Computational Studies on Energetics of Protein Fold

Distinctive Roles of Solvent Effect and Side-chain Packing

Akira KINJO
Doctor of Philosophy

Department of Genetics
School of Life Science
The Graduate University for Advanced Studies

2000

To Rinsho Kadekaru and my parents

Abstract

Proteins not only fold and unfold, but they are manifolds: they carry information about the evolutionary history of life, they function in the living cell, and they are molecules. In this thesis, I mostly treat the last aspect of the protein structure although the first aspect may be also relevant. A protein is a molecule consisting of thousands of atoms which are arranged in a specific configuration in space under the physiological condition. The problem of how proteins adopt their specific three-dimensional conformation should be addressed in terms of physics and chemistry. On the other hand, a number of proteins had diverged from a common ancestral protein, resulting in their similar three-dimensional structures. In other words, different proteins consisting of different sequences of amino acid residues may adopt similar three-dimensional conformations. As more and more structures of proteins became available, it has been recognized that a set of proteins with similar three-dimensional structures can be grouped into one category which is now termed as "fold." From the definition, fold is conceptual and is not a solid physical object. Even if their three-dimensional structures are similar and they share a common fold, different proteins are different. This fact poses a difficult problem since different proteins compose different physical systems which are not directly comparable. By what physical principles is a protein *fold* stabilized? I attempt to answer this question by computational means, namely by classifying structural properties and associating them with corresponding energy terms.

The first chapter is a general introduction to the problem addressed in this study. I summarize the physics of the molecular system and some known structural aspects of proteins relevant to this study. Some definitions of terminology are also presented. Emphases are put on the solvent effect and side-chain packing which are later shown important for interpreting the stabilization mechanism of the native structure and native fold.

Chapter 2 describes the methods used in the present studies such as molecular mechanics, implicit solvent models, continuum electrostatics and the "near-native" structure. The near-native structure of a protein is a structure with almost the same backbone conformation as the native structure, but with different side-chain conformations. Since the backbone conformation is the basis of the definition of fold, the near-native structure is the most crucial idea in the present study, which made possible to extract the energy components important for stabilizing native fold.

Chapter 3 is the main body of the thesis in which various energy components of the native, near-native, and intentionally misfolded structures are compared for arbitrarily selected seven proteins of various structural classes. It is shown that the solvent effects such as electrostatic shielding, the Born energy, and hydration are important for the stabilization of the near-native structures rather than the native structures themselves. The native structures are stabilized to a great extent by the packing energy. The stabilization of the native structure by the packing energy is shown to be far larger than the stabilization of the near-native structure by the solvent effect. Since the near-native structure has the same fold as the native structure, it is suggested that the stabilization of proteins with the same fold is attained by the solvent effect. Implications of these results are discussed with respect to protein structure prediction.

Homologous proteins that are found in thermophilic and mesophilic organisms show similar three-dimensional structures but thermophilic proteins are far more stable than their mesophilic homologs. In Chapter 4, I have applied the methodology developed in Chapters 2 and 3 to the investigation of the stabilization mechanism of thermophilic proteins relative to that of mesophilic ones. In the current structural database, five families of thermophilic proteins and their mesophilic homologs that were of high resolution crystal structures and of moderate sizes were available. The energies of the native (n), near-native (m) and unfolded (u) structures were calculated for the five thermophilic proteins as well as for their mesophilic homologs. The energy difference between the native and unfolded structures (ΔE_{n-u}) was decomposed into the ones between the near-native and unfolded structures (ΔE_{m-u}), and between the native and near-native structures (ΔE_{n-m}). It was found that the sum of electrostatic and hydration energies was consistently lower for the thermophilic proteins than for the mesophilic homologs. This trend was observed not only in ΔE_{n-u}, but also in ΔE_{m-u}. No conspicuous tendency was found in ΔE_{n-m}. This result indicates that, relative to their mesophilic homologs, the thermophilic proteins are stabilized by the solvent effect that determines the approximate native fold rather than the side-chain packing that determines the precise native structure itself. A consequence of this conclusion is discussed.

Finally, the results in the thesis are summarized in Chapter 5. The use of the near-native structure and other related problems are also discussed.

Acknowledgements

First of all, I thank Prof. Ken Nishikawa for suggesting the main themes of my thesis, for fruitful and stimulating discussions and advice. Dr. Akinori Kidera gave me a lot of advice regarding molecular mechanics calculations, and his critical comments helped the work developed in Chapter 3. Prof. Haruki Nakamura kindly taught me electrostatic calculations and provided me his computer program. I am also grateful to Drs. Takeshi Kawabata and Motonori Ota for extensive discussions on many subjects. Especially, attending the CASP3 meeting with them in 1998 was a great opportunity to expose myself to the current state of protein structure prediction. Dr. Satoshi Fukuchi gave me some information about thermophilic proteins. Dr. Yasuo Shirakihara gave me advice on how to validate the quality of experimental structures. I thank Profs. Yoshio Tateno and Hideaki Sugawara for comments on my progress reports. Most of the computations were carried out on the VPP500 supercomputer at National Institute of Genetics. I am grateful to the system engineers who maintained the computer, especially Yuichi Kawanishi who also helped me with programming. Tomoko Narita helped me with her secretaryship. Playing the sanshin once a week with my "disciples", Keiko Ichikawa, Yukiko Daito, Sachiko Nagira, and Yumiko Sato was relaxing and refreshing. Gisela Hoschek cultivated my interests to biological sciences some ten years ago. The lectures on statistical mechanics by Prof. Nobuhiro Go enlightened me on theoretical studies of protein structure. Finally, I am grateful to Hiroe Hentona, Emma, Lazlo, Chiho Watanabe, Miyoko Kishimoto, Sukrishna Ishii, Donya and the musicians at *Café de Schiele* for encouraging and/or feeding me, and the Japan Society for the Promotion of Science and the tax payers in Japan for financial support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Structure and Forces: Introduction

1.1 The fundamentals

Protein is the object treated in this thesis. By "protein", I mean a globular (soluble) protein in aqueous solution. Not to mention the famous remark by Feynman [25] "all things are made of atoms." This *atomic hypothesis* is especially true for proteins which are, in spite of the large size, microscopic objects, that is, molecules. In his presentation of the atomic hypothesis, Feynman described atoms as "little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another" [25]. This qualitative behavior of atoms should be quantitatively described by the non-relativistic Schödinger equation [101] of the system of interest:

$$\hat{H}\Phi = E\Phi \tag{1.1}$$

where $\hat{H}$,Φ, E are the Hamiltonian operator, wave function, and energy of the system. In our case, the system consists of a protein molecule surrounded by a large number of water molecules. Since atoms are made of nuclei and

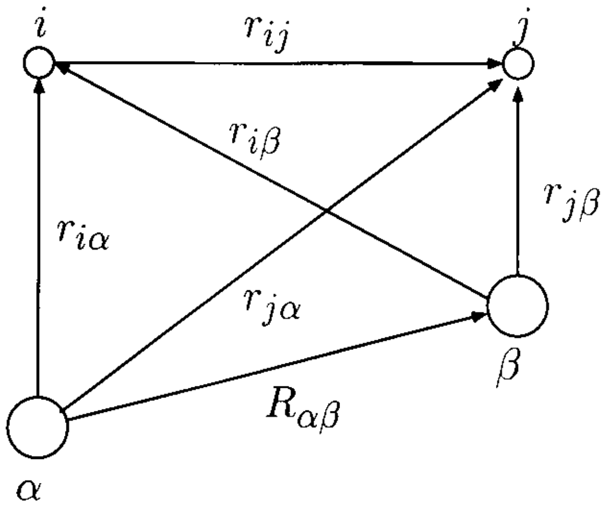


Figure 1.1: A molecular coordinate system: i, j = electrons; α, β = nuclei.

electrons, the Hamiltonian $\hat{H}$ in atomic unit is given as:

$$\begin{aligned} \hat{H} &= -\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2 - \sum_{\alpha=1}^{M}\frac{1}{2M_\alpha}\nabla_\alpha^2 - \sum_{i=1}^{N}\sum_{\alpha=1}^{M}\frac{Z_\alpha}{r_{i\alpha}} \\ &\quad + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{\alpha=1}^{M}\sum_{\beta>\alpha}^{M}\frac{Z_\alpha Z_\beta}{R_{\alpha\beta}} \end{aligned} \tag{1.2}$$

where M_α and Z_α are the mass and charge of the nucleus α relative to those of the electron. $r_{i\alpha}$, r_{ij} and $R_{\alpha\beta}$ are the relative orientations of nuclei and electrons as depicted in Figure 1.1.

In principle, it is possible to investigate the energetics of protein structure based on Equations 1.1 and 1.2. In reality, of course, it is impossible because of the large number of freedom involved in the Hamiltonian $\hat{H}$ (Equation 1.2) even with the advances in modern quantum chemistry [101, 74]. Furthermore, even if the whole Schödinger equation could be solved, it is doubtful whether we can claim that we have understood the principle of protein structure. The system we are concerned with is very large and complicated. This in turn leads to the solution to the Schödinger equation of the protein-solvent system too complex for us to interpret. Therefore, some approximations are not only necessary but indeed essential for our understanding.

In this thesis, various levels of approximations are applied. In fact, the main theme of the thesis is focused on a way to approximate the protein structure, namely the concept of protein *fold*, which is discussed in the following sections. Another approximation involved is the use of the molecular mechanics method. In stead of solving Equation 1.1, a classical Hamiltonian is defined and used for evaluation of energy of protein molecules. Molecular mechanics and related methods used in the present study are briefly reviewed in Chapter 2. The energy of a system implies the structure of the system. But their relation is not trivial, especially for such a complex system as the protein-solvent system. Hence another level of approximation is introduced to link the energetics with structures. A long history of studies on protein structure has revealed the importance of two phenomena: the solvent effect [41, 20] and tight side-chain packing [84, 85]. I will pay special attention to these phenomena. These are briefly described later in this chapter, and are the main subject in Chapter 3.

1.2 Some definitions

The main title of this thesis is "Computational studies on energetics of protein *fold*". It is not "Computational studies on energetics of protein *structure*". The phases "protein fold" and "protein structure" are often used interchangeably. If I had followed the common practice, the latter phase could have been in the title. In this thesis, I give distinctive meanings to the terms protein *structure* and protein *fold* in order to more clearly define the problem dealt in the present study. Therefore, it should be useful to give definitions to some terms which may otherwise be confusing.

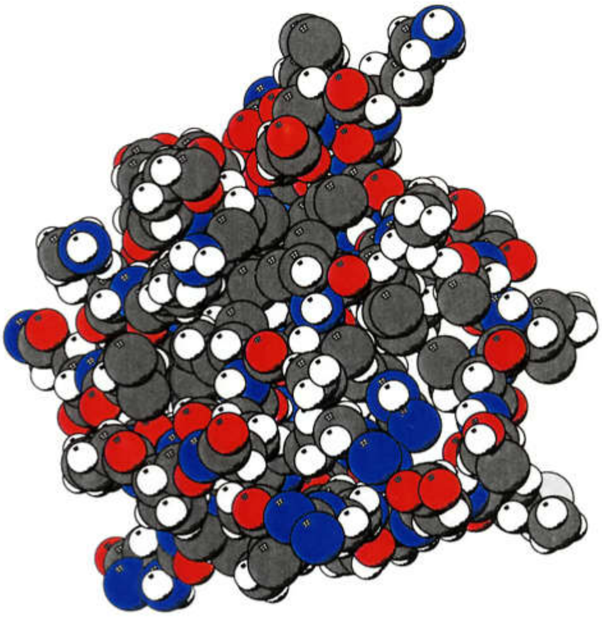


Figure 1.2: The native *structure* of a protein is defined by both backbone and side-chain atoms. The example is the PDB [8] entry 2ci2, barley chymotrypsin inhibitor 2 [57]. The figure was drawn with MolScript [44].

Structure of a protein Let a protein consist of N atoms (atom 1, atom 2, $\cdots$, atom N). Then a structure of the protein is defined by a particular set of the coordinates of *all* the atoms in the protein: $\{\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_N\}$. The phrase "protein structure" is often used to refer to "native structure".

Native state (of a protein) The native state of a protein is a set of structures of the protein under the physiological condition.

Native structure The native structure of a protein is one particular structure which represents the native state of the protein. Whenever we use the term "native structure", the structure is defined by the atomic coordinates of both backbone atoms and side-chain atoms. An illustrative example of a native structure is shown in Figure 1.2.

Fold Since only some 100 experimentally determined protein structures were available, it was already known that "the chain fold is an efficient criterion for protein classification" [92]. Today, the number of known protein structures are enormous and a number of databases exist that provide the classification of proteins based on the folds. According to the structural classification of proteins (SCOP) database [62], the term "fold" is defined as follows:

> **Fold**: *Major structural similarity*
> Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.

One major characteristics of the concept of fold in this definition is that it is a relative one, rather than an absolute one. In other words, the fold of a protein cannot be defined unless other proteins of similar structures are given. Another characteristics is that detailed side-chain conformations and intricate loop structures are not taken into account. The term fold merely refers to the global feature of the backbone conformation.

However, the classification of protein structures has advanced so far that nowadays the concept of fold appears independent of the classification, that

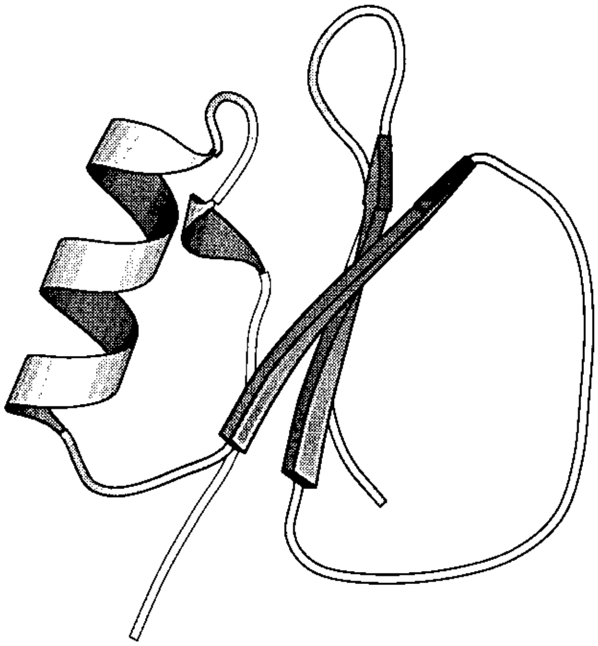


Figure 1.3: The native *fold* of a protein is defined by the backbone topology irrespective of side-chain conformations. The example is again the PDB [8] entry 2ci2, barley chymotrypsin inhibitor 2 [57]. The figure was drawn with MolScript [44].

is, it seems possible to define the fold of a protein independent of other proteins. In this case, when we use the term fold, it means the sequence of secondary structure elements and its relative three-dimensional arrangement, which is also called "chain topology" or simply "topology". In this thesis we use the term fold and (chain) topology interchangeably. An illustrative example of a fold is shown in Figure 1.3.

Although the native structures of proteins are implicitly assumed in the above definition, I enlarge the definition so that the term fold can refer to protein-like but non-native structures. The adjective "protein-like" is used for structures which contain as many secondary structures, and are as compact as the native structures. If protein structures (in this case, not necessarily the native ones) have similar backbone conformations, they are said to share the common fold.

1.3 The protein folding problem

The first theory of protein folding was a theory of denaturation proposed by Wu in 1931 [111]. At that time, no structures of any proteins were known, but the denaturation and renaturation of proteins as well as the crystallization of some proteins were. Based on a series of careful experiments by himself and others and on close examination of the results, Wu [111] proposed a "theory" which was stated in the following two sentences:

> The compact and crystalline structure of the natural protein molecule, being formed by virtue of secondary valences, is easily destroyed by physical as well as chemical forces. Denaturation is disorganization of the natural protein molecule, the change from the regular arrangement of a rigid structure to the irregular, diffuse arrangement of the flexible open chain.

It is remarkable that, to a large extent, this theory still holds true. This theory of Wu states that protein folding-unfolding transition is essentially an order-disorder transition. Theories of protein folding since then have been trying to explain how this order-disorder transition occurs and what forces determine the ordered state, *i.e.*, the native state. One defect in Wu's theory is that he assumed the native structure to be like regular symmetrical crystal [111]. It must have been difficult to imagine such complicated structures of native proteins as we know today. In fact, when the first low-resolution structure of a protein, namely sperm whale myoglobin, was solved by Kendrew and coworkers in 1958, they mentioned the structure as follows [42]:

> Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and it is more complicated than has been predicted by any theory of protein structure.

It is interesting to note that the very first attempt to predict protein structure failed, which seems to imply the later development of structure prediction.

The fundamental physical principle of the folding problem is now stated as the "thermodynamic hypothesis" which was proposed by Anfinsen and coworkers [4]. Anfinsen defined the hypothesis as follows [4]:

> This hypothesis states that 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 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.

Based on this hypothesis, the protein folding problem can be stated in two ways. One is to predict the native structure of a protein given its amino acid sequence. The other is to elucidate the pathways along which a protein folds from its denatured state to its native state. Both aspects of the protein folding problem have been studied extensively by experimental means as well as theoretical ones. Here, I concentrate mostly on the theoretical aspects of the problem.

1.3.1 Protein structure prediction

Because of the hierarchical nature of protein structures [1], there are multiple levels in protein structure prediction. Although the ultimate level of prediction is the tertiary (or quaternary) structure, earlier attempts were limited to secondary structure predictions. In the mid 1970's, a blind prediction test was already performed in which the secondary structure of adenyl kinase was predicted by various methods and compared with later solved experimental structure [93].

The increase in the number of experimentally determined protein structures revealed rather limited variations of folds. This observation led to estimation of underlying principles of protein folds from the physical [26] and phylogenetic [92, 12] points of view. It should be noted that already in 1981 when some 150 protein structures were known Schulz [92] estimated the number of basic protein folds to be about 200. Furthermore, the explosive increase of sequence data led Chothia [12] to estimate that the number of protein families is no more than one thousand. Each protein family corresponds in general to a single protein fold. Therefore, if all the protein folds are known, prediction of a protein fold reduces to selecting from the structure database an appropriate fold which is compatible with the given amino acid sequence. Searching the sequence-structure compatibility is assumed to be equivalent to finding the global minimum of the Gibbs free energy of the protein-solvent system. The practical method for tertiary structure prediction exploiting the structural database appeared in the early 1990's. The first successful method was developed by Bowie *et al.* [9]. Their method was subsequently modified and refined by others (e.g., [37, 67]). These methods are now called "fold recognition" or "threading". A case study on protein structure prediction by the fold recognition method is presented in

Appendix A. Although many research groups have tried to improve the fold recognition method, its reliability is still limited. The major source of the limited reliability is the ignorance of the determinants of protein folds. In the fold recognition method, one has to define a set of functions that measure the compatibility of the sequence of interest with structures in the database. This set of functions is conceptually analogous to the energy function of the protein-solvent system although rigorous justification is absent at present. A rigorously physical treatment of sequence-structure compatibility would be possible if the protein-solvent system is represented in the fully atomic detail based on Equation 1.1, but in this case the direct relationship between fold and energetics will be obscured. The concept of fold is related more to the human way of perception than to the physical existence. The physical principles that govern protein folds are not trivial, and they are the main subject in the present thesis.

1.3.2 Folding mechanism *per se*

Recent advances in understanding the folding mechanism have been brought by experimentalists and theoreticians, and their cooperative efforts [2]. Most of the theoretical methods were based on the model introduced by Gō in the 1970's [29]. From the analyses of various interactions in the native structures of proteins, Gō found that these interactions were apparently consistent with each other. Subsequently, he put forward his argument [29]:

> The consistency discussed above cannot be perfect in real proteins. However it is useful to view real proteins as deviating from hypothetical idealized proteins in which the consistency is perfect. The situation is analogous to the concept of an ideal gas. No real gases are ideal. Yet, it is useful to view real gases as

> deviating from the hypothetical ideal gas. The concept of an idealized protein with perfect consistency should be useful because various energy terms in real proteins are indeed consistent with each other in the first approximation. I propose to call this fact the consistency principle in protein folding structure.

Bryngelson and Wolynes [10] applied spin glass theory to study the mutually competing energetic factors in protein structures, and subsequently reformulated the consistency principle as the principle of minimal frustration. The models to which the consistency principle is applied (so called "Gō-models") assume that there exist only those interactions which stabilize the native structure. In order to construct such a model, the native structure of the protein of interest must be known *a priori*. Therefore, the Gō-models are not able to predict the native structure. Nevertheless, recent years have seen many variants of the Gō-model used to study the folding process of real proteins, and thus obtained theoretical results were actually compared with experiments with good correlations [103]. It should be noted that all variants of the Gō-model use coarse-grained representation of protein structures with no atomic detail. One of the most outstanding results was obtained by Plaxco *et al.* [79]. They defined a quantity called "relative contact order" which is the average sequence separation between contacting residues in the native state, normalized by the length of the protein chain. They have found that the relative contact orders of single-domain, fast-folding proteins show a good correlation with the folding rates [79]. It was also shown experimentally that proteins of the same topology fold in similar manners [30]. These studies led to a simple conclusion: the fold (or topology) of the protein determines the folding mechanism [5]. The folding mechanism seemingly does not depend on the atomic details of protein structures.

The success in predicting folding mechanism using the Gō-models enforces the importance of the concept of fold. However, as far as *a priori* knowledge of the native fold is required as in the Gō-model, the problem remains: what determines the fold?

1.4 Forces in protein structure

It is obvious from the form of the Hamiltonian of the protein-solvent system (Equation 1.2) that all the forces involved in the system are essentially electrostatic in nature. However, as mentioned in Section 1.1, it will be easier for us to classify the force into various classes in an approximate and intuitive way. Instead of giving detailed physical formulation, I present more schematic pictures of the forces, mostly based on statistical analyses of protein structures.

1.4.1 Local and non-local interactions

Local interactions are interactions between atoms near by along covalent bonds. This class of interactions includes bond stretching, bond bending, and torsional motions. Standard bond lengths and bond angles are well known. Torsion angles of the polypeptide backbone show distinctively skewed distribution which was first analyzed extensively by Ramachandran and coworkers [82]. They analytically calculated stereochemically allowed regions of the backbone ϕ and ψ angles and presented graphically in the so-called $\phi - \psi$ map, or the Ramachandran plot[1]. Statistical analyses of protein structures have validated the analytical result of Ramachandran and coworkers. Distribution of $\phi - \psi$ angles of twenty types of amino acid residues in a structural

[1] According to Sarma [91], the Ramachandran plot should be called as Ramachandran-Sasisekharan-Ramakrishan ($\phi - \psi$) diagram, signifying equal contributions of the three people.

database of proteins [46] is shown in Figure 1.4.

It can be seen glycines and prolines have distinct features in the distribution compared to other amino acid residues. A closer examination reveals that there are differences among all types of amino acid residues. These differences are caused by the steric hindrance due to the different shapes of the side-chain groups. The propensity of each amino acid residue for α helix [75] or for β sheet [76] is at least partly attributed to these differences in steric hindrance.

Plots similar to the Ramachandran plot can be obtained for the χ angles of side-chain groups (Figure 1.5). Compared to the distribution of $\phi - \psi$ angles, that of $\chi_1 - \chi_2$ seems to show clearer peaks. Ponder and Richards [80] exploited this fact and represented all the side-chain conformation in only 67 rotameric states. They applied their rotamer library to prediction of side-chain packing given the correct backbone structure. Since then, many groups have developed different rotamer libraries as high-resolution structures of proteins become available [53].

It is remarkable that side-chain conformations can possibly be represented as rotamers to a great accuracy in spite of other non-local interactions involved to achieve tight side-chain packing in the native structure. In his derivation of the consistency principle, Gō [29] based one of his arguments on this fact.

Non-local interactions in proteins are the interactions other than the local interactions described above. Non-local interactions include Coulomb interactions due to the partial atomic charges, van der Waals interactions due to instantaneous polarization of atoms [77]. In the case when we are interested in thermodynamically stable structures, entropic force may play a role.

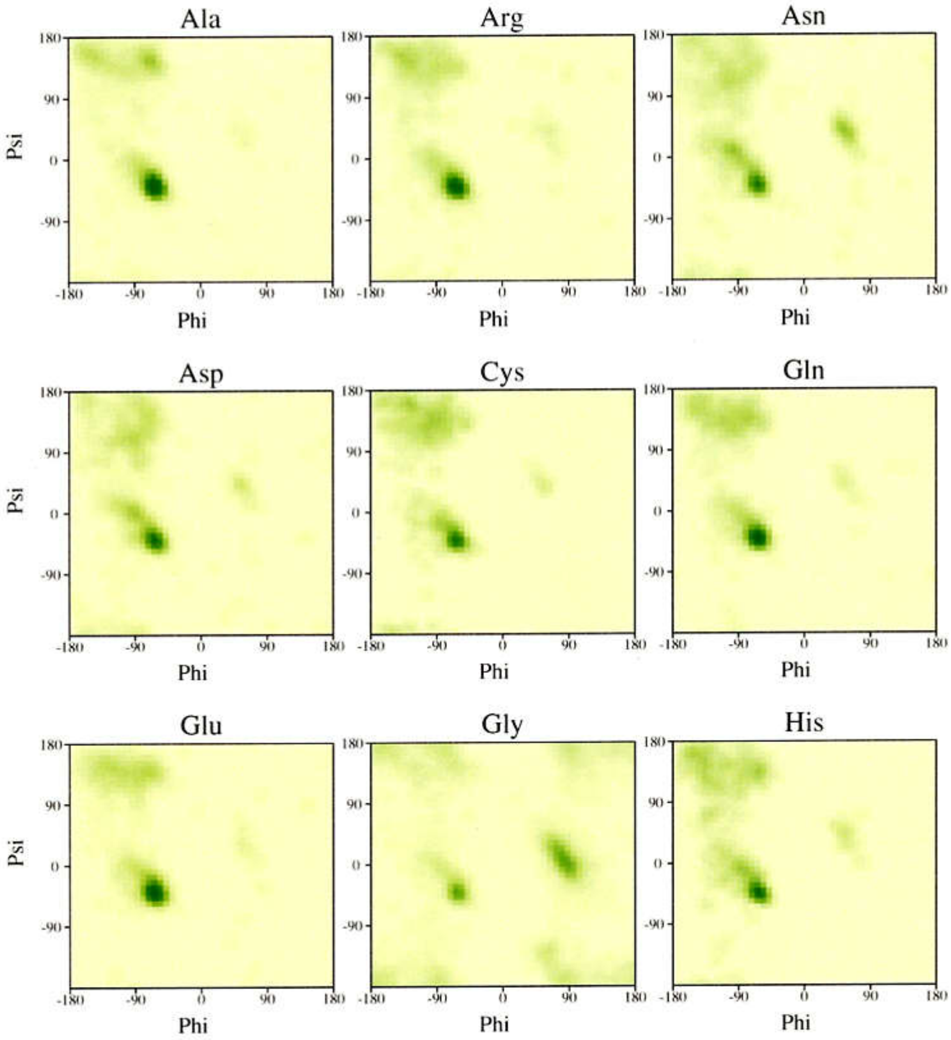


Figure 1.4: The Ramachandran plots for all residue types. Shading shows favorable backbone conformations as obtained from an analysis of 163 structures at resolution 2.0 Å or better. This figure was taken from the output of the PROCHECK program [46].

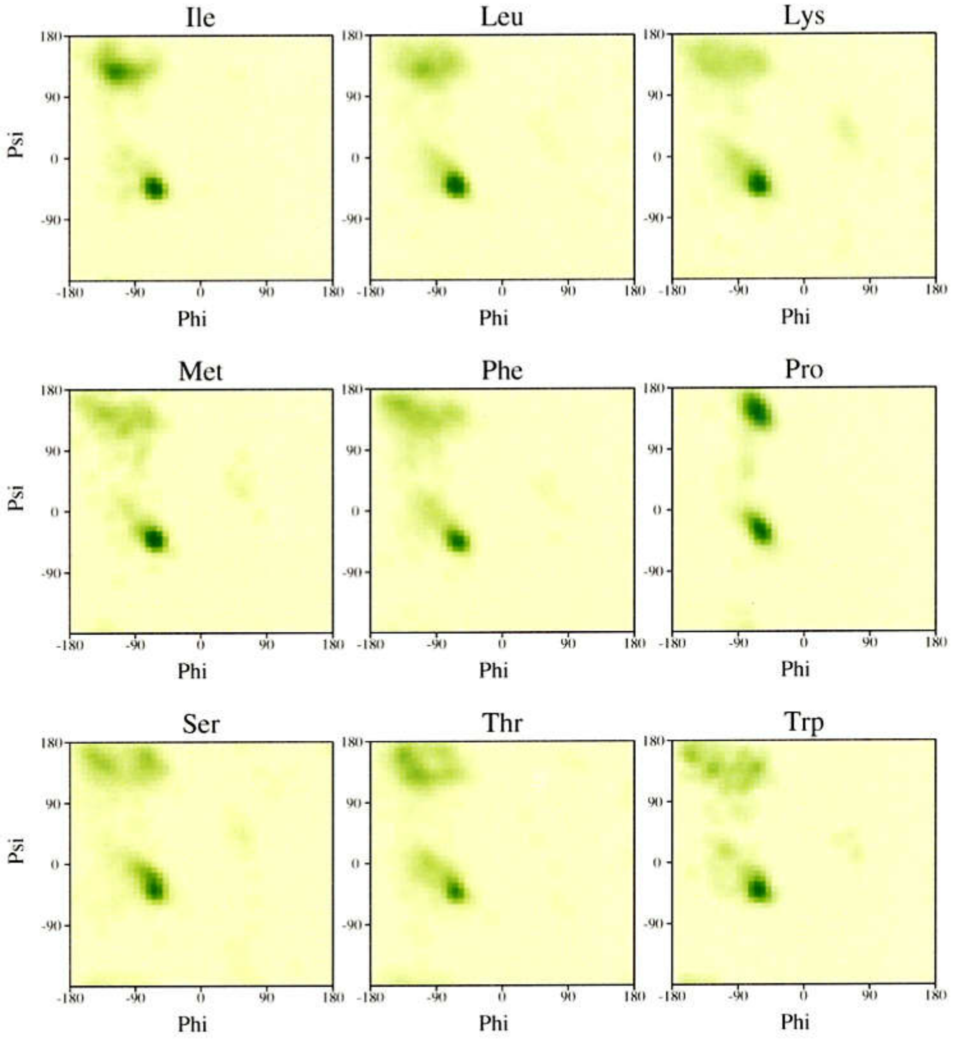


Figure 1.4: The Ramachandran plots for all residue types. (continued)

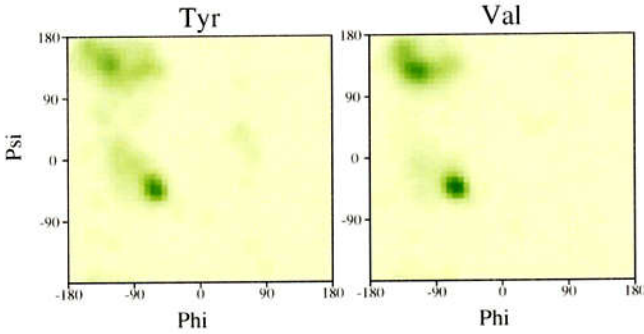


Figure 1.4: The Ramachandran plots for all residue types. (continued)

1.4.2 Solvent effect

In this thesis, I actually mean two effects by "solvent effect". One is the electrostatic effect, the other, the hydrophobic effect. The former is relatively easy to understand. The native structure of a protein is a compact globule with a relatively rigid core. To a first approximation, the compact globule can be regarded as a low dielectric medium. On the other hand, the surrounding solvent is a high dielectric medium with a dielectric constant of about 80. One of the electrostatic effects caused by the high dielectric solvent is the electrostatic shielding caused by reorganization of the dipolar water molecules and salt ions around the protein molecule. Another important electrostatic effect is the Born energy which is associated with the work needed to push a charge into the low dielectric medium. Since a charge is more stable in the high dielectric solvent than in the low dielectric protein, very few charged amino acid residues are observed inside proteins. Also polar or charged groups such as those in the peptide group cannot exist inside a protein without forming hydrogen bonds [65].

The hydrophobic effect is rather difficult to understand. Historically, Kauzmann [41] first introduced the concept of the hydrophobic effect which

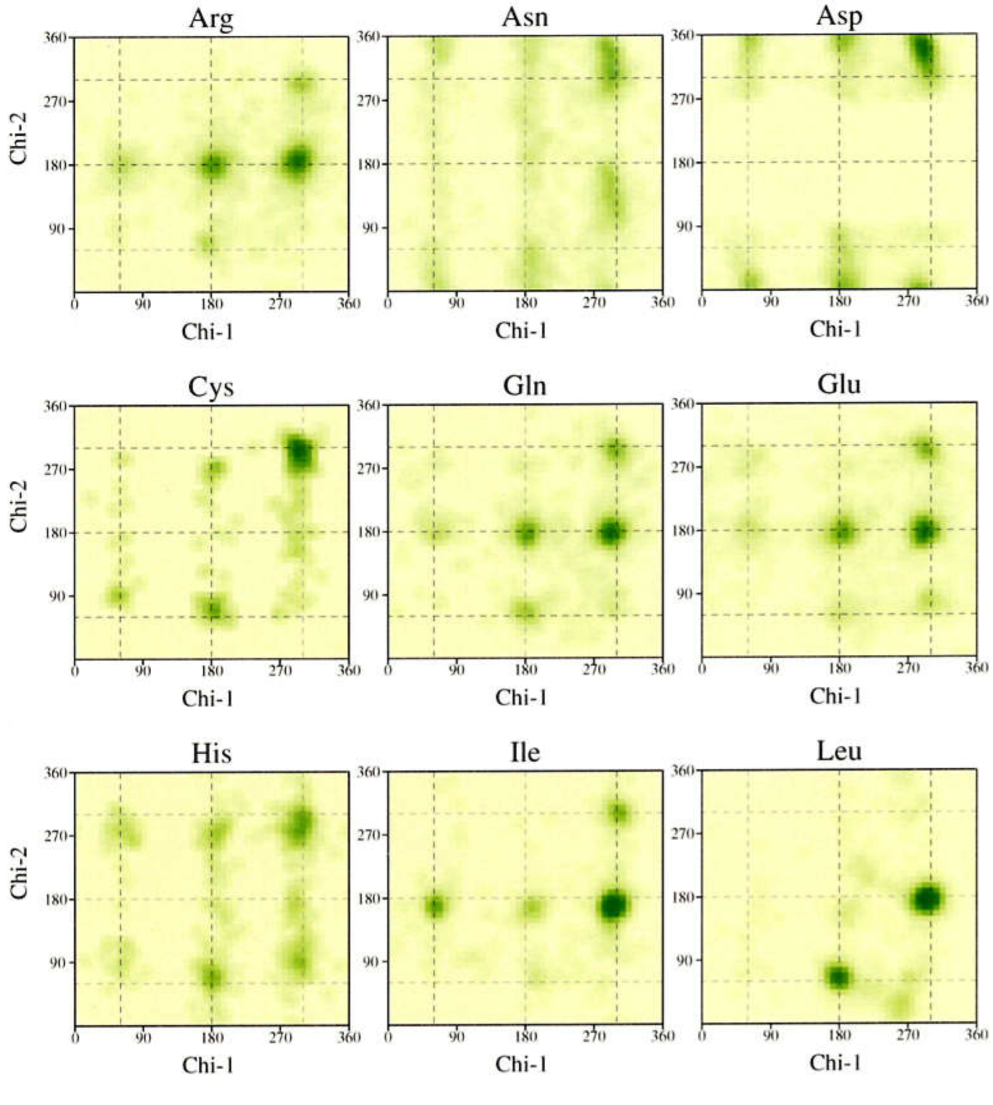


Figure 1.5: The $\chi_1 - \chi_2$ plots. Shading shows favorable side-chain conformations as obtained from an analysis of 163 structures at resolution 2.0 Å or better. This figure was taken from the output of the PROCHECK program [46].

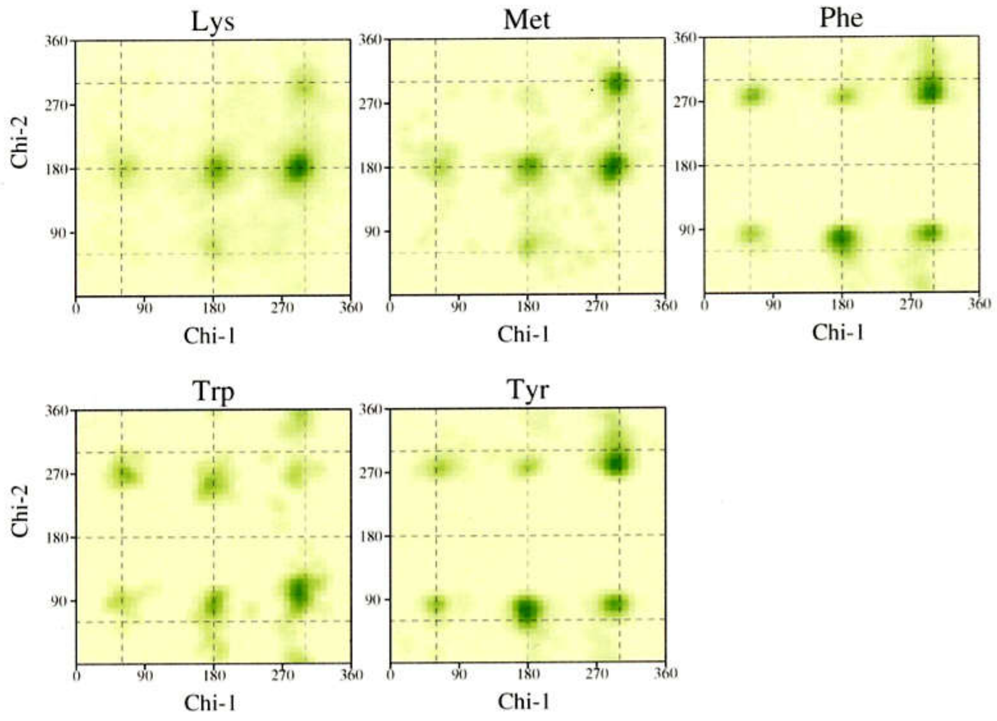


Figure 1.5: $\chi_1 - \chi_2$ plot. (continued)

he called hydrophobic bonding:

> Since the non-polar side chains have a low affinity for water, those polypeptide chain configurations in proteins which bring large numbers of these groups into contact with each other, and hence tend to remove them from the aqueous phase, will be more stable than other configurations, other things being equal. One can consider that the side chains of the above mentioned amino acids[2] (and perhaps others) will form intramolecular "micelles" analogous to the micelles known to occur in aqueous solutions of soaps and detergents. This tendency of the non-polar groups of

[2] This refers to valine, leucine, isoleucine, phenylalanine, as well as proline, alanine and tryptophan.

> proteins to adhere to one another in aqueous environments has been referred to as *hydrophobic bonding.*

On one hand, Kauzmann's notion is right: most non-polar side-chains are indeed in contact with each other and are buried inside the native structures of proteins. On the other hand, it is wrong: the native structures of proteins are not like micelles, they are precisely and tightly packed, which is more like organic crystals [84]. Furthermore, Kauzmann concluded that the origin of the hydrophobic effect was the ordering of water molecules, or *iceberg*, around exposed non-polar groups, which would lead to a large loss of entropy. In other words, the native structure forms so as to minimize the loss of entropy of water molecules. This point was strongly criticized by Makhatadze and Privalov [55]:

> If protein folding results in an overall increase of entropy, why are proteins denatured on heating? The fact that all proteins denature on heating means that the enthalpic factors prevail over the entropic factors in protein stabilization. This experimental fact, however, is ignored in the most theoretical considerations of the mechanism of protein folding.

Based on their supersensitive calorimetric measurements of the folding (unfolding) free energy of many proteins as well as on the hydration and sublimation free energies of small organic molecules, Makhatadze and Privalov [54, 81, 55] have concluded that the van der Waals interactions between the non-polar groups and weak polar interactions between the aromatic groups are the main contributors to the hydrophobic effect, hydration playing relatively a minor role.

1.4.3 Side-chain packing

The Wu theory predicted that the native structure of a protein would be compact and crystalline. As soon as the first protein structure was solved, it was apparent that the protein structure, at least that of myoglobin, was indeed compact, but it was so in spite of its irregularity [42]:

> ··· chains pursue a complicated course, turning through large angles and generally behaving so irregularly that it is difficult to describe the arrangement in simple terms; but we note the strong tendency for neighbouring chains to lie 8–10 Å apart in spite of the irregularity.

The resolution of the first myoglobin structure was 6 Å, therefore even the backbone trace was not complete. Still, the above observation by Kendrew *et al.* was indicative of tight side-chain packing which was revealed in the 1970's by the geometric analysis of Richards [83].

In order to quantitatively analyze the atomic packing in protein structures, Richards [83] introduced a measure called packing density which was defined as the ratio of the volume enclosed by the van der Waals envelope of a given molecule or atom to the actual volume of space it occupies. The van der Waals envelope can be determined once the van der Waals radii of the atoms, and their covalent geometry are given. The "actual volume of space" each atom occupies can be defined by the Voronoi polyhedra (Figure 1.6; [98]). Richards developed an approximate method of the Voronoi tessellation to handle the heterogeneity of the protein structure which consists of various atoms of different radii. He then calculated the packing density for the structures of lysozyme and ribonuclease S, and found that their packing densities were about 0.75 which is very close to the values of the crystals of

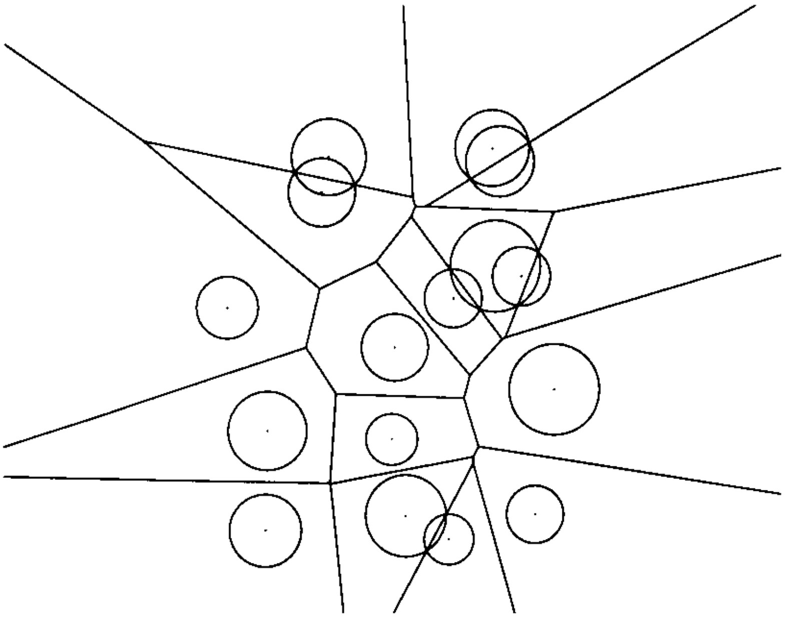


Figure 1.6: An example of the Voronoi tessellation in two-dimensional space. Circles represent atoms, polygons are the associated Voronoi polyhedra (polygons in this case). Note that the Voronoi polyhedra cannot be defined for peripheral atoms.

small organic molecules. Since then, the packing density has been calculated for many other protein structures, and the finding of Richards [83] seems to have been confirmed [85]. However, Richards did not fully take into account the covalent structure, he used instead the average values of the van der Waals volumes for each atom type [83]. Therefore, the value of packing density such as 0.75 may be an overestimate [85]. Nevertheless, the native structure of a protein is well packed as organic crystals. Makhatadze and Privalov [54, 81, 55] used this fact to estimate the energy associated with packing in protein structures from sublimation energy of organic crystals.

Packing itself is a purely geometrical concept. The approach taken by Makhatadze and Privalov is an attempt to link the geometry to energetics. Regarding the present work, it is useful to clarify what good packing is. First, good packing should be associated with a large number of atomic contacts which are located nearly at the minima of the van der Waals interactions. This requirement would be relatively easily achieved for monatomic molecules, but would be rather difficult for polymeric molecules as proteins. Then, the second requirement is less distorted covalent geometry. The exact definition of the packing energy used in this study will be given in Chapter 3.

Chapter 2

Strategy for Studying Protein Energetics

2.1 Molecular mechanics

It should be possible to study the energetics of protein fold by solving Equation 1.1 in principle, which is, needless to say, impossible in practice. Therefore, I employ the molecular mechanics method which treats the physical system in the framework of classical mechanics. The molecular mechanics method consists of various components: the energy function or force-field, construction of atomic models, optimization procedures such as energy minimization or molecular dynamics. In the present study, I used the molecular mechanics program EMBOSS [63] which was kindly provided by Biomolecular Engineering Research Institute, Osaka, Japan. EMBOSS was originally developed for experimental determination of protein structure by nuclear magnetic resonance spectroscopy. One major feature of EMBOSS is that it can perform simulated annealing mass-weighted molecular dynamics in four-dimensional space [34, 63]. Earlier versions of EMBOSS included only the distance geometry force-field. The distance geometry force-field includes only the local geometry terms and soft-repulsion terms apart from the terms for distance and dihedral angle restraints. The local geometry terms are the ones to keep local geometry such as bond lengths, bond angles, chiral volumes, *etc.* to their ideal values. The soft repulsion term is used to avoid van der Waals clashes. The repulsion energy between the atoms i and j are defined as

$$k_{soft}(r_{ij}^2 - r_{ij,c}^2)^2, \quad \text{if } r_{ij} < r_{ij,c} \tag{2.1}$$

$$0, \quad \text{if } r_{ij} \geq r_{ij,c} \tag{2.2}$$

where k_{soft} is the weight of the term, r_{ij} is the distance between two interacting atoms i and j,and $r_{ij,c}$ is the sum of their van der Waals radii. This term is the only non-local interaction term in the distance geometry force-field. As can be seen, this force-field is actually a set of penalty functions that penalize deviations from the ideal geometry of polypeptide chains, and therefore, it is suitable to construct molecular models solely by geometrical restrictions, independent of more "physical" molecular mechanics energy function. This feature of the distance geometry force-field is exploited in the studies given later in this thesis.

The version 5.0 of EMBOSS includes the AMBER all-atom force-field [109] which is described in the next section. I have added a subroutine to calculate the hydration free energy based on the solvation model of Ooi *et al.* [70] in the minimization routine of EMBOSS, which will be described in Section 2.3.

2.2 Energy function

It is difficult, if not impossible, to know the true energy function of the protein in water. Among various energy functions so far developed, the ones for molecular mechanics calculations should be the best possible choice because of their relatively clear physical background. Commonly used molecular mechanics energy functions are composed of several terms:

$$\begin{aligned} E &= \sum_{\alpha} k_{\alpha}(b_{\alpha}-b_{\alpha}^{0})^{2}+\sum_{\beta} k_{\beta}(\theta_{\beta}-\theta_{\beta}^{0})^{2}+\sum_{\gamma} V_{\gamma}(1+\cos(n_{\gamma}\phi_{\gamma}+\delta_{\gamma})) \\ &\quad +\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}\right)+\sum_{i,j}\frac{q_{i}q_{j}}{\epsilon r_{ij}}+\sum_{i,j}\left(\frac{C_{ij}}{r_{ij}^{12}}-\frac{D_{ij}}{r_{ij}^{10}}\right) \end{aligned} \tag{2.3}$$

where the first three terms represent the energies associated with bond lengths (b_{α}), bond angles (θ_{β}) and torsion angles (ϕ_{γ}), and the last three terms represent the non-bonded energy such as van der Waals and Coulomb interactions, and hydrogen bonds, respectively. The hydrogen bond term may be absent in more recent force-fields (e.g., [18]). The parameters such as k_{α}, k_{β}, V_{γ}, A_{ij}, B_{ij}, and q_i *etc.* are determined either by experiments or by *ab initio* quantum mechanics calculations. Most molecular mechanics energy functions ignore the polarization effect of atoms. There are a number of molecular mechanics energy functions proposed. In the present study, we use the AMBER all-atom energy function (force-field) [109].

The most realistic molecular mechanical treatment of protein-water system will include the protein itself with a large number of explicit water molecules around it (Figure 2.1). In this case, the equation (2.3) takes into account the interactions between protein atoms, between protein and water atoms, and between water molecules. Although such calculation is possible and has been actually carried out, it is computationally very heavy and prohibits comprehensive studies. Therefore, some kind of approximation is needed as far as it does not lose the structural details of the protein molecule.

There are several possible ways to approximate the protein-water system. The most simple approximation is to ignore water molecules. If we are interested only in the native structure of the protein, this approximation works rather well. It is suggested that the protein molecules are extremely stable *in*

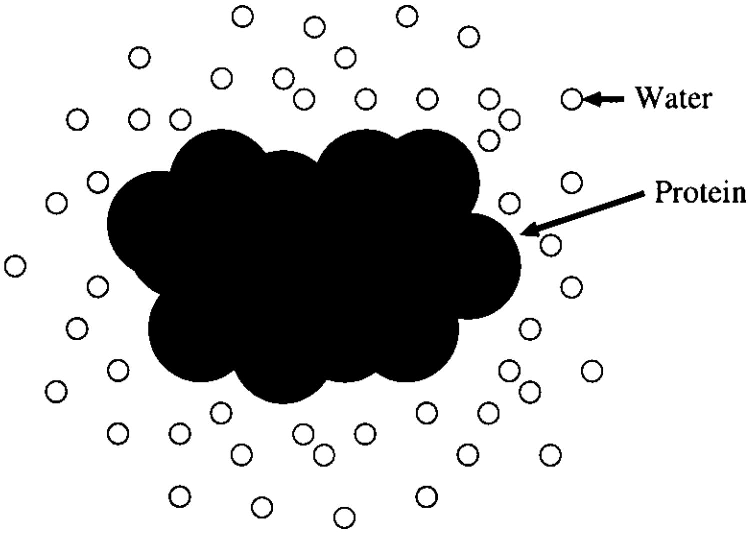


Figure 2.1: Schematic picture of the protein-water system

vacuo [95]. However, if we are interested in other states of the protein such as the unfolded state, ignoring water molecules is not appropriate. Novotný *et al.* [68] showed that the vacuum energy function could not discriminate the native structure from a deliberately misfolded structure unless some kind of solvent effect was taken into account [69]. Since we are interested in the stability of the protein structure, it is crucial to consider states other than the native one. Therefore, the effect of water molecules on the protein cannot be ignored. Then the next step of approximation is to treat the water molecules implicitly, which is discussed in the next section.

2.3 Implicit solvent model

2.3.1 Theoretical background

I follow the formulation by Lazaridis and Karplus [48] in this sub-section. If the Hamiltonian H of the protein-water system is additive, it is expressed as follows:

$$H = H_p + H_{pw} + H_{ww} \tag{2.4}$$

where H_p, H_{pw}, and H_{ww} represent the intramolecular interactions within the protein, protein-water interactions, and water-water interactions, respectively. The configurational canonical partition function Z of the protein-water system can be calculated as

$$Z = \int \exp(-\beta H) d\mathbf{r} d\mathbf{R} \tag{2.5}$$

where $\beta = k_B T$, and $\mathbf{r}$ and $\mathbf{R}$ are the water and protein degrees of freedom, respectively. We define the potential of mean force W by integrating out the Boltzmann factor $\exp(-\beta H)$ by the water degrees of freedom:

$$\exp(-\beta W) = \frac{\int \exp(-\beta H) d\mathbf{r}}{Z_{ww}} \tag{2.6}$$

$$Z_{ww} = \int \exp(-\beta H_{ww}) d\mathbf{r}. \tag{2.7}$$

Using Equation (2.4), we obtain

$$\exp(-\beta W) = \exp(-\beta H_p) \frac{\int \exp(-\beta H_{pw} - \beta H_{ww}) d\mathbf{r}}{Z_{ww}} \tag{2.8}$$

The last part of Equation (2.8) can be regarded as the ensemble average of $\exp(-\beta H_{pw})$ in the pure water system:

$$\frac{\int \exp(-\beta H_{pw}) \exp(-\beta H_{ww}) d\mathbf{r}}{Z_{ww}} = \langle \exp(-\beta H_{pw}) \rangle_0 \tag{2.9}$$

This ensemble average is related to the standard solvation (in the present case, hydration) free energy ΔG^{solv}.

$$-k_B T \ln \langle \exp(-\beta H_{pw}) \rangle_0 = \Delta G^{solv} \tag{2.10}$$

Now, Equation (2.8) becomes

$$\exp(-\beta W) = \exp(-\beta H_p) \exp(-\beta \Delta G^{solv}) \tag{2.11}$$

or

$$W(\mathbf{R}) = H_p(\mathbf{R}) + \Delta G^{solv}(\mathbf{R}) \tag{2.12}$$

Thus, the effective energy function of the protein-water system is represented by the sum of the intramolecular energy of the protein H_p and the solvation free energy ΔG^{solv}. H_p is simply the conformational energy of the protein *in vacuo*, which is readily calculated. The problem now is the realization of the hydration free energy ΔG^{solv}.

2.3.2 Practical representation

There are a number of implicit solvent models to date. Basically all the implicit solvent models are based on the decomposition of the solvent effect to atomic contributions.

$$\Delta G^{solv}(\mathbf{R}) = \sum_{atom\ i} \Delta G_i^{solv}(\mathbf{R}) \tag{2.13}$$

The next assumption is that the contribution of each atom ΔG_i^{solv} can be represented as follows:

$$\Delta G_i^{solv}(\mathbf{R}) = \sigma_i A_i(\mathbf{R}) \tag{2.14}$$

where σ_i is a constant which depends on the atom type of the protein atom i, and A_i is the measure of "accessibility" of the protein atom i to the solvent which is defined by the global structure of the protein. Differences in implicit solvent models emerge from the difference in the definition of the accessibility A. I briefly review a few implicit solvent models.

The excluded volume model The excluded volume model was introduced by Gibson and Scheraga [28]. This model was later called the hydration shell model [39]. In this model, the accessibility A_i is defined as the number of water molecules around the protein atom i. If other protein atoms exist near the atom i, water molecules will be excluded from the neighborhood of the atom i. The number of water molecules excluded from the neighborhood is assumed to be proportional to the total volume of the protein atoms occupying the nearest neighbor of the atom i. Kang *et al.* [39] refined the model by developing an analytical method for the calculation of the excluded volume.

The accessible surface area model The accessible surface area (ASA) introduced by Lee and Richards [50] is the most popular measure of accessibility. Chothia [11] found a linear relationship between the hydrophobicity of amino acid side-chain groups and their accessible surface area which justifies the assumption of Equation 2.14. Eisenberg and McLachlan [21] derived the atomic solvation parameters σ_i from experimentally measured free energy change of amino acid side-chains from octanol to water to estimate the hydrophobicity of the atomic groups. In a similar manner, Ooi *et al.* [70] derived their atomic solvation parameters for seven atomic groups from experimentally measured free energy change of many small rigid molecules from the gas phase to water to complement the molecular mechanics energy function *in vacuo*. Wesson and Eisenberg [110] derived the atomic solvation parameters similar to Ooi *et al.* [70], but for five atomic groups including charged ones.

Various methods for calculating ASA, either numerically or analytically, have been proposed. The analytical method was proposed and proved by Richmond [86]. Basically all the analytical method for the calculation of ASA is based on the Gauss-Bonnet theorem [43] which was used by Richmond himself [86]. Although the analytical method is computationally more demanding than numerical ones, it has an advantage that the derivative of the ASA with respect to the atomic coordinates can be calculated. The derivative of the ASA is useful for energy minimization and molecular dynamics simulations. Wesson and Eisenberg [110] actually carried out molecular dynamics simulations using an ASA-based implicit solvent model.

The solvent contact model The solvent contact model introduced by Colonna-Cesari and Sander [17] is computationally less demanding than other models described above. The accessibility measure A_i of this model is simply the number of contacts that the protein atom i makes with other protein atoms subtracted from the maximum possible number of contacts the atom i can make. The maximum possible number of contacts are derived empirically from, for example, a database of protein structures. The residual number of contacts is supposed to correspond to the number of contacts with water molecules. Since the calculation of the accessibility involves simply counting the number of contacts, this model is easy to implement and to compute. Stouten *et al.* [97] employed this model to simulate a solvated bovine pancreatic trypsin inhibitor. Lazaridis and Karplus [48] developed an implicit solvent model that is a combination of the excluded volume and solvent contact models. They derived the parameters σ_i from the experimental results of Makhatadze and Privalov [54, 81, 55]. Their model also involves modification of the partial charges of protein atoms and a distance-dependent dielectric constant.

For my study, I have employed the implicit solvent of Ooi *et al.* [70] which is based on the accessible surface area of protein atoms. The excluded volume model is very complicated and computationally complex. In this respect, the solvent contact model seems attractive, but it did not give much improvement in side-chain packing prediction (Haruki Nakamura and Akinori Kidera, personal communication). One fundamental difficulty in the solvent contact model is that the interactions with water is reduced to essentially a two-body problem. In other words, the solvent contact model cannot represent the many-body nature of the solvent effect. The accessible surface area model can take into account at least four-body interactions, and it is computationally more feasible than the excluded volume model. Therefore, it is reasonable to employ the accessible surface model.

2.4 Validation of the energy function

In order to test the general ability of the molecular mechanics energy function, I have applied it to the problem of discriminating the native structure out of hundreds of decoys. The four-state reduced decoy sets created by Park and Levitt [73] were obtained from a web site (http://dd.stanford.edu/). Each decoy set consists of more than 600 decoy structures. The structures of the Park and Levitt decoy sets were first energy-minimized for 300 steps using the AMBER potential function, excluding electrostatic and hydrogen bond terms, with the positional restraints on the backbone heavy atoms. Another 100 steps of conjugate gradient energy minimization were carried out using AMBER with the implicit solvent term of Ooi *et al.* (OONS) [70]. In order to take into account the solvent shielding effect, we used a distance-dependent dielectric constant $\epsilon = 2r$ where r is the distance between two interacting atoms.

I used only three sets (1ctf, 1r69, 3icb) out of the seven sets by eliminating those having disulfide bonds or iron-sulfur cluster in order to make the comparison of different structures possible. Also the decoy set for 2cro (434 Cro protein) was not used because it is similar to 1r69 (434 repressor).

Figure 2.2 shows the minimized energy values (AMBER + OONS) for the decoy sets. For all the three cases, the native structure has the energy

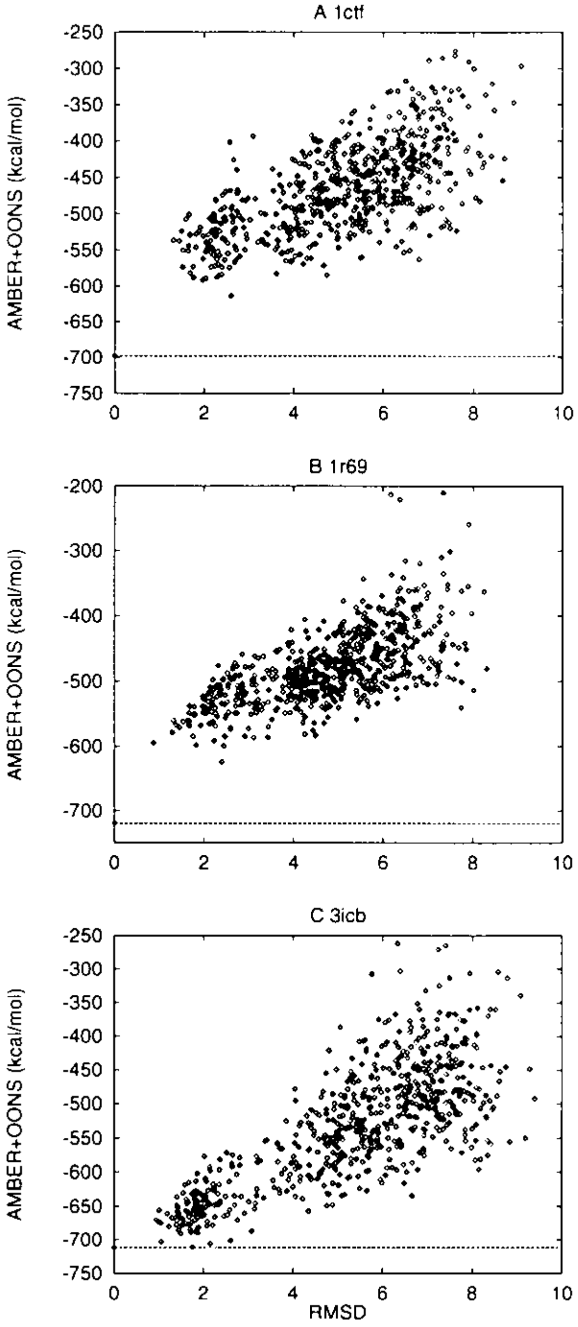


Figure 2.2: Results for the Park and Levitt decoy sets. The horizontal axis is the RMSD (Å) of the decoy from the native structure. The vertical axis is the minimized energy of the decoys. The dashed line indicates the energy of the native structure.

lower than any other decoys. For 3icb, there is one decoy which has only 0.7 kcal/mol higher energy than the native. However, this structure is actually very close to the native one with C_α root mean square deviation (RMSD) of less than 2 Å. Therefore, the energy function in the present study is proved to be sufficiently good for the use in discrimination of the native structure from misfolded structures.

2.5 Electrostatics

Although discrimination of the native structure out of hundreds of decoys was successful, it is simply a necessary condition that an ideal energy function must satisfy. We used the AMBER energy function with a distance-dependent dielectric constant $\epsilon = 2r$ combined with the hydration term of Ooi *et al.* One of the deficiencies of the present energy function is the lack of appropriate treatment of the solvent shielding effect. Also there is no penalty for buried charges which should be given as the Born energy [65]. Although the hydration term of Ooi *et al.* takes into account the direct interactions with the water molecules in the first hydration shell, it cannot represent interactions of more indirect nature such as solvent shielding and the Born energy. As pointed out in an earlier section, fully molecular treatment of the protein-water system (Figure 2.1) is impractical at present. This is especially true for the treatment of long-range electrostatic interactions. One convenient way to calculate the electrostatic potential of the protein-water system is to treat the system as a continuum dielectric medium and to solve the classical Poisson equation [6] or the generalized (linear) Poisson-Boltzmann equation [58]. The continuum dielectric model of the protein-water system (Figure 2.3) treats the protein as a low dielectric medium and the water as a high dielectric one.

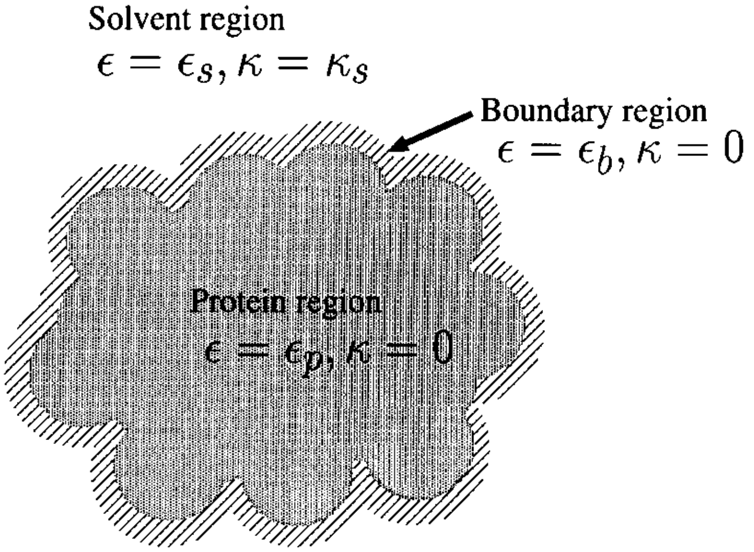

Solvent region
$\epsilon = \epsilon_S, \kappa = \kappa_S$
Boundary region
$\epsilon = \epsilon_b, \kappa = 0$
Protein region
$\epsilon = \epsilon_p, \kappa = 0$

Figure 2.3: Schematic picture of continuum dielectric model of the protein-solvent system.

In order to apply the continuum electrostatics, the protein-solvent system is first divided into three distinct regions: the protein, solvent, and boundary regions. The protein region is defined by the region enclosed by the molecular surface which in turn is defined by the van der Waals radii of protein atoms and the radius of water molecule (1.4 Å). The boundary region is defined as the region between the molecular surface and accessible surface of the protein molecule. The remaining region outside the accessible surface of the protein is defined to be the solvent region (Figure 2.3). Dielectric constant and ionic strength are assigned to each region as schematically shown in Figure 2.3.

The solution to the following generalized linear Poisson-Boltzmann equation gives the electrostatic potential $\phi(\mathbf{r})$ of the protein-solvent system [65].

$$\nabla \cdot \epsilon(\mathbf{r}) \nabla \phi(\mathbf{r}) = -4\pi \sum_i q_i \delta(\mathbf{r} - \mathbf{r}_i) + \epsilon_s \kappa^2 \phi(\mathbf{r}) \tag{2.15}$$

where $\epsilon(\mathbf{r})$ and $\phi(\mathbf{r})$ are the dielectric constant and the electrostatic potential at the position $\mathbf{r}$, respectively. q_i and $\mathbf{r}_i$ are the charge and the position of the i-th protein atom. $\epsilon(\mathbf{r})$ is set to ϵ_p in the protein region, and to ϵ_s in the solvent region. κ is the Debye's shielding parameter which is meaningful only in the solvent region. The atomic charges and van der Waals radii were taken from the AMBER force-field in this study. The dielectric constants of the solvent region (ϵ_s) and the boundary region (ϵ_b) were set to the macroscopic value of water, 80. The dielectric constant ϵ_p of the protein is not trivial because of the heterogeneity of the protein molecule. Nevertheless, it is a common practice to set it to a value between 1 and 20 throughout the protein region. The values $\epsilon_p = 10$ and 4 were used in Chapters 3 and 4, respectively. More specifically, the Poisson equation

$$\epsilon_p \nabla^2 \phi(\mathbf{r}) = -4\pi \sum_i q_i \delta(\mathbf{r} - \mathbf{r}_i) \tag{2.16}$$

is solved in the protein region, and the Laplace equation

$$\nabla^2 \phi(\mathbf{r}) = 0 \tag{2.17}$$

is solved in the boundary region, and the Poisson-Boltzmann equation

$$\nabla^2 \phi(\mathbf{r}) = \kappa^2 \phi(\mathbf{r}) \tag{2.18}$$

is solved in the solvent region with appropriate boundary conditions to smoothly connect the electrostatic potential of the different regions.

The partial differential equation (Equation 2.15) can be readily solved by numerical methods. For numerical integration, a large box containing the protein molecule is set which is partitioned into a large number of small cubes. The side length of each cube is set to 1 Å to retain the numerical accuracy. The self-consistent boundary condition [66] is imposed at the edges of the box so that the electrostatic potential approaches zero at infinitely distant points.

In order to avoid the divergence of the electrostatic potential due to the self Coulomb energy of the point charges of protein atoms, we solve the above equation twice, one for the protein-solvent system ($\epsilon_s = 80$), the other for the protein in vacuum ($\epsilon_s = \epsilon_p$). Thus we obtain two solutions ϕ^{sol} and ϕ^{vac}, respectively. By taking their difference, we obtain the reaction field, $\phi^{react} = \phi^{sol} - \phi^{vac}$ from which the self energies and Coulomb interaction energies are eliminated. Finally, the electrostatic energy of the protein-solvent system (E_{p-s}) is given by

$$E_{p-s} = \frac{1}{2}\sum_i q_i \phi^{react}(\mathbf{r}_i) + {\sum_{i<j}}' \frac{q_i q_j}{\epsilon_p |\mathbf{r}_i - \mathbf{r}_j|} \tag{2.19}$$

The second sum in the right hand side of this equation is the Coulomb interaction energy, which is calculated only for 1-4 and 1-5 interacting pairs of atoms as is done for usual calculations with the AMBER force-field. In this case, no distance cut-off is applied for the Coulomb interaction term.

The electrostatic energy E_{p-s} given by Equation (2.19) includes, in addition to intramolecular Coulomb interaction energy, the solvent shielding effect and the Born energy. However, it should be noted that the heterogeneity of the protein molecule is neglected in the continuum dielectric model. Furthermore, dielectric constants are originally defined for macroscopic objects instead of microscopic ones such as protein molecules. Therefore, the electrostatic energy E_{p-s} obtained by the continuum dielectric model should be taken as a first approximation.

2.6 "Near-native" structures

A protein is made of atoms. A protein structure is defined by the atomic coordinates. Whereas the fold of a protein is an artificial concept in which atomic details are neglected, calculation of molecular mechanics energy requires the atomic details of the protein structure. Then how can we study the energetics of protein fold? Comparative study of the structures of homologous or analogous proteins have been a powerful method to elucidate the determinants of protein fold. For example, the series of works by Chothia and coworkers [7, 32, 13] have shown that a set of the *key residues* shared by the members of each superfamily of proteins are the determinants of their fold. Murzin [61] showed that this comparative approach was extremely powerful in structure prediction of proteins with no homologous structures in the database. However the "determinants" in their sense are evolutionary ones, rather than energetic ones. In this study, I try to solve the conceptual conflict between protein structure and protein fold by introducing a class of atomic models called *near-native* structures.

If there should be any physicochemical factors that determine the fold of a protein, then these factors may be recognizable even when the precisely determined native structure is perturbed as far as the perturbed structure shares the common fold with the native structure. Since the fold is determined by the backbone conformation irrespective of the side-chain conformation, the perturbation is applied to the side-chain conformation while the native backbone conformation is retained. I call thus constructed structure the near-native structure. An example of the near-native structure is shown in Figure 2.4.

Near-native structures as depicted in Figure 2.4 are constructed by minimizing atomic overlaps with positional restraints on the backbone atoms so that the fold is the same as, while the side-chain conformations are different from the native structure. The positional restraints are imposed as a penalty function E_{pos}:

$$E_{pos} = \frac{1}{2}k_B T_0 {\sum_i}' |\mathbf{r}_i - \mathbf{r}_i^0|^2 \tag{2.20}$$

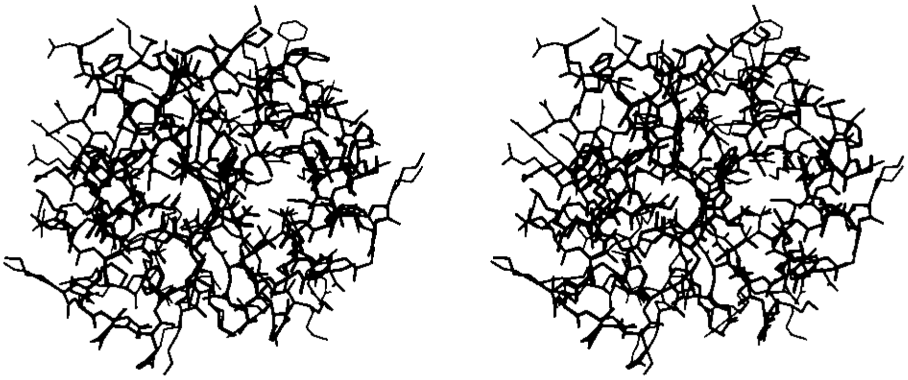


Figure 2.4: Stereoview of superposed native and near-native structures. The thick line indicates the native structure of the PDB entry 1tmy [105]. The thin line indicates a near-native structure based on 1tmy. The figure was drawn with MolScript [44].

where k_B is the Boltzmann constant and T_0 is an absolute temperature set to 300 K, and $\mathbf{r}_i$ and $\mathbf{r}_i^0$ are the atomic coordinates of the i-th atom of the near-native and crystal structures, respectively. The prime ($'$) on the summation indicates that the sum is taken for the backbone heavy atoms only. The near-native structure is an operationally defined, artificial object. In nature, the probability that the near-native structure exists will be extremely low. Conformational states like the near-native structure have not been observed to date. Nevertheless, I will show that the near-native structure is methodologically useful for the study of the energetics of protein fold. Since the backbone conformations of the native and near-native structures are almost the same, energetic factors shared by them strongly suggest themselves to be the determinants of the fold. The details on how to construct near-native structures will be given in later chapters (Chapters 3 and 4). It should be stressed that the near-native structures are constructed solely from geometric restraints based on the backbone conformation of the native structure and the intrinsic steric hindrance of protein atoms using soft repulsive terms described in Section 2.1. Hence, it is expected that the near-native structures should be independent of any particular force-field employed to evaluate their energy.

Chapter 3

Energetic Determinants of Protein Fold

3.1 Introduction

This chapter is the main body of this thesis. The problem regarding the energetics of protein fold is directly addressed in an intuitive manner. A formal and abstract formulation of the problem is presented in Appendix B.

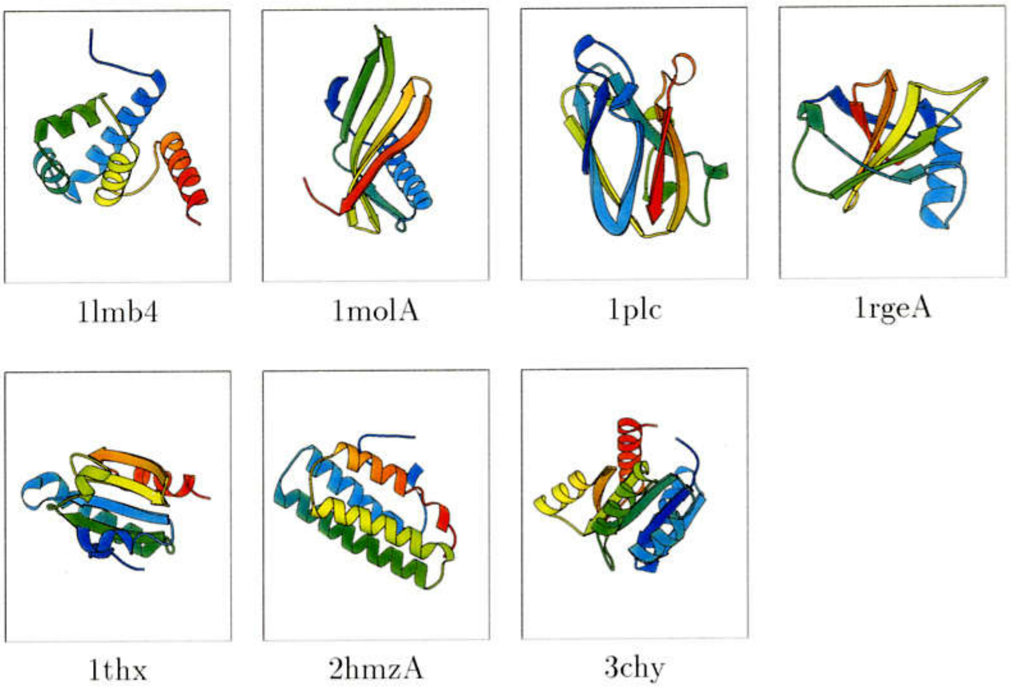
The phrases like "protein structure prediction is one of the most important but unsolved problems in the field of biophysics" are already classical, but they still hold true. These statements refer to the cases when there is no apparent homology between the sequence of the protein of unknown structure and sequences of known structures. The ultimate goal of the protein structure prediction problem is the prediction of the native structure which is defined by both the backbone and side-chain conformations. It is believed that the native structure of a protein is of the lowest energy conformation. Therefore, given the correct energy function, predicting the native structure is equivalent to finding the conformation with the lowest energy. Since finding the global minimum conformation seems very difficult, one often tries to predict the native "fold" of a protein. Although the definition of a "fold" is in general artificial, it is usually interpreted as a set of similar backbone structures of proteins. A problem arises when one tries to find the native fold because a structure with the native fold is not necessarily the same as the native structure, and therefore there is no known physical factors that determine the native fold. For the prediction of the native fold, coarse-grained protein models, in which residue-wise interactions are taken into account, are often used. Then the fold at the minimum of such residue-wise interaction potential is assumed to be the native fold. Threading in general and some *ab initio* methods utilize such a strategy but their success is currently limited. Typical residue-wise interaction potentials are derived from structural databases [96]. Not only because the structural databases contain the sources of errors and noises, but also because the functional form of the potential depends on one's intuition, it is often difficult to identify what is wrong with the potentials, which in turn makes it difficult to improve them. It is therefore preferable to use more physically well-grounded potential functions. Although the "true" potential function is difficult to know, the conventional molecular mechanics force-fields should be a best possible choice.

Since the pioneering work by Novotný *et al.* [68], it has widely been believed that the molecular mechanics energy function is unable to discriminate the native structure from the misfolded ones. However, some authors recently have begun to claim that a molecular mechanics energy function combined with a hydration term can discriminate the native structure from the misfolded. Janardhan and Vajda [36] investigated the use of a molecular mechanics energy function combined with solvation and entropic terms for selecting near-native structures among homology-based models. They showed that the solvation and molecular mechanics energy terms are useful for selecting models with good side-chain packing and well-built loops, respectively. Vorobjev *et al.* [107] developed an elaborate method to calculate conformational free energy of proteins combining molecular dynamics simulation with explicit and implicit solvent models. Their method incorporates conformational entropy as well as the solvation free energy. They applied the method to the native and misfolded structures of 9 small proteins and found that the native structures always gave lower conformational free energies than the misfolded. Lazaridis and Karplus [48] have developed an effective energy function that combines the conformational potential function with a simple solvent model. Their effective energy function was successful in discriminating the native structures from misfolded structures and hundreds of decoys [47].

In this chapter, I investigate the physicochemical factors that are necessary for discriminating between the native folds and misfolds. To do so, I construct "near-native" models as representatives of the native (i.e., correct) fold. The backbone structure of a near-native model is almost the same as that of the native structure, but its side-chain conformation is different from that of the native structure. These near-native models are constructed by the same procedure as misfolded models. From the comparison of the energy components of the native structure, and near-native and misfolded models, it became possible to identify the factors that are necessary for the prediction of the native fold.

3.2 Generation of correct and incorrect folds

I arbitrarily selected seven proteins of various structural classes (all α, all β, α/β and $\alpha+\beta$) which are composed of approximately 100 amino acid residues. These proteins are called "target" proteins (Figure 3.1 and Table 3.1) and I construct correctly and incorrectly folded models of these



1lmb4 1molA 1plc 1rgeA

1thx 2hmzA 3chy

Figure 3.1: Structures of the target proteins

Table 3.1: Target proteins

Target[a]	Nres[b]	class[c]	name	reference
1lmb4	92	all α	lambda repressor	[16]
1molA	94	$\alpha + \beta$	monellin	[104]
1plc	99	all β	plastocyanin	[31]
1rgeA	96	$\alpha + \beta$	ribonuclease Sa	[94]
1thx	108	α/β	thioredoxin	[88]
2hmzA	113	all α	hemerythrin	[35]
3chy	128	α/β	CheY protein	[106]

[a] The Protein Data Bank (PDB) codes of the target proteins. The fifth letter, if present, indicates the chain identifier. [b]Number of residues of the target protein. [c]Structural classification according to the SCOP database [62].

target proteins. All the protein structures were obtained from the Protein Data Bank (PDB) [8] and the names of all the proteins in this chapter are referred to by their Protein Data Bank (PDB) codes. The target proteins do not have disulfide bonds in the structural core. For those having a disulfide bond (i.e., lrgeA and lthx), I deleted the disulfide bonds so that all the cysteine residues are treated in their reduced form. The backbone structures of correctly or incorrectly folded models of a target protein were selected using a conventional threading method in the following manner. Each target sequence was threaded through structures whose number of residues are larger than that of the target in a fold library without gaps using a threading program S3 [71]. The program S3 employs a sequence-structure compatibility function which were statistically derived from a structural database. The compatibility function consists of four terms: side-chain packing, hydration, local conformation, and hydrogen bonds. These terms are similar to those of Matsuo *et al.* [56] but local structures are treated in more detail [71]. The candidates for the predicted structures were selected from those giving the top 10 scores. Those candidates are called "templates" and are listed in Table 3.2. In order to avoid non-compact structures, only those proteins whose sequences are longer than the target at most by 5 residues were used and the obtained template structures are indeed compact (Table 3.2).

Since only the backbone atoms are represented explicitly in the threading, we next have to build the side-chain conformations in order to apply the molecular mechanics energy function. The model building was done as shown in Table 3.3 using the program EMBOSS [63] which was originally developed for the structure determination by nuclear magnetic resonance spectroscopy. EMBOSS can perform efficient conformational sampling by the simulated annealing mass-weighted molecular dynamics in four dimensional space [34,

Table 3.2: Templates obtained by ungapped threading.

Target	Correct			Incorrect	R_g ratio[e]
	NN[a]	HM[b]	MA[c]	misfolded[d]	
1lmb4	1lmb4			1frrA 1molA 1nsgB 1pdr 1rgeA 1ris 1sxl 2hgf 2rgf	0.93 (0.05)
1molA	1molA			1hsbB 1nct 1plc 1prtF 1ris 1stfI 1sxl 2acy 2rgf	0.95 (0.03)
1plc	1plc			1ag2 1audA 1bftA 1dcpC 1hsbB 1lktA 1lt5D 1tlk 1tul	1.06 (0.04)
1rgeA	1rgeA			1aps 1audA 1hsbB 1iuz 1nct 1pcs 1stfI 1tiiD 2ncm	1.04 (0.04)
1thx	1thx	2trxA	1tof	1bcpD 1cewI 1csyA 1kpeA 1picA 1rblM 1rtu	1.06 (0.06)
2hmzA	2hmzA		2mhr	1cd8 1dutA 1hcd 1kb5A 1pbk 1rot 1tvdA 2rspA	1.02 (0.04)
3chy	3chy			135l 1a25A 1adl 1aizA 1bbhA 1bff 1cpq 1ftpA 1kuh	1.06 (0.07)

[a]The templates for "near-native" models. [b]The template for a "homologous" model. [c]The templates for "misaligned" models. [d]The templates for "misfolded" models. [e] Average ratio (and its standard deviation in parentheses) of the radius of gyration of constructed models to that of the native structure.

63]. A random coil was given as the initial conformation. After the simulated annealing in four dimensional space, the weight k_{4D} of the 4-th dimensional energy is increased to compress the fourth coordinate and to obtain the three-dimensional structure. From stages 1 to 4 in Table 3.3, the distance geometry force-field was used, which consists of only local geometry terms and long-range soft repulsion terms, but with no attractive term. Positional restraints on all the backbone atoms (Equation 2.20 on page 37) were imposed throughout the optimization steps so that the backbone structure of the model becomes the same as the coarse-grained model used in the threading. Note that side-chain conformations are determined by only the repulsive

Table 3.3: Protocol of simulated annealing optimization.

	A random coil is generated as the initial structure
stage 1	500-steps conjugate gradient minimization ($k^a_{4D} = 0.05$) Distance geometry force-field
stage 2	5,000-steps molecular dynamics at 1000 K ($k_{4D} = 0.05$) Distance geometry force-field atomic mass 1,000 Da, step size 50 fs, coupling constant 40 ps
stage 3	100,000-steps molecular dynamics to 1 K ($k_{4D} = 0.05$) Distance geometry force-field atomic mass 1,000 Da, step size 50 fs, coupling constant 40 ps cooling rate 1 K/100 steps
stage 4	3,000-steps conjugate gradient minimization ($k_{4D} = 10.0$) Distance geometry force-field
stage 5	500-steps conjugate gradient minimization AMBER force-field with OONS[b]

Positional restraints on all the backbone atoms (Equation 2.20) are imposed throughout the stages. [a]The weight of 4-th dimension. [b]For the calculations shown in Figure 3.3B, OONS was not included.

term, that is, they are determined simply by minimizing the atomic overlaps with each other and with backbone atoms. A residue-based cutoff scheme was applied. The cutoff length of 6 Å was used through the stages 1 to 4 in Table 3.3, and 12 Å was used when a minimization involved the AMBER force-field. The interaction tables were updated every 100 steps through stages 1 to 4, and every 20 steps for all other cases. For comparison, the native structure of each target determined experimentally was also minimized for 500 steps of the conjugate gradient method with positional restraints on the backbone atoms (Equation 2.20). By the minimization, the native structures deviated from the experimental structures by RMSD (all heavy atoms) of at most 0.3 Å.

Computations were done on a VPP500 (Fujitsu) with vector processors and AP3000 (Fujitsu) workstations with Ultra SPARC-II (296MHz) CPUs.

A typical modeling procedure from the stage 1 to 4 took about 1 hour for each model on the VPP500. 100 steps of conjugate gradient minimization with AMBER and the implicit solvent term took about 3 minutes for each model of 1thx (108 residues) on the AP3000.

3.3 Evaluation of correct and incorrect folds

The conventional ungapped threading search was performed for each target in Table 3.1 using the program S3 [71]. The ten structures giving the best ten scores of S3 were selected as the template structures for the target, and were subject to all-atom modeling and the energy minimization by the method described in Section 3.2. In all the cases I tried, the native conformations were ranked at the top as expected [38]. The selected structures are summarized in Table 3.2. All the templates are of protein-like structures in that they are compact and contain significant amount of secondary structures. A model based on the native backbone structure itself is called "near-native" model. Note that the side-chain conformations of the near-native models are completely reconstructed in the same manner as other models. Structural differences between the native and reconstructed near-native structures are shown in Table 3.4. For the target 1thx, two homologous structures (2trxA and 1tof) were incidentally found (Table 3.2). The alignment of 1thx and 2trxA is correct. We call the model "1thx-2trxA" (the model of the target "1thx" based on the template "2trxA") the "homologous" model. The alignment of 1thx and 1tof is partly incorrect (Figure 3.2A), thus we call the model 1thx-1tof the "misaligned" model (Table 3.2). Based on the alignments by the threading, the sequence identity of the templates 2trxA and 1tof with the native sequence of 1thx is 42.6% and 17.6%, respectively. The backbone RMSD of the models 1thx-2trxA and 1thx-1tof from the native

Table 3.4: The difference between reconstructed near-native models and the native structures.

model	RMSD (Å)		χ correct (%) [c]	
	backbone[a]	all[b]	χ_1	χ_2
1lmb4-1lmb4	0.136	1.62	78.6	80.0
1molA-1molA	0.116	1.90	76.5	72.7
1plc-1plc	0.121	1.45	70.8	22.2
1rgeA-1rgeA	0.104	1.51	82.4	83.3
1thx-1thx	0.130	1.57	64.3	88.9
2hmzA-2hmzA	0.117	2.01	65.5	79.2
3chy-3chy	0.123	1.55	75.0	85.7
Average	0.121	1.66	73.3	73.1

[a] RMSD for backbone heavy atoms between the native and near-native model.
[b] RMSD for all the heavy atoms between the native and near-native model.
[c] The fraction of correctly predicted χ angles of buried residues. Residues are defined to be buried if the solvent accessible surface area is less than 10% of extended Gly-X-Gly conformation. χ angles with deviation from the native less than 40 degrees are considered to be correct.

structure (1thx) is 1.2 Å and 3.7 Å, respectively. For the target 2hmzA, one homologous structure (2mhr) was found by threading and their alignment is partly incorrect (Figure 3.2B) with the sequence identity of 33.6%, thus the model 2hmzA-2mhr is another misaligned model. Its backbone RMSD from the native (2hmzA) is 3.4 Å. The native, near-native, homologous and misaligned models are defined to be "correct" models. Other models are of totally different fold from the native structure with RMSD of more than 8 Å, thus they are called "misfolded" models, accordingly, they are defined to be "incorrect" models (Table 3.2).

After the all-atom modeling and the energy minimization, the electrostatic energy based on the continuum dielectric model was calculated for each model. The results of the energy calculations are shown in Figure 3.3A. In this figure, the total energy is defined as the sum of the AMBER energy

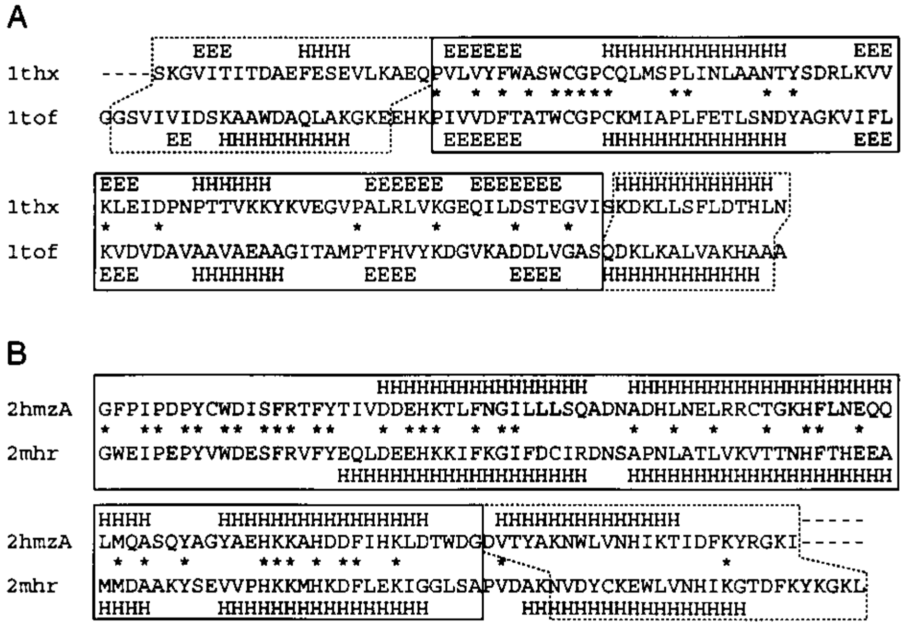

A
EEE HHHH EEEEEE HHHHHHHHHHHHHH EEE
1thx ----SKGVITITDAEFESEVLKAEQPVLVYFWASWCGPCQLMSPLINLAANTYSDRLKVV
* * * * ***** ** * *
1tof GGSVIVIDSKAAWDAQLAKGKEEHKPIVVDFTATWCGPCKMIAPLFETLSNDYAGKVIFL
EE HHHHHHHHHH EEEEEE HHHHHHHHHHHHHH EEE
EEE HHHHHH EEEEEE EEEEEEE HHHHHHHHHHHH
1thx KLEIDPNPTTVKKYKVEGVPALRLVKGEQILDSTEGVISKDKLLSFLDTHLN
* * * * * *
1tof KVDVDAVAAVAEAAGITAMPTFHVYKDGVKADDLVGASQDKLKALVAKHAAA
EEE HHHHHHH EEEE EEEE HHHHHHHHHHHH
B
HHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHHHHHH
2hmzA GFPIPDPYCWDISFRTFYTIVDDEHKTLFNGILLLSQADNADHLNELRRCTGKHFLNEQQ
* ** ** ** *** ** * *** * ** * * * * ** *
2mhr GWEIPEPYVWDESFRVFYEQLDEEHKKIFKGIFDCIRDNSAPNLATLVKVTTNHFTHEEA
HHHHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHHHHHH
HHHH HHHHHHHHHHHHHHHH HHHHHHHHHHHHHH -----
2hmzA LMQASQYAGYAEHKKAHDDFIHKLDTWDGDVTYAKNWLVNHIKTIDFKYRGKI-----
* * * *** * ** * * *
2mhr MMDAAKYSEVVPHKKMHKDFLEKIGGLSAPVDAKNVDYCKEWLVNHIKGTDFKYKGKL
HHHH HHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHHH

Figure 3.2: The alignments of misaligned models obtained by ungapped threading. A: The alignment for the model 1thx-1tof. B: The alignment for the model 2hmzA-2mhr. The region bound by a solid box coincides with the correct alignment. The region bound by a dotted box indicates incorrectly aligned sites. The secondary structures ("H" for α helices, "E" for β strands) are also shown. The alignment sites with identical residues are marked with asterisks.

without the Coulomb interaction terms, and the OONS hydration free energy, and the electrostatic energy E_{p-s} obtained by Equation (2.19). Large energy gaps of more than 100 kcal/mol were found between the native structures and the near-native or any misfolded models. Although the near-native structures have much higher energies than the native, their energies are nevertheless lower than those of the misfolded ones (Figure 3.3A). Even the misaligned models have fairly low energies compared to the other misfolded models. It is because their backbone topologies are similar to those of the native as mentioned above and in Figure 3.2. The homologous model 1thx-2trxA has higher energy than the near-native model 1thx-1thx, and lower energy than the misaligned model 1thx-1tof. This trend is reasonable considering the structural similarity of these three models to the native structure of 1thx. This result shows that the present energy function is also able to discriminate the correct folds from incorrect ones. In the following section, I examine combinations of energy terms in search of physicochemical determinants of the native structure and the correct fold.

3.4 Solvent effects

In order to see the effect of the solvent, I carried out the minimization without the hydration term (Figure 3.3B). In this case, the compared energy is the AMBER force-field with a distance-dependent dielectric constant ($\epsilon = 2r$). The native structures always have lower energies than any other near-native or misfolded structures. Although the near-native structures have lower energies than the misfolded structures, the energy difference is small (less than 15 kcal/mol) for targets such as 1lmb4 and 1molA. Also, the homologous model 1thx-2trxA has higher energy than the misaligned model 1thx-1tof in contrast to the result for the total energy discussed above (Figure 3.3A). It seems that the solvent effect is important for stabilizing near-native structures rather than the native structure. The solvent effects are further discussed in the following paragraphs.

The "hydrophobic interaction" is believed to be a dominant factor in stabilizing protein structures [20]. In the present study, the hydration effect is taken into account in terms of the implicit solvent model of Ooi *et al.* [70]. Figure 3.4A shows that the correct structures, including the native, do not necessarily have lower hydration free energy than incorrect structures. This trend is also observed by others [108, 107, 47]. Makhatadze and Privalov [55] showed that the main contributors to the hydrophobic interactions are van

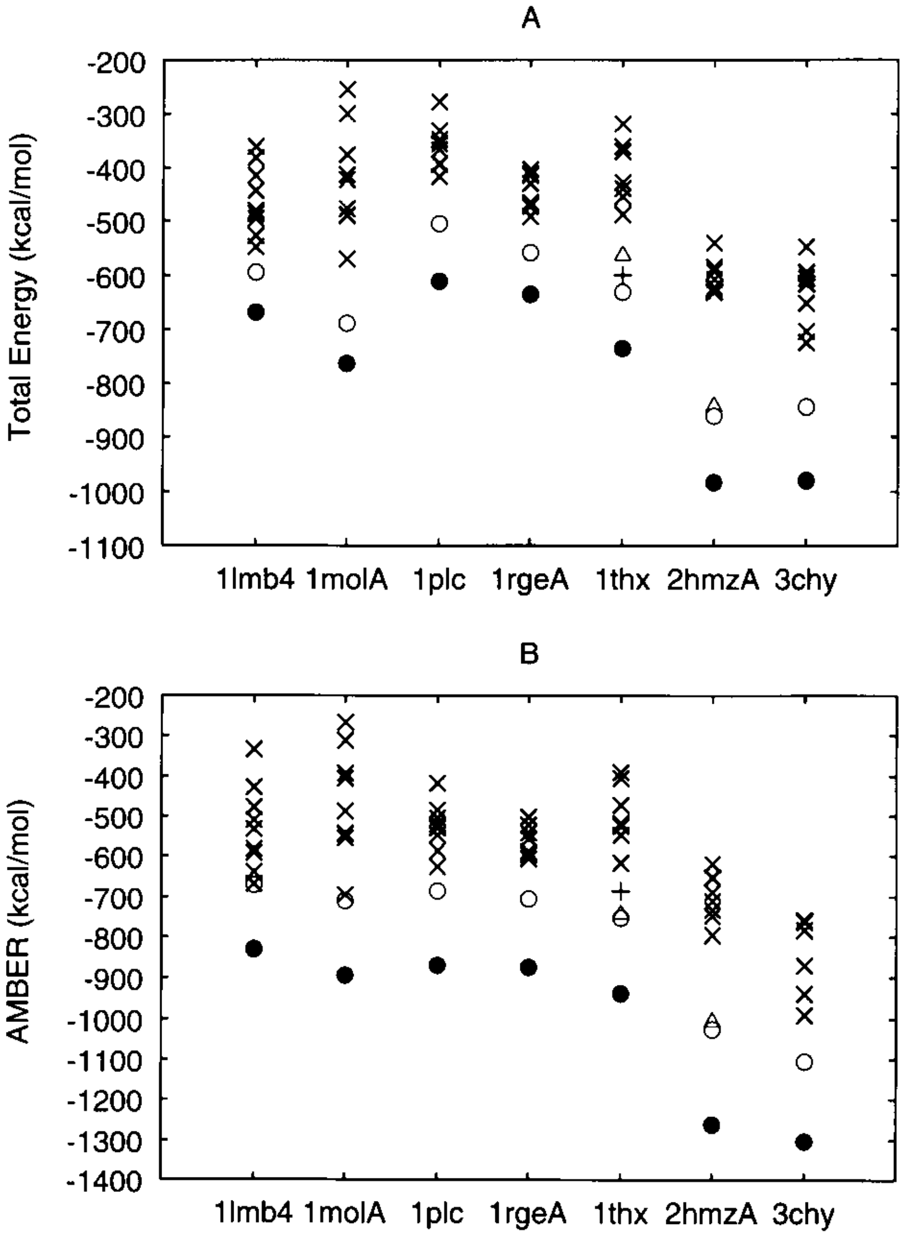


Figure 3.3: Total energies of the native structures, correct and incorrect models. The horizontal axis indicates target proteins. The symbols are defined as follow: "•", native structures; "o", near-native models; "+", homologous model; "△", misaligned models; "×", misfolded models. See Table 3.2 for the nomenclature of the models. A: The sum of the bond length, bond angle, torsion angle, improper torsion angle, 1-4,1-5 van der Waals and hydrogen bond terms of the AMBER force-field, and the OONS hydration free energy, and the continuum electrostatic energy. B: The AMBER force-field energy with the dielectric constant $\epsilon = 2r$.

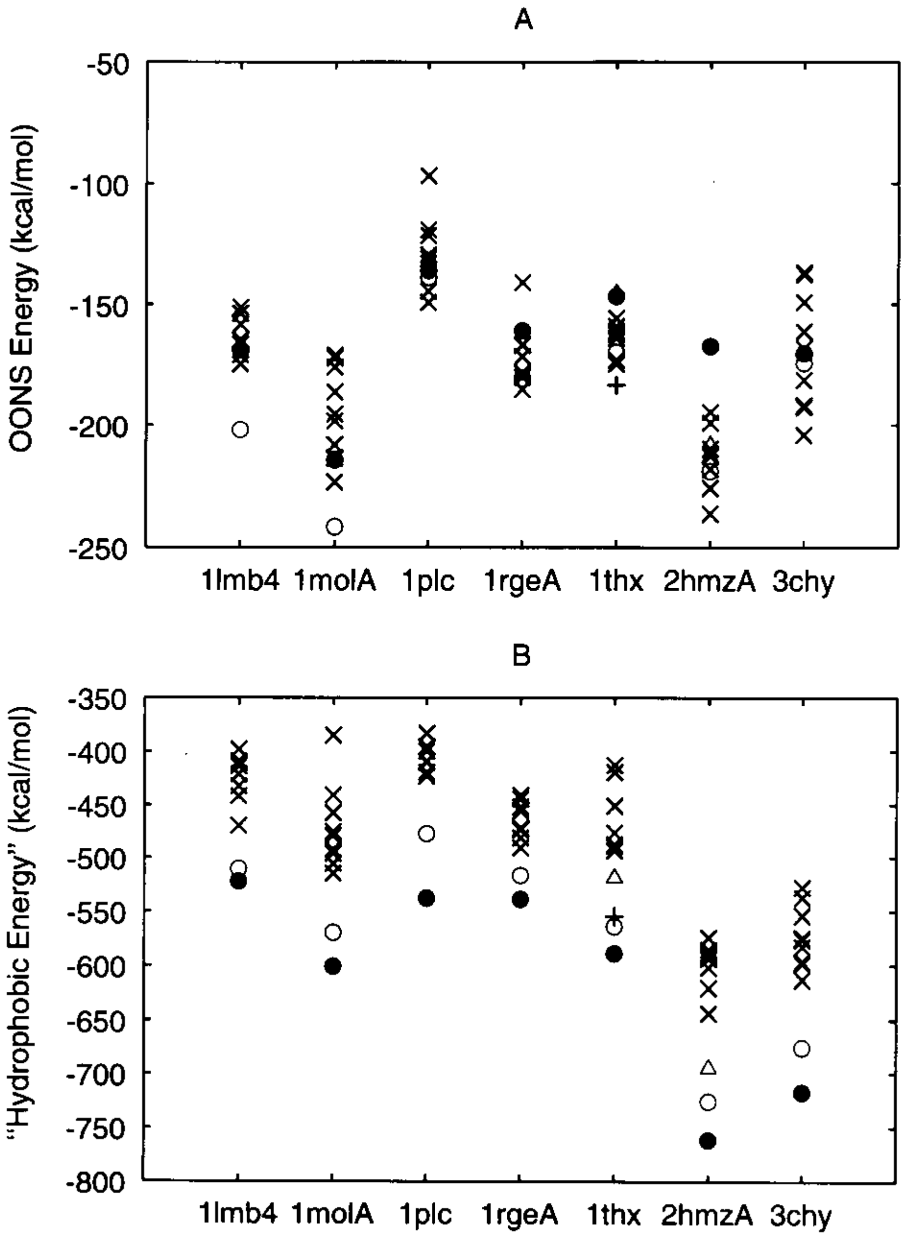


Figure 3.4: A: The hydration free energy. B: The "hydrophobic energy" which is the sum of van der Waals energy and hydration free energy. The symbols are defined as in Figure 3.3.

der Waals interactions in addition to the hydration effect. Therefore, I compared the sum of van der Waals energy and hydration free energy as the "hydrophobic energy". Figure 3.4B shows that all the correct structures can be readily discriminated from their misfolded counterparts. The difference of the hydrophobic energy between the native structure and near-native model is in most cases less than 50 kcal/mol, which is small compared to the difference of the total energy (compare Figure 3.4B with Figure3.3A). This fact suggests that the hydrophobic interaction stabilizes near-native structure more than the native structure itself.

The electrostatic energies E_{p-s} based on Equation (2.19) are shown in Figure 3.5A. Except for the target 2hmzA, the native structure has the lowest electrostatic energy and for most cases, the near-native model has the next lowest electrostatic energy. For the target 2hmzA, there are buried glutamates which bind to irons in the experimental structure. Since these irons are ignored in the calculation, the electrostatic energy of the native structure of 2hmzA is not the lowest. This exception demonstrates the importance of the Born energy which penalizes the buried charges. To see the importance of the solvent-shielding and the Born energy, I also examined the Coulomb energy as calculated with a distance-dependent dielectric constant, $\epsilon = 2r$ (Figure 3.5B). In this case, although all the native structures have the lowest Coulomb energy, the near-native models show the energy value close to, or even higher than, their misfolded counterparts. The stabilization by the solvent-shielding and the Born energy, in other words, the contribution of the solvent to the electrostatic energy, seems more important for the near-native models than for the native structures.

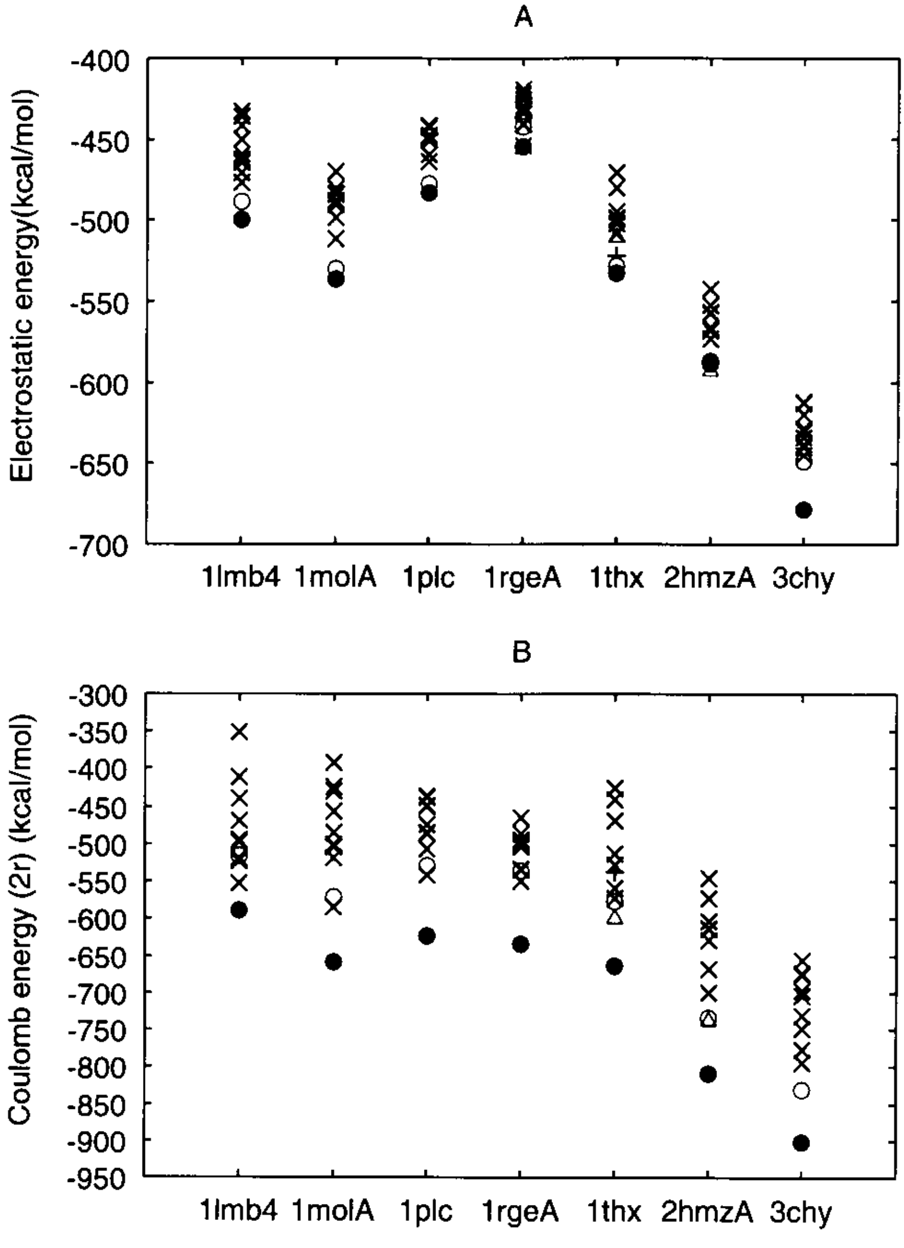


Figure 3.5: A: The electrostatic energy calculated by solving the Poisson equation of protein-solvent system (See the section 2.5). B: The electrostatic energy calculated with a distance-dependent dielectric constant ($\epsilon = 2r$). The symbols are defined as in Figure 3.3.

3.5 Side-chain packing

Vorobjev *et al.* [107] reported that "packing energy", that is, the sum of local geometry terms (bond lengths, bond angles and torsion angles) and van der Waals energy was in favor of the native structure. The present results are consistent with their observation (Figure 3.6A). The native structures have about 100 kcal/mol more stable packing energies than other models. This amount of stabilization is significant compared to the hydrophobic energy (Figure 3.4B) and electrostatic energy (Figure 3.5A). We also find that the near-native structures have lower packing energy than any other misfolded structures, but the energy difference is marginal in a few cases. While the native structures always have significantly low van der Waals energy, van der Waals energy alone cannot necessarily discriminate near-native from misfolded structures (Figure 3.6B). Therefore, local geometry and van der Waals energies are consistent only for the correct models, but this consistency alone is not enough to give the correct models significantly lower packing energy than the misfolded ones. In other words, the packing energy is a good index for discriminating the native structure, but it is not so for discriminating near-native structures. The native structure of a globular protein shows specific and close packing of side-chains [85]. The constructed near-native models have the side-chain conformations similar to the native structures (Table 3.4), but Figure 3.3 and Figure 3.6 show that the differences are significant from the energetics' point of view. Note that the difference of van der Waals energy between the native structure and near-native models are about 50 kcal/mol, which is small compared to the difference of the packing energy. This shows that not only the close packing, but also the less distorted local geometry contributes significantly to the stabilization of the native structure.

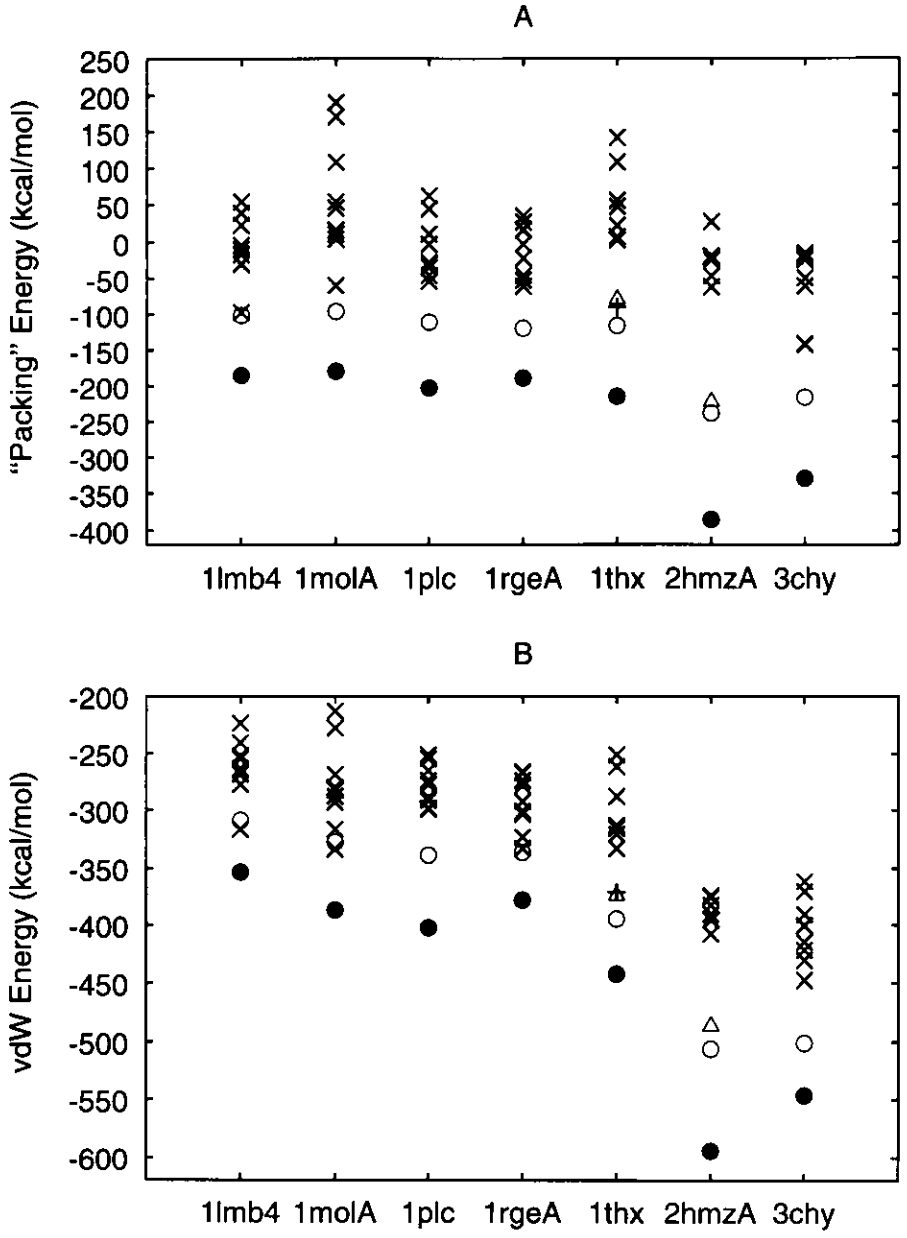


Figure 3.6: A: The "packing energy". In our case, the packing energy is defined as the sum of the terms for bond lengths, bond angles, proper and improper torsion angles, and 1-4 and 1-5 van der Waals interactions. B: The sum of 1-4 and 1-5 van der Waals energy terms. The symbols are defined as in Figure 3.3.

Janardhan and Vajda [36] reported that the solvation term was important, but the molecular mechanics energy was useless, for selecting near-native models with good packing. But their molecular mechanics energy did not include the van der Waals term, and their electrostatic energy did not incorporate the solvent shielding and the Born energy. Therefore, their result for the molecular mechanics energy is more or less similar to Figure 3.3B in our case. In the mean time, they used the atomic solvation parameters whose reference state was an organic solvent to complement the absence of the van der Waals term, whereas the reference state of the parameters by Ooi *et al.* [70] is the vacuum. Hence, their solvation free energy corresponds to the hydrophobic energy in our case (Figure 3.4B). The results of Janardhan and Vajda [36] are actually consistent with the present observation.

Sahasrabudhe *et al.* [89] applied a homology modeling method [52] to model the structure of an RNA-binding protein using two kinds of template proteins, a ferredoxin-like fold as the correct fold and cold shock protein A as an incorrect fold. They found that the former had a plausible negative value of the energy, while the latter showed an unrealistically high value [89]. In addition to the molecular mechanics energy in vacuum, they used structural restraints derived from both the backbone and side-chain conformations of the template. Consequently, the side-chains of their incorrect model were forced into unrealistic conformations, hence the unrealistically high energy. In our case, no restraints were imposed on side-chains. Therefore, the side-chains of the incorrect folds can adopt relatively relaxed conformation compared to the case of Sahasrabudhe *et al.* [89]. Nevertheless, the present result (Figure 3.6A) indicates that structures with good packing (i.e., the native structures) can be readily discriminated even in the absence of any artificial restraints for side-chain conformations.

3.6 Implications for structure prediction

I have shown that the stabilization by the solvent effects such as "hydrophobic energy", solvent shielding and the Born energy are good indexes for the discrimination of near-native models from misfolded ones. Also the homologous model 1thx-2trxA was found more stable than the misaligned model 1thx-1tof when solvent effects were included (Figures 3.3A, 3.4B and 3.5A). These results indicate that the solvent effects does not depend on structural details, but depend on more global features such as the pattern of hydrophobic and hydrophilic residues in the three-dimensional space. Hydrophobicity and the Born energy are effectively taken into account in the compatibility functions for threading [65]. This is one of the reasons for the success of threading in predicting approximately correct folds.

Although the hydrophobic energy can well discriminate the global fold of the native structure (Figure 3.4B), its contribution is rather small compared to that from the packing energy (Figure 3.6A). In fact, the large energy gap between the native structure and near-native model comes mainly from the specific side-chain packing (Figure 3.3A and Figure 3.6A). Consequently, the lack of detailed treatment of side-chain packing may be the reason for the false positives often found by threading. In fact, some attempts have been made that incorporate more or less detailed representation of side-chain packing into statistical potentials and their results show significant improvements in recognizing the native folds [56, 90]. However, since threading involves alignment of the target sequence to structures, exact modeling of side-chain conformations is in principle impossible. Also, it is difficult to model the side-chain conformations based on a distantly related template structure whose backbone conformation differs to some extent from that of the native structure of the target protein [14]. Therefore, we cannot adopt any conventional algorithms for side-chain packing prediction which require the rigorously fixed backbone conformation [51]. Instead, we have to allow the backbone to move to the extent at which correct side-chain packing can be achieved. The present study treats the backbone conformation restrained to a given template, but allows it more or less to move. In order to solve the more general side-chain packing problem allowing the backbone movement, and to reach the true native from a near-native structure, it will be necessary to employ powerful conformational sampling techniques such as generalized ensemble methods (e.g., [64, 99]). The computation might be accomplished by sufficient and adequate conformational sampling in the optimization process, as far as the native structure is located at the global minimum of the energy surface of the protein molecule.

3.7 Concluding remarks

For all of the seven protein sequences examined in the present study, the native conformation, minimized from the x-ray structure, was always the lowest in total energy among those conformations selected by threading. This fact, together with the test runs performed for decoy models (see Figure 2.2 and Section 2.4), validates the energy function employed. Misfolded conformations always have relatively higher energies, and correct models including the near-native and homologous models, situated inbetween the native and misfolded conformations (Figure 3.3A). The near-native model, reconstructed according to the native backbone as the template, has an almost identical backbone conformation to the native (Table 3.4), but different side-chain conformations which were attained by minimizing atomic overlaps. Therefore, a distinct energy difference between the native and the near-native conformations (Figure 3.3A) is mainly attributed to the difference of the side-chain conformations between them. However, the difference is little (Figure 3.5A) or small (Figure 3.4B) in electrostatic and hydrophobic energies, respectively, indicating that these energy terms are relatively insensitive to the side-chain conformation. On the other hand, the packing energy alone seems to yield the net change between the native and near-native structures (Figure 3.6A), although distinction of the near-native from misfolded models becomes somewhat unclear in the packing energy than in the total energy. Taking all these results into account, the energetic contributions to a native protein are summarized into two categories: one mainly depending on the topology or backbone conformation of a protein molecule (i.e., electrostatic and hydrophobic terms), and the other depending on the detailed side-chain conformation (i.e., packing energy). Importance of the both contributions to the protein stability delineates the limit of the threading treatment in which the side-chain is in principle simplified as a norm and therefore detailed side-chain packing must be totally neglected. Given an approximately correct topology (backbone conformation) alone, the next step for us to go is realizing the true native conformation for whole protein atoms. This would be one of the necessary steps toward *ab initio* predictions.

Chapter 4

Stabilization Mechanism of Thermophilic Proteins

4.1 Introduction

This chapter presents an application of the near-native structure as a useful tool to analyze protein stability. The stabilization mechanism of proteins from thermophilic organisms have attracted interests for many years. Recently, the accumulation of both sequence and structural data of thermophilic proteins as well as their mesophilic homologs has made it possible to study structural properties that are likely to stabilize the thermophilic proteins (e.g., [102, 45, 40]). Although comprehensive, most of these studies are limited to statistical analysis of some structural parameters such as amino acid compositions, the number of ion pairs, the number of cavities, etc. Thus, in these studies, the energetic bases of the thermostability were not directly addressed. However, since early statistical studies indicated the importance of ion pairs in some thermophilic proteins, the electrostatics of thermophilic proteins were studied intensively (e.g., [112, 22]). These studies confirmed the importance of the electrostatic interactions in thermophilic proteins although other energetic factors remained to be clarified. Lazaridis *et al.*[49] investigated the thermostability of the rubredoxin from the hyperthermophilic archaeon *Pyrococcus furiosus* and the one from mesophilic *Desulfovibrio vulgaris* by unfolding molecular dynamics simulations at various temperatures. Although the computational complexity of the molecular dynamics simulation inhibited extensive analysis to draw definitive conclusions, some interesting insights were obtained regarding the energetic factors and the rigidity of the hyperthermophilic protein [49].

In this chapter, I investigate various energy components of thermophilic and mesophilic proteins by the molecular mechanics method. In order to find the factors that are common to different thermophilic proteins, we have employed a computationally more feasible method than Lazaridis *et al.* [49], and have investigated five protein families of reasonable sizes. Basically, I have analyzed the energy difference between the native and unfolded structures of proteins. In addition, I also calculated the energy of artificially constructed near-native structures (Section 2.6). The near-native structures are so called because they have almost the same backbone conformation as the native one, but their side-chain packing is significantly distorted. From the study in Chapter 3, the near-native structures are shown to be useful to extract the dominant factors that stabilize the the native fold irrespective of the detailed side-chain packing. To summarize the results in Chapter 3, the sum of local geometry and van der Waals terms is the dominant factor that stabilizes the native structure which is defined by both the backbone topology and precise side-chain packing, whereas the electrostatic and hydration terms are the factors that stabilize the native topology or fold relatively independent of the details of packing (Chapter 3). I exploit the artificially constructed near-native structures to analyze in more detail the energy difference between the native and unfolded structures. Operationally, the energy difference between the native and unfolded structures can be divided into two parts: one between the native and near-native, the other between the near-native and unfolded structures. As will be shown later, this strategy made it possible to understand more clearly the stabilization mechanism of thermophilic proteins.

From statistical analyses, it is suggested that hyperthermophilic proteins from archaea are stabilized by different mechanisms than thermophilic eubacteria (Szilágyi and Závodszky 2000). Moreover, the distant evolutionary relationship between archaea and other organisms would make the comparison difficult. Therefore, I restrict the present analysis to the proteins from thermophilic eubacteria and their homologs from mesophilic organisms in the present study.

4.2 Energy function and decomposition of energy difference

The energy function for the final evaluation of structures is composed of the AMBER all-atom force-field [109] together with the hydration term of Ooi *et al.* [70] and the electrostatic contribution from the solvent. The electrostatic contribution from the solvent (reaction field) was calculated based on a continuum dielectric model of the protein-solvent system [66, 65]. The dielectric constants were set to 4 in the protein region, and to 80 for the solvent and boundary regions. The ionic strength was set to 0. The details of the calculation procedure are given in Chapter 2

In order to make the comparison of proteins of different sizes easier, the energy values were normalized by the molecular weight of each protein. Hence the energy unit cal/g is used instead of more often used kcal/mol.

Thus calculated energy values were analyzed as follows. Let E_n be the

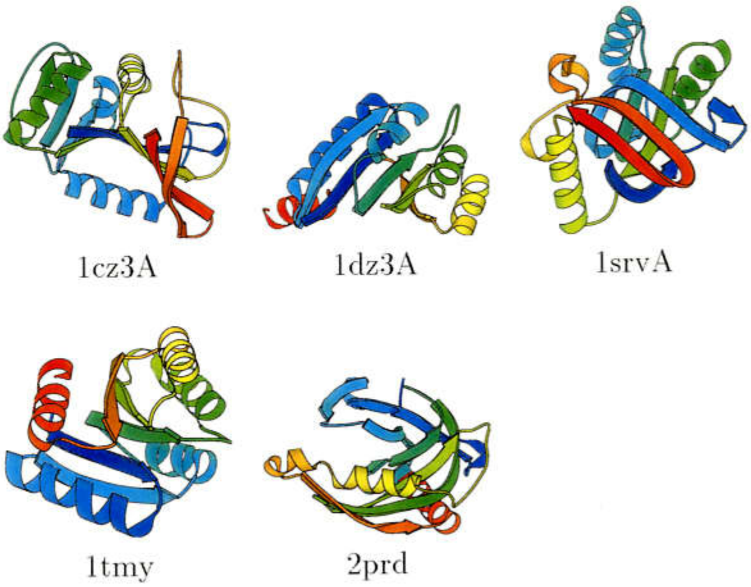


Figure 4.1: Structures of the thermophilic proteins.

energy of the native structure of a protein, and E_u be the average energy of unfolded structures of the same protein. The first quantity we investigate is the energy change between these states: $\Delta E_{n-u} = E_n - E_u$ which corresponds to the enthalpy change during folding transition. Next, we define E_m as the average energy of the near-native structures of the protein of interest, and $\Delta E_{n-m} = E_n - E_m$, $\Delta E_{m-u} = E_m - E_u$. Accordingly, we decompose ΔE_{n-u} into two terms: $\Delta E_{n-u} = \Delta E_{n-m} + \Delta E_{m-u}$. We can regard ΔE_{m-u} as the energy change associated with the formation of an approximate native fold, and ΔE_{n-m}, the one with the formation of specific packing to reach the precise native structure.

4.3 Preparation of structures

All the protein structures were extracted from the Protein Data Bank (PDB) [8]. I have selected proteins from thermophilic eubacteria and their mesophilic

Table 4.1: Target proteins from thermophilic bacteria.

PDB[a]	name	species	R[b]
1cz3A	Dihydrofolate reductase	*Thermotoga maritima*[c]	2.10
1dz3A	SPO0A	*Bacillus stearothermophilus*[d]	1.65
1srvA	GroEL apical domain	*Thermus thermophilus*[e]	1.70
1tmy	CheY	*Thermotoga maritima*[c]	1.9
2prd	Inorganic pyrophosphate	*Thermus thermophilus*[e]	2.0

[a]PDB codes and chain identifier. [b]Resolution of the crystal structure (Å). Optimum growth temperatures are [c]80 °C, [d]52.5 °C, [e]75 °C according to ref [102].

homologs that are solved by X-ray crystallography with 2.5 Å or better resolution, and are composed of less than 200 residues. Proteins from thermophilic eubacteria were selected by database search with the keyword 'THERM' in the SOURCE field of the PDB file. Thus found thermophilic proteins were used as queries for the BLAST [3] search against the PDB sequence database. Only those targets whose appropriate mesophilic homologs existed in the PDB were retained. After all, we obtained five thermophilic proteins (Figure 4.1 and Table 4.1) and their mesophilic counterparts (Table 4.2). Note that in the present study thermophilic proteins from archaea are not treated, but all the target thermophilic proteins are from eubacteria.

For the analysis of energy components, we calculated energies for three states of each protein: the native, near-native, and unfolded structures. The native structure was obtained from the X-ray structure and its energy was minimized by 500 steps of conjugate gradient method with the distance geometry force-field followed by 500 steps of conjugate gradient method the AMBER force-field without 1-5 electrostatic term to remove close contacts. The distance geometry force-field includes only the local geometry terms and non-local soft repulsion terms but no attractive term [63] (see also the sec-

Table 4.2: Homologs of the target proteins.

Thermophile[a]	Mesophile[a]	species	ID%[b]	R[c]
1cz3A (164)	1aoeA (192)	*Candida albicans*	52.7	1.60
	1dyjA (159)	*Escherichia coli*	53.5	1.85
1dz3A (123)	1srrA (119)	*Bacillus subtilis*	64.6	1.9
1srvA (145)	1kid (193)	*Escherichia coli*	82.6	1.7
1tmy (119)	2chf (128)	*Salmonella typhimurium*	60.7	1.8
	3chy (128)	*Escherichia coli*	59.5	1.66
2prd (174)	1obwA (175)	*Escherichia coli*	66.5	1.9

[a] PDB code and chain identifier with chain length in the parentheses. [b] Percent sequence identity with the thermophile. [c] Resolution of the crystal structure in Å.

tion 2.1). The positional restraints on the backbone atoms (Equation 2.20) were imposed so that the minimized structure did not deviate much from the experimental coordinates.

The procedure for generating the near-native structure is shown in Table 4.3. First a random coil was generated which was minimized for 500 steps of conjugate gradient method. Then the random coil was subject to 55000 steps of a simulated annealing molecular dynamics in four dimensional space using the computer program EMBOSS [63]. The temperature was set to 500 K for the first 5000 steps, then cooled exponentially to 1 K. Next, 3000 steps of conjugate gradient minimization was applied with an increased weight for the fourth dimensional energy to compress the fourth dimension and to obtain the three dimensional structure. Finally, 500 steps of conjugate gradient minimization with the AMBER force-field without the 1-5 electrostatic energy term completed the generation of a near-native structure. Except for the final stage, the distance geometry force-field was used. Throughout the stages, the positional restraints (Equation 2.20) were imposed on the backbone atoms

Table 4.3: Generation of near-native structure[a]

A random coil generated as the initial structure				
stage	N steps[b]	protocol[c]	FF[d]	k_{4D}[e]
1	500	mini	DG	0.05
2	5000	MD (T = 500 K)	DG	0.05
3	50000	SAMD (T = 500 → 1 K)	DG	0.05
4	3000	mini	DG	10.0
5	500	mini	Am	∞

[a]Positional restraints on the backbone atoms are imposed throughout the stages. [b]Number of optimization steps. [c]Optimization protocol: mini, conjugate gradient minimization; MD, molecular dynamics at constant temperature; SAMD, simulated annealing molecular dynamics (temperature is decreased exponentially). [d]Force-field: DG, the distance geometry force-field; Am, the AMBER force-field without 1-5 Coulomb term. [e]The weight of 4-th dimensional energy.

so that the near-native model has the same backbone structure as the native structure. Since the main optimization stages employ the distance geometry force-field which does not include any attractive term, side-chain conformations are determined solely by steric hindrance. Other characteristics of the near-native structure are presented in Chapter 3. Twenty near-native structures were generated with different initial conditions (random coils and initial velocities) for each protein and the average energy and energy components were used for the analysis given below.

The unfolded structure was generated by almost the same procedure as the near-native structure (Table 4.4). The differences are that the starting structure was the native structure, that the simulated annealing molecular dynamics was performed in the three dimensional space, that the minimization following the simulated annealing lasted only for 1000 steps, and that the restraints were imposed so that the structure deviated from the native one by the root mean square deviation (RMSD) between 8 Å and 12 Å [23].

Table 4.4: Generation of unfolded structure[a]

The native structure is given as the initial structure				
stage	N steps	protocol	FF	k_{4D}
1	500	mini	DG	∞
2	5000	MD (T = 500 K)	DG	∞
3	50000	SAMD (T = 500 → 1 K)	DG	∞
4	1000	mini	DG	∞
5	500	mini	Am	∞

[a] RMSD restraints (Equation 4.1) were imposed on the backbone atoms throughout the stages.

The RMSD restraint [23] is given as a set of penalty functions E_{RMSD}:

$$E_{RMSD} = \begin{cases} \frac{1}{2}k_B T_0 M(\rho_l - \rho(\{\mathbf{r}\}))^2 & (\text{for } \rho < \rho_l) \\ 0 & (\text{for } \rho_l \leq \rho \leq \rho_u) \\ \frac{1}{2}k_B T_0 M(\rho_u - \rho(\{\mathbf{r}\}))^2 & (\text{for } \rho > \rho_u) \end{cases} \tag{4.1}$$

where k_B is the Boltzmann constant, T_0 is an absolute temperature set to 300 K, M is the number of backbone atoms, and ρ_l and ρ_u are the lower and upper limit of the instantaneous backbone RMSD $\rho(\{\mathbf{r}\})$ between the unfolded and native structures. In the present case, ρ_l and ρ_u are set to 8 Å and 12 Å, respectively. $\rho(\{\mathbf{r}\})$ is given by

$$\rho(\{\mathbf{r}\}) = \sqrt{\frac{\sum_i{}' |\mathbf{r}_i - \mathbf{r}_i^1|^2}{M}} \tag{4.2}$$

where $\mathbf{r}_i$ and $\mathbf{r}_i^1$ are the atomic coordinates of the i-th atom of the unfolded and native structures, respectively. The coordinates of the native structure is superposed onto the instantaneous unfolded structure by the method of Diamond [19] every time E_{RMSD} is calculated so that the coordinate deviation arising from the translational and rotational degrees of freedom is removed. Thus constructed unfolded structure shows a global chain topology somewhat similar to the native structure, but is significantly expanded and contains no residual secondary structures. An example is shown in Figure 4.2. Again,

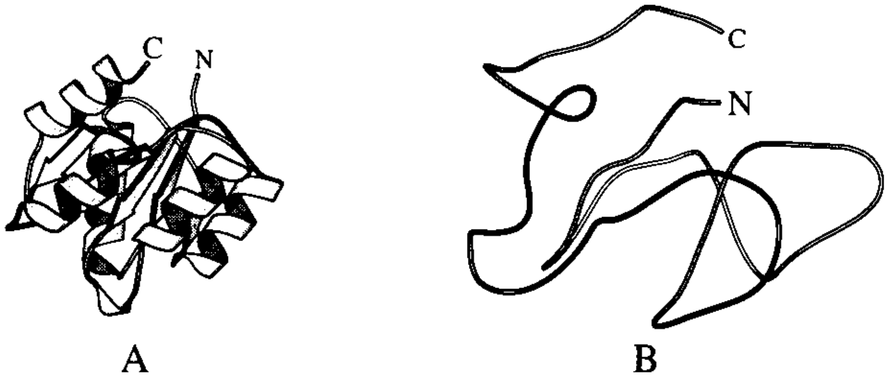


Figure 4.2: An example of the unfolded structure. A: the native structure of the PDB entry 1tmy [105]. B: An unfolded structure of 1tmy. The amino and carboxyl termini are marked with N and C, respectively. The figure was drawn with MolScript [44].

twenty structures were generated with different initial velocities for each protein and the average energy and its components were used for the analysis below.

4.4 Energy difference between the native, near-native and unfolded structures

Table 4.5 summarizes the average radius of gyration (R_g) calculated for all heavy atoms of each protein together with RMSD of near-native and unfolded structures from the native one. It can be seen that the near-native structures have slightly larger R_g because of their imperfect packing. The unfolded structures show much larger R_g than the native structure indicating that they are indeed unfolded. The RMSD of the unfolded structures are all close to 12 Å which is the upper limit of the RMSD restraints imposed (see above).

Table 4.6 shows ΔE_{n-u} (the energy difference between the native and unfolded structures) and its components, and some combinations of the com-

Table 4.5: Radius of gyration calculated for all heavy atoms (Å).

protein[a]	native	near-native[b]	unfolded[c]
t1cz3A	15.3	15.5 (0.16, 1.81)	22.5 (11.8)
m1aoeA	16.2	16.3 (0.14, 1.88)	23.4 (11.9)
m1dyjA	15.1	15.3 (0.17, 1.99)	21.9 (11.7)
t1dz3A	16.5	16.6 (0.11, 1.66)	22.9 (11.7)
m1srrA	13.0	13.1 (0.16, 1.69)	20.5 (11.8)
t1srvA	13.8	13.9 (0.15, 1.59)	21.1 (11.7)
m1kid	16.5	16.6 (0.13, 1.60)	23.9 (11.8)
t1tmy	12.7	12.8 (0.14, 1.52)	20.3 (11.7)
m2chf	13.1	13.4 (0.13, 1.76)	21.0 (11.8)
m3chy	13.1	13.3 (0.12, 1.65)	21.0 (11.8)
t2prd	15.0	15.2 (0.15, 1.75)	22.3 (11.9)
m1obwA	15.1	15.2 (0.13, 1.74)	22.2 (11.8)

[a] The PDB code and chain identifier with the first letter "t" or "m" indicating that the protein is either thermophilic or mesophilic, respectively. [b] The numbers in the parentheses are RMSD (Å) from the native structure; the first number was calculated for backbone atoms, the second for all heavy atoms. [c] RMSD (Å) from the native structure calculated for backbone atoms.

ponents. Thermophilic proteins show lower total energy changes than their mesophilic counterparts in four out of five groups of proteins (the only exception is 1tmy). To the contrary, only one thermophilic protein, 2prd, shows lower energy in vacuo than its mesophilic homolog. These observations suggest that the solvent effect may be crucial for the stabilization of thermophilic proteins. The electrostatic energy has been shown important for the stabilization of thermophilic proteins [112]. In our case, the electrostatic energy is defined as the sum of the Coulomb and reaction field terms in Table 4.6. Although most of the thermophiles show lower electrostatic energy than their mesophilic homologs, 1cz3A does not. However, ΔE_{n-u}(EH) (the sum of the electrostatic and hydration terms in ΔE_{n-u}) is consistently lower for all the thermophiles than in the mesophiles. This observation confirms the

Table 4.6: ΔE_{n-u} and its components (cal/g)[a]

protein[b]	tot	loc	vdW	Cou	HB	Oo	Rea	vac	pac	EH
t1cz3A	**-13.53**	**0.05**	-28.91	-18.03	**-2.65**	**8.29**	27.72	-49.54	**-28.85**	**17.98**
m1aoeA	-10.30	0.83	-28.16	-25.08	-2.49	10.60	34.00	-54.90	-27.33	19.52
m1dyjA	-10.81	0.40	-29.10	-17.41	-2.56	11.11	26.76	-48.68	-28.70	20.45
t1dz3A	**-15.79**	**0.25**	-25.43	-12.68	-2.61	**6.01**	**18.67**	-40.47	-25.18	**12.00**
m1srrA	-14.88	1.94	-29.68	-19.56	-2.86	7.61	27.67	-50.16	-27.74	15.72
t1srvA	**-15.57**	1.11	**-29.23**	**-21.66**	-2.69	**7.03**	**29.87**	-52.47	-28.12	**15.24**
m1kid	-14.72	-0.50	-29.08	-21.64	-2.90	8.79	30.61	-54.12	-29.58	17.76
t1tmy	-15.60	1.17	-27.66	**-25.35**	-3.03	**7.03**	32.25	-54.88	-26.49	**13.93**
m2chf	-12.87	0.65	-30.89	-22.41	-3.05	9.16	33.66	-55.69	-30.24	20.41
m3chy	-18.04	-1.18	-32.03	-21.72	-3.23	9.19	30.93	-58.16	-33.21	18.40
t2prd	**-12.36**	**-0.39**	-30.29	**-22.26**	**-2.76**	10.16	33.18	**-55.70**	**-30.68**	**21.08**
m1obwA	-8.19	-0.22	-30.35	-9.36	-2.53	9.86	24.42	-42.46	-30.57	24.92

[a] The energy value of thermophiles is typed in **boldface** if it is lower than the corresponding values of any other mesophiles. [b] See the caption of Table 4.5. The notation of the energy components are the following: tot, total energy; loc, local energy which is the sum of bond length, bond angle, torsion, and improper torsion terms; vdW, the sum of 1-4 and 1-5 van der Waals terms; Cou, the sum of 1-4 and 1-5 Coulomb interaction terms; HB, hydrogen bond term; Oo, hydration free energy term of Ooi *et al.* [70]; Rea, electrostatic contribution from the solvent, i.e. reaction field energy; vac, total energy in vacuum, which is equal to tot - Oo - Rea; pac, packing energy which is the sum of local and vdW terms; EH, the sum of Cou, Rea and Oo terms.

importance of the solvent effect.

In Chapter 3, I have shown that the "packing energy", i.e., the sum of bond length, bond angle, torsion angle and van der Waals terms [107], is the dominant factor for determining the precise native structure. Table 4.6 shows that the packing energy, ΔE_{n-u}(pac), of each protein is of a large negative value but only two thermophiles, 1cz3A and 2prd, show lower ΔE_{n-u}(pac) than their mesophilic counterparts. Furthermore, the difference of the ΔE_{n-u}(pac) of these two thermophiles from their corresponding homologs are small (-0.1 and -0.15 cal/g, respectively) compared to the

Table 4.7: ΔE_{m-u} and its components (cal/g)[a]

protein	tot	loc	vdW	Cou	HB	Oo	Rea	vac	pac	EH
t1cz3A	**-3.19**	5.41	**-23.13**	-13.44	**-2.30**	**6.87**	23.39	-33.46	-17.72	**16.83**
m1aoeA	-0.66	5.75	-22.43	-17.27	-2.04	8.65	26.69	-35.99	-16.68	18.07
m1dyjA	-2.33	5.01	-22.81	-12.07	-2.23	8.58	21.19	-32.11	-17.80	17.70
t1dz3A	**-8.59**	**3.29**	-21.75	-9.93	-2.39	**4.86**	**17.33**	-30.78	-18.46	**12.26**
m1srrA	-5.62	4.84	-25.32	-13.97	-2.48	6.47	24.83	-36.92	-20.47	17.33
t1srvA	-5.30	5.41	-24.94	-16.21	-2.53	**6.67**	**26.30**	-38.28	-19.53	**16.76**
m1kid	-6.30	4.63	-25.53	-17.64	-2.66	7.68	27.23	-41.20	-20.90	17.27
t1tmy	**-8.03**	**3.13**	-24.85	**-17.79**	-2.69	6.52	27.64	**-42.19**	-21.72	**16.37**
m2chf	-5.85	4.83	-25.52	-12.52	-2.60	6.10	23.87	-35.81	-20.69	17.44
m3chy	-6.90	4.07	-26.29	-14.07	-2.73	7.01	25.12	-39.02	-22.22	18.05
t2prd	**-2.19**	4.70	-24.58	**-14.07**	**-2.39**	**7.32**	26.82	**-36.33**	-19.88	**20.07**
m1obwA	-0.24	3.67	-24.67	-8.82	-2.23	8.15	23.65	-32.04	-20.99	22.98

[a] See Table 4.6 for explanation of the columns.

difference of ΔE_{n-u}(EH) (-1.5 to -3.8 cal/g, respectively). Thus, the relative stability of the thermophilic proteins are dominated by the difference in ΔE_{n-u}(EH). The electrostatic and hydration energies have been shown to be important for determining the native fold regardless the specific packing, rather than the native structure itself (Chapter 3). Therefore, the trend of ΔE_{n-u}(pac) and ΔE_{n-u}(EH) in Table 4.6 suggests that the thermophilic proteins are stabilized irrespective of detailed packing. This point is further discussed in the following paragraphs.

Table 4.7 shows ΔE_{m-u} (the energy difference between the near-native and unfolded structures). Also shown are their components and some combinations of the components. Thermophilic proteins from four out of five groups show lower total ΔE_{m-u} values than the corresponding mesophiles. The thermophile 1tmy has lower ΔE_{m-u} than any other homologs although it has higher ΔE_{n-u} than its mesophilic homolog 3chy (Table 4.6). The opposite trend is found for 1srvA which has higher ΔE_{m-u} than 1kid. As implied in the analysis of ΔE_{n-u}, ΔE_{m-u}(EH) is consistently lower for the thermophiles than for the mesophiles. None of the thermophiles shows ΔE_{m-u}(pac) lower than the mesophiles. Therefore the relative stability of the thermophiles associated with the formation of the fold is also dominated by ΔE_{m-u}(EH).

The stabilization by the formation of the specific packing can be seen in ΔE_{n-m} (Table 4.8). The stability of thermophilic proteins are not so conspicuous in ΔE_{n-m} as in ΔE_{n-u} and ΔE_{m-u}. Three out of five thermophiles show lower total ΔE_{n-m} than their mesophilic homologs. Only two thermophiles, 1cz3A and 2prd, show lower ΔE_{n-m}(pac) than their mesophilic homologs. The ΔE_{n-m}(pac) of thermophilic 1tmy is significantly higher than those of 2chf and 3chy. This observation shows that although good packing stabilizes some thermophilic proteins, it is not the universal mechanism for the thermostability of thermophiles, which confirms the statistical analysis by Karshikoff and Ladenstein [40]. ΔE_{n-m}(EH) are lower for the thermophiles than for the mesophiles, except for 1dz3A. The ΔE_{n-m}(EH) of 1dz3A is less stable than the ΔE_{n-m}(EH) of 1srrA by 1.35 cal/g. But this difference is overwhelmed by the difference in ΔE_{m-u}(EH) (-5.07 cal/g).

The side-chain packing of the near-native structures are largely distorted, and the electrostatic and hydration energy components are known to stabilize the correct fold even in absence of the detailed packing (Chapter 3). Thus we can conclude that the thermophilic proteins are stabilized by the force that stabilizes the correct fold or topology rather than by the force that stabilizes the detailed packing.

Table 4.8: ΔE_{n-m} and its components (cal/g)[a]

protein	tot	loc	vdW	Cou	HB	Oo	Rea	vac	pac	EH
t1cz3A	**-10.34**	**-5.36**	**-5.78**	-4.59	-0.35	**1.41**	**4.33**	-16.08	**-11.14**	**1.15**
m1aoeA	-9.64	-4.93	-5.72	-7.81	-0.45	1.95	7.30	-18.90	-10.65	1.45
m1dyjA	-8.48	-4.61	-6.29	-5.34	-0.33	2.53	5.57	-16.57	-10.90	2.76
t1dz3A	-7.20	**-3.04**	-3.68	-2.75	-0.22	1.15	**1.34**	-9.68	-6.72	-0.26
m1srrA	-9.26	-2.90	-4.36	-5.59	-0.38	1.14	2.84	-13.24	-7.26	-1.61
t1srvA	**-10.27**	-4.30	**-4.28**	**-5.45**	-0.16	**0.35**	3.57	**-14.19**	-8.58	**-1.52**
m1kid	-8.42	-5.13	-3.55	-4.00	-0.23	1.11	3.38	-12.92	-8.68	0.50
t1tmy	-7.56	-1.97	-2.81	-7.56	-0.35	**0.51**	**4.61**	-12.68	-4.78	**-2.44**
m2chf	-7.03	-4.18	-5.37	-9.89	-0.45	3.06	9.80	-19.88	-9.55	2.97
m3chy	-11.15	-5.25	-5.74	-7.65	-0.50	2.18	5.81	-19.14	-11.00	0.34
t2prd	**-10.17**	**-5.09**	**-5.72**	**-8.19**	**-0.37**	2.84	6.36	**-19.37**	**-10.81**	**1.01**
m1obwA	-7.95	-3.89	-5.69	-0.54	-0.30	1.70	0.77	-10.42	-9.58	1.94

[a] See Table 4.6 for explanation of the columns.

4.5 Discussion

From Tables 4.6 and 4.7, we can see that ΔE_{n-u}(pac) and ΔE_{m-u}(pac) are of large negative values whereas ΔE_{n-u}(EH) and ΔE_{m-u}(EH) are of large positive values. This means that while the packing energy stabilizes the native and near-native structures, the EH energy destabilizes them. Therefore, the conspicuous tendency that ΔE_{n-u}(EH) and ΔE_{m-u}(EH) are always lower for the thermophilic proteins than for the mesophilic ones indicates that the relative energetic stability of the thermophilic proteins are due to destabilization of their unfolded state. This tendency is not seen in ΔE_{n-m}(EH) and not all the thermophiles show lower ΔE_{n-m}(EH). Therefore the main difference of the thermophilic from the mesophilic proteins resides in between the near-native and unfolded states. The near-native structures are a set of hypothetical structures which represents a set of structures somewhere between the native and unfolded states, in which overall chain topology or fold is well formed but the side-chain packing is not yet fully achieved. On one hand, since the overall native topology is well formed in the near-native structure, there should be a large entropy loss with respect to the unfolded structures. On the other hand, the energetic stabilization of the native structure is mostly attained by the precise packing, ΔE_{n-m}(pac), which is not yet realized in the near-native structure. Therefore the near-native structures would be thermodynamically unstable. These characteristics of the near-native structures are qualitatively similar to the transition-state structure of proteins [24, 95]. If we assume that the near-native structures are a representative set of the transition-state structures, then the stability of the thermophilic proteins can be explained as follows (Figure 4.3).

First, the less positive values of ΔE_{m-u}(EH) of the thermophilic proteins (Table 4.7) indicate less stable unfolded state, hence lower $\Delta G_f^\ddagger$ (the activation free energy of folding). Second, since there is not a significant difference in ΔE_{n-m} or ΔE_{n-m}(pac) (Table 4.8), $\Delta G_u^\ddagger$ (the activation free energy of unfolding) should not be different between thermophiles and mesophiles. These result in larger negative value of overall folding free energy ΔG of the thermophilic proteins. This scheme for the stabilization of thermophilic proteins implies that these proteins fold faster. The faster folding of thermophilic proteins is actually suggested by a study of amino acid compositions (S. Fukuchi and K. Nishikawa, unpublished results).

That the thermophilic proteins show less stable ΔE_{m-u}(pac) suggests their near-native structures structurally more flexible or entropically more stable compared to the mesophilic proteins. The entropic stabilization of a thermostable protein was observed by the hydrogen exchange and neutron scattering experiments [27]. At a glance, this conclusion contradicts the results of Lazaridis *et al.* [49]. However, the protein they investigated was from

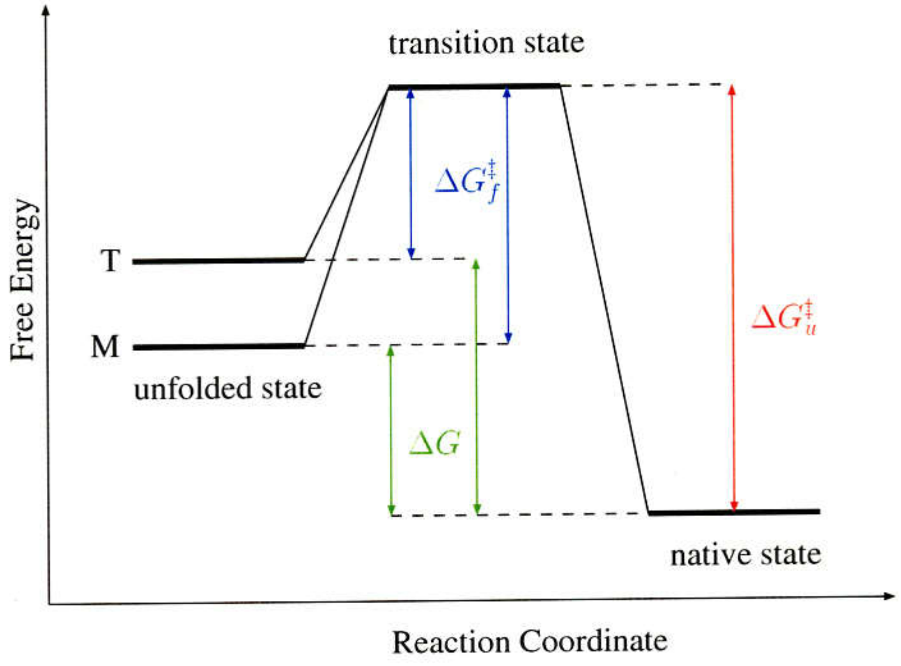


Figure 4.3: The suggested stabilization mechanism of thermophilic proteins. $\Delta G_f^\ddagger$, the activation free energy of folding; $\Delta G_u^\ddagger$, the activation free energy of unfolding; ΔG, the free energy of folding and unfolding. "T" and "M" stand for the unfolded state of thermophilic and mesophilic proteins, respectively.

an archaeon. It is suggested that one of the factors that stabilize the proteins from archaea is a better packing, but it is not the case for the proteins from thermophilic eubacteria (Szilágyi and Závodszky 2000). In any case, the present conclusion is limited to the proteins from thermophilic eubacteria. Since the precise side-chain packing will be difficult to maintain at a high temperature because of the thermal fluctuation, it may be possible that the thermophilic proteins have evolved so that they can rapidly recross the transition state to recover the native state. Perl *et al.* [78] compared the folding and unfolding kinetics of thermophilic cold shock proteins with their mesophilic homolog. In that study, the large difference between the thermophiles and mesophile was observed in the unfolding rather than in the folding, which contradicts our prediction. However, the transition state of a cold shock protein is known to be extremely anomalous [79]. Clarke *et al.* [15] studied the folding process of divergent proteins sharing a common fold, namely the immunoglobulin-like β-sandwich fold. They found that there was a strong correlation between ΔG_{N-D} (N and D stand for the native and denatured state respectively) and the refolding rate whereas no correlation was found between ΔG_{N-D} and the unfolding rates [15]. Although the proteins Clarke *et al.* [15] studied are all presumably mesophilic, the mechanism as depicted in Figure 4.3 may be present. Similar experimental studies on other proteins are required to validate the results of this chapter.

Chapter 5

Energetics of Protein Fold: Conclusion

5.1 The role of solvent effect

It was shown in Chapter 3 that the electrostatic and hydrophobic energies are the main contributors to the discrimination of the correct (native) fold. Both of these energy terms include the solvent effect, namely, the reaction field energy in the former and the hydration (free) energy in the latter. To summarize, the solvent effect stabilizes the correct fold, whereas the side-chain packing stabilizes the native structure. This trend is schematically shown in Figure 5.3.

It should be noted, however, that the "pure" solvent effect does not stabilize the native fold. In Figure 3.4A, the hydration energy (OONS) is not always lowest for the near-native structure. Also, we can see from Table 4.6 that the contributions from the reaction field and hydration actually destabilize the native fold with respect to the unfolded structures. The native fold and structure are undoubtedly stabilized by the intramolecular attractions in the protein. The intramolecular attractions are the Coulomb and van der Waals interactions in terms of the present study. The magnitudes of these interactions are quite large as can be seen in Figures 3.5B (page 55) and 3.6B (page 57). As a result, the native structure of a protein is expected to be extremely stable *in vacuo*. However, in the absence of the precise side-chain packing of the native structure, the intermolecular attractions by themselves are not able to discriminate the correct fold, they seems to stabilize any compactly folded structures, even misfolded ones (Figures 3.5B and 3.6B on pages 55 and 57, respectively). The fact that adding the solvent contributions to these attractions improves the discrimination of the correct fold (Figures 3.5A and 3.4B on pages 55 and 53, respectively), together with the fact that the solvent effect actually destabilizes compact structures (Tables 4.6 and 4.7 on pages 73 and 74, respectively), suggests that the solvent effect helps folding of the protein by destabilizing incorrect folds to a larger extent than the correct folds (Figure 5.1).

5.2 The role of side-chain packing

The results obtained in Chapter 3 clearly shows that the packing energy, which is the sum of local geometry and van der Waals terms, stabilizes the native structure to such a great extent that the total energy difference between the native and misfolded structures (Figure 3.3A on page 52) may be almost completely attributed to the difference in the packing energy (Figure 3.6A, page 57). This conclusion is illustrated in Figure 5.2.

In contrast with the solvent effect, we can see from Table 4.6 on page 73 that the packing energy actually stabilizes the native structure with respect to the unfolded structure. As discussed in the subsection 1.4.3, a good side-chain packing is characterized by both a large number of atomic contacts and less distorted local geometries. In the unfolded structure, it is easy for local structures to satisfy their ideal geometries because the influence from other

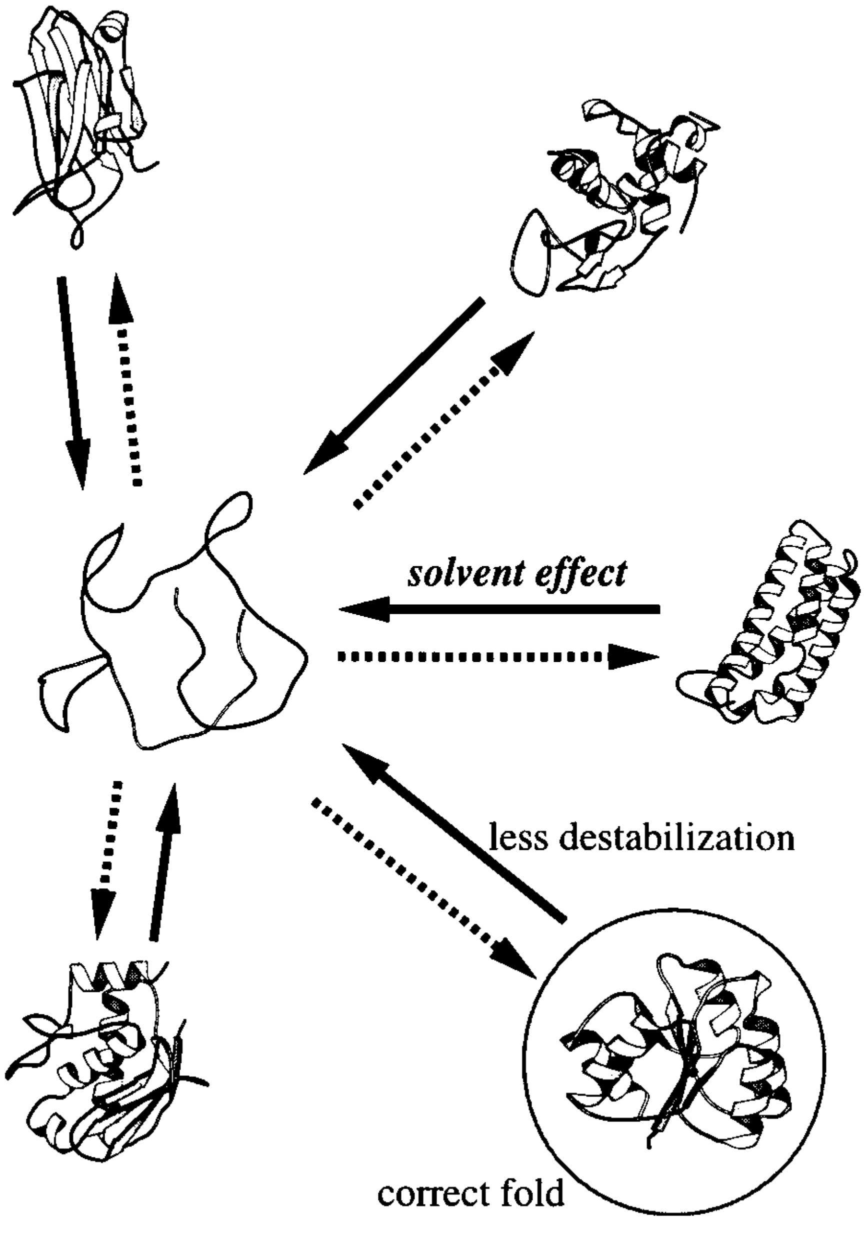


Figure 5.1: A schematic picture of how solvent effect stabilizes the correct fold. The intramolecular attractive forces (dotted arrows) stabilizes compact structures while the solvent effect (solid arrows) favors the unfolded structure. The correct fold (in the circle) is less destabilized than other compact folds.

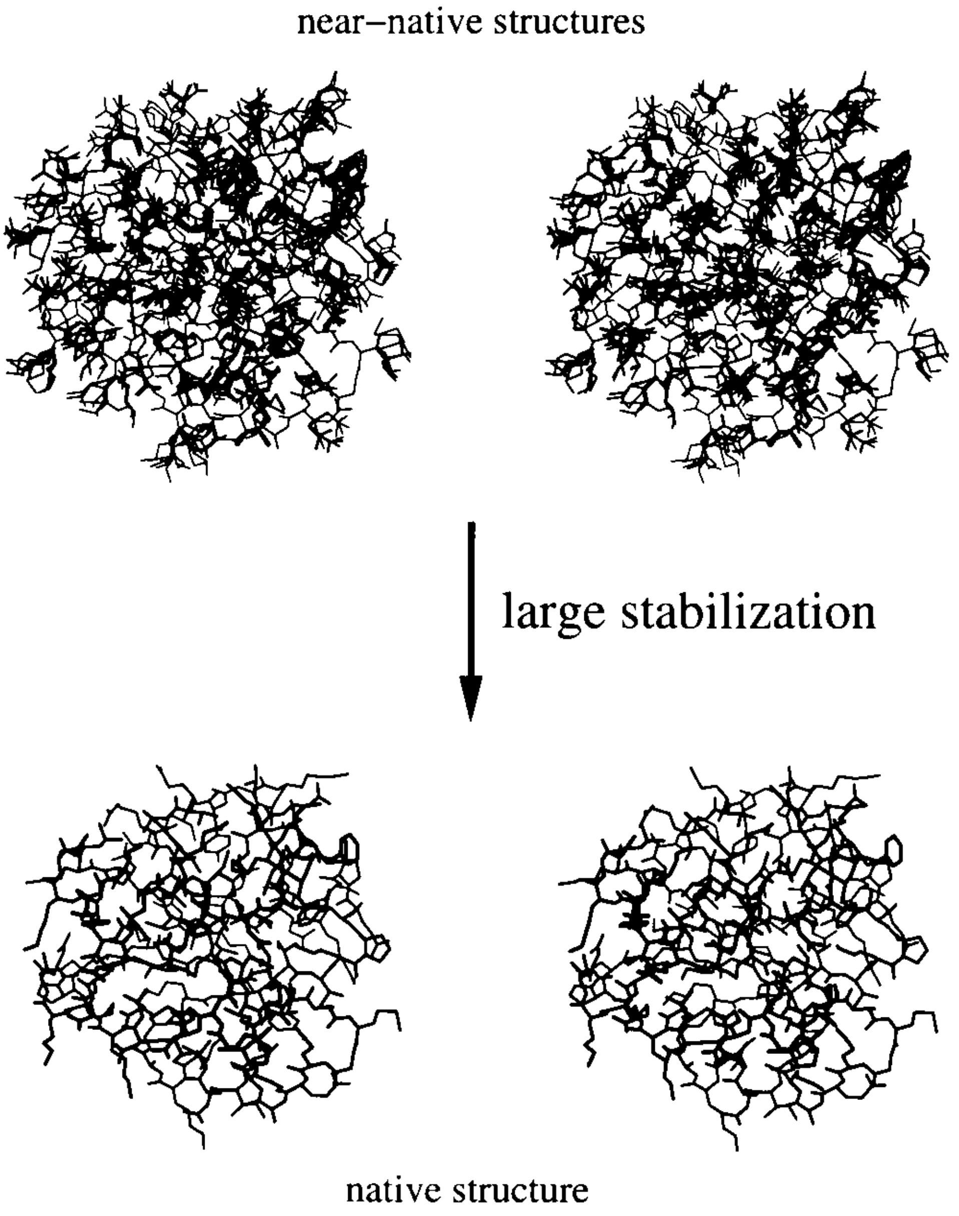


Figure 5.2: The side-chain packing significantly affects the stability of the native structure. Small differences in the side-chain packing may accompany a large energy difference. The upper figure is a stereo diagram of many near-native structures (side-chain bonds only) superposed on the native structure (the lower figure also, all bonds shown) of the PDB entry 1tmy [105]. This figure was drawn with MolScript [44].

non-local interactions is expected to be minimal in the unfolded structure. The small absolute values of ΔE_{n-u}(loc) in Table 4.6 indicate that the energy value associated with local geometry in the native structure is actually not so much different from those of the unfolded structures. On the other hand, the values of ΔE_{n-m}(loc) in Table 4.8 on page 76 are rather large, indicating that the local structures in the near-native structures are significantly distorted from the ideal geometry. The destabilization caused by these distortions are indeed comparable to that caused by insufficient atomic contacts which is associated with ΔE_{n-m}(vdW) in Table 4.8. These observations show that the local and non-local interactions in the native structure are consistent with each other. This is nothing but a concrete example of the consistency principle proposed by Gō [29] (see page 10).

5.3 Summary of energetics of protein fold

Thus, the major conclusion of the present study is that the energetic determinants of the native structure of proteins can be divided into two categories (Figure 5.3): one is attributed to the solvent effect which stabilizes the native fold (or approximate native structure), the other is associated with the side-chain packing which significantly stabilizes the native structure itself. As a consequence, it became possible to compare the relative contributions of these two categories of energy components. It was shown that the contribution of the solvent effect is small compared to that of the side-chain packing.

At this point, it is interesting to relate the above conclusions to the structural properties shared by the proteins of the same fold. Russell and Barton [87] compared structural properties such as side-chain to side-chain contacts, secondary structures, and residue-based solvent accessibility in pro-

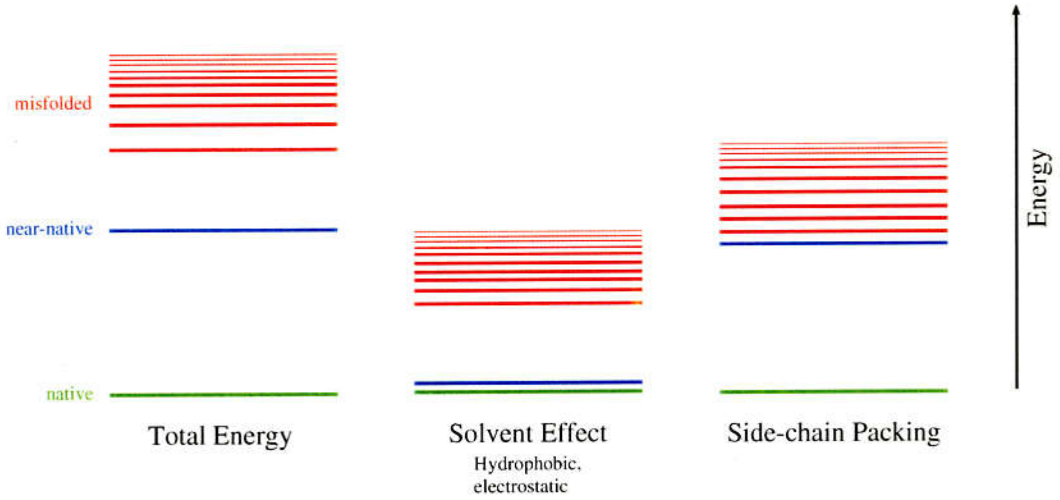


Figure 5.3: Summary of the energetics of protein fold. The solvent effect stabilizes the correct fold, the side-chain packing stabilizes the native structure. The stabilization by the latter is comparable to the total stabilization of the native structure.

teins sharing common folds. They found that virtually none of the specific structural properties were conserved in proteins with the same fold. How well a particular sequence adopts a particular fold was found to be governed by general features such as "hydrophobic residues buried in the core of proteins, and polar residues on the surface" [87]. This finding means that detailed structural properties such as side-chain to side-chain contacts are the particular characteristics of a particular native structure, whereas general structural properties such as buried hydrophobic residues and exposed polar residues are the general characteristics of folds. The results of Russell and Barton [87] exactly correspond to the conclusion of this thesis. The structural and energetic properties of the factors that determine the native fold and native structure are summarized in Table 5.1.

Table 5.1: Summary of structural and energetic determinants of the native fold and structure.

native structure

Structural: Tight and precise side-chain packing.

Energetic: Less distorted local geometry, and many favorable van der Waals contacts.

Remark: Extensive stabilization and specificity attained.

native fold

Structural: Buried hydrophobic and exposed hydrophilic residues.

Energetic: Reaction field, hydration together with intramolecular attractions.

Remark: Only loosely specific, selected by less destabilization relative to misfolds.

5.4 On the use of the near-native structure

Throughout the thesis, the key idea is the use of the near-native structure. As already mentioned in Chapter 2, a near-native structure is an artificial object *in silico* which is unlikely to exist in nature. A question may arise whether it is appropriate to use such an object to study the energetics of proteins. The answer to this question is a pragmatic one.

The very definition of protein fold is subjective. It is based more on the human way of perception than on the physical reality. In this sense, the energetics of protein fold is an attempt to interpret the abstract concept in terms of physicochemical interactions. Therefore, as far as protein folds are related to energy functions, the use of the near-native structure should be justified.

Furthermore, in Chapter 4, the use of the near-native structure actually helped interpreting the stabilization mechanism of thermophilic proteins. In such a study, the basic method will be to compare the energy difference between the native and unfolded structures which is all the information we can obtain. By using the near-native structure, we were able to extract more information about not only energetic factors, but also structural factors regarding the thermostability of the proteins. This, in turn, was possible because the relationship between the protein structure and energy components has been elucidated as summarized in Figure 5.3 and Table 5.1. The use of the near-native structure thus provide us with a wealth of information regarding structures and energetics.

Although the near-native structure has been proved useful, we must be cautious of their physical meaning. In Chapter 4, the near-native structure was related to the transition state of the protein folding process by a qualitative argument. It is only qualitative. At present, by no means the near-native structure can be quantitatively related to any physical existence. Some modifications and extension of the near-native structure are required to meet them with reality. The near-native structures constructed in the present study were forced to adopt almost the same backbone conformations as the native structure (Table 3.4 on page 49 and Table 4.5 on page 72). One possible modification will be to loosen the restraints on the backbone atoms. This can be achieved, for example, by imposing the RMSD restraints [23] so that the backbone conformation can deviate from the native structure by several Å. Such a modification may give more realistic representation of physical objects such as the transition-state structure. It will be also useful for the structure prediction problem in which an approximate fold of the target protein is given and the true native structure is to be searched, provided that an extremely powerful method for conformational search is available (see Section 3.6).

5.5 What are missing

So far, we have been concerned with the energetics in its most limited sense, that is, energy is the only quantity treated throughout the thesis. However, as discussed in Chapter 1, the fundamental physical principle of the protein structure is the thermodynamic hypothesis [4]. Recalling the notion by Anfinsen on page 8, not only the energy but the *Gibbs free energy* of the protein-solvent system which matters. Since protein volumes change very little upon folding [33], the Gibbs free energy under a normal pressure is expected to be well approximated by the Helmholtz free energy $A = E - TS$ where E, T and S are the energy, absolute temperature and the entropy of the whole system. In the present study, only E was fully taken into account. Although the entropy arising from the solvent's degree of freedom was implicitly treated (Sections 2.3 and 2.5), the chain entropy of the protein was completely neglected. Since the enthalpic factors prevail over the entropic factors in protein stabilization [55] as mentioned on page 19, the neglect of the chain entropy seems adequate to the first approximation. Regarding the study in Chapter 3, the chain entropy may not affect much. Vorobjev *et al.* [107] reported that there was no distinctive difference between the chain entropy of the native structure and that of misfolded ones. However, regarding the study in Chapter 4 where the native as well as unfolded structures are taken into account, treatment of chain entropy remains to be a problem to be overcome. Exact calculation of the chain entropy requires exhaust enumeration of all possible structures of the protein chain, which is impossible.

Thus, some great simplicifation is needed. One possible way is to select a set of representative structures of each thermodynamic state of the protein and to perform extensive conformational sampling around these representative structures. The latter will be possible in the near future with recently developed extremely powerful conformational sampling techniques such as the multicanonical algorithm [64], the replica exchange algorithm [99], or algoritms combining both [100]. The former seems more difficult and will require further experimental studies on protein folding and unfolding.

Appendix A

The CASP3 Experiment: A Case Study on Structure Prediction

This appendix presents a case study on protein structure prediction by the fold recognition method. In 1998, the third meeting on the critical assessment of techniques for protein structure prediction (CASP3) [59] was held. This meeting is often called the "structure prediction contest." In CASP, amino acid sequences of soon-to-be-solved protein structures are provided to predictors who submit predicted models of those protein structures before the structures are experimentally solved, and the models are subsequently evaluated after the experimental structures become available. There were three categories of prediction in CASP3: comparative or homology modeling, fold recognition, and *ab initio* prediction. In the comparative or homology modeling category, the amino acid sequences of target proteins (i.e., proteins whose structures are to be predicted) show clear homology to the proteins of known structures so that the main emphasis is put on detailed prediction of side-chain packing and loop structure building. In the fold recognition category, proteins which show structural similarity with the proteins of known structures in spite of no apparent homology are predicted. What *ab initio* prediction means is quite confusing, but at the time of CASP3, it meant prediction methods other than comparative modeling or fold recognition methods.

My colleagues (Motonori Ota, Takeshi Kawabata and Ken Nishikawa) and I organized the team UNAGI and participated in the fold recognition category of CASP3 [71]. We employed an "cooperative approach" [71], that is, we used multiple methods to predict the folds of the target proteins together with biological knowledge from the literature. By combining various sources of information regarding the target proteins, we decided the final predictions. We submitted 56 model structures for 25 target proteins. The whole results is listed in Table A.1.

To summarize the results in Table A.1, we successfully predicted the folds of 9 out of 21 proteins whose experimental structures became available. In addition, 3 new folds were correctly predicted to be so. However, the critical assessment [60, 71] revealed that most of the predicted three-dimensional models were of poor quality even though the correct fold was assigned. This was mostly due to the inaccurate alignment of sequence to structure. We noticed that different methods often yielded totally different predictions. If those different predictions are effectively managed, we may predict at least the correct folds of proteins. Nevertheless, this observation reveals the immaturity of the prediction methods. The main programs for fold recognition we employed different sequence-structure compatibility functions [67, 56, 72, 71]. As was discussed in Chapter 3, development of these compatibility functions embodies intrinsic difficulties such as the noises in structural database of proteins, or functional forms of the functions which heavily depend on one's intuition. One of the discussions in Chapter 3 aims at elucidation of the physical background of sequence-structure compatibility functions.

Table A.1: Summary of the UNAGI's predictions

Target	Len	M1	M2	M3	M4	M5	3D	C	SE	AE
T0043(HPPK)	158	1cus	1ble	2fx2			√	F	w	w
T0044(RTCA)	347	1asyA	1atiA	3hsc	1scuB	2admA	√	F	w	w
T0045(YBAK)	158	**none**	1hrdA	1tmy	1btmB		√	F	N	N
T0046(ADG)	119	**2mcm**	1tul	1fivA	1slcA		√	F	R	F
T0051(GLME)	483	**1reqB**					√	F	R	D
T0052(CV-N)	101	**none**	1pczAB				√	F	N	N
T0053(CBIK)	264	**1akl**					√	F	R	B
T0054(VANX)	202	none	1lbu	1vhh			√	F		?
T0056(DNAB)	114	**none**	1bmfD	1hulA			√	F	N	N
T0061(HDEA)	89	1ngr	1am3	**none**			√	F	N	w
T0062(UBIB)	232	2cnd								
T0063(IF5A)	138	1rsy	2snv				√	F	w	w
T0067(PBP)	187	none	1hurb				√	F	w	w
T0068(PGL2)	376	**1rmg**					√	C	R	
T0071(ADAC)	238	none	**1tf4A**	1slcA	**6fabH**		√	F	R	w
T0072(CD5)	110	1vfaA	1noal							
T0074(EPS15)	98	**2scpA**					√	C	R	
T0075(ETS-1)	110	none	1aoa				√	F	w	w
T0077(L30)	105	1tmy	1div				√	F	w	+
T0078(TESB)	288	none								
T0079(MARA)	129	**1pdnC**	1sfe				√	F	R	F
T0080(3MG)	219	none	1a0i				√	F	w	w
T0081(MGSA)	152	1rnl					√	F	?	•
T0083(CYNS)	156	**1r69**	none				√	F	R	E
T0085(C554)	211	**1fgjA**					√	F	R	D

Len: Number of amino acid residues. M1 ~ M5: five models submitted as prediction. 3D: Whether the experimental structure was available at the time of evaluation (√) or not (blank). C: Category of prediction: F, targets for fold recognition; C, targets which eventually turned out to be those for comparative modeling. SE: Self-evaluation of the models (R, correct model predicted; w, wrong prediction; N, correct "prediction" of new fold; ?, unsure.) AE: The evaluation by the assessor [60] (The quality of correct models are ranked in the order of A (good model) to F(poor, but correct, model); +, may be correct model; •, almost correct; N, new fold correctly "predicted.") The correctly predicted models are typed in **boldface**.

Appendix B

An Abstract Formulation of the Problem

In this appendix, I present an abstract formulation of the main problem of the thesis, the energetics of protein fold. What is investigated in the present study is not specific determinants of particular folds, but universal determinants of various folds. In order to clarify what is meant by the above statement, let us define three types of sets or spaces (Table B.1).

The term "fold" in Table B.1 is defined as in the common practice, for example, as in the SCOP database (page 5, [62]). In this definition, fold is a set of native structures with similar backbone structures. Each native structure corresponds to one amino acid sequence, and *vice versa*. Therefore, although the explicit definition of fold is based on a set of structures, it is implicitly defined on an amino acid sequence space. On the other hand, a confor-

Table B.1: Sequence space, conformation space, energy component space

S_I	Set of amino acid sequences of proteins which adopt fold I.
C_i	Set of conformations a protein i.
e_i	Vector of energy components of a conformation of a protein i.
E_i	Set of e_i of a protein i. (Energy component space.)

mation space of a protein is based on one particular protein by definition. An energy component vector is defined by a set of values of energy components: $e_i = (b_1, \cdots, a_1, \cdots, \theta_1, \cdots, r_1, \cdots, A_1, \cdots)$, which is a set of values of particular bond length, bond angle, torsion angle, non-bonded, hydration energy terms, *etc.* Accordingly, the energy component space is defined as $E_i \equiv \{e_i^0, e_i^1, e_i^2, \cdots\}$. An element s of S_I corresponds to a particular element c_s^0 of C_s which is the native structure of the protein s:

$$s \in S_I \longmapsto c_s^0 \in C_s \tag{B.1}$$

Hereafter, the superscript "0" represent the native structure. An element c of C_i is related to an element of E_i through the energy function f_i of the protein i:

$$f_i : \; c_i \in C_i \longmapsto e_i \in E_i \tag{B.2}$$

Given a particular fold I, the energetic determinants of the fold may be found by studying the set of the energy components of native structures:

$$\mathcal{E}_I^0 \equiv \{e_a^0, e_b^0, \cdots, e_m^0 | \; a, b, \cdots, m \in S_I\} \tag{B.3}$$

On the other hand, the study of the structural determinants of the fold corresponds to the study of the set

$$\mathcal{C}_I^0 \equiv \{c_a^0, c_b^0, \cdots, c_m^0 | \; a, b, \cdots, m \in S_I\} \tag{B.4}$$

However, since the proteins $a, b, \cdots m$ compose different physical systems (in general, $|e_a^0| \neq |e_b^0|$) , what the set $\mathcal{E}_I^0$ means is not trivial. The origin of the difficulty is that fold is defined on the sequence space. Thus, in this study, I have substituted the problem with the one defined on conformation space, dealing with different conformations of one particular protein. This is equivalent to studying the conformation space:

$$\mathcal{C}_i \equiv \{c_i^0, c_i^1, \cdots | \; d_c(c_i^0, c_i^k) < \epsilon_F, \; \forall k\} \subset C_i \tag{B.5}$$

Here, a similarity measure (distance) between two backbone conformations of c_i^a and c_i^b of the same protein is given as $d_c(c_i^a, c_i^b)$, and the threshold that the two backbone conformations can be regarded as sharing a common fold is given as ϵ_F. In the present study, c_i^1 *etc.* are the near-native structures.

Now, how can we study the energetic determinants of fold? We first define a similarity measure (distance) between two energy component vectors, $d_e(e_i^a, e_i^b)$, which is the absolute value of total energy difference between the two structures c_i^a and c_i^b. If c_i^1 is a near-native structure such as ones in Chapter 3, $d_e(e_i^0, e_i^1)$ is about 100 kcal/mol which is quite large. Next, we define an operator g_i^α defined by a characteristic α. The operator g_i^α transforms one e_i to another:

$$g_i^\alpha : \; E_i \longrightarrow G_i^\alpha \subset E_i \tag{B.6}$$

g_i^α is supposed to "project" some energetic properties represented in E_i onto the subspace G_i^α. An element ξ_i^a of G_i^α is related to an element c_i^a of C_i by $\xi_i^a = g_i^\alpha \circ f_i(c_i^a)$, where f_i is the energy function of the system defined in Equation B.2. Finding the energetic determinant of fold can be ideally formulated as follows:

> **Problem A** Let $\mathcal{C}_i$ be a set of structures that share a common fold with the native structure of protein i. Find the transformation g_i^α that minimizes $d_e(\xi_i^0, \xi_i^k)$ for $c_i^k \in \mathcal{C}_i$ and that maximizes it for $c_i^k \notin \mathcal{C}_i$.

Thus found g_i^α, and therefore the characteristics α, is defined to be an energetic determinant of the fold of the protein i. Furthermore, since we are interested in universal determinants, the determinant g_i^α obtained for one particular fold should be also a determinant of other folds. That is, we imposed a condition that g_i^α of one protein i is equivalent to g_j^α of another protein j in the sense that g_i^α and g_j^α are both based on the characteristics α and they satisfy the conditions stated in **Problem A** above:

$$g_i^\alpha \sim g_j^\alpha \tag{B.7}$$

In Chapter 3, it was shown that the electrostatic energy and hydrophobic energy were likely to determine protein folds, and the side-chain packing determines the precise native structure. Therefore, α of g_i^α in the last section is either the "electrostatic energy" or the "hydrophobic energy." On the other hand, the packing energy may be regarded as a function g_i^α that maximizes $d_e(c_i^0, c_i^k)$ for $c_i^k \in C_i \cap \overline{\{c_i^0\}}$.

Bibliography

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. *Molecular Biology of the Cell*, chapter 3. Garland Publishing, New York & London, third edition, 1994.

[2] E. Alm and D. Baker. Matching theory and experiment in protein folding. *Curr. Opin. Struct. Biol.*, 9:189–196, 1999.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[4] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[5] D. Baker. A surprising simplicity to protein folding. *Nature*, 405:39–42, 2000.

[6] V. D. Barger and M. G. Olsson. *Classical Electricity and Magnetism: A Contemporary Perspective*, chapter 4. Allyn and Bacon, Newton, Massachusetts, USA, 1987.

[7] D. Bashford, C. Chothia, and A. M. Lesk. Determinants of a protein fold. *J. Mol. Biol.*, 196:199–216, 1987.

[8] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[9] J. U. 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.

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

[11] C. Chothia. Hydrophobic bonding and accessible surface area in proteins. *Nature*, 248:338–339, 1974.

[12] C. Chothia. One thousand families for the molecular biologist. *Nature*, 357:543–544, 1992.

[13] C. Chothia, I. Gelfand, and A. Kister. Structural determinants in the sequences of immunoglobulin variable domain. *J. Mol. Biol.*, 278:457–479, 1998.

[14] S. Y. Chung and S. Subbiah. A structural explanation for the twilight zone of protein sequence homology. *Structure*, 4:1123–1127, 1996.

[15] J. Clarke, E. Cota, S. B. Fowler, and S. J. Hamill. Folding studies of inmunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway. *Structure*, 7:1145–1153, 1999.

[16] N. D. Clarke, L. J. Beamer, H. R. Goldberg, C. Berkower, and C. O. Pabo. The DNA binding arm of lambda repressor: Critical contacts from a flexible region. *Science*, 254:267–270, 1991.

[17] F. Colonna-Cesari and C. Sander. Excluded volume approximation to protein-solvent interaction: The solvent contact model. *Biophys. J.*, 57:1103–1107, 1990.

[18] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[19] R. Diamond. A note on the rotational superposition problem. *Acta Cryst.*, A44:211–216, 1988.

[20] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–7155, 1990.

[21] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986.

[22] A. H. Elcock. The stability of salt bridges at high temperatures: Implications for hyperthermophilic proteins. *J. Mol. Biol.*, 284:489–502, 1998.

[23] P. Ferrara, J. Apostolakis, and A. Calflisch. Computer simulation of protein folding by targeted molecular dynamics. *Proteins*, 39:252–260, 2000.

[24] A. R. Fersht. Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci. USA*, 97:1525–1529, 2000.

[25] R. P. Feynman, R. B. Leighton, and M. Sands. *The Feynman Lectures on Physics*, volume 1, chapter 1. Addison-Wesley, Reading, Massachusetts, USA, 1963.

[26] A. V. Finkelstein and O. B. Ptitsyn. Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. molec. Biol.*, 50:171–190, 1987.

[27] J. Fitter and J. Heberle. Structural equilibrium fluctuations in mesophilic and thermophilic α-amylase. *Biophys. J.*, 79:1629–1636, 2000.

[28] K. D. Gibson and H. A. Scheraga. Minimization of polypeptide energy, I: Preliminary structures of bovine pancreatic ribonuclease S-peptide. *Proc. Natl. Acad. Sci. USA*, 58:420–427, 1967.

[29] N. Gō. Theoretical studies of protein folding. *Ann. Rev. Biophys. Bioeng.*, 12:183–210, 1983.

[30] D. P. Goldenberg. Finding the right fold. *Nature Struct. Biol.*, 6:987–990, 1999.

[31] J. M. Guss, H. D. Bartunik, and H. C. Freeman. Accuracy and precision in protein crystal structure analysis: Restrained least-squares refinement of the crystal structure of poplar plastocyanin at 1.33 angstroms resolution. *Acta Crystallogr. B*, 48:790–811, 1992.

[32] Y. Harpaz and C. Chothia. Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J. Mol. Biol.*, 238:528–539, 1994.

[33] Y. Harpaz, M. Gerstein, and C. Chothia. Volume changes on protein folding. *Structure*, 2:641–649, 1994.

[34] T. F. Havel. An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Prog. Biophys. Molec. Biol.*, 56:43–78, 1991.

[35] M. A. Holmes and R. E. Stenkamp. Structures of met and azidomet hemerythrin at 1.66 Å resolution. *J. Mol. Biol.*, 220:723–737, 1991.

[36] A. Janardhan and S. Vajda. Selecting near-native conformations in homology modeling: The role of molecular mechanics and solvation terms. *Protein Sci.*, 7:1772–1780, 1998.

[37] D. T. Jones, W. R. Taylor, and J. M. Thornton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

[38] D. T. Jones and J. M. Thornton. Potential energy functions for threading. *Curr. Opin. Struct. Biol.*, 6:210–216, 1996.

[39] Y. K. Kang, G. Némethy, and H. A. Scheraga. Free energies of hydration of solute molecules. 1. improvement of the hydration shell model by exact computations of overlapping volumes. *J. Phys. Chem.*, 91:4105–4109, 1987.

[40] A. Karshikoff and R. Ladenstein. Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. *Protein Eng.*, 11:867–872, 1998.

[41] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, 14:1–63, 1959.

[42] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, 181:662–666, 1958.

[43] S. Kobayashi. *Differential Geometry of Curves and Surfaces*, chapter 4. Shokabo, Tokyo, Japan, 1995. (in Japanese).

[44] P. J. Kraulis. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, 24:946–950, 1991.

[45] S. Kumar, C.-J. Tsai, and R. Nussinov. Factors enhancing protein thermostability. *Protein Eng.*, 13:179–191, 2000.

[46] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26:283–291, 1993.

[47] T. Lazaridis and M. Karplus. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.*, 288:477–487, 2 1999.

[48] T. Lazaridis and M. Karplus. Effective energy function for proteins in solution. *Proteins*, 35:133–152, 1 1999.

[49] T. Lazaridis, I. Lee, and M. Karplus. Dynamics and unfolding pathways of a hyperthermophilic and a mesophilic rubredoxin. *Protein Sci.*, 6:2589–2605, 1997.

[50] B. Lee and F. M. Richards. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, 55:379–400, 1971.

[51] M. Levitt, M. Gerstein, E. Huang, S. Subbiah, and J. Tsai. Protein folding: The endgame. *Annu. Rev. Biochem.*, 66:549–579, 1997.

[52] H. Li, R. Tejero, D. Monleon, D. Bassolino-Klimas, C. Abate-Shen, R. E. Bruccoleri, and G. T. Montelione. Homology modeling using simulated annealing of restrained molecular dynamics and conformational search calculations with CONGEN: Application in predicting the three-dimensional structure of murine homeodomain Msx-1. *Protein Sci.*, 6:956–970, 1997.

[53] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. The penultimate rotamer library. *Proteins*, 40:389–408, 2000.

[54] G. I. Makhatadze and P. L. Privalov. Contribution of hydration to protein folding thermodynamics. I: The enthalpy of hydration. *J. Mol. Biol.*, 232:639–659, 1993.

[55] G. I. Makhatadze and P. L. Privalov. Energetics of protein structure. *Adv. Protein Chem.*, 47:307–425, 1995.

[56] Y. Matsuo, H. Nakamura, and K. Nishikawa. Detection of protein 3D-1D compatibility characterized by the evaluation of side-chain packing and electrostatic interactions. *J. Biochem. (Tokyo)*, 118:137–148, 1995.

[57] C. A. McPhalen and M. N. James. Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. *Biochemistry*, 26:261–269, 1987.

[58] D. A. McQuarrie. *Statistical Mechanics*, chapter 15. HarperCollins, New York, USA, 1976.

[59] J. Moult, T. Hubbard, K. Fidelis, and J. T. Pedersen. Critical assessment of methods of protein structure prediction (CASP): Round III. *Proteins*, Suppl. 3:2–6, 1999.

[60] A. G. Murzin. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins*, Suppl. 3:88–103, 1999.

[61] A. G. Murzin and A. Bateman. Distant homology recognition using structural classification of proteins. *Proteins*, Suppl. 1:105–112, 1997.

[62] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[63] T. Nakai, A. Kidera, and H. Nakamura. Intrinsic nature of the three-dimensional structure of proteins as determined by distance geometry with good sampling properties. *J. Biomol. NMR*, 3:19–40, 1993.

[64] N. Nakajima, H. Nakamura, and A. Kidera. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J. Phys. Chem. B*, 101:817–824, 1997.

[65] H. Nakamura. Roles of electrostatic interaction in proteins. *Q. Rev. Biophys.*, 29:1–90, 1996.

[66] H. Nakamura and S. Nishida. Numerical calculations of electrostatic potentials of protein-solvent systems by the self consistent boundary method. *J. Phys. Soc. Jpn.*, 56:1609–1622, 1987.

[67] K. Nishikawa and Y. Matsuo. Development of pseudoenergy potentials for assessing protein 3D-1D compatibility and detecting weak homologies. *Protein Eng.*, 6:811–820, 1993.

[68] J. Novotný, R. Bruccoleri, and M. Karplus. An analysis of incorrectly folded protein models: Implications for structure predictions. *J. Mol. Biol.*, 177:787–818, 1984.

[69] J. Novotný, A. A. Rashin, and R. E. Bruccoleri. Criteria that discriminate between native proteins and incorrectly folded models. *Proteins*, 4:19–30, 1988.

[70] T. Ooi, M. Oobatake, G. Némethy, and H. A. Scheraga. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. USA*, 84:3086–3090, 1987.

[71] M. Ota, T. Kawabata, A. R. Kinjo, and K. Nishikawa. Cooperative approach for the protein fold recognition. *Proteins*, Suppl. 3:126–132, 1999.

[72] M. Ota and K. Nishikawa. Assessment of pseudo-energy potentials by the best-five test: A new use of the three-dimensional profiles of proteins. *Protein Eng.*, 10:339–351, 1997.

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

[74] R. G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules.* Oxford University Press, New York, USA, 1989.

[75] L. Pauling and R. B. Corey. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proc. Natl. Acad. Sci. USA*, 37:235–240, 1951.

[76] L. Pauling and R. B. Corey. The pleated sheet: A new layer configuration of polypeptide chains. *Proc. Natl. Acad. Sci. USA*, 37:251–256, 1951.

[77] L. Pauling and E. B. Wilson, Jr. *Introduction to Quantum Mechanics with Applications to Chemistry*, chapter 14. Dover, New York, USA, 1985.

[78] D. Perl, C. Welker, T. Schindler, K. Schröder, M. A. Marahiel, R. Jaenicke, and F. X. Schmid. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nature Struct. Biol.*, 5:229–235, 1998.

[79] K. W. Plaxco, K. T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, 277:985–994, 1998.

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

[81] P. L. Privalov and G. I. Makhatadze. Contribution of hydration to protein folding thermodynamics. II: The entropy and Gibbs energy of hydration. *J. Mol. Biol.*, 232:660–679, 1993.

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

[83] F. M. Richards. The intepretation of protein structures: Total volume, group volume distributions and packing density. *J. Mol. Biol.*, 82:1–14, 1974.

[84] F. M. Richards. Areas, volumes, packing, and protein structure. *Ann. Rev. Biophys. Bioeng.*, 6:151–176, 1977.

[85] F. M. Richards and W. A. Lim. An anaysis of packing in the protein folding problem. *Q. Rev. Biophys.*, 26:423–498, 1993.

[86] T. J. Richmond. Solvent accessible surface area and excluded volume in proteins: Analytical equations for overlapping spheres and implications for the hydrophobic effect. *J. Mol. Biol.*, 178:63–89, 1984.

[87] R. B. Russell and G. J. Barton. Structural features can be unconserved in proteins with similar folds: An analysis of side-chain to side-chain contacts, secondary structure and accessibility. *J. Mol. Biol.*, 244:332–350, 1994.

[88] M. Saarinen, F. K. Gleason, and H. Eklund. Crystal structure of thioredoxin-2 from anabaena. *Structure*, 3:1097–1108, 1995.

[89] P. V. Sahasrabudhe, R. Tejero, S. Kitao, Y. Furuichi, and G. T. Montelione. Homology modeling of an RNP domain from a human RNA-binding protein: Homology-constrained energy optimization provides a criterion for distinguishing potential sequence alignments. *Proteins*, 33:558–566, 1998.

[90] R. Samudrala and J. Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, 275:895–916, 1998.

[91] R. Sarma. *Ramachandran*. Adenine Press, Schenectady, New York, USA, 1998.

[92] G. E. Schulz. Protein differentiation: Emergence of novel proteins during evolution. *Angew. Chem. Int. Ed. Engl.*, 20:143–151, 1981.

[93] G. E. Schulz, C. D. Barry, J. Friedman, P. Y. Chou, G. D. Fasman, A. V. Finkelstein, V. I. Lim, O. B. Ptitsyn, E. A. Kabat, T. T. Wu, M. Levitt, B. Robson, and K. Nagano. Comparison of predicted and experimentally determined secondary structure of adenyl kinase. *Nature*, 250:140–142, 1974.

[94] J. Sevcik, I. Zegers, L. Wyns, Z. Dauter, and K. S. Wilson. Complex of ribonuclease Sa with a cyclic nucleotide and a proposed model for the reaction intermediate. *Eur. J. Biochem.*, 216:301–305, 1993.

[95] E. I. Shakhnovich and A. V. Finkelstein. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers*, 28:1667–1680, 1989.

[96] M. J. Sippl. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, 5:229–235, 1995.

[97] P. F. W. Stouten, C. Frömmel, H. Nakamura, and C. Sander. An effective solvation term based on atomic occupancies for use in protein simulations. *Mol. Simul.*, 10:97–120, 1993.

[98] K. Sugihara. *Computational Geometry Programming in Fortran.* Iwanami Shoten, Tokyo, Japan, 1998. (in Japanese).

[99] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.

[100] Y. Sugita and Y. Okamoto. Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chem. Phys. Lett.*, 329:261–270, 2000.

[101] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory.* Dover, Mineola, New York, USA, 1996.

[102] A. Szilágyi and P. Závodszky. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: Results of a comprehensive survey. *Structure*, 8:493–504, 2000.

[103] S. Takada. Gō-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. USA*, 96:11698–11700, 1999.

[104] M. T. Tomic, J. R. Somoza, D. E. Wemmer, Y. W. Park, J. M. Cho, and S. Kim. ^{1}H resonance assignments, secondary structure and general topology of single-chain monellin in solution as determined by ^{1}H-2D-NMR. *J. Biomol. NMR*, 2:557–572, 1992.

[105] K. C. Usher, A. F. A. de la Cruz, F. W. Dahlquist, R. V. Swanson, M. I. Simon, and S. J. Remington. Crystal structures of CheY from *Thermotoga maritima* do not support conventional explanations for the structural basis of enhanced thermostability. *Protein Sci.*, 7:403–412, 1998.

[106] K. Volz and P. Matsumura. Crystal structure of *Escherichia coli* CheY refined at 1.7-Å resolution. *J. Biol. Chem.*, 266:15511–15519, 1991.

[107] Y. N. Vorobjev, J. C. Almagro, and J. Hermans. Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. *Proteins*, 32:399–413, 1998.

[108] H. Wako. Monte Carlo simulations of a protein molecule with and without hydration energy calculated by the hydration-shell model. *J. Protein Chem.*, 8:733–747, 1989.

[109] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.*, 7:230–252, 1986.

[110] L. Wesson and D. Eisenberg. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.*, 1:227–235, 1992.

[111] H. Wu. Studies on denaturation of proteins XIII: A theory of denaturation. *Chinese J. Physiol.*, 5:321–344, 1931. Reproduced in Adv. Protein Chem. Vol. 46, pp. 6–26, (2000).

[112] L. Xiao and B. Honig. Electrostatic contributions to the stability of hyperthermophilic proteins. *J. Mol. Biol.*, 289:1435–1444, 1999.