ISAS - INTERNATIONAL SCHOOL
FOR ADVANCED STUDIES

Effective Potentials for Protein Folding Models

Thesis submitted for the degree of
Doctor Philosophiæ

Condensed Matter Sector

CANDIDATE
Cecilia Clementi

SUPERVISOR
Prof. Amos Maritan

October 1998

SISSA - SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI
TRIESTE
Via Beirut 2-4

TRIESTE

SISSA ISAS

SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI
INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

Effective Potentials for Protein Folding Models

Thesis submitted for the degree of
Doctor Philosophiæ

Condensed Matter Sector

CANDIDATE
Cecilia Clementi

SUPERVISOR
Prof. Amos Maritan

October 1998

To my dear grand father, Aldo

(“ ... speriamo di no ?! ”)

Contents

CHAPTER 1

INTRODUCTION

Proteins are heteropolymer chain molecules, built of various combinations of the 20 naturally occurring amino acids. They differ only in the sequence in which the amino acids are assembled into the polymeric chain. The number of amino acids of a protein can vary from few tens to several hundreds.

Proteins control most biological functions in living organisms. Virtually, every property that characterizes a living organism is affected by proteins: they store and transport a variety of particles ranging from macromolecules to electrons, transmit information between specific cells and organs, control the passage of molecules across cell membranes, have a function in the immune system of complex organisms, control genetic expression, are the crucial components of muscles, etc...

Protein activity mostly depends on their three-dimensional shape, that is stable against slight variations of the environment. Under normal physiological conditions (i.e. aqueous solvent, near neutral pH and temperature at $20^o - 40^o$) a protein spontaneously [1] folds into a well defined spatial structure, the *native state*.

[1] There are indications that some very large proteins need the presence of *molecular chaperons* (or *chaperonins*) to fold into their correct native state. The role of chaperonins in the folding process would be in binding transiently to certain incomplete proteins and prevent them from engaging in illicit

At the end of '50, Anfinsen [1] showed that the three–dimensional structure of a protein is entirely encoded by its sequence; in particular the stable fold of a protein does not rely on any special biological machinery.

The pioneering work of Anfinsen naturally led to the formulation of the following questions:

- ⇨ How is a specific three dimensional structure encoded in the amino acid sequence of a protein?
- ⇨ Given a three dimensional structure, which are the amino sequences that admit it as their native state?

These two simple but extremely important questions are known as the "Protein Folding Problem" and the "Inverse Folding Problem" (or "Design Problem"), respectively, and remain unanswered after 3 decades of intensive studies. Answering these questions is so vital and fascinating that the problem has become interdisciplinary and it is challenging biologists and physical chemists, as well as physicists and mathematicians.

The folding and the inverse folding problem are different formulations of the same problem, and both call for the understanding of the relationship between amino acid sequences and native structures.

Learning about the physical process underlying protein folding would be of great importance for biomedicine: in designing novel proteins with desired biological functionality, in designing new drugs, and in predicting the protein function from the knowledge of the mere amino acid sequence. In fact, obtaining amino acid sequences is much simpler than obtaining the structures of proteins. The data base of sequences is already huge: more than 50000 sequences are known, and the number of new sequences is approximately doubling every year. Amino acid sequences are obtained by biochemical methods, either by direct determination from the protein or, indirectly, but more rapidly, from the nucleotide sequence of the corresponding gene or DNA. On the other hand, determining the protein structures is a long process. About 7500 proteins are known at atomic resolution, most of them from *x-ray crystallography*, some of the smaller ones from methods involving *nuclear magnetic resonance* (NMR). In general, this latter method gives structural informations at very low resolution while high enough resolution (better than $\sim$ 2.5 Å) is provided from x-ray crystallography. However, a prerequisite for the utilization of crystallographic method is the crystalliza-

associations.

tion of a protein (i.e. the packing of many identical protein molecules into a regular, repeating array), that is usually quite difficult to achieve.

In principle, one could explore the protein behavior by numerically integrating motion laws of each protein degree of freedom, using interaction energies obtained from experiments, and including all the details of the protein-solvent environment. However this is beyond the current computational capabilities. Supercomputers can currently simulate up to nanosecond of real-time protein dynamics with time steps appropriate to simulate harmonic motions of bonded atoms. This time scale is still too small with respect the $10^{-3} - 10$ seconds typically required to fold real proteins. This difficulty is thought to be related to the existence of a huge number of local energy minima (even in the neighborhood of the native state, that is commonly considered the global minimum) which prohibits rapid conformational sampling of phase space. Because of these limitations in exploring the correspondence between amino acid sequences and native structures atomistic models are usually abandoned in favor of coarse–grained models, where amino acids are represented in a simplified way, averaging over suitable degrees of freedom.

Plan of this thesis

Here we will address to the problem of reproducing the essential features of real proteins by simplified models. The models we will adopt require a definition of an effective potential function that is a suitable approximation to the actual potential. The problem of determining the effective interaction potential is discussed in the second Chapter. We will present a strategy and test our approach on simple lattice models, where the true interactions are known *a priori*.

The virtue of lattice models is that their ground state can be known exactly and, at least in principle, one may take into account all possible conformations. Nevertheless, using virtual bonds of fixed length and forming fixed angles, lattice models are often too simple in order to reproduce quantitatively real protein behavior.

In Chapter 3 we will extend the method of the previously mentioned approach to off–lattice situations.

Chapter 4 is devoted to a comparison between our off–lattice model and another different approach to the protein folding problem: the contact map representation.

In Chapter 5 we will discuss an improved and a more realistic off–lattice model. It will be shown that the model can reproduce the real structure of a real protein

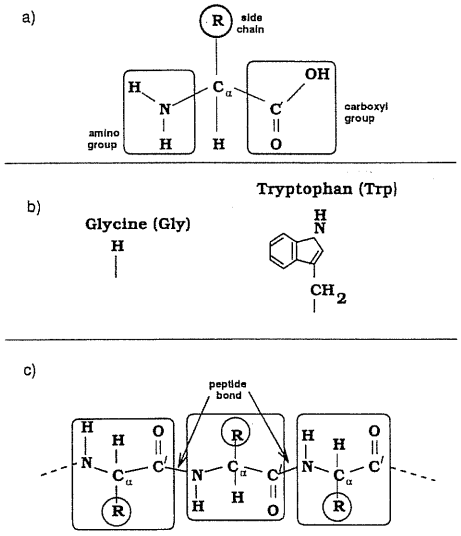

Figure 1.1: A schematic diagram of an amino acid is shown on the top (a). Chemical structure of two side chain: Glycine and Tryptophan (b). On the bottom (c), repeating units are bonded in a polypeptide chain, forming peptide bonds.

(crambin) within the limits of experimental errors.

Finally, Chapter 6 is focused on some possible applications of previously discussed models and techniques, to predict the characteristic features of protein aggregation in forming multimeric structures. These studies are currently under investigations, but the preliminary results seem very promising.

In the remainder of the present Chapter we present an overview of the experimental and theoretical framework that stays at the basis of the work discussed in this thesis.

1.1 Protein structures

All the 20 amino acids have a similar chemical structure. They all have a central carbon atom (usually indicated as C_α) to which are attached a hydrogen atom, an amino group (NH_2), and a carboxyl group ($COOH$), as shown in Figure 1.1(a). The carbon atom of the carboxyl group is usually indicated as C'. These common atoms represent the *main chain* (or *backbone*) atoms of a protein. Different species of amino acids are distinguished by a different side–chain R, bonded to the C_α atom. The chemical structures of Tryptophan (the biggest side chain) and Glycine (the smallest side-chain) are shown in Figure 1.1(b).

The polypeptide chain is made by peptide bonds, linking the C' of one amino acid to the nitrogen atom of the next. One water molecule is eliminated in this process. The

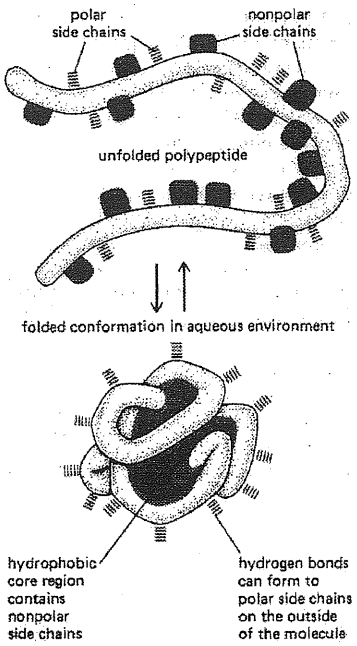

Figure 1.2: The polar amino acid side chains tend to gather on the outside of the protein, where they can interact with water; the non polar amino acid side chains are buried on the inside to form a hydrophobic core that is "hidden" from water.

resulting chain is called *primary structure* of the protein.

Depending on the chemical nature of the side chain, the amino acids are usually divided into different classes. Some amino acids have a net charge. The most important differentiation among the 20 amino acids is between *hydrophobic*, if they have an unfavorable interaction with water, or *polar*, in the opposite case. Indeed, the main driving force for folding water-soluble protein molecules is to pack hydrophobic side chains into the interior of the molecule [1, 2, 3, 4], creating a *hydrophobic core* shielded from the solvent by a *hydrophilic* surface, as shown schematically in Figure 1.2.

Proteins, in their native state, are more tightly packed than almost any other organic matter. The packing efficiency ratio of internal proteins is roughly what is expected for the close-packing of hard spheres (0.74).

There are regular features present in protein structures. Small portions of the protein chain, typically consisting in about 10 amino acids, can be organized in local substructures, called *secondary structures*. Secondary structures are one of two types: *alpha helices* or *beta sheets*. Both types are characterized by having the main chain

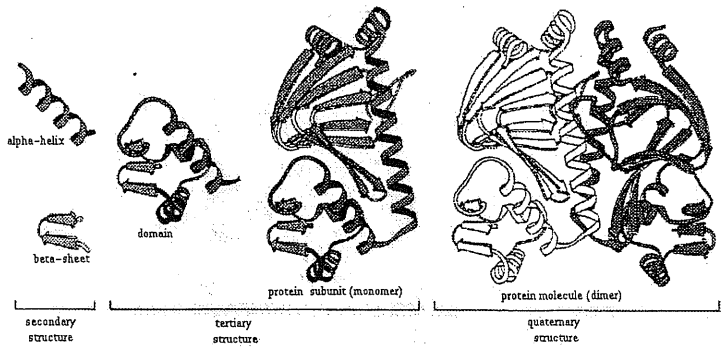

Figure 1.3: The three dimensional structure of a protein can be described in terms of different levels of folding, each of which is constructed from the preceding one in hierarchical fashion. These levels are illustrated in this picture using the *catabolite activator protein (CAP)*, a bacterial gene regulatory protein with two domains.

NH and $C'O$ groups connected by hydrogen bonds. Most protein structures are built up from combinations of secondary structures elements which are connected by *loop regions* of various lengths and irregular shape. Simple combinations of a few secondary structure elements with a specific geometric arrangement have been found to occur frequently in protein. These units have been called *supersecondary structures* or *motifs*. Several motifs usually combine to form compact globular structures, which are called *domains*. *Tertiary structure* is a common term both for the way motifs are arranged into domain structures and for the way a single polypeptide chain folds into one or several domains. Many protein molecules have only one chain and are called *monomeric proteins*. However a fairly large number have several identical polypeptide chains that coalesce into a multimeric molecule with a specific *quaternary structure*. Figure 1.3 summarize this structural hierarchy.

Based on simple considerations of connected motifs, protein structures can be classified into three main group [5]: α domains, when the protein core is build up exclusively from α helices, β domains, when the core comprises β sheets, and α/β domains, when the protein structure is made from combinations of α and β motifs. In addition to these groups, there are a number of small proteins, rich in disulfide bonds or metals, that do not exhibit secondary structure. The structure of these proteins seem to be strongly influenced by the presence of these metals or disulfides and often look like distorted version of more regular proteins.

1.2 The protein folding energy landscape

The energy of a protein is the sum of antagonistic interactions between amino acids and chain topological constraints, so that protein are characterized by frustration. This fact produces a complex connectivity pattern between low energy states and the folding landscape of proteins is necessarily rugged.

May, then, proteins be thought as random objects ?
In a simple model one can assume that, when non-native contacts are made the energy contributions are random, so that these contributions to the protein's energy could be treated just as for a random heteropolymer. In this spirit, many ideas (see for example [6, 7, 8, 9]) are borrowed from the statistical mechanics of disordered system, in particular from the Random Energy Model (REM) [10].

Random heteropolymers have an underlying driving force to collapse, but adopt three–dimensional structures which depend too much on the specific initial conditions, because of the conflict of different interaction energies. This is in contrast with the common wisdom that evolution has selected sequences thermodynamically stable in their native state and capable to reach it rapidly and almost independently on the initial conditions, by overcoming energy barriers at the relevant (folding) temperature. Because native contacts and local conformation energies are more stable in proteins than expected from random heteropolymer, recent studies supposed [11, 12, 13, 14, 15, 16, 17, 18, 19, 20] a smooth overall slope of the energy landscape towards the native structure. In this hypothesis the folding landscape of a protein should resemble a partially rough funnel riddled with traps where the protein can transiently reside and proteins are thought as heteropolymer with minimally frustrated sequences [21].

In order to have a funnel-like energy landscape, an ideally designed folding sequence should have the energy of its conformations proportional to a parameter Q, that measures the similarity between this native structure and any other conformation, plus some roughness introduced by the non-native contacts. This correlation between energy and structure introduces a bias that favors not only the native configuration, but that biases all non-native conformations according to their degree of similarity to the folded state. This correlation is responsible for the funnel shape of the landscape. Even conformations that are completely different but have similar Q (native parts are different) could have similar energies. A random sequence would display no such correlation between energy and structure, leading to a rough landscape without any globally preferred structure. Minor variations in pH, temperature, denaturant concen-

tration or mutations will affect the native configuration by favoring other low–energy structures, that in a funnel-like landscape are very similar to the original one. This competition between energetic bias towards native conformation and roughness seems to be fundamental in determining the folding mechanism, and in leading to a diversity of folding scenarios [22]. In this viewpoint, what is most important for understanding the folding process is a global overview of the protein energy surface. In fact, folding seems to occur through organizing an ensemble of structures rather than only a few uniquely defined structural intermediates, in such a way that there is no unique pathway but a multiplicity of convergent folding routes towards the native state.

CHAPTER 2

EFFECTIVE INTERACTION POTENTIALS

A major problem in protein science is that although several structures are experimentally known relatively accurately (typically with a 2 Åresolution) the interaction stabilizing them have escaped quantitative estimation. On the other hand, the prediction of the three–dimensional structure of the native state of proteins from the knowledge of their amino acid sequences can only be achieved if the interaction potentials among various parts of the peptide chain in the presence of solvent molecules are known to some extent. The native states of globular proteins, in fact, correspond to the conformations which are global minima of the free energy [1], then the knowledge of the energy of a sequence in a given conformation would be an important step towards the complete solution of both the folding problem, and the inverse one, i.e. the design of a sequence of amino acids that rapidly folds into a desired conformation.

2.1 Forces determining protein structure

In principle, in order to evaluate the energy of a sequence in a given conformation, it is necessary to take into account non covalent forces occurring between atoms in the protein chain and between each amino acid atoms and the solvent. The (covalent) peptide bond may be omitted from the calculation of the energy since it cancels out in the evaluation of energy difference between different conformations.

Non–covalent energies are three orders of magnitude smaller than covalent binding energies. All interactions arise from a limited set of non–covalent forces, namely:

- ✔ *short-range repulsions* between any pair of atoms as they approach each other. As they come near enough for their electron orbitals to overlap significantly, the effective repulsion increases enormously as a consequence of the Pauli exclusion principle;
- ✔ *electrostatic forces* between partially charged atoms according to Coulomb's law. These interaction are modulated by the dielectric constant of the surrounding medium;
- ✔ *van der Waals interactions* as a results of mutual interactions due to induced–polarization effects;
- ✔ *hydrogen bonds* occurring when two electronegative atoms compete to be bonded to the same hydrogen atom.

Moreover, in a rigorous treatment one should take into account simultaneously the state of the protein and its environment, typically an aqueous solution. The interaction of water with ions, dipoles, and hydrogen bond acceptor and donors is highly effective in reducing the forces that occur among such groups in vacuum or in a non–polar solvent. In particular, water induces an effective force between non–polar atoms, this effect has come to be known as the *hydrophobic interaction*. This interaction is a major factor in the stabilization of proteins and results in a tendency of non–polar atoms to cluster together in order to minimize the area exposed to the solvent (as schematically shown in Figure 1.2).

2.2 Coarse–grained models

The most rigorous approach to the interaction potential determination would be to derive the forces between amino acids [23, 24] from quantum mechanical calculations

and from spectroscopic experimental data.

Even if one were able to calculate the exact contribution of each energetic term among the huge number of interacting atoms constituting the protein, such a rigorous approach from "first principles" would be not practical and beyond actual computational possibilities.

An alternative approach consists of introducing a coarse-grained description, keeping into account just one degree (or few degrees) of freedom to represent an entire amino acid. In this spirit, *lattice models* have been widely used in the recent literature for various goals, ranging from folding dynamics to thermodynamic properties of folded states of proteins (see e.g. [25, 26, 9]) . In lattice models a peptide chain is represented as a self–avoiding walk whose nodes represent extremely simplified amino acids. The advantages are that it is possible to do several rigorous tests for most commonly techniques used in the literature for the study of protein folding. Results of lattice test have shown that commonly used techniques are often unjustified or based on some misunderstood principle of statistical mechanics (see below for a more detailed discussion).

One of the main difficulties with coarse–grained representations of protein chains is the fact that one needs to introduce an *effective* Hamiltonian that captures the essential features of the problem under study. In the most commonly used model Hamiltonian, effective two-body forces between neighboring (in space but not in sequence) amino acids are the only interactions taken into account [27, 28]. These forces have to represent the global effect of all pairwise interactions mentioned before among each couple of atoms as well as solvent induced interactions.

Probably the simplest model is the H-P model [28] where there are only two classes of amino acids, mimicking hydrophobic (H) or polar (P) tendency. The $H-P$ Hamiltonian for a sequence S in a spatial conformation Γ is:

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

where $\mathbf{r}_i$ denotes the lattice position of the i–th amino acid in the configuration Γ and the contact matrix $\Delta_\Gamma(\mathbf{r}_i - \mathbf{r}_j)$ is 1 if $\mathbf{r}_i$ and $\mathbf{r}_j$ are nearest neighbor sites of Γ that are not occupied by consecutive amino acids and zero otherwise. The amino acid s_i can be either H or P and $B(H,H) = \epsilon_{HH}$, $B(H,P) = B(P,H) = \epsilon_{HP}$ and $B(P,P) = \epsilon_{PP}$ are the interaction parameters. To favor the collapse of an hydrophobic core the interaction between two hydrophobic amino acids are chosen to be attractive

($\epsilon_{HH} < 0$) and stronger than the other interactions ($\epsilon_{HH} < \epsilon_{HP}$, ϵ_{PP}).

Once a model Hamiltonian has been chosen one has to fix parameters entering it in order to reproduce as good as possible the complex interactions between amino acids. In the case of $H-P$ model parameters to be fixed are the strength of interactions ϵ_{HH}, ϵ_{HP} and ϵ_{PP}.

A more realistic model could be defined by eq. (2.1) where s_i labels one of the 20 different species of amino acids. In this case B is a 20×20 symmetric matrix involving 210 interaction parameters to set appropriately.

2.3 Extraction of energy parameters

Customarily, the potential energies of the effective interactions are derived from pairing frequencies of amino acids observed in the native structures contained in the Protein Data Bank (PDB) [27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40]. In this approach, real proteins are represented as chains of beads where a residue is represented by a suitable chosen centroid. Within this scheme two non consecutive amino acids are said to be in contact if their space separation does not exceed 6.5 Å.

The Hamiltonian of the system is taken to be like the expression (2.1), and parameters $B(s_i, s_j)$ are estimated from the relative abundance of contacts between amino acid of species s_i and s_j observed in static structures of different proteins and invoking a Boltzmann distribution. We recall the method with some details in § 2.4.3.

This method, known as the quasi-chemical approximation, is widely used and relatively easy to implement in such a difficult context.

In important recent work, Thomas and Dill [41] have rigorously tested the underlying assumptions and approximations of the quasichemical method in a lattice model of the type (2.1). Using an a priori assigned interaction potential, one is able to construct a Model Protein Data Bank (MPDB) identifying proteins as amino acid sequences having a unique ground state conformation (native state) among all possible conformations (this is accomplished by exhaustive exact enumeration for sufficiently small values of the protein chain length and/or the number of amino acid classes). Figure 2.1 summarizes the construction of a MPDB in the lattice.

Applying the quasichemical method to several of these exact cases, Thomas and Dill [41] demonstrated the inadequacies of the method and identified its possible weak points. Indeed, the interaction parameters could, in the worst cases, be off by as much

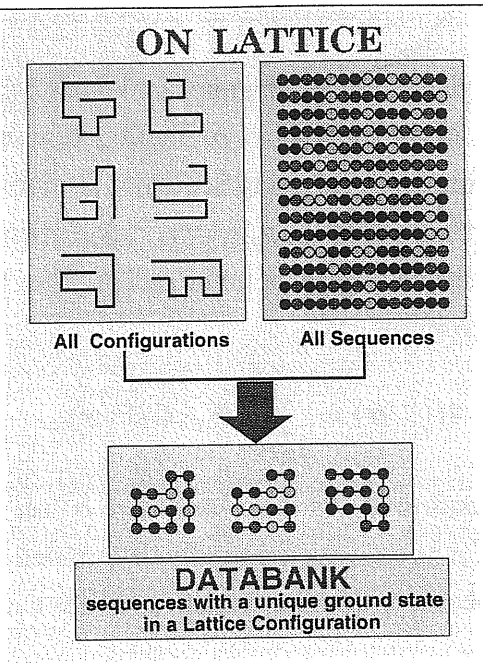

Figure 2.1: Construction of a toy-model PDB on the lattice. Sequences having a unique ground state in a lattice configuration are chosen to represent model proteins.

as a factor of two and the native states of protein sequences of the MPDB could be correctly identified only for 84% of the cases, a rather poor result given the simplicity of the model.

At the beginning of the 90's Crippen [42] proposed a novel approach to the determination of effective interaction potential. His method is based on the requirement that the energy of a sequence in its native state is lower than the energy of the same sequence in any other spatial configuration.

Taking inspiration from Crippen works, we recently developed [43, 44] a new optimization method for this problem. Our technique will be analyzed in depth in the next sections, where we assess its performance against existing techniques (for both 2 and 4 classes of amino acids). In § 2.4.3 we discuss the Crippen method, comparing it to our method.

2.3.1 The model

We consider a set of N_s sequences $\Omega_\sigma = \{\boldsymbol{\sigma}_s\}_{s=1,..,N_s}$ each comprising of N amino acids $\boldsymbol{\sigma}_s = \{\sigma_i^s\}_{i=1,..,N}$. Each sequence $\boldsymbol{\sigma}_s$ is postulated to have a unique native state (assumed to be the ground state) in a known spatial conformation, Γ_s. The set of all native conformations is denoted by $\Omega_\Gamma = \{\Gamma_s\}_{s=1,..,N_s}$.

We assume that for a given number of amino acid types N_a, the effective interaction potentials can be written in the form of a symmetric pairwise interaction matrix $E_{\mu\nu}$, where $\mu, \nu = 1, .., N_a$. Similarly, for every sequence-conformation pair we introduce a symmetric contact matrix $C(\Gamma, \boldsymbol{\sigma})_{\mu\nu}$, $\mu, \nu = 1, .., N_a$, giving the (effective) number of contacts between the different types of amino acids. Thus the energy of a sequence $\boldsymbol{\sigma}$ in conformation Γ is given by

$$\mathcal{H}(\Gamma, \boldsymbol{\sigma}) = \sum_{\mu \leq \nu = 1}^{N_a} E_{\mu\nu}\, C(\Gamma, \boldsymbol{\sigma})_{\mu\nu} \,. \tag{2.2}$$

The number of spatial conformations that a sequence $\boldsymbol{\sigma}$ can take is gigantic and will be indicated as Γ while $\Gamma_{\boldsymbol{\sigma}}$ is the experimentally determined native state structure. At a temperature T, the probability that the sequence $\boldsymbol{\sigma}$ is in a conformation Γ is simply given by [45, 46, 47]

$$P_\Gamma(\boldsymbol{\sigma}) = \exp\left[-(\mathcal{H}(\Gamma, \boldsymbol{\sigma}) - F(\boldsymbol{\sigma}))/T\right] \,, \tag{2.3}$$

where the Boltzmann constant k_B was set equal to 1, and $F(\boldsymbol{\sigma})$ is the free energy, defined as

$$F(\boldsymbol{\sigma}) = -T \log \left(\sum_{\Gamma'} \exp[-\mathcal{H}(\Gamma', \boldsymbol{\sigma})/T] \right) \,. \tag{2.4}$$

This definition of the free energy ensures that the probability $P_\Gamma(\boldsymbol{\sigma})$ of eq. (2.3) is properly renormalized.

Because the experimentally observed structure of the sequence $\boldsymbol{\sigma}$ is the conformation $\Gamma_{\boldsymbol{\sigma}}$, the value of $P_{\Gamma_{\boldsymbol{\sigma}}}(\boldsymbol{\sigma})$ must be large ($> \frac{1}{2}$) at temperatures below the folding transition temperature. Indeed, $P_{\Gamma_{\boldsymbol{\sigma}}}(\boldsymbol{\sigma})$ should be equal to 1 at zero temperature, if the ground state is non-degenerate.

In [48], Crippen has suggested that even with the knowledge of the exact contact potential from which the folding sequences and their unique native conformations are determined, one may not be able to correctly select which sequences fold to a desired target structure. Indeed if one attempts to design a sequence in a given structure by minimizing the energy $\mathcal{H}(\Gamma, \boldsymbol{\sigma})$, then the resulting sequence could be not the right one, i.e. the one having its native state in the target structure. The resolution [46] of this puzzle is that the right "score" to be maximized in the inverse folding problem is (2.3), i.e. $\mathcal{H}(\Gamma, \boldsymbol{\sigma}) - F(\boldsymbol{\sigma})$ has to be minimized, and not just the energy $\mathcal{H}(\Gamma, \boldsymbol{\sigma})$. A key feature of this score is that $F(\boldsymbol{\sigma})$, the free energy, does not depend on the specific target structure, but formally it depends only on the sequence being considered. Thus,

the determination of the exact contact potential is a valuable first step for an attack on the protein design problem.

In what follows, we describe a zero temperature version (which is appropriate in most instances) of such a procedure to extract the exact potentials. Furthermore, we restrict ourselves to models where each conformation is a self-avoiding walk consisting of joined nearest neighbor sites of a d-dimensional hypercubic lattice ($d = 2$ in the present applications). Amino acids are placed at lattice sites, and contacts are defined to be 1 for non–consecutive, neighboring sites, and 0 otherwise.

2.3.2 The method

Instead of starting immediately with a cost function that has to be minimized, we concentrate for a moment on the space spanned by the interaction potentials. Since all energies scale linearly with the amplitude of the interaction potentials, we have to keep e.g. the first parameter fixed to set a scale.

Introducing the vector $\vec{E} \equiv (E_{11}, E_{12}, \ldots, E_{N_a N_a})$ and $\vec{C}(\Gamma, \boldsymbol{\sigma}) \equiv (C(\Gamma, \boldsymbol{\sigma})_{11}, C(\Gamma, \boldsymbol{\sigma})_{12}), \ldots, (\Gamma, \boldsymbol{\sigma})_{N_a N_a})$ of $D \equiv N_a(N_a + 1)/2$ components, we can rewrite eq. (2.2) as

$$\mathcal{H}(\Gamma, \boldsymbol{\sigma}) = \vec{E} \cdot \vec{C}(\Gamma, \boldsymbol{\sigma}) \tag{2.5}$$

The fact that a sequence has a lower energy in its native conformation than in any alternative conformation, provides a linear inequality (or hyperplane) in the parameter space separating the space into allowed and forbidden halfspaces. If we define:

$$\vec{X}(\Gamma, \boldsymbol{\sigma}) \equiv \vec{C}(\Gamma, \boldsymbol{\sigma}) - \vec{C}(\Gamma_{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) \tag{2.6}$$

the allowed points in parameter space have to satisfy the linear inequality

$$\vec{E} \cdot \vec{X}(\Gamma, \boldsymbol{\sigma}) > 0 \,. \tag{2.7}$$

for all $\Gamma \neq \Gamma_{\boldsymbol{\sigma}}$. The left term of eq. (2.6) represents the energy gap of $\boldsymbol{\sigma}$ in the conformations Γ and $\Gamma_{\boldsymbol{\sigma}}$.

Repeating this operation for all the sequences in Ω_σ and for all the alternative conformations Γ, and retaining only the allowed part of the parameter space that satisfies all the inequalities, we obtain a convex *cell* around the target parameters. All the points in this cell correspond to potential parameters that yield the correct native conformation as the unique ground state for each of the sequences in Ω_σ. In the test model, we

have generated the set Ω_σ using the energy function (2.2), and therefore, the existence of the cell is guaranteed and the problem is well posed. For real proteins the form of the energy function has to be postulated. The validity of the ansatz is tested by the (non)existence of a finite cell.

Each inequality corresponds to a hyperplane in the $D-1$ dimensional parameter space (spanned by $\vec{E}'$s with the 11-component fixed) separating allowed and forbidden half-spaces. The distance of any point $\vec{E}$ in parameter space to this hyperplane, is related to the energy gap between the two configurations leading to this inequality (at the value of parameters given by $\vec{E}$) by the following linear equation:

$$\mathrm{dist}(\vec{E},\vec{X}) = \frac{\vec{E}\cdot\vec{X}}{\sqrt{\vec{x}^2}}\,, \tag{2.8}$$

where $\vec{x}$ is the vector obtained from $\vec{X}$ by setting the $_{11}$-component to zero. Using all the information in the data set, the cell is maximally reduced. A selection procedure is needed in order to isolate an optimal point within the cell. To be really predictive, the selection procedure has to be maximally robust to the modification of the cell. That is to say, the point selected using a restricted set of informations would resist to the addition of informations.

The optimal interactions are chosen such that the smallest gap among all the sequences in the training set is as large as possible. The cost function ($F_{\mathrm{gap}}(\vec{E})$) is hence taken to be minus the smallest gap, i.e.

$$\begin{aligned} F_{\mathrm{gap}}(\vec{E}) &= -\min_{\boldsymbol{\sigma}\in\Omega_\sigma}\min_{\Gamma\neq\Gamma_{\boldsymbol{\sigma}}}\{E(\Gamma,\boldsymbol{\sigma})-E(\Gamma_{\boldsymbol{\sigma}},\boldsymbol{\sigma})\} \\ &= -\vec{E}\cdot\vec{X}(\Gamma^*,\boldsymbol{\sigma}^*) \end{aligned} \tag{2.9}$$

where $\boldsymbol{\sigma}^*$ and Γ^* are the sequence and the alternative conformation respectively that yield the minimum gap. In other words, the interaction potentials are chosen in such a way that the maximum minimum (mxm-) gap is obtained.

This cost function has two major advantages over previous attempts. First, it ensures automatically that all sequences have their unique ground states in the correct structures. In fact, a negative mxm-gap would imply that the a priori assumption of the form of the energy function (2.2) is incompatible with the data in the training set.

Second, it does not suffer from an unphysical bias due to statistical fluctuations that were present in the cost functions proposed in [43]. These cost functions not only make use of all inequalities, but also of the number of occurrences of each inequality over the

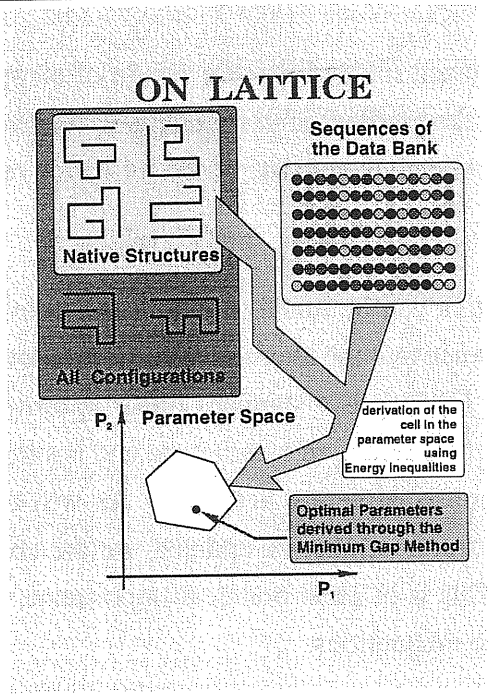

Figure 2.2: Extraction of potential parameters from the knowledge of sequences and relatives structures from the MPDB. We select a set of parameters in the parameter space that are able to assign the right structure to each PDB sequence. In this set we select the point for which the smallest gap among all the sequence is the largest possible.

training set. Therefore, it may happen that inequalities that occur more frequently, push the optimal parameters away from their true values. One may expect all sequences in the training set to satisfy the minimum conditions that make them good folders, which implies that each inequality is equally important, irrespective of the number of times it occurs. In realistic cases it may be important to rescale the energy gap associated with a given sequence with respect to its ground state energy. Because, in this work, we have sequences of the same length in a given training set, all ground state energies are roughly the same, and rescaling of the gap is not necessary.

Figure 2.2 is a schematic summary of this extraction procedure.

Since the width of the energy gap is not equal to the distance in parameter space (eq. 2.8), inequalities that do not contribute to the boundaries of the cell may still influence the mxm-gap. Nevertheless, for all the cases that we have encountered, using only the inequalities contributing to the cell, we obtain exactly the same optimal parameters as the ones derived by using all inequalities. This simplifies the maximization procedure. The sequences have to be mounted on every alternative conformation only once before the optimization procedure, and we retain only those inequalities that contribute to the cell. Then we start our optimization procedure with only these relevant, few in-

equalities. Given the relevant inequalities the design of an optimization procedure is straightforward, because the gradient of each inequality is given by $\vec{x}$. Therefore, each step in the optimization procedure consists of the following replacement

$$\vec{E}' = \vec{E} + \gamma\, \vec{x}(\Gamma^*, \boldsymbol{\sigma}^*)\ , \tag{2.10}$$

where γ is a parameter that can be tuned to obtain fast convergence, and $\boldsymbol{\sigma}^*$ and Γ^* are respectively the sequence and the alternative conformation that yield the minimum gap at given parameters $\vec{E}$. This is a standard linear optimization procedure (named *perceptron learning*) for the interaction parameters, and is similar to the one introduced in [49, 50, 51] in the context of neural networks. As will become clear in the next section however, after a small number of updates, we are able to select and save the important inequalities. This is of great practical importance, given the computational cost of (re)calculating the inequalities.

2.4 Results

The method has been tested on models with an increasing number of interaction parameters to check the dependence on the dimensionality (N_p) of the parameter space. The test has been done on the standard H-P model [28], $N_a = 2$, with nearest neighbor (nn) interactions, i.e. with 2 free parameters ($N_p = 2$), and some variations like considering next nearest–neighbor interactions, to check the robustness of the method.

Furthermore, the method has been applied to models with 4 types of amino acids ($N_p = 9$) and nn interactions. For the latter we have also studied the dependence of the quality of the obtained parameters on the size of the MPDB and of the set of alternative structures.

Although the potential extraction is still feasible up to parameter numbers as high as $N_p = 9$, increasing the number of interactions and consequently the dimension of the cells, leads to very lengthly calculations for the cell corners (even retaining only relevant inequalities).

With increasing dimension, the number of inequalities contributing to the cell grows linearly, while the number of corner points of the cell grows exponentially. Therefore, one may have to opt for a hybrid method. The first step consists of a rough optimization recalculating the inequalities at each update (eq. 2.10). Once a point in parameter space satisfying all constraints (i.e. in the cell) has been obtained, the inequalities at a distance less than some tolerance parameter from this point are selected

and saved. The number of such inequalities is relatively low and it grows linearly with the number of parameters. Then, the full optimization is done only with respect to these inequalities (eq. 2.10). An implementation of this method on the model with $N_p = 9$ for different choices of parameters shows that roughly 20-30 updates for the first step are needed. Then, typically about 100 inequalities have to be saved to perform the second step. This hybrid method is very efficient because it uses the insights in the geometry of the cells in parameter space and avoids unnecessary time losses.

2.4.1 Results for the H-P model

As a benchmark of our method we studied extensively the well known H–P model of Lau and Dill [28], which has 2 types of amino acids. To fix the energy scale, we have chosen to fix the parameter E_{HH} equal to -1. Hence, we are left with two independent parameters (E_{PP} and E_{HP}) and a 2 dimensional parameter space, which allows us to clarify our reasoning with instructive pictures. We have considered three types of target interaction parameters:

$$(E_{HH}, E_{PP}, E_{HP}) = (-1, 0, 0),\ (-1, -1/\sqrt{2}, 0),\ (-2, -2, -1) \qquad (2.11)$$

and seven distinct groups of amino acid chains each of length: $N = 10, .., 16$.

To reduce the required computer time, we restricted our search to all semi-compact two dimensional conformations on a square lattice. By semi-compact, we mean that we restrict the conformations to a box of size 5×5 (tests with all conformations of a certain length on a square lattice show that the results are unaltered).

As alternative conformations, we considered both the set of encodable conformations $\Omega_\Gamma^{\text{good}}$ (having at least one sequence that has its unique native state in it), and the set of all conformations $\Omega_\Gamma^{\text{all}}$ (obtained by complete enumeration and also used to generate the good sequences). Although good results can be obtained considering only $\Omega_\Gamma^{\text{good}}$, it is not excluded (and is indeed observed) that additional information can be gained (in the sense of new inequalities, further reducing the cell) by also considering new conformations from $\Omega_\Gamma^{\text{all}}$. However, when the cells are closed, in all cases we obtain exactly the same set of optimized parameters. Since some inequalities are very rare, it is difficult to say how many alternative conformations are needed to maximally reduce the cell. Disappointingly, so far we have not found a criterion to determine beforehand whether a certain sequence and a given alternative conformation gives rise to a "tight" inequality (contributing to the boundaries of the final cell) or not. Therefore,

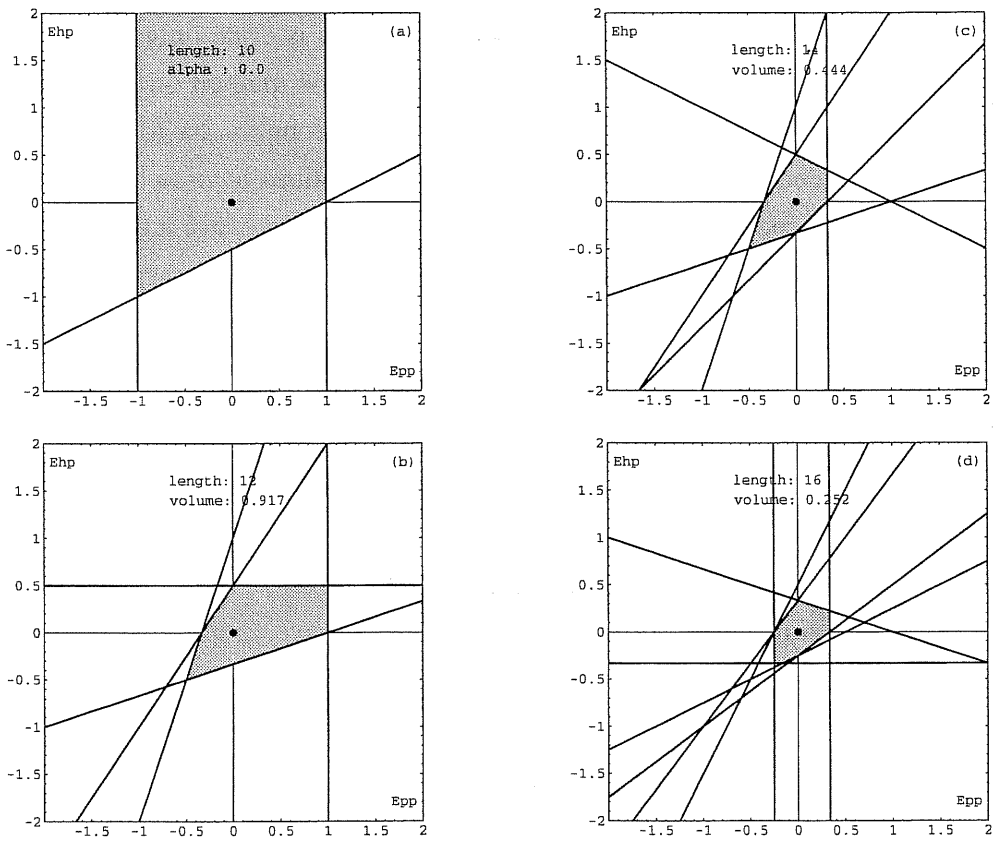

Figure 2.3: The 2 dimensional cells for the target parameters ($E_{HH} = -1, E_{PP} = 0, E_{HP} = 0$) (in units of k_BT), using all the structures as alternatives, for different sequence lengths. Indicated are the cell (shaded area), the target parameters (fat point), the sequence length (10, 12, 14, 16), the volume (if the cell is closed) or the opening angle alpha of the allowed area.

although the obtained parameters are relatively stable to changes in the shape of the cell, the best strategy seems to be to use as much information as is available, or as is numerically feasible. It cannot be ruled out that regenerating all the good sequences with the newly obtained parameters, would add some new "good" sequences to the set.

Tables 2.1, 2.2, 2.3 and Figure 2.3, show that the volume of the cells tends to decrease monotonically with increasing sequence length. The only exceptions are observed with length $N = 13$, but are probably due to finite–size effects. In two cases, a segment of a line of points in parameter space yields the same mxm-gap.

In all the cases that we have considered (where the ratios of the target potentials are rational numbers made up out of small integers) the maximization of the minimum gap

Table 2.1: Results for the H-P model fixing E_{HH} to its true value to set the energy scale (in units of k_BT) for the first set of target interaction parameters. The Table shows the sequence length, the obtained interaction parameters, the true minimum gap with the target parameters, the obtained mxm-gap, the volume of the cell (or opening angle in cases in which the cell is not closed) both using all conformations and only the "good" ones as alternative conformations. The success rate in the prediction of native conformations of the "good" sequences with the obtained parameters is 100% in all cases that the cell is closed.

target	0.0	0.0				
N	E_{PP}	E_{HP}	truegap	mxmgap	$\mathrm{vol}(\Omega_\Gamma^{\mathrm{all}})$	$\mathrm{vol}(\Omega_\Gamma^{\mathrm{good}})$
10	/	/	1.0	/	0.0^o	2.034444^o
11	0.0	0.0	1.0	1.0	1.062500	0.643501^o
12	0.0	0.0	1.0	1.0	0.916667	0.643501^o
13	0.0	0.0	1.0	1.0	0.625000	0.0^o
14	0.0	0.0	1.0	1.0	0.444444	0.0^o
15	0.0	0.0	1.0	1.0	0.272817	1.203704
16	0.0	0.0	1.0	1.0	0.252315	0.611111

Table 2.2: As in Table 2.1, but for the second set of target interaction parameters.

target	-0.707107	0.0				
N	E_{PP}	E_{HP}	truegap	mxmgap	$\mathrm{vol}(\Omega_\Gamma^{\mathrm{all}})$	$\mathrm{vol}(\Omega_\Gamma^{\mathrm{good}})$
10	/	/	0.12132	/	0.0^o	0.321751^o
11	-5/7	0.0-0.10	0.12132	0.142856	0.034722	0.034722
12	-5/7	0.0-0.04	0.12132	0.142856	0.017361	0.029514
13	-5/7	0.0	0.12132	0.142856	0.030556	0.030556
14	-5/7	0.0	0.12132	0.142856	0.015129	0.019097
15	-5/7	0.0	0.12132	0.142856	0.007955	0.007955
16	-5/7	0.0	0.12132	0.142856	0.007955	0.007955

Table 2.3: As in Table 2.1 and 2.2, but for the third set of target interaction parameters.

target	-2.0	-1.0				
N	E_{PP}	E_{HP}	truegap	mxmgap	$\mathrm{vol}(\Omega_\Gamma^{\mathrm{all}})$	$\mathrm{vol}(\Omega_\Gamma^{\mathrm{good}})$
10	/	/	1.0	/	0.0^o	1.249046^o
11	-2.0	-1.0	1.0	1.0	0.733333	0.785398^o
12	-2.0	-1.0	1.0	1.0	0.733333	0.785398^o
13	-2.0	-1.0	1.0	1.0	0.900000	0.566729^o
14	-2.0	-1.0	1.0	1.0	0.733333	0.566729^o
15	-2.0	-1.0	1.0	1.0	0.357143	0.554762
16	-2.0	-1.0	1.0	1.0	0.215320	0.334641

recovers the exact potentials. Furthermore, we observe that all the obtained parameters are rational, even if the "true" parameters were not, due to the fact that in our model only an integer number of contacts is possible. It also explains the fact that for the target parameters $(-1, -1/\sqrt{2}, 0)$, the obtained parameter E_{PP} is invariably $-5/7$ for all sequence lengths $N = 11, .., 16$ although the cell changes dramatically. One would have to consider (much) longer sequences to get contact numbers high enough to generate a rational number closer to $-1/\sqrt{2}$. This insensitivity may be lifted in cases where the number of contacts is no longer integer, e.g. for real space proteins.

For this set of target parameters, we have also considered taking only those good sequences with a minimum gap larger than certain thresholds (i.e. 0.5 and 0.75), and although the obtained cells are larger (it scales with $(\text{mingap})^{N_p}$), the obtained parameters are unaltered until the cell ceases to be closed.

To get an idea of the performance of the algorithm as the dimension of parameter space increases, we did some checks on the following variations:

- ✔ a model with $N_p = 3$, with 2 kinds of amino acids as before ($N_a = 2$), but including a next to nearest neighbor (nnn) interaction for a H-P contact,
- ✔ a model with $N_p = 5$, with 2 kinds of amino acids ($N_a = 2$) and nn and nnn contacts,
- ✔ a model with $N_p = 5$, with 3 kinds of amino acids ($N_a = 3$) and only nn-contacts, and
- ✔ models with $N_p = 9$, with 4 kinds of amino acids ($N_a = 4$) and only nn-contacts (see § 2.4.2).

The quality of the obtained parameters is always as good as those shown in Tables 2.1, 2.2, 2.3 and does not depend on N_p.

To check the sensitivity of the method to a wrong choice of energy function, we have generated good sequences and structures using 6 interaction parameters ($N_p = 5$, both $N_a = 2$, nn- and nnn-contacts and $N_a = 3$, nn-contacts), and tried to satisfy all inequalities using fewer parameters, e.g. ignoring nnn H-P contacts. The method immediately indicated that the cell does not exist, and thus that the number of parameters was insufficient. On the other hand, putting in more free parameters than were used to generate the good sequences, the irrelevance of these parameters is immediately recognized and their obtained values are (very close to) 0.

2.4.2 Results for the 4 amino acids problem

The $E_{\mu\nu}$ matrix has 10 parameters (9 independent parameters, fixing the energy scale) in this case. Our tests have been carried out for four different sets of parameter values where each parameter is generated independently from a Gaussian distribution with mean -2 and variance 1. The length of the chain was 14. For each set of parameters, we have generated a MPDB of about 600 sequences and their corresponding (unique) native states. Furthermore, the sequences have an energy gap, Δ (the energy difference between the first excited state and the ground state) greater then 0.5. Indeed it is thought [25] that real proteins in order to have thermodynamical stability and short folding times should possess a pronounced global minimum on the potential surface. A comparison with one case where $\Delta > 2$ is also presented. The trial Hamiltonian is parameterized as the true one and we have chosen the energy scale by fixing to its exact value one of the most negative $E_{\mu\nu}$'s. The remaining 9 parameters are then determined maximizing the minimum gap using the method explained above. We have also verified that simulated annealing techniques are quite efficient for this case and give the same set of extracted potentials as the method used in § 2.4.1. Figures 2.4(a),(b),(c) and (d) show the extracted potentials versus the true ones. The extracted potentials are then tested for *new sets* of "good" sequences for each of the four cases to determine their ground state configurations over all possible self–avoiding chains of length 14. For all the sequences in the MPDB, we get full success (Figure 2.4).

Indeed, since the maximum gap has been calculated on a restricted set of conformations there is no guarantee that the good sequences used in the optimization procedure recognize their own native state among all possible conformations. Thus, it is important to test the extracted potentials using a new independent set of good sequences. In all four cases that we studied, at most 2 out of 628 do not found their original native state. The percentages of the correct determination of the native states using the extracted potentials are indicated in the Table.

We have tested the performance of the method as the size of the MPDB is decreased. Figure 2.5 shows how the percentage of success depends on the number N of sequences contained in the MPDB. Only three of the four cases are shown for clarity (the fourth case has the same behavior as the other three). The minimum N used is 14. Note that full success is almost reached for $N \sim 200 - 300$. For the first set of potential parameters, we have also generated a MPDB with an energy gap $\Delta > 2$. The results are shown in Figure 2.5 , and saturation is reached at about $N \sim 100$.

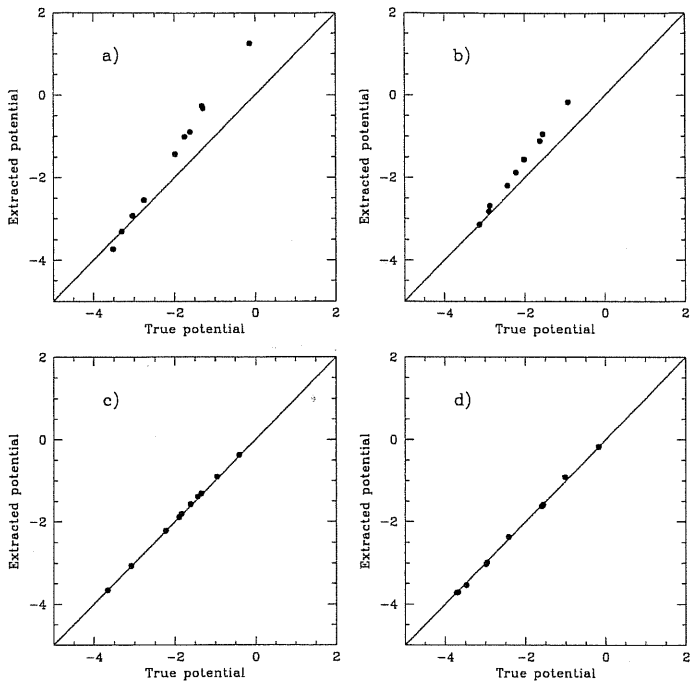

Figure 2.4: Derived potential versus true potential for the 4 amino acid problem (in units of k_BT). Results for the parameter set 1 (a), 2 (b), 3 (c) and 4 (d).

Table 2.4: Results for the 4 amino acid model. For each parameter set, the Table shows the size of the MPDB used for the derivation of the potential and the success rate in the correct prediction of the native state for each of the training set sequences.

parameter set	size of the MPDB	success
1	628	99.7%
2	716	99.9%
3	840	100%
4	798	100%

2.4.3 Comparison with other methods

The quasi-chemical method

The quasi-chemical method [27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41] is widely used in various forms for obtaining the effective potential between amino acids and to provide "scores" for candidate protein structures. Briefly, the procedure is as follows: from the data bank, one compiles the normalized probability density, $f_{A,B}(r)$, that two given amino acids are at a distance r from each other. The basic idea

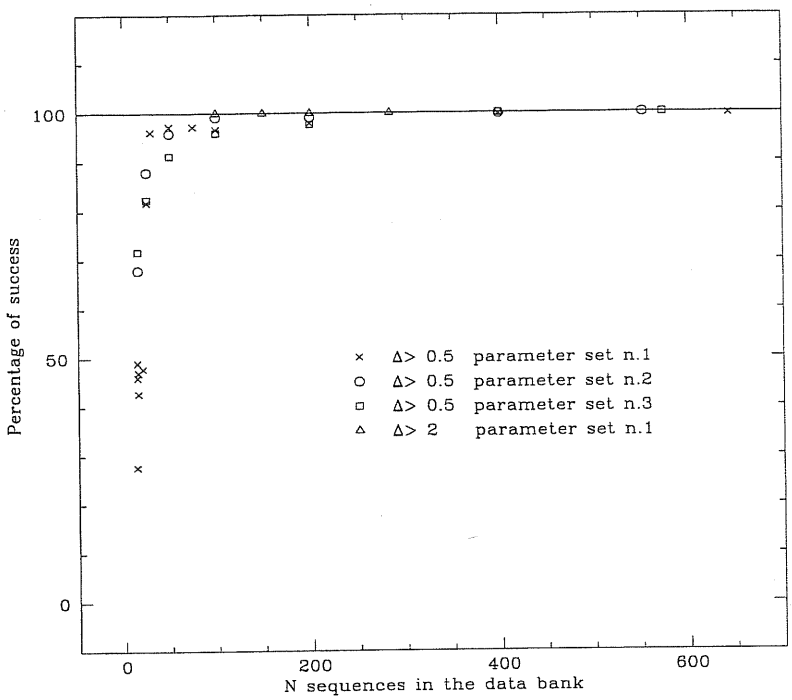

Figure 2.5: Effect of the MPDB size on the success rate using the extracted parameters to determine the ground state configurations for new sets of sequences, for different minimum energy gaps Δ (in units of k_BT). For the case with $\Delta > 2$ (open triangles), the cell is not closed for small N. When the cell is closed, the success rate is almost 100%.

is that if A and B attract each other strongly, they are more likely to be near each other compared to a random reference state of a non-interacting gas of amino acids with the same number of A and B residues. Conversely, if A and B dislike each other, they avoid each other a bit more than what one would expect from random considerations. This idea is then quantified in the form

$$E_{A,B}(r) \propto -kT \ln[f_{A,B}(r)] \,. \tag{2.12}$$

Additional considerations pertaining to how far apart two amino acids are along the sequence are sometimes introduced in order to capture the secondary structure propensities. The derived quantities such as $E_{A,B}$ are now interpreted as the energies of interaction between the amino acid pairs and used for determining which structure among many alternatives yields the most favorable value of the energy. As said before, Thomas and Dill [41] tested the quasi-chemical method for the H–P model and the results are shown in Table 2.5. They have to be compared with the corresponding cases in Tables 2.1, 2.2, 2.3 obtained by our method.

Table 2.5: Summary of the results of TD [41] using the Miyazawa-Jernigan scheme [29], for a sequence length of 14 monomers. The Table shows the true parameters, the parameters obtained by TD and the success rate in the prediction of native conformations of the "good" sequences with the obtained parameters.

True			TD test			
E_{HH}	E_{HP}	E_{PP}	E_{HH}	E_{HP}	E_{PP}	success
-5	-4	-1	-5	-3.0	+0.8	74%
-5	-1	-2	-5	-1.1	-2.1	100%
-5	-5	-1	-5	-3.7	+1.4	84%
-5	-3	+1	-5	-2.6	+2.5	96%
-5	-3	-1	-5	-2.4	0.0	64%

The explanation for deriving the interaction energy from the observed pairing frequency invokes the Boltzmann's principle and has also been called the Boltzmann device. Boltzmann statistics pertains to the occupation probabilities for the energy levels of an individual system. Thus, if a system has energies E_0, E_1, E_2, etc., the probability that the system has an energy E_i is proportional to $\exp[-\frac{E_i}{kT}]$.

There are several critiques that one can move to the quasichemical method. First, the native structures of distinct sequences of amino acids do not correspond to the excitations of a single system. Instead, each of the sequences is a separate system, whose native state structure is known from experiment. Thus, the basic premise of the method is wrong. Second, even making the assumption that Boltzmann statistics did hold, there is no simple relationship between the observed pairing frequency and the energy of interaction as envisaged by expression (2.12). The role of temperature in eq. (2.12) is unclear, because the native states of each of the sequences correspond to their ground states or equilibrium states at $T = 0$. Third, the quasichemical method relies on a reference state – the observed pairing frequencies are compared to those expected in this reference state in order to determine whether two amino acids like each other and by how much. Often, this reference state is chosen as a noninteracting gas made up of all the amino acids constituting all the sequences with known native structure. This does not seem to have a physical basis, because the sequences are all distinct entities and do not originate from a common soup of amino acids.

These difficulties with the quasichemical method, which were already partially recognized in the literature (see [41] and references therein), are avoided in our strategy. The sequences whose structures are known are analogous to quenched variables in statistical mechanics while the conformations that a given sequence can adopt, are

the analog of annealed variables. A thermodynamic average can be performed over the annealed variables but not over the quenched ones. We used Boltzmann statistics but for each sequence separately. We dealt with the energies directly and not with a derived quantity such as the pairing frequency. Indeed, our strategy embodies the complete information in the system and, in principle, has information not only about pairing frequencies but also triplet and higher order correlations. Our method does not rely on a reference state and the role of temperature is well-defined.

Mirny and Shakhnovich's method

Recently, Mirny and Shakhnovich (MS) [52] have proposed to use the Z-score, which is a measure of how pronounced the energy minimum corresponding to the native state is, to carry out potential derivation. The Z-score is given by:

$$Z(\sigma) = \frac{\mathcal{H}(\Gamma_\sigma, \sigma) - \langle \mathcal{H} \rangle}{\mathrm{var}(\mathcal{H})} \,. \tag{2.13}$$

where the average of $\mathcal{H}$, $\langle \mathcal{H} \rangle$, and its variance, $\mathrm{var}(\mathcal{H})$, are computed for a set of alternative (decoy) conformations. The method of [52] entails the minimization of the cost function (harmonic mean)

$$\langle Z \rangle_{\mathrm{harm}} = \Big(\sum_{\sigma \in \Omega_\sigma} Z(\sigma)^{-1} \Big)^{-1} \,. \tag{2.14}$$

Indeed, real potential are believed to distinguish the native structure by making its energy much lower than energy of all other conformations. Therefore it is reasonable that an essential property of the folding potential is that the energy of a native sequence folded into its respective native conformation should be much lower than the energy of this sequence in every alternative, misfolded, conformation [25, 53, 54].

For each conformation Γ of the ensemble Ω_σ, the average $\langle \mathcal{H} \rangle$ and the variance $\mathrm{var}(\mathcal{H})$ are calculated in an ensemble of phantom conformations with the same number of residue-residue contacts as in Γ and with the assumption that these contacts occur independently of each other. This approximation will be discussed further later on. One can repeat these calculations for our cases.

We have implemented this method using expressions (2.13) and (2.14), calculating $\langle \mathcal{H} \rangle$ and $\mathrm{var}(\mathcal{H})$ exactly for each sequence using the structures of the MPDB.

In order to have a finite minimum of $\langle Z \rangle_{\mathrm{harm}}$, it is necessary to fix the variance, or equivalently one of the interaction parameters like we did. MS also fix the average

Table 2.6: Results for the H-P model, using the Z-score of MS [52], fixing both the variance and the average of the interaction parameters to their true values. The Table displays the sequence length, the derived interaction parameters, the total number of good sequences, the number of sequences for which the predicted ground state is wrong, and the success rate.

true	-1.0	-0.707107	0.0			
N	E_{HH}	E_{PP}	E_{HP}	$\#_{tot}$	$\#_{wrong}$	success
12	-0.969868	-0.747827	0.010588	728	4	99.45%
13	-0.990930	-0.719754	0.003577	750	0	100.0%
14	-0.977091	-0.738392	0.008377	2005	26	98.70%
15	-0.963963	-0.755398	0.012254	4302	77	98.21%
16	-0.972729	-0.744113	0.009735	8892	151	98.30%

Table 2.7: Same as Table 2.6, but only fixing the variance of the interaction parameters.

true	-1.0	-0.707107	0.0			
N	E_{HH}	E_{PP}	E_{HP}	$\#_{tot}$	$\#_{wrong}$	success
12	-0.655207	-0.330015	0.474504	728	124	82.97%
13	-0.841471	-0.317697	0.568875	750	198	73.60%
14	-0.745093	-0.317782	0.526300	2005	675	66.33%
15	-0.657765	-0.327509	0.477554	4302	1754	59.23%
16	-0.740636	-0.328271	0.517735	8892	3274	64.00%

potential, which requires more information, which, a priori, one does not have. In the case of many interaction parameters, however, this might not be so crucial. Furthermore, in their implementation, MS do not explicitly require that all the $Z(\boldsymbol{\sigma})$'s be negative, since $\langle\mathcal{H}\rangle$ and var$(\mathcal{H})$ are approximated. For the H-P model, we first require that $Z(\boldsymbol{\sigma}) < 0$ for all $\boldsymbol{\sigma}$ in the MPDB, and then we minimize $\langle Z\rangle_{\rm harm}$.

For the 4 amino acid case, the minimization of eq. (2.14) leads to spurious minima corresponding to $\sum_{\boldsymbol{\sigma}} Z(\boldsymbol{\sigma})^{-1} \simeq 0$, since both positive and negative $Z(\boldsymbol{\sigma})$'s appear. This happens both with the exact $\langle\mathcal{H}\rangle$ and var$(\mathcal{H})$, and with the approximations of MS [52].

Since the search of the parameter domain where all the $Z(\boldsymbol{\sigma})$'s are negative was impractical in this case, we have modified eq. (2.14) to $\langle|Z|\rangle_{\rm harm} = (\sum_{\boldsymbol{\sigma}} |Z(\boldsymbol{\sigma})|^{-1})^{-1}$The physic meaning of the procedure is unchanged by this modification, since the harmonic mean is used in order to weight more the terms with the smallest Z scores, and this is accomplished also using the Z score absolute values.

Tables 2.6, 2.7 show the results for one of the H-P cases we have considered before, i.e. $E_{HH} = -1$, $E_{HP} = -1/\sqrt{2}$ and $E_{PP} = 0$. Table 2.6 corresponds to the case where we have fixed both the variance and the average of the E's, leaving only one

parameter to be determined. Table 2.7 shows the results when only the variance of the interaction parameters is fixed to set the energy scale, and two parameters are left to be determined, as in our method.

With the parameters obtained from the minimization we checked how many of the good sequences of the MPDB still have their unique ground state in the correct conformation among all the possible conformations obtained by exact enumeration. In contrast to our method, not all of the sequences in the MPDB find the correct native state, as can be seen in the Tables. Note that neither the success rate, nor the values of the potentials are monotonic as a function of the chain length, at least within the small range of lengths used.

Table 2.8 shows the results for the 4 amino acid case for the same four sets of potential parameters used to test our method. The variance and the average potential have been fixed to their exact values (thus there are 8 free parameters and not 9 as in our case). For one parameter set we have also implemented the MS optimization method. MS use the following expressions for $\langle \mathcal{H} \rangle$ and var($\mathcal{H}$) (see the discussion following eq. (2.14)):

$$\mathcal{H} = \sum_{i<j} E_{\mu_i,\mu_j} \langle \Delta_{i,j} \rangle \tag{2.15}$$

$$\mathrm{var}(\mathcal{H}) = \sum_{i<j} \sum_{k<l} E_{\mu_i,\mu_j} E_{\mu_k,\mu_l} T_{ij,kl} \tag{2.16}$$

with the average density of contacts between residues i and j

$$\langle \Delta_{ij} \rangle = \frac{n}{n_{tot}} \tag{2.17}$$

and the contact correlator

$$T_{ij,kl} = \langle \Delta_{ij}\Delta_{kl} \rangle - \langle \Delta_{ij} \rangle \langle \Delta_{kl} \rangle = \begin{cases} \frac{1}{n_{tot}^2} & (i,j) \neq (k,l) \\ \\ \frac{1}{n_{tot}} - \frac{1}{n_{tot}^2} & (i,j) = (k,l) \end{cases} \tag{2.18}$$

where n is the number of contacts in the native conformation, n_{tot} is the total number of the topologically possible contacts and the indices $i, j, \ldots$ run from 1 to the length of the chain (14 in our case). Assuming that contacts in the alternative conformations are distributed independently and uniformly and that the number of contacts is the same as in the native conformations (MS hypothesis), we found a different expression for $T_{ij,kl}$.

The second term of $T_{ij,kl}$ is trivially:

$$\langle \Delta_{ij} \rangle \langle \Delta_{kl} \rangle = \langle \Delta_{ij} \rangle^2 = \frac{n^2}{n_{tot}^2}. \tag{2.19}$$

The evaluation of the first term of $T_{ij,kl}$ is easily done by estimating in two different way the sum over each pair lm:

$$\sum_{lm} \langle \Delta_{ij} \Delta_{kl} \rangle = n \langle \Delta_{ij} \rangle \tag{2.20}$$

and

$$\sum_{lm} \langle \Delta_{ij} \Delta_{kl} \rangle = \langle \Delta_{ij} \Delta_{ij} \rangle + \sum_{lm \neq ij} \langle \Delta_{ij} \Delta_{kl} \rangle = \langle \Delta_{ij} \rangle + \sum_{lm \neq ij} \langle \Delta_{ij} \Delta_{kl} \rangle \tag{2.21}$$

By equalling eq. (2.20) and eq. (2.21) the expression for $T_{ij,kl}$ is obtained:

$$T_{ij,kl} = \begin{cases} \frac{n(n-1)}{n_{tot}(n_{tot}-1)} - \frac{n^2}{n_{tot}^2} & (i,j) \neq (k,l) \\ \\ \frac{n}{n_{tot}} - \frac{n^2}{n_{tot}^2} & (i,j) = (k,l) \end{cases} \tag{2.22}$$

The results corresponding to both assignments, eqs. (2.22) and (2.18), are also reported in Table 2.8 and should be compared with the results of our method in Table 2.4. Figures 2.6(a), (b), (c) and (d) are the analogs of Figure 2.4(a), (b), (c) and (d) for the MS method. Figure 2.6(a) shows the extracted potentials using both the exact $\mathcal{H}$ and var($\mathcal{H}$) and the approximation of eqs. (2.17) and (2.22) (which according to Table 2.8 works better than the MS one, i.e. eq. (2.17) and (2.18)) for parameter set 1.

Crippen's method

The method introduced by Crippen in [42], as explained in detail in [55], entails in the search of interaction parameters satisfying a set of inequalities like those in eq. (2.6), but with the right hand side substituted by a suitable positive number M, e.g.:

$$\vec{E} \cdot \vec{X}(\Gamma, \boldsymbol{\sigma}) > M \tag{2.23}$$

for each sequence in the MPDB and all the alternative conformations. Expression (2.23) is the requirement that a sequence has lower energy in its native conformation than in any alternative conformation (as already explained in § 2.3.1).

Our first claim is that the parameter M is useless and can be chosen equal to zero. Indeed if $(\vec{E})$ satisfies all the inequalities (2.23) with $M = 0$ and $|F_{\text{gap}}(\vec{E})|$ denotes

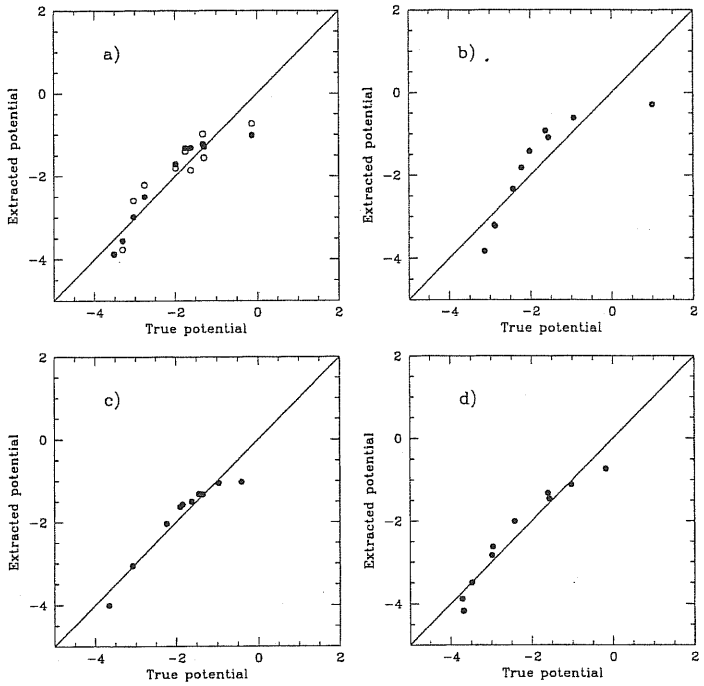

Figure 2.6: Derived potential versus true potential (in units of k_BT), for the 4 amino acid problem using the MS method. Results for the parameter set 1 (a), 2 (b), 3 (c) and 4 (d). Figure (a) also shows the results using the approximation of eqs. (2.17) and (2.22) (open circles).

the minimum value of the left hand side of eq. (2.23) (i.e. $|F_{\rm gap}(\vec{E})|$ is the smallest gap introduced in eq. (2.9)), then the interaction parameters:

$$\vec{E}' = k\frac{M}{|F_{\rm gap}(P)|}\vec{E} \tag{2.24}$$

satisfy eq. (2.23) for any $k > 1$ and $M > 0$. Thus the parameter M does not play any role in the procedure of potential extraction. As explained in § 2.3.2, one can then fix the energy scale fixing one parameter as done in the previous section.

Now we rediscuss explicitly the example of Crippen [55] with two amino acids on a two dimensional square lattice with $E_{11} = -1$, $E_{12} = 0$ and $E_{22} = 1$. Fixing $E_{11} = -1$ inequalities (2.6) for various chain lengths n up to 10 give the cells shown in Figure 2.7. All the cells are open and there is no way that using eq. (2.23) one can choose a particular point inside them to represent the true interaction parameters. The maximum-minimum gap method (as explained in § 2.3.2), for chain length from 8 to 10 gives $E_{12} = 0$ and $E_{22} \geq 1$ (straight line in Figure 2.7). This is due to the fact

Table 2.8: Results for the 4 amino acid model, using the MS method [52], fixing both the variance and the average of the interaction parameters to their true values. The Table shows the identification of the parameter set, the size of the MPDB used for the derivation of the potentials, the number of failures in the prediction of the correct native state, and the success rate. In the top 4 lines, the exact $\langle E \rangle$ and var(E) are used, whereas for the fifth line the approximations (2.17) and (2.22), and for the sixth line the approximations (2.17) and (2.18) are used.

parameter set	$\#_{tot}$	$\#_{wrong}$	success
1	628	64	89.8 %
2	716	80	88.2 %
3	840	14	98 %
4	798	96	88 %
1 (using eq. 2.22)	628	71	88 %
1 (using eq. 2.18)	628	105	83 %

that no contact between amino acid of type 2 can be found in the native state of chains of length up to 10. Thus, the extracted values reported by Crippen [55] are the result of the particular method applied to solve the set of inequalities (2.23). Indeed, his iterative procedure stops as soon as a set of parameters satisfying eq. (2.23) is found. In such a way the extracted values will crucially depend on the starting condition of the search. Of course, if all the data (and hence all the possible inequalities) of the (M)PDB have been used, any point inside the remaining cell, as determined by any method, is, a priori, equally good. In realistic cases, however, when the complete data is not available (e.g. in the case that not all the alternative (decoy) conformations can be generated), the point inside the cell, as determined by our method seems to be the best guess.

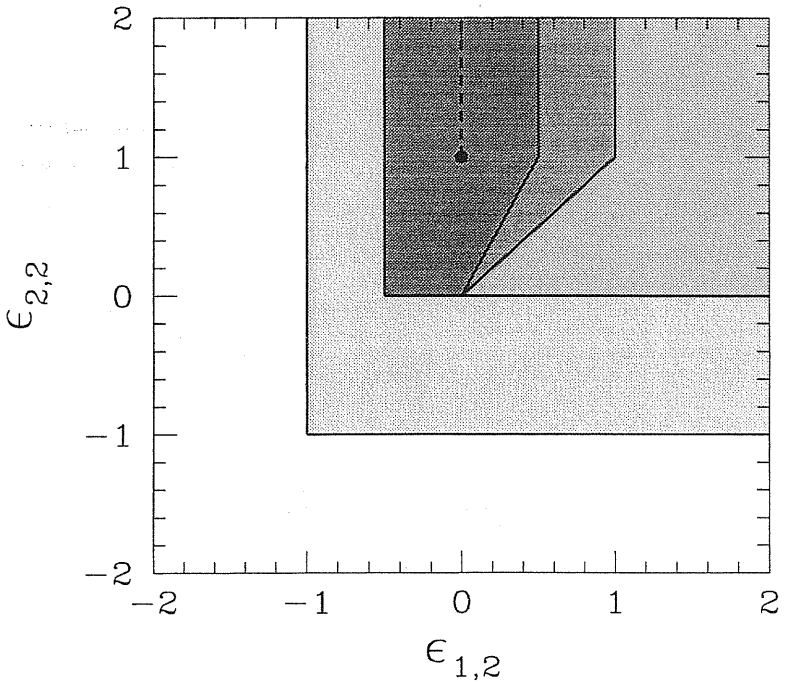

Figure 2.7: Results obtained solving exactly the inequalities for different chain length n with Crippen's parameters. The value of the parameter $\epsilon_{1,1}$ is fixed to his true value. For length $n = 4$ the parameters $E_{1,2}$ and $E_{2,2}$ are completely undetermined, each point on the plane satisfies the full set of inequalities. For length $n = 5$ the MPDB is empty. For length $n = 6$ and $n = 7$ each point in the region $E_{1,2} > -1$, $E_{2,2} > -1$ satisfies (lightest gray region). For length $n = 8$ the region becomes the one included in $E_{1,2} > -0.5$, $E_{2,2} > 0$ (darker gray region). For $n = 9$ each point such that $-0.5 < E_{1,2} < 1$, $E_{2,2} > 0$ and $E_{2,2} > E_{1,2}$ is a solution (even darker gray region). Then, for $n = 10$ the region shrinks to that included in $-0.5 < E_{1,2} < 0.5$, $E_{2,2} > 0$ and $E_{2,2} > E_{1,2}$ (darkest gray region). The dot in the picture represents the true value of the parameters. The dashed line is the solution for $n = 10$ using our method (maximization of minimum gap). The value obtained for the parameter $E_{1,2}$ is the exact one whereas $E_{2,2}$ remains still undetermined, e.g. all $E_{2,2} > 1$ are completely equivalent as far as MPDB sequences have to recognize their own native states.

CHAPTER 3

OFF–LATTICE DYNAMICS OF MODEL HETEROPOLYMERS

In the previous Chapter we presented the problem of the determination of effective potential among various parts of the peptide chain and we discussed studies of this problem performed on–lattice.

Lattice models, using virtual bonds of fixed distance and fixed angle, are a more efficient way to perform a first test of a method than it is working on real proteins [25, 56, 57, 53, 58, 26, 28, 59, 60]. This was the motivation of the studies on extraction of interaction potentials previously discussed. In lattice models protein chains are modeled as self–avoiding walks on a lattice and contact between amino acids are defined for neighboring sites in space (but not subsequent along the sequence). The inter–amino acid interaction potential is, thus, completely determined when the values of the energy gain $\epsilon_{i,j}$, for the contact between each pair of amino acids, i and j, are known. Then, a finite set of parameters $\epsilon_{i,j}$ defines the interaction potential on the lattice. Moreover, at least in principle, one may take into account all the possible conformations as alternative structures. For sufficiently small values of the chain length,

or for sufficiently long computational time, this can be accomplished by exhaustive exact enumeration.

For real proteins the situation is more complicated. If one examines the distribution of distances between two real amino acids types, i and j, this is far from being a delta function, as it is assumed in lattice models. Moreover, shape and width of the distribution are strongly dependent on the types of amino acids involved in the contact. There is a range of possible contact distances between amino acids. There are several reasons for this behavior. The first is that the interaction potential between amino acids is a function defined on the continuous 3-d space. The shape of the potential function depends on the types of amino acid involved. For instance, one can anticipate that the depth of the minimum of the potential function is related to the hydrophobicity of amino acids (that is, roughly speaking, a measure of the tendency of an amino acid to bury itself among the others), and that the equilibrium distance of the bond is related to the volume of amino acids involved in a contact.

This Chapter presents a summary of the results of a comprehensive and unified study of the above issues for an off–lattice model of heteropolymers. We discuss here the first step to the final goal of applying the method for extracting the interaction potential of real proteins. Our study provides an important test of the feasibility of the implementation of various strategies in a realistic, albeit simplified framework. Starting from these results, in Chapter 5 we will present an improved, more realistic off–lattice model and we will apply it to the determination of effective interaction potential for a real protein (crambin).

3.1 The off–lattice model

The simplest off–lattice protein model can be obtained by assuming that a protein may be represented by a chain of N beads (see for instance [61, 62, 63]). For a real protein, the beads may, for example, represent the C_α atoms of the amino acids.

A conformation of a chain made up of N residues is defined by the coordinates $\mathbf{r}_1, \ldots, \mathbf{r}_N$ of beads in the three-dimensional space. We consider only effective two–body forces between amino acids obtained by integrating out the degrees of freedom associated with the internal coordinates of each residue and the solvent. A simple choice for the interaction potential is:

$$V_{ij} = \delta_{i,j+1} f(r_{i,j}) + \eta((\frac{\sigma}{r_{ij}})^{12} - (\frac{\sigma}{r_{ij}})^{6}), \tag{3.1}$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the inter-residue distance. The parameter η entering this equation controls the energy scale, whereas σ determines the interaction length between monomers. The values of σ and η have to be adjusted to fit both the complex interactions between the various groups of amino acids and the interactions with the solvent. Furthermore, these parameters could depend on the different types of amino acids involved in the interaction.
The energy function of the C_α–C_α virtual bond is chosen to be

$$f(x) = a(x - d_0)^2 + b(x - d_0)^4, \tag{3.2}$$

with a and b taken to be 1 and 100 respectively. We add a quartic term to the usual quadratic one [64] because a plain harmonic potential could induce energy localization in some specific modes, significantly increasing the time needed for equilibration. We have chosen the simplest generic interactions that maintain the connectivity of the chain, provide for an attractive interaction between monomers and yet respect self-avoidance.

The parameter d_0, which represents the equilibrium distance of the nearest neighbors along the chain is set equal to 3.8 Å, the experimental value for the mean distance between nearest neighbor C_α atoms along the chain in real proteins, as determined from the PDB.
The Hamiltonian is given by:

$$H = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2} + \sum_{i=1}^{N} \sum_{j>i} V_{i,j}. \tag{3.3}$$

The first term is the classical kinetic energy of the chain, where the $\mathbf{p}_i$'s are the canonical variables conjugated to the $\mathbf{r}_i$'s.

We have used Molecular Dynamics (MD) (entailing the integration of Newton's laws of motion on a computer) for simulating the kinetics of the chains. We employed an efficient and precise symplectic algorithm [65], in which one varies the energy density $\epsilon = E/N$, which is related to the temperature [66].

3.2 Construction of a simple Data Bank

On lattice, one usually selects a compact conformation and attempts to design a sequence that has this structure as a thermodynamically stable ground state. Off–lattice, there are an infinite number of conformations, almost all of which are not designable

(i.e. there is no sequence which has them as its ground state). Our first goal was to generate a number of compact, designable conformations. To accomplish this task, we started by considering an homopolymer model (just one kind of amino acid) with overall attractive interactions between pairs of monomers. For the homopolymer case, we fixed the parameters $\eta = 40$ and $\sigma = 6.5$ Å.

Such a value for σ ensures that, in practice, two monomers significantly interact when their distance is smaller then 9 Å. This distance threshold is conventionally used for the bond between amino acids and is determined by the requirement that the average number of C_α–C_α contacts for each amino acid is roughly equal to the respective number obtained with the all–atom definition of contacts [67]. Starting with different initial conditions we find different low energy configurations for the chain, since the minimum energy state of an homopolymer is largely degenerate. At the initial time we generated random configurations of beads in the space, constraining the distance between two neighboring beads to be equal to $d_0 + \Delta d\,(\Delta d < d_0)$. Random configurations of this kind were used to initialize the numeric integration of Hamiltonian equations of motion.

For a three dimensional homopolymer made up of 30 identical monomers, twenty compact conformations (the radii of gyration varied between 7.52 and 7.59 Å, as it is shown in Table 3.1) with low-lying energies were obtained performing MD simulations in a slow-cooling way. The energies were typically distributed over a range of $1.5 \times \eta$ and the energy spectrum does not show any peculiar difference with respect to the one obtained for the Lennard–Jones clusters [68, 69].

The system was equilibrated for 8000 time steps after successive cooling on lowering the temperature each time by a factor of 0.8, as customary in simulated annealing procedure [70].

The compact conformations were chosen so that they had little structural overlap with each other. The distance D between two 3-dimensional conformations is given by

$$D = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_i^{\cdot})^2}, \tag{3.4}$$

where one structure is translated and rotated to get a minimal D [71]. Two conformations were assumed to be different if the D between them exceeded 1 Åbased on the experimental resolution of protein structures [72].

The mean distance between a given conformation and the 19 others ranged between

Table 3.1: Gyration radius R_g, mean distance $\langle D \rangle$ to the others for each and energy E of the 20 compact configuration obtained as homopolymer low energy states.

Config. number	R_g	$\langle D \rangle$	E
1	7.55	7.00	-1591.28
2	7.53	7.01	-1592.98
3	7.52	6.72	-1613.72
4	7.59	6.94	-1571.50
5	7.53	6.45	-1585.34
6	7.57	7.49	-1551.46
7	7.53	6.67	-1592.80
8	7.56	7.04	-1588.29
9	7.57	7.34	-1572.96
10	7.53	7.33	-1568.36
11	7.56	6.69	-1611.00
12	7.52	6.71	-1613.72
13	7.56	6.86	-1612.20
14	7.58	6.86	-1583.46
15	7.58	6.79	-1550.76
16	7.53	7.21	-1592.97
17	7.59	6.80	-1555.87
18	7.54	7.30	-1575.28
19	7.58	6.95	-1561.95
20	7.57	7.33	-1588.63

6.67Åand 7.49Å. For each of the 20 compact configuration we report in Table 3.1 the value for the gyration radius (defined as $R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \sum \frac{\mathbf{r}_j}{N})^2}$), the mean distance $\langle D \rangle$ with respect to the others and the energy.

All the resulting structures show secondary motifs, especially helices (see Figure 3.1). The appearance of secondary motifs is not a general phenomenon but it is linked to the our choice of the length parameters σ and d_0. The relation between the ratio σ/d_0 and the nature of the conformation of low-lying energy states is at present under investigation, and will be discussed elsewhere [73].
For the homopolymer chain, the 20 compact configurations obtained in the first step (§ 3.2) are energetically equivalent.

3.2.1 The design

Next we switched to consider an heteropolymer model employing different types of amino acids and we selected a set of suitable sequences, each of them having one of

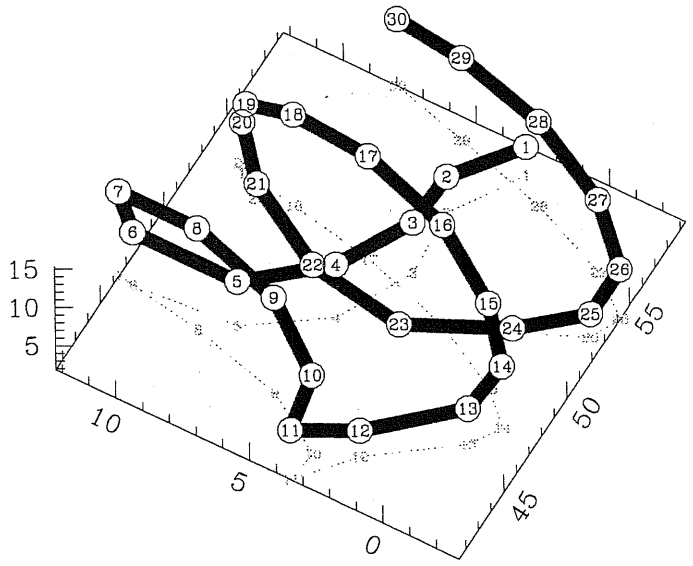

Figure 3.1: One of the compact structures obtained for the homopolymer model. Note that the structure exhibits a helix having a complete turn in 10–12 beads, whereas in naturally occurring proteins, helices have 3.6 residues per turn. The designed sequence for this conformation is: 3 1 1 1 1 4 4 4 2 3 4 4 3 1 2 2 1 2 2 2 4 4 3 4 3 3 3 3 3 4. There are 65 contacts having a σ value for equilibrium distance equal to 6.25 Å, 41 with σ equal to 6.5 Å, 12 with 7 Å, 15 with 7.5 Å, and 48 with 8 Å. The buried positions in the structure are occupied by the hydrophobic monomers.

these configurations as its native state, i.e. as a structure much more energetically convenient than all possible alternatives.

We used four kinds of amino acids, two of which were strongly hydrophobic and thus mutually strongly attractive (on integrating the solvent degrees of freedom) while the other two were weakly attractive. The composition of the heteropolymer was constrained: each sequence comprises 6 amino acids for each of the two more hydrophobic types, and 9 amino acids of the third and fourth kind.

The interaction potential has the same analytic form as before, but to account for all possible interactions among the four different hydrophobicity types the Lennard–Jones interaction is characterized by a set of ten parameters, that we chose by hand, imposing the constraint that the inequalities $2\eta_{i,j} < \eta_{i,i} + \eta_{j,j}$, $(i \neq j)$, were satisfied. In this way the clustering of amino acids of the same hydrophobicity degree is favored

Table 3.2: Values of parameters η of the Lennard Jones interaction potential used in the heteropolymer model for the four types of amino acids.

residue type	1	2	3	4
1	40	30	20	17
2	30	25	13	10
3	20	13	5	2
4	17	10	2	1

[74]. The values of matrix elements $\eta_{i,j}$ are reported in Table 3.2.

Small variations in the Lennard–Jones length parameter are now permitted, considering 5 possible values of σ equal to 6.25, 6.5, 7.0, 7.5 and 8.0 Å, to take approximately into account the diversity of sizes of amino acids. This is in contrast with the homopolymer model that was amino acid-type independent (only the 6.5 Åvalue was employed), and plays an important role in the stabilization of the target native states.

A simplified design procedure due to Shakhnovich and Gutin (SG) [75] was carried out and entailed an optimal assignment of amino acid type to each monomer and independently a choice of the σ parameter for each $i-j$ pair with $|i-j|>1$. The SG procedure consists in performing a Monte Carlo algorithm in sequence space. Starting with a random sequence, at fixed composition, one has to perform random permutations accepting or rejecting new sequences with respect to the Boltzmann factor $P=\exp(-\Delta E/k_BT)$, where ΔE is the energy variation due to permutation and T is the "temperature" parameter of the Monte Carlo optimization scheme. By slowly lowering the parameter T the sequence that has the lowest energy in the target conformation is selected.

Figure 3.2 is a schematic summary of the construction of a off–lattice model data bank.

The design procedure was carried out for each of the twenty conformations obtained from the homopolymer model, and was validated by detailed simulations which showed that the designed sequences do indeed have the target conformations as their ground states.

We slowly cooled each designed sequence several times (typically 50) starting from different random initial conditions. From this procedure, we confirmed that the target conformations are indeed the lowest energy structures. These cooling simulations also generate a set of alternative, higher energy, metastable conformations ($2-5$ for each sequence) that, when perturbed, "decay" to the global minimum conformation (the

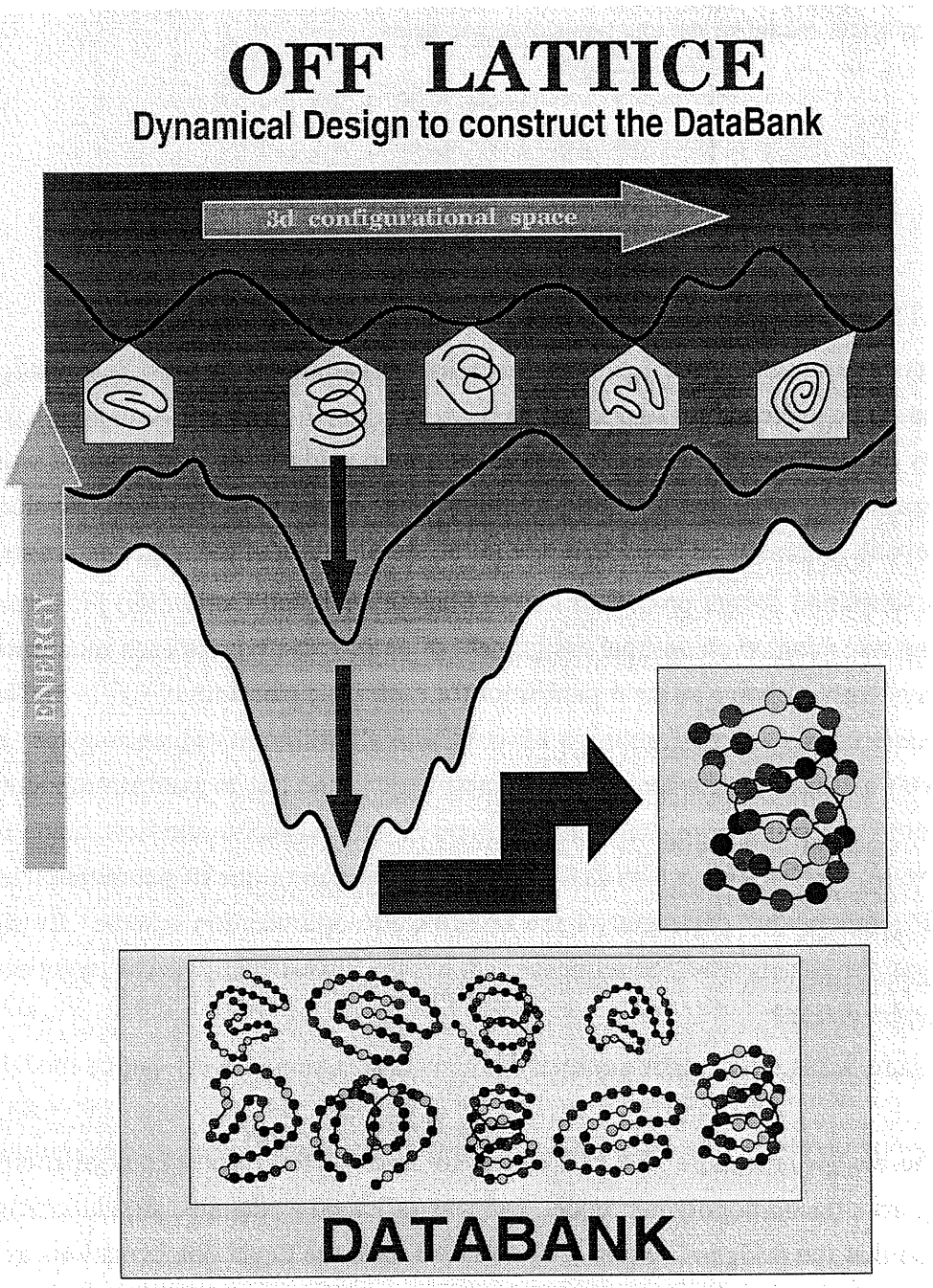

Figure 3.2: Construction of a toy-model PDB off lattice. We initially select compact configurations, energetically equivalent, then we design an appropriate sequence on each configuration, having for it a deep energy minimum.

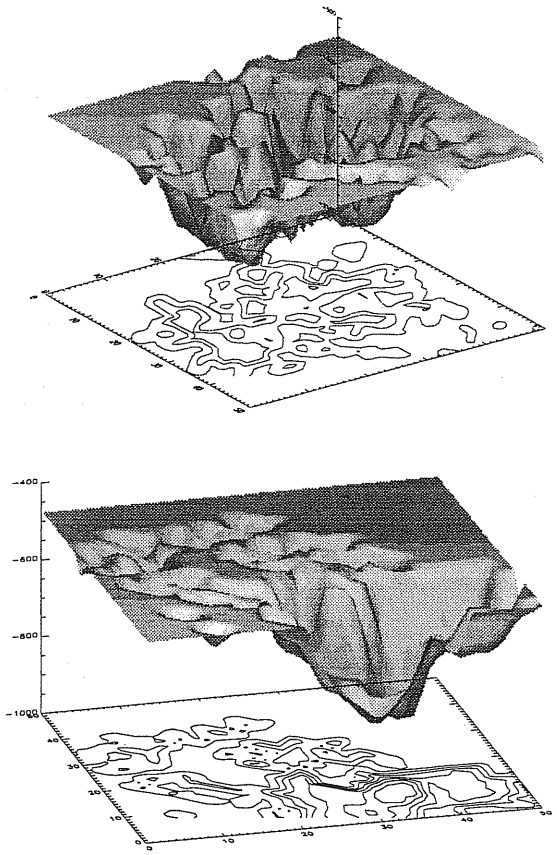

Figure 3.3: The energy landscape of the homopolymer (a) and of one designed sequences (b). Pictures are derived from the conformations obtained during numerous dynamical runs of slow cooling. The energy of each conformation is plotted as a function of its distance (see eq. 3.4) from two fixed "reference" conformations.

target structure). The energy landscape is modified by the design procedure and a folding funnel, that promotes thermodynamic stability and kinetic accessibility, is created. Figure 3.3 shows the energy landscape for the two cases of an homopolymer and one of the designed sequence. At variance with the homopolymer case, where the landscape has a multi–minimum structure [76], the designed sequence exhibits a unique pronounced minimum with a funnel shape [17, 18, 22, 19, 20, 14].

Let us stress that the different values allowed for the interaction equilibrium distance (i.e. parameter σ) has been an important ingredient on the construction of an

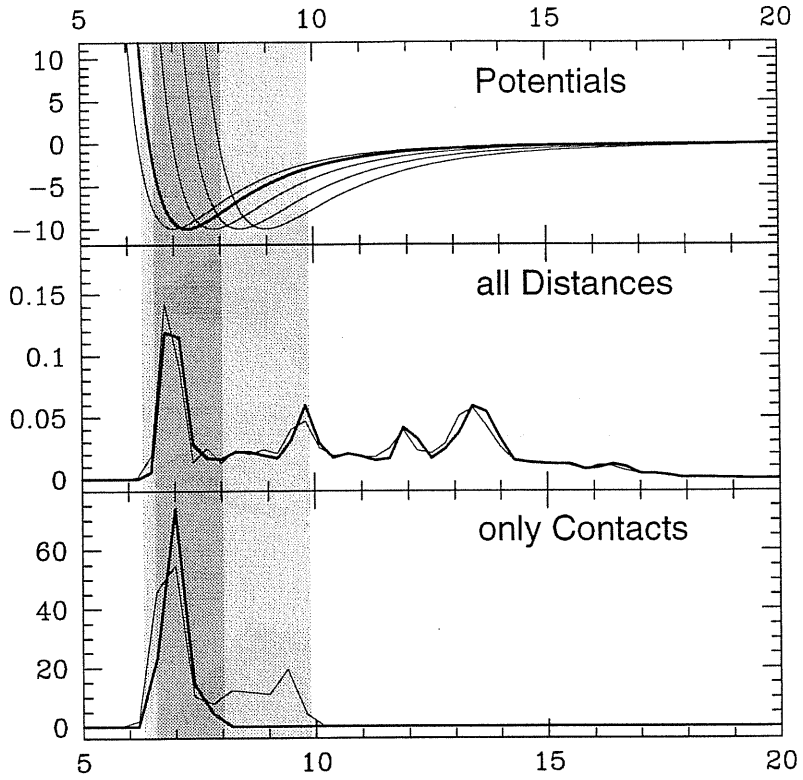

Figure 3.4: In the bottom panel of the Figure the original distance distribution computed for each pair of beads in low energy states of homopolymer (thick line) and in the configurations obtained – by MD simulations – as the heteropolymer native states (thin line) are compared. They are very similar because the native states are very close to the initially selected compact configuration of homopolymers. Indeed the design procedure was explicitly built up in order to have a target configuration as native state. Nevertheless, in the homopolymer case only those pairs lying in a narrow range around the unique allowed equilibrium distance give a significant (negative) contribution to the energy, while the variety of equilibrium distances for heteropolymers allows a greater number of pairs to stay on their maximally convenient energetic state. In the top panel of Figure 3.4 the potential shape for the homopolymer (thick line) is shown together with all the variations allowed in the heteropolymer (thin lines).

optimally designed sequence (i.e. with a funnel-like energy landscape). By moving the equilibrium distances, the resulting final sequences (i.e. after completion of the design procedure) were allowed to have a larger number of contacts (i.e pairs lying close to the bottom of their interaction potential well) with respect to the case of the homopolymer potential, for the same spatial configuration. This in turn stabilize the designed native state, by making it even more convenient. This is summarized in Figure 3.4.

3.3 Determination of potential parameters

We then set about to determine the effective parameters of the potential of interaction between the four kinds of amino acids, using the knowledge of the test bank of sequences and their known native structures.

In the previous Chapter we exposed a procedure for the determination of effective potential energies of interaction and extensively tested it on lattice models. The basic idea of the procedure is to require that the energy of the designed sequences is less in their ground states than it is in all the alternative conformations. This is simply a consistency requirement for the definition of a ground state (native state).

The method selects the optimal parameters such that the smallest energy gap (chosen among the set Ω_s of sequences $\{\sigma_s\}_{s=1,...,20}$ in the training set) between the energy of the sequence in its native conformation Γ_s^n and the (higher in energy) minimum energy alternative one is as large as possible. This additionally promotes thermodynamic stability.

This condition may be implemented by defining a cost function F_{gap}:

$$F_{gap} = -min_{(\sigma_s \in \Omega_s)} min_{(\Gamma \neq \Gamma_s^n)} \frac{E(\Gamma, \sigma_s) - E(\Gamma_s^n, \sigma_s)}{|E(\Gamma_s^n, \sigma_s)|} \tag{3.5}$$

and by choosing the parameter values of the potential in order to minimize eq. (3.5). Note that this is a slightly modified version of equation (2.9) in Chapter 2. The key new feature is the presence of the denominator $|E(\Gamma_s^n, \sigma_s)|$ which serves to rescale the energy gap associated with a given sequence with respect to its ground state energy. By introducing this denominator, the cost function (3.5) is now adimensional.

As the trial potential function for the interactions between amino acids we adopted the generic Lennard–Jones form, with unspecified parameters η and σ. One of the η values was fixed to its true value in order to set the energy scale. The use of the Lennard–Jones function is a great simplification. On the other hand a Lennard–Jones potential is a reasonable approximation to represent intermolecular interaction ([77], we will tell more about it in the next Chapter).

Iterative procedure

In lattice models amenable to exact enumeration, all conformations other than that of the native state, can be conveniently used as alternative or decoy conformations. In an off–lattice model, there are potentially an infinite number of such decoy conformations.

A key ingredient for the success of the potential extraction procedure is the use of decoy conformations that are significant competitors to the native state in housing the given sequence.
In order to answer this problem we devised the following iterative technique:

- ✔ We started by taking as a set of alternative trial structures, the compact ones derived in the study of homopolymer landscape, and the metastable structures collected during the cooling of the heteropolymers. We thus used 90 different decoy conformations: 19 of the 20 basic structures obtained from the homopolymer model, (we exclude the native structure itself), and 71 from the alternatives generated by the repeated cooling process (as described above).

- ✔ We proceeded to determine rough values of the potential parameters, minimizing the cost function (3.5) with respect to these set of decoy structures.

- ✔ To add more relevant conformations to the decoy set, we slowly cooled each sequence about 5 times, using the potential parameter values extracted in the previous step.
  Initially, when non-optimal values of parameters are used, the simulations lead to lowest energy conformations that differ from the true ones for almost all the sequences. We added them to the set of used decoy structures, increasing the number of alternatives, enabling us to iteratively refine the parameters of the potential.

We iterated the procedure until it converged self–consistently, i.e. until a cooling simulation with extracted parameters leads to the true ground state within a precision of 1 Åper bead.

3.3.1 Results

The procedure converges very nicely and yields values in excellent agreement with the "true" potential parameters (Figure 3.5). Convergence is achieved using a decoy set of 1631 structures (i.e. 19 of the 20 basic structures and 1612 alternative ones).

Taken together, these steps lead to a unified and entirely self-consistent description of possibly the simplest off–lattice model of heteropolymer chains and opens the way for similar studies of small segments of real proteins.

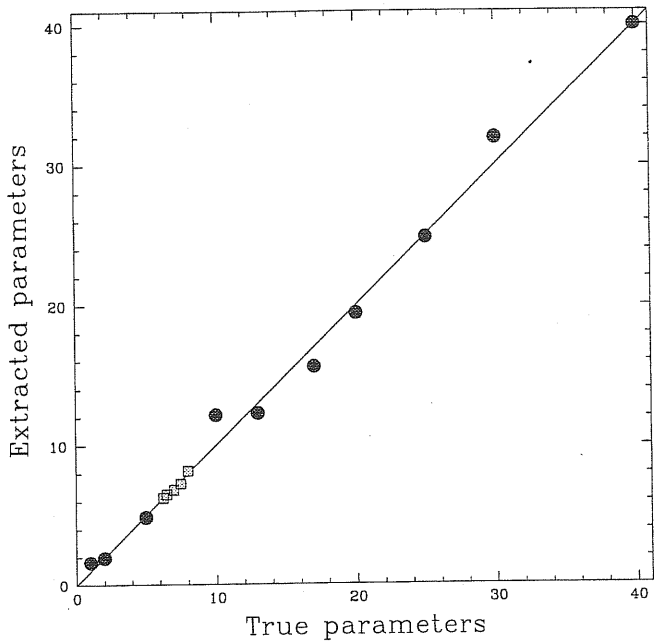

Figure 3.5: Extracted parameters of the potential versus the true parameters as obtained in the last iteration. Dark circles represent the η parameters, whereas the light squares denote the σ parameters. The correlation coefficient among two sets of parameters (true and extracted) is 0.997.

An important feature of the study is that a simple known potential was used for the design studies and therefore facilitated the verification of the potential parameters that were subsequently determined. This luxury is not present for similar studies on real proteins, for which the potential energies of interactions between amino acids are truly unknown.

In Chapter 5, building on the results reported here, we will tackle the same problem but for real proteins.

CHAPTER 4

FOLDING OF MODEL PROTEINS BY CONTACT MAP DYNAMICS

To tackle the protein folding problem from a theoretical point of view, one has to choose:

① a representation of protein structure;

② an approximation for the energy;

③ an algorithm to select conformations which minimize the energy.

In this Chapter we compare two promising approximate ways of accomplishing these three points, that have been proposed by two groups: the Weizmann group and our group at Sissa.

4.1 Contact Map representation meets Molecular Dynamics

The Weizmann's group has developed the *contact map approach* to protein folding. The contact map [78, 79, 80, 81, 67] of a protein with N residues is an $N \times N$ matrix S, whose elements are defined as S_{ij}=1 if residues i and j are in contact, and 0 otherwise. One can define "contact" between two residues in different ways: for example, two amino acids could be considered in contact when they are not nearest neighbor along the chain and their two C_α atoms are closer than some threshold (they used 8.5 Å[67]). Alternatively one can define contact $S_{ij} = 1$ when a pair of heavy (all but not hydrogen) atoms belonging to the two residues i and j is closer than 4.5 Å[81, 82].

The central premise of the contact map approach is that the search, to identify maps of low energy, executed in the space of possible contact maps for a fixed sequence, has an important computational advantage with respect to the search in the space of all possible configurations of a polypeptide chain. Changing a few contacts in a map induces rather significant large-scale coherent moves of the corresponding chain [83]. Secondary structures are easily detected from contact map. Alpha helices appear as thick bands along the main diagonal, since they involve contacts between one amino acid and its four successors. The signature of parallel or anti-parallel beta sheets are thin bands, parallel or perpendicular to the main diagonal. On the other hand, the overall tertiary structure is not easily discerned. It is possible to associate a map to any possible conformation of a polypeptide chain but the vice versa is not true: one of the most problematic aspects in the contact map representation is that by performing an unconstrained search of the minimum energy map in the contact map space one obtains maps with a very low energy, but without any physical meaning since they do not correspond to any realizable spatial conformation of the chain.

Vendruscolo *et al.* in [84] presented a method to execute an efficient search in the contact map space and in [67] provided a method to project any map onto a nearby one corresponding to a physically realizable conformation.

As detailed in the previous Chapter, we have pursued a different strategy by developing minimalist off-lattice models of proteins. We studied off-lattice toy protein models able to reproduce the essential features of the real folding process in terms of stability and accessibility – first of all the very peculiar and typical "funnel" structure for the energy landscape, that recent works [17, 18, 22, 19, 20, 14] have shown to play

a critical role). We found (see Chapter 3) that the case of a C_α chain with 4 species of amino acids, equipped by a suitable design procedure, is effective in representing global features of the energy landscape [85]. Off–lattice "toy models" of proteins are an essential tool to suitably test the potential extraction techniques [43, 44], and on the other hand the study of the dynamical properties can provide important insights in the real mechanisms acting in real proteins.

In the following, we will show that the contact map representation is a suitable approximation of the Lennard–Jones (LJ) model of proteins. The structures that are stabilized by a Lennard–Jones potential, can also be *approximately* stabilized by a contact potential. We generated a model data bank by using Molecular Dynamics technique (MD), and then applied Contact Map tools (CM) developed by the Weizmann group to recover the model structures. The contact map dynamics is able to provide, for a given choice of contact parameters, not just a unique structure for each sequence corresponding to the lowest energy state, but a set of candidates. We will show that an hybrid method, obtained by a suitable fusion of Contact Map Dynamics and Molecular Dynamics, seems to be very promising.

4.2 Generation of native–like structures

We used the same procedure exposed in Chapter 3 to generate several different compact structures, by cooling an homopolymer chain equipped with the interaction potential function:

$$V_{ij} = \delta_{i,j+1} f(r_{i,j}) + \eta((\frac{\sigma}{r_{ij}})^{12} - (\frac{\sigma}{r_{ij}})^{6}), \tag{4.1}$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ are the interparticle distances and the energy bond function $f(x)$ is :

$$f(x) = a(x - d_0)^2 + b(x - d_0)^4. \tag{4.2}$$

We set $d_0 = 3.8$, $\sigma = 6.5$ and $\eta = 40$ for the homopolymer case. The chain is made up of 30 monomers. The obtained compact structures are like the one shown in Figure 3.1.

A design procedure similar to that of § 3.2.1 was used to select suitable sequences for each configuration. Four types of amino acids were used ($a_i \in [1, \cdots, 4]$, $i \in [1, \cdots, 30]$) with fixed composition as in § 3.2.1, and values of parameters η as reported in Table 3.2. The only difference from the previously exposed procedure is that here the parameter σ is kept fixed to the homopolymer value ($\sigma = 6.5$) for each interacting pair of heteropolymers. The aim here is just to compare the Lennard–Jones molecular

dynamics with the contact map representation. Contact map approximation, as defined in [67, 84, 86] is characterized by a single length scale. The introduction of different values for the equilibrium distances is an crucial in order to get a funnel–like energy landscape. Without allowing the parameter σ to change not all compact configurations (obtained as homopolymer low energy state) are well designable. Nevertheless we were able to select 6 structures with associate sequences $a = (a_1, a_2, \cdots, a_{30})$ having the conformations as their unique ground state (at least within the dynamically accessible states).

4.3 Derivation of a set of contact energy parameters

The 6 sequences and native conformations, derived as described above, are used as a database to derive pairwise contact energy parameters.
The important question to answer is the following:

⇨ Given a set of sequences and native conformations obtained from the Lennard–Jones molecular dynamics, is it possible to assign a set of contact energy parameters $\mathbf{w}$ in order to obtain, *by using a contact map representation*, the same structures as the native states of the same sequences ?

A definition of *contact* is needed in order to assign a contact map S to each sequence. We defined in *contact* two beads on a C_α chain having a distance between R_L and R_U. Threshold values R_L and R_S identify a square well contact potential:

$$S_{i,j} = \begin{cases} 1 \quad \text{if} \quad R_L < r_{ij} < R_U \\ \\ 0 \quad \text{otherwise} \end{cases} \tag{4.3}$$

Suitable values for R_L and R_U were selected by the procedure exposed in § 4.3.1.
In the *pairwise contact approximation* the energy is written as:

$$E^{pair}(\mathbf{a}, \mathbf{S}, \mathbf{w}) = \sum_{i<j}^{N} \mathbf{S}_{ij} w(a_i, a_j). \tag{4.4}$$

That is, if there is a contact between residues i and j ($\mathbf{S}_{ij} = 1$) the parameter $w(a_i, a_j)$, which represents the energy gained by bringing amino acids a_i and a_j in contact, is added to the energy.

To derive the set w of pairwise contact energy parameters we required the following set of conditions to be satisfied for all.the proteins in the database:

$$E^{pair}(\mathbf{a}, \mathbf{S}_N, \mathbf{w}) < E^{pair}(\mathbf{a}, \mathbf{S}_\mu, \mathbf{w}) \; . \tag{4.5}$$

Here $\mathbf{S}_N$ is the contact map of the native conformation of sequence $\mathbf{a}$ and S_μ ($\mu = 1, \ldots N_D$) are contact maps taken from a set of N_D alternative conformations, generated by the procedure described in § 4.3.2.

As explained in § 2.3.2, we selected an optimal parameter set w such that for each decoy structure S_μ the inequality (4.5) is satisfied and the *gap* G, defined as

$$G = E^{pair}(\mathbf{a}, \mathbf{S}_\mu, \mathbf{w}) - E^{pair}(\mathbf{a}, \mathbf{S}_N, \mathbf{w}) \tag{4.6}$$

is as large as possible (see § 2.3.2 and refs. [43, 44, 86] for details).

4.3.1 Square well approximation of the Lennard–Jones potential

A successful approximation of the LJ potential by a contact potential depends critically on the choice of the upper and lower thresholds, R_U and R_L, in the definition of the contact (eq. 4.3). In the following the procedure adopted to derive R_U and R_L is explained in detail.

Derivation of R_U

Using MD, we analyzed the stability of the folded conformations against a cut in the tail of the LJ potential: we substituted the full Lennard–Jones potential

$$V_{LJ}(r) = \eta((\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^{6}), \tag{4.7}$$

with a potential equal to eq. (4.7) only for r *smaller* than R_U, and equal to zero for r *larger* than R_U:

$$V_{LJ}(r) = \begin{cases} \eta((\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^{6}) & \text{if } r < R_U \\ \\ 0 & \text{if } r > R_U \end{cases} \tag{4.8}$$

For a given R_U, we determined the ground state conformations of the 6 proteins in the database. This was done by performing n runs (with n about 50) of Molecular Dynamics minimization (by slow cooling, as described in § 3.2). For each protein, each minimization ended in a slightly different structure $\mathbf{r}_i^j(R_U)$ ($i = 1, \ldots, 6$, $j = 1, n$).

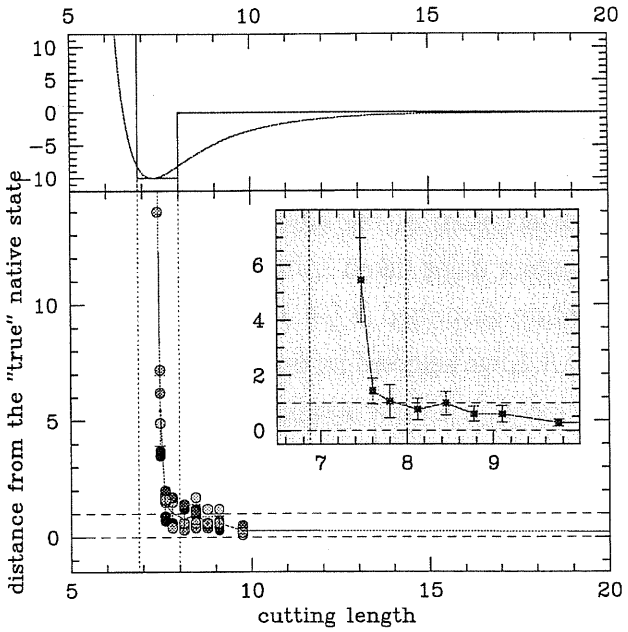

Figure 4.1: Mapping the LJ potential onto the contact potential. Distances of ground state conformations obtained for several values of R_U from the native state are shown for each sequence. In the inset the average RMS distance versus the cut–off values R_U is shown.

We measured the average root mean square (RMS) distance $\langle D(R_U)\rangle$ (defined as in refkab), averaged over the 6 proteins in the database, between the native conformations and the conformations found as minima of the potential of eq. (4.8).

$$\langle D(R_U)\rangle = \frac{\sum_{i=1}^{6}\sum_{j=1}^{n} D(\mathbf{r}_i^*, \mathbf{r}_i^j(R_U))}{6n} . \tag{4.9}$$

In Figure 4.1 the results for each protein are shown. In the inset the average distance $\langle D(R_U)\rangle$, with its standard deviation as a function of the cutting length R_U, is reported. Cutting the LJ tail very far from the abscissa of the minimum does not change the results at all, but cutting the LJ potential inside the attractive well leads to not trivial results. Remarkably, it is possible to cut the LJ potential inside the attractive well down to R_U=8 Å, keeping the average distance (4.9) below 1 Å. Therefore, we fixed the upper threshold $R_U = 8$ Å.

Derivation of R_L

To derive R_L we used an entirely different method. For $R_U = 8$ Å, and for different choices of R_L, we generated by contact map dynamics a set of $N_D = 60000$ decoys, 10000 for each sequence in the database, as explained in § 4.3.2.

We found that for $R_L < 6.9$ Å, the derivation of a set of contact energy parameters (by using the technique exposed in Chapter 2) is an *impossible* task. This can be proved rigorously by using the standard linear optimization procedure (i.e. the *perceptron learning* of ref. [51]). Instead for $R_L > 6.9$ Å, it was possible to find a set $\mathbf{w}$ to satisfy all the inequalities.

An intuitive explanation is the following. In the contact energy representation, monomers can move freely in the flat bottom of the well. There is an overall gain in energy if, on average, they get closer, since by making the structure more compact, more contacts are realized and the energy is lowered. When $R_L < 6.9$ Å, there are too many compact conformations that are in competition with the native state and have to be penalized by a suitable choice of the energy parameters. It can be proved that in this case there is not a solution to eq. (4.5) for large enough N_D.

From the previous discussion it might seem convenient to choose R_L as large as possible. However, in the following we explain why this is not the case.

Starting from the native conformations $\mathbf{r}_i^*$ we prepared the contact maps by defining a contact whenever two monomers were closer than $R_U = 8$ Å. Given a contact map it is in general possible to recover a conformation having that contact map as its map [67]. When a lower cutoff R_L is introduced in the reconstruction procedure, it turns out that it is possible to reconstruct exactly the native contact map only for $R_L < 6.8$ Å. If monomers are not allowed to get closer than $R_L = 6.8$ Å, the true native map is non physical. In other words, there is no chain which can realize that contact map. However, at $R_L = 6.9$ Å, it is possible to reconstruct conformations whose maps are "close" to the native one and that are physical. Such physical maps typically differ by a very few contacts from the native map. Moreover, the corresponding three dimensional structures are on average at a RMS distance less that 3 Åfrom the native conformations.

We decided to choose $R_L = 6.9$ Å, as the best possible compromise between the requests of solving eq. (4.5) and to assure that the native maps are physical.

4.3.2 Generation of alternative conformations

Our aim was to generate maps of energy low enough to "compete" with the native contact map. To obtain a large ensemble of these decoy maps we performed dynamics in contact map space collecting maps along the trajectory. The method is extensively explained in [84]. Here we outline the details that are relevant in the present implementation. Our algorithm is divided in four steps.

① We started from an existing map and performed large scale *"cluster" moves*. Clusters represent small groups of contacts between amino acids that are well-separated along the chain. The clusters are identified on a given map by laying down bonds that connect neighboring contacts on the map and identifying groups of contacts that are connected by such bonds [84]. Some of the existing clusters of contacts were removed and some other groups were restored elsewhere. The energy of the resulting coarse map was evaluated and a low energy map was retained. At this stage, no attempt was made to preserve physicality. The contact map which is obtained by this procedure is typically uncorrelated to the initial one.

② The resulting map was refined by using *local moves*. This refinement procedure consists in turning on or off (mostly one at a time) contacts that are in the vicinity of existing ones, following the rules introduced in [81].

③ We restored physicality by projecting the map obtained from the second step onto the physical subspace. Indeed, a generic contact map is not guaranteed to correspond to any real conformation of a polypeptide chain in three dimensional space. We used an efficient Monte Carlo reconstruction algorithm, developed by *Vendruscolo et al.* ([67] and [86]), that checks whether any given target map is physical or not. The method consists in moving around the beads of a C_α chain, without tearing the chain and without allowing one bead to invade the space of another. A "cost function", vanishing when the contact map of the chain coincides with the target contact map, and increasing when the difference between the two maps increases, controls the motion of the chain. The procedure ended up with a chain configuration whose contact map is physical by definition, and close to the target map.

④ We performed a further optimization by an energy minimization in real space using a standard Metropolis crankshaft technique [84].

Using this algorithm, we generated a set of N_D=60000 alternative conformations, 10000 for each of the 6 sequences in the database. As discussed in [84], the contact maps that are obtained by contact map dynamics (CMD) are uncorrelated and the N_D decoys form a representative set of low energy competitors for the native state. The important requirement is to generate a large set of uncorrelated decoys, since a good energy function must stabilize the correct structure and destabilize all the others.

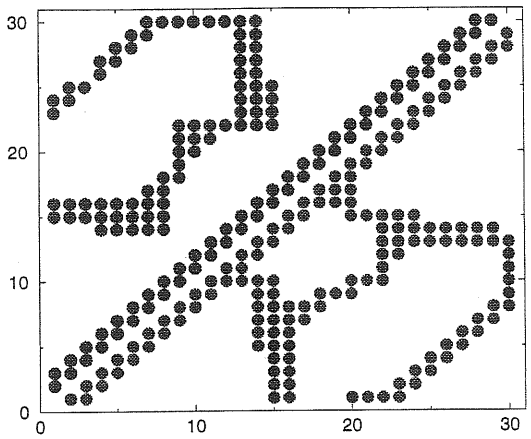

Figure 4.2: Native contact map of sequence 5 (below) and lowest energy contact map found with our procedure (above).

This selection of candidates was performed using the set of LJ parameters $w_{ij} = \eta_{ij}$ as a first guess for the contact parameters.

4.4 Folding in contact map space

The first question we posed is

⇨ Is it possible to fold the sequences of the database in contact map space?

A complete success in folding a protein in contact map space is marked by the exact recovery of the native map. As already observed, using $R_L = 6.9$ Åand $R_U = 8.0$ Å, the contact map associated to the true native structure is non physical. Nevertheless it is possible to find conformations of the chain whose maps are physical and close to the native ones (see Figure 4.2). In the present study we adopted a less stringent criterion for folding, namely the prediction in the contact map space is successful if the structure related to the predicted map is closer than 3 Åto the native structure. This criterion is supported by the evidence that a structure closer than 3 Åto the native one goes very quickly to the true native structure when is used as initial condition for Molecular Dynamics simulations, in a slow cooling mode. In other words, structures closer than 3 Åfrom the native one lie inside the attractive basin of the latter (as Figure 4.4 shows).

A successful folding is achieved since the procedure to derive energy parameters smoothes the shape of the energy landscape as Figure 4.3 shows for the contact maps generated in the simulation of sequence 5.

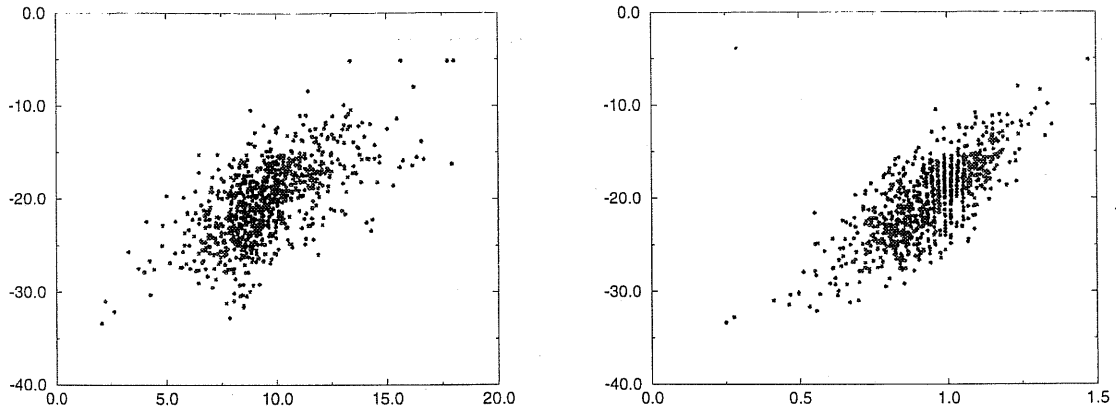

Figure 4.3: Correlation between energy and RMS distance to the native state(upper Figure), and between energy and Hamming distance (lower Figure).

The second and complementary question we posed is then:

⇨ Are the low energy conformations generated by contact map dynamics good starting points for Molecular Dynamics minimization ?

We analyzed again sequence 5. As shown in Figure 4.4, the 10 lower energy conformations generated by contact map dynamics are correctly ranked by Molecular Dynamics. Eight out of ten are able to reach the native state, while two find a final configuration with higher energy. From this fact, the answer to the question seems to be *yes*. This means that the utilization of Contact Map Dynamics as a first–step in a minimization procedure in the configurational space could be very useful.

The computational advantage of the Contact Map allows a quick first selection on the configurational space. Then, in this limited set of configurations, Molecular Dynamics, using a more realistic function to represent the inter amino acids potential, could be successfully employed.

4.5 Conclusions

In this Chapter we have shown that it is possible to approximate the LJ potential with a pairwise contact energy. This is the first case in which a non trivial energy potential is shown to admit such approximation. The result is grounded on the fact that both potentials are characterized by a single length scale.

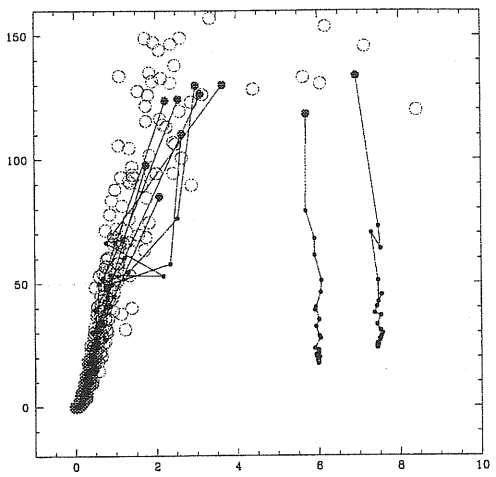

Figure 4.4: MD folding of sequence 5. Each trajectory is obtained starting from one of 10 conformations randomly chosen from the 100 conformations of lowest energy obtained by Contact Map Dynamics. Open dots are configurations collected during the generation of native structures (as detailed on § 4.2)

We also showed that a dynamics in contact map space is a suitable tool to perform energy minimization, when a correct parameterization of the energy is possible. Contact map dynamics provides, for a given choice of contact parameters a limited set of structures as good starting point for a MD minimization.

When the list of predictions obtained from CMD is submitted to a MD minimization, with the original LJ form of the energy restored, this procedure leads to the correct ground states. Contact map dynamics is then very useful to make a "first selection" on the conformational space, in order to perform a "fine tuning" with a Lennard–Jones potential. This result suggests that it could be very powerful to work with an hybrid method (as resulting from a "fusion" of these two technique of investigation) combining them in such a way that we take advantage of the their respective positive features, that in some sense are complementary.

CHAPTER 5

DYNAMICS OF PROTEINS: A REALISTIC TWO-BEADS MODEL

Coarse–grained descriptions of protein, mainly based on lattice models, have been widely used for various goals in protein research (see e.g. [9, 25, 26, 61, 62, 63]). Also Chapter 3 of this thesis is devoted to a very simple off–lattice model, where a protein is represented by a chain of N beads. Nevertheless the approximations involved in very simplistic models do not allow to completely reproduce the dynamics and the thermodynamics of real proteins. On the other hand, the Molecular Dynamical behaviors of specific real proteins are generally investigated by detailed atomistic approaches [23, 87]. Such calculations, however, are extremely demanding, limiting the scope of studies that can be performed. Experimental techniques on structural biology is a field in great expansion. At present there are a lot of experimental results that no kind of approach is able to reproduce. Yip and collaborators [88], for instance, were able to directly measure the forces involved in Insulin monomer-monomer interactions. A

theoretical prediction of this kind of measurements is beyond actual possibilities of "first principles" approaches.

Several models, going beyond the simple C-α one (discussed in Chapter 3), have been already developed during the last years (see for example [89, 90, 91, 92]). These models aim to provide a way to simulate the full folding process of real protein–chains, by looking for the best compromise between a high degree of simplification and the possibility of representing accurately the actual geometry of protein conformations.

In the same spirit, we present in this Chapter a realistic off–lattice model as an alternative to the already existing ones. The model is explicitly thought to investigate the dynamics of short protein-chains up to time significantly greater than the typical scale available by force field simulations. The representation of a protein chain by interacting "springs and beads" is relatively easy to be equipped with an Hamiltonian dynamics.

5.1 The model

The model is based on the representation of amino acid by a system of two spheres.

All amino acids can be regarded as being made of two parts:

- ✔ one, constituting the protein backbone, that is the same in all of them, and
- ✔ one representing the diversity of the amino acids, constituting the protein side–chain.

The role of the side–chain seems to be crucial in the secondary structure formation [14, 93]. The representation proposed here is the natural development from simpler models such the C-α one exposed in details in Chapter 3 (see also [61, 62, 63, 85]). In C-α models, proteins are regarded as chains of beads in three–dimensional space that in the folded form should have the beads at the α carbon positions. In the two–beads model we kept this representation to reproduce the backbone chain while we added a bead linked to each C-α to reproduce the related side–chain. Beads constituting the side–chain hold the information about the different kind of amino acids. Note that in this scheme Glycine is representable just as a sphere (the backbone sphere), without adding any side–chain sphere.

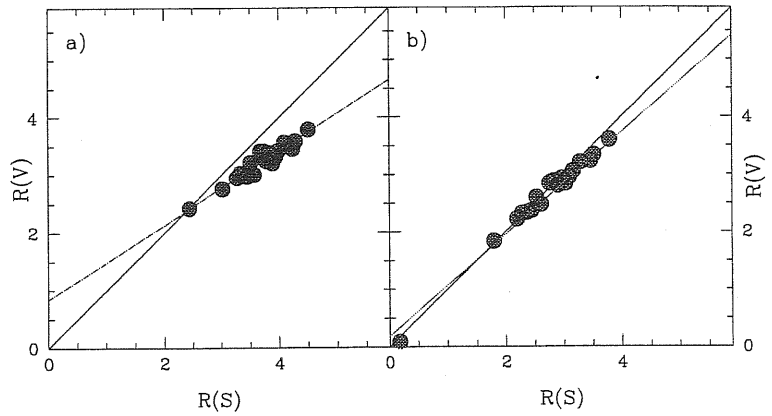

Figure 5.1: Estimates of radii obtained from experimental values of volume versus estimates of radii obtained from values of surface area of amino acids (a). The continuous line is the linear fit to the data, dashed line corresponds to the relation (5.1). (b) is the same as in (a), but with volume and surface area values of side chain effective spheres (i.e. the experimental values minus the values corresponding to Glycine sphere).

5.1.1 Preliminary considerations about a two beads model

A relevant aspect of real proteins is that each amino acid is not simply formed by a single atom but it consists of a molecule of several atoms, with a complex spatial structure showing fixed planes and preferential directions. To represent each amino acid with a single bead on a chain may be misleading if the goal is to reproduce the features of real proteins. Representations such as the C-α ones imply a spherical symmetry of each amino acid. On the contrary, in real proteins the different and particular spatial properties (as the asymmetries) of each amino acid play an important role in the folding process and heavily affect torsion and bond angles along the chain.

One can quickly estimate the improvement obtained giving up the spherical symmetry switching from a C-α model to a two-beads model by looking at the experimental values of volumes and surface area of amino acids. If the approximation of amino acids as a sphere were a good one, the following relationship should hold:

$$\frac{3}{4\pi}\sqrt[3]{V} \sim \frac{1}{4\pi}\sqrt[2]{S} \tag{5.1}$$

where V indicates the residue volume, and S the surface area. Figure 5.1(a) shows the plot of $\frac{3}{4\pi}\sqrt[3]{V}$ versus $\frac{1}{4\pi}\sqrt[2]{S}$ obtained from the experimental data [94] listed in Table 5.1. By linear fitting these values we obtain a slope of 0.6, an intercept of 0.8 Åand a correlation coefficient of 0.96.

A much better approximation of eq. (5.1) can be obtained by subtracting the values of volume and surface area of Glycine from all other amino acids.

Residue Type	Residue Volume (Å^3)	Surface Area (Å^2)	Side-Chain Effective Radius (Å)
Gly G	60.1	75	2.4 (backbone)
Gly G	60.1	75	0.
Ala A	88.6	115	1.8
Ser S	89.0	115	1.8
Cys C	108.5	135	2.2
Thr T	116.1	140	2.3
Pro P	112.7	145	2.3
Asp D	111.1	150	2.4
Asn N	114.1	160	2.5
Val V	140.0	155	2.6
Gln Q	143.8	180	2.8
Glu E	138.4	190	2.8
Leu L	166.7	170	2.8
Ile I	166.7	175	2.9
Met M	162.9	185	2.9
His H	153.2	195	3.
Lys K	168.6	200	3.
Phe F	189.9	210	3.2
Arg R	173.4	225	3.2
Tyr Y	193.6	230	3.3
Trp W	227.8	255	3.6

Table 5.1: Experimental data for volumes and surface area of amino acids (columns 2 and 3) and side–chain radii (column 4) as roughly estimated from the argument exposed in § 5.1.1.

In fact, one can assume that amino acids could be represented as two spheres, and that the Glycine is substantially equal to the backbone sphere present in all amino acids. Then, by subtracting the Glycine values from volume and surface area of all other amino acids, it is possible to define side–chain volumes and side–chain surface areas for which the relationship (5.1) might be more satisfactorily reproduced. Figure 5.1(b) shows the plot of side–chain radii, defined from the volume versus those defined from the surface area, derived under the above assumption. Linear fit of these values yields a slope of 0.9, an intercept of 0.1 Åand a correlation coefficient of 0.97.

This argument could suggest a way to estimate the radius of each effective side–chain and backbone spheres.

Nevertheless, looking at these estimates (reported in columns 4 of Table 5.1) and at the values of the distances between side–chain and backbone effective sphere centers (computed as described in § 5.1.2, and reported in columns 3 of Table 5.2) one can

notice that there is a superposition of backbone and side–chain spheres. One should then correct previous estimates by subtracting the shared areas and volumes from the total areas and volumes. We overcame this difficulty by keeping the representation of an amino acid as "a two center object", but computing the values of interaction distances directly by analyzing the distance distribution among the centers of each couple of interacting sphere, as shown in § 5.1.2. In this way any estimate of the effective sphere radii is not directly involved in the definition of the model.

5.1.2 Definition of the two–beads model

As in C-α models, only effective two body forces between beads, obtained by integrating out the degrees of freedom associated with the internal coordinates of each residue and the solvent, are considered.
In a two–beads model there are three kinds of allowed interactions:

① between backbone and backbone, if they are not nearest neighbors along the chain;

② between side–chain and backbone, if they are not constituting the same residue;

③ between side–chain and side–chain.

To define the model a choice of the interaction potentials between each pair of beads is needed.

Backbone/backbone interaction

The energy function of the virtual bond between C-α along the backbone chain is chosen to be a simple harmonic function:

$$f(x) = \frac{1}{2}a(x - d_0)^2 \tag{5.2}$$

with a taken to be 500. The parameter d_0, which represents the equilibrium distance of the nearest neighbors along the chain, is set equal to 3.8 Å, i.e. the experimental value for the mean distance between nearest neighbor C-α atoms along the chain in real proteins, as determined from the PDB.

Backbone/side–chain and side–chain/side–chain interactions

Also the interaction between each backbone bead and the corresponding side–chain bead is chosen to be an harmonic function,

$$f(x) = \frac{1}{2}a(s_i)(x - b(s_i))^2 \tag{5.3}$$

with parameters $a(s_i)$ and $b(s_i)$ differing for each kind of amino acid s_i. In order to fix suitable values for these parameters we analyzed the distribution of distances between side–chain and corresponding backbone in the PDB. The center of the side–chain is assumed to coincide with the *center of volume* of each side–chain, where we define *center of volume* the average of each side–chain atom position weighted on the respective volumes (experimental data for volume of atoms constituting the side–chain are taken from the work by Richards [95]). As said above, we assumed that the center of the backbone is located on the C-α atom of each amino acid. We measured the mean value $\overline{d_0(s_i)}$ and the variance $\sigma^2(s_i)$ of the distribution of distances between backbone and side–chain centers of each kind s_i of amino acids ($d_0 \equiv 0$ for Glycine). It is known (e.g. [96]) that amino acids do not have always the same fixed form, but can assume different conformations depending on the particular values of the angles among internal sub–units, that can vary in number for different kind of amino acids. There are amino acids (e.g. Thr, Cys, Pro) with a global form only slightly varying and others (e.g. Arg, Met, Lys) with two or more possible conformations significantly different from each other. Therefore the distribution of distances between the two centers looks very narrow for the former, and broader for the latter ones. Figure 5.2 shows a distribution of distances between side–chain and backbone centers for Arg and Thr.

We set the equilibrium distance $b(s_i)$ entering in equation (5.3) for a backbone/side–chain bond of a kind s_i of amino acid equal to the average value $\overline{d_0(s_i)}$.

Considering that the bond potential between each side–chain and backbone has to be much stronger than interactions among non bonded backbone and side–chain, and approximating the distribution of distances of the bonded backbone-side–chain as a gaussian, the following relationship holds:

$$\exp\left(-\frac{(r - \overline{d_0(s_i)})^2}{2\sigma^2(s_i)}\right) \propto \exp\left(-\frac{a(s_i)(r - \overline{d_0(s_i)})^2}{2T}\right) \tag{5.4}$$

where r is the side–chain backbone distance for an amino acid s_i and T is a temperature around the folding temperature.

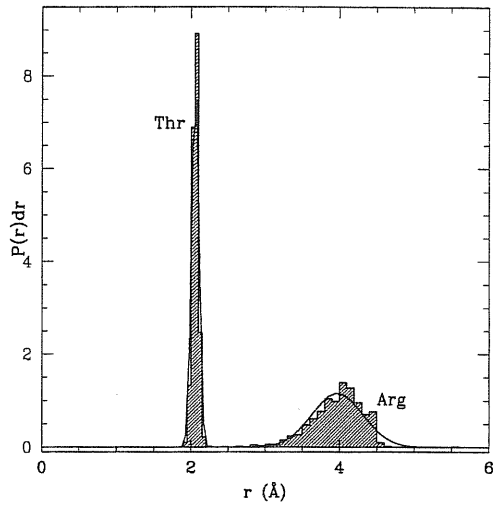

Figure 5.2: Distribution of distances between centers of side–chain and backbone for Thryptophan (left) and Arginine (right), as obtained from data of 163 proteins of PDB. A gaussian with the same mean and variance of each distribution is shown.

Residue Type (1)	$\frac{1}{\sigma^2(s_i)}$ (Å^{-2}) (2)	$\overline{d_0(s_i)}$ (Å) (3)
Ala	175.	1.53
Glu	20.	2.8
Gln	15.	2.85
Asp	16.	2.26
Asn	161.5	2.34
Leu	135.5	2.74
Gly	0.0	0.0
Lys	11.	3.57
Ser	121.	1.82
Val	360.	2.1

Residue Type (1)	$\frac{1}{\sigma^2(s_i)}$ (Å^{-2}) (2)	$\overline{d_0(s_i)}$ (Å) (3)
Arg	8.5	3.97
Thr	400.	2.06
Pro	705.	1.87
Ile	23.	2.5
Met	10.	3.1
Phe	182.5	3.3
Tyr	133.	3.63
Cys	491.	2.1
Trp	28.	3.75
His	173.5	3.

Table 5.2: Average distance (column 3) and inverse of the variance (column 2) of the distribution of distances between backbone and side–chain centers for each of 20 amino acids

We can thus take $a(s_i)$ proportional to $\frac{1}{\sigma^2(s_i)}$. Since we fixed the value of a in the energy function (5.2) to 500, comparing the width of the distribution of bonded C-α – C-α distances of PDB with the distribution of side–chain/backbone distances, we fix also $a(s_i)$ values in eq. (5.3) for each kind of amino acid ($a \equiv 0$ for Glycine). Table 5.2 shows the values of these parameters.

In this way the side–chain of an amino acid with a slightly varying shape is represented as a bead stiffly connected to the backbone, while the variation of shape of a side–chain is taken into account by a greater flexibility of the side–chain/backbone bond.

To completely define the model, a choice of a function for the energy interaction among non bonded side–chain and backbone is needed. The Lennard–Jones function is a suitable choice to reproduce the interaction between atoms and molecules and it is largely used in all atom computations [23, 24]. In fact, we employed it in previous works [85]. Nevertheless a Lennard–Jones function implies a very narrow range where the interaction between beads is effective: two beads are "bonded" (or strongly interacting) if their distance is approximatively between $0.9 \times r_m$ and $1.1 \times r_m$, where r_m is the abscissa of the minimum of the function [97].

Our two–beads representation is not precise enough to use a well defined interacting distance for each couple of amino acids. It is therefore convenient to use the following function:

$$V(r) = \begin{cases} \eta(\exp(\alpha(r - L_1)^2) - 2) & \text{if} \quad r < L_1 \\ -\eta & \text{if} \quad L_1 < r < L_2 \\ 4\eta((\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^6) & \text{if} \quad r > L_2 \end{cases} \tag{5.5}$$

that reproduces the attractive part of the Lennard–Jones function but allows a broader range of strong interactions. The exponential repulsive part avoids the superposition of amino acids. In some sense this potential is a combination between a square well and a Lennard–Jones shape.

The parameter α entering equation (5.5) is taken to be $\alpha = 1500$, while the parameter η has to be adjusted for each different couple of amino acid to fit the global effect of any kind of interaction between the various groups of amino acids and the interactions with the solvent. We assumed that the Glycine sphere is fully equivalent to the backbone sphere not only in terms of linear dimensions, but also in terms of strength of interaction. Then, twenty kinds of different spheres, interacting among each other with different strength, lead to 210 values for the parameter η entering eq. (5.5).

We will explain the procedure to estimate the values of η in § 5.2.1.

In equation (5.5) the lower cutoff L_1 and upper cutoff L_2 have also to be fixed for each couple of spheres.

If the effective sphere radii were estimable with some accuracy, one could assume $L_1(i, j)$ and $L_2(i, j)$, for amino acids of type i and j, proportional to the sum of their

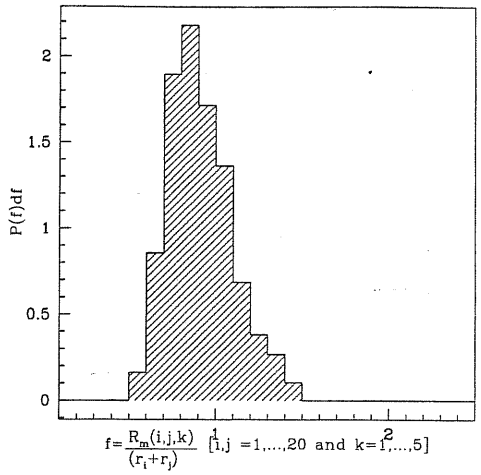

Figure 5.3: Distribution of values for the ratio $f = \frac{R_m(i,j,k)}{(r_i+r_j)}$ calculated for each value of (i,j,k). $R_m(i,j,k)$ are the minimum distances between interacting centers of amino acids of kind i and j ($k = 1, \cdots, 5$ marks the different range of distances along the chain), and $(r_i + r_j)$ is the sum of a-priori calculated radii as listed in Table 5.1.

respective radii, r_i and r_j, i.e.

$$L_1(i,j) = \beta_1 \times (r_i + r_j) \tag{5.6}$$

$$L_2(i,j) = \beta_2 \times (r_i + r_j) \qquad \text{with} \quad \beta_1 < \beta_2. \tag{5.7}$$

As discussed at the end of § 5.1.1, the estimates of radii that one can get from experimental data on volumes and surface areas are not precise enough.

We thus analyzed the distribution of distances $R(i,j)$ between each pair of interacting centers i and j, as computed from 163 real proteins of PDB. By examining the minimum distances $R_m(i,j)$ between each pair i,j it is possible to notice that amino acids closer than 4 residues along the chain have minimum distances smaller than more distant ones. One can argue that the closer amino acids are along the chain, the more they feel the deviation from the spherical symmetry that we are assuming in the interaction. In order to correct for this effect we took into account the possibility of having five values for the minimum radii $R_m(i,j,k)$ $k = 1, \cdots, 5$ for each pair of amino acid i,j, respectively distant 1, 2, 3, 4, and more than 4 residues along the chain.

Figure 5.3 is a plot of the distribution of the ratio between the minimum distances and the sum $(r_i + r_j)$ of radii of Table 5.1. The mean value of this distribution is equal to 0.93, with variance 0.24.

We fixed $L_1(i,j,k) = R_m(i,j,k)$, and $L_2(i,j,k) = L_1(i,j,k) + 1$ Å, since 1 Åis a common resolution RMS error on experimental determination of protein structures.

5.2 Application to crambin

An ideally perfect protein model should be able to reproduce the essential features of most of known proteins, primarily by finding the correct three dimensional structure of each foldable sequence. This implies that in a model with an adjustable set of parameters (as in our case the strength of interactions $\eta(s_i, s_j)$ $s_i, s_j \in [1, \cdots, 20]$) it should exist an universal choice of $\eta(s_i, s_j)$, for which the model works for all proteins.

Here we propose a more simplified test, by looking for a set of parameters able to stabilize the structure of a real protein: the crambin.

Crambin[1] is a plant seed protein, from Abyssinian Cabbage (Crambe Abyssinica) seed. It has a long history as test protein in protein folding simulation investigation (see for instance [101, 102, 86, 103]), because it is one of the smallest proteins (46 residues) with a significant amount of secondary structure (it comprises 1 sheet, 2 strands, 3 helices, 1 beta hairpin, 3 disulphide bridges).

The goal of the restricted test is, first of all, a feasibility study for our model. Indeed, if even a single protein could not be represented by a model, such a model is completely useless. On the other hand, a single protein study could be, if successful, a starting point to perform tests climbing back in the hierarchical classification of protein domain structures in order to define the largest set of proteins reproducible by a model. For instance, crambin is classified as belonging to:

- ✔ the class of *alpha-beta* proteins,
- ✔ the architecture of *2-Layer Sandwich*,
- ✔ the topology of *Alpha-Beta Plaits*, and
- ✔ the homologous superfamily of *1cbn*.

Therefore, the next step in this study could be the search of parameters able to stabilize each protein of the *1cbn* homologous superfamily, and so on.

5.2.1 Derivation of energy parameters

It is well known from experiments that real proteins are able to regain their native states after denaturation (by heating, changing PH, etc..), if the denaturation parameter (variation of temperature, of PH, etc...) does not exceed some upper limit. This means

[1] PDB ID code **1crn** [98, 99, 100])

that the native structure of a protein represents a deep minimum in the configurational space.

Analogously, we can say that a certain set of parameters is a good set for our model if, by performing molecular dynamics simulations, it is possible to find again the native structure of crambin, when perturbed.

In Chapter 2 we proposed an optimization method for the determination of effective potential energies of interaction which we extensively tested on lattice models [43, 44]). In Chapters 3 and 4 we applied this method to different toy protein models. We briefly recall that the application of the method requires the knowledge of a collection of N_Γ alternative conformations Γ_i, $i \in [1, \cdots, N_\Gamma]$ (decoys) for a given sequence σ. The method selects the optimal parameters such that the smallest energy gap between the energy $E(\Gamma, \sigma)$ of the sequence σ in its native conformation Γ_n and the (higher in energy) minimum energy alternative one (among the decoys) is as large as possible.

The procedure may be implemented by defining a cost function F_{gap}:

$$F_{gap} = -min_{(\Gamma_i \neq \Gamma_n)}(E(\Gamma, \sigma) - E(\Gamma_n, \sigma)) \tag{5.8}$$

and by choosing the parameter values of the potential that minimize this cost function. Notice that now we have a single sequence σ to stabilize (and not many as in Chapters 2, 3, 4). Working with only one sequence, the denominator introduced in § 3.3 is not needed.

In Chapter 3 we presented an iterative procedure to select significant competitors to the native state in housing the given sequence in order to extract potential parameters for an off–lattice model. The off–lattice iterative procedure is model independent and we repeated an identical strategy in order to stabilize the native state of crambin using the two beads model described above.

Initially, in order to collect a first trial set of decoy structures, we performed 10 Molecular Dynamics simulations –combined with a slow-cooling procedure, as in Chapter 3– of the crambin-like sequence, in a two beads representation.

The initial interaction parameters $\eta(i, j)$ entering eq. (3.1) are set at random in the interval (0.1 – 20). We employed the same algorithm [65] of Chapters 3 and 4 to numerically integrate the equation of motion. In each run the system was equilibrated for about 1000 time steps after successive cooling obtained by lowering the temperature each time by a constant factor (< 1).

We proceeded to determine rough values of the potential parameters that minimize the cost function (5.8) with respect to the initial set of decoy structures. The expres-

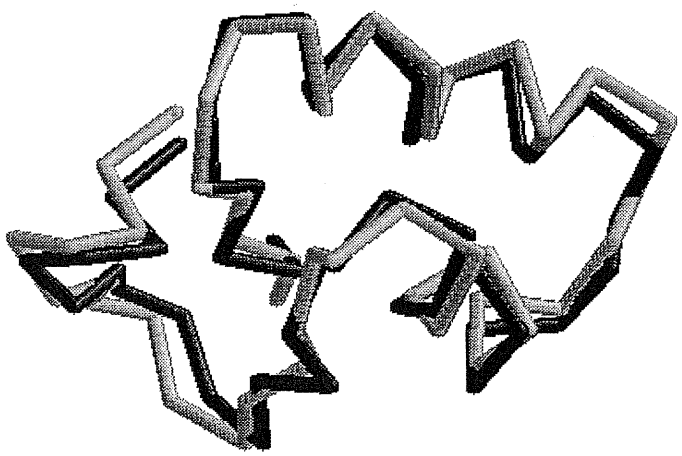

Figure 5.4: Native structure of crambin (light grey) and the structure obtained from MD simulations with extracted parameters, using the two beads model (dark grey). Only bonds connecting the adjacent alpha carbons of each amino acid are shown for clarity.

sion (5.5) with unspecified parameters η was used as the trial potential function for the interactions between non-bonded beads, while for bonded beads expressions (5.2) and (5.3) were used with fixed parameters (determined as described in § 5.1.2). Values of parameters η are constrained to lie in the interval (0.1–20).

We used extracted values of η to perform several runs (about 5, with different initial conditions) of molecular dynamics simulations in a slow-cooling mode. The initial configuration of each run is chosen as a perturbation of the crambin native state. We used the lowest energy configurations, collected during these simulations as additional decoy structures, in order to iteratively refine the parameters of the potential. We iterated the procedure until it converged self-consistently, i.e. until a cooling simulation with extracted parameters led to low energy structures that do not change the parameters when added to the decoy set.

5.2.2 Results

The crambin structure, as taken from PDB, has a RMS experimental determination error of 1.5 Å, and therefore we did not retain as alternative conformation those closer than 1.5 Åfrom the crambin native structure. Crambin sequence is composed by 15 of 20 amino acids and 4 of these appear only once. Then, parameters η effectively entering the energy are 116.

The procedure of § 5.2.1 converged to the parameters set reported in Table 5.3.

	Ala	Glu	Gln	Asp	Asn	Leu	Gly	Lys	Ser	Val	Arg	Thr	Pro	Ile	Met	Phe	Tyr	Cys	Trp	His
Ala	0.13	8.27	—	6.92	12.07	19.36	0.10	—	19.87	0.13	0.53	0.79	0.11	1.95	—	0.39	0.91	2.32	—	—
Glu	8.27	—	—	0.30	19.98	19.35	6.39	—	19.94	19.99	7.71	5.66	19.90	0.12	—	11.71	19.57	0.13	—	—
Gln	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Asp	6.92	0.30	—	—	19.70	3.65	5.20	—	19.97	20.00	0.11	0.12	0.13	19.51	—	0.30	2.93	19.81	—	—
Asn	12.07	19.98	—	19.70	19.99	19.96	0.55	—	20.00	19.85	19.96	10.41	10.01	3.81	—	19.94	19.87	0.11	—	—
Leu	19.36	19.35	—	3.65	19.96	—	2.89	—	19.65	19.84	19.70	0.18	19.98	9.91	—	16.10	19.65	0.27	—	—
Gly	0.10	6.39	—	5.20	0.55	2.89	0.14	—	0.10	0.10	2.96	1.42	0.11	4.45	—	0.65	0.83	0.10	—	—
Lys	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Ser	19.87	19.94	—	19.97	20.00	19.65	0.10	—	19.90	19.91	19.94	0.11	19.98	7.21	—	19.98	0.15	0.13	—	—
Val	0.13	19.99	—	20.00	19.85	19.84	0.10	—	19.91	2.83	19.83	5.36	1.04	19.27	—	19.03	19.96	20.00	—	—
Arg	0.53	7.71	—	0.11	19.96	19.70	2.96	—	19.94	19.83	18.56	0.10	19.89	17.94	—	19.48	19.98	0.15	—	—
Thr	0.79	5.66	—	0.12	10.41	0.18	1.42	—	0.11	5.36	0.10	5.00	0.11	1.29	—	6.27	1.94	0.69	—	—
Pro	0.11	19.90	—	0.13	10.01	19.98	0.11	—	19.98	1.04	19.89	0.11	15.53	18.27	—	0.17	0.14	0.10	—	—
Ile	1.95	0.12	—	19.51	3.81	9.91	4.45	—	7.21	19.27	17.94	1.29	18.27	13.92	—	0.10	2.71	9.24	—	—
Met	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Phe	0.39	11.71	—	0.30	19.94	16.10	0.65	—	19.98	19.03	19.48	6.27	0.17	0.10	—	—	0.13	14.37	—	—
Tyr	0.91	19.57	—	2.93	19.87	19.65	0.83	—	0.15	19.96	19.98	1.94	0.14	2.71	—	0.13	0.12	9.19	—	—
Cys	2.32	0.13	—	19.81	0.11	0.27	0.10	—	0.13	20.00	0.15	0.69	0.10	9.24	—	14.37	9.19	0.10	—	—
Trp	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Hys	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

Table 5.3: Parameters for 116 amino acids interacting pairs of crambin. Values reported in the Table as involving *Glycine* mean values involving the backbone.

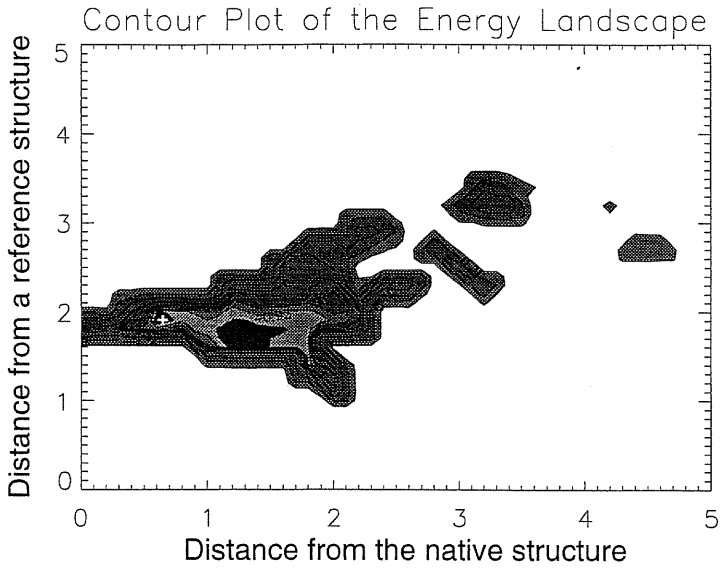

Figure 5.5: Contour plot of the energy landscape of crambin as obtained from the conformations obtained during numerous dynamical runs of slow cooling. The energy of each conformation is plotted as a function of its distance from the native structure and from a fixed "reference" conformation. Minimal energy conformation is located at distance 0.6 Åfrom crambin native state. An other relative minimum structure is located at distance 1.3 Åfrom the native state. Both structures are within the limits of experimental error.

Using these values of η parameters MD simulations led to a stable minimum energy structure for crambin differing 0.6 Åfrom the real one. The real native structure and that simulated by the two beads model are shown in Figure 5.4.

Figure 5.5 shows the contour plot of a three dimensional representation of the crambin energy landscape, obtained from final values of parameters in the two beads representation. On the horizontal axis the distance from the native structure is reported. The minimum energy structure obtained with the extracted parameters is at distance 0.6 Åfrom the real one. One other broad minimum appears at a distance of about 1.3 Åfrom the native structure.

5.3 Conclusions

In this Chapter a realistic, simple off–lattice model has been presented, in which the amino acids are represented as a couple of interacting centers with suitable equilibrium bond lengths. Energy parameters entering the Hamiltonian of the model are calculated in order to reproduce the stability of the crambin structure. Our procedure is able,

employing the extracted parameters, to recognize the native state of crambin as a –at least relative– minimum, within the limits of experimental error.

An application of this model to reproduce experimental results – like that shown in Ref. [88]– seems eminently feasible.

CHAPTER 6

CONCLUSIONS AND PERSPECTIVES

6.1 Conclusions

In this thesis we have presented models and techniques concerning the modeling of proteins, introducing in this study innovative tools and ideas.

First of all we exhaustively treated the problem of determination of the effective interaction potentials among the amino acids of a protein. The knowledge of these potentials is crucial for computational studies of folding and protein design, and it is at the basis of any realistic protein model. A new, promising, strategy to tackle this problem has been presented (Chapter 2) and compared with already existing methods.

Supported by the acquisition of this extraction potential technique, we have then developed powerful, albeit simple, off–lattice models: initially (Chapter 3) by using the simplest assumption that a protein may be represented by a chain of N beads, then (Chapter 5) by considering a more realistic representation of amino acid by a system of two spheres. The possibility of an hybrid technique (essentially a fusion of Molecular

Dynamics and Contact Map Dynamics tools) to increase the efficiency in the protein energy investigation has been also presented (Chapter 4).

A very stringent test for models and techniques presented in this thesis would be the prediction of certain protein properties yet to be investigated, both experimentally and theoretically.

In this spirit, we chose to address the problem of protein aggregation. In this Chapter we first present a short overview about multimeric proteins, and then we discuss how this problem could be tackled within a very simplified protein representation. Finally preliminary results of a study currently under investigation are described.

6.2 Multimeric protein aggregation

A large number of proteins are formed by several identical polypeptide chains that coalesce into a multimeric molecule. There are still many crucial and outstanding questions about the detailed mechanism of aggregation of different monomers into a multimeric protein.

Insulin, for instance, is a dimeric protein comprised of a 21–amino–acid chain (A) and a 30–amino–acid chain (B) linked to each other. However, it can also exist in a monomeric form, and as monomer it interacts with its transmembrane receptor to increase the glucose transport and utilization. Figure 6.1 gives a backbone representation of the dimeric state of insulin, as taken from the PDB.

The efficacy of insulin medications for the treatment of insulin-dependent diabetes could be boosted by a better understanding of insulin association and dissociation mechanism. Recently, direct force measurements of insulin monomer–monomer interactions have been performed by Yip *et al.* [88]. They revealed the complexity of the insulin dimer dissociation and suggested the disruption of discrete molecular bonds at the monomer–monomer interface.

Many multimeric proteins gain biological activity only by the association of subunits. A representative example is the HIV–protease, a dimeric protein of the Human Immunodeficiency Virus. This protein is composed by two chains symmetrically disposed in the native state, and it acts only in the dimeric form. A deep understanding of the association–dissociation mechanism could be important to the development of an antiviral therapy. Figure 6.2 shows the native state of the HIV-protease.

Several experimental results were collected during last ten years by G.Weber and

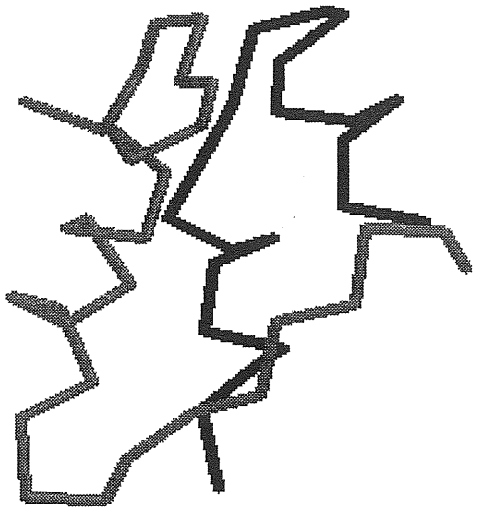

Figure 6.1: Dimeric native structure of insulin. Only bonds among C_α are shown for clarity. The darker chain is the 21 amino acids A chain, while the lighter one is the 30 amino acids B chain.

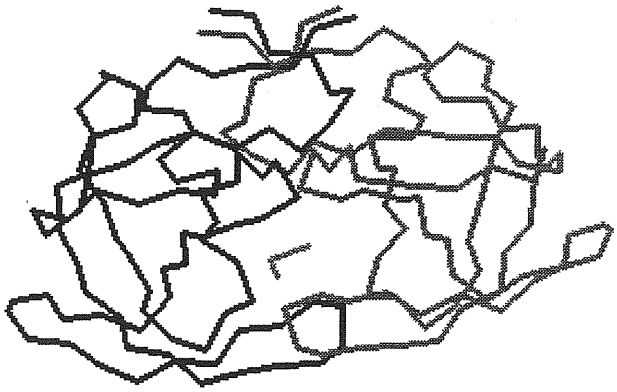

Figure 6.2: Dimeric native structure of a complex of HIV–protease with a dihydroethylene–containing inhibitor (short segment inside the two chains). Only bonds among C_α are shown for clarity. The structure is composed by two 99 amino acids chains symmetrically disposed.

coworkers [104] about dissociation of native dimers and tetramers. They showed that conformational changes of the protein subunits (monomers) occur when these become separated from each other by application of hydrostatic pressure, low temperature, or simple dilution. Conformational changes are supposed to be a consequence of the exposure of hydrophobic side–chain in the dissociate monomer structure, whereas in the native structure they are buried in the interface between monomers. In spite of these structural rearrangements, several hydrophobic side–chain remain exposed to water in the dissociated structure. This means that part of the tertiary structure should be stabilized by the interaction between subunits, so that the native state stability could depend to a large extent on the interaction between subunits. These features suggested that a dissociated subunit could be a molten globule.

In the following we present preliminary studies to approach the aggregation mechanism of multimeric proteins. Although these studies are accomplished by using a very simple (off–lattice) toy model, the results can reproduce, at least in a qualitative way, some of the experimental features described above. We defer the detailed comparison with the case of a real dimeric protein to future work.

6.3 A preliminary study

6.3.1 The model

In order to perform a feasibility study of the problem, we considered a toy–model dimeric protein. We chose to work initially with the simplest model, the C_α representation introduced in Chapter 3.
The interaction potential of the model is given by:

$$V = V^A + V^B + V^{AB}, \tag{6.1}$$

where V^A (V^B) is the potential energy of N_A (N_B) interacting beads constituting chain A (B):

$$V^A = \sum_{i<j \;\; i=1,N_A} \{\delta_{i,j+1} f(r_{i,j}) + \eta((\frac{\sigma}{r_{ij}^A})^{12} - (\frac{\sigma}{r_{ij}^A})^6)\}, \tag{6.2}$$

and V^{AB} is the potential energy given by the interaction of beads of chain A with beads of chain B:

$$V^{AB} = \sum_{i=1,\cdots,N_A \;\; j=1,\cdots,N_B} \{\eta((\frac{\sigma}{r_{ij}^{AB}})^{12} - (\frac{\sigma}{r_{ij}^{AB}})^6)\}. \tag{6.3}$$

In eq. (6.2) $r_{ij}^A = |\mathbf{r}_i^A - \mathbf{r}_j^A|$ represents the distance between beads i and j of chain A, whereas in eq. (6.3) $r_{ij}^{AB} = |\mathbf{r}_i^A - \mathbf{r}_j^B|$ is the distance between bead i of chain A and bead j of chain B.

The parameters η and σ entering these equations determine, respectively, the energy scale and the interaction range between monomers.

The function $f(x)$ in eq. (6.2) represents the energy of the virtual C_α–C_α peptide bond and it is equal to:

$$f(x) = a(x - d_0)^2 + b(x - d_0)^4, \tag{6.4}$$

with a and b taken to be 1 and 100 respectively, and d_0 set equal to 3.8 Å. The effect of $f(x)$ is to act as a "soft clamp" to keep subsequent residues at nearly the typical distance observed in real proteins.

As in Chapter 3, in order to select compact, designable configurations, we initially dealt with a homopolymer model, setting parameters η and σ to constant values $\eta = 40$ and $\sigma = 6.5$ Åin both eqs. (6.2) and (6.3).

6.3.2 Construction and design of a dimeric structure

Most of dimeric native structures present some symmetry, one of the most common being the $C2$ symmetry. The HIV–protease, for instance, exhibits this property. We therefore chose to impose a $C2$ symmetry to the native state of the dimeric toy–protein.

Construction

In order to select a compact, low–energy configuration of the dimer with a given symmetry $\mathcal{S}$, we modified the homopolymer dynamics. In practice we considered the motion of chain A only under the potential:

$$V = V^A + \frac{1}{2} V^{A\,\mathcal{S}(A)}, \tag{6.5}$$

where $\mathcal{S}(A)$ is the chain configuration obtained from of the application of the symmetry transformation $\mathcal{S}$ to the chain A. The potential $V^{A\,\mathcal{S}(A)}$ depends only on the coordinates of chain A. In our case, if (x_i, y_i, z_i) $i = 1, \cdots, N_A$ are the coordinates of beads of chain A, then $(-x_i, -y_i, z_i)$ $i = 1, \cdots, N_A$ are the coordinates of those in chain $\mathcal{S}(A)$ and the expression of $V^{A\mathcal{S}(A)}$ is:

$$V^{A\,\mathcal{S}(A)} = \sum_{i,j=1,\cdots,N_A} \{\eta((\frac{\sigma}{\tilde{r}_{ij}})^{12} - (\frac{\sigma}{\tilde{r}_{ij}})^6)\}. \tag{6.6}$$

where $\tilde{r}_{ij} = \sqrt{(x_i + x_j)^2 + (y_i + y_j)^2 + (z_i - z_j)^2}$.

Starting by a randomly generated swollen configuration of a chain made up of 50 monomers and by using Molecular Dynamics simulations combined with a slow cooling procedure (as described in details in Chapter 3), we generated a target conformation for the dimer native state.

Design

The design procedure of Chapter 3 was used to assign a suitable sequence to the selected structure. We used four kind of amino acids $a_i \;\; i = 1, \cdots, 4$, with the values of the matrix elements $\eta_{a_i,a_j} \;\; i,j = 1, \cdots, 4$ as given in Table 3.2. The composition of both the sequences (of chain A and B) were constrained so that there were 10 amino acids of the first and second types and 15 amino acids of the third and fourth kinds. We did not impose any constraint about equality or symmetries between sequences A and B. In fact, looking at real dimeric proteins of PDB, one can note that a symmetric native state not necessarely implies strictly symmetric sequences of the two subunits.

Indeed, sequences obtained from the design procedure are different, even if very similar as a consequence of the symmetry of the structure.

Repeating the procedure of Chapter 3, small variations in the Lennard–Jones length parameter were allowed. Here 6 possible values of σ (equal to 6, 6.25, 6.5, 7.0, 7.5 and 8.0 Å) were permitted, both for potential V^A (V^B) and V^{AB}.

Figure 6.3 shows the dimer structure. Different colors of the beads represent different degrees of hydrophobicity. Type 1 beads are the most hydrophobic ones, and are buried in the monomer cores and on the monomer–monomer interface.

6.3.3 Properties of the dimer model

We confirmed that the symmetric target structure was the lowest–energy state for the two interacting chains equipped with the designed sequences, A and B, by slowly cooling the chains several times from different random initial conditions.

We collected several configurations of the interacting chains during MD simulations in order to study the energy landscape of the dimer. Figure 6.4 shows the energy of each configuration of the two chains plotted against the distance from the dimer native state. The distance among configurations is measured by using the Kabsch expression [71] already used in Chapters 3, 4 and 5.

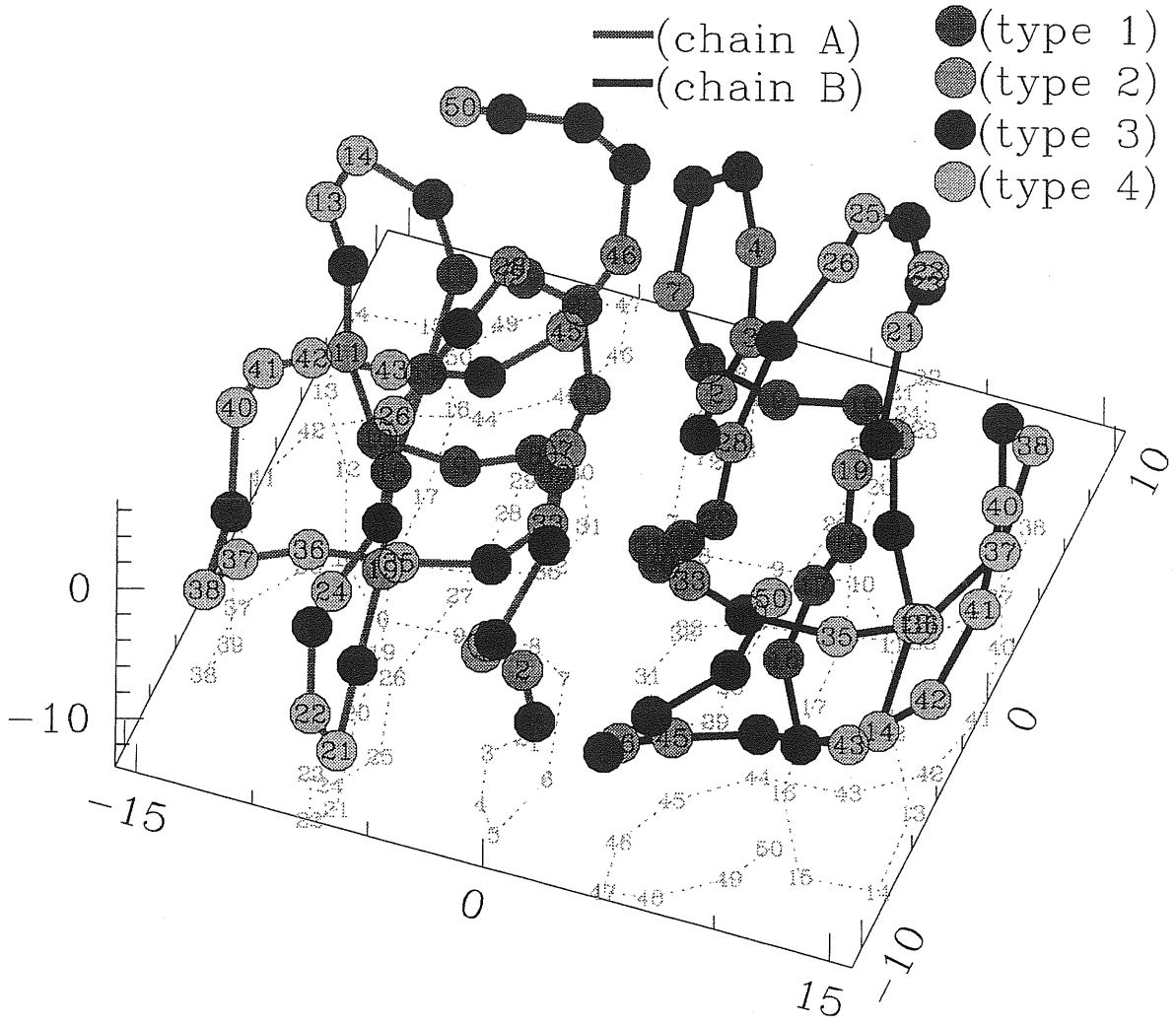

Figure 6.3: Native structure of the toy–model dimer. The two chains are represented in different colors, violet (chain A) and blue (chain B). The red beads are the most hydrophobic ones (type 1), the following in our hydrophobicity scale are the green beads (type 2), then the blue (type 3) and the azure ones (type 4).

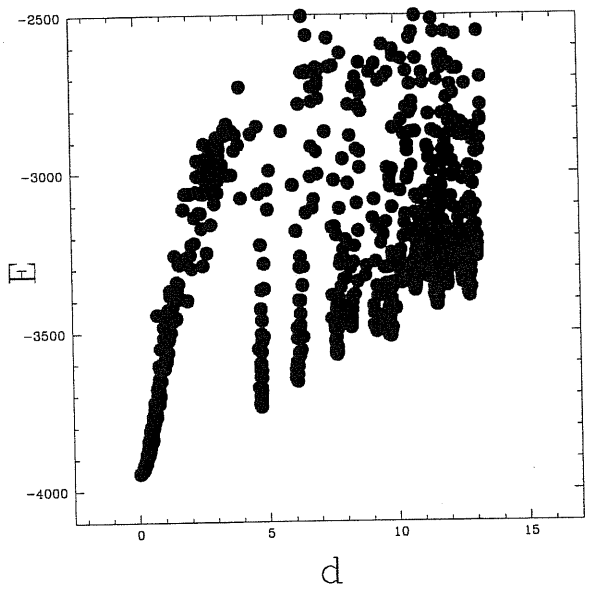

Figure 6.4: Energy of the dimer versus the distance from the dimeric native state.

MD simulations showed that two interacting chains are able to find the lowest minimum in a reasonable time (e.g. in a time observable by a MD simulation) only if starting from some initial conditions. Obviously, if all beads of chain A were initially located too far –with respect to the Lennard–Jones interacting distance σ– from any bead of chain B, the two chains would evolve independently, and the dimer structure could not be reached.

Figure 6.4 shows the results of several simulations, started from initial condition allowing the chains A and B to "feel" each other.
Interesting features arose from these dynamical simulations as we discuss in the following.

Monomer–monomer force

Static measurement We measured the interacting force between the subunits of the dimer upon varying their distance. We placed the two monomers at increasing distance, while "freezing" the shape that they have in the dimer native state; then we computed numerically the total force acting in a monomer, along the line joining the monomer centers of mass. The behavior of this force is shown in Figure 6.5. Large points in the Figure represents the computed values for the force $F(r)$ at fixed values r of distance between monomer centers of mass. The continuous line

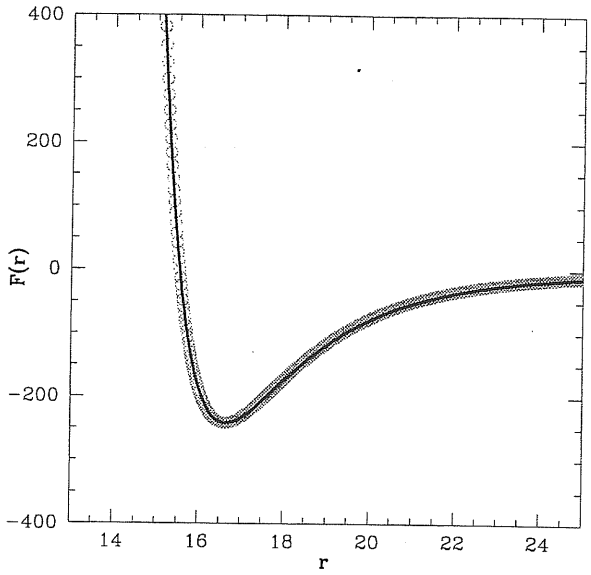

Figure 6.5: Result of the "static" evaluation of the inter–monomer force at varying the initial separation between centers of mass.

is the best fit $F^{fit}(r)$ of the data in the form:

$$F^{fit}(r) = \epsilon\{(\frac{\rho}{r})^{\alpha_1} - (\frac{\rho}{r})^{\alpha_2}\}. \tag{6.7}$$

The best fit is obtained for parameters values $\epsilon = 693.343$, $\rho = 15.5$ Å, $\alpha_1 = 22.8$ and $\alpha_2 = 8.5$, and the correlation coefficient among fitted and measured values is equal to 0.9998.

Such a measurement is a "static" estimate of the force between monomers: we measured the force acting on the monomers at the first instant, when they start to relax after separation. During the successive time, if the monomer beads are free to move, they rearrange the structure in order to decrease this force.

"Free fall" measurement Experimental measurement of the force acting between dimer subunits are performed in a finite, albeit short, time interval. The dynamical behavior of the force need also to be taken into account when comparing experimental results and theoretical predictions.

We measured the behavior of the force during the motion of monomers, when they are free to move towards each other. This is shown in Figure 6.6 during the relaxation motion of the monomers. MD trajectories quickly move close to the dimer native state. Then the inter–monomer force oscillates very fast around the equilibrium (zero temperature) dimer value (i.e. $F(r_{native}) = 0$). The final

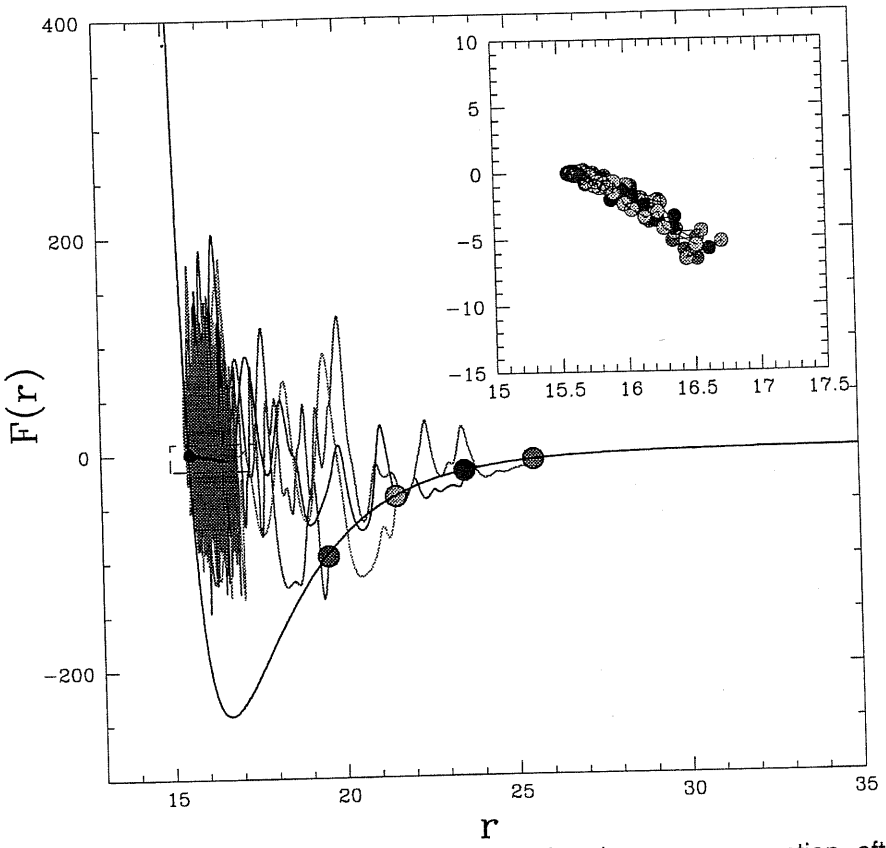

Figure 6.6: Behavior of the inter–monomer force during the monomer motion, after a separation. Big "colored" points at right side are the starting points of the MD trajectories. Four different trajectories are represented by four "colored" lines. Black point at the left side is the final state (native state) reached by all trajectories. The static force behavior is shown for reference (black line). The region marked by the sketched rectangle is blown up in the inset to show the behavior of the average force in the fast oscillations region.

structure of all these simulations is the dimer native structure (black point on the left side). Simulations starting from inter–monomer distance larger than 26 Å(i.e. the rightmost initial distance in Figure 6.6) are not able to find the dimer native state, since the initial inter–monomer attractive force becomes too weak.

Constrained measurement Experimentally, in order to measure the monomer–monomer force, some part of the two monomers must be attached to opposing surfaces at different distances [88], so that the two monomers are not allowed to move towards each other during force measurement.

In order to mimic this feature (the anchoraging of monomers to substrates) we performed also numerical calculations using a Constrained Molecular Dynamics

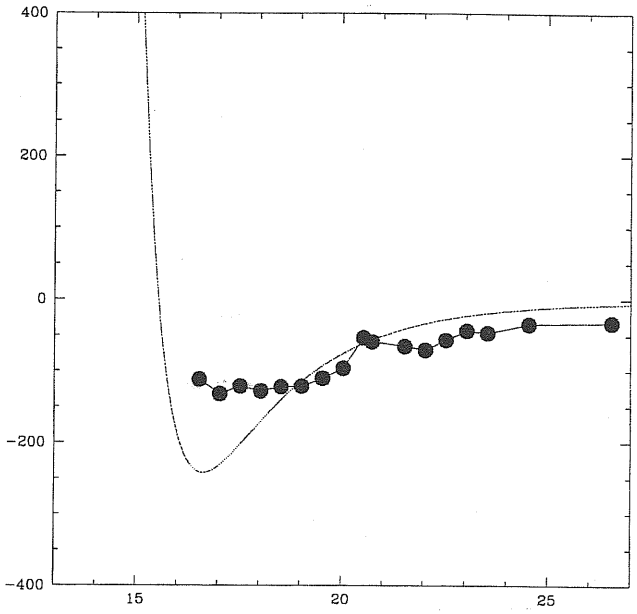

Figure 6.7: Inter monomer force as resulting from Constrained MD simulations (filled dots). The static force behavior is also shown for reference (continuous line).

algorithm. In particular, we used the application of the "Rattle" algorithm [105] (that is a slightly modified version of the standard "Shake" algorithm [106]) to the MD symplectic algorithm [65] already used in Chapter 3, 4, 5.

In this way we obtained a simple and efficient method for integrating the equations of motion for molecules subject to intra–molecular constraints. As a constraint, we chose to immobilize the center of mass for both monomers, at a given distance between each other.

Figure 6.7 shows the results of these simulations. Filled points represent the inter–monomer force measured when the center of mass of both monomers are fixed at a certain position, versus this separation distance. Surprisingly the measured behavior of the force versus the inter–monomer distance exhibits multiple vertical jumps, comparable with those observed in force curves acquired experimentally [88, 107, 108, 109, 110, 111].

The dimer toy–model, in spite of its extreme simplicity, was able to account for a peculiar, interesting feature shown by real dimer.

It has been hypothesized [107, 108, 109, 110, 111] that the multiple jump behavior of inter–monomer force was a consequence of the disruption of a hierarchy of dif-

ferent kinds of molecular bonds (hydrogen bonds, electrostatic forces, van der Waals interactions...) at the monomer–monomer interface.

Within our model we reproduced the same jump behavior, by using only a simple Lennard–Jones potential to mimic the complex interactions among amino acids and by designing the dimeric structure to be strongly favored from the energetic point of view. This result suggests that the characteristic feature of inter–monomer force curves could be simply a consequence of a design procedure optimizing the dimer structure with respect to the alternatives. In order to support this conclusion, we have begun a series of detailed comparative studies among well–designed and undesigned dimeric structure.

Conformational drift

We performed several MD simulations starting from a monomer–monomer initial distance larger than the one allowing interactions between the chains. Starting from swollen configurations spatially far from each other, the two monomers fold separately, as two independent chains. We collected several configurations during MD simulations of these independent monomers.

We found that the minimum energy conformations of the couple of non–interacting monomers is composed by very similar A and B monomeric structures –0.7 Åof RMS difference– and not too different from the monomer structures constituting the dimer native state (about 2 Å). The similarity between the minimum–energy conformations of chains A and B folded separately could be predicted since their sequences are similar, although not identical. The similarity between the independently folded monomer minimum energy structure and the monomer native structure in the dimer agrees with the experimental results, showing a conformational drift, as described before (§ 6.2). Indeed, as Figure 6.8 illustrates, the single monomer minimum–energy structure is just a rearrangement of the monomer side pertaining to the monomer–monomer interface in the dimer native structure.

Figure 6.9 shows the energy landscape of the full dimer (light grey dots) in comparison with the energy landscape of the two non–interacting monomers (black dots). The full dimer landscape exhibits a well pronounced funnel shape, while the landscape of the two non–interacting monomers has a less pronounced one. In order to compare data of interacting and non–interacting monomers, in Figure 6.9 a different definition of distance between two configurations is adopted: here the distance $D(\Gamma_{AB}, \Gamma^n_{AB})$ be-

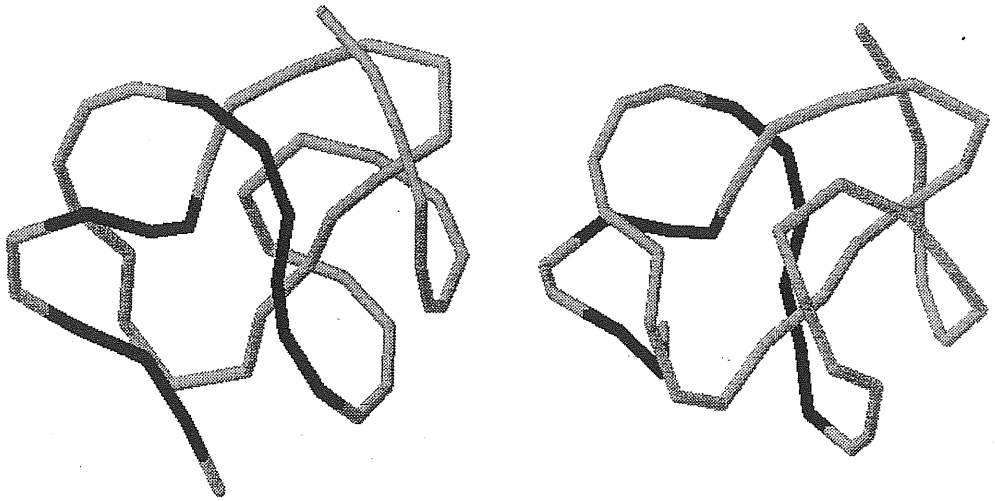

Figure 6.8: Structure of a monomer in the dimeric native state (a). The some monomer find a similar structure as lowest energy conformation (b) when folded independently from the other subunit. Structure (a) and (b) are differing each other for about 2 Å. The main difference between the two structure is in the monomer side exposed at the interface in the dimer native state.

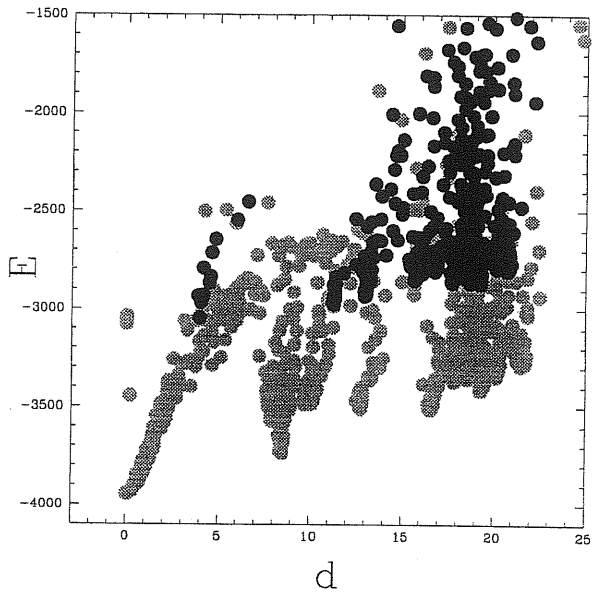

Figure 6.9: Energy of the dimer versus the distance from the dimeric native state (grey dots). Black dots represent the energy of the two non–interacting monomers (i.e. folded at a sufficiently far from each other). Among the non–interacting monomer structures, the lowest energy one is located at the closest point to the native dimer structure. Distances between two structure are measured here as explained in the text (§ 6.3.3).

tween a given configuration Γ_{AB} of chains A and B and the dimer native state Γ^n_{AB} is defined as the sum $D = D^A + D^B$ of the distances of each monomeric structure Γ_A and Γ_B to the corresponding structure Γ^n_A and Γ^n_B in the dimer native state.

6.3.4 Results and perspectives

Using an extremely simplified model we were able to recover characteristic properties of a dimeric protein, such as the multiple jump behavior of inter–monomer force and the conformational drift effect. We thus expect that a more careful and detailed study of the protein aggregation process could give very interesting results. In order to perform a more realist study of the issues presented in this Chapter we are currently working with a C_α model including 20 kinds of amino acids, using the Miyazawa-Jernigan [29] values for the potential parameters. Improved models (as the two beads model discussed in Chapter 5) will be also used in a near future, in order to develop techniques able to predict experimental results of some specific multimeric proteins.

Acknowledgments

My deepest thank is to Amos, my supervisor during these three years in Sissa. He introduced me to this field of research and during these years he has been providing a lot of fruitful opportunities to improve my work, my attitude to research, and myself. His suggestions have been always precious. I hope that our collaboration will continue, despite of the distance.

I am mostly indebted to Erio Tosatti. He always transmitted me enthusiasm towards research and, each time I talked to him, he gave me important advice and increased my confidence. I am really sorry that I have been not able to interact more with him during my stay in Sissa.

Jayanth Banavar is one of the kindest person I met during the work. I learned a lot from the collaboration with him and he has often been a reference person during these years.

I am very grateful to Michele, who, besides being a dear friend, has always offered me an important support in the work. His peace of mind in tackling problems has often been an help and an example for me.

I would like to acknowledge Eytan Domany for his kind hospitality at Weizamnn Institute of Science and for the fruitful discussions we had there.

I want to thank Flavio Seno for being a guide at the beginning of my work and Paolo Carloni for having supported me during the last stage.

The collaboration with Jort van Mourik was important, albeit short, especially for

stimulating me in my research.

Regretfully I did not fully exploit the presence of Cristian Micheletti. His reliability and efficiency in the work has been a great example and stimulus for me.

I am also sorry that the interaction with the "new conscription" working in Protein Folding Problem (i.e. Gianni, Andrea, Giulia and Stefania) has been so limited. I hope there will be room for it in the future.

A special thank to Enrica Galli Fossati for having sent me a lot of interesting papers (by fax!).

Acknowledgments are also due to all the SISSA staff, secretaries and system managers for their kindness and efficiency.

I am deeply indebited to a huge number of people, not directly involved in my work, for giving me the possibility to enjoy the time I spent in Trieste. They managed (someone voluntary, someone involuntary) to transform these last three years into a fundamental experience of my life, not only from the scientific point of view. I prefer to show them my gratitude and friendship elsewhere and otherwise.

Bibliography

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

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

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

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

[5] M. Levitt & C. Clothia, Nature **261**, 552 (1976).

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

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

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

[9] T. Garel, H. Orland and D. Thirumalai, New Developments in Theoretical Studies of Proteins, R. Elber (ed.), World Scientific, Singapore, (1996).

[10] B. Derrida, Phys. Rev. Lett. **45**, 79 (1980); B. Derrida, Phys. Rev. **B 24**, 2613 (1981).

[11] N. Gō, J. Stat. Phys. **2**, 413 (1983).

[12] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).

[13] P.E. Leopold, M. Montal and José N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).

[14] J.N. Onuchic, Z. Luthey-Schulten and P.G. Wolynes, Ann. Rev. Phys. Chem. **48**, 545 (1997).

[15] K. A. Dill & H. S. Chan, Nature Struct. Biol. **4**, 10 (1997). .

[16] Z. Guo & D. Thirumalai, J. Mol. Biol. **263**, 323 (1996).

[17] H. Frauenfelder, S.G. Sligar and P.G. Wolynes, Science **254**, 1598 (1991);.

[18] M.R. Betancourt and J.N. Onuchic, J. Chem. Phys. **103**, 773 (1995);.

[19] J.N. Onuchic, P.G. Wolynes and N.D. Socci, Proc. Natl. Acad. Sci. U.S.A. **92**, 3626 (1995);.

[20] N.D. Socci, J.N. Onuchic, and P.G. Wolynes, J. Chem. Phys. **104**, 5860 (1996);.

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

[22] J.D. Bryngelson, J.N. Onuchic and P.G. Wolynes, Proteins, Struct., Funct. and Genetics **21**, 167 (1995);.

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

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

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

[26] K.A. Dill, S. Bromberg, S. Yue, K. Fiebig, K.M. Yee, P.D. Thomas and H.S. Chan, Protein Science **4**, 561 (1995).

[27] S. Tanaka and H.A. Scheraga, Macromolecules **9**, 945 (1976).

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

[29] S. Miyazawa and R.L. Jernigan , Macromolecules **18**, 534 (1985); J. Mol. Biol. **256**, 623 (1996).

[30] M.J. Sippl, J. Mol. Biol. **213**, 859 (1990).

[31] M. Hendlick, P. Lackner, S. Weitckus , H. Floeckner, R. Froschauerer, K. Gottsbacher, G. Casari and M.J. Sipp, J. Mol. Biol **216**, 167 (1990).

[32] D.A. Hinds and M. Levitt, Proc. Natl. Acad. Sci. USA **89**, 2536 (1992).

[33] D.T. Jones, W.R. Taylor and J.M. Thorton, Nature **358**, 86 (1992).

[34] A. Godzik and J. Skolnick , Proc. Natl. Acad. Sci. USA **89**, 12098 (1992).

[35] S.H. Bryant & C.E. Lawrence, Proteins: Struct. Funct. Genet. **16**, 92 (1993).

[36] M. Wilmanns and D. Eisenberg, Proc. Natl. Acad. Sci. USA **90**, 1379 (1993).

[37] M. Pellegrini , S. Doniach, Proteins: Struct. Funct. Genet. **15**, 436 (1993).

[38] K. Nishikawa and Y. Matsuo, Protein Eng. **6**,811 (1993).

[39] M.J. Sippl, J. Comput. Aided Mol. Des. **7**, 473 (1993); Proteins **17**, 355 (1993); M.J. Sippl, M. Jaritz, in *Protein structure by distance analysis*. Edited by H. Bohr and S. Brunak , (1994) Amsterdam IOS press.

[40] H. Flockner, M. Braxenthaler, P. Lackner, M. Jaritz, M. Ortner and M.J. Sippl, Proteins: Structure, Function and Genetics **23**, 376 (1995).

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

[42] G.M. Crippen, Biochemistry **30**, 4232 (1991).

[43] F. Seno, A. Maritan and J.R. Banavar, Proteins: Struct. Funct. Genet. **30**, 244 (1998).

[44] J. van Mourik, C. Clementi, A. Maritan, F. Seno and J. R. Banavar, Determination of Interaction Potentials of Amino Acids from Native Protein Structures: Test on Simple Lattice Models, J. Phys. Chem., submitted (1998) [Preprint cond-mat/9801137].

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

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

[47] M.P. Morrisey and E.I. Shakhnovich, Folding and Design **1**, 391 (1996).

[48] G.M. Crippen, Proteins: Struct. Funct. Genet. **26**, 167 (1996).

[49] Rosenblatt, F. (1962). *Principles of neurodynamics*, Spartan books, New York.

[50] Minsky, M. L. & Papert, S. A. (1969). *Perceptrons*, MIT press, Cambridge MA.

[51] Krauth, W. & Mezard, M. (1987). J. Phys. A. **20**, L745-L752.

[52] Mirny, L. A. & Shakhnovich, E. I. (1996). J. Mol. Biol. **264**, 1164-1179.

[53] E. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).

[54] A.M. Gutin, V. Abkevich and E. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **92**, 1282 (1995).

[55] G. M. Crippen, J. Mol. Biol. **260**, 497 (1996).

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

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

[58] V. Abkevich, A.M. Gutin and E. Shakhnovich, J. Chem. Phys. **101**, 6052 (1994); J. Mol. Biol. **252**, 460 (1995).

[59] T. Garel, H. Orland and D. Thirumalai, New Developments in Theoretical Studies of Proteins, R. Elber (ed.), World Scientific, Singapore (1996).

[60] Goldstein, R. A., Luthey-Shulten, Z. A. & Wolynes, P. G. (1992). Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 9029-9033.

[61] H. Nymeyer, A. E. García and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A., in press (1998).

[62] J.E. Shea, Y.D. Nochomovitz, Z. Guo and C. L. Brooks III, J. Chem. Phys., submitted (1998).

[63] Z. Guo, D. Thirumalai and J. D. Honeycutt, J. Chem. Phys, **97**, 525 (1992).

[64] T. Veitshans, D.K. Klimov and D. Thirumalai, Folding & Design **2**, 1 (1996); R. M. Scheek, W.F. Van Gunsteren and R. Kaptein, Methods in Enzymology **177**, 204 (1989).

[65] L. Casetti, Physica Scripta **51**, 29 (1995).

[66] J. L. Lebowitz, J. K. Percus and L. Verlet, Phys. Rev. **153**, 250 (1967).

[67] Vendruscolo, M., Kussell, E. & Domany, E. Folding & Design **2**, 295 (1997).

[68] J.D. Honeycutt and H.C. Andersen, J. Phys. Chem. **91**, 4950 (1987).

[69] Y. Zhou, M. Karplus, J.M. Wichert and C.K. Hall, J. Chem. Phys. **107**, 10691 (1997).

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

[71] W. Kabsch, Acta Cryst. **A 32**, 922 (1976); Acta Cryst. **A 34**, 827 (1978).

[72] see, for example: Z. Dauter, L.C. Sieker, K.S. Wilson, Acta Cryst. **B 48**, 42 (1992).

[73] C. Clementi, A. Maritan, J.R. Banavar, in preparation.

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

[75] E. Shakhnovich and A.M. Gutin, Protein Eng. **6**, 793 (1993); E. Shakhnovich, Phys. Rev. Lett **72**, 3907 (1994).

[76] The case of a random (not designed) sequence has been already studied in G. Iori, E. Marinari, G. Parisi, J. Phys. **A 24**, 5349 (1991); G. Iori, E. Marinari, G. Parisi and M. V. Struglia, Physica **A 185**, 98 (1992), and corresponds to an energy landscape similar to a spin glass.

[77] In chapter 4 we show that results obtained using Lennard-Jones potential are not too sensitive on the particular Lennard-Jones shape: the same results can be obtained using a suitable square-well potential.

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

[79] A. Godzik, J. Skolnik and A. Kolinski, Protein Eng. **6**, 801 (1993).

[80] Holm, L. & Sander, C. J. Mol. Biol. **233**, 123 (1993).

[81] L. Mirny, and E. Domany, Proteins: Struct., Funct. and Gen. **26**, 391 (1996).

[82] Hinds, D. A. & Levitt, M. J. Mol. Biol. **243**, 668 (1994).

[83] Vendruscolo, M. & Domany. E. Efficient Dynamics in Contact Maps Space, Fold. Des. **3**, xxx (in press) (1998).

[84] Vendruscolo, M. & Domany. E. (1998). Efficient Dynamics in Contact Maps Space, in preparation.

[85] C. Clementi and A. Maritan and J.R. Banavar, Folding, Design and Determination of Interaction Potentials Using Off-Lattice Dynamics of Model Heteropolymers, Phys. Rev. Lett., in press (1998) [Preprint cond-mat/9802269].

[86] M. Vendruscolo and E. Domany, Contact Potential are Unsuitable for Protein Folding, J. Chem. Phys., submitted (1998).

[87] , F. B. Sheinerman and , C. L. Brooks III, Proc. Natl. Acad. Sci. USA **95**, 1562 (1998).

[88] C.M. Yip, C.C. Yip and M.D. Ward, Biochemistry **37**, 5439 (1998).

[89] B. Park & M. Levitt, J. Mol. Biol. **249**, 493 (1995).

[90] A. Monge, E.J.P. Lathrop, J.R. Gunn, P.S. Shenkin and R.A. Friesner, J. Mol. Biol. **247**, 995 (1995).

[91] B. Park & M. Levitt, J. Mol. Biol. **258**, 367 (1996).

[92] A. Godzik, A. Kolinski and J. Skolnick, J. Mol. Biol. **227**, 227 (1996).

[93] J. G. Saven and P. G. Wolynes, J. Mol. Biol. **257**, 199 (1996).

[94] A.A. Zamyatin, Prog. Biophys. Mol. Biol. **24**, 107 (1972); C. Chotia, J. Mol. Biol. **105**, 1 (1975).

[95] F.M. Richards, J. Mol. Biol. **82**, 1 (1974).

[96] J.M. Ponder and F.M. Richards, J. Mol. Biol. **193**, 775 (1987).

[97] C. Clementi, M. Vendruscolo, A. Maritan and E. Domany, Contact map Dynamics Meets Molecular Dynamics, in preparation (1998).

[98] M.M.Teeter, Proc. Nat. Acad. Sci. USA, **81**, 6014 (1984).

[99] W.A.Hendrickson, M.M.Teeter, Nature, **290**, 107 (1981).

[100] M.M.Teeter,W.A.Hendrickson, J. Mol. Biol.,**127**, 219 (1979).

[101] C. Wilson and S. Doniach, Proteins **6**, 193 (1989).

[102] A. Kolinski and J. Skolnick, Proteins **18**, 353 (1994).

[103] M.H. Hao and H.A. Sheraga, Proc. Natl. Acad. Sci. USA **93**, 4984 (1996).

[104] G. Weber, Biochemistry **25**, 3626 (1986).

[105] H.C. Andersen, J. Comput. Phys. **52**, 24 (1983).

[106] J.P. Ryckaert, G. Ciccotti and H.J.C. Berendsen, J. Comput. Phys. **23**, 327 (1977).

[107] P. Hinterdorfer, W. Baumgartner, H.J. Gruber, K. Schilcher and H. Schindler, Proc. Natl. Acad. Sci. USA **93**, 3477 (1996).

[108] S. Allen, X. Chen, J. Davies, M.C. Davies, A.C. Dawkes, J.C. Edwards, C.J. Roberts, J. Sefton, S.J.B. Tendler and P.M. Willimans, Biochemistry **36**, 7457 (1997).

[109] E.L. Florin, V.T. Moy and H.E. Gaub, Science **264**, 415 (1994).

[110] V.T. Moy, E.L. Florin and H.E. Gaub, Science **266**, 257 (1994).

[111] G.U. Lee, L.A. Chrisey and R.J. Colton, Science **266**, 771 (1994).