all the former and present colleagues for creating such a nice atmosphere in the institute. I cannot forget the numerous friends with whom I shared my life outside work, in particular I spent a lot of fantastic time together with Daniele (aka DM^3) Alejandro (aka Johan), Andrea, Tristan and Miriam (aka Lomendil), Franca and Eduardo. In my Acknowledgments I cannot ignore the persons that long before and during the period in Amsterdam, represent a reference point in my life. My family, my source, the origin of me as person. So far the experience gained during my life did changed me, but only as a variation over the *Coluzza*'s theme. Beatrice, su cui posso sempre contare e la cui dolcezza compensa per le amarezze che ho dovuto ingoiare. Davide, and Mattia that are my link with the real world outside physics and always shared with me the *Important things in life* list.

Ivan Coluzza was born on June 18 1978 in Rome, Italy. After finishing the high school in Switzerland, he studied physics in the University "La Sapienza" in Rome. During the period of his studies he also joined, for three months, the lab of Prof Elisha Moses at the Weizmann Institute where he participated to an experiment of protein incorporation in lipid vesicles. He graduated in Rome in September 2000 with a thesis on “Molecular Dynamics of rare events with multiple constraints” based on the research conducted in the group of Prof. G. Ciccotti. After graduating he moved to Amsterdam where he started as a PhD in the group of Prof. D. Frenkel. The 23rd of June he defended the present thesis.