ISAS - INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

Statistical Mechanics Approach to Protein Design

Thesis submitted for the degree of
Doctor Philosophiæ

Candidate:
Andrea Rossi

Supervisors:
Prof. Amos Maritan
Dr. Cristian Micheletti

October 2000

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

TRIESTE

Contents

Introduction 3

1 Protein folding and design 9
- 1.1 The building blocks of proteins 10
- 1.2 Determining the protein structure 15
- 1.3 The protein folding problem 18
- 1.4 The protein design problem 21
- 1.5 Summary 23

2 Statistical mechanics approach to protein design 25
- 2.1 A coarse-grained model for proteins 26
- 2.2 Which potentials to use? 30
- 2.3 Designing sequences by energy minimization 33
- 2.4 A rigorous approach to protein design 38
- 2.5 The first order cumulant approximation 39
- 2.6 A Monte Carlo estimate for the conformational free energy . . 41
- 2.7 Conclusions 44

3 An iterative method for protein design 47
- 3.1 Formulation of the design strategy 48
- 3.2 The iterative strategy 50
- 3.3 Implementation and test of the iterative procedure 52
- 3.4 Results and discussion for the iterative method 54
- 3.5 A first geometrical criterion: maximum overlap 57
- 3.6 Similarity of conformations in terms of burial 60

3.7 Encodability, designability and burial degeneracy 63

4 Knowledge-based approach to protein design **73**

4.1 Protein modeling 74

4.1.1 Two- and three-body energy functions 74

4.1.2 Partitioning the 20 amino acids into classes 76

4.2 Learning the interaction potentials 77

4.2.1 A new theoretical approach 77

4.3 Designing PDB structures 81

4.3.1 The design strategy 81

4.3.2 Homologous sequences and comparison of similarities . 84

4.3.3 Are extremized sequences the best? 85

4.3.4 Homologous sequences and conserved sites 87

4.3.5 Data for barnase and chymotrypsin inhibitor 92

4.4 Summary 95

Conclusions **97**

A Determination of the weight functions $\Delta^{(2)}$ **101**

A.1 Two-body energy 101

A.2 Three body energy 103

B Perceptron learning of the optimal potentials **105**

Bibliography **109**

Introduction

Proteins belong to the group of biopolymers, which also comprises nucleic acids (DNA, RNA) and polysaccharides. While the latter are evolved to perform a particular task—i.e. information storage for nucleic acids and energy storage for polysaccharides—, proteins can cover an unlimited and amazing number of different functions in living beings. In addition to catalyzing almost all the biochemical reactions, proteins are responsible for the transport of ions, can form structures as skin and hairs, control and repair genetic material, regulate the transcription of DNA determining molecular biosynthesis and have a fundamental role in macroscopic mechanisms as muscle contraction.

One of the secrets of this functional diversity and omni-presence of proteins in the organism is the complex relationship between the *sequence* of amino acids forming the polypeptide chain and the associated three-dimensional *structure*. The sequence of amino acids characterizes uniquely the protein from a chemical point of view and corresponds to the order in which amino acids are assembled together during the biosynthesis process. This order is always the same for a given protein and it is well defined and encoded in a specific segment of DNA, called *gene*. The sequence of a protein is the translation of the associated gene and for each protein in the organism there is a gene that codify for it. All the information that an organism needs to synthesize a protein, and hence to switch on the associated biological activity, is encoded in the gene and might be inferred by the only knowledge of the sequence of nucleotides forming the gene. There are four different kinds of nucleotides, i.e. Adenine, Guanine Cytosine and Thymine for DNA; in RNA Thymine is replaced by Uracil The translation code to determine a protein

sequence starting from the sequence of bases corresponding to the gene is called *genetic code* and it is clearly understood: to each codon, i.e. triplet of nucletides, corrispond a well defined amino acid. Though the details of translation and the cell machinery involved in protein synthesis may be complex, it is at least algorithmically simple to imagine sequential processing in which codons are read one by one and the corresponding amino acids are added to the growing protein chain.

When the protein is synthesized, it is not yet biologically active. In order to become active the polymeric chain has to fold into a unique and specific three-dimensional structure. It is the stability of the protein in this (native) conformation to guarantee the correct effects of the protein in the cell. Usually, such effects occur thanks to specific interactions between the protein and other molecules or macromolecules. It is the geometrical shape of the protein, together with the specific biochemical interactions in some special region of the protein, to confer a specific effect, the biological function, to the protein. Some famous diseases are just caused by an incorrect folding of the involved protein, which is no more able to work in the desired manner (Mediterranean anemia, mad cow disease, and so on). To summarize, information contained in the gene are expressed through a three-step mechanism: the gene encodes the sequence, which specifies the structure, which, finally, determines the biological function. A schematic representation of gene expression is represented in fig. 1.

In order to understand a particular biochemical reaction in the cell, it is necessary to know all the reagents participating to the reaction. Up to now, there is no standard method in structural biology to predict the structure of a known protein sequence. Experimental determination of structures, basically X-ray crystallography and NMR, are hard and expensive [11, 4]. In contrast, amino acid sequences can be determined very fast and the number of sequences that have been determined is at least an order of magnitude greater than the number of protein structures [23]. Furthermore, the genomic revolution promises to increase the gap between known sequences and known structures and will make available, in few years, a large number of protein sequences. It follows that the intermediate step sequence→structure is a sort

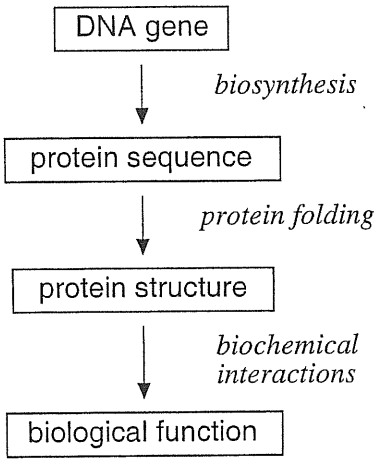


Figure 1: Gene expression occur through many intermediates: the gene is translated in an amino acid sequences, which will be biologically active only after reaching the three-dimensional conformation.

of bottleneck in the understanding how genes are expressed in the cell.

How proteins can find their three-dimensional structure, is interesting even from a theoretical, and not only applicative, point of view. The number of conformations that could house a protein sequence is exponentially large and it is intriguing [34] that proteins can find their native conformation in a very short time (1ms to few seconds). It has been shown [2] that at least a large variety of proteins can unfold and re-fold in their three-dimensional structure without the assistance of any biological machinery (this is not true for biosynthesis, for example, that is working thanks to a complex biological mechanism). These studies confirm, instead, the thermodynamic hypothesis: the native structure is the conformation in which the protein has a free energy minimum. According to this discovery, proteins fold in their three-dimensional structure through a spontaneous and, to some extent, reversible physical process, until the lowest free energy conformation is reached. Proteins in their native conformations, together with the acqueous solvent in which they are embedded, constitute a system in thermodynamic equilibrium. It follows that the key to understand *protein folding* is not a well-defined code, like the genetic code, controlled and regulated by biological

mechanisms in the cell, but it can be understood by the laws of physics.

The above considerations suggest that physics should be able to answer a lot of questions that are crucial to understand the mechanism of the protein folding: How does it fold a protein? Are there some regions of the chain that have a key-role in the folding process? All the amino acids contribute with the same weight to the protein stability? Are there some crucial intermediate conformations between the new-formed polypeptide chain and the final native conformation? In principle, one could attempt to answer these question by integrating numerically the equations of motion for proteins in the solvent and observing the evolution of the polypeptide chain. However, all atoms simulations are very expensive and cannot be applied, up to now, to macromolecules as large as proteins.

For this reason, a lot of simplified models have been proposed for studying protein folding, both by computer simulations [20] and analytical calculations [6]. Some of these models arose from the comparison of proteins with other systems that have been widely studied in physics: in particular spin glasses and random heteropolymers. Frustration, a key concept in spin glasses, is realized in physical polymeric systems like proteins, basically, by the chain connectivity. Such constraints prevents the chain to reach unphysical lower energy conformations, where each amino has favorable interaction with its neighbors. Proteins can be considered as special cases of random heteropolymers, whose glassy dynamics is strongly influenced by the chain connectivity. When the temperature is lowered, random heteropolymers are frozen in a unique conformation; however, in contrast with proteins, the conformation is random and depends on an uncontrollable number of conditions In proteins, instead, the final conformation is completely defined by the sequence, the folding is fast and the transition between the unfolded and the folded state is more abrupt with respect to random heteropolymers. It is conceivable that protein sequences have been selected during evolution for fast folding. This selection principle is also called "principle of minimal frustration": among all heteropolymers which it is possible synthesize using the twenty amino acids, protein sequences are the ones with the smallest frustration in their native state [5]. The energy (or free energy) landscape has been smoothed

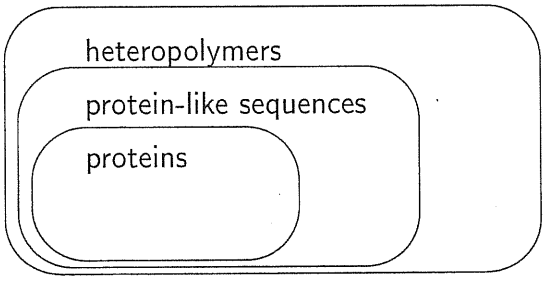


Figure 2: Among all the possible sequences only a fraction of them are protein-like, i.e. fold rapidly in a unique and well-defined three-dimensional structure. Natural proteins are an extremely small subset of all the protein-like sequences. A selection procedure should find the correspondence among protein-like sequences and the structure which they encode.

by the selection procedure carried on by natural evolution and traps and kinetic barrier have been, in part, eliminated in order to aid fast folding [51]. Furthermore interresidue interactions have been optimized through changes in the amino acid sequence in order to optimize the thermodynamic stability of the protein chain.

Once the principle of minimal frustration has been accepted, a lot of questions remain unsolved: How has Nature selected protein sequences? How many heteropolymer sequences show protein-like features? How large is the fraction of proteins that fold quickly and reversibly into a given structure? Which kind of sequences fold into a given target structure? The last two questions, in particular, are extremely important. They are interesting not only from a theoretical point of view, but also from an experimental and technological one. In experimental *protein design* the goal to synthesize proteins that are able to fold into a desired stable conformation, has very few successful examples in the literature, at least in comparison with the performed efforts [52, 56, 30, 83, 1]. In computational all-atoms protein design, where the prediction is usually followed by an experimental validation test, the main focus has been on problems of optimal packing of amino acids [55, 21, 13, 22, 75, 74].

Approaches to protein design by using analytical tools or simulations on

coarse-grained models clarified, at least in our opinion, several aspects of protein design.

In this thesis, the problem of finding an optimal sequence selection procedure, that is a central problem in protein design, is faced. Strategies for protein design based on rigorus statistical mechanics principles will be implemented both on simplified models of proteins and on more realistic models. The original part of our work is grouped in chapters 3 and 4.

The plan of this thesis is the following. In the first chapter we introduce the most important and basic concept related to protein folding and design. In the second chapter, after introducing some standard lattice models of proteins and heteropolymers, the most important methods of protein design present in the literature are described. In the third chapter we will introduce a novel iterative procedure for protein design and it will be applied to lattice protein models [58]. A different approach based on geometrical criterion [49] will also be presented. In the fourth chapter, we will implement an approximated approach in order to design real protein structures [59]. In this case, it has been possible to compare our designed sequences with real sequences, whose native states are known. The good correlation between natural sequences and designed sequences indicates that the method is very promising.

Chapter 1

Protein folding and design

One of the secrets of the biological diversity and omni-presence of proteins in the organism is the complex relationship between the *sequence* of amino acids forming the polypetide chain and its three-dimensional *structure*.

This relationship can be studied from two opposite and complementary points of view. On one hand it is of fundamental importance to determine the structure of proteins present in the organisms, since it would help greatly in understanding the role of the proteins. This problem, generically called *protein folding*, is based on the Anfinsen's discovery: a protein can fold fast and reversibly into a unique three dimensional structure, the native state, corresponding to a free energy minimum. The basic idea is that the structure of a protein could be, in principle, obtained by integrating the equation of motion for the sequence in presence of the solvent. Direct applications of this method are not convenient, since proteins are large macromolecules with many degrees of freedom whose mutual interactions, due to the presence of the acqueous solvent are too complicated for the present computational capabilities.

The second approach is known as "inverse protein folding" and aims of finding the sequences whose native state is preassigned. This problem, although apparently involves a search in sequence space rather than in conformation space, is even more complex. Given a structure, only a microscopic fraction of all the viable sequences fold on that, and finding them is like searching a needle in a haystack. The problem of finding one or more se-

quences able to fold into a given structure is named *protein design*, and has fundamental applications in drug design.

The plan of this chapter is the following. In the first section we will briefly describe structural properties of amino acids, the building blocks of proteins, and some regular arrangements of them, known as secondary structures. The description of a protein through its sequence of amino acids is the simplest one and is sometimes referred as primary structure of the protein. However, the sequence does not contain information, by itself, on the three-dimensional structure of the protein, that is more conveniently described in terms of secondary structures. In the second section, the problem to determine the three-dimensional structure of proteins will be analyzed. As we shall see, theoretical predictions of protein structures succeed only in particular cases and experimental tools (X-ray and NMR) are often the only (expensive) methods to obtain information about the three-dimensional structures. In the last two sections, we will briefly review protein folding and protein design.

1.1 The building blocks of proteins

Proteins are macromolecules ranging from 1000 to more than 5000 atoms without apparent symmetries or regularities. Describing such large objects at the atomic level seems a quite discouraging, and in some cases useless, task. Fortunately, since 1958, when the first protein structure has been determined by X-ray crystallography, a number of recurrent structures and motifs have been discovered. In some cases, description of protein properties by these motifs is helpful and simplifies concepts, while, in other cases, a resolution at the atomic level is necessary. For example, secondary structures (see below) are very convenient to describe the architecture of proteins and to associate a protein to a family (channels, enzymes, antibody domain and so on).

At the lowest level of this hierarchy, there are the 20 amino acids, whose covalent structure is the base for the structure of proteins. Amino acids are bonded together to form a linear chain, through the peptide bond, which constitutes the backbone of the structure. Though the polymeric chain is flexible and can adopt, in principle, many different conformations, the in-

teractions among the different regions of the chain are such that only one conformation will be adopted by the protein under physiological conditions (temperature, pressure, pH). The order of amino acids placed along the chain is of fundamental importance, since changing it will dramatically change the interactions, destabilizing the native conformation.

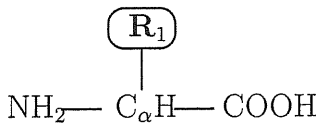
The sequence, i.e. the order in which amino acids are placed along the protein backbone, is the first level of complexity. It can be in fact represented by a one-dimensional string, where each letter is associated to one of the twenty types of amino acids (see table 1.1). For example the string

```
APLEPEYPGDNATPEQMAQYAAELRRYINMLTRPRY
```

is the amino acid sequence of the Bovine Pancreatic Hormone, a protein involved in the digestion of cow. Here we have used the single letter representation, though in other circumstances is more convenient the three-letter representation (see table 1.1).

The string above, does not tell us very much about the Bovine Pancreatic Hormone. Of course, to each letter we have to associate the covalent structure of every amino acid in the sequence. By doing so, we obtain a polymeric chain, that can assume, in principle, many different conformations compatible with steric constraints. To understand which conformations are allowed and which not, one has to know the covalent structures of amino acids and how they bind together to form the peptide chain.

To the α carbon atom, are bonded the aminic and carboxylic groups (NH_2 and COOH, respectively), the residue **R** and a hydrogen atom:



(1.1)

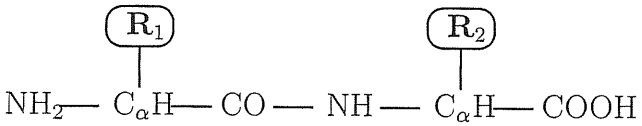
Apart from proline, whose carbon atom in the carboxilic group is bonded to the residue itself, the other amino acids differ only in the nature of the residue R. The number of atoms forming residues can vary from one—in glycine R is just one hydrogen atom—to a maximum of eighteen for Tryptophane and

Residue type			Frequency in proteins(%)
Alanine	ALA	A	8.3
Arginine	ARG	R	5.7
Asparagine	ASN	N	4.4
Aspartic acid	ASP	D	5.3
Cysteine	CYS	C	1.7
Glutamine	GLN	Q	4.0
Glutamic acid	GLU	E	6.2
Glycine	GLY	G	7.2
Histidine	HIS	H	2.2
Isoleucine	ILE	I	5.2
Leucine	LEU	L	9.0
Lysine	LYS	K	5.7
Methionine	MET	M	2.4
Phenylalanine	PHE	F	3.9
Proline	PRO	P	5.1
Serine	SER	S	6.9
Threonine	THR	T	5.8
Tryptophane	TRP	W	1.3
Tyrosine	TYR	Y	3.2
Valine	VAL	V	6.6

Table 1.1: List of the twenty amino acids with their frequency in proteins (data from [11]). Amino acids can be identified by a three-letter code or a one-letter code (second and third column, respectively.)

Arginine. Basically residues are formed by different combination of carbon and hydrogen atoms, but for some amino acids oxygen is present too. In Cysteine and Methionine a sulfide atom is present and it is responsible for stabilizing three-dimensional structures through a disulfide bridge.

While the residue characterizes the chemical-physical properties of the amino acid type, the aminic and the carboxilic groups have an important role to connect amino acids in a polymeric chain. When two amino acids are hydrolised, the aminic group and the carboxylic group of different amino acids form a covalent bond following the reaction $NH_2+COOH \rightarrow NHCO+H_2O$.



(1.2)

The bond between the carbon and the nitrogen is called peptide bond and, since it is a partial double-bond, rotations along this axis are very rare (except 180°). Rotations are, instead, allowed along the single bonds between C_α and N and between the two carbon atoms, as for as steric clashes do not occur. As illustrated in figure 1.1, rotations along this axis are represented by two torsional angles called ϕ and ψ, respectively. Since bonds between nearest neighbours atoms are not aligned, these rotations cause a conformational change in the polypeptide chain.

Though ϕ and ψ, also called dihedral angles, can in principle assume all the values in $[-\pi, \pi]$, some values are more likely than others. In particular, some values are never allowed due to steric reasons, since they would correspond to an overlap of atoms of the residue, or side-chain, with atoms of the backbone. The permitted values of ϕ and ψ were first determined by Ramachandran and co-workers [57], using hard-sphere models of the atoms and fixed geometries of the bonds. The permitted values of ϕ and ψ are usually indicated on a two-dimensional map of the (ϕ, ψ) plane, what has come to be known as a Ramachandran plot. Since the size of the residue depend strongly on the amino acid type, Ramachandran plots of different amino acids, are different. In particular, glycine, which has the smallest

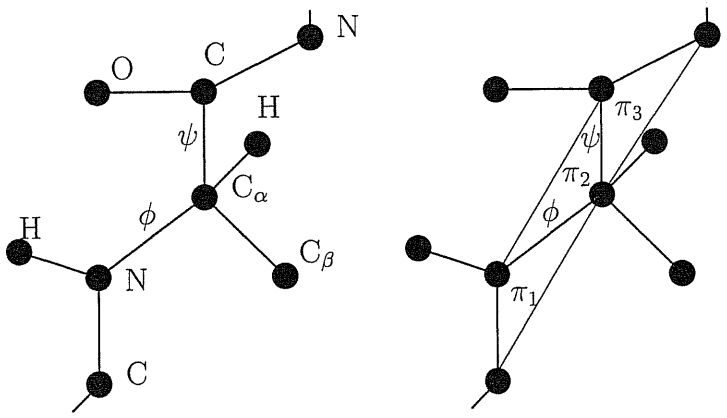


Figure 1.1: Protein flexibility is due to the presence of single bonds along the main chain between the nytrogen atom and C_α (ϕ angle) and between C_α and carbon atom (ψ angle). Angles ϕ and ψ are defined as the dihedral angles formed by planes π_1 and π_2 and by π_2 and π_3, respectively.

residue, has a Ramachandran plots with several allowed regions, indicating a flexibility unknown to other amino acids.

In real proteins, not all the allowed regions of the Ramachandran plots are equally likely. A statistical analysis of protein structures shows that some regions of the (ϕ, ψ) plane are more populated than others. The most populated region corresponds to angles around $(-60°, 50°)$. Several amino acids with such values of (ϕ, ψ), takes part to helical structures This is called α-helix and it is a motif quite recurrent in proteins. Each turn in the helix is formed on average by 3.6 amino acids, the i-th amino acid being in spatial contact with the $(i+3)$-th and with $(i+4)$-th one.

Another recurring secondary structure is the extended conformation, or β-strand, that is associated to the angles $(\phi, \psi) \approx (-130°, +120°)$. Extended conformations are frequently found associated together to form β-sheets. It is possible to distinguish between parallel and anti-parallel sheets. In the first case if the i-th and j-th amino acids are in contact then $i+1$-th and

$j+1$-th will be still in contact. In the second case it will be true for the $i+1$-th and $j-1$-th ones.

Secondary structures are assembled together to form more complex structures by turns and loops. In the first case, an amino acid, usually glycine, that is the smallest residue, makes a tight turn which changes completely the direction of the backbone. In loops, the change of direction is more gradual, being distributed over several amino acids. Finally, random coil are protein regions for which amino acids do not have a definite value.

Secondary structures contain information on the structure of the protein and can be represented by a one-dimensional string. For example, for the Bovine Pancreatic Hormone, to the string representing the sequence we can associate a string representing the secondary structure amino acids belong to:

```
APLEPEYPGDNATPEQMAQYAAELRRYINMLTRPRY
RRRRRRRRSSRSSTTHHHHHHHHHHHHHHHHHRRRRR
```

where H=helix, S=bend or loop, T=turn and R=random coil ([29, 62]). In fig. 1.1 secondary structures are highlighted by the cartoon scheme of visualization.

1.2 Determining the protein structure

One of the most important challenges in understanding biological reactions occurring in organisms is the determination of the structure of the molecules participating to the reaction. This is true especially for proteins, for which the structure has been selected by evolution for a specific biological task. In some cases, it is just the geometrical shape that contains important information on the function, especially when a cavity in the structure is complementary to the geometrical shape of another macromolecule ligand (docking problems). More often, the geometrical shape can give only generic indications where the binding site is located, and only detailed electrostatic calculations can solve the docking problem.

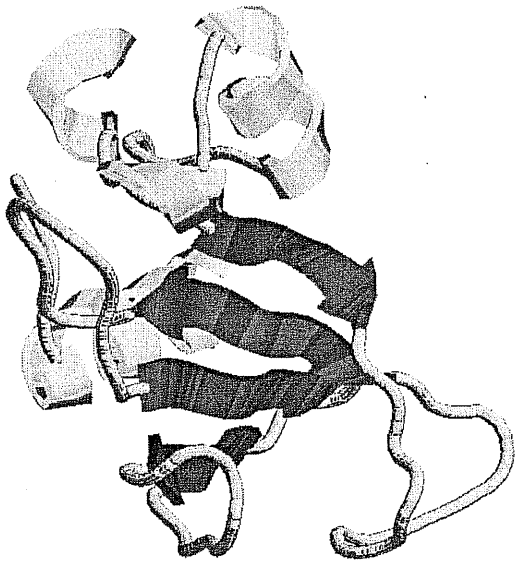


Figure 1.2: A "cartoon" representation for the protein 1a2p. Secondary structures are highlighted by a special pictorial scheme: helices are represented by a light-grey cartoon, strands by dark-grey arrows and random coil and loops by a tube.

While the structure is important for understanding the function of the protein, its experimental determination is difficult and expensive. By contrast, it is very easy to determine the sequence of amino acids by experimental measurements (sequencing) or by translating the associated gene. The number of sequences that have been determined up to now is almost an order of magnitude larger than the number of structures, and the number of sequences that will be acquired per day is destinate to increase with the genomic revolution. It follows that one of the most important research field in bioinformatics and biophysics is the prediction of the structure of already known sequences. In principle this problem can be solved by following the dynamics of the protein embedded in the solvent (which has a fundamental role in driving the folding of the protein) on a computer and finding the lowest free energy conformation. However, the complexity of the atomic structure of a protein and the time scale on which the folding occurs, do not allow to

integrate the equations of motion.

A possible way out to overcome this kind of problems might be to use simplified descriptions of proteins in which amino acids interactions and steric constraints are described in an effective way. These models have received a lot of interest in the community of physicists. However, because of their lack of atomic details and their approximate description of interactions, simplified models are still far to be successfully applied in structure prediction.

Among the methods more reliable for structure prediction there are those based on homology modeling. Homology modeling deals with the problem to detect an homology, i.e. an evolutionary relationship, between the protein, for which the structure has to be determined, and proteins of known structure. Usually, homology is detected by aligning and comparing the target sequence —i.e. the sequence for which the structure has to be predicted— with the sequences of proteins of known structure. Such structures can be used as templates to make a "model", that will be carefully refined. A similar procedure has many advantages. First, it can be automated allowing many scientists to access model structures for proteins, whose structure has not yet experimentally determined. Second, for good model structures, usually when a high homology has been detected, an experimental determination could be not necessary. Third, it can be used on a large number of known protein sequences: it has been estimated that it is currently possible to model with useful accuracy significant part of approximately one third of all known protein sequences. Furthermore, the number of proteins of known structures is destined to increase.

The basic idea of homology modeling is that similar sequences, likely, have similar structures. Just to give an idea, similarity above 25% can be enough to produce a good model of the unknown structure. Usually, structures are more conserved than sequences by the evolution. This implies that two related sequences can share similar structures [10]. However, deciding on the base of sequence similarity if two structures are similar, is a very hard task.

For these reasons and for the importance that the structure has for molecular biologists, structure data have been collected in a unique big database, called Protein Data Bank (PDB). Since 1975, when PDB has been builded up,

a lot of structures of macromolecules (basically proteins) have been collected. In 1992 there were about one thousand of macromolecules structures. At the moment (June 2000) of the 12474 macromolecules structures (proteins, nucleic acids, carbohydrates) there are more than 11059 of protein structures. Most of the protein structure data were obtained by X-ray crystallography (9233) and by solution nuclear magnetic resonance (NMR) (1583) and only 243 are obtained by theoretical modeling. Structures deposited on PDB constitute an important source for molecular biologists and for people working in bioinformatics and biophysics.

1.3 The protein folding problem

In the higher living beings proteins are synthesized in the cytoplasm through a complex mechanism of biosynthesis. Once the sequence is synthesized the protein is not yet active. To become biological active it has to fold into a specific three-dimensional conformation, i.e. the native state. In principle, there are a lot of different conformations that the sequence can adopt. Assuming there are 3 different coarse-grained conformations per amino acid, the number of possible distinct conformations, for a protein with 100 amino acid (a relatively small protein), should be $3^{100} \approx 10^{48}$. Some of these conformations are not accessible, due to steric reasons. Nevertheless, even taking into account this observation, the number of physical conformations is enormous and the protein should take, to fold, a time larger than the universe age, to find the native state by a random exploration. How can the protein find his native conformation among a gigantic number of conformations? This question is known as Levinthal paradox, from the first that arose it ([34]).

A first attempt to answer this question, comes from the Nobel laureate Anfinsen and coworkers ([2]). Before their studies, the nature of the sequence-structure relationship was completely unknown: is the structure written in the sequence as a physical-chemical message or there is a biological machinery similar to the complex of enzymes regulating biosynthesis? Anfinsen's studies on the re-folding of ribonuclease showed clearly that protein sequences under physiological conditions can automatically find their

native state minimizing the free energy. In other words, proteins with their solvent constitute a physical system that is thermodynamically stable only in their native conformation. This "thermodynamic hypothesis" excludes the possibility, that proteins adopt their native conformation thanks to a complex biological machinery. Anfinsen itself, writes about the thermodynamic hypothesis ([2]):

> ...the three-dimensional structure of a native protein in its normal physiological milieu (solvent, pH, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, and other) is the one in which the Gibbs free energy of the whole system is the lowest; that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment.

The impact that this discovery had on molecular biology has been so strong that this paper is one of the most cited in the field. If the native state of a protein is the global minimum of the free energy, consequently it is possible to predict its structure just by simulating its dynamics by standard tools of physics. This is up to now, one of the fundamental still unsolved problems in biophysics. The complexity of the problem is mainly due to the size of proteins —proteins range from 100 to 500 amino acids, i.e. from 1000 to 5000 atoms— and to the difficulty to treat the solvent accurately. Integrating the Newton's equation numerically is possible by using supercomputers only for time intervals of few picoseconds. A striking result in this field has been obtained in 1998 by Duan and Kollman, who were able to follow the folding of a small protein for 1μs [17]. The time necessary for the simulation was 4 months on a 256 processors supercomputer corresponding to a total CPU time of 80 years! At the end of the simulation they observed the presence of an intermediate state in the folding pathways.

The result of Duan and Kollman, if it is important in the history of force field simulation, shows that we are still far from solving the protein folding problem by brute force, i.e. simulating the real dynamics of proteins until the global free energy minimum is found. Most of the proteins are much longer

than the one used in [17] and they would fold in a CPU time that is 10^3–10^6 million times much longer. Furthermore following biophysical processes by simulating them in full details does not necessarily mean to understand them.

In order to understand protein folding better a lot of simplified models have been proposed. Usually, simplified models have not, at the moment, the aim to find the native state of protein sequences, but focus on the dynamics and thermodynamics of such physical systems. This is the case, for example, of the Go-model [20], in which the knowledge of the native structure is the input of the model itself. The main idea is that and energetic bias towards the native state, without any realistic description of physical interactions, allow to study protein dynamics. In the simplest form, the energy function is defined as the negative of the number of contacts present in the native conformation. Two amino acids are said to be in contact, if the distance of their C-α's is less than a given cutoff, usually taken in the interval 6–8Å, and if they are not consecutive along the peptide chain. To each conformation is, then, associated an energy equals to the negative of the number of contacts shared with the native conformation. Since the energy is well specified and the native state is by definition the state with the lowest free energy, the only problem is how to define the dynamics. The basic feature of this model is "ultra-specificity", since the native conformation is by definition the ground state and an ergodic dynamics will reach always the native state. However, such model is interesting because it allows an analisys of the folding process of a well designed sequence and how dynamics is controlled by the topology of the native state.

Models for random heteropolymers can be used to study protein-like features. Usually, models that allow analytical calculations are too simple to capture the sequence-structure relationship, peculiar of proteins. Instead, models apt for numerical implementation, like the one described in next chapter, capture some of general features of proteins and allow a deeper study of protein folding and design. Proteins, in fact, can be considered special heteropolymers that have evolved for fast-folding into a unique and thermodynamically stable conformation [51, 67, 5, 31]. By contrast with the

Go-model, such models are completely unbiased, since fast folding and other protein-like features have to emerge through a suitable sequence selection [61, 67, 70, 8, 15]. In fact, heteropolymers at low temperature behave differently from proteins. They show a glassy dynamics and the state in which they fold depend on the initial conditions [5, 53]. Kinetic and energetic barriers prevent a easy access to the ground state and the search of the global minimum is more similar to that prospected by Levinthal [34].

This is not the case for sequences that show protein-like features: they fold through a two-state mechanisms rapidly and reversibly in the native state [15]. It follows, that a rigorous study of protein folding has to be preceded by a suitable optimization of the sequence: only for such a sequence physical properties, for example the type of transition in the native state, could be compared with real proteins. For example, the optimization of the thermodynamic gap between the native state and a generic random coil conformation seems to be a sufficient condition for fast folding [60]. Such optimization will be reviewed in next chapter, where different methods for protein design on simple models will be presented and discussed.

1.4 The protein design problem

Protein design deals with the problem to find how many and which kind of sequences fold on a given target structure. In principle the problem can be solved enumerating all the possible sequences and attempting to solve the protein folding problem for each of them. Obviously, a similar strategy is not practicable. The number of sequences that can be mounted on a structure formed by 100 amino acids is 20^{100}. Even having a crude algorithm able to decide if a sequence is compatible with the given structure in a very short time—let's say 1ms—it would take 10^{37} seconds to solve the problem. Furthermore, as we have seen previously, the structure prediction problem is far from a general solution even in an approximated way.

A simple but approximated solutions to protein design, has been proposed and verified by experiments by Hecht and coworkers [30, 83]. The method is based on the assumption that the hydrophobic force is the main

force driving proteins into their native conformation. The hydrophobic force is the propensity of hydrophobic amino acid to dislocate themself in buried regions and polar ones in regions exposed to the polar solvent. By specifying explicitly the sequence locations of hydrophobic and polar amino acids it should be possible to design sequences, whose stability in the target conformation is large. This strategy is based on the premise that formation of stably folded structures does not require the explicit design of specific inter-residue contacts; the precise packing of the three-dimensional jigsaw puzzle need not be specified a priori. This strategy is different from others used in computational protein design where the goal is the optimization of the packing of side-chains especially in the protein core.

Such a strategy has been applied, first, to a four helix bundle [30]. The four helix bundle is a common fold among natural protein and has been the target structure also in previous efforts in de novo protein design. The collection of protein sequences have been produced by generating a degenerate family of synthetic genes. Each gene encoded a different sequence, but all the sequences shared the same periodicity of polar and non-polar residues . Positions requiring a non-polar residue were filled by Phe, Leu, Ile, Met or Val, whereas positions requiring a polar amino acid were filled by Glu, Asp, Lys, Asn, Gln, or His. Since in the four-helix bundle there are 24 buried positions and 32 surface positions the sequence degeneracy can be simply evaluated: $5^{24} \times 6^{32} = 4.7 \times 10^{41}$. This is a number extremely small with respect to the number of possible sequences, when nobody restriction is used, i.e. $20^{24+32} = 7.2 \times 10^{72}$. The design strategy is based on the premise that a substantial fraction of the sequences fitting this pattern will actually fold into proteins that are compact, stable, and α-helical. By contrast, if one does not use any design criterion to select sequences, a very small (negligible) fraction of sequences will be expected to fold to the target structure, neither to a similar one.

In order to test this strategy, 48 sequences were generated by binary patterning amino acids following this criterion. Experimental determinations of the protein structure, as we have seen in previous sections, are long and expensive. For these reasons, a direct determination of their three-dimensional

structure has not been attempted. However, the ability of 29 of the 48 sequences to resist to proteolytic degradation suggests that they fold into stable globular structures. Other tests done on three of the remaining 29 sequences show that they possess a content of alpha-helix comparable with the four-helix boundle. In conclusion, at least a large fraction of the designed sequences 60% fold into globular α-helical folds.

If hydrophobic-polar patterning of sequences is not a sufficient condition for a successful design, it seems to be a good filter to reduce the gigantic number of sequences. On the other hand, a statistical analysis of protein structures show that only a fraction of amino acids (70%, [77, 43]) are binary patterned in protein sequences. It follows that burial of amino acids cannot be the only criterion to select sequences; in other words hydrophobic-polar patterning is not even a necessary conditions for protein design. There are two reasons to this, in part obvious, result. First, hydrophobic interactions are not the only ones. Stabilization in the native state is increased by hydrogen bond, polar forces between amino acids, van der Waals forces, and so on. Furthermore, the different size of side-chain can play an important role in discarding otherwise energetic favorable conformations. Second, optimization of sequences on the target conformations, in this case by binary patterning amino acids on the basis of information of the target structure, cannot be the most convenient strategy. It has been shown, with the use of simple models, that alternative conformation can play an important role in destabilizing sequences optimized to fold on the target conformations. This problem, called negative design, will be reviewed in next chapter. It will be shown how it works in simple models and it will be one of the main issues of this thesis.

1.5 Summary

In this chapter, we have introduced several concepts of fundamental importance to understand the motivations of this work. Here, we would like to summarize some of the most important concepts that will be useful in the following chapters.

protein sequence: is the sequence of amino acids, that uniquely defines the biological function of the protein.

protein structure: is the native conformation of the protein, i.e. the conformation in which the protein is biologically active. Protein structures can be determined by X-ray crystallography or by solution NMR and experimental data can be retrieved from PDB [23].

secondary structure: several consecutive amino acids can be arranged in highly regular structures, basically alpha-helix or beta-strand.

structure prediction: the possibility to predict the three-dimensional structure of proteins is of fundamental importance in molecular biology. An algorithm for structure prediction reads as input a protein sequence and returns as output the associated protein structure.

protein folding: globular proteins are physical systems that fold spontaneously and reversibly into globular conformations, the protein structure. The study of protein folding through standard methods of physics has attracted many physicists.

protein design: only a microscopic fraction of all the possible sequences fold on a preassigned protein structure. Giving a rule to determine which kind of sequences adopt a given target structure would be of fundamental importance in drug design.

Chapter 2

Statistical mechanics approach to protein design

In this chapter we will discuss the protein design problem in the contexts of simplified protein models. Such models treat a protein structure as a self-avoiding walk on a cubic lattice. Each vertex of the self-avoiding walk corresponds to an amino acid of the polypeptide chain. Interactions between amino acids are described by some coarse-grained contact potentials that can be extracted from real proteins [47, 48, 37, 71, 72, 76, 79, 82, 63, 64, 80, 16, 9, 38, 46] or assigned according to heuristic observations [33, 35]. During the coarse-graining procedure all the internal degrees of freedom have been integrated out, while the dependence on the coarse-grained conformation has been retained. By eliminating some "marginal" problems like packing of different side-chains or the detailed description of amino acid interactions—that are of fundamental importance in a experiment-based approach to protein design— one can focus on the major problem in protein design: the negative design.

This problem arises from the observation that sequences optimized for being stable in a target conformation, can be even stable in alternative conformations known as decoys. This competition between decoys and target structure can be solved by decreasing the fitness of sequences with respect to decoys structure, rather than optimize it on the target one. Negative

design has been understood only recently on the bases of statistical mechanics [14, 65]. Its solution relies on writing the right fitness or scoring function as a non-local energy function, i.e. depending on the target conformation as well on the alternative ones. Previous approximated methods proposed by Dill for the HP model (see below) [8, 77] and by Shakhnovich [69, 67, 70], give only a partial and unsatisfactory answer to negative design.

In this and in the following chapters, negative design will have a primary role. In the following sections we will review the main approaches to protein design on coarse-grained models of proteins and we will introduce a rigorous scoring function for protein design. In the next chapter, we will describe an iterative method to approximate in an efficient way such scoring function. Finally, in the last chapter, an approximated scoring function will be applied to real protein and designed sequences will be compared with sequences selected by evolution.

2.1 A coarse-grained model for proteins

In this section we describe a protein model that has been widely used in the literature for understanding global properties of proteins and that has been a key ingredient in my studies. A lot of questions have been faced by the use of this model, ranging from protein folding to protein design: How proteins fold? Which and how many sequences fold into a given structure? Why some structures are more encodable (or designable) than others? Though extremely simplified with respect to real proteins, such models contains a lot of features that can be identified in proteins and more in general in random heteropolymers.

In the model we are considering, interactions are described in an effective way as two-body contact interactions. Two amino acids are said to be in contact if the distance between their C-α's is less than a cutoff value, d_0,ranging between 6 and 8Å. For example, in our studies on protein design in chapter 4, we have done the choice $d_0 = 8$Å. Contacts are conveniently described in terms of a contact map. To a given conformation Γ, we can associate the

contact map $\Delta(\Gamma)$, such that:

$$\Delta_{ij}(\Gamma) = \begin{cases} 1 & d_{ij} < d_0 \\ 0 & \text{otherwise} \end{cases} . \tag{2.1}$$

where i and j are two sequence indices ($i = 1, \ldots, L$, where L is the number of amino acids of the sequence) and d_{ij} is their Euclidean distance in space. Contact maps are two-dimensional representations of protein structures and are frequently used by people determine or analyze protein structures. In fig. 2.1 there is a graphical representation of the contact map of barnase, whose structure is represented in fig. 1.1.

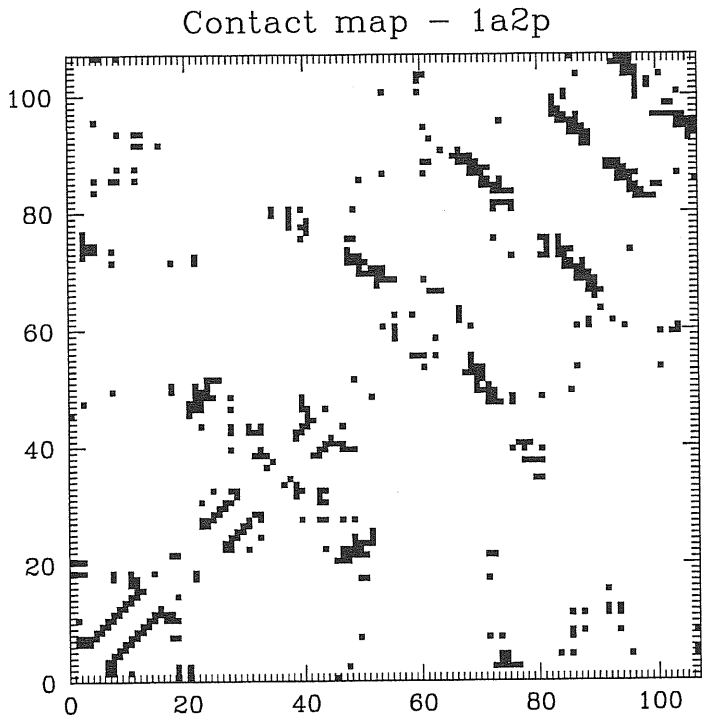


Figure 2.1: Contact map for the protein 1a2p. Both on the horizontal and vertical axis there is the sequence index. Contacts between amino acids are visualized by dark squares. Secondary structures appear as clusters parallel (α-helices or parallel β-sheets) or perpendicular (antiparallel β-sheets) to the diagonal.

Two amino acids that are in contact, contribute to the energy with an

amount that depends on the type of amino acids involved in the contact. Since in nature there are twenty different amino acids, the contact potentials can be stored in a 20×20 matrix, the symmetric contact potentials matrix B; thus there are 210 distinct interaction potentials. Let's σ be the sequence and σ_i and σ_j the amino acid types at position i and j. Then, the contribution to the energy from the i-j contact is $B(\sigma_i, \sigma_j)$. To a particular sequence σ, adopting a given conformation Γ, is associated an energy that is the sum over all the possible two-body interactions

$$H_\sigma(\Gamma) = \sum_{i<j} \Delta_{ij}(\Gamma) B(\sigma_i, \sigma_j) \ , \tag{2.2}$$

where $\sum_{i<j}$ is a shorthand for $\sum_{j=1}^{L} \sum_{i=j+1}^{L}$. Two different conformations, generally speaking, will have two different contact maps and, probably, two different energy values.

Since we are not considering internal degrees of freedom of amino acids, the energy function H has to be considered equivalent to the free energy (sometimes called internal free energy) mentioned in section the previous chapter. In the coarse-graining procedure, internal degrees of freedom are integrated out through a (partial) statistical average. In the present chapter and in the next ones we will call energy function, or Hamiltonian, the free energy associated to a coarse-grained conformation and we will distinguish it from the conformational (or global) free energy.

In the coarse-grained representation that we are considering, proteins are modeled as chains of dimensionless beads and sticks. In order to implement self avoidance due to the three-dimensional nature of amino acids, one has to introduce constraints on the angles formed by two consecutive sticks. We observe that, without such constraints, the model protein would collapse on itself, minimizing the energy in a trivial and unphysical way. The simplest way to introduce these constraints is to allow amino acids to lay only on the sites of a lattice. In particular, we will adopt the cubic lattice, that is simple to use and respects the correct peptide bond length (see fig. (2.2)). Two amino acids, consecutive along the chain, lay on two nearest neighbour sites on the lattice. In order to implement self-avoidance, one site will be occupied

by no more than one amino acid. Then, a generic conformation or protein structure is represented by a self-avoiding walk on the cubic lattice like the one in fig. (2.2) with a number of steps equal to the number of bonds in the polypeptide chain. Two self-avoiding walks are different if they cannot be super-imposed by any roto-translation and correspond to two different protein structures.

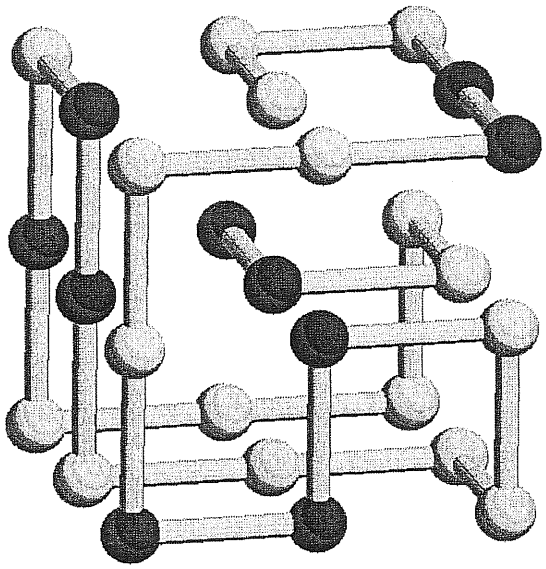


Figure 2.2: An example of protein conformation in the cubic lattice model. The 27 monomers are connected by 26 bonds and occupy a $3 \times 3 \times 3$ cube. The sequence is represented by the differently colored monomers.

The angles formed by two consecutive sticks are not too realistic since the protein backbone trace does not follow a cubic zig-zag. Furthermore the different size of side-chain can be important in the folding of protein sequences [41], while the lattice constant is independent of the type of amino acid. The cubic lattice is quite poor from this point of view. In more sophisticated calculations a higher lattice resolution, without necessarily introducing an off-lattice model, can improve the description of the conformation space. For example, it has been shown that using a Face Centered Cubic lattice, a real structure can differ by less than 2Å from the best-fit conformation [54].

2.2 Which potentials to use?

The interactions of amino acids in proteins are much more complex than the contact potentials shown in the previous section. First of all, real proteins can fold only in their natural environment. This means that the presence of water is of fundamental importance to drive the protein into its native conformation and to stabilize it, whereas, in the model we deal with, the solvent is absent. Then, interactions between amino acids are the result of several different kind of interaction (electrostatic ones, van der Waals forces, hydrogen bonds) and vary in a complex way with the distance. Finally, the interactions between two amino acids do not depend only on the kind of amino acid but even on the environment in which they are embedded: several amino acids can have a hydrophobic or polar behaviour depending on their position on the protein. Why one should use then such a simplified model for interactions?

There are many reasons. First, models with two-body interactions have an important role in statistical physics and, in some cases, it is possible to find analytical results [73]. Second, exact calculations of energetic interactions require quantum mechanics and are not practicable for objects as large as proteins, whose native conformation is stabilized by the interactions with thousands of water molecules. Last, but not least, even if the dependence of the interaction with the distance is specified a priori by the definition of the contact map, the contact potentials B's can be optimized by fitting data obtained by the Hamiltonian (2.2) using a method generally called *knowledge-based potential extraction*. The free parameter in eq. (2.2) allow to compensate, though only partially, the simplifications described above.

The problem to fit/optimize the two-body contact potentials with the unknown real ones is a fundamental subject in using coarse-grained Hamiltonians. Many different groups have used several different methods to accomplish this formidable task. Here we are interested especially in the so-called knowledge-based methods that are relatively simple to use and have the advantage to take into account all the external factors (like the interactions with the solvent, the different frequency of amino acids in proteins, and so

on) [47, 71, 37, 48, 46, 9, 80]. Such methods, instead of facing the difficult task to coarse-grain the quantum interactions between atoms by an effective mean force parameters, aim to find optimal parameters by using in a smart way the already known sequence-structure relationship. Here we will review briefly some of them.

One of the most used (and criticized) methods for extracting contact potentials, is the quasi-chemical approximation [47, 48, 71]. The basic assumption of the quasi-chemical method is that two kinds of amino acids, let's say A and B, that are frequently found in contact in protein structures, likely, will attract each other strongly. By analyzing a set of protein structures, it is possible to estimate the frequency f_{AB} of contacts between A and B—defined as the number of contacts A-B divided by the total number of contacts—and the frequency f_A and f_B of the amino acids A and B, respectively. The contact potential $B(A,B)$ is given by

$$B(A,B) = -T \log \frac{f_{AB}}{f_A f_B} , \tag{2.3}$$

where the denominator in the logarithm represents the expected frequency for independent events. If the measured frequency is larger than the expected frequency the potential will be negative and the associated interaction attractive.

Another knowledge-based method for extracting potentials is the one introduced by Maiorov and Crippen [37, 80]. From PDB it is possible to retrieve data for M sequences together with their associated structure, (σ_1, Γ_1), (σ_2, Γ_2), ... ,(σ_M, Γ_M). By threading (a technique widely used in this field) or by other tools it is possible to generate, for each protein, N alternative conformations not related to the native state. Since the native state is the conformation in which the free energy has a global minimum, we can write the following system of inequalities:

$$\begin{array}{lcl} H(\sigma_1,\Gamma_1) & < & H(\sigma_1,\Gamma_1^{(j)}) \\ H(\sigma_2,\Gamma_2) & < & H(\sigma_2,\Gamma_2^{(j)}) \\ & \cdots & \\ H(\sigma_M,\Gamma_M) & < & H(\sigma_M,\Gamma_M^{(j)}) \end{array} \tag{2.4}$$

where $j = 1, \ldots, N$. This set of NM inequalities can be solved by standard methods of linear programming or with more efficient techniques. The potentials obtained in this way are, at least in principle, more reliable, since they are based on physical principles. However, the goodness of such parameters depends in a crucial way by the method used to generate alternative conformations and, in part, by the parametrization used for the energy function.

Simplified models can be interesting tools even when one does not possess optimal potentials. In this case, one renounces to reproduce some partial results in structure prediction of natural sequences, in protein folding or in protein design —for example, the prediction of key-sites of a specific protein, like folding nucleus or conserved residues—but use them to understand better general features of proteins. For example, one could use lattice models to answer the following general questions: Can the typical sequence-structure relationship of proteins be reproduced in simple models? How does it depend on the potentials used? In this case, the use of a twenty-letter alphabet is not necessary, whereas using a smaller alphabet can help in exploring the sequence space.

The simplest protein model (we are not considering ultra-specific models, like the Go-model [20]) is the HP model [33]. The twenty amino acids are grouped in two classes, Hydrophobic and Polar, and, consequently, B becomes a 2×2 (symmetric) matrix. The Hamiltonian can be rewritten in the following form:

$$H_\sigma(\Gamma) = N_{HH}B_{HH} + N_{HP}B_{HP} + N_{PP}B_{PP} \tag{2.5}$$

where N_{HH}, for example, is the number of hydrophobic contacts in the structure Γ. The standard potentials introduced by Lau and Dill are:

$$B_{HH} = -1 \quad B_{HP} = B_{PH} = B_{PP} = 0 \ . \tag{2.6}$$

Then the energy function is simplified in:

$$H_\sigma(\Gamma) = N_{HH}B_{HH} \ . \tag{2.7}$$

This energy function implement the observation that proteins in nature form a compact hydrophobic core, while polar amino acids are prevalently located

at the surface. In fact, a generic sequence will adopt as maximally stable conformation the one in which the number of H-H contacts is as large as possible.

We observe that the presence of the solvent, even in this simple Hamiltonian, is not neglected: instead, it is taken into account in an effective way, since monomers that maximizes the contact with other monomers, necessarily minimize the exposition to water molecules.

Li et al. in [35] preferred to use a slightly different potentials, i.e.:

$$B_{HH} = -2.3 \quad B_{HP} = B_{PH} = -1.0 \quad B_{PP} = 0 \ . \tag{2.8}$$

They observed, in fact, that these potentials satisfy some physical constraints, not satisfied by the standard potentials:

1. compact shapes have lower energy than non-compact shapes;
2. H monomers are buried as much as possible, which is expressed by the relation $B_{HH} < B_{HP} < B_{PP}$;
3. different types of monomers tend to segregate, which is expressed by $2B_{HP} > B_{HH} + B_{PP}$.

Conditions 2. and 3. were derived from analysis of the real-protein data contained in the Miyazawa-Jernigan matrix of inter-residue contact energies.

Finally, another set of potentials that has been used for testing design algorithms is the so called AB model. The two kinds of monomers are here referred with the letters A and B, since potentials are not aimed to implement hydrophobicity. Potentials for AB model are:

$$B_{AA} = -1 \quad B_{AB} = B_{BA} = 0 \quad B_{BB} = -1 \ . \tag{2.9}$$

In this model, monomers tend to segregate but no hydrophobic core is formed.

2.3 Designing sequences by energy minimization

In this section we will discuss one of the first methods of protein design inspired by statistical mechanics studies. The goal is to give a strategy for

selecting sequences showing protein-like features, i.e. sequences able to fold fast into a unique and specific protein structure. The (target) structure, that is specified before the selection procedure, will be chosen randomly or, when possible, selected so that there exist many sequences folding on it. This requirement, on one hand, guarantees that the set of solutions to protein design is not empty and, on the other hand, makes more stringent the correspondence with real problems of protein design. Indeed, it is strongly believed that real protein structures admit many different sequences folding on them [35]; this provides a wide basin for selecting the sequences most appropriate to their biological function.

In order to test rigorously and efficiently the proposed strategy, simplified models of proteins will be used. For such models, indeed, folding algorithms can be, at least for short proteins, efficiently implemented and, in some cases, an exhaustive enumeration in the conformation space is possible. An algorithm of protein folding reads in input a sequence and returns in output the structure with the lowest energy (native structure or ground state). An algorithm of protein design reads in input a target structure and returns in output a sequence, such that will fold on the given structure. By piping the folding algorithm to the design algorithm, it is possible to verify the goodness of the design algorithm: if the target structure and the native structure are the same, the design algorithm will be considered successful; otherwise will be considered unsuccessful (see fig. 2.6).

The method that will be reviewed in this section is also called energy minimization [69, 67, 70]. It has been proposed by Shakhnovich and it is based is based on previous studies of the energy spectrum of random heteropolymers. Approximated statistical mechanics studies of random heteropolymers have shown that their energy spectrum consists of two parts: the *continuous* part, to which the majority of random conformations belong, and the *discrete* part, representing a few conformations with best-fit contacts. In the continuous part energy levels are highly degenerate with an exponentially large number of conformations associated to each such level. More important, in this part the model is self-averaging, so that its features do not depend on the specific order of the amino acids but rather on the global properties of the sequence

as his composition. By contrast the bottom part of the spectrum is very sequence-sensitive, so that different sequences deliver significantly different energies to their low energy states.

The above discussion suggests a simple way for selecting good folder sequences. The basic idea is that thermodynamic stability can be, on average, a necessary and sufficient condition for fast folding. This is not guaranteed *a priori*, since dynamical properties of statistical systems could depend on the dynamical accessibility in the configuration space too. However, some studies on protein folding suggest that the ability for a protein to fold rapidly depends mostly on the thermodynamic gap [60, 61, 78]. Experimental studies, furthermore, suggest that proteins would be optimized by Nature not for fast folding, but for stabilizing the native conformation [31]. Hence, the problem is selecting sequences with high thermodynamic stability. In fig. 2.3 is shown a possible energy spectrum for a random heteropolymer and a protein-like sequence. Since the continuous part of the spectrum is composition dependent, optimization of the energy with respect to sequences by keeping constant the amino acid composition will lead to a consequential optimization of the thermodynamic gap.

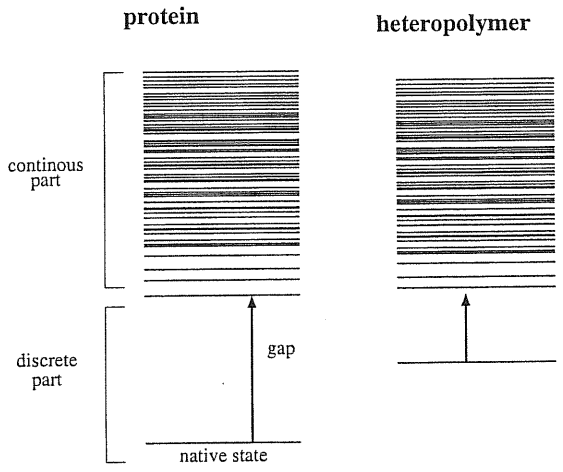


Figure 2.3: Energetic spectrum for a protein sequence: the ground state is a pronounced energetic minimum.

The technical aspect of searching in the sequence space for sequences with

low energy does not create particular difficulties. Shakhnovich and Gutin, for example overcame this problem by optimizing the energy of the sequence by exploring the sequence space with a Monte Carlo method at low temperature. Moves in the sequence space are efficiently proposed by swapping two amino acids at the same time. Such moves generate new sequences with the same composition of the old ones. The probability for accepting the proposed moves is the standard Metropolis probability:

$$P(\sigma \to \sigma') = \begin{cases} \exp\left[-\Delta H_\sigma(\Gamma)/T_{\text{sel}}\right] & \text{if } \Delta H_\sigma(\Gamma) > 0 \\ 1 & \text{otherwise} \end{cases} \tag{2.10}$$

where T_{sel} is a control parameter.

The meaning of the probability (2.10) is the following: if $\Delta H < 0$, then the move will be always accepted—accepted with probability 1—otherwise will be accepted with probability depending on $\Delta H/T$. By the previously described rule for proposing moves and with the Metropolis acceptance probability, one can generate a path, or Markov chain, in the sequence space, i.e.:

$$\sigma^{(0)} \to \sigma^{(1)} \to \sigma^{(2)} \to \cdots \to \sigma^{(n)} \ . \tag{2.11}$$

Sequences present in this chain, have, on average low energy and, hence, a high thermodynamic gap. (This is not necessarily true for the initial sequence, $\sigma^{(0)}$, that, usually, is generated randomly, and for sequences close to it along the Markov chain, since they are still correlated to it.) For a high value of T_{sel} the algorithm is not enough selective and sequences with a large value of $H_\sigma(\Gamma)$ can be visited as well. At an intermediate value of T_{sel} the algorithm visits low energy sequences and, at the same time, fluctuations in the energy allow to overcome small energetic barriers. Finally at very low T_{sel} the algorithm is not able to overcame global and local energetic barriers and low energy sequences becomes hard to be retrieved. In this case one can use more sophisticated selection algorithms like simulated annealing or simulated tempering.

The method described above has been applied by Shakhnovich and coworkers to many different models—always in the context of two-body Hamiltonian but for different interaction parameters and chain length. For example he

tried to design 27-mer self-avoiding walks with AB interactions and 80-mer with HP and MJ interactions, described in the previous section. In the first case the relatively shortness of the chain and the further constraints to consider only maximally compact conformation allows an exhaustive enumeration to assess the performance of the design strategy.

In the second case an exhaustive enumeration is out of reach and a folding algorithm is used to verify the ability of the designed sequences to find the target structure. The success of the design procedure is only partial since it works only for the MJ interactions, while designed sequences in the framework of the HP model are unable to find the target conformation. Shakhnovich suggests that for some protein models, viable sequences folding on a given structure could not exist. Miyazawa and Jernigan potentials should describe real-world interactions and should be used as a prototype for protein folding simulations. His statement is that the number of amino acids classes for reproducing a sequence-structure relationship isomorphous to the one occurring in proteins has to be larger than 5 or 6.

He takes as a support to his observation, analytical calculations for random heteropolymers. From this calculation the critical energy E_c which separates the continuous part of the spectrum from the discrete one is given by:

$$E_c = E_0 - JN\sqrt{2\ln\gamma} \tag{2.12}$$

where E_0 is the average of the distribution of the energy levels, J its standard deviations and γ the number of conformations per monomer—about 4 for the cubic lattice. For a particular sequence composition Shakhnovich estimated both E_0 and J by using 1000 random energy states. Furthermore he calculated the energy E_N for a generic low energy designed sequence and compared with E_c. He found for the HP model $E_N > E_c$ while $E_N < E_c$ for the MJ parameters and concluded that "the HP model is not specific enough to have unique [native] structures".

In the next chapter we will use procedures for protein design with a reduced number of amino acid classes as well as with twenty classes. In fact, it is likely that just two classes (Hydrophobic and Polar) are enough

to reproduce the correct sequence-structure relationship present in the real world. As we have seen in the previous chapter, protein design experiments by binary patterning amino acid, produced good results [30, 83]. Such results show essentially two things:

- a two-letter alphabet allows for the unique encoding of structures.
- the hydrophobic force is the main force driving protein to the native state.

For these reasons understanding the sequence-structure relationship for a reduced number of classes seems an interesting challenge both from a conceptual and technological point of view.

2.4 A rigorous approach to protein design

In the energy minimization method explained above there are a lot of unsatisfactory aspects. First, one has to choose *a priori* the sequence composition. Consequently, sequence selection occurs by exploring an infinitesimally small subspace formed by all the sequences with such composition. Second, it is not generally applicable: in fact for models with a reduced number of classes it gives wrong answers [86]. Third, it is based on the empirical consideration that for heteropolymers with the same composition the energy gap should depend only on the discrete part of the spectrum. Even if the method is very simple and does not consume CPU time, a direct improvement is not feasible: the solution obtained by minimizing the energy can fold or not fold in the target structure but it is not possible to decide which of these are the best solutions.

The correct method was envisaged independently by Deutsch and Kuroski [14] and Seno *et al.* [65], who replaced the energy scoring function with a more complex scoring function. Here, we will follow the approach given in [65], that follows directly from elementary statistical mechanics considerations.

The probability to find a sequence σ on a particular conformation Γ^* and

at a given temperature T, is given by the Boltzmann probability

$$P_\sigma(\Gamma^*) = \frac{\exp(-H_\sigma(\Gamma^*)/T)}{\sum_\Gamma \exp(-H_\sigma(\Gamma)/T)} \ . \tag{2.13}$$

The conformational free energy for a sequence σ at the temperature T is given by $F_\sigma = -T \ln \sum_\Gamma \exp(-H_\sigma(\Gamma)/T)$. Then, the probability P_σ can be rewritten as:

$$P_\sigma(\Gamma^*) = \exp\left[-\left(H_\sigma(\Gamma^*) - F_\sigma\right)/T\right] \ . \tag{2.14}$$

At high temperature $P_\sigma(\Gamma)$ will be almost 0 for all the sequences and for all the structures. However, at lower value of T some sequences show a preference for lower energy structures. If the temperature is still decreased the probability to find the sequence in its ground state will be almost 1 and it will be 0 for all the other structures. Since the F_σ is independent of Γ^* the conformation with the largest probability is the ground state. However, in protein design problems the Boltzmann probability is not directly correlated with $H_\sigma(\Gamma^*)$. In particular, it is clear that sequences with the highest probability to fold on a given conformation are those which minimize $H_\sigma - F_\sigma$.

2.5 The first order cumulant approximation

Deutsch and Kuroski apply this strategy to the 27-mer lattice model with interaction of type AB and HP. In order to evaluate F_σ they introduce two important simplifications. The first one is to consider just the maximally compact conformations. The justification is that globular proteins have folds with a high degree of compactness. In the framework of protein lattice models this is realized by adding to the Hamiltonian a negative interaction to all the contacts, independently on the type of amino acids that are interacting. This excludes the possibility that for some sequences the native state is a non-maximally compact conformation.

The second simplification they apply is the high-temperature expansion on the free energy. This approximation is not completely justified. The temperature at which native states are highly populated has to be small compared with the average energy separation between low energy conformations.

The first order expansion gives

$$F_\sigma \approx \langle H_\sigma \rangle + \text{const.} \tag{2.15}$$

so that the approximate scoring function to minimize for the contact interaction Hamiltonian is:

$$K_\sigma(\Gamma) = \sum_{i<j} [\Delta_{ij} - \langle \Delta_{ij} \rangle] B(s_i, s_j) \tag{2.16}$$

This scoring function is very similar to the energy function of eq. (2.2). The geometrical properties of the conformations are contained in an "effective" contact map, $\Delta - \langle \Delta \rangle$. Since all the entries in $\langle \Delta \rangle$ range in the interval $[0, 1]$, $\Delta_{ij} - \langle \Delta_{ij} \rangle$ is positive (negative) for a native (non-native) contact. The average $\langle \cdots \rangle$ has to be done over all the maximally compact conformations for which an exact enumeration is feasible. Depending on the self-avoidance of the walks not all the contacts have the same probability. For some short chains (27-mers) one can assume a mean-field approximation where the probability for a contact is equal for all the contacts allowed by geometrical properties of the lattice. In this case

$$\langle \Delta_{ij} \rangle = \frac{N_{\text{contacts}}}{N_{\text{possible contacts}}} \tag{2.17}$$

for all the contacts i, j allowed by the geometric topology of the cubic lattice. For longer chains, a dependence of Δ on $|i - j|$ should also be considered. A discussion about such problem can be found in chapter 4 and appendix B, whereas in fig. 2.4 we show a graphical representation of an average contact map, where the $|i - j|$ dependence has been taken into account

It is possible to show that the energy minimization approach to protein design is a consequence of the score function of eq. (2.16). Mutations in the sequence during the Monte Carlo procedure has effect on the average contact map only for sequences with different composition. It is easy to observe that, if $\langle \Delta_{ij} \rangle = \Delta_0$ is independent on i and j, the second term in eq. (2.16), i.e. $\Delta_0 \sum_{ij} B(\sigma_i, \sigma_j)$ is a function of the composition of the sequence.

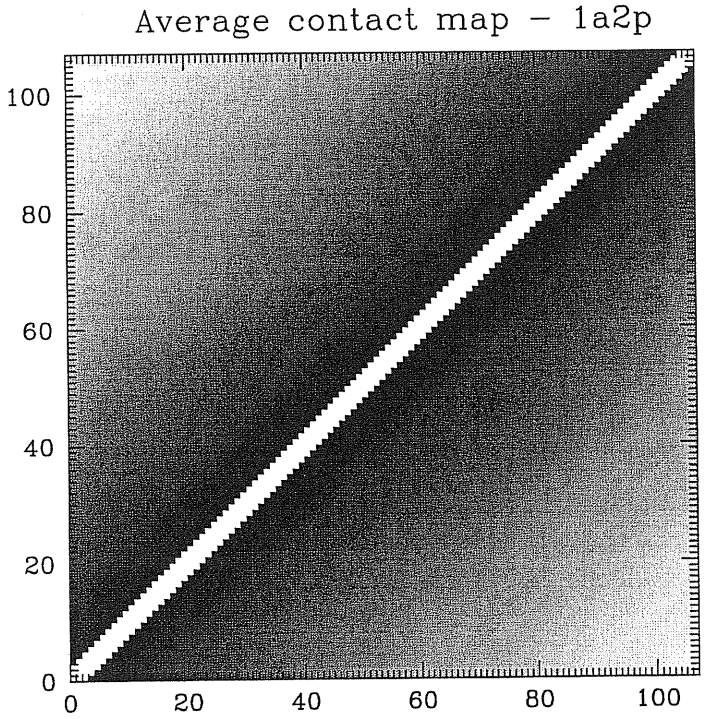


Figure 2.4: A visual representation of an average contact map evaluated according to criteria explained in appendix B. Dark (light) colors correspond to large (small) values of average contacts. To be compared with the map in fig. 2.1.

2.6 A Monte Carlo estimate for the conformational free energy

The first-order high-temperature expansion in eq. (2.15) is the first attempt to take into account contributions coming from the free-energy, i.e. from the denominator of eq. (2.13). Approximations of the free energy through a cumulant expansion like in [50], do not lead to a significant improvement with respect to the first order cumulant expansion (Seno et al., unpublished results).

A substantial improvement with respect to mean field approximation, has been introduced in reference [65]. There, it is recognized that high-energy

expansion of the free energy is not a good approximation for lattice models of protein and, instead, it is applied a standard method from polymer theory to estimate it. The method is based on a stochastic generation of self-avoiding walk to sample the configuration space in a smart way.

First one recognizes that the partition function, that is related to the conformational free energy by $Z_\sigma = \exp(-F_\sigma/T)$, can be evaluated as a thermal average, i.e.

$$Z_\sigma = \frac{C_{\text{tot}}}{\langle e^{H_\sigma/T} \rangle} \,. \tag{2.18}$$

Here C_{tot} is the total number of conformations and $\langle \cdots \rangle$ denotes the canonical-ensemble average, i.e.

$$\langle e^{H_\sigma/T} \rangle = \frac{\sum_\Gamma e^{-H_\sigma/T} e^{H_\sigma/T}}{\sum_\Gamma e^{-H_\sigma/T}} \tag{2.19}$$

Then one uses an importance sampling Monte Carlo growth scheme to estimate the average in eq. (2.18). Conformations are generated in a step by step manner and are completely decorrelated. At each step a direction is selected in a stochastic way and a new bond is added to the walk in such direction. Directions carrying the smallest amount of energy are selected with the highest probability, while directions that don't preserve self-avoidance are never selected. This guarantees self-avoidance and compactness, speeding up determination of the average. However, since walks are generated with a growing mechanism, the distribution is different from the desired Boltzmann one and a re-weighting mechanism is necessary for a correct estimate. Let p_i be the probability for the selected displacement at the i-th step, the probability that a generic conformation Γ is sampled is $p(\Gamma) = \prod_i p_i$. Finally, we get the following estimate of the denominator of eq. 2.18:

$$\langle e^{H_\sigma/T} \rangle \approx \frac{\sum_\Gamma (e^{-H_\sigma/T}/p(\Gamma)) e^{H_\sigma/T}}{\sum_\Gamma (e^{-H_\sigma/T}/p(\Gamma))} \tag{2.20}$$

where the sum is done over the sampled structures.

This method to estimate the partition function has been applied in the context of HP lattice model to the 48-mer Harvard structures. Such structures are some 48-mer maximally compact conformations that have been

used to test folding design algorithms. Shakhnovich, representing the Harvard group, proposed to design some HP sequences that,according to his method described above, would fold to pre-selected three-dimensional structures of his choice (the Harvard structures). Then he gave the sequences to the UCSF group, coordinated by Dill, and the UCSF group attempted to fold the sequences, i.e. to find their lowest energy structures. If structures with energy equal to or lower than the energy of the target structure were found, then the design algorithm would have been unsuccessful. The blind test (the UCSF group didn't know the target structures selected by Shakhnovich) is sketched in fig. 2.6.

For the folding process the UCSF group used the CHCC algorithm (Constraint-based Hydrophobic Core Construction), based on a systematic assembly process using discrete geometry. For the design algorithm the Harvard group used the energy minimization (see section (2.3)) with fixed composition using a Monte Carlo procedure to explore low energy sequences. As a result of the test, in 9 cases over 10 the energy for the low energy conformations obtained by CHCC were systematically much lower than the energy on the relative target structures. Only in one case the two energy were equal. While both the groups conclude that the energy minimization strategy does not work for the HP model, really interesting are the different explanation that they give about this failure. While the Harvard group point out the poorness of the HP model and states that for a twenty class model energy minimization should work, the UCSF group believes that there may be a problem in the Harvard procedure. In particular, they insist on the problem of negative design: a method based on a mean-field theory of heteropolymers cannot have enough information for designing out bad conformations.

The statement that the HP model is not enough protein-like, we think, it is not well supported. Since the hydrophobic force has been recognized as the main force stabilizing protein in their native state, a two class model appear realistic enough to capture the main essence of the sequence-structure relationship. Furthermore, attempts to design by patterning amino acids in two classes has been successful more than once. In principle, the structures used by Shakhnovich could be not encodable for the HP model. Since the

number of sequences N_{seq} is 2^{48}, while the number of structures N_{str} is a number of the order μ^{48} with $\mu \approx 4$, the average designability [35] N_{seq}/N_{str} is unfavorable, while it should be favorable for a twenty class model. This estimate for the average encodability is not realistic because we know that too open conformations have not to be considered, since they are not low energy conformations.

In fact, an answer to the Harvard-San Francisco problem has been found later by Micheletti et al. [44]. Though the sequences proposed had not a unique ground state, the CHCC algorithm couldn't found structures for all ten proposed structures with an energy lower than the energy on the target ones. The method used by Micheletti et al. [44] was a combination of the energy minimization and the importance sampling Monte Carlo in conformational space as described above. Energy minimization has been used to filter 75 viable sequences with different composition, since estimate of free energy with importance sampling is quite expensive. For each sequence 100000 unrelated conformations were generated in order to estimate the score $H_\sigma - F_\sigma$ (see eq. 2.14). Finally, the sequence with the lowest score was folded by the CHCC algorithm.

Nevertheless, sequences selected by Micheletti et al. [44], are high degenerate, since CHCC estimates from 1000 to 10^6 or more ground state for each of them. An exact solution for protein design should not be degenerate at all. Actually, we don't know if a solution for the Harvard structure exists or not, and hence how reliable the design method is. For sure, the ten structures chosen randomly by E. Shakhnovich, are not the most designable, and hence are not protein-like. In fact, in nature, protein structures are the most encodable, since sequence degeneracy, i.e. the fact that many sequences can fold on the same structure, it has been suggested it is important for evolutionary reasons.

2.7 Conclusions

The success of the method used by Micheletti et al. [44] shows that energy minimization and minimization of energy gap are not enough selective for

protein design, even for simplified protein models. A more refined method for estimating the free energy has to be used. However, generating low energy conformations for a given sequence can be computationally very expensive and, for low value of T, not too accurate. Micheletti et al. overcome this problem isolating a reduced set of sequence by energy-minimization, and applying to their method on these filtered sequences.

In my first studies on protein design, we tried to apply *ab initio* the Rosenbluth Monte Carlo for estimating the conformational free energy, i.e. by not limiting in any way the search in sequence space. Since neither the conformation space nor the sequence space, at least for long chains can be exhaustively enumerated, the basic idea is to nest a stochastic procedure in conformation space in a stochastic procedure in sequence space. This idea, that has been applied by adopting a different formalism on short lattice proteins by Peterson and coworkers [26, 27], is not doable for longer chains. If one uses the Rosenbluth method for sampling the conformation space and a standard Monte Carlo procedure in the sequence space, the computational cost is extremely large. Furthermore, the approximation for the free energy that is allowed for computational reasons, is not enough accurate and this selection procedure, though computational expensive, does not improve previous results.

For these reasons, we renounced to follow the idea of using a nested Monte Carlo and we focused on a different strategy. Instead of generating low energy conformation for every sequence visited by our stochastic procedure in sequence space, we will build up a small but accurate set of conformations and the conformational free energy will be evaluated by using only this set. In next chapter we will give a procedure for selecting such conformations and we will show that the approximation is very good even by using a very small set.

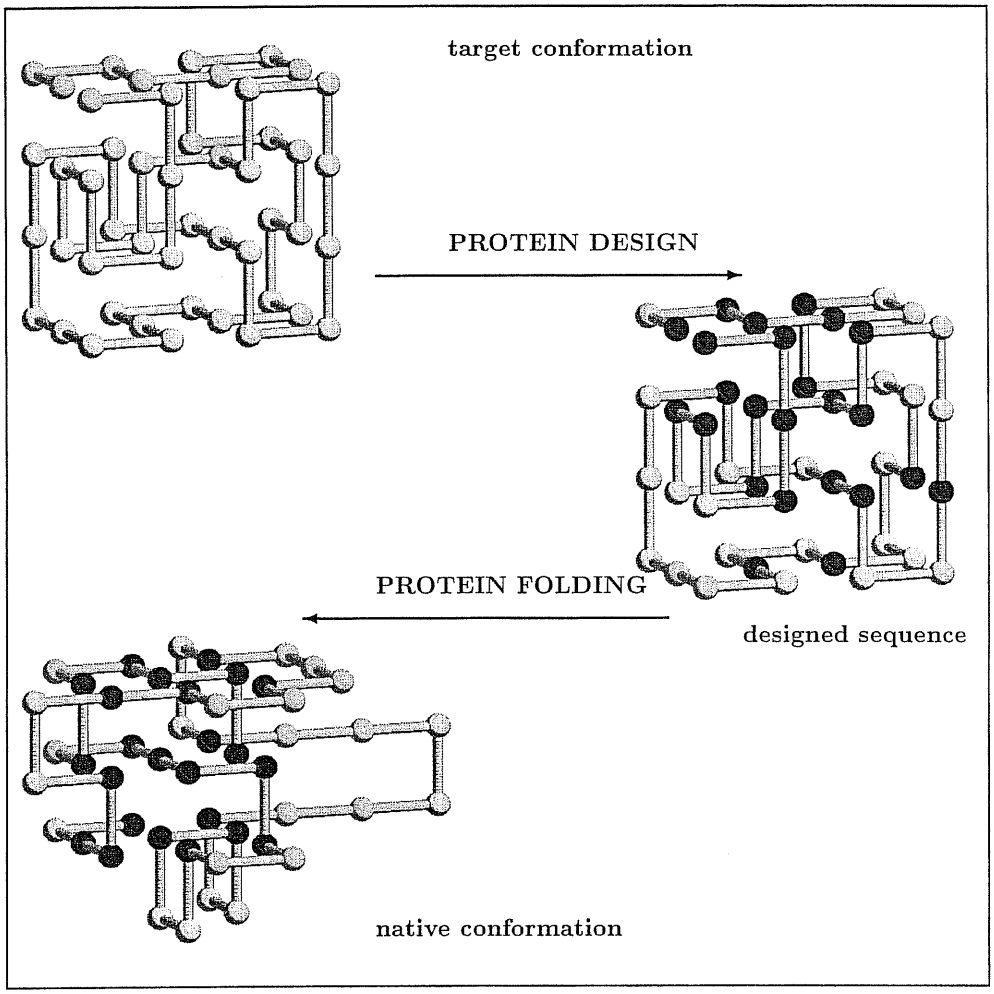


Figure 2.5: An example of blind test in protein design. The sequence selected by Shakhnovich, aimed to stabilize the target structure, is "folded" by the UCSF group, finding a lower energy conformation. Ideally, the target structure would have to be recovered by the UCSF group.

Chapter 3

An iterative method for protein design

In experimental studies on protein design, sequences optimized to be stable on a particular conformation, unfold in an ensemble of unwanted conformations or decoys. The problem to decrease the fitness of designed sequences with respect to such decoys is a main theme in experimental protein design.

In this chapter we will introduce and test on lattice models some strategies for recognizing such decoys and taking into account them in protein design. Such strategies are based on a stochastic minimization of the scoring function $H_\sigma - F_\sigma$ in the sequence space, where the conformational free energy is estimated by using only an optimized subset of decoy conformations.

The first strategy we will present is an iterative procedure for selecting conformations highly-competitive with the target structure and reducing at the minimum the decoy set. In a spirit of trial and error, design attempts are followed by validation tests through a folding procedure and failures in designing sequences are exploited by enriching the decoy set by new decoys.

The iterative method is extremely efficient in terms of number of generated decoys. However, it is based on a folding procedure that can be used, up to now, only on simple models like lattice proteins. Implementing such method to design real protein structures, would require an enormous computational effort.

In the attempt to overcome this problem, we introduce some geometric methods, which are based on criteria of similarity of decoy conformations with the target one. Geometrical criteria, though are not performing like the iterative method, are simple to be implemented and in many cases might be an efficient tool for designing real structures.

3.1 Formulation of the design strategy

The design procedure, as described in several works [44, 14, 65, 45] and in the previous chapter, consists in maximizing the probability that a sequence s is found in a target conformation Γ_0 at a preassigned temperature T ($k_B = 1$)

$$P_\sigma(\Gamma_0) = \exp\left\{-\left(H_\sigma(\Gamma_0) - F_\sigma\right)/T\right\} , \tag{3.1}$$

where $H_\sigma(\Gamma_0)$ is the energy of s in Γ_0 and the conformational free energy F_σ is defined by:

$$\exp(-F_\sigma/T) = \sum_{\{\Gamma\}} \exp\left(-\beta H_\sigma(\Gamma)\right) . \tag{3.2}$$

Here $\{\Gamma\}$ represents the ensemble of all the possible conformations. If Γ_0 is the native state of a sequence s then there must exist a temperature T_F below which

$$P_\sigma(\Gamma_0) > \frac{1}{2} , \tag{3.3}$$

i.e. Γ_0 is the ground state of s and below T_F there is a macroscopic probability to occupy it. In terms of eq. (3.1), eq. (3.3) can be rewritten

$$K_\sigma(\Gamma_0) \equiv H_\sigma(\Gamma_0) - F_\sigma < T \ln 2 . \tag{3.4}$$

All sequences satisfying inequality (3.4) are solutions of the design problem on the conformation Γ_0 and have folding temperature greater than T. The main problem of inequality (3.4) is that in order to evaluate F using eq. (3.2), one would need, in principle, all possible alternative conformations that can house the sequence σ.

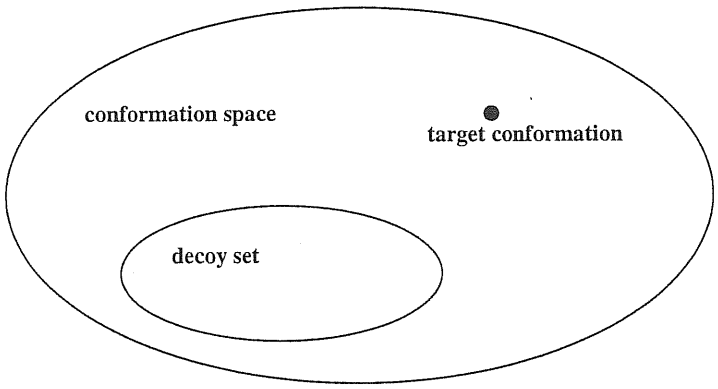


Figure 3.1: The approximate conformational free energy is evaluated in this chapter by using a small subset of the conformation space: this subset is represented by the decoy conformations and the target one.

The result of the following analysis is that F can be well approximated restricting the sum in eq. (3.2) to a manageable set, $\mathcal{D}$, of conformation Γ. In other words, if we define the approximated free energy $\mathcal{F}$ by

$$\exp(-\mathcal{F}_\sigma/T) = \exp(-H_\sigma(\Gamma_0)/T) + \sum_{\Gamma\in\mathcal{D}} \exp(-H_\sigma(\Gamma)/T) \qquad (3.5)$$

we will show that in the design procedure we can replace the exact free energy F with $\mathcal{F}$, without affecting too much the final results. Sequences satisfying eq. (3.4), will satisfy the following inequality too:

$$\mathcal{K}_\sigma(\Gamma_0) \equiv H_\sigma(\Gamma_0) - \mathcal{F}_\sigma < T \ln 2 \; . \qquad (3.6)$$

In fig. 3.1 is shown in a schematic way the decoy set as subset of the conformation space.

In order to test in a rigorous way the validity of our design methods, we have made calculations on a lattice model, which has been quite commonly used in the literature [8, 70, 35, 14, 44]. The model we will treat has an energy functional

$$H_\sigma(\Gamma) = \sum_{i<j-1} \Delta_{ij}(\Gamma) B(\sigma_i, \sigma_j) \qquad (3.7)$$

where Γ represents a self avoiding chain with nodes $i = 1, \ldots, N$ in a simple cubic lattice, $\Delta_{ij}(\Gamma) = 1$ if i and j are non-consecutive nearest neighbors

nodes, i.e. they form a contact, and zero otherwise and $B(\sigma, \sigma')$ is the energy associate to a contact between amino acid type s and s'. Furthermore, if we restrict to $N = 27$, the set $\mathcal{C}$ of all the maximally compact conformations, i.e. conformations filling a $3 \times 3 \times 3$ cube, can be exhaustively enumerated [35].

When all the B's are negative enough and maximally compact target native states are considered, then the search for the most competitive decoys can be restricted to $\mathcal{C}$. We will consider only two classes of amino acids, hydrophobic, H, and polar, P, since the hydrophobic forces are the main forces driving the folding of proteins ([33]). There are many reasons we choose to use two classes. First, the number of possible sequences that is possible to mount on a 27-mer structure is easily enumerable and this allows for stringent tests. Second, the HP model is the hardest benchmark for protein design, as already recognized in the blind test on the energy minimization design algorithm ([86]). Finally, theoretical studies on two-letter code models of proteins are justified by the repeated successes in experimental attempts to design by binary patterning amino acids ([30, 83]).

3.2 The iterative strategy

Selecting a set with the most competitive conformations is a combinatorial problem. If you have N possible conformations and you want select the M most competitive, you have to choose among $\binom{N}{M}$ possible sets. In our case $N = 103346$ and M will range between 100 and 1000. It is clear that an exhaustive enumeration of all the possible sets is not feasible. Furthermore, we don't have a simple criterion to establish which set is better in recognizing good sequences (i.e. sequences satisfying eq. (3.4)). For every set we should perform several design test, verifying if the designed sequences satisfy eq. (3.4).

The iterative strategy overcome this hindrance by updating the set $\mathcal{D}$ iteratively. The basic idea is to implement several design attempts followed by a validation procedure, i.e. a folding procedure, to check the proposed solutions. After each attempt, sequences recognized as wrong solutions are

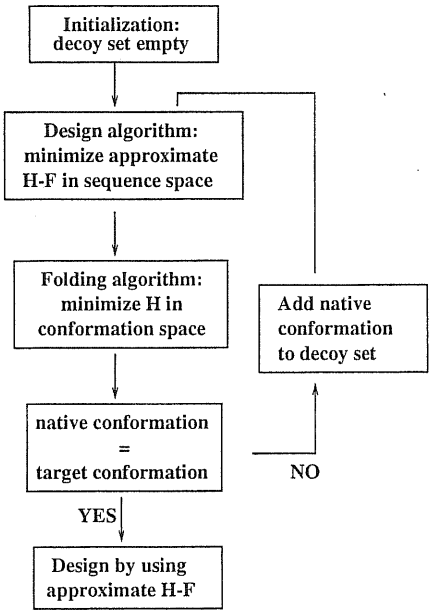


Figure 3.2: Schematic representation of the iterative procedure.

discarded and, most importantly, conformation with energy less than the target structure are used to improve $\mathcal{D}$. An optimal decoy set could in principle not exist, i.e. could be very as large as the whole conformation space, and the typical size of this set will depend on the protein model used. However, we will show that a similar procedure will converge very rapidly to the optimal decoy set, for three different lattice model.

The iterative procedure is schematically represented in fig. 3.2 and can be described as follows.

1. At the beginning $\mathcal{D}$ contains the target structure and other optional conformations. These conformations can be selected randomly or with a heuristic procedure. For exmple, in our tests on lattice models we have selected no optional conformation. The free energy for a given sequence is evaluated by using this primitive decoy set.

2. A simulated annealing in the sequence space is implemented to find at least a sequence s^* satisfying eq. (3.6).

3. A folding procedure is used to validate our putative solution s^*. Such

procedure could be a real dynamics in the conformation space, a stochastic algorithm, or an experimental structure determination. In our studies, we have used both an exhaustive search and a deterministic algorithm like CHCC (see below) to find the ground state of s^*.

4. If the retrieved conformation, let's say Γ^*, is just the target structure Γ_0, then s^* is a real solution to protein design (supposing the folding procedure is enough reliable). Otherwise s^* is not a solution of eq. (3.4) meaning that the decoy set has to be improved.

5. The improvement is obtained by adding the conformation Γ^* to $\mathcal{D}$.

6. The new decoy set allow a more efficient sequence selection procedure. The simulated annealing will be applied as in point 2., but now the estimate of $\mathcal{K}$ will be more accurate.

3.3 Implementation and test of the iterative procedure

Step 3. of the iterative procedure was carried out in two distinct ways. In a first attempt we found the true lowest energy state of s by exhaustive search. In a second attempt we tried to mimic the difficulty of finding the ground state in a realistic context and hence carried out a random partial exploration of the structure space. Although the first method was expected to be more efficient than the second one, their performance turned out to be almost identical, as we discuss below.

The four target conformations used to test the procedure are given in Table 3.1.

We used three possible choices for the B's. First, we adopted the standard 2-class HP model with $B_{\mathrm{HH}} = -1 - \alpha$ and $B_{\mathrm{HP}} = B_{\mathrm{PP}} = -\alpha$. α is a suitable constant ensuring that native conformations are compact. Since all conformations considered here have the same number of contacts the value of α is irrelevant and will be omitted from now on. This model reduces to

	relative structures	Enc.
Γ_1	URDDLLFFRBRFULLBBUFFRRBDLU	25
Γ_2	UURFLFDBRBDFLFRRBBUUFFLDRB	337
Γ_3	UURRFDLULDDFUURRDDBBULDFFU	1224
Γ_4	UURRDLFFRBULLDDRBRFFLLUURR	1303

Table 3.1: The four structures used for benchmarking the design strategy. The conformations are encoded in bond directions: U, up; D, down; L, left; R, right; F, forward; B, backward. The encodability in the rightmost column is defined as the number of sequences admitting the corresponding structure as their unique native state (HP interactions are assumed).

	1	2	3	4	5	6
1	-50.00	-20.49	-38.20	-6.65	-43.65	-10.63
2	-20.49	-14.91	-18.13	-4.00	-15.56	-3.81
3	-38.20	-18.13	-35.75	-5.07	-23.96	-26.02
4	-6.65	-4.00	-5.07	-1.65	-5.17	-9.47
5	-43.65	-15.56	-23.96	-5.17	-43.71	-18.63
6	-10.63	-3.81	-26.02	-9.47	-18.63	-26.70

Table 3.2: Energy parameters for the 6-class model. Parameters obey the segregation principle [35].

the standard HP model described in chapter 2 (see eq. (2.6)). The second case is a 6-class model and the B's are shown in Table 3.3.

For the last case we considered the full repertoire of 20 amino acids used the Miyazawa and Jernigan energy parameters given in Table 3 of ref. [48]. With the standard HP parameters, structures $\Gamma_1 - \Gamma_4$ have various degree of designability. The latter is defined as the number of sequences admitting them as unique ground states [35] . Hence, the encodability of Γ_1 and Γ_2 is poor and average respectively, while Γ_3 and Γ_4 have very large encodability. It was shown that the degree of encodability is mainly a geometrical property of the structure and not too sensitive to the number of amino acid classes or the values of interaction parameters [35, 41, 42]. For this reason we expect

that the relative encodability of $\Gamma_1 - \Gamma_4$ remain different when using all the three sets of parameters.

3.4 Results and discussion for the iterative method

The "dynamical" performance of the iterative algorithm can be seen in Figs. 3.3a-c. The plots show the number of solutions retrieved as a function of the number of iterations at a "physiological" temperature equal to 0.1, 10.0 and 0.7 for 2, 6 and 20 classes of amino acids, respectively. The different values of the physiological temperature are related to the different energy scales of the interactions.

It can be seen that, after an initial transient, the performance of the method (given by the slope of the curves) is very high. In particular, for a large number of classes, it is nearly equal to 1 for all structures. Table 3.4 provides a quantitative summary of the performance of the method. For the HP model, first column of Table 3.4, the method was iterated until it could not find further solutions with (estimated) folding temperature greater that 0.1 . For the cases of 6 and 20 classes, a very large number of solutions exist. Hence, we stopped the procedure after 1000 or 500 iterations, depending on the number of classes.

An appealing feature is that the extracted solutions show no bias for sequence composition (see Fig. 3.4) or ground-state energy. This can be seen in Fig. 3.5, where we have plotted the energies of 1000 designed solutions of fixed composition for the 6-classes case. Solutions do not exhibit packing around the minimum energy (≈ -830) and their energy spread is fairly wide (the estimated maximum energy is ≈ -170). Furthermore, for each extracted sequence we also calculated its folding temperature, to compare it with T. As we remarked, if all the significant competitors of Γ_0 were included in $\mathcal{D}$, then sequences satisfying eq. (3.4) should have folding temperatures greater than T. As shown in the typical plot of Fig. 3.6 this is almost always the case, ensuring that solutions can be extracted with a desired thermal stability.

	HP		6 classes		20 classes	
	N_{it}	N_{sol}	N_{it}	N_{sol}	N_{it}	N_{sol}
Γ_1	62	8	1000	895	500	388
Γ_2	722	337	1000	891	500	419
Γ_3	1898	1219	1000	906	500	423
Γ_4	1719	1297	1000	911	500	457

Table 3.3: Number of extracted solutions, N_{sol}, after N_{it} iterations of the design procedure. For the HP model N_{it} is the number of iterations at which the iterative scheme stopped. It was verified that the 1297 extracted solutions for structure Γ_4 have a folding temperature between 0.15 and 0.6.

An alternative measure of the thermal stability connected to the cooperativity and rapidity of the folding process is the Z_{score}. For a sequence, s, designing structure Γ, the Z_{score} is defined as [3]:

$$Z_{score} = \frac{\langle H_s \rangle - H_s(\Gamma)}{\sigma_s}, \tag{3.8}$$

where $\langle H_s \rangle$ is the average energy over the maximally compact conformations and σ_s the standard deviation of the energy in this ensemble. Fig. 3.7 shows a scatter plot of extracted solutions for target structure Γ_1 for the 20-letter case. It can be seen that there exist solutions with very high Z_{score} throughout the displayed energy range. This proves the usefulness of the novel design technique which has no bias in native-state energy. In fact, it allows to collect equally good folders with a wide range of native-state energy (and hence very different sequences). This ought to be useful in realistic design contexts, where among all putative design solutions one may wish to retain those with specific amino acids in key protein sites. The ability to select sequences across the whole energy range highlights the efficiency of the technique. In fact, as shown in Fig. 3.8, away from the lowest energy edge, the fraction of good sequences over the total ones with the same energy is minuscule (note the logarithmic scale). Our method is able to span across the whole energy range without restricting to those of minimal energy, which are a negligible fraction of the total solutions.

Finally, we analyzed the degree of mutual similarity between extracted solutions. For the 6-class model, the sequence similarity between solutions was rather low, being around 20%, as can be seen in Fig. 3.9. This rules out the possibility that solutions correspond to few point mutations of a single prototype sequence.

One of the most significant features of the novel design procedure is that the number of decoys, used to calculate the approximate free energy, (3.5), can be kept to a negligible fraction of the total structures and yet allow a very efficient design. This is proved even more strikingly by a further test of our design strategy in the whole space of both compact and non-compact conformations. We carried out a design of structure Γ_2 by using the HP parameters with the constant α set to 0. This amounts to allow for non-compact conformations to be native states. Since it is unfeasible to explore this enlarged structure space, folding was carried out with a stochastic Monte Carlo process, as described in refs. [65, 44], which generated dynamically growing low-energy conformations at a suitable fictitious Monte Carlo temperature. The correctness of the putative solutions was carried out by using an algorithm known as Constrained Hydrophobic Core Construction (CHCC)[85, 84]. The algorithm relies on an efficient pruning of the complete search tree in finding possible low energy conformations for a sequence. At the heart of the algorithm is the observation that the most energetically convenient conformations for the hydrophobic monomers is to form a compact, cubic-like, core. This ideal situation may not be reachable for arbitrary sequences, due to frustration effects; these are taken systematically into account to build a compact core with a number of cavities sufficient to expose P singlets (i.e. a P flanked by two H monomers in the sequence) on the surface, which is energetically more effective than burying them in the core. Then, exhaustive search algorithms are used to check the compatibility of a sequence with cores of increasing surface area (i.e. decreasing energy). A detailed description of the method can be found in Refs. [85, 84]. The time required by CHCC to find the ground state energy of a sequence increases significantly, on average, with the increase of the number of H residues. For this reason we limited the search for design solutions to sequences with $n_H = 13$. The

solutions, obtained in about one hundred iterations, appears in table (3.4). All the 23 extracted solutions had Γ_2 as the unique ground state among the compact structures, and 17 of them retained Γ_2 as ground state even when non-compact structures are considered. Given the vastity of the enlarged structure space this represent a remarkable result.

3.5 A first geometrical criterion: maximum overlap

The iterative method is a powerful tool to build up a set of decoy conformation, when amino acids interactions are well specified, like in the case of model proteins, and when an approximate but efficient folding algorithm can be implemented. In this section and in the next one, we will introduce other design methods that don't require a folding procedure to recognize decoy conformations. Such methods are instead based on a criterion of geometrical similarity between decoys and the target conformation.

In order to search for a decoy set determined by only geometrical information of the target conformation, we will study geometrical properties of the iterative decoy set. We define the overlap between two conformations Γ and Γ' as:

$$O(\Gamma,\Gamma') = \sum_{i<j-1} \Delta_{ij}(\Gamma)\Delta_{ij}(\Gamma') \tag{3.9}$$

and it will be used as a parameter to quantify the degree of similarity of two conformations. In the case we are dealing with (i.e. only compact conformations of 27 beads) $O(\Gamma,\Gamma') \leq 28$. Fig. 3.5 shows the distribution of the overlap of the conformation Γ_2 with the whole set of decoys obtained with the iterative method and with an equal number of random conformations. Similar distributions are obtained also for the other target conformations. Notice that the tail at high overlap in the latter case (random set) is much lower than in the former one. This fact suggests that similarity in terms of overlap with the target could be a good geometrical criterion for selecting decoys.

Correct solutions
010111001110110001010100001
000011011100111101000100101
000010011100111101000101101
000010000111100101110101101
000010010100101111000110111
000010000110100111010110111
000010110100100111000101111
000010000110100111010110111
000010110100100111000101111
000010010100101111000101111
000010000110100111010101111
000010110100100101000111111
000010010100101101000111111
000010100100100111000111111
000010010100100111000111111
000010000110100101010111111
000010000100100111010111111
Incorrect solutions
110010001110110001010101001
010011001110110100010100101
110010001100110101000101101
100011001100110101000101101
000010100100100111010101111

Table 3.4: Extracted solutions for structure Γ_2. The design attempt was carried out in the whole space of conformations with arbitrary degree of compactness.

interactions: $\vec{B} = -1, 0, 0$				
	A	**R**	**B**	**C**
Γ_1	86	3	12	3
Γ_2	85	1	21	16
Γ_3	57	1	13	3
Γ_4	87	1	35	11
average	**79**	**1**	**20**	**8**

Table 3.5: Number of distinct sequences correctly designed over a bunch of 100 sequence predicted with our design procedure and using a set of 250 decoys. The decoy sets are determined using different criteria: **A**=iterative method; **R**=conformations chosen randomly; **B**=method of maximum overlap; **C**=method of minimum burial distance. In order to implement a rigorous test we used four highly encodable target conformations (the number of sequences having as ground states the four conformations is $\Gamma_1 = 337$, $\Gamma_2 = 1303$, $\Gamma_3 = 1224$ and $\Gamma_4 = 1310$).

Column **B** in table 3.5 shows the design performance obtained when the set $\mathcal{D}$ is made up with 250 conformations with high overlap with the target one. In fact, by comparing these results with those shown in columns **A** and **R**, we can conclude that the overlap criterion captures some features of the optimal decoys set, i.e. the one obtained with the iterative method, whose performance is much better. This observation leads us to introduce a second geometrical criterion.

3.6 Similarity of conformations in terms of burial

In this section we define a new criterion to determine decoy conformations. Given a conformation Γ, let z_i be the number of nearest neighbour non-consecutive nodes of Γ around the i-th node. For a self avoiding chain in a simple cubic lattice $2 \leq z_i \leq 5$ for the extrema (i=1,27) whereas $1 \leq z_i \leq 4$ for the remaining nodes. Thus for each conformation a burial vector $\vec{z} = z_1, \ldots, z_{27}$ is defined. For interaction potentials favoring the burial of the most hydrophobic amino acids and exposition of the most hydrophilic ones, we might expect that the most competitive decoys with a given target conformation Γ_0 are those with a burial vector similar to Γ_0. We define a distance between Γ and Γ' in terms of their burial vectors $\vec{z}$, as

$$B(\Gamma, \Gamma') = \left(\sum_i [z_i(\Gamma) - z_i(\Gamma')]^2 \right)^{1/2} \tag{3.10}$$

The minimum value of B is zero. For some conformations, Γ_0, with high encodability we found that there are no Γ's, apart from the target itself, such that $B(\Gamma_0, \Gamma) = 0$. This is the case for Γ_2 and Γ_4, whereas Γ_1 and Γ_3 have *siblings*, that is conformations for which the score function B is equal to zero (respectively 2 and 3).

Given a target Γ_0, we built a decoys set $\mathcal{D}$ made up of the 250 closest conformations (in terms of burial distance) to calculate the approximate free energy in eq. (3.5) The performance obtained in designing is summarized in

table 3.5, column **C**. The results show that this geometrical criterion seems to be worse than the overlap method: this might depend on the fact that our HP Hamiltonian does not correctly account for a burial effect, i.e. only H-H contacts give a contribution to the energy.

In order to gain a more complete insight, we have used the same four decoy sets as above with another set of interaction parameters, i.e. $\vec{B} = (B_{HH}, B_{HP}, B_{PP}) = (-2.3, -1, 0)$ ([35]). The results are presented in table 3.6; they are quantitatively similar to the previous ones, except for the burial decoys, whose performance remarkably increases. We note that both criteria

interactions: $\vec{B} = (-2.3, -1, 0)$					
	A	**R**	**B**	**C**	**M**
Γ_1	69	0	5	8	15
Γ_2	100	2	25	51	77
Γ_3	86	3	18	18	21
Γ_4	99	4	38	56	68
average	**88**	**2**	**21**	**33**	**45**

Table 3.6: The same as in table 3.5, but with different interaction potentials. With this interaction potentials geometrical criteria, in particular **C**, are quite efficient. The mixed set (column **M**) is obtained using 50 high conformations and 200 low burial distance conformations.

(B and C) capture some geometrical features of the optimal decoys. So we created another set by combining the high overlap and low burial distance decoys. The results (see table3.6, column **M**) are remarkably good for two targets, Γ_2 and Γ_4. However they fail with Γ_1 and Γ_3. This is not surprising for Γ_1 which has an encodability much lower than Γ_2, Γ_3, Γ_4 (see caption Tab. 3.5). However it is rather surprising for Γ_3. As we will suggest below, this is probably due to the different number of siblings of these conformations.

An interesting feature of any geometrical method is that the temperature T can be varied in order to obtain solutions with the desired stability. By applying the strategy outlined in sec. II, if the free energy is approximated

with the proper decoys set $\mathcal{D}$, then the solutions that can be found are likely to have a folding temperature T_F higher than the temperature T. We have tested the robustness of our geometrical methods by rising the design temperature (from $T = 0.1$ to $T = 0.5$). It turned out that conformations with siblings (Γ_1 and Γ_3) are hardly designable with any decoy set, including the iterative one (we just could not find 100 sequences satisfying Eq. (3.4)), whereas the performance of the geometrical methods on Γ_2 and Γ_4 is even better than at lower temperature. The efficiency of the mixed geometrical method (method **M**) has been verified also for other nine conformations without burial siblings,for which the true encodability was unknown: here the average success obtained is 75.5%, that is comparable with Γ_2 and Γ_4. This results shows that the number of siblings is a stringent criterion to determine the designability of a conformation, at least within our design scheme.

interactions: $\vec{B} = (-2.3, -1, 0)$, $T = 0.5$					
	A	**R**	**B**	**C**	**M**
Γ_1	0	1	6	5	0
Γ_2	100	10	54	70	91
Γ_3	9	1	8	2	30
Γ_4	100	4	30	82	95
average	**52**	**4**	**24**	**40**	**55**

Table 3.7: The same as in table 3.6, but with higher design temperature ($T = 0.5$ instead of $T = 0.1$). In some cases less than 100 sequences were found: there we reported the true fraction of solutions over sampled sequences instead of the percentage.

3.7 Encodability, designability and burial degeneracy

In design problems, at least for protein models, the aptitude of a target conformation to be designed with stochastic sampling in the sequence space is related to its encodability, i.e. the number of sequences having as unique ground state the target conformation. However, it is likely also other characteristics of the sequence space affect significantly this aptitude. For some conformations the landscape of $H_s - F_s$ in the sequence space could be smooth with a very large funnel in the neighborhood of good sequences, facilitating the search by stochastic methods. In other cases the landscape is very rough with good sequences spread out in all the sequence space. In the last section we noted that there are two highly designable conformations (Γ_2 and Γ_4) which have a non-degenerate burial vector, i.e. no other conformation has the same burial vector. We have looked at the possibility that the high designability of the conformations we have used in the previous sections is related to the fact that these conformations have a small number of burial siblings, i.e. their burial vector has a low degeneracy. To do this, we first investigated, in an analogous spirit to [36], the relationship between encodability and burial degeneracy. A random sampling of 10^6 sequences, each folded to its ground state, permitted to identify some encodable conformations: these conformations were ranked into different sets, depending on the number of design solutions found during the sampling; this ranking is related to the exact encodability of the conformations. Then 100 conformations at random were picked out from each set; for each conformation the burial degeneracy d_B was calculated. Then, was obtained the average value $\langle d_B \rangle$ for each set (see Table 3.8).

Table 3.8 shows that $\langle d_B \rangle$ decreases as the estimated encodability increases. This data is confirmed by fig. 3.5, where distribution of d_B for two sets of conformations with high and low encodability is shown.

Only 120 out of the 103346 compact conformations of length 27 have no compact conformations with the same burial vector. We have verified that among them there are the most designable conformations with both the

Encod.	$\langle d_B \rangle$
0	31.2
1	19.2
2	19.8
3	18.7
≥ 10	11.1

Table 3.8: Conformations are ranked as a function of the number of the estimated encodability. The encodability has been estimated with a random a sampling of 10^6 sequences. In particular conformations have been gathered in five sets (0,1,2,3, $\geq$ 10), referring to the number of sequences having it as unique ground state.In the second column the average number of siblings within each set is calculated. This average decreases as the encodability increases.

interaction parameters we used before: this means that they are likely to be highly encodable, but also that designing them with a geometrical method is quite easy.

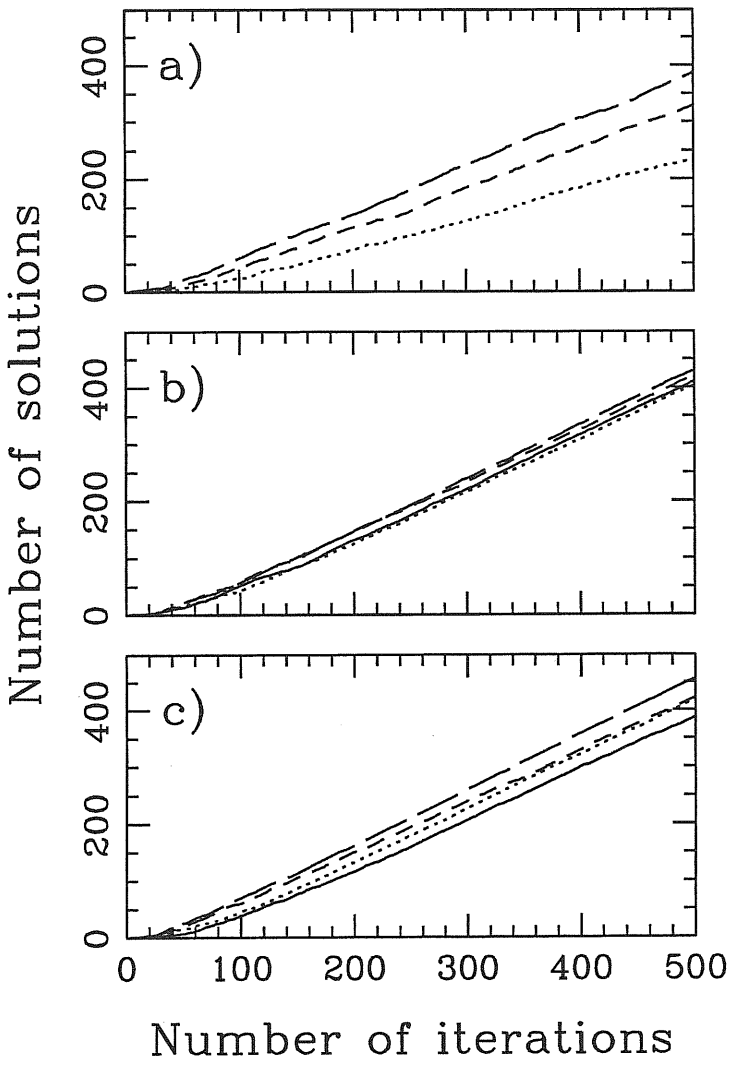


Figure 3.3: Number of extracted solutions versus the number of iterations for (a) HP interactions, (b) 6 amino acid classes and (c) 20 classes. The ideal curve, corresponding to efficiency 1, should have slope 1. Plots referring to structures Γ_1, Γ_2, Γ_3, Γ_4 are denoted with continuous, dotted, dashed and long-dashed lines respectively.

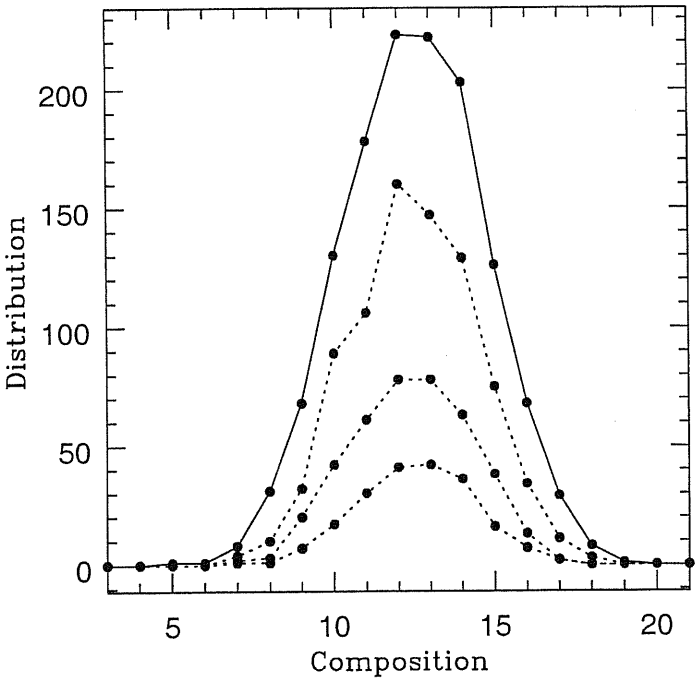


Figure 3.4: Histogram of the number of extracted solutions as a function of sequence composition (HP model). Curves pertain to an HP-design attempt on structure Γ_4 at different values of N_{it} : $200, 400, 800, 1719$. It can be seen that the efficiency of the design technique is independent of the sequence composition.

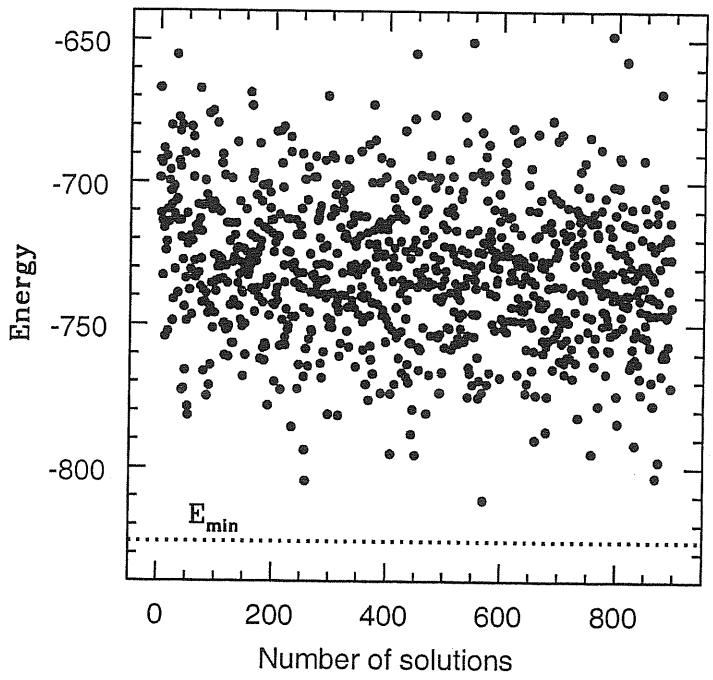


Figure 3.5: Energy of the solutions found for structure Γ_4 (6-class model) at fixed composition $(4, 4, 4, 5, 5, 5)$

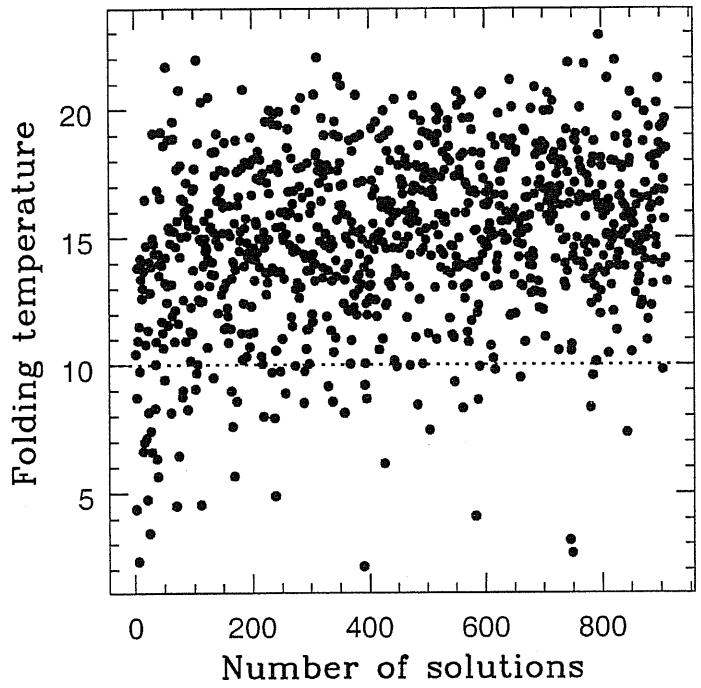


Figure 3.6: Folding temperatures of solutions designing structure Γ_4 (6-class model) as a function of the order of extraction. Very few solutions turn out to have a folding temperature below the simulation temperature $T = 10$ (shown with a dotted line).

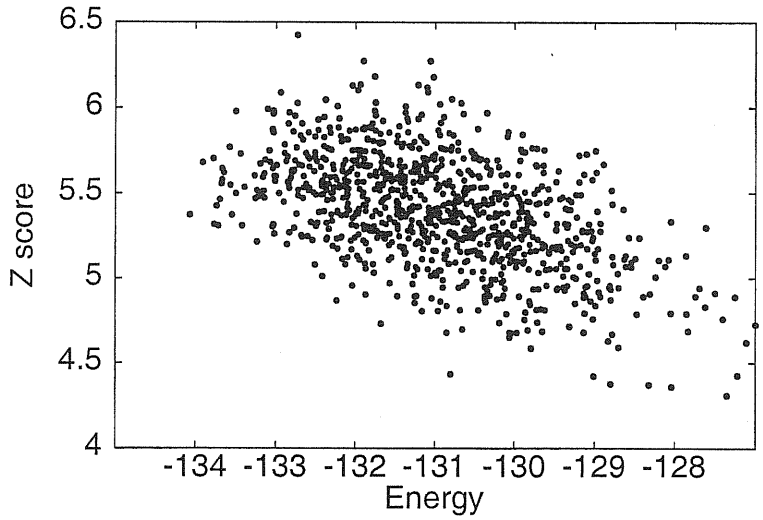


Figure 3.7: Scatter plot of the Z_{score} against native-state energy of extracted solutions designing structure Γ_1. The data are for a 20-letter alphabet of amino acids at fixed and nearly uniform composition.

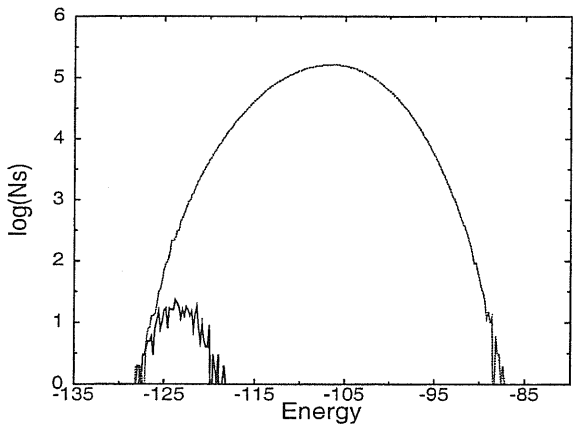


Figure 3.8: Solid line: distribution (in arbitrary units) of solutions (good sequences) to the design problem on structure Γ_1 (20 letter alphabet). The dotted line denotes the distribution containing bad sequences. The data was obtained by randomly sampling 10^7 sequences with fixed composition.

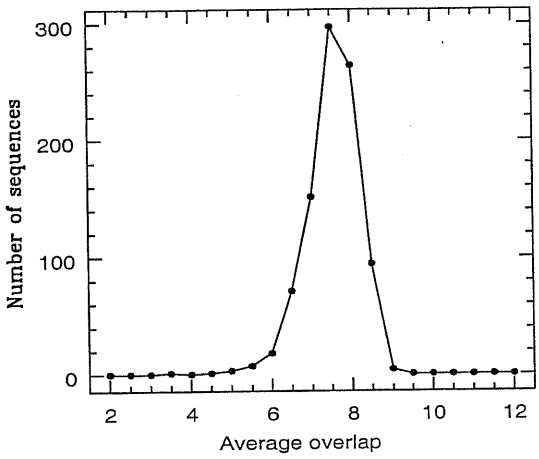


Figure 3.9: Histogram of the average overlap (sequence identity) of solutions for Γ_4 (6-class model). For a given sequence the average overlap is calculated over all other extracted solutions.

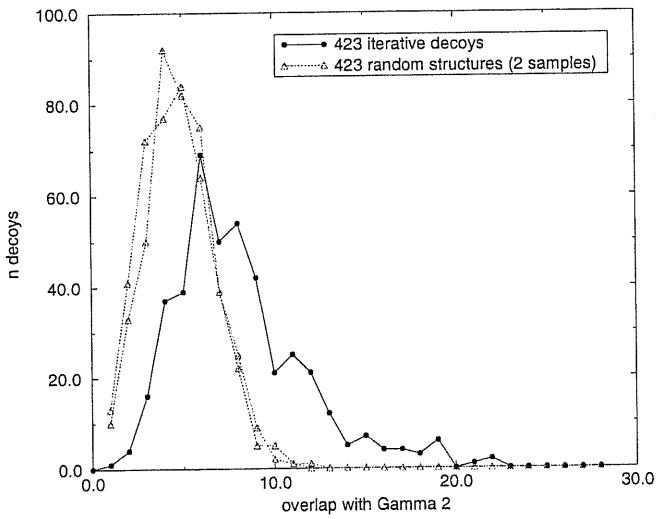


Figure 3.10: Distribution of the decoy conformations as a function of the overlap with the target conformation Γ_2 (solid line: iterative set; dashed line: random set). The decoy set is biased towards conformations with high overlap.

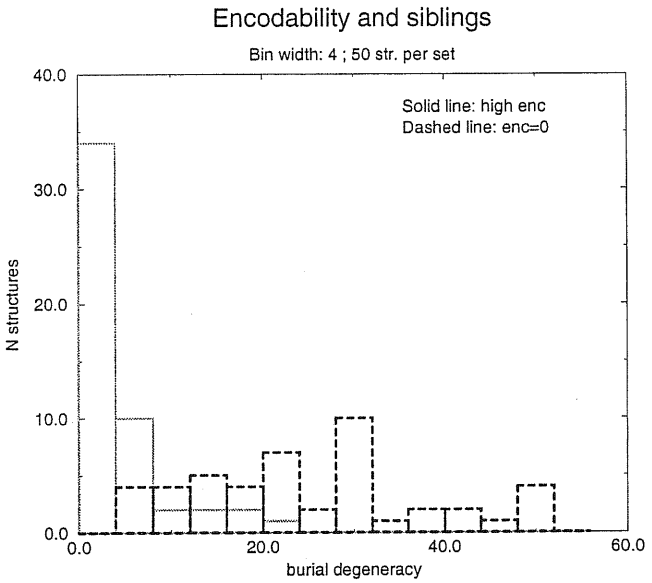


Figure 3.11: Histogram of number of conformations with a given number of burial siblings $N(\Gamma)$. In the set of highly encodable conformation (solid line) there are a lot (34) of conformations with less than five siblings each, whereas in the set of low encodable conformations there is no one.

Chapter 4

Knowledge-based approach to protein design

Protein design, having obvious practical and evolutionary significance, has attracted considerable attention and effort of experimentalists and theorists, especially in the last years [52, 56, 70, 65, 14, 50, 13, 64, 44, 43, 45, 75, 83, 87]. The difficulty of the protein design problem is enormous because, in principle, a rigorous approach [65, 45] would entail a simultaneous exploration of both the family of viable sequences and the family of physical conformations. Progresses in protein design obtained using model proteins have been remarkable in the last ten years. However, despite several efforts [69, 77, 43], the extension of this machinery to the design of natural proteins has not yet reached maturity. The reasons are mainly two:

- the difficulty in giving a reasonable functional form of $H_s(\Gamma)$ [81],
- the impossibility to verify whether the predicted sequence really folds in the desired conformation, without performing an expensive real experiment.

These two obstacles are absent in simplified models where $H_s(\Gamma)$ is assigned *a priori* and the exact solution can be rigorously found. In this chapter we investigate the degree of accuracy one can reach when designing natural structures (taken from the Protein Data Bank (PDB)) by using simple functional form of $H_s(\Gamma)$ and a limited number of classes of amino acids. The

unknown parameters defining $H_s(\Gamma)$ are determined with a strategy [12, 63] based on the observation that physical forms of the energy ought to guarantee that any amino acid sequence should recognize its native state as the conformation with minimum energy-score and maximum thermodynamic stability. We use such optimized energy functions to design PDB protein conformations by applying some of the above-mentioned theoretical techniques. Finally we check the quality of our predicted sequences not only through a mere comparison with the naturally folding amino acid sequences (retrieved from the PDB) but performing a statistical analysis of our results with respect to the full set of homologous sequences (e.g. sequences folding to the selected protein or in homologous conformations) [18]. In this way we try to establish which amino acids are important to stabilize the sequence in the target structure and we compare these sites with sites important for the folding process, i.e. sites belonging to the folding nucleus [68]. Furthermore, we show how it is possible to give a degree of reliability to any design attempt.

The chapter is organized as follows: in section 1 the schematic representation of protein structures is illustrated together with the energy functions and the classification of amino acids that have been used. In section 2 the new strategy to estimate interaction potentials is derived while section 3 is devoted to explain the design procedure and to discuss the results, which are summarized in section 4, while technical details are given in the Appendices.

4.1 Protein modeling

4.1.1 Two- and three-body energy functions

As is customary in many numerical approaches to folding and design strategies we shall also adopt a simplified protein backbone representation that neglects amino acid rotameric degrees of freedom. In fact, we shall use the common coarse-grained model of PDB proteins in which each amino acid unit is represented by a centroid placed on the β carbon (for glycine the coordinates of the centroid can be estimated by the local geometry of the backbone [54]). According to this procedure any protein conformation, Γ,

obtained by a sequence of N amino-acid is specified through the $3N$ Cartesian coordinates:

$$\Gamma \equiv \left(\vec{r}_1^{\,C_\beta}, \vec{r}_2^{\,C_\beta}, \ldots, \vec{r}_N^{\,C_\beta}\right) \ . \tag{4.1}$$

This simplification is mainly dictated by the necessity to deal only with the main protein degrees of freedom, but, as we shall mention, it is also particularly appropriate in design contexts. Furthermore, we shall also partition the 20 types of amino acids into a restricted number of classes. This simplification is not dictated by the numerical convenience of dealing with a restricted sequence space (in fact, the design strategy outlined in section III can be straightforwardly applied to 20 amino acids classes). Rather, the choice follows from the need to have a sound statistical basis for estimating the free-energy contribution of interacting amino acid classes and also from the observation that most amino acids in natural proteins can be substituted without disrupting native folds [30]. Hence, within the present design scheme we aim at predicting the classes of amino acids designing a given structure. As in ref. [75], the putative solution could, in principle, be fine-grained into 20 amino acids alphabet by using steric packing and solvation constraints.

Finally, the last ingredient of our strategy is the introduction of a suitable (free) energy scoring function. The most popular choice adopted in simplified models is the pairwise-interaction form

$$H_s^{(2)}(\Gamma) = \sum_{i<j} \Delta_{ij}^{(2)}(\Gamma) B_2(s_i, s_j) \ , \tag{4.2}$$

where i, j are the positions along the sequence of the of amino acids and the sum is taken over all possible pairs. $B_2(s_i, s_j)$ represents the interaction strength of the amino acid pair s_i and s_j. However, only amino acids that are close enough will interact in a non-negligible way. This is enforced with a suitable weight function, or contact map, $\Delta_{ij}^{(2)}(\Gamma) \equiv f(x = |\vec{r}_i - \vec{r}_j|)$, where:

$$f(x) = \frac{1}{2}\tanh(a_0 - x) + \frac{1}{2} \tag{4.3}$$

and a_0 is a cutoff value that we choose equal to 8Å.

Hydrophobic	Neutral	Charged
Alanine	Asparagine	Arginine
Isoleucine	Cysteine	Histidine
Leucine	Glutamine	Lysine
Methionine	Glycine	Aspartic acid
Phenylalanine	Serine	Glutamic acid
Proline	Threonine	—
Tryptophane	Tyrosine	—
Valine	—	—

Table 4.1: The three columns contain the three classes of amino acid we have used in the design strategy.

In addition to this scoring function (4.2), and to assess possible design improvements, we shall adopt also one including three-body interactions:

$$H_s^{(3)}(\Gamma) = H_s^{(2)}(\Gamma) + \sum_{i<j<k} \Delta_{ijk}^{(3)}(\Gamma) B_3(s_i, s_j, s_k) \ , \tag{4.4}$$

where $\Delta_{ijk}^{(3)}(\Gamma) \equiv \Delta_{ij}^{(2)}(\Gamma)\Delta_{jk}^{(2)}(\Gamma)\Delta_{ki}^{(2)}(\Gamma)$. The matrix B_3 represents the effective three-body interactions among the different classes of amino-acids. Indeed, it has been recently suggested that pairwise energies [81] may be unsuitable to describe effective amino acid interactions in proteins. Hence, the introduction of three-body terms might be regarded as the first correction term to (4.2) in an expansion scheme where all many-body interactions are included.

4.1.2 Partitioning the 20 amino acids into classes

In order to estimate the interaction-potential matrices B_2 or B_3 appearing in Eq. (4.2) and (4.4), we introduce a suitable classification of the 20 types of amino acids. In an attempt to go beyond previous studies [77, 43] where the two letter code was used, we decided to subdivide amino acids into three classes (table 4.1).

Although many other subdivisions could be possible, adopting the one

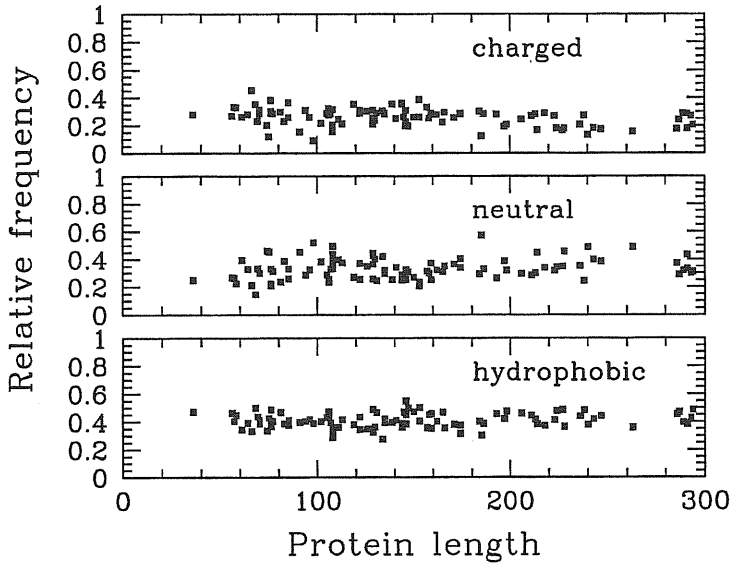


Figure 4.1: Typical relative frequency of the three classes in which amino acids are partitioned, as a function of the protein length.

followed here has the advantage that, besides clustering amino acids according to their chemical similarities, it creates classes which are almost equally populated (see Fig. 4.1). Since the B matrices are symmetric, the number of entries to be determined is 6 and 10 for B_2 and B_3 respectively.

4.2 Learning the interaction potentials

4.2.1 A new theoretical approach

An efficient way to estimate the effective potentials B_2 and B_3 was pioneered by Crippen [37] (and recently optimized and used [80, 16]). This scheme aims at finding a set of potentials so that, given a protein sequence s, its native state Γ is recognized as having energy substantially below that of any other equally long conformations Γ' (assumed to be outside the native basin of Γ [25]). For a generic energy function $H_s(\Gamma)$ this requires:

$$H_s(\Gamma) < H_s(\Gamma') \tag{4.5}$$

A key difficulty in turning this idea into a powerful automated scheme is the choice/generation of physically viable decoy structures, Γ'. In many instances the decoys are generated by taking compact "chunks" of suitable length from a bank of proteins (gap-less threading). Such decoys may not be physical for certain sequences (for example due to steric clashes) so that the inequalities (4.5) may enforce rather loose or unrealistic constraints on the extracted potentials.

The first goal in this chapter is to propose a strategy to overcome this difficulty. Our idea is based on the fact that the thermodynamic stability requirement, (4.5), should be simultaneously satisfied as much as possible for a whole set of conformation Γ_c which compete significantly with the native state.

This thermodynamic requirement can be accomplished by imposing that

$$H_s(\Gamma) \ll \langle H_s \rangle \ , \tag{4.6}$$

where the average $\langle \cdots \rangle$ is carried out over all the set Γ_c. In a more mathematical spirit, equation (4.6) can be derived as follow: Eq. (3.1) gives the statistical probability that a given sequence s is in a specific conformation Γ at temperature T. If Γ is the native state of s, below the folding temperature only the conformations present in Γ_c give a not vanishing contribution to Z_s. By writing $Z_s = \exp(\log Z_s)$ and taking the first order term in its cumulant (high-temperature) expansion , the condition of maximizing $P_s(\Gamma)$ yields Eq. (4.6).

Due to the linear dependence of the energies H_2 and H_3 on the contact maps (the only factors that contain geometric information about structures), the r.h.s. of Eq. (4.6) can be re-casted into the following forms:

$$\langle H_s^{(2)} \rangle = \sum_{i<j} \langle \Delta_{ij}^{(2)} \rangle B(s_i, s_j) \ , \tag{4.7}$$

and:

$$\langle H_s^{(3)} \rangle = \sum_{i<j} \langle \Delta_{ij}^{(2)} \rangle B_2(s_i, s_j) + \sum_{i<j<k} \langle \Delta_{ijk}^{(3)} \rangle B_3(s_i, s_j, s_k) \ . \tag{4.8}$$

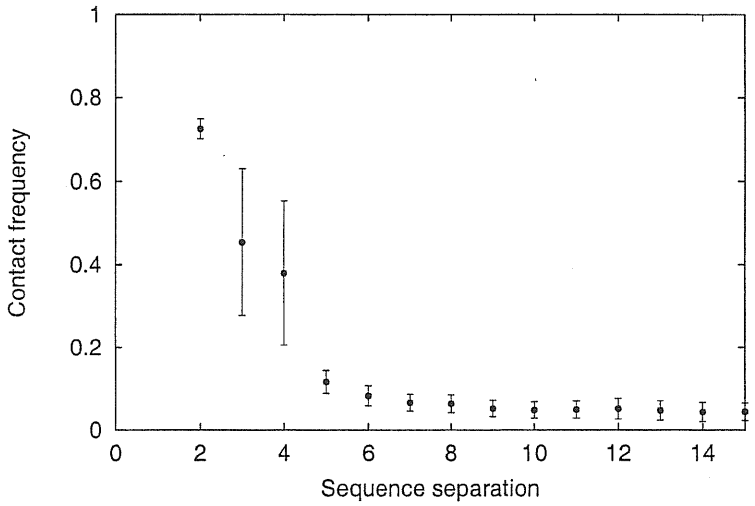


Figure 4.2: $\langle \Delta_{ij}^{(2)} \rangle$ for small values of $k = |i - j|$. For $k = 3, 4$ big errorbars are due to the presence of α and non-α proteins in our protein set.

Notice that both $\langle H_s^{(2)} \rangle$ and $\langle H_s^{(3)} \rangle$ depend on the sequence s and no more on the structure Γ. A detailed technical description of how the averages in equations (4.7) and (4.8) are obtained, is presented in Appendix A. To summarize, the functional dependence of $\langle \Delta^{(2)}(i,j) \rangle$ was determined by inspecting its behaviour as a function of i, j. The main difficulty was to find a form suitable to represent the behaviour of $\langle \Delta^{(2)} \rangle$ for a variety of protein lengths and families. A very satisfactorily "collapse" of data from many structures could be obtained by assuming that $\Delta^{(2)}(i,j)$ merely depends on i and j, irrespective of the chain lengths, for $|i - j| < 16$, as shown in Fig. 4.2.

This is reasonable, since the frequency of "local" contacts is not expected to be influenced by the overall protein shape or length. Contacts between residues with sequence separation larger than 16 are rather rare hence were modeled by assuming a constant frequency of occurrence, $\Delta_2^{(0)}$. The value of $\Delta_2^{(0)}$ is regarded as a free parameter that is to be tuned separately for each protein length so that the average number of overall contacts, $\sum_{i,j} \Delta_{i<j}^{(2)}$, matches the number observed in nature. An analogous procedure was followed for the three-body weight function, whose functional form is shown

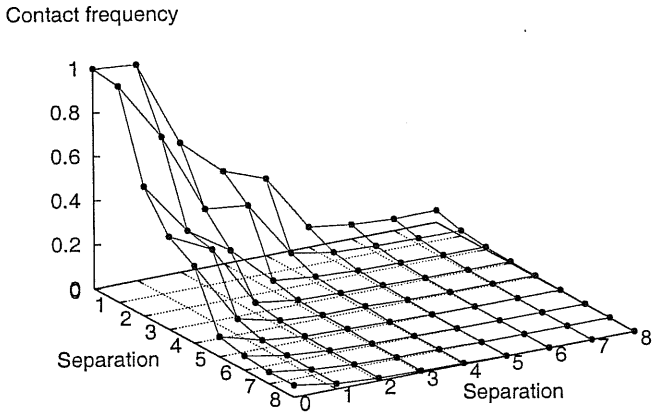


Figure 4.3: $\Delta^{(3)}(k1, k2)$ for small values of $k1 = |i - j|$ and $k2 = |j - k|$. Fluctuations are of the same order of the previous figure.

in Fig. (4.3). For determining the potentials we consider a set of 31 non redundant proteins listed in Tab. 4.2. Hence, through Eq. (4.6) and (4.2), (4.7) (or (4.4), (4.8)) we obtained one inequality for each protein in the set (that we shall term *training set*) The determination of the potentials, B, was done by employing an efficient algorithm, called perceptron, that is guaranteed to provide the best solution for a whole set of inequalities. The method is outlined in Appendix B. In our case, we have one inequality for each of the training proteins. Clearly, by suitably choosing the B's it is possible to make arbitrarily large each inequality individually. the perceptron procedure allows to find the best B's that make all inequalities as large as possible simultaneously. There is no guarantee, however, that the inequalities can be all satisfied. Indeed, as a rule of thumb, when the number of inequalities greatly exceeds the number of parameters, no solution can be found if the functional form of it and/or the approximations involved are not too satisfactory. In our case we dealt with 5 (or 15) parameters and succeeded to find physical solutions to the problem. This suggests that the adopted form of the energies were reasonable, otherwise the problem would have been unlearnable. A further proof of this is that, by using a different set of training proteins,

SET 1					
1acp	77	1beo	98	1cei	94
1coo	81	1cty	107	1erv	105
1fd2	106	1fkb	107	1fna	91
1fow	76	1kum	108	1mit	69
1opd	85	1pdr	99	1rro	108
SET 2					
1shg	57	1tul	108	1who	96
1yat	113	1yeb	108	2c2c	112
2fxb	81	2imm	114	2mcm	112
2mhr	118	2rhe	114	351c	82
3b5c	93	3ssi	113	3wrp	108
9rnt	104	-		-	

Table 4.2: PDB name of the 31 proteins used in our design scheme with their respective number of amino acids.

nearly the same optimal parameters were obtained, a fact that corroborates the robustness of the potential extraction procedure.

4.3 Designing PDB structures

4.3.1 The design strategy

Once the potentials were determined, the energy scoring function of any desired conformation can be computed within the energies defined in Eq. (4.2) or in Eq. (4.4). In order to tackle our ultimate goal, the design of protein conformations , it is necessary to define the design procedure. It has been discussed in the introduction that a rigorous, but unpractical way, of pursuing this objective consists of finding, for a given conformation Γ^*, the sequence (or sequences) s^* maximizing the occupation probability $P_s*(\Gamma^*)$. In the previous section we have however shown that, for the correct energy parameters, the desired sequence should satisfy the inequality:

$$W(s,\Gamma^*) = H_s(\Gamma^*) - \langle H_s \rangle \ll 0 \ , \tag{4.9}$$

Therefore , since we have obtained a reliable estimate of $\langle H_s \rangle$ we can use Eq. (4.9) to perform protein design. In practice, given the target conformation we search for the sequence that minimizes the function $W(s,\Gamma^*)$ where all the quantities are calculated with the above determined potentials. The optimal solution is identified by a stochastic procedure (simulated annealing) in sequence space, the elementary move being the random mutation of a fraction of residues from one class to another. Generally, the most stringent way to test the reliability/validity of the extracted parameters would be to apply them to design proteins unrelated to the training set. However, due to the "nearly perfect" independence of the parameters from the training set (a result reflecting the benefit of the coarse-graining into three amino acid classes) this precaution is unnecessary in our case. Therefore we shall use the extracted potentials to design the training proteins. Finally, the representation described by Eq. (4.1) ensures the absence of any information that could distinguish different types of amino-acids thus allowing an unbiased test.

As in refs [43, 77], the success rate of the design procedure is defined as the fraction of correctly predicted amino-acid classes with respect to those of naturally occurring sequences (as found in the PDB) for the chosen configuration. The success rate for a randomly designed sequence where each residue is assigned randomly to one of the three classes would be 33%.

For all the considered conformations (see fig. 4.4) we obtained a success rate between 40% and 55%.

This success rate can be compared with optimized success rates for two amino acid classes [43] which is, on average, $\sim$ 75%. Clearly, increasing the number of classes makes the problem more difficult, hence a reduced success rate. It is interesting, however, to note that the success rate of the optimal design strategy remains above the random-guessing threshold by about 20%, as for the two-letter case. It is also interesting to notice that this rate does not improve (see Fig. 4.4) by working with the concentration of amino-acid biased towards the composition of the wild-type sequence or even by using

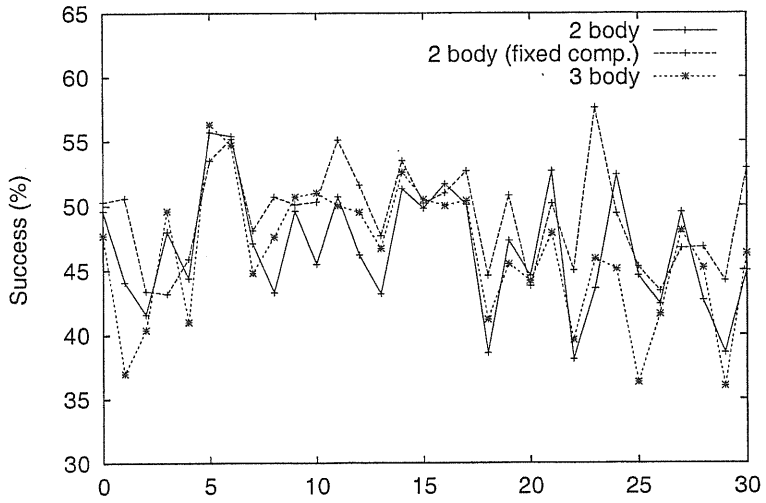


Figure 4.4: The success is here defined as the similarity between the designed sequence and the wild-type sequence as retrieved from the PDB file. The designed sequence has been obtained by a minimization of W (simulated annealing) and the success has been obtained as an average over ten independent minimizations. The three curves refer to the design using Eqs. (4.2), (4.6) and (4.7) with arbitrary or fixed composition –i.e. exploring only sequences with composition not too different with respect to the composition of the wild-type sequence–, and using Eqs. (4.4), (4.6) and (4.8).

the three body energy. This possibly suggests that important features of real proteins have been equally neglected by all these kinds of energy function.

On the other hand, the one-to-one comparison between the designed sequence (defined as the one that minimize $W(s)$) and naturally occurring ones could not be the best check to do. The reasons are twofold:

- homologous sequences, e.g. sequences which roughly fold in the same native state, can differ by up to 70% (similarity) of their amino-acidic composition. A one-to-one comparison (although averaged over many sequences) could not be sufficient to verify if our wrong predictions are involving the most important amino-acids or only the marginal ones;

- naturally occurring proteins may not have necessarily evolved to max-

imize the occupation probability but also to ensure a fast folding process ([68, 39, 24]). Therefore to select only the sequences that minimize $W(s)$ could be a too drastic selection criteria, especially considering that we are working with unperfectly parameterized energy-scoring functions.

In order to estimate the importance and the effects of these two arguments we performed the analysis, discussed below.

4.3.2 Homologous sequences and comparison of similarities

It has been shown by Clothia *et al.* [10] that naturally occurring sequences with a very low degree of similarity ($\sim$ 30%, but this rate is very dependent on the length of the alignment) can be homologous, that is they adopt almost the same three-dimensional structure [18]. This kind of analysis has been performed by considering all the 20 kinds of amino acids. We decided to re-analyze the set of protein sequences treated in ref. [62] by using the same three-letter classification of amino acids employed in our design procedure. In such context, the similarity between two sequences of classes of amino-acids is defined in the same way as the design success score, that is we check whether each amino acid pair in optimally aligned homologous sequences (data from the HSSP database [62]) belong to the same class. By definition, this similarity cannot be smaller then the 20-letter one.

The results for a specific protein, 1acp, are given in Fig. 4.5. It turns out that, on average the homology threshold of 30% for the full amino acid alphabet corresponds to 55% when the three letter code is used. This value is remarkably close to the best design scores achieved with our procedure. This does not imply automatically that our solutions are viable. Site-directed mutagenesis experiments have shown that a small proportion of protein sites do not tolerate any substitutive mutation at all (otherwise the native state would be destabilized). It should then be checked whether such key residues, which are conserved in homologous proteins, are conserved also by our design strategy. In one of following sub-sections we shall examine this issue in

connection with heavily investigated proteins, such barnase and ci2, and we will show that, as a by-product of the design procedure, the location of such sites can be easily predicted with high reliability. This is not a proof that our design solutions, although different from the native one, are correct, too, but it sheds a new light about their validity.

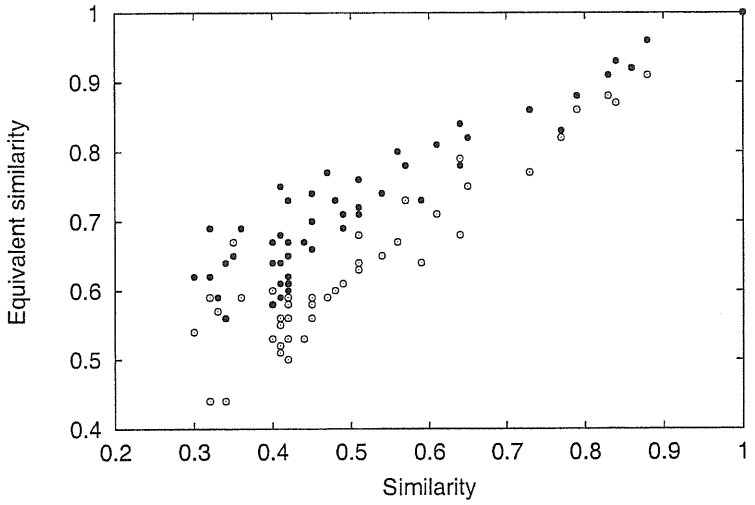


Figure 4.5: Equivalent similarity evaluated for two different classifications: the one used in this study (dark circle) and another random (open circle) as a function of the similarity defined for the 20 amino acids. The figure refers to comparison between protein sequence 1acp with 51 sequence of homologous protein.

4.3.3 Are extremized sequences the best?

The design analysis we have described so far was based on the selection of sequences that minimize $W(s)$, i.e. on the maximization of the gap between the energy of the sequence in the target conformation and the average energy $\langle H_s \rangle$. However, it is presumable that the evolutionary pressure towards rapid and reliable folding [40] has not taken the maximization of inequality (4.9) to the extreme, but to a lower threshold sufficient for biological purposes. For this reason we chose to test the success rate not only for the minimum value of

$W(s)$, but also for other sequences. In particular it is interesting to compare all the sequences s with $W(s) < W(s^*)$, where s^* is the wild-type sequence. For each annealing temperature we extract 100 decorrelated sequences and make statistical analysis on this sequence set. We evaluate the average of $W(s)$ for this set and a "super-sequence" by applying a pointwise majority rule to this set. In other words for each site we assign the most frequent amino acid class observed in this sequence set at the given location. Fig. 4.6 shows the data pertaining to such design attempts on five different proteins. It appears that, indeed, the highest matching with the native sequence, is not obtained for the lowest value of W, but for higher ones.

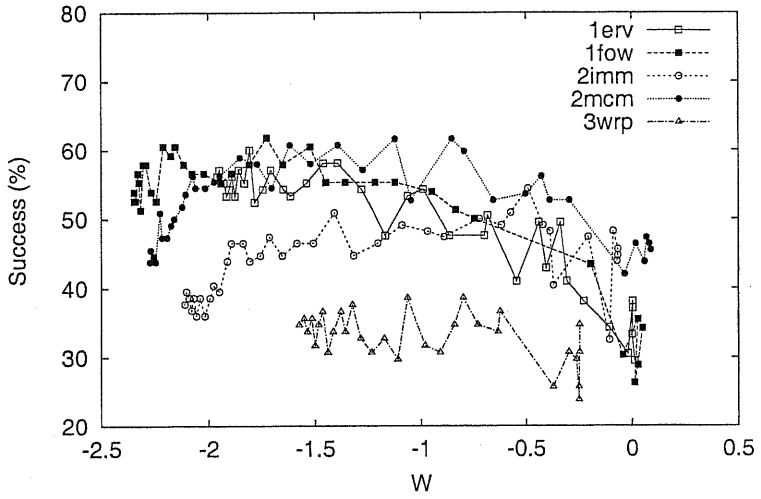


Figure 4.6: The success as a function of the cost function $W(s, \Gamma_t) = H_s(\Gamma_t) - \langle H_s \rangle$ per site. Success is defined here by majority rule on a sampling of hundred (decorrelated) sequences. The value of the cost function for the respective wild-type sequences is between -0.48 and -0.78.

This fact suggests a powerful way to improve the reliability, of the design strategy: we can select as putative solutions a wider range of protein sequences and then process the statistical information contained in them to yield a single "super-sequence". Furthermore, one can decide to make a prediction only for those sites where a class has an occurrence frequency bigger than some suitable threshold f_0. The number of sites N_s for which we make

such prediction is a decreasing function of f_0 and for a given f_0 depends on the fictitious temperature (at low temperature all the sites are locked). Fig. (4.7) shows success rate over the N_s betted sites for different values f_0 (data pertain to protein 1erv, other proteins produce analogous plots).

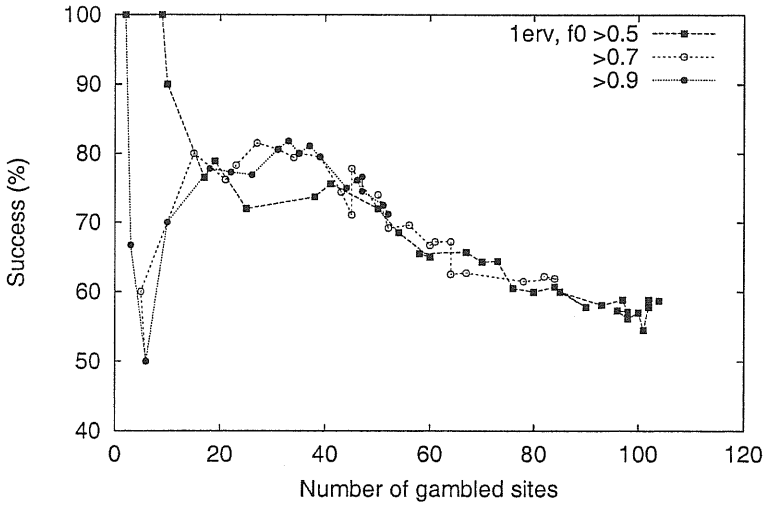


Figure 4.7: Success as a function of the number of betted sites for the protein 1erv. Betting over the 40 most locked sites it is possible to obtain almost an 80% of success. Note that success is almost independent on the frequency threshold f_0.

It is evident that, when N_s is small, the design procedure is very reliable: retaining the first 40 sites gives the impressive success rate of 80% . It is tempting to conjecture that the residues that are assigned with very little uncertainty by our design procedure (conserved design residues) could also correspond to conserved residues in nature. In the next section we shall examine in detail this possibility, and conclude that there is a significant correlation between the two.

4.3.4 Homologous sequences and conserved sites

It is well known [62] that homologous sequences present conserved sites, e.g. sites where the type of the amino-acid remains unalterated throughout

the full set of sequences. In Fig. 4.8 this fact is graphically elucidated (and even enforced) by analyzing the homologous sequences of protein 1erv with our tripartite classification of amino-acids. To each site we assign a color reflecting the conservation of the most frequent class observed in that position. A full conservation of H, N and C type is denoted with a saturated green, red and blue color respectively, while the lowest possible conservation of the most frequent class, 1/3, is associated with the white color. According to this scheme, sites with high variability will correspond to lighter nuances.

A visual inspection of the colors assigned to protein 1erv (top panel of Fig. 4.8) reveals that about 30% of the sites are highly conserved. We want to elucidate whether there exist a connection between such conservation of amino acids found in nature and the one emerging in the putative solution obtained from our design procedure.

To do this we performed a simple analysis of the design solutions at different values of the conservation threshold, W. In each batch of 100 design runs, the target value of W was fixed (in a stochastic way) by varying a suitable control parameter, T (by analogy, if we identify W as an energy cost function, T plays the role of the temperature). Finally, for each value of T we analyze the conservation of residues in the designed sequences and color them with the same scheme described above. The results are shown in the large box of Fig. 4.8.

For high values of T (high W) all the color intensities are very low indicating an uniform (random) distribution of the classes, but upon decreasing the temperature some of them start to be selected with higher and higher frequency. At very low temperature all the sites are locked in a particular class. This trivial situation is not shown in Fig. 4.8 which, instead, concentrates on the more relevant range of intermediate temperatures. The comparison of the native colored panel and the designed one strongly confirm the hypothesis that sites locking early (at high values of T) are related to the naturally conserved ones. This connection is examined in a more circumstantiated context in the next section where we consider two specific protein instances: barnase and chymotrypsin inhibitor.

An even more quantitative analysis of the correlation between designed

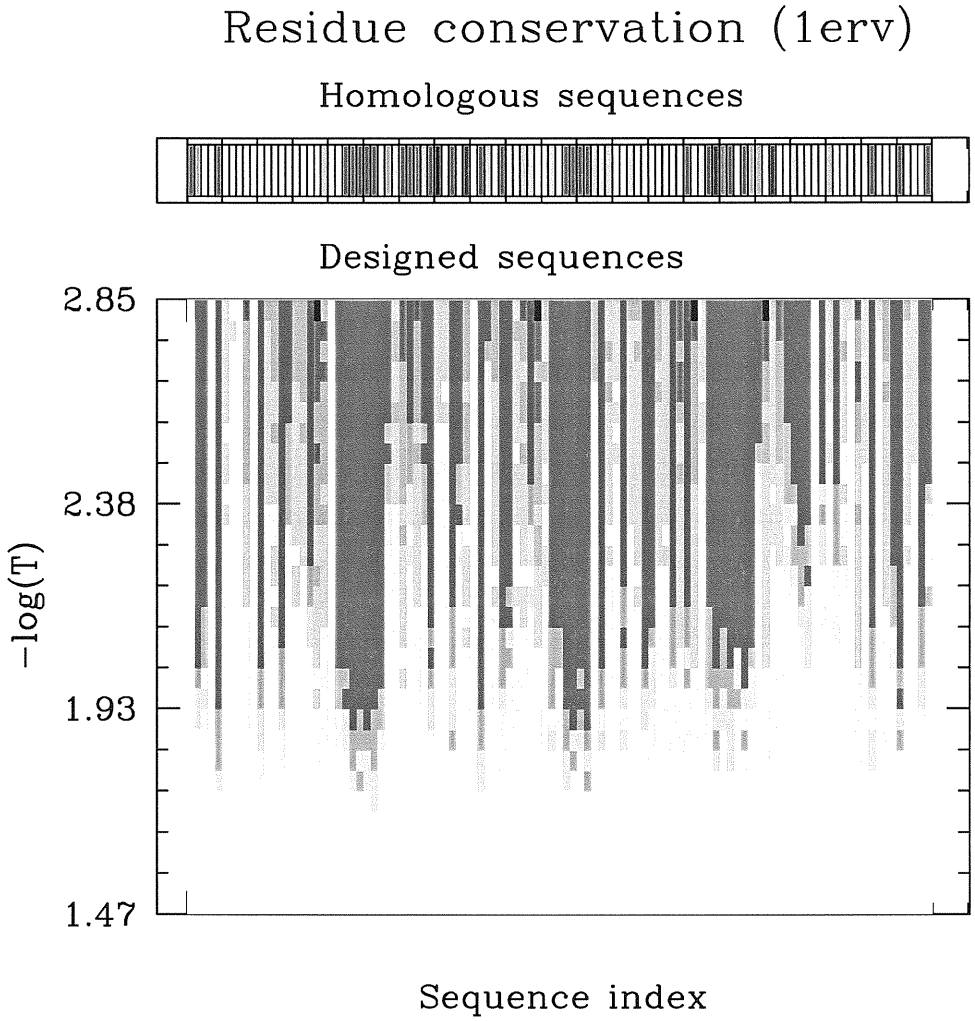


Figure 4.8: Color-coded conservation of residues in 1erv in natural context (top panel) and in putative solutions obtained with our design procedure. The color code, described in the text, assigns lighter colors to highly variable sites. The conservation in the natural context was obtained from the analysis of the HSSP database [62].

and homologous sequences can be obtained by a simple geometrical construction. For each amino acid located at site i in a given protein, a three-dimensional vector is constructed whose components are the frequencies with which the three classes appear: in the design sequences (and we term the vectors $\vec{f}_i^D$) or in the homologous sequences ($\vec{f}_i^N$). To make the comparison meaningful, the design procedure was carried out at a value of T chosen so that the fraction of conserved residues was similar to the one observed in nature. The vector of a site which is conserved in a specific class of amino acids is aligned with the associated axis whereas the vector of a non-conserved site has at least two non-vanishing components.

The angle θ_i formed by the two vectors $\vec{f}_i^D$ and $\vec{f}_i^N$ provides a quantitative measure of the correlation between residue conservation in the natural and design contexts. This angle is zero if the agreement is perfect, while it attains the maximum value of $\pi/2 \approx 1.5$ if a residue is maximally conserved in nature and minimally conserved in design (or the other way around).

In Fig. 4.9, we plot (for four different proteins) the histogram of these correlation angles (light grey). Remarkably , for all the proteins, the highest entries correspond to small angles and they represent a considerable fraction (1erv=24, 2imm=18, 2ci2=12 and 1a2p=20) of all sites, thus highlighting a highly significant agreement. To validate the design scheme is then crucial to verify if the highest agreement (small angles) is observed in correspondence of sites highly conserved in nature. This is indeed the case: in the same figure we plot, for each angle bin , the number of sites which are naturally highly conserved (dark grey), i.e. that have a conservation entropy, evaluated as in HSSP data bank [62], lower than ln(1.5) (ln(1) and ln(3) correspond respectively to the minimum and the maximum values for the entropy when only one class is assigned or all the three classes are assigned with equal probability). Almost all the sites with vanishing correlation angle satisfy this property!

We can then conclude that amino acids which, in our design scheme, are designed with a higher confidence strongly correlate with those that are conserved in natural sequences.

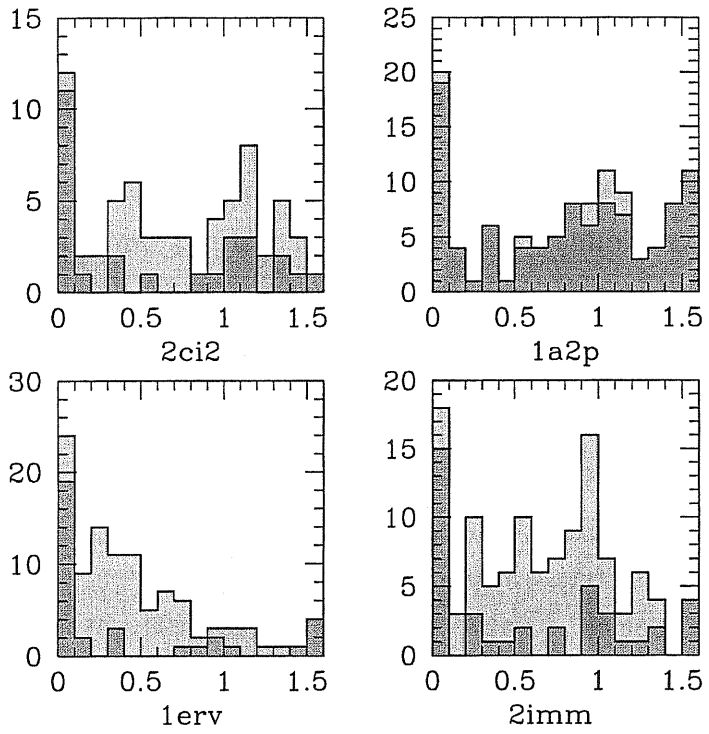


Figure 4.9: Distribution of the angles (in radians) between amino acid frequency vectors for designed sequences and aligned sequences for all the sites (light grey) and for conserved sites (dark grey). For this plot we considered conserved sites that with entropy less than ln(1.5).

4.3.5 Data for barnase and chymotrypsin inhibitor

In this last section we shall apply the design strategy to two proteins, whose folding process has been heavily investigated experimentally. With a series of key measurements [19, 28], Fersht and coworkers-workers have identified a restricted set of residues, the folding nucleus, which play a key role in the folding process in proteins such as barnase (1a2p) and chymotrypsin inhibitor (2ci2). While, generally speaking, naturally occurring proteins can tolerate a fair degree of amino acid substitutions without disrupting the native state, random mutations of sites in the folding nucleus will impair the folding process dramatically. Indeed, recent theoretical studies [40] have shown that key sites in the folding process nucleus are part of a bottleneck in the folding kinetic, which is mainly dictated by the native state topology. Overcoming such a bottleneck can occur only through a careful selection of the type of involved amino acids [7]. This novel arguments confirm and explain the observation already present in the literature [68] that sites involved in folding nuclei should have been conserved during the evolutionary process. Hence, our goal in this section is to design the backbone of 1a2p and 2ci2 and compare the set of residues, which are conserved in our design strategy with those in the folding nucleus. As already seen in the previous section, we identify the conserved residues by monitoring the frequency with which a given residue is assigned to one of the three classes during the lowering of W controlled by suitably changing the temperature-like parameter, T introduced in the previous section. As we said before, the tendency of one site to prefer one class over the others grows stronger as T is reduced, (e.g. minimizing W). However, not all sites show this preference at the same value of T as visible in Fig. 4.10 where we have shown the intensity with which protein sites in barnase are locked in the H, N and C classes. The most conserved residues are those for which the class-locking occurs at very high temperature. It turns out that the sites involved in the locking process occupy buried positions and are consistently assigned to the hydrophobic class. A visual inspection of Fig. 4.10 reveals that sites that are first locked in barnase correlate well with the hydrophobic *core1* which Fersht identified as the initiator of the

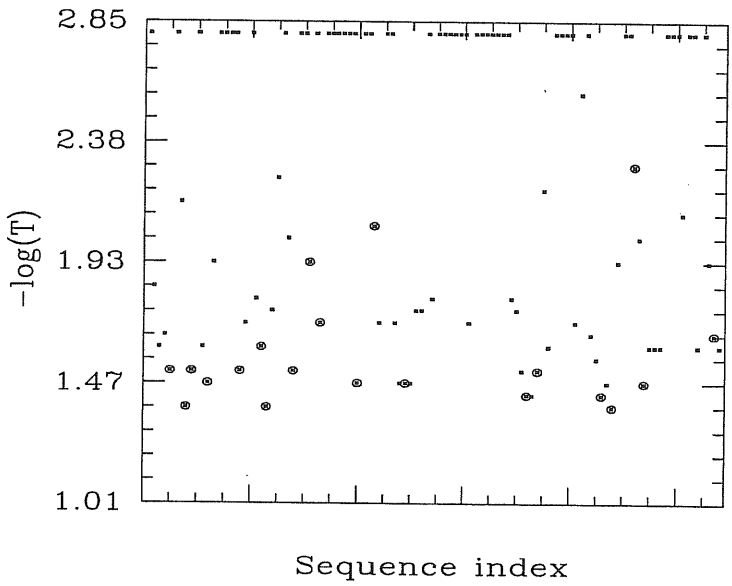


Figure 4.10: Quenched index versus sequence index for barnase. Quenched index is here defined as the first index for which the relative frequency for the hydrophobic class is greater than 0.5. Circled dots represent sites belonging to *core1*, *core2* or *core3* [66].

folding transition.

An excellent agreement with experimental findings is also observed for 2ci2, where key sites have been pinpointed through mutagenesis experiment and measurements of ϕ-values [28]. The key sites have been identified as those positions which are the highest rank in order of early locking. As visible in Fig. 4.11 the most conserved sites in our design scheme include those found to be crucial in the folding process. Again, these striking results serve a two-fold purpose. On one hand they confirm the validity of the present design approach; on the other they also show some of its possible applications, in connection with the prediction of folding nucleus.

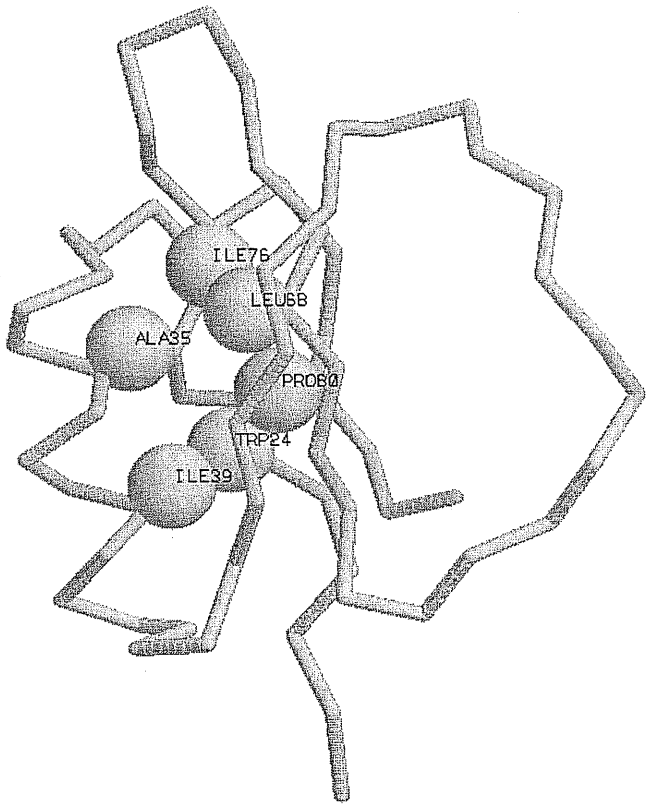


Figure 4.11: Backbone for the CI2 with the 6 most conserved residues in our design attempts. Three of them (ALA35, ILE76, LEU68) are indicated by Itzhaki *et al.* [28] as the most important in the folding process.

4.4 Summary

To summarize, we carried out automated protein design attempts over some PDB conformations by introducing several novel strategies to identify optimal energy-cost functions and select putative design solutions. A mere comparison of designed sequences with the PDB ones gives a success rate between 40% and 55% when working with three classes of amino acids: a value well above the random-guessing threshold. This success rate is not improving by introducing more sophisticated energy functions, suggesting that other important features of real proteins are neglected by short range Hamiltonians. Nevertheless, a statistical analysis of a wider set (non-extremal) of possible solutions, shows how the design procedure could be used to correctly predict, with a high confidence, at least a sub-set of protein sites. These residues can be related to the conserved sites obtained by a statistical analysis of naturally occurring homologous sequences. Moreover, for two specific proteins (barnase and chymotrypsin inhibitor), these highly predictable sites correspond with a very good precision, to the folding nucleus, which is crucial for the folding process.

Conclusions

In chapter 3 we have shown the iterative method, which is a powerful tool for protein design at least for lattice models. This method allows to select conformations, that are highly competitive, with respect to the target conformation, in being the ground state of bad designed sequences. It is quite different from the more natural strategy of nested Monte Carlo, where a stochastic search in conformation space for estimating correctly the conformational free energy is nested to a stochastic exploration of sequences space. In fact, the iterative method is based on a preliminary iterative procedure for collecting good decoys and a subsequent sequence selection procedure. Once the first step is performed, the second is a very efficient procedure.

Though the iterative method is powerful on lattice models, on real proteins is not yet applicable, at least in the form discussed in this thesis. Folding algorithms on real proteins are too expensive, in terms of CPU time, and do not allow an efficient implementation of the iterative procedure. Furthermore, in real proteins interactions are not so simple and well-defined as those used in lattice models. Further tests on model proteins are necessary, in order to understand how to implement the iterative strategy on proteins.

Even if the iterative approach is to complex to be applied on real systems and to be compared with experimental data, other methods are becoming accessible. In chapter 4 we have shown a design procedure that uses previous results obtained on lattice models. In particular we use a scoring function similar to the one introduced by Deutsch and Kuroski (see eq. 2.16). Difficulties met in the evaluation of the average contact map, due basically to the different length of protein structures, are described in appendix A. The other more crucial problem, the selection of optimal potentials, has been

solved by using a partially novel technique that has been improved in a self-consistent way in the design procedure. The optimization procedure exploits experimental data retrieved from the PDB and is based on energetic considerations, that have been already used, though in other contexts like protein folding and with a different formalism.

Once the scoring function has been optimized by tuning the free parameters on real proteins, the design procedure can be performed. A rigorous evaluation of the performances of our procedure can be done through an accurate comparison between designed sequences and homologous natural sequences. Designed sequences are similar to natural sequences up to 60% and on average $50 - 55\%$. This does not guarantee, of course, that our solutions are correct; however the result shows that the approximate scoring function used is, to some extent, correct.

Such results are particularly surprising if one makes the following considerations.

1. The method is completely automated and no extra-information on the protein is put by hand in the design procedure.

2. The mathematical description of interactions is extremely simplified if compared to the one used by standard programs of molecular dynamics.

3. Both the training set and the test set are quite heterogeneous: proteins belonging to such sets have in general completely different structures and different functions.

4. Similarity between designed and natural sequences could be improved by putting some phenomenological constraints like propensity in secondary structures.

The last two points need a further explanation. Regarding point 3., it is possible that the design procedure could be improved by considering proteins sharing a similar structure or, at least, a similar fold. For example, the content of α-helices and β-sheets is completely variable both in the training set and test set. At least in principle, the performances of our method might improve by restricting the sets to only α proteins.

Another possibility to improve our results is introducing in the Hamiltonian some phenomenological terms. The optimization procedure, since it is knowledge-based, can be successfully applied independently of the nature of the terms in the Hamiltonian. Furthermore, since it is self-consistent, the new optimized Hamiltonian can be immediately used to design structures in the test set. Preliminary studies are proceeding along these direction.

Appendix A

Determination of the weight functions $\Delta^{(2)}$

A.1 Two-body energy

We estimated the average contact maps $\langle \Delta^{(2)}_{ij} \rangle$ and $\langle \Delta^{(3)}_{ijk} \rangle$ by considering as a set of possible competing configurations an ensemble of structures extracted form the PDB. We analyzed $N = 116$ proteins (with length ranging from 36 to 296) and for each conformation, Γ_n, we computed the corresponding value of the contact matrix $\Delta^{(2)}_{ij}(\Gamma_n)$. If the structures had the same length, let's say L, $\langle \Delta^{(2)}_{ij} \rangle$ could be estimated by simple averaging:

$$\langle \Delta^{(2)}_{ij} \rangle = \frac{1}{N} \sum_{n=1}^{N} \Delta^{(2)}_{ij}(\Gamma_n). \tag{A.1}$$

where $i, j < L$. However, since we are working with proteins of different length, we can expect a dependence of $\langle \Delta^{(2)}_{ij}(\Gamma_n) \rangle$ on the length of the chains. To investigate this possibility we first notice that $\langle \Delta^{(2)}_{ij}(\Gamma_n) \rangle$ mainly depends on the sequence separation $k = |j - i|$ (at least for small k) between the amino-acids along the chain more than from the position along the chain and from the length of the protein (see Fig. A.1)

Let us compute now the average $\langle \Delta^{(2)}_{k} \rangle$ value of this contact frequencies

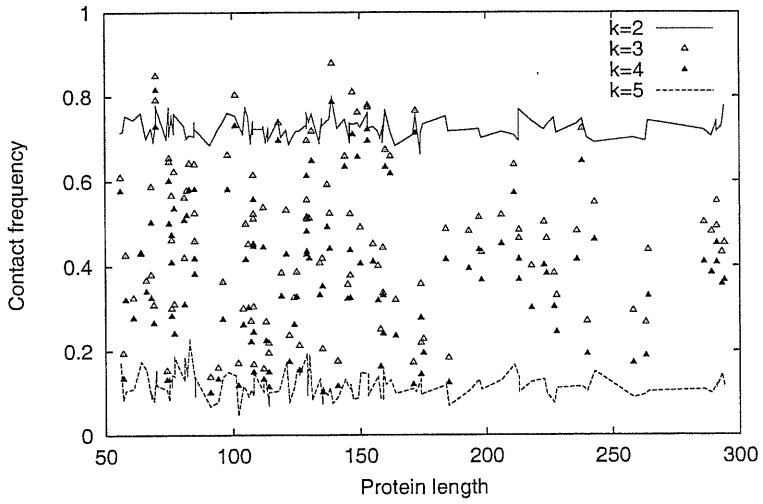


Figure A.1: Contact frequency for different values of the amino acid separation k as a function of the length the protein. For $k = 3$ and $k = 4$ the fluctuations are large and depend on the protein family (α or β)considered α-protein or β-protein). For all the k's there is no significative dependence on the protein length.

according to :

$$\langle \Delta_k^{(2)} \rangle = \frac{1}{N} \sum_{n=1}^{N} \langle \Delta_{i,j}^{(2)}(\Gamma_n) \rangle_{i-j=k} \ , \tag{A.2}$$

where $\langle \cdots \rangle_{i-j=k}$ represents the arithmetic average over all the contacts with a given sequence separation k for a given protein. Then, we notice that it is a rapidly decaying function of the chemical distance, k (see Fig. 4.2).

We can then estimate $\langle \Delta_{ij}^{(2)} \rangle$ according to the rules:

$$\langle \Delta_{ij}^{(2)} \rangle = \begin{cases} \langle \Delta_k^{(2)} \rangle & k < k_0 \\ \Delta_0^{(2)} & k \geq k_0 \ . \end{cases} \tag{A.3}$$

where k_0 is a cut-off distance that we fixed equal to 16. The value $\langle \Delta_k^{(2)} \rangle$ can be estimated numerically from the data bank through eq. A.2 whereas $\Delta_0^{(2)}$ should be determined according to the length of the chain.

Indeed the dependence of the total number of contacts, $\sum_{i<j}\Delta^{(2)}_{ij}(\Gamma_n)$, is well approximated by a linear function of the length –or number of amino acids–, L_n, of Γ_n. Thus, using this linear dependence on L_n and Eq. (A.3) we are able to determine $\Delta^{(2)}_0$.

A.2 Three body energy

The average contact map $\langle\Delta^{(3)}_{ijk}\rangle$ can be determined in an analogous way.

For a conformation Γ we define the total number of 3-body contacts as

$$N^3_c(\Gamma) = \sum_{i<j<k}\Delta^{(3)}_{ijk}(\Gamma). \tag{A.4}$$

Similarly to the former case this number of contacts can be fitted by a linear relation.

In this larger parameter space $\langle\Delta^{(3)}_{ijk}\rangle$ will depend on two indexes $k1 = |j-i|$ and $k2 = |k-j|$:

$$\langle\Delta^{(3)}_{ijk}\rangle = \Delta^{(3)}(k1,k2)\ . \tag{A.5}$$

For $k1, k2 < k_0$ (that we choose on the basis of the statistical analysis to be $k_0 = 6$)

$$\Delta^{(3)}(k1,k2) = \frac{1}{N}\sum_{n=1}^{N}\langle\Delta^{(3)}_{ijk}(\Gamma_n)\rangle_{j-k=k_2,\ i-j=k_1} \tag{A.6}$$

while for $k_1 \geq k_0$ or $k_2 \geq k_0$ we assume a constant value. Here, $\langle\cdots\rangle_{j-k=k_2,\ i-j=k_1}$ represents the arithmetic average over all the contacts with given sequence separation k_1, k_2.

The average contact map for a generic protein will be

$$\langle\Delta^{(3)}_{ijk}\rangle = \begin{cases} \Delta^{(3)}(k1,k2) & k1,k2 < k_0 \\ \Delta^{(3)}_0 & \text{otherwise}\ . \end{cases} \tag{A.7}$$

Using, again, that $\sum_{i<j<k}\langle\Delta^{(3)}_{ijk}\rangle$ is well interpolated by a linear function of L_n we can determine $\Delta^{(3)}_0$ in Eq. (A.7), after $\Delta^{(3)}(k_1,k_2)$ for $k_1, k_2 < k_0$ have been evaluated.

Appendix B

Perceptron learning of the optimal potentials

A convenient way to find the optimal potentials that satisfy inequality constraints such those of Eq. 4.6, is the use of the perceptron algorithm for the optimization of a set of linear inequalities[32, 80].

For instance, in the case of the two body Hamiltonian, Eq. (4.6) can be written, using the result of Eq. (4.7), as:

$$\sum_{i>j=1}^{L} (\langle \Delta_{ij}^{(2)} \rangle - \Delta_{ij}^{(2)}(\Gamma)) B_2(s_i, s_j) > 0 \qquad \text{(B.1)}$$

where L is the length of the protein. If $n_{kl}(\Gamma)$ denotes the number of contacts in the conformation Γ involving amino acids of types k and l, and $\langle n_{kl}^{(2)} \rangle$ the corresponding average computed on the set of competing configurations by using Eq.A.3, Eq. B.1 can be rewritten as:

$$\sum_{k>l=1}^{3} (\langle n_{kl}^{(2)} \rangle - n_{kl}(\Gamma)) B_2(k,l) = \sum_{k>l=1}^{3} a_{kl}(\Gamma) B_2(k,l) = \mathcal{F}_\Gamma(\vec{B}) \qquad \text{(B.2)}$$

where the vector $\vec{B}_2$ is defined as:

$$\vec{B} \equiv (B(1,1), B(1,2), B(1,C), B(2,2), B(2,3), B(3,3)) \qquad \text{(B.3)}$$

Given the native state Γ and the sequence s the six entries of a_{kl} depends only on the average properties of the decoy structures.

For a given set of M inequalities to be satisfied simultaneously, it is convenient to identify the one (related to the conformation Γ_s) that, with a given set of trial potentials is the worst satisfied one , e.g.:

$$\mathcal{F}_{\Gamma_s}(\vec{B}) < \mathcal{F}_k(\vec{B}) \quad k = 1, \ldots, M \quad k \neq s \tag{B.4}$$

Once Γ_s has been determined, one updates the trial potentials adding a quantity proportional to $a_{kl}(\Gamma_s)$ where the proportionality constant is chosen to be much smaller than one. With this new choice of the potentials, each inequality is re-valuated and the updating cycle is repeated until $\mathcal{F}_{\Gamma_s}(\vec{B})$ (stability) reaches the maximum possible value. One is allowed to fix the scale of B's by requiring $|\vec{B}| = 1$ where the $|\cdot|$ is the usual Euclidean norm. This method can be shown to converge to an optimal solution, $\mathcal{F}^*$, which can be of either sign. If it is negative , it means that no set of potentials can be found that consistently satisfied all inequalities in the set. Otherwise the problem is learnable and the optimal potentials are identified with those giving the highest stability.

We have extracted potentials by using the perceptron scheme with $M = 31$ globular proteins. The related set of inequalities has turned out to be learnable in all cases, with two or three body energy terms.

For the two body energy we have extracted a first set of potentials using the 15 proteins and a second one with the remaining 16. The two set of potentials are plotted one versus the other in Fig. B showing an extremely good correlations.

This validate the conclusion that an interaction matrix B depending only on 6 parameters can be determined with a dozen of non redundant globular proteins. Similar results have been obtained with the three body energy.

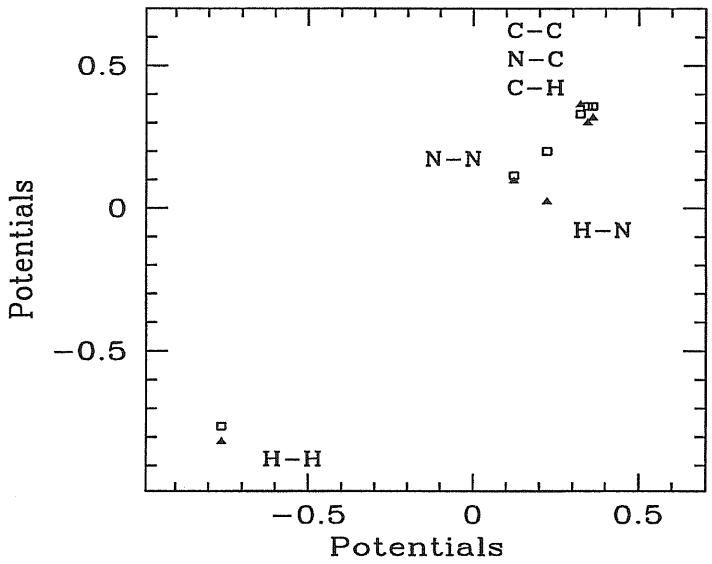


Figure B.1: The potentials $\vec{B}$ determined using a set of 15 proteins and another set of 16 proteins (see table 4.2) are plotted versus the same potentials determined by the whole set of 31 proteins. The correlation between the potentials obtained with the two sets and the biggest one is nearly perfect (ideally points should lie on the diagonal). Using the whole set of table (4.2) we found $\vec{B} = (0.12, 0.22, 0.36, -0.76, 0.35, 0.32)$. Potentials are here sorted as in eq. (B.3) where 1,2,3 refer respectively to classes P,H,C.

Bibliography

[1] M. Altamirano, M. Blackburn, C. Aguayo, and A. Fersht. Directed evolution of new catalytic activity using the α/β-barrel scaffol. *Nature*, 403:617–622, 2000.

[2] C. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–239, 1973.

[3] J. Bowie, R. Lüthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170, 1991.

[4] C. Branden and J. Tooze. *Introduction to protein structure.* Garland Publishing, New York, 1991.

[5] J. Bryngelson, J. Onuchic, N. Socci, and W. P.G. Funnels, pathways and the energy landscape of protein folding: A synthesis. *Proteins*, 21:167–195, 1995.

[6] J. Bryngelson and P. Wolynes. Spin glasses and the statistical of protein folding. *Proc. Natl. Acad. Sci. USA*, 84:7524–7528, 1987.

[7] F. Cecconi, C. Micheletti, P. Carloni, and A. Maritan. The structural basis of antiviral drug resistence. *SISSA preprint*, 2000.

[8] H. Chan and K. Dill. Sequence space soup of proteins and copolymers. *J. Chem. Phys.*, 95:3775, 1991.

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

[10] C. Clothia and A. M. Lesk. The relation between the divergence of sequences and structures in proteins. *EMBO J.*, 5:823–826, 1986.

[11] T. Creighton. *Proteins, structure and molecular properties.* W.H.Freeman and Company, New York, second edition, 1993.

[12] G. Crippen. Prediction of protein folding from amino acid sequence over discrete conformation space. *Biochemistry*, 30:4232–4237, 1991.

[13] B. Dahiyat and S. Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–87, 1997.

[14] J. Deutsch and T. Kurosky. New algorithm for protein design. *Phys. Rev. Lett.*, 76:323–326, 1996.

[15] K. A. Dill, S. Bromberg, K. Yue, K. Fiebig, D. Yee, and P. Thomas. Principles of protein folding - a perspective from simple exact models. *Protein Science*, 4:561–602, 1995.

[16] R. Dima, G. Settanni, C. Micheletti, J. Banavar, and M. A. Extraction of interaction potentials between amino acids from native protein structures. *J. Chem. Phys.*, 112:9151–9166, 2000.

[17] Y. Duan and P. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in water solution. *Science*, 282:740–744, 1998.

[18] A. Fersht. *Structure and mechanism in proteinscience: a guide to enzyme catalysis and protein folding.* W.H. Freeman, New York, 1999.

[19] A. R. Fersht. Optimization of rates of protein folding - the nucleation condensation mechanism and its implications. *Proc. Natl. Acad. Sci. USA*, 92:10869–10873, 1995.

[20] N. Go and H. A. Scheraga. *Macromolecules*, 9:535–542, 1976.

[21] P. Harbury, J. Plecs, B. Tidor, T. Alber, and P. Kim. High-resolution protein design with backbone freedom. *Science*, 282:1463–1467, 1998.

[22] H. Hellinga and F. Richards. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. USA*, 91:5803–5807, 1993.

[23] H.M.Berman, J.Westbrook, Z.Feng, G.Gililland, T.N.Bhat, H.Wessig, I.N.Shindyalov, and P.E.Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[24] T. Hoang and M. Cieplak. Molecular dynamics of folding of secondary structures in go-like models of proteins. cond-mat/9911488.

[25] E. Huang, P. Koehl, M. Levitt, R. Pappu, and J. Ponder. Accuracy of side-chain prediction upon near-native protein backbones generated by ab initio folding methods. *Proteins: Struct. Funct. and Gen.*, 33:204–207, 1998.

[26] A. Irback, C. Peterson, F. Potthast, and E. Sandelin. Monte-carlo procedure for protein design. *Phys. Rev. E*, 58:R5249–R5252, 1998.

[27] A. Irback, C. Peterson, F. Potthast, and E. Sandelin. Design of sequences with good folding properties in coarse-grained protein models. *Structure with folding and design*, 7:347–360, 1999.

[28] L. S. Itzhaki, D. E. Otzen, and A. R. Fersht. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engeeniring methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.*, 254:260–288, 1995.

[29] W. Kabsch and C. Sander. Dictionary of protein secondary structure:pattern recognition of hydrogen-bonded and geometricl features. *Biopolymers*, 22:2577–2637, 1983.

[30] S. Kamtekar, J. Schiffer, H. Xiong, J. Babik, and M. Hecht. Protein design by binary patterning of polar and nonpolar amino acids. *Science*, 262:1680–1685, 1993.

[31] D. E. Kim, G. Hongdi, and D. Baker. The sequence of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl. Acad. Sci. USA*, 95:4982, 1998.

[32] W. Krauth and M. Mezard. Learning algorithms with optimal stability in neural networks. *J. Phys. A*, 20:L745–L752, 1987.

[33] K. F. Lau and K. A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22:3986–3997, 1989.

[34] C. Levinthal. Are there pathways for protein folding? *J. Chem. Phys.*, 65:44–45, 1968.

[35] H. Li, R. Helling, C. Tang, and N. S. Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273:666, 1996.

[36] H. Li, C. Tang, and N. S. Wingreen. Are protein folds atypical? *Phys. Rev. Lett.*, 95:4987–4990, 1998.

[37] V. N. Maiorov and G. M. Crippen. Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.*, 227:876–888, 1992.

[38] V. N. Maiorov and G. M. Crippen. Learning about protein folding via potential functions. *Proteins: Structure Function and Genetics*, 20:173–176, 1994.

[39] A. Maritan, C. Micheletti, and J. Banavar. Role of secondary motifs in fast folding polymers: a dynamical variational principle. *Phys. Rev. Lett.*, 2000.

[40] C. Micheletti, J. Banavar, A. Maritan, and F. Seno. Protein structures and optimal folding from a geometrical variational principle. *Phys. Rev. Lett.*, 82:3372–3375, 1999.

[41] C. Micheletti, J. Banavar, F. Seno, and A. Maritan. Steric constraints in model proteins. *Phys. Rev. Lett.*, 80:5683, 1998.

[42] C. Micheletti, A. Maritan, and J. R. Banavar. A comparative study of existing and new design techniques for protein models. *J. Chem. Phys.*, 110:9730–9738, 1999.

[43] C. Micheletti, F. Seno, A. Maritan, and J. Banavar. Design of proteins with hydrophobic and polar amino acids. *Proteins: Structure Function and Genetics*, 32:80, 1998.

[44] C. Micheletti, F. Seno, A. Maritan, and J. Banavar. Protein design in a lattice model of hydrophobic and polar amino acids. *Phys. Rev. Lett.*, 80:2237, 1998.

[45] C. Micheletti, F. Seno, A. Maritan, and J. R. Banavar. Strategies for protein folding and design. *Annals of Combinatorics*, 3:439–458, 1999.

[46] L. Mirny and E. Shakhnovich. How to derive a protein folding potential? a new approach to an old problem. *J. Mol. Biol.*, 264:1164–1179, 1996.

[47] S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal-structures - quasi-chemical approximation. *Macromolecules*, 18:534–552, 1985.

[48] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, 256:623–644, 1996.

[49] G. Morra and A. Rossi. Geometrical caracterization of conformations which compete with protein native states. Submitted to Phys. Rev. E.

[50] M. Morrisey and E. Shakhnovich. Design of proteins with selected thermal properties. *Folding and design*, 1:391–405, 1996.

[51] J. N. Onuchic, P. G. Wolynes, L.-S. Z., and N. D. Socci. Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl. Acad. Sci. USA*, 92:3626–3630, 1995.

[52] C. Pabo. Designing proteins and peptides. *Nature*, 301:200, 1983.

[53] V. Pande, A. Grosberg, and T. Tanaka. Heteropolymer freezing and design: Towards physical models of protein folding. *Rev. Mod. Phys.*, 72:259–314, 2000.

[54] B. Park and M. Levitt. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.*, 258:367–392, 1996.

[55] J. Ponder and F. Richards. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, 193:775–791, 1987.

[56] T. Quinn, N.B.Tweedy, R. Williams, J. Richardson, and D. Richardson. De-novo design, synthesis and characterization of a beta sandwich protein. *Proc. Natl. Acad. Sci. USA*, 91:8747–8751, 1994.

[57] G. N. Ramachandran and V. Sasisekharan. Conformation of polypeptides and proteins. *Adv. Protein Chem.*, 23:283–437, 1968.

[58] A. Rossi, C. Micheletti, and A. Maritan. A novel iterative strategy for protein design. *J. Chem. Phys.*, 112:2050–2055, 2000.

[59] A. Rossi, C. Micheletti, F. Seno, and A. Maritan. Self-consistent knowledge based approach to protein design. Submitted to Biophys. J.

[60] A. Sali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369:248–251, 1994.

[61] A. Sali, E. Shakhnovich, and M. Karplus. Kinetics of protein folding: a lattice model study for the requirements for folding to the native state. *J. Mol. Biol.*, 235:1614–1636, 1994.

[62] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68, 1991.

[63] F. Seno, A. Maritan, and J. Banavar. Interaction potentials for protein folding. *Proteins: Structure Function and Genetics*, 30:224–248, 1998.

[64] F. Seno, C. Micheletti, A. Maritan, and J. Banavar. Variational approach to protein design and extraction of interaction potentials. *Phys. Rev. Lett.*, 81:2172, 1998.

[65] F. Seno, M. Vendruscolo, J. Banavar, and A. Maritan. Optimal protein design procedure. *Phys. Rev. Lett.*, 77:1901–1904, 1996.

[66] L. Serrano, J. T. Kellis, P. Cann, A. Matousheck, and A. R. Fersht. Substructure of barnase and the contribution of different interactions to protein stability. *J. Mol. Biol.*, 224:783–804, 1992.

[67] E. Shakhnovic and A. Gutin. Engineering of stable, and fast folding sequences of models proteins. *Proc. Natl. Acad. Sci. USA*, 90:7195–7199, 1994.

[68] E. Shakhnovich, V. Abkevich, and O. Ptitsyn. Conserved residues and the mechanism of protein folding. *Nature*, 379:96–98, 1996.

[69] E. Shakhnovich and A. Gutin. A new approach to the design of stable proteins. *Protein Engineering*, 6:793–800, 1993.

[70] E. I. Shakhnovich. Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett.*, 72:3907–3910, 1994.

[71] M. Sippl. Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213:859–883, 1990.

[72] M. Sippl. Knowlwdge based potentials for proteins. *Curr. Opin. Struct. Biol.*, 5:229–235, 1995.

[73] M. Skorobogatiy, H. Guo, and M. J. Zuckermann. A deterministic approach to the protein design problem. *Macromolecules*, 30:3403–3410, 1997.

[74] A. Street, D. Datta, D. Gordon, and L. S. Mayo. Computational protein design. *Structure with folding and design*, 7:R105–109, 1999.

[75] A. G. Street and L. S. Mayo. Computational protein design. *Structure with folding and design*, 7:R105–109, 1999.

[76] S. Sun. Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.*, 2:762–785, 1993.

[77] S. Sun, R. Brem, R. Chan, and K. Dill. Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng.*, 8:1205–1213, 1995.

[78] C. Tang. Simple models of the protein folding problem. *cond-mat/9912450*, 1999.

[79] P. Thomas and K. Dill. An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. USA*, 93:11628–11633, 1996.

[80] J. van Mourik, C. Clementi, A. Maritan, F. Seno, and J. Banavar. Determination of interaction potentials of amino acids from native protein structures: test on simple lattice models. *J. Chem. Phys.*, 110:10123–10133, 1999.

[81] M. Vendruscolo and E. Domany. Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.*, 109:11101–11108, 1999.

[82] A. Wallqvist and M. Ullner. A simplified amino acid potential for use in structure prediction of proteins. *Proteins: Structure Function and Genetics*, 18:267–280, 1994.

[83] M. W. West, W. Wang, J. Patterson, J. D. Mancias, J. R. Beasley, and M. H. Hecht. *De novo* amyloid proteins from designed combinatorial libraries. *Proc. Natl. Acad. Sci. USA*, 96:11211–11216, 1999.

[84] K. Yue and K. Dill. Sequence-structure relationship in proteins and copolymers. *Phys. Rev. E*, 48:2267–2279, 1993.

[85] K. Yue and K. A. Dill. Forces of tertiary structural organization in globular proteins. *Proc. Natl. Acad. Sci. USA*, 92:146–150, 1995.

[86] K. Yue, K. Fiebig, P. Thomas, H. Chan, E.I.Shackhnovich, and K. Dill. A test of lattice protein-folding algorithms. *Proc. Natl. Acad. Sci. USA*, 92:325–329, 1995.

[87] J. Zou and J. G. Saven. Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. *J. Mol. Biol.*, 296:281–294, 2000.

Aknowledgemnts

First of all, I would like to thank Cristian, who had a fundamental role in my education during these three years; he always gave me precious suggestions.

A fundamental role, of course, had Amos, whose style to look at science, I will always remember.

During my Ph.D., I have had the pleasure to collabote with Flavio Seno, whose wrong believes in football did not obstacolate friendly discussions, and with Giulia Morra.

I am grateful to my room mates, Antonio, Gianni and Gianluca for cultural and scientific exchanges and for having borne me, so long time.

A particular thank to Fabio and Alessandro, whose showman attitude will remain always in my heart.

I am particularly grateful to the students of the Condensed Matter Sector for having helped me in bad moments, at the end of the first year.

A particular thank is reserved to the Abdus Salam International Center for the kind hospitality in the last two years.

Finally, I would like to thank my parents for their patient encouragement and Yannet for having changed my life.