IDENTIFICATION OF NATIVE CONFORMATIONS OF PROTEINS USING A LOW RESOLUTION MODEL

by

Ş. BANU ÖZKAN
B.S. in Chemical Engineering, Boğaziçi University, 1995

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science
in
Chemical Engineering

Boğaziçi University
1997

ACKNOWLEDGMENTS

I would like to express my gratitude to my thesis advisor, Prof. Dr. Ivet Bahar, for her guidance, helpful criticisms and stimulating encouragement throughout my study. I am also grateful to Prof. Dr. Burak Erman for his interest, help and constructive suggestions in my work.

Special thanks are due to Doç. Dr. Türkan Haliloğlu for her help whenever I felt hopeless and for the time she has devoted to reading and commenting on my thesis.

I wish to remember Doç. Dr. Ali Rana Atılgan for his suggestions and moral support.

I am much indebted to Özlem Keskin, Melik Cumhur Demirel, Mehmet Sayar who were always there to support me. I would like to thank Neşe Kurt, Başak Samur, Taner Şen for their patience and help especially during the preperation of this thesis. Surely, I want to thank Berna Sarıyar (the one who made the bad days more than enjoyable), Leyla Özkan, Elif Özkırımlı, Seza Orçun.

ABSTRACT

Exhaustive enumeration of accessible conformations is performed for eight sample proteins by using a low resolution model. The secondary structure elements of the proteins are assumed to behave as rigid blocks. Large ensembles of decoy structures (10^{12}) are generated by rotating the virtual bonds selected at the flexible regions of the protein.

Several constraints are used to ensure the diversity and proper sampling of native-like conformations. These include the excluded volume requirement, the radius of gyration constraint, and the need for the occurrence of a sufficiently strong attractive interaction to hold the amino acids in a stable, coherent form. Accordingly, conformations that are not compact enough or have an overlapping segment are eliminated, as well as those subject to a weak overall potential of mean force. The number of accepted conformations is thus reduced to 1500-5000 decoys for each protein. In order to distinguish the most native-like fold among these conformations, two criteria are used: Root-mean-square (RMS) deviations with respect to x-ray structure and energies of the generated conformations. Energies are evaluated on the basis of knowledge-based potentials. In the case of four of the examined proteins, the lowest energy conformation obtained in simulations also exhibits the lowest RMS deviation from the corresponding x-ray structure. This observation indicates the suitability of the low resolution model and parameters for distinguishing the native fold. Furthermore, in all of the studied proteins, the RMS deviation between the coordinates of the lowest energy conformation and those of x-ray structures is found to remain lower than 2.0 Å, except for one protein, trypsin inhibitor. The latter is found to be correctly predicted upon proper choice of flexible regions.

The dihedral angle preferences of the flexible bonds in the set of low energy conformations are examined by obtaining the torsional state probability distribution curves for the rotatable bonds. It is observed that some of the flexible bonds exhibit a strong tendency to assume well-defined rotational angles in most of the low energy conformations. This indicates the existence of a stable region in the neighborhood of these bonds, which may possibly act as a nucleus in the sequential folding of the protein.

ÖZET

Sekiz protein için basitleştirilmiş bir modelin alabileceği tüm konformasyonlar arasından, kristal yapıya en yakın olanlar simulasyon yardımı ile bulunmuştur. Proteinin ikincil yapıları simulasyon sırasında sabit yapılar olarak kabul edilmiştir. Proteinin sabit olmayan, hareketli bölgelerindeki sanal bağların 30 derecelik açılarla döndürülmesi sonucu her protein için, yaklaşık 10^{12} konformasyon üretilmiştir.

Elde edilen konformasyonların çeşitliliğini ve gerçek yapıya benzerliğini sağlamak amacı ile birtakım kıstaslar getirilmiştir. Bunlar: i) proteinin boyuna uygun bir hacmi kaplaması, ii) birbirine bağlı olmayan iki rezidünün yakınlığının 2 Å' ı geçmemesi, ve iii) proteinin kararlı yapıda bulunması için gerekli enerjiye sahip olması şeklinde sıralanabilir. Bu özelliklere sahip olmayan konformasyonların elenmesiyle, üretilen konformasyonların sayısı 1500-5000'e düşmüştür. Bu konformasyonların arasında gerçek yapıya en yakın konformasyonu belirlemek için iki özellik incelenmiştir: i) x ışını kristallografisi metodu ile hesaplanmış yapıdan ortalama karekök sapma (RMS) değeri, ii) konformasyonların enerjileri. Enerjiler, bilgi esaslı potensiyellere göre hesaplamıştır. İncelenen proteinlerin dördünde, hesaplamalar sonucu en düşük enerjiye sahip konformasyon, ayni zamanda kristal yapıdan en düşük RMS sapması göstermiştir. Bu bulgu, kullanılan basitleştirilmiş modelin ve enerji parametrelerinin uygun olduğunu göstermektedir. Ayrıca tripsin inhibitör haricindeki tüm proteinlerin en düşük enerjili konformasyonları, kristal yapılarına oranla 2 Å 'dan daha düşük sapmalar göstermiştir. Hareketli bölgelerin doğru belirlenmesiyle, tripsin inhibitör proteini içinde gerçeğe en yakın yapılar doğru olarak tahmin edilmiştir.

Hareketli bölgelerdeki bağların dönme açılarına göre olasılık dağılım eğrilerini hesaplamak için, tüm düşük enerjili konformasyonlar incelenmiştir. Bu bölgelerde, bazı dönme açılarının belirli değerleri daha çok tercih ettiği gözlenmiştir. Bu durum, bu bağların bulunduğu bölgede çok kararlı bir yapı olduğunu kanıtlamaktadır.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

A	Singular value decomposition matrix
ΔA	Elastic free energy change
$E(\phi_i)$	Potential associated with the torsion of the i^{th} flexible virtual bond
$E(\theta_i\text{-},\phi_i)$	Potential associated with coupling between bond angles and torsions of the i^{th} flexible virtual bond
E_{BB}	Potential of mean force between two backbone units
E_{LR}	Long-range interaction energy
E_{SB}	Potential of mean force between a side group and a backbone unit
E_{SR}	Short-range interaction energy
E_{SS}	Potential of mean force between two side groups
E_{total}	Total interaction energy
N	Number of residues
R_{ij}	Separation between the i^{th} and j^{th} C^{α} atoms
R_g	Radius of gyration
ΔR_{ij}	Fluctuations in the separation between the i^{th} and j^{th} C^{α} atoms
$\{\Delta R\}$	3N dimensional column vector of fluctuations
S_i	Sidechain atom
T_i	Transformation matrix for the i^{th} backbone bonds
T_i^s	Transformation matrix for the i^{th} sidechain
U	Orthonormal matrix of left singular vector
V	Orthonormal matrix of right singular vector
W	Distribution function
Z	Configurational partition function
k_B	Hookean force constant
l_i	Position vector of the i^{th} backbone bond in its local frame
l_i^s	Position vector of the i^{th} sidechain in its local frame
l_i	Length of i^{th} backbone bond
l_i^s	Length of i^{th} sidechain bond
n	Number of residues
r_{ij}	distance between i^{th} and j^{th} interaction sites
u_i	i^{th} left singular vector of SVD matrix

v_i	i^{th} right singular vectors of SVD matrix
$\overrightarrow{r_N^{\alpha}}$	Position vector of the N^{th} backbone bond in the main frame
$\overrightarrow{r_N^{s}}$	Position vector of the N^{th} sidechain in the main frame
ϕ_i	Torsional angle of i^{th} backbone bond
$\phi_i{}^s$	Torsional angle of i^{th} sidechain
γ^*	Counterpart of the single parameter in Hookean potential
Γ^{-1}	Stiffness matrix
Γ	Kirchoff matrix
Λ	Diagonal matrix
λ_i	Eigenvalues of diagonal matrix
θ_i	Bond angle between i^{th} and $i\text{-}1^{th}$ backbone bonds
$\theta_i{}^s$	Bond angle between i^{th} backbone bond and i^{th} sidechain
σ_i	Singular values of SVD matrix

ABBREVIATIONS

Ala	Alanine
Arg	Arganine
Asn	Asparagine
Asp	Aspartic acid
B-B	Backbone-backbone
Cys	Cysteine
Gln	Glutamine
Glu	Glutamatic acid
Gly	Glycine
His	Histidine
Ile	Isoleucine
Lys	Lysine
Met	Methionine
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
Phe	Phenylalanine
Pro	Proline
RMS	Root-mean-square
S-B	Sidechain-backbone
S-S	Sidechain-sidechain
Ser	Serine
SVD	Singular Value Decomposition
Thr	Threonine
Trp	Tryptophan
Tyr	Tyrosine
Val	Valine
1ctf	Ribosomal Protei
1sn3	Scorpion Neurotoxin
1ubq	Uboquitin
1r69	434 Repressor

2cro	434 Cro Protein
3icb	Calcium Binding Protein
4pti	Trypsin inhibitor
4rxn	Rubredoxin

1. INTRODUCTION

Two major difficulties faced during the search for the most favorable conformation of proteins are: (i) the exponentially large number of possible protein conformations, and consequently the low probability of generating a sufficient number of compact conformations, and (ii) the lack of effective criteria for differentiating between correct (native) and incorrect folds. A possible way of overcoming the first difficulty is to adopt coarse-grained, or low resolution models. After generating coarse-grained conformations, an important question is to assess whether an energy function exists that is able to distinguish the correct fold from all other possible conformations [1]. Empirical energy functions derived as potentials of mean force from the database of known structures are utilized for this purpose. Levitt [2] generated potentials of mean force by averaging energies over all relative orientation of side-chains. Huang et al. [3] devised a potential which uses only a simple classification of different residues as hydrophobic or hydrophilic, reminiscent of the theoretical energy models of Yue and Dill [4]. Mairov and Crippen [5] found a potential function by an optimization procedure whose aim is to maximize the difference in energy between correct and incorrect protein conformations.

Yet, it is recognized that for design purposes and dynamic simulations, more precise expressions including both the distance dependence of residue-residue potentials, and the conformational preferences of the polypeptide backbone should be developed. Park and Levitt [1] showed that the efficiency of the potential in discriminating x-ray and near native folds from amongst an ensemble of decoy conformations increased when an artificial distance dependence was introduced into the potentials [6]. Monge et al. [7] emphasized the importance of improving the inter-residue potential functions. In the recent study of Bahar and Jernigan [8], each residue was represented by two interaction sites, one on the backbone, and the second on the amino acid sidechain. Distance-dependent sidechain-sidechain (S-S) interaction potentials extracted at 0.4 Å resolution were used to estimate the effective contact potentials operating over different distance ranges. These potentials are used to evaluate the energies of conformations generated in the present study.

Generally, the potentials developed for low-resolution models have been tested by trying to predict the observed tertiary structure of a protein from its primary structure in two ways: i) Folding simulations where the conformation of the polypeptide chain is changed to obtain spatial arrangements with lowest possible potential energy using energy minimization, and Monte Carlo techniques [9-13] or genetic algorithms [14-17]; ii) Inverse folding method, in which the aim is to seek the sequence that will favor the particular fold by using a known three-dimensional structure [18, 19]. Instead of changing the conformation, the sequence is threaded in this case through template conformations [14, 20-24].

Other approaches involving the generation of sets of decoy protein conformations are performed as an alternative effective method for testing potentials. For example, all lattice conformations within a volume having the correct fold were generated by Covell and Jernigan [25] for several small proteins. They found that the conformation nearest the native was always within the top 1% of the conformations ranked according to their overall potential. Williams et al. [26] generated near native conformations by using Monte Carlo techniques to investigate the usefulness of different solvation models. Monge et al. generated enumeration conformations within the range of 4 to 10 Å deviations from X-ray structures to examine the effectiveness of all-atom and reduced representation of energy functions. Hinds and Levitt [27, 28] generated exhaustive sets of decoys on a diamond lattice. Wang et al. [29] generated small ensembles of conformations 2.8 to 7.8 Å RMS deviation from the X-ray structure by using molecular dynamics simulations.

The aim of the present study is to test the usefulness of a recently developed model and its geometry and energy parameters [30] for recognizing the tertiary structure of globular proteins, knowing the secondary structure and the flexible regions of the protein. The approach is similar in spirit to that of Park and Levitt [1]. Basically, the packing of rigid structural elements connected by flexible strings of amino acids is explored. Some segments of the protein are therefore implicitly assumed to possess sufficient stability on a local scale to maintain their structure during the three-dimensional organization of the molecule. These rigid structural elements may be associated with early folding parts of proteins.

Such a sequential scheme for structure formation goes back to the original proposals of Ptitsyn and Rashin [31] for α-helices, and Cohen, Sternberg and Taylor [32] for β-strands. The same type of hierarchical folding mechanism proved useful in a number of recent coarse-grained simulations [33, 34, 7, 4] aiming at predicting tertiary structures of proteins with known secondary structure. Obviously the method is useful if (i) precise knowledge of secondary structure, is available, and (ii) a sequential folding mechanism is applicable to the investigated protein.

The study is composed of three parts: (i) generating a complete set of conformations for small globular proteins by rotating the virtual C^{α}-C^{α} bonds that are located in the flexible regions of the protein, (ii) filtering out the lowest energy conformation(s) on the basis of empirical contact potentials, and (iii) comparing the predicted structures with those already measured by x-ray or NMR experiments.

In addition to an assessment of the discriminative ability of knowledge-based potentials, the present analysis is expected to give insights into several questions concerning protein simulations on a coarse grained scale. For example, what might be an appropriate resolution, or step size, for exhaustively searching the conformational space, in order to capture the native fold? Stated in other terms, is the energy minimum at the native state deep and broad enough to be distinguished despite moving at large steps over the energy landscape? How much specificity is needed in order to recognize the native fold? To what extent is a sequential folding mechanism applicable? Is it possible to identify regions that invariably fold into the correct tertiary structure for a given protein, and therefore act as nuclei or core regions underlying the stability of the overall tertiary structure?

The plan of the present thesis is as follows: In the following section, an outline of structural characteristic of proteins will be presented. The protein folding problem, recent studies on the protein folding problem and the evaluation of thermal fluctuations will be discussed in the second and the third chapter, respectively. The model and the method that will be used for predicting the native tertiary structure of a protein will be presented in the fifth chapter. The results of the calculations will be shown in the sixth chapter. Conclusion and recommendations will be presented in the seventh chapter.

2. PROTEIN STRUCTURE : GENERAL INFORMATION

2.1. Definition of Proteins

Proteins are chemically high polymers which play crucial roles in virtually all biological processes. They are very large molecules; their molecular weights are often in ten thousands. The basic component of these molecules is the polypeptide chain, an unbranched polymer consisting of a sequence of amino acid residues. There are 20 commonly occurring amino acids, and a typical chain will contain a few hundreds of these elementary structural units. Protein molecules consist of one or a small number of such polypeptide chains, complemented in some cases by one or more prosthetic groups (e.g., metal ions or special organic molecules).

All of the 20 amino acids have in common a central carbon atom (C^{α}) to which are attached a hydrogen atom, an amino group (NH_2), and a carboxyl group (COOH). The sidechain that is attached to the C^{α} through its fourth valency distinguishes one amino acid from another. (Figure 2.1)

During protein synthesis the carboxy group of a given amino acid condenses with the amino group of the next amino acid along the chain sequence to eliminate water and yield a peptide bond. This process is repeated as the chain elongates. The amino group of the first amino acid of a polypeptide chain and the carboxy group of the last amino acid remain intact, and the chain is said to run from its amino terminus to carboxy terminus. The formation of a succession of peptide bonds generates a "main chain" or "backbone" from which the various sidechains are projected.

The main-chain atoms of each amino acid are the carbon atom C^{α} to which the sidechain is attached, the NH group bound to C^{α} , and the carbonyl group C′= O, where the carbon atom C′ is attached to C^{α}. These unit residues are linked into a polypeptide by a peptide bond between the C′ atom of one residue and the nitrogen atom of the next (Figure 2.2)

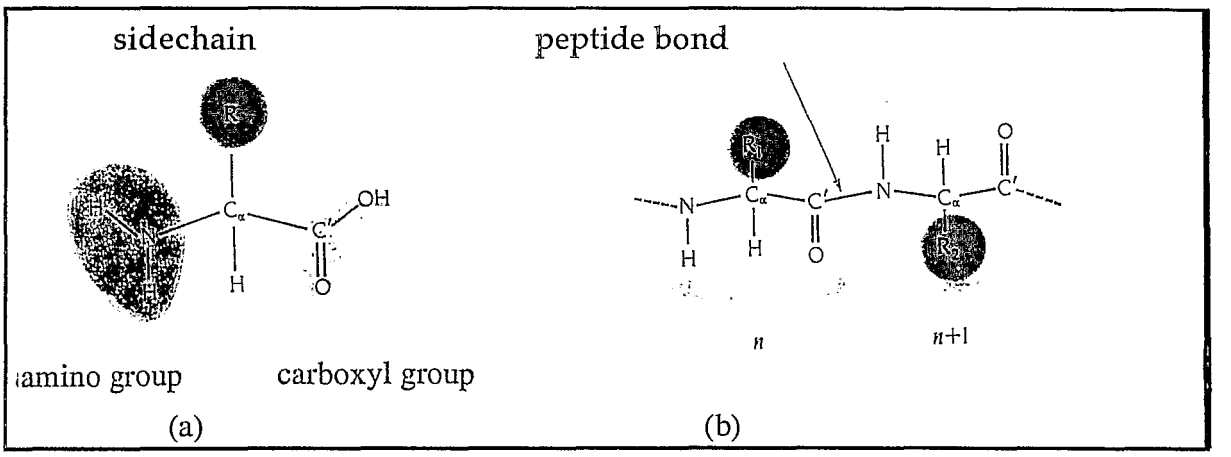


FIGURE 2.1. Schematic diagrams of an amino acid and a polypeptide chain. (a) Representation of an amino acid which consists of a central atom C^{α} bonded to an amino group NH_2, a carboxy group C′OOH, hydrogen atom H, and a side chain, R. (b) Representation of polypeptide chain in which the carboxy group of amino acid n has formed a peptide bond, C-N, to the amino group of amino acid n+1, with the elimination of one water molecule.

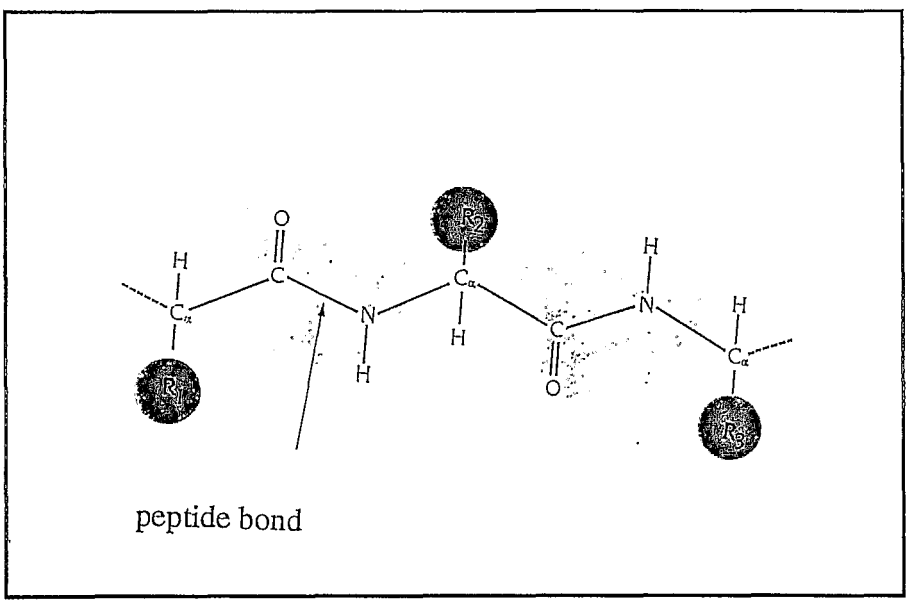


FIGURE 2.2. Schematic representation of a part of a polypeptide chain divided into peptide units. Each unit consists of the C^{α} and the C′=O group of residue n, and the NH group and C^{α}atom of residue n+1.

The identity of the sidechain distinguishes the different amino acids. The names of these twenty different amino acids are abbreviated with both a three letter and one letter code as shown in Table 2.1.

TABLE 2.1. The abbreviation of 20 amino acids.

Name	Three-letter code	One-letter code
Glycine	Gly	G
Alanine	Ala	A
Valine	Val	V
Isoleucine	Ile	I
Leucine	Leu	L
Serine	Ser	S
Threonine	Thr	T
Aspartatic acid	Asp	D
Asparagine	Asn	N
Glutamatic acid	Glu	E
Glutamine	Gln	Q
Lysine	Lys	K
Arginine	Arg	R
Cysteine	Cys	C
Methionine	Met	M
Phenylalanine	Phe	F
Tyrosine	Tyr	Y
Tryptophan	Trp	W
Histidine	His	H
Proline	Pro	P

The amino acids are usually classified in three groups, depending on the chemical nature of the sidechain. The first group consists of those with hydrophobic sidechains Ala (A), Val (V), Leu (L), Ile (I), Phe (F), Pro (P), and Met (M): The four charged residues Asp (D), Glu (E), Lys (K) and Arg (R) form

the second class. The third class contains those with polar side chains Ser (S), Thr (T), Cys (C), Asn (N), Gln (Q), His (H), and Tyr (W). The amino acid glycine Gly (G) is considered either to form a fourth class or to belong to the first class.

The four groups attached to the C^{α} atom are chemically different for all the amino acids except glycine where two H atoms bind to C^{α} atom. All amino acids except glycine are chiral atoms which can exist in two mirror-image forms, called the L-isomer and the D-isomer (Figure 2.3). Only L-amino acids are constituents of proteins.

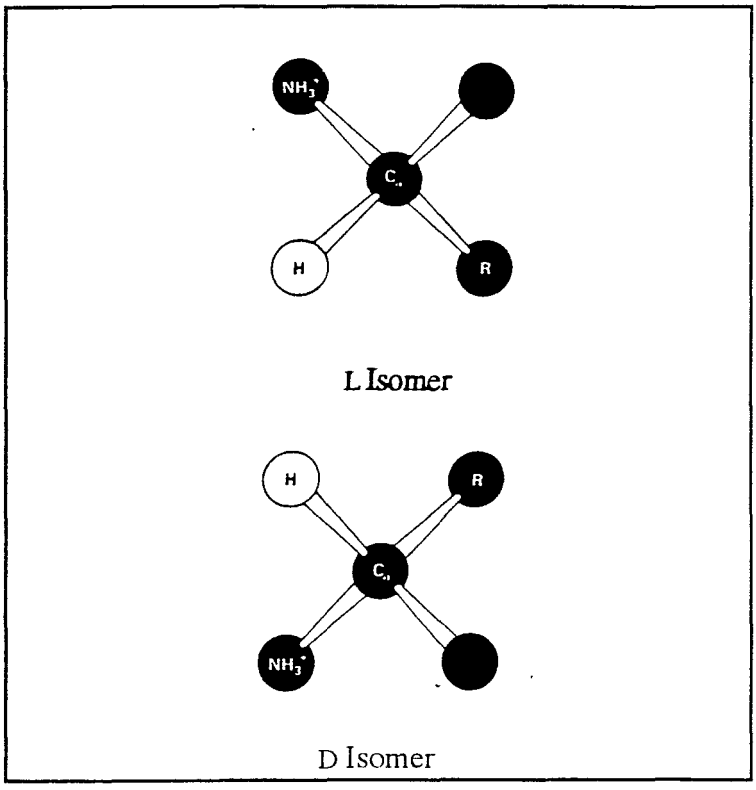


FIGURE 2.3. The L and D isomers of amino acids. They are the mirror images of each other.

The conformation of a protein is specified by rotational angles of all the single bonds of its covalent structure. Different conformations of the same protein have the same covalent structure, and they are interconverted by the rotations around these C^{α}-C′ and the N-C^{α} bonds. The only exception to this is

the disulfide bond, which can be formed reversibly between two Cys residues [35].

The polypeptide chain may be divided into peptide units that extend from one C^{α} atom to the next. Each C^{α} atom except the first and last thus belongs to two such units. All the atoms in such a unit are fixed in a plane with the bond lengths and bond angles being very nearly the same in all units. Since the peptide units are effectively rigid groups that are linked together by covalent bonds at C^{α} atoms, the only degrees of freedom they have are rotations around virtual C^{α}-C^{α} bonds.

Four levels of protein structure are defined (Figure 2.4). The amino acid sequence and location of disulfide bridges, if there are any, is called the primary structure. Secondary structure refers to the regular organization of amino acid residues that are close to one another in the linear sequence such as α-helices and β-strands. Tertiary structure is formed by packing of secondary structures into one or several compact globular domains. Proteins that contain more than one polypeptide chain display an additional level of structural organization, namely quaternary structure, which refers to the way in which the chains are packed together. Each polypeptide chain in such a protein is called a subunit. The primary structure of a protein determines its tertiary structure; in other words, the final specific structure of a polypeptide chain is an unique function of its amino acid sequence [36].

2.1.1. Secondary Structures

It has been noticed by Kendrew [36] that the amino acids in the interior of the protein have mostly hydrophobic side groups, and the hydrophilic groups are generally located on the surface of the protein. The main driving force for folding compact water-soluble globular protein molecules is to bury hydrophobic side chains into the interior of the molecule, thus creating a hydrophobic core and a hydrophilic surface.

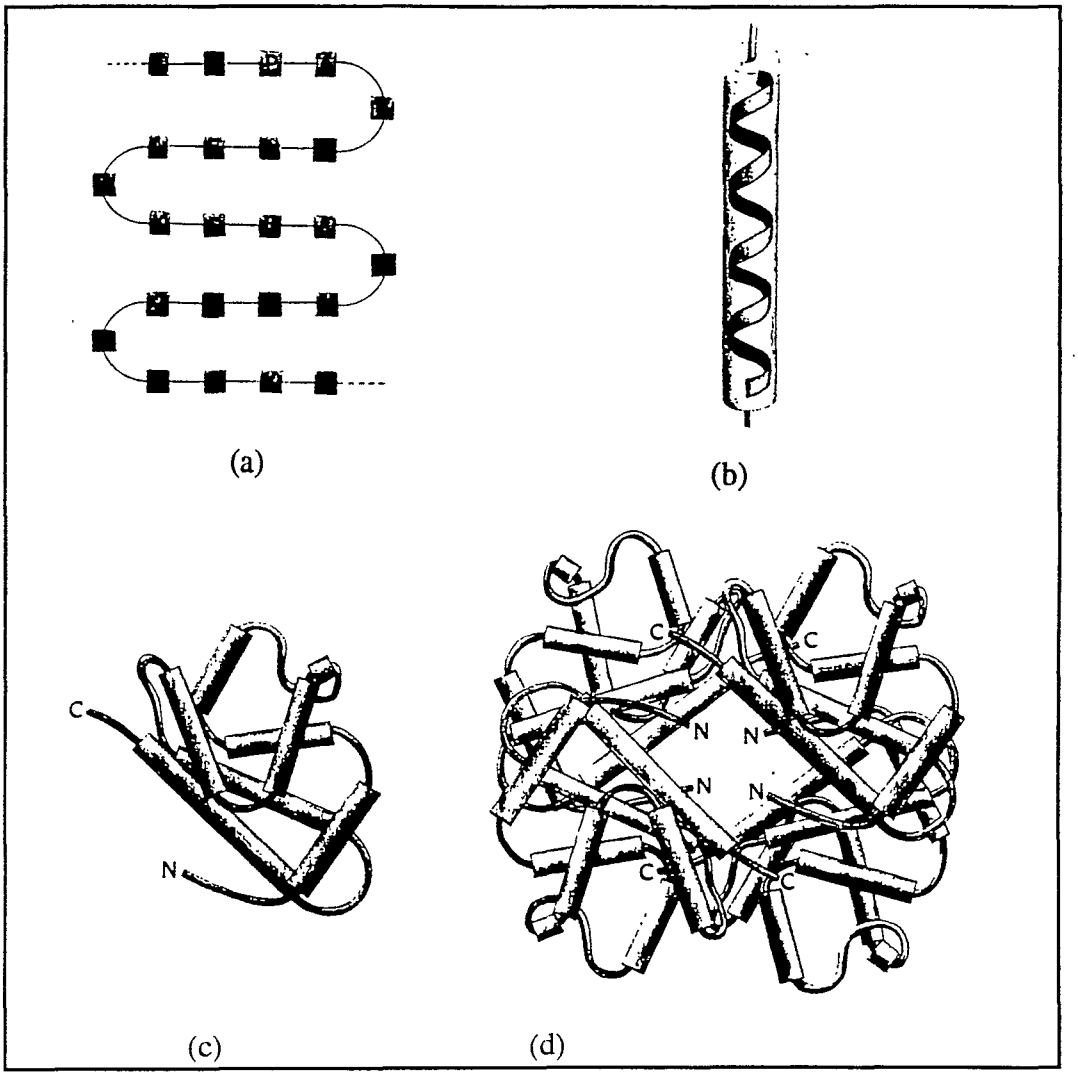


FIGURE 2.4. Four different levels of protein structure; (a) primary, b) secondary, (c) tertiary, and (d) quaternary structures.

In order to bring the sidechains into the hydrophobic core, the main chain must also participate in the core. However, the main chain is highly polar and therefore hydrophilic. Its polar groups must be neutralized by hydrogen bond formation in a hydrophobic environment. In order to form hydrogen bonds, regular secondary structures such as α–helices and β–sheets are formed in the interior of the protein. Both types of secondary structures are characterized by having the main chain NH and C'O groups participating in hydrogen bonds.

2.1.1.1. Alpha(α) Helices. The α-helix is the classic element of the protein structure. It was first described by Linus Pauling in 1951 [37]. The α-helix is a rodlike structure. The inner part of the rod is formed by tightly coiled polypeptide main chain, and the sidechains extend outward in a helical array. (Figure 2.5) The α-helix is stabilized by hydrogen bonds between the C'O and NH residues i and i+4, respectively.

Each residue is related to the next one by translation of 1.5 Å along the helix axis and a rotation of 100° , which gives 3.6 amino acids per turn of helix. α-helices are found when a stretch of consecutive residues all have the phi (ϕ), psi (ψ) angles of - 60° and -50° respectively, using the rotation $\phi = \psi = 180°$ for the *trans* state and 60° and 300° for the *gauche*$^-$ and *gauche*$^+$ states respectively. The pitch of the α-helix is 5.4 Å. α-helices vary considerably in length in globular proteins ranging from 4-5 amino acids to over 30 amino acids. The average length is around 10 residues, corresponding to three turns. Since the rise per residue of an α-helix is 1.5 Å along the helical axis, this corresponds to about 15 Å from one end to the other of an average α-helix. The screw-sense of α-helix can be either right-handed or left-handed; the α-helices found in proteins are right-handed [35, 36]. Ala, Glu, Leu, and Met are good α-helix formers, while Pro, Gly, Tyr, and Ser do not favor helical formation. The most common location for an α-helix in a protein structure is along the outside of the protein with one side of the helix facing the solution and other side facing the hydrophobic core (Figure 2.5.).

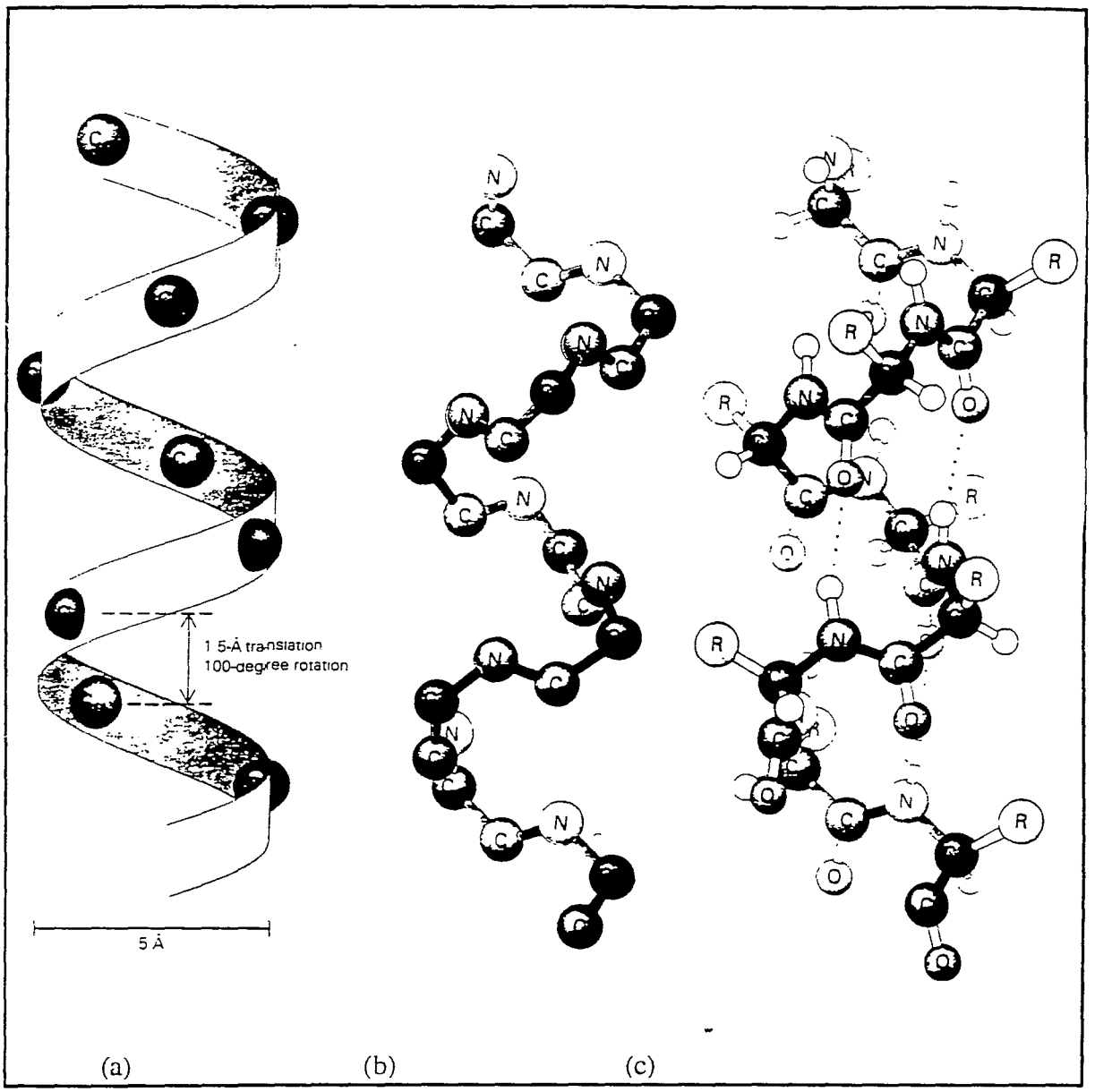


FIGURE 2.5. Three models of a right-handed α-helix. (a) Only the C^{α} atoms are shown. (b) Only the backbone nitrogen (N), α-carbon (C^{α}), and carbonyl carbon (C') atoms are shown. (c) The entire helix is shown. The dots, '.', represent the hydrogen bonds between the NH and CO groups.

2.1.1.2. Beta (β) strands and sheets. The second type of major structural element found in globular proteins is the β-sheet formed by two or more β-strands. β-sheets are built up from a combination of several sequentially distant regions of the polypeptide chain, in contrast to α-helices, which are built up from adjacent amino acids along the sequence. β-strands are usually five to ten residues long, and are in an almost fully extended conformation. They are aligned adjacent to each other such that hydrogen bonds can form between the C'O groups of one strand and the NH groups of an adjacent strand and vice versa. (Figure 2.6)

There are two different forms of β-sheets:

1. The **antiparallel pleated sheet,** in which neighboring hydrogen bonded strands run in opposite directions,

2. The **parallel pleated sheet,** in which the hydrogen bonded strands run in the same direction [38].

β-strands are much more distorted than α-helices, showing gross twisting and kinking, plus the formation of β-bulges. The formation of β-sheets requires long-range interactions, whereas α-helices involve short-range interactions only [39, 40].

2.1.1.3. Nonrepetitive Structures (Loops). The regular secondary structures α-helices and β-sheets comprise around half of residues in globular proteins. The remaining polypeptide segments are said to have a coil or loop conformation. That is not to say, however, that these nonrepetitive secondary structures are any less ordered than α-helices or β–strands, they are simply irregular and hence more difficult to describe [38]. Most proteins are built up from combinations of secondary structures, which are connected by loops.

Loop regions exposed to solvent are rich in charged and polar hydrophilic residues. This has been used in several prediction schemes, and it has proved possible to predict loop regions from amino acid sequence with a higher degree confidence than α-helices and β-strands, which is ironic since the loops have irregular structures.

Since proteins that exhibit sequence homology have similar core structures in general, it is apparent that the specific arrangement of secondary structure elements in the core is rather insensitive to the lengths of the loop

regions. In addition to their functions as connecting units between secondary structure elements, loop regions frequently participate in forming binding sites and enzyme active sites. Loop regions that connect two adjacent antiparallel β-strands are called hairpin loops.

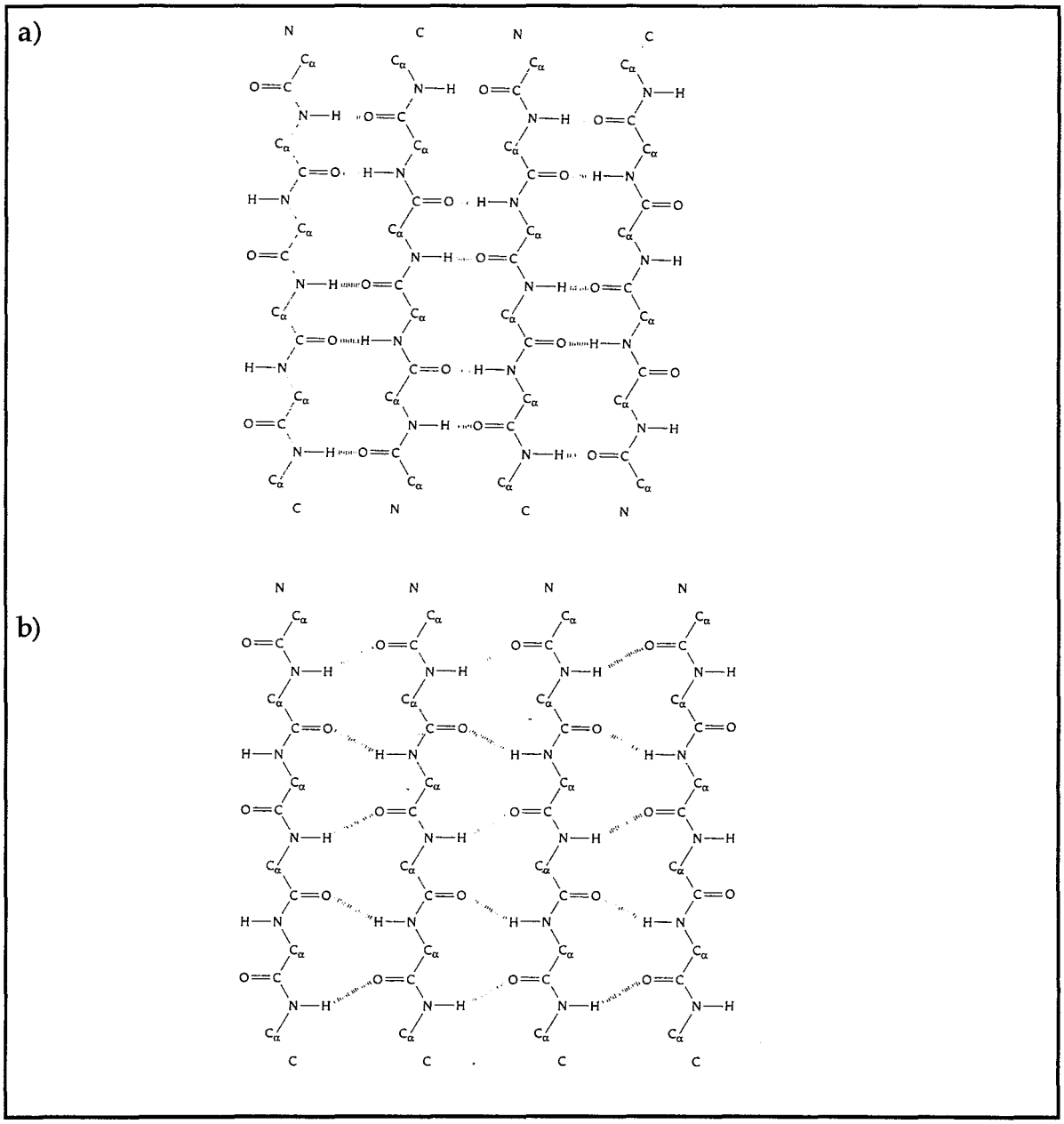


FIGURE 2.6. Schematic diagram showing the hydrogen bond pattern in β-sheets a) antiparallel β-sheet b) parallel β-sheet

3. PROTEIN FOLDING

3.1. Introduction

The biological function of each protein emerges directly from the details of its unique and highly specific three-dimensional structure. By a process known as folding, the long and rather uninteresting organic polymer that emerges from a ribosome spontaneously assumes its biologically active native structure. In many respects, protein folding represents, at the molecular level, the step where life begins, where chemistry makes jump to biology [41].

The prediction of protein folds from their amino acid sequences is an impressively long-standing challenge in molecular biology and biophysics. For over a quarter of a century, ideas on protein folding have been dominated by two inter-related concepts: the Levinthal paradox; and a necessity for folding intermediates. Levinthal argued that, because there is an astronomical number of conformations open to the denatured state of a protein, an unbiased search through these would take "an eternity". It was thus a short logical step to argue that there must be defined pathways to simplify the choices in folding.

Theoretical studies of protein folding have focused on a number of issues: first, what are the sequence requirements for proteins to fold rapidly and be stable in their native conformations? Second, what are the thermodynamic mechanism(s) of protein stabilization and the kinetic mechanism(s) of folding? Third, are there special native structures (structural motifs) that are more likely to correspond to the native structures of foldable proteins? Fourth, what is the best approximation for protein-folding energetics (potentials)? These are inter-related topics which aim at solving the protein folding problem.

3.2. Dominant Forces in Protein Folding

In proteins, the following noncovalent interactions exist: van der Waals interactions, interactions between charged groups (salt links), between polar groups (hydrogen bonds) and between nonpolar groups, so- called hydrophobic interactions.

In aqueous solution, hydrogen bonds are formed not only between the protein polar groups but also between these groups and water. Since the number of charged groups in proteins is low, it is generally assumed that interaction between charged sidechains play a minor role in the stabilization of the native structure, although charged groups might be important in directing protein folding [42]. The stabilization of the native structure is primarily due to the hydrophobic interactions between nonpolar groups. There are many such groups in a given protein and they are clustered together forming the nonpolar core of the protein.

3.3. Classical and New Views on Protein Folding

In the classical view, folding is assumed to proceed via a set of intermediates. In the new view, on the other hand, a protein is considered to be a system with many degrees of freedom, for which entropy is an essential factor in free-energy balance and kinetics. The classical view reflects a chemical understanding of protein folding as being a complex reaction that proceeds by sequential mechanism. In contrast, the new view envisages folding as proceeding via a 'statistical pathway' that features a sequence of multiply populated, kinetically distinguishable macrostates: the unfolded state; intermediate states (if any); the transition state and finally, a more unique (lower entropy) native state. The formation of a transition state results from statistical fluctuations and can occur in an enormous number of ways. The descent from a nucleus transition state to the native conformation can be a deterministic process with a markedly favored pathway.

In contrast to the classical assumption of the necessity of intermediates for the solution of 'Levinthal 's paradox', recent experimental evidence and theoretical analyses suggest that, while there may be cases in which intermediates facilitate folding, they do not necessarily represent a vital feature of protein-folding dynamics, at least for small proteins.

The legitimate question remains of how this understanding of folding kinetics helps to solve the most renowned aspect of protein folding problem-tertiary structure prediction. The answer to this is that the best model which combines the right energetics with computational tractability will enable to find the native structure of the protein.

3.4. Protein Folding Models

The theoretical study of protein folding has relied heavily on computer simulations, although important analytical studies have been carried out as well. Early efforts to model the folding of protein were numerical studies, in which the interactions between amino acids were included in the greatest possible detail. These can be described as 'top-down approaches'. Protein folding occurs, however, on time scales that are computationally unreachable via top-down simulations; therefore, such detailed models cannot be used to study folding, either now or in the foreseeable future.

An alternative approach which proceeds from the bottom up is used to overcome the computational barrier. It starts from the simplest model that still bears some resemblance to a protein, while being complex enough to pose nontrivial theoretical questions and having the potential to reproduce certain fundamental aspects of protein folding. Examples of this strategy are analytical studies using heteropolymer beads on a string models, coarse-grained simulations using lattice or off-lattice models [43]. In some applications these studies yield a visible similarity between the actual folds of small proteins, but usually they face the formidable problem of multiple energy minima and very slow transitions between them.

At the next level simplification, the kinetic approach leads to a sequential scheme of protein structure prediction. One predicts the α helices, β–strands, or other internally more or less stable elements and then tries to pack these "pre-existing" blocks. Within the approximation of this model, the problem apparently reduces to the determination of the optimal organization of the already predicted secondary structural elements in space. The validity of this model depends both on the accurate knowledge of these secondary structural elements and on the sequential folding mechanism associated with the secondary structure content.

An inherent drawback of all the sequential models to study folding is the necessity of discarding many alternative pathways on the initial or intermediate stages of folding without a good guarantee that some advantage obtained later cannot compensate the disadvantage (e.g. lack of stability) of the discarded folding intermediate.

Recent progress in protein-fold recognition is connected with the techniques that search for the native fold as being the most stable one, or, perhaps more appropriately, as being the most probable one.

4. EVALUATION OF THERMAL FLUCTUATIONS IN PROTEINS

4.1. Singular Value Decomposition (SVD) Technique

This is a useful matrix decomposition technique which solves a variety of problems. Among other things, it can determine the numerical rank of a matrix, and can be used to construct a best low rank approximation to a given matrix. The SVD technique has been used in a wide range of applications ranging from satellite image compression to separating out components of kinetics data. Most of these rely upon the basis set the SVD constructs for the data. In fact, the SVD is central to the technique of principal component analysis in statistics.

The basis set that the SVD constructs can be thought of as a set of orthogonal vectors spanning an m-dimensional space, where m is the number of rows of the matrix being decomposed. For a protein, this basis set describes the most principal axes of its structure defined in a least-square sense.

The SVD of a matrix, **A**, is defined as

$$\mathbf{A} = \mathbf{U}\,\Sigma\,\mathbf{V}^{\mathbf{T}} \tag{4.1}$$

where **U** and **V** are orthonormal matrices. The columns of **U**, or $\mathbf{u_i}$, are called the left singular vectors of **A**, the columns of **V**, or $\mathbf{v_i}$, are called the right singular vectors of **A**, and Σ is a nonnegative diagonal matrix whose diagonal elements, σ_i, are the singular values of **A** [44].

4.2. Kirchhoff Adjacency Matrix for Non-bonded Interactions

Correlations between residue pairs can be determined by analyzing the Kirchhoff adjacency matrix of non-bonded interactions. The basic postulate is that the protein in the folded state is equivalent to a three-dimensional elastic network. In the classical theories, the junctions of the network undergo Gaussianly distribution fluctuations under the potential of the pendant chains. The C^{α} atoms are identified with the junctions of the network; and they fluctuate under the potentials of their near-neighbours. Thus, the interaction between closely located pairs of C^{α} atoms substitute for the harmonic potentials constraining the end-to-end separation of the classical network chains.

Following the model network of randomly fluctuating junctions, the fluctuations ΔR_{ij} in the separation $R_{ij} = |R_i - R_j|$ between the i^{th} and j^{th} C^{α} atoms in the folded protein are assumed to obey the Gausssian distribution

$$W(\Delta R_{ij}) = (\gamma^*/\pi)^{3/2} \exp(-\gamma^* \Delta R_{ij}^2) \tag{4.2}$$

Here the normalization constant γ^* is the counterpart of the single parameter in the Hookean potential adopted by Tirion [45]. In the statistical theory of polymer networks the distribution function $W(\Delta R_{ij})$ is substituted into the expression $\Delta A = - k_B T \ln W(\Delta R_{ij})$ for the elastic free energy change associated with the fluctuation ΔR_{ij}. This yields the harmonic potential $k_B T \gamma^* \Delta R_{ij}^2$ and Hookean force constant equal to $2k_B T\gamma^*$. The configurational partition function for a protein of N residues may then be expressed, with analogy, to the theory of random Gaussian networks, as:

$$Z_N = K \exp(-\{\Delta R^T\} \Gamma \{\Delta R\}) \tag{4.3}$$

Here $\{\Delta R\}$ is the 3N dimensional column vector formed by the fluctuations $\{\Delta R_1, \Delta R_2, \Delta R_3, ...\Delta R_n\}$ of the C^α atoms, the superscript T denotes the transpose, K is a constant and Γ is a symmetric matrix, known as the Kirchhoff or valency adjacency matrix in graph theory. The elements of Γ are given by

$$\Gamma_{ij} = \left\{ \begin{array}{c} -\gamma^* \text{ if } i \neq j \text{ and } R_{ij} \leq r_c \\ 0 \text{ if } i \neq j \text{ and } R_{ij} \geq r_c \\ \\ -\sum_{i \neq j} \Gamma_{ij} \quad \text{if } i = j \end{array} \right\} \tag{4.4}$$

The summation for evaluating Γ_{ij} is performed overall off-diagonal elements on the i^{th} column (or row). r_c is the cuttoff separation defining the range of non-bonded contacts. The equilibrium correlations between the fluctuations of two sites k and l are obtained from

$$\langle \Delta R_k . \Delta R_l \rangle = (1/Z) \int \Delta R_k . \Delta R_l \, e^{-V/kT} \, d\{\Delta R\} \tag{4.5}$$

$$= \int \Delta R_k . \Delta R_l e^{-V/kT} d\{\Delta R\} / \int e^{-V/kT} d\{\Delta R\} = [\Gamma^{-1}]_{kl} \tag{4.6}$$

where $V = \{\Delta R^T\}\Gamma \{\Delta R\}$ is the potential associated with the vibrations of the C^α atoms, Z is the partition function, the angular brackets designate the ensemble average over all fluctuations, $d\{\Delta R\} = d\Delta R_1 d\Delta R_2 .....d\Delta R_m$, and $[\Gamma^{-1}]_{kl}$ is the kl^{th} element of the inverse of Γ. Γ may be regarded as the atomistic counterpart of the stiffness matrix in the analysis of elastic bodies. The mean-square fluctuations of the C^α atoms are readily evaluated from the diagonal

elements of Γ^{-1} using $\langle R^2 \rangle = [\Gamma^{-1}]_{kk}$ and the cross-correlations between the fluctuations of the C^{α} atoms are found from the off-diagonal entries of Γ^{-1}.

The determinant of the Kirchhoff matrix is equal to zero, hence the matrix cannot be inverted directly. Instead, the matrix is subjected to a similarity transformation, and reconstructed after eliminating the zero eigenvalue. The eigenvalue decomposition of the Kirchhoff matrix Γ of a protein of n residue reads

$$\Gamma = U \Lambda U^T \tag{4.7}$$

$$\Gamma^{-1} = U(\Lambda^{-1})U^T \tag{4.8}$$

Here U is an orthogonal matrix whose columns u_i, $1 \leq i \leq n$, are the eigenvectors of Γ and Λ is a diagonal matrix whose elements are the squared eigenvalues (λ_i), usually organized in ascending order $\lambda_1 = 0 < \lambda_2 < ... < \lambda_n$. λ_i represents the frequency of the i^{th} mode of the relaxation [30].

When the mean-square fluctuations of the C^{α} atoms and the cross-correlations between the fluctuations of the C^{α} atoms are obtained, according to the method mentioned above, it is observed that the more localized atoms exhibit relatively low amplitude fluctuations while chain termini and loop regions undergo slower and larger amplitude fluctuations. On the basis of this knowledge and method, the flexible parts of the proteins are identified in the present study and will be used as input in simulations as will be presented in the Chapter 5.

5. METHOD FOR THE ESTIMATION OF THE TERTIARY STRUCTURE OF PROTEINS

5. 1. General Approach

In this study, the aim is to generate sets of decoy conformations for specific proteins and to identify the correct native structure among them. A basic postulate underlying the whole approach is that native structure is the most stable state conformation, i.e. the one corresponding to the lowest free energy. Knowledge-based potentials which are extracted from a database of known protein structures, the Brookhaven Protein Data Bank [46], are used in calculations.

The success of this approach depends critically on the quality and quantity of the decoys. Decoy structures must: 1) include structures that are close to native x-ray structure; 2) be native-like in all properties of the real polypeptide chain except the overall folded conformations, otherwise they can be easily distinguished by trivial tests; 3) be diverse so as to sample all possible arrangements and 4) be numerous for more sensitive testing.

Large ensembles (hundreds of thousands) of test structures will be generated here for eight small proteins of $54 < N < 76$ residues. These proteins have a variety of different folds, and include all the well-defined secondary structures. The decoys are native-like in that they are all constrained to have native secondary structure. With this approach, the protein folding problem reduces to the determination of optimal organization of the already predicted secondary structural elements in space.

After generating the decoy structures, the objective is to identify the lowest energy conformation of the proteins using knowledge based potentials.

5. 2. Model and Method

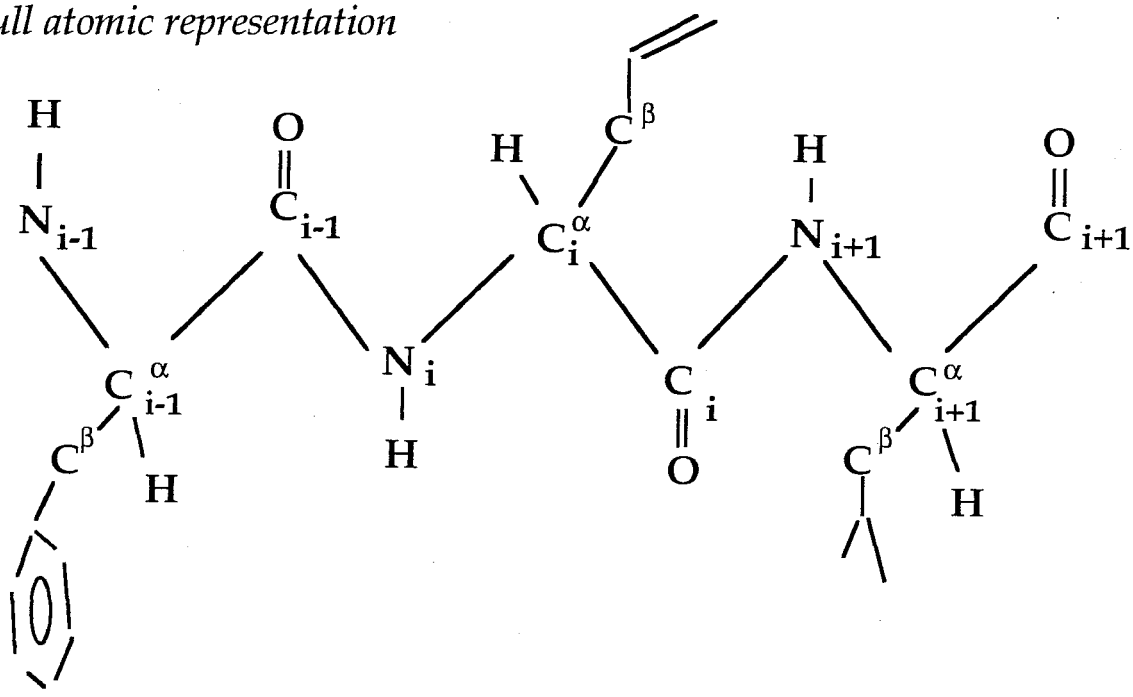
5.2.1. Dataset of Tested Proteins

The set of proteins used in calculations is presented in Table 5.1. The set comprises eight proteins, which were previously considered by Park and Levitt [1] in their exhaustive enumeration of protein conformations. The 3-dimensional structures of these proteins were determined at more than 1.2 Å resolution by X-ray crystallography. The Brookhaven Protein Data Bank (PDB) identifier, the size and the structural class of each protein is given in the table. All of these proteins are structurally distinct, except for 434 Cro protein (2cro) and 434 Repressor (1r69) which are homologous but distant in sequence.

5.2.2. Low Resolution Model

The backbone of each protein is represented by the virtual bond model originally proposed by Flory and collaborators [47]. Accordingly, the backbone of a protein of n residues consists of n-1 virtual bonds connecting successive α-carbons. l_i designates the bond vector of magnitude l_i pointing from the i^{th} α-carbon, C^{α}_i, to the $(i+1)^{th}$, C^{α}_{i+1}; θ_i is the bond angle between l_i and l_{i+1}, and ϕ_i is the torsional angle defining the rotation about bond l_i, as illustrated in Figure 5.1. The sidechain of each residue i, on the other hand, is represented by a single site, S_i, which is specific to the type of the amino acid. S_i is found from the centroid of either all sidechain atoms, or a few specific atoms, depending on the hydrophobicity or polarity of the amino acid, in conformity with the descriptions given in a recent study [8, 48]. The sidechain virtual bond l^s_i connects S_i to C^{α}_i. θ^s_i is the bond angle between l_i and l^s_i, and ϕ^s_i is the sidechain dihedral angle defined by the three consecutive bonds l_{i-1}, l_i and l^s_i. So, the set $\{l_i, \theta_i, \phi_i, l^s_i, \theta^s_i, \phi^s_i\}$ defines the position of the i^{th} residue in space.

(a) Full atomic representation



(b)Virtual bond representation

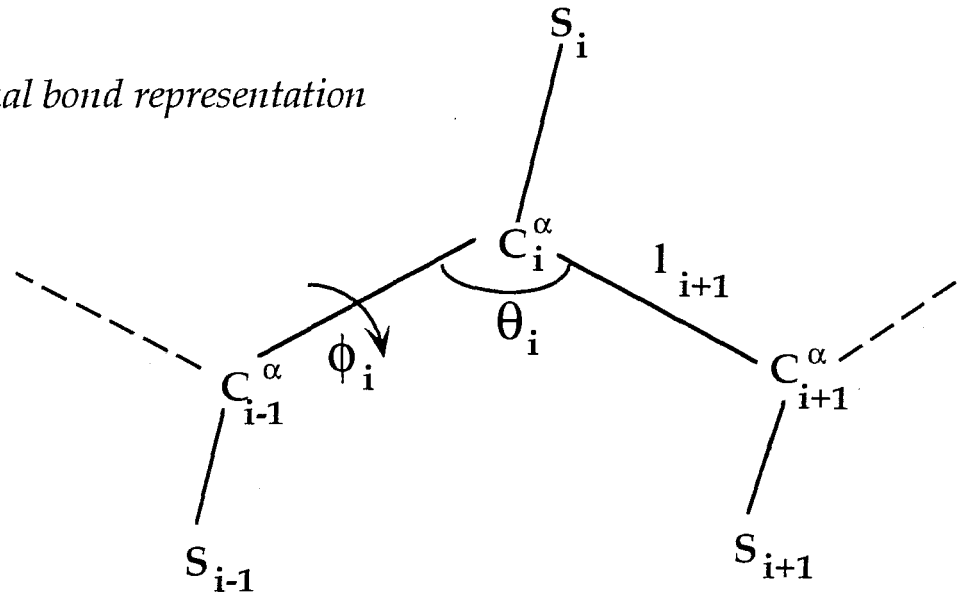


FIGURE 5.1. Full atomic and virtual bond representations of a protein segment of three amino acids. In part (a) the full atomic representation is displayed , (b) In the virtual bond model two points per residue are considered: the alpha carbon (C^{α}_i) as the backbone site, and the centroid of the side chain, S_i, as the sidechain interaction site. l_i is the virtual bond connecting C^{α}_iandC^{α}_{i-1}, ϕ_i is the torsion angle of bond l_i, θ_i is the bond angle between virtual bonds i and i+1.

TABLE 5.1 Information about the proteins considered in the present study.

PDB code	Protein Name	Resolution (Å)	Number of Residues	Structural Class	Reference
4rxn	Rubredoxin	1.2	54	--	[49]
4pti	Trypsin inhibitor	1.5	58	α+β	[50]
1r69	434 Repressor	2.0	69	α	[51]
2cro	434 Cro Protein	2.35	71	α	[52]
1sn3	Scorpion Neurotoxin	1.8	65	α+β	[53]
1ctf	Ribosomal Protein	1.7	74	α+β	[54]
3icb	Calcium Binding Protein	2.3	75	α	[55]
1ubq	Ubiquitin	1.8	76	α+β	[56]

5.2.3. Determination of ResiduePositions

In their local frames, the position vectors of backbone (C^{α} atoms) and sidechain sites for all residues are denoted as respectively

$$\mathbf{l}_i = \begin{bmatrix} l_i \\ 0 \\ 0 \end{bmatrix} \qquad \mathbf{r}_i^s = \begin{bmatrix} l_i^s \\ 0 \\ 0 \end{bmatrix} \tag{5.1}$$

The local frame of the first bond of the backbone chain (l_2) is taken as the main frame and all position vectors of the rest are expressed with respect to this frame by using transformation matrices T_i. The transformation matrices used in the present calculations are [57, 58]

$$\mathbf{T_i} = \begin{bmatrix} \cos\theta_i & \sin\theta_i & 0 \\ \sin\theta_i\cos\phi_i & -\cos\theta_i\cos\phi_i & \sin\phi_i \\ \sin\theta_i\sin\phi_i & -\cos\theta_i\sin\phi_i & -\cos\phi_i \end{bmatrix} \tag{5.2}$$

$$\mathbf{T_i^s} = \begin{bmatrix} \cos\theta_i^s & \sin\theta_i^s & 0 \\ \sin\theta_i^s\cos\phi_i^s & -\cos\theta_i^s\cos\phi_i^s & \sin\phi_i^s \\ \sin\theta_i^s\sin\phi_i^s & -\cos\theta_i^s\sin\phi_i^s & -\cos\phi_i^s \end{bmatrix} \tag{5.3}$$

Here T_i is the transformation matrix that converts vectorial quantities of the $i+1^{th}$ bond into their representation in the frame of the i^{th} bond, and T_i^s is the transformation matrix which is used to express the position of the side chain of the residue S_i with respect to the frame of the i^{th} backbone bond. The torsional angle of the *trans* state is defined here as 0^o.

The first C^α atom is placed on the origin of the main frame, so the position vector of the first backbone atom is defined as:

$$\overrightarrow{r_1^\alpha} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{5.4}$$

and the position of the second backbone bond is:

$$\overrightarrow{r_2^{\alpha}} = l_1 + T_1 l_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + T_1 \begin{bmatrix} l_2 \\ 0 \\ 0 \end{bmatrix} \tag{5.5}$$

where T_1 and T_2 are transformation matrices and l_1 and l_2 are the position vectors of the first and the second backbone bonds in their local frame. So the position of the N^{th} backbone site is computed using :

$$\overrightarrow{r_N^{\alpha}} = 1_1 + T_1 l_2 + \ldots\ldots\ldots\ldots + [T_1 T_2 \ldots\ldots T_{N-1}]\, l_N \tag{5.6}$$

In the same manner, the position vector of each sidechain S_i is expressed with respect to the main frame by using the serial multiplication of transformation matrices. So the position vector of the third and N^{th} sidechains, for example, are found from

$$\overrightarrow{r_3^{s}} = l_1 + T_1 l_2 + T_1 T_2\, T_3^{s}\, l_3^{s} \tag{5.7}$$

$$\overrightarrow{r_N^{s}} = 1_1 + T_1 l_2 + \ldots\ldots\ldots\ldots + [T_1 T_2 \ldots\ldots T_{N-1} T_N^{s}]\, l_N^{s} \tag{5.8}$$

5.2.4. Conformational Sampling Technique

The aim is to determine the lowest energy structure in a set of coarse-grained conformations generated by an exhaustive enumeration technique. In the present study, conformations are generated by rotating the backbone virtual bonds at fixed intervals within their full range $-180 \leq \phi_i \leq 180°$. We note, however, that intervals of 30°, for example, lead to $\nu = 12$ rotational states per residue, and consequently $\nu^{n-2} \approx 12^{52}$ states to be enumerated even for the smallest protein of 54 residues in our set, which is not computationally possible.

A significant reduction in conformational space is achieved by assuming that some protein segments, such as those participating in secondary structures, are rigidly formed at early stages of folding and thereby possess sufficient internal stability to be held rigid in simulations. This approach has proven useful in recent simulations [1, 4, 7, 34]. About ten flexible residues, generally located in loop regions, may conveniently be chosen to enumerate $\nu^{12} \approx 10^{12}$ conformations for each protein. Since this approach is the same in spirit as that of Park and Levitt [1], the set of flexible residues adopted in their analysis will also be used here. Accordingly, each protein is composed of five or six linear segments, separated by four or five hinges, as schematically shown in Figure 5.2.

In an alternative method, we determined the most flexible residues of each protein by a Gaussian dynamics method described in Chapter 4, and all calculations are repeated upon adopting the new set of flexible residues. The Gaussian dynamics method identifies the residues which undergo the largest amplitude RMS fluctuations in a given protein in native state, after filtering out all vibrational modes, except a few slowest ones. The resulting set of flexible residues are displayed in Figure 5.3. The flexible residues are generally located at the loop regions or turns between α-helices or β-strands, or at the termini of helices.

In order to reduce further the ensemble size, we applied two constraints before proceeding to the energetic evaluation of the conformations. (1) Conformations that violate the excluded volume principle are discarded. A

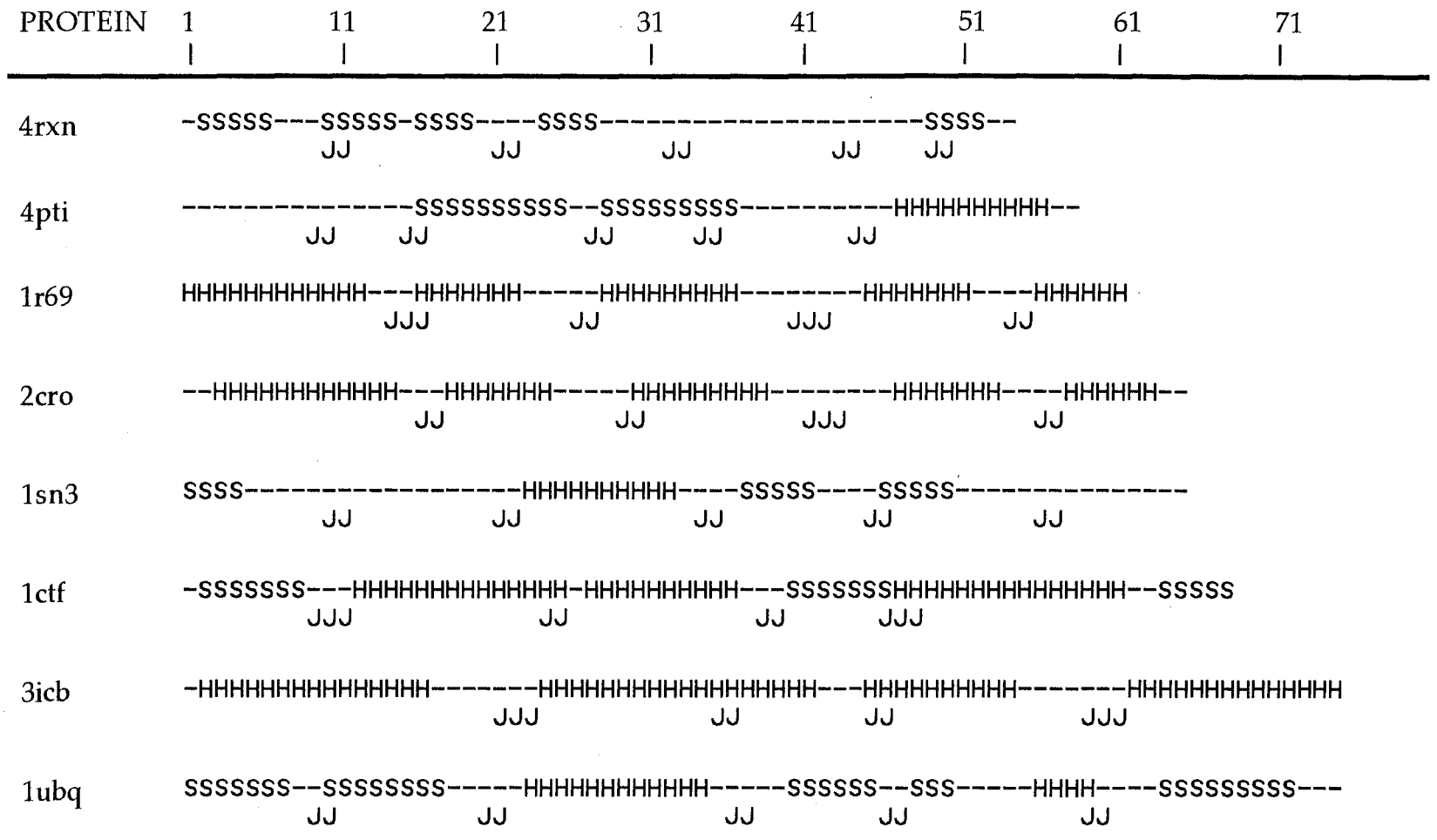


FIGURE 5.2 Secondary structures and flexible residues of the proteins. The secondary structure is indicated by S: β-strand, H: α-helix or - : neither. Hinge residues are marked with J below the secondary structure. Flexible residues are taken from the study of Park and Levitt [1].

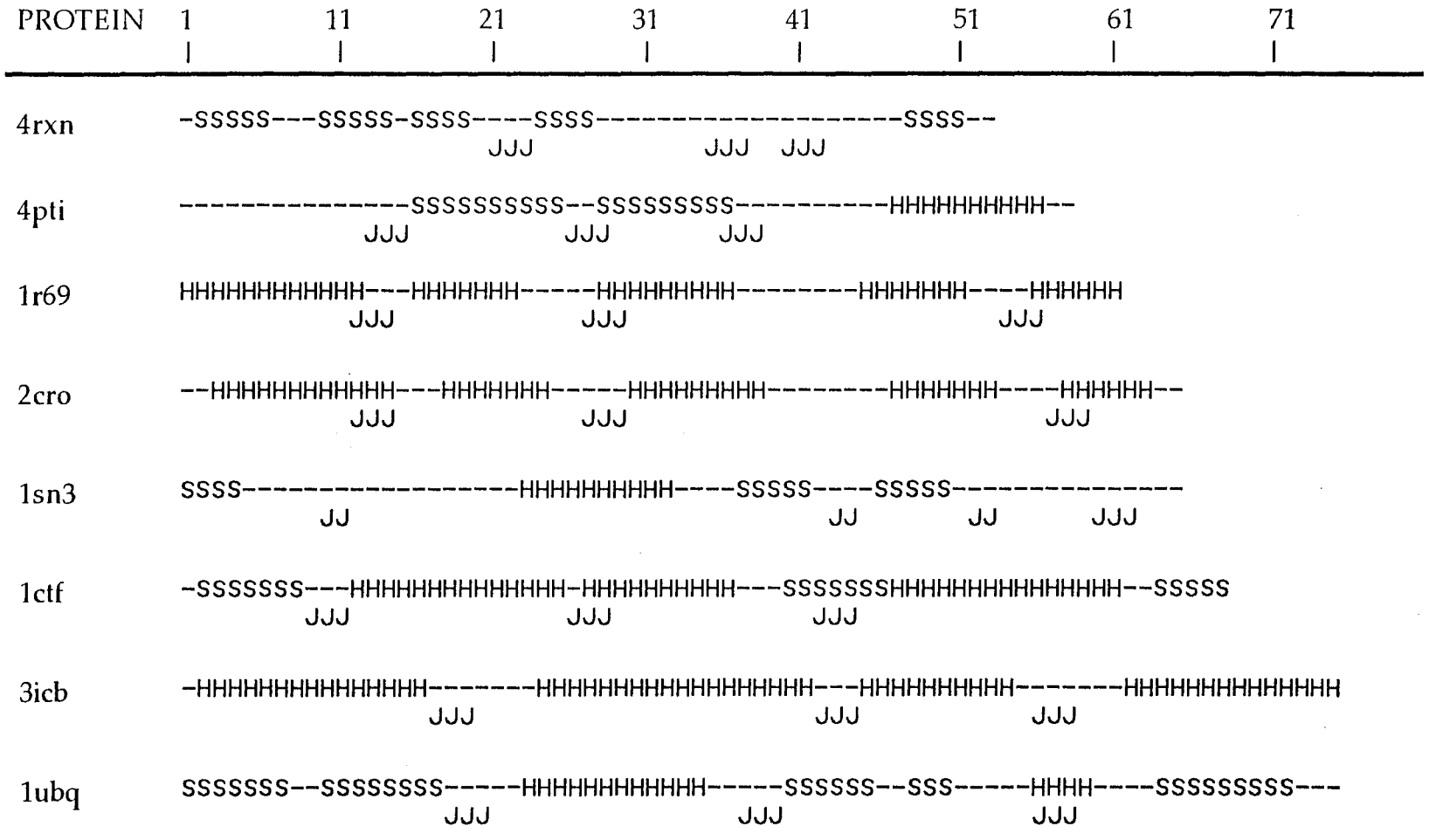


FIGURE 5.3 Secondary structures and flexible residues of the proteins. The secondary structure is indicated by S : β-strand, H: α-helix or - : neither. Hinge residues are marked with J below the secondary structure. Flexible residues are found by the Gaussian dynamics method.

threshold value of 2.0 Å is adopted for the closest distance of approach between two interaction sites, conformations giving rise to closer interactions being eliminated. (2) The radius of gyration R_g of the generated protein segments is verified to obey the expression [8]:

$$\log R_g^2 = (2/3) \log n + 0.92 \tag{5.9}$$

with an error limit of $\Delta[\log R_g^2] = \pm 0.2$ Å. Here n is the number of residues. Conformations violating this criterion are eliminated.

5.2.5. Energy Evaluation

The effective potential is taken from recent studies of Bahar and Jernigan [8], in which 302 non-homologous proteins from the PDB were used for extracting empirical parameters. The potential consists mainly of two types of contributions: (i) long-range interactions (LR). These take place between interaction site separated by five virtual bonds, (ii) short-range (SR) interactions. These include the interactions between near neighbours along the chain.

$$E_{total} = E_{LR} + E_{SR} \tag{5.10}$$

The three contributions of E_{LR} are given by :

$$E_{LR} = \sum_{i=1}^{N-5} \sum_{j=i+5}^{N-5} E_{BB}(r_{ij}) + \sum_{i=1}^{N-3} \sum_{j=i+3}^{N} E_{SS}(r_{ij}) + \sum_{i=1}^{N-4} \sum_{j=i+4}^{N} E_{SB}(r_{ij}) \tag{5.11}$$

Where, $E_{BB}(r_{ij})$ is the potential of mean force between two backbone units $C\alpha_i$ and $C\alpha_j$, r_{ij} is the distance between these groups; $E_{SS}(r_{ij})$ is the potential between side groups S_i and S_j; and $E_{SB}(r_{ij})$ is the one between the side group and the backbone interaction centers of the i^{th} and j^{th} residues.

The short-range conformational energy E_{SR} is calculated from

$$E_{SR} = \sum_{i=1}^{10} [E(\phi_i^-)/2 + E(\phi_{i-1}^+)/2 + [E(\phi_i^-, \phi_i^+)] + \sum_{i=1}^{10} [\Delta E(\theta_i,\phi_i^-) + \Delta E(\theta_i,\phi_i^+)] \qquad (5.12)$$

where the first three terms are the potential associated with the torsion of the flexible virtual C^α bonds. The fourth, and the fifth terms account for the coupling between the bond angles and torsions. Potentials of the bond angle distortions are not considered since bond angles are kept constant. The short-range interaction energy is evaluated only for the flexible residues, that is to say only for ten specific residues. Because of this reason, the contribution of the short-range interaction energies to the overall energy is expected to be relatively small.

5.2.6. Application of the Procedure to Rubredoxin (4rxn)

As mentioned before, the eight proteins are divided into rigidly held segments. The consecutive segments are connected by hinge regions composed of two or three flexible residues. First, complete enumeration is performed over all conformations of these segments. For rubredoxin (4rxn), the first segment lies between residues 2 and 21; Gly10 and Tyr11 form the flexible region. The number of possible conformations for this segment is 12^2, however the ensemble size decreases to 21 when two filters (R_g and hard core) are applied. The same procedure is performed for each segment of the protein

and independent ensembles of different conformations for every segment are obtained.

The exhaustive enumeration for each segment is performed using large intervals such as 30°. In order to get the native-like structure, the computations are repeated with smaller intervals (10°) in the neighborhood of the values obtained from the previous computations as a second step. The virtual bonds are rotated around their most favorable dihedral angles within the range of $-15 < \Delta\phi < +15$ by 10° intervals. In this way, nine choices are obtained for the pair of residues at the hinge region. Among these nine choices, the one which leads to a radius of gyration closest to the theoretically expected one (see eq. 5.9) is accepted. Table 5.2 shows the different segments and the number of possible conformations which satisfies the R_g and hard core restrictions for rubredoxin (4rxn).

TABLE 5.2. The segments, the flexible residues, and the number of possible conformations surviving after the hard core and radius of gyration filters for rubredoxin (4rxn)

Segment	Residues	Flexible residues	no of tested conformations	no of retained conformations
1	2-21	Gly 10, Tyr 11	144	21
2	11-33	Asp 21, Asp 22	144	27
3	22-43	Ile 33, Pro 34	144	16
4	34-49	Gly 43, Val 44	144	61
5	44-54	Phe 49, Glu 50	144	138

At the third step, segments are gradually combined. Initially, the first and the second segments are bonded. It is like a puzzle game, we seek for the best-fit conformation which obeys the R_g and hard core restrictions by

combining the most favorable independent conformations of the first and the second segments. In 4rxn, the number of possible conformations for the first and second segments were 21, 27 respectively. When these two segments are joined, the number of generated conformations becomes 21x27=567. However, among these 567 conformations, those which satisfy the hard core and radius of gyration filters are retained, such that finally 105 conformations are obtained. Then, the third segment is united to the combined segments 1&2, in the same manner. However, as the successive segments are added, the number of accessible conformations increases rapidly and the method becomes computationally inefficient. An energy barrier is used to decrease the number of possible conformations, when the fourth and fifth segments are added. Accordingly, the conformation whose overall energies are higher than -2.0nRT are not accepted. Finally, an ensemble of decoy conformations is obtained for the protein. The conformation with the lowest energy is sought in this set. The number of retained conformations during the gradual combinations of the segments are presented in Table 5.3.

TABLE 5.3. The number of tested and retained conformations of rubredoxin (4rxn) during the gradual combination of the segments.

Segment	Residues	no of tested conformations	no of retained conformations
1&2	2-33	24 x 27	105
12&3	2-43	105 x 16	651
123&4	2-49	651 x 61	113
1234&5	2-54	113 x 138	3570

In the second part of the study, flexible regions are determined using the connectivity and correlation matrices. Then, the same procedure is used for generating decoy conformations rotating the new flexible bonds. However, the optimization procedure in which the enumeration is repeated for the smaller intervals in the neighborhood of the values obtained from the previous

enumerations is neglected. Instead of rotating the flexible virtual C^{α} bonds by 30^{o} intervals, 10^{o} intervals are used, so that each bond has 36 possible torsional states. Since the ensemble size is larger than the ones with 30^{o} intervals, more severe radius of gyration filter is applied. In this part, an error limit of $\Delta[\log Rg^2] = \pm 0.05$ is used.

6. RESULTS AND DISCUSSIONS

6. 1. Evaluation of Generated Conformations

The aim of this study was to generate a large set of protein structures for each of the eight test proteins using a low resolution model and identify the lowest energy conformation. As explained in the preceeding chapters, exhaustive enumeration technique was performed for nine or ten carefully chosen flexible residues for each protein. The flexible residues were generally found in the regions between or at the ends of secondary structures. The size of these ensembles were reduced by using two filters: radius of gyration and excluded volume (hard core).

After generating a complete set of conformations which satisfies the hard core and radius of gyration constraints, the correct and incorrect folds are differentiated on the basis of knowledge-based potentials. The knowledge-based potentials are taken from recent studies of Bahar and Jernigan [8]. In order to distinguish the most native-like fold in the set of different conformations, two criteria are used: Root-mean-square (RMS) deviations with respect to native state and energies of the generated conformations.

To quantify the similarity between the generated structures of the proteins and those already measured by x-ray and NMR, the following root-mean-square deviation (RMS) value is used:

$$RMS = \left[\frac{\sum_{i=2}^{N} |r_i - r_i^o|^2}{N-1} \right]^{1/2} \qquad (6.1)$$

where r_i is the position of the atom i in the test structure and r_i^o is its counterpart in the native structure, given that the two have been optimally superimposed [59].

The low energy conformations obtained for each protein were sorted for two purposes: i) to discriminate the conformation with the lowest energy, since it is often assumed that the native structure is at a global free-energy minimum [60], and ii) to determine the rank of the most native-like structure, i.e. the one with the RMS deviation with respect to experimentally determined structure

6.2 Simulation Results and Discussions

6.2.1. Comparison of Lowest Energy Conformations Obtained in Simulations with X-ray Structures

In order to examine which kind of interaction energy plays a crucial role in discriminating the correct folds from incorrect folds, three types of interaction energies are considered: i) Overall energy including both long-range and short-range interactions (E_{total}); ii) Only long-range interaction energies (E_{LR}); iii) non-specific backbone-backbone (E_{BB}) energies alone among the three types, S-S (sidechain-sidechain), S-B (sidechain-backbone), and B-B (backbone-backbone) of long-range potentials.

The native energies, the lowest energy attained by the coarse-grained chain generations for each protein and their RMS deviations with respect to x-ray structures are presented in Tables 6.1-6.3. Energies are evaluated on the basis of total interactions (including long-range and short-range interactions), long-range interactions and non-specific backbone-backbone interactions in the three respective tables. The flexible residues are taken from the study of Park and Levitt [1].

The most striking result is that the lowest energy conformation obtained for each protein exhibits a rather low (< 2.0 Å) RMS deviation with respect to x-ray conformations, except trypsin inhibitor (4pti). Examination of the lowest energy conformations evaluated according to total, long-range and non specific backbone-backbone energies shows that the effect of short-range interaction energies is negligible. In fact, the lowest energy conformations with respect to total and long-range interactions are identical for all of the examined proteins except 4pti.

TABLE 6.1 Energy and RMS deviations of the conformations which have lowest energy for each of the eight proteins. Both long-range and short-range contributions are included. Flexible parts are adopted from the study of Park and Levitt [1].

PDB name	Energy (RT) of native-state	Lowest energy (RT) obtained in simulations	RMS deviation (Å) with respect to x-ray structure
4rxn	-182.8	-168.5	1.91
4pti	-159.1	-121.4	7.91
1r69	-263.6	-193.3	0.99
2cro	-239.8	-197.8	1.17
1sn3	-188.6	-166.8	1.77
1ctf	-322.9	-261.8	1.05
3icb	-280.6	-241.8	0.76
1ubq	-328.1	-240.4	1.40

TABLE 6.2. Energy and RMS deviations of the conformations with the lowest energy. Only long-range interaction energies are considered The flexible parts are taken from the study of Park and Levitt [1].

PDB name	Energy (RT) of Native-state	Lowest energy (RT) obtained in simulations	RMS deviation (Å) with respect to x-ray structure
4rxn	-231.4	-204.8	1.91
4pti	-230.7	-182.8	6.63
1r69	-345.9	-309.1	0.99
2cro	-337.1	-289.4	1.17
1sn3	-284.7	-252.7	1.77
1ctf	-358.6	-297.5	1.05
3icb	-368.2	-322.4	0.76
1ubq	-381.1	-301.5	1.40

TABLE 6.3. Energy and RMS deviations of the conformation with the lowest energy determined on the basis of backbone-backbone interaction energies, (E_{BB}). The flexible parts are adopted from the study of Park and Levitt [1].

PDB name	Energy (RT) of Native-state	Lowest energy (RT) obtained in simulations	RMS deviation (Å) with respect to x-ray structure
4rxn	-92.8	-99.1	1.91
4pti	-111.9	-111.3	2.19
1r69	-144.3	-153.1	8.63
2cro	-157.8	-155.5	2.09
1sn3	-133.1	-128.4	6.59
1ctf	-153.8	-147.2	1.00
3icb	-168.0	-165.7	0.76
1ubq	-159.5	-149.5	3.19

According to Table 6.3, it is clear that the non-specific backbone-backbone energies are not able to distinguish the correct fold from the incorrect folds, since the RMS deviations of the conformations with the lowest backbone-backbone energies are relatively high, except for the protein 4pti. For 4pti, among the lowest energy conformations evaluated on the basis of E_{total}, E_{LR}, E_{BB} the relatively low RMS deviation is reached when backbone-backbone energies (E_{BB}) are taken into consideration. The positions of the sidechains play a crucial role in energy evaluation. When the conformations of the sidechains are not close to the native conformations as in the case of 4pti, the RMS deviation between the generated conformation and the native structures becomes larger. It is interesting to note that the lowest energy conformations identified for ribosomal protein (1ctf), and calcium binding protein (3icb) on the basis of backbone-backbone interactions are identical to those already listed in Tables 6.1 and 6.2.

6.2.2 Examination of Conformations Exhibiting the Lowest RMS Deviations from X-ray Structure

The total (E_{total}), long-range (E_{LR}), and non-specific backbone-backbone (E_{BB}) energies; the ranks, and the RMS deviations for the most native-like structures (which have the lowest RMS deviation from x-ray structure) are presented in Tables 6.4.-6.6 respectively. The flexible parts are adopted from the study of Park and Levitt [1].

The most fascinating result that is observed from these tables is that the ranks of the proteins 434 Repressor (1r69), calcium binding protein (3icb), ribosomal protein (1ctf), ubiquitin (1ubq) are 1. This means that the conformation with lowest energy is exactly the most native-like structure. This also proves the usefulness of the model and the energy parameters. By this method, the tertiary structures of the globular proteins can be predicted provided that the secondary structures and flexible regions are precisely known.

TABLE 6. 4. Rank and energies of the conformations with the lowest RMS deviation from x-ray structure.

PDB name	RMS deviation (Å) of the nearest-native conformation	Rank	Total energy (RT) of the nearest-native conformation
4rxn	0.77	50	-141.6
4pti	1.65	25	-103.9
1r69	0.99	1	-193.3
2cro	1.1	3	-186.4
1sn3	0.47	32	-131.4
1ctf	1.05	1	-261.8
3icb	0.76	1	-241.8
1ubq	1.40	1	-240.4

TABLE 6. 5. Rank and energies of the conformations with the lowest RMS deviation from x-ray structure.

PDB name	RMS deviation (Å) of the nearest-native conformation	Rank	Long-range energy (RT) of the nearest-native conformation
4rxn	0.77	11	-204.3
4pti	1.65	35	-165.1
1r69	0.99	1	-309.1
2cro	1.1	3	-276.6
1sn3	0.47	25	-227.5
1ctf	1.05	1	-297.5
3icb	0.76	1	-322.4
1ubq	1.40	1	-301.5

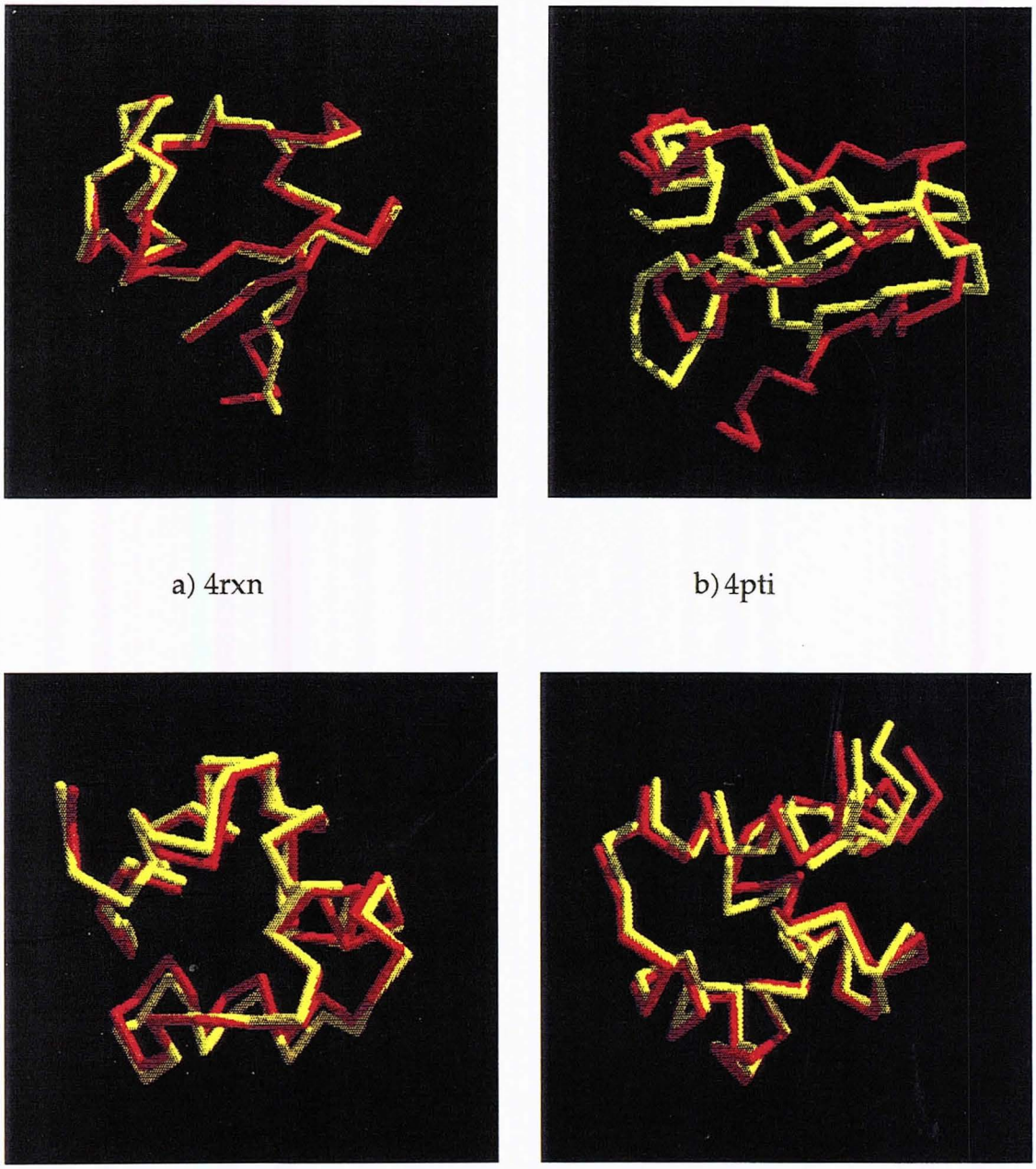
TABLE 6. 6. Rank and energies of the conformations with the lowest RMS deviation from x-ray structure.

PDB name	RMS deviation (Å) of the nearest-native conformation	Rank	Backbone-backbone energy (RT) of the nearest -native conformation
4rxn	0.77	6	-95.6
4pti	1.65	3	-109.8
1r69	0.99	44	-138.9
2cro	1.1	406	-134.8
1sn3	0.47	5	-123.2
1ctf	1.05	18	-141.7
3icb	0.76	1	-165.7
1ubq	1.40	12	-129.9

The conformations with the lowest energy (on the basis of long-range interactions) and the corresponding native state which are optimally superimposed are presented for all proteins in Figures 6.1-6.2. These figures illustrate how well the knowledge-based potentials differentiate the correct folds, since the structures of the conformations with the lowest energy almost identically coincide with the native structure, except for trypsin inhibitor (4pti)

6.2.3. Calculations Using Flexible Regions Identified by Gaussian Dynamics

In the second part of the study, flexible regions are identified by the Gaussian dynamics method described in Chapters 4 and 5, and same computations are performed for each protein using the newly determined flexible regions. Energies of the conformations are again evaluated according to the total, long-range, and non-specific backbone-backbone potential energies, respectively. Results are presented in Tables 6.7-6.12. Tables 6.7-6.9 displays the lowest energy conformations and their RMS deviations from x-ray structure. Tables 6.10-6.12 list the conformations with the lowest RMS deviation and their corresponding rank and energies.

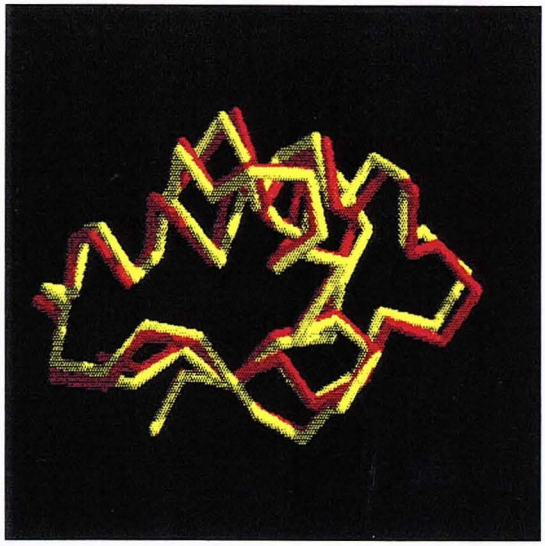


a) 4rxn b) 4pti

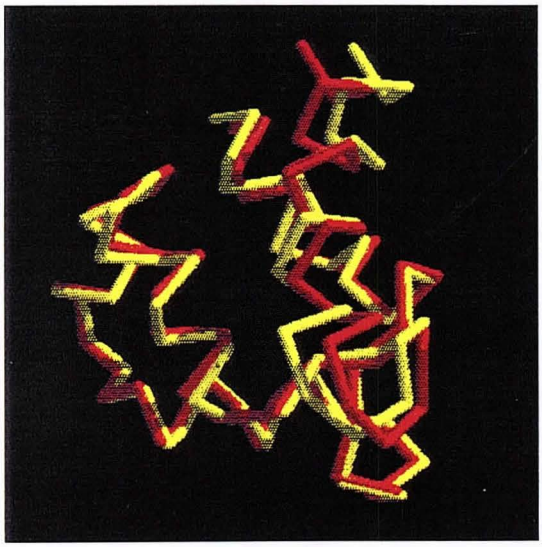
c) 1r69 d) 2cro

FIGURE 6.1. The superimosed structure of the proteins a) 4rxn, b) 4pti, c) 1r69, d) 2cro. The yellow structure is the native structure and the red one is the predicted tertiary structure. The 3 - D structure of the backbone is shown. Flexible parts are taken from Park and Levitt [1].

a) 1sn3



b) 1ctf



d) 3icb

d) 1ubq

FIGURE 6. 2. The superimosed structure of the proteins a) 1sn3, b) 1ctf, c) 3icb, d) 1ubq. The yellow structure is the native structure and the red one is the predicted tertiary structure. The 3 - D structure of the backbone is shown. Flexible parts are adopted from Park and Levitt [1].

TABLE 6.7. Energies and RMS deviations of the conformations with the lowest total energy (E_{total}). The flexible parts are calculated using the Gaussian Dynamics method.

PDB name	Energy (RT) of Native-state	Lowest energy (RT) obtained in simulations	RMS deviation (Å) with respect to x-ray structure
4rxn	-209.5	-167.1	6.92
4pti	-185.9	-156.4	1.53
1r69	-335.3	-312.3	0.78
2cro	-303.4	-290.9	0.75
1sn3	-222.2	-181.3	3.51
1ctf	-289.1	-280	0.70
3icb	-280.6	-171.9	1.41
1ubq	-354.0	-240.1	1.36

TABLE 6.8. Energies and RMS deviations of the conformations with the lowest long-range interaction energies (E_{LR}). The flexible parts are calculated using the Gaussian Dynamics method.

PDB name	Energy (RT) of Native-state	Lowest energy (RT) obtained in simulations	RMS deviation (Å) with respect to x-ray structure
4rxn	-231.4	-175.9	8.08
4pti	-230.7	-196.9	1.53
1r69	-345.9	-328.3	0.63
2cro	-337.1	-318.7	0.75
1sn3	-284.7	-222.9	2.92
*1ctf	-358.6	-355.2	0.89
3icb	-368.2	-293.5	2.26
1ubq	-381.1	-274.3	1.36

TABLE 6.9. Energies and RMS deviations of the conformations with the lowest backbone-backbone interaction energies (E_{BB}). The flexible parts are determined by the Gaussian Dynamics method.

PDB name	Energy (RT) of Native-state	Lowest energy (RT) obtained in simulations	RMS deviation (Å) with respect to x-ray structure
4rxn	-92.8	-96	7.62
4pti	-111.9	-109.5	10.35
1r69	-144.3	-146.8	9.9
2cro	-156.9	-162.8	3.23
1sn3	-133.1	-128.4	6.59
1ctf	-153.8	-163.1	2.66
3icb	-168.0	-174.3	9.45
1ubq	-159.5	-162.4	8.01

TABLE 6.10. Rank and energies of the conformations with lowest RMS deviation from x-ray structure. The flexible parts are determined by the Gaussian Dynamics method.

PDB name	RMS deviation (Å) of the nearest-native conformation	Rank	Total energy (RT) of the nearest-native conformation
4rxn	1.76	748	-111.9
4pti	1.03	357	-113.4
1r69	0.52	15	-304.7
2cro	0.48	3	-287.7
1sn3	2.92	1	-181.3
1ctf	0.70	1	-280.0
3icb	1.41	1	-242.1
1ubq	1.36	1	-240.1

TABLE 6. 11. Rank and energies of the conformations with lowest RMS deviation. The flexible parts are determined by the Gaussian Dynamics method.

PDB name	RMS deviation (Å) of the nearest-native conformation	Rank	Long-range energy (RT) of the nearest-native conformation
4rxn	1.76	129	-143.2
4pti	1.05	49	-165.1
1r69	0.52	17	-319.3
2cro	0.48	2	-276.6
1sn3	2.92	1	-222.9
1ctf	0.70	2	-349.5
3icb	1.41	2	-242.1
1ubq	1.36	1	-274.3

TABLE 6. 12. Rank and energies of the conformations with lowest RMS deviation with. The flexible parts are determined by the Gaussian Dynamics method.

PDB name	RMS deviation (Å) of the nearest-native conformation	Rank	Backbone-backbone energy (RT) of the nearest -native conformation
4rxn	1.76	220	-78.0
4pti	1.03	4027	-87.4
1r69	0.52	2458	-137.5
2cro	0.48	1446	-151.8
1sn3	2.92	1735	-110.1
1ctf	0.70	77	-153.8
3icb	1.41	533	-142.2
1ubq	1.36	390	-118.3

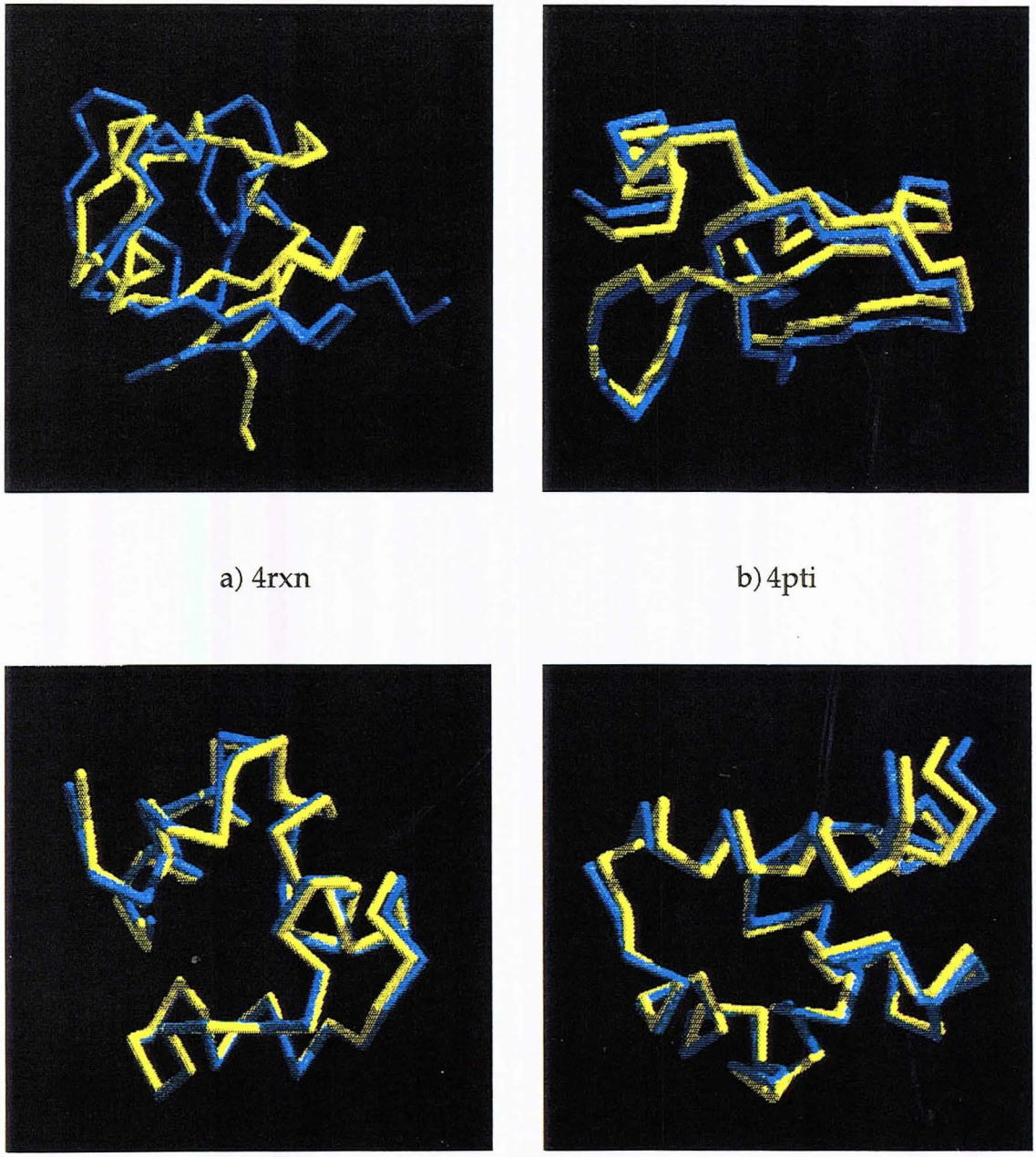
When Tables 6.7 and 6.8 are compared, the RMS deviations of the lowest energy conformations found on the basis of total energy, E_{total} are observed to be lower than those found from long-range energy, E_{LR}. The results of Table 6.9 show that non-specific backbone-backbone energies again fail to predict to correct tertiary structures. When these tables are compared Tables 6.1-6.3, it is noticed that the results for 434 repressor (1r69), 434 cro protein (2cro), and ribosomal protein (1ctf) are now improved. The lower RMS deviations presently attained are due to the fact that the rotations around the flexible C^{α} bonds are performed using smaller intervals, 10^{o} instead of 30^{o}.

The superimposed structures of the native states and the lowest energy states for the simulations in which the flexible residues are determined by the Gaussian dynamics method are presented in Figures 6.3-6.4. Except for rubredoxin (4rxn), the predicted conformations are almost identically superimposable on the native structure, which illustrates how well the method and knowledge-based potentials are applicable to the recognition of tertiary structures of the proteins.

For rubredoxin (4rxn), RMS deviations of the lowest energy conformations obtained using the flexible residues proposed by Park and Levitt [1], and those determined by the Gaussian dynamics method, are 1.91 and 8.08 Å, and for trypsin inhibitor (4pti), these are 6.63 and 1.53 Å. This difference can be observed from the comparison of the structures for these two proteins shown in Figures 6.1, and 6.3. Obtaining two different results for the same protein indicates the importance of choice of the flexible regions.

6.2.4. Comparison with Previous Simulations

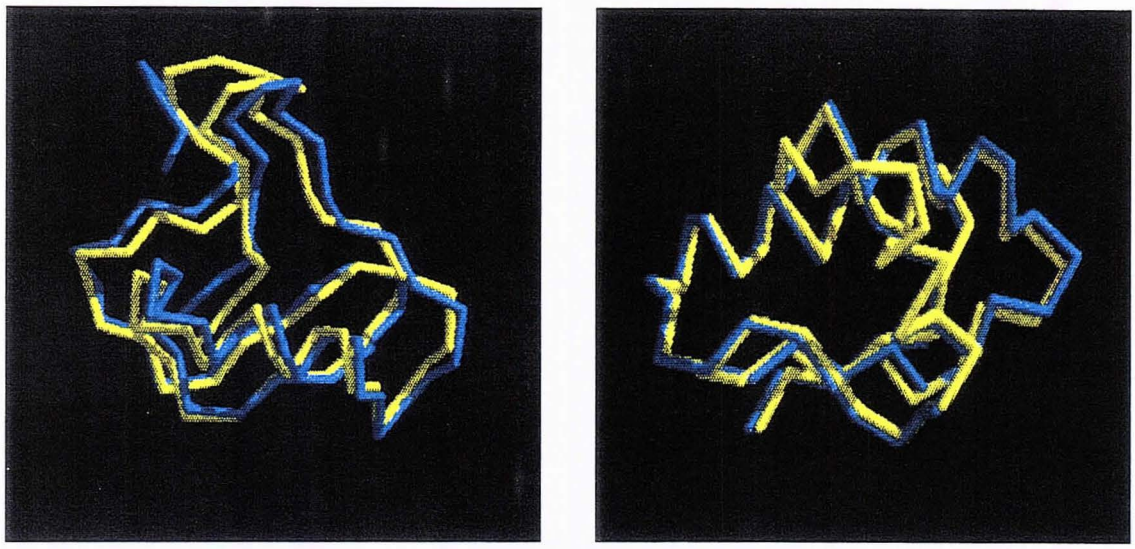
Hinds and Levitt [27] generated exhaustive sets of decoys within the range of 5 to 7 Å on a diamond lattice. Monge et al. [34] used Monte Carlo algorithm to generate conformations with RMS deviations of 4 to 10 Å from x-ray structures. Park and Levitt [1] generated complete sets of conformations using their four-state model for the same proteins as those considered in the present study. They evaluated the energies of the generated conformations



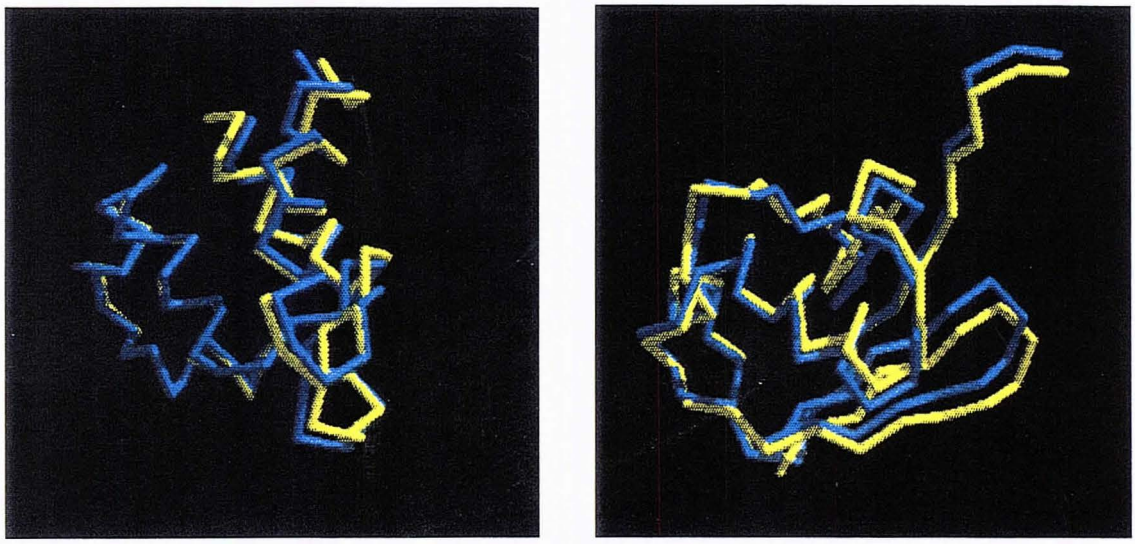
a) 4rxn b) 4pti

c) 1r69 d) 2cro

FIGURE 6.3. The superimosed structure of the proteins a) 4rxn, b) 4pti, c) 1r69, d) 2cro. The yellow structure is the native structure and the blue one is the predicted tertiary structure. The 3 - D structure of the backbone is shown. The flexible regions are calculated by the Gaussian Dynamics method.



a) 1sn3 b) 1ctf



c) 3icb d)1ubq

FIGURE 6.4 The superimosed structure of the proteins a) 1sn3, b) 1ctf, c) 3icb d) 1ubq. The yellow structure is the native structure and the blue one is the predicted tertiary structure. The 3 - D structure of the backbone is shown. The flexible regions are calculated by a Gaussian Dynamics method.

with respect to contact, van der Waals, surface area and histogram energy functions. The RMS deviations of the lowest energy conformations, from the x-ray structures were relatively high (between 5.7 and 14.2 Å). The lowest energy conformations of the present study, on the other hand, which are calculated on the basis of the knowledge-based potentials of Jernigan and Bahar [8] exhibit rather low RMS deviations (< 2 Å). This shows the capabilities of these energy parameters in differentiating the native-like structures. The results of the present study are satisfactory, when they are compared with the studies of Hinds and Levitt [27] and Monge et al. [34]. The method is simpler and more efficient than Monte Carlo algorithm, since the exhaustive enumeration is performed by simply rotating the virtual C^{α}-C^{α} bonds with 30^{o} intervals and also fairly small ensembles can be generated by using Monte Carlo techniques [34].

The conformations with the lowest RMS deviations are always within the top 0.01 per cent of the generated low energy conformations. We note that Covell and Jernigan [27] also generated all lattice conformations for several proteins, and showed that a comparable accuracy level is obtained using on-lattice simulations. They found that the conformation closest to the native structure was always within the top 1 per cent of conformations. However, their simulations were in the form of threading calculations, rather than generating new conformations.

The validity of the present approach depends on how accurately the secondary structures and the location of flexible residues are known. The secondary structural elements are assumed to be rigid blocks, so the model and the methods will fail, if the secondary structure elements are not completely correct. Bohr et al. [61] predicted the protein structures from the incomplete knowledge of disulfide bridges, surface and secondary structure assignments and additional distance constraints by using a minimization algorithm for six small proteins, but they could not obtain conformations with RMS deviation lower than 2 Å.

6.2.5. Identification of the Most Stable Regions

Finkelstein argues that [62] the structural features common to all conformations can be identified by viewing predicted tertiary structures from an ensemble of conformations instead of examining a single conformation. Following this approach, in order to identify the regions of the proteins which exhibit a strong tendency to fold correctly, the most favorable torsional states of the flexible residues were determined Thus, the sets of low energy conformations generated for each protein were examined.

The probability distribution curves for the rotational angles (ϕ_i) of the flexible bonds are obtained for all proteins. These are presented in Figures 6.5-6.12. As seen from the figures, some residues exhibit a strong preference for certain rotational states, their probabilities being much higher than expected for random distribution. For example, in rubredoxin (4rxn), among the ten rotatable bonds, Asp 21, Asp 22, Ile 33, and Pro 34 are distinguished by their strong preferences for the torsional states 60°, 40°, 50°, and 110°, respectively. This indicates a stable region covering the residues $12 \leq i \leq 42$. Except for 434 repressor (1r69) and ubiquitin (1ubq), the most stable regions of each protein are found by this analysis. The number of conformations considered for the statistical analysis and the stable regions of each protein are presented in Table 6.13. Figures 6.13-6.14 present the stable regions superimposed on the native structure. It is observed that the stable regions which are obtained by evaluating the most probable state of the flexible residues are identically superimposable on those parts of the native structures.

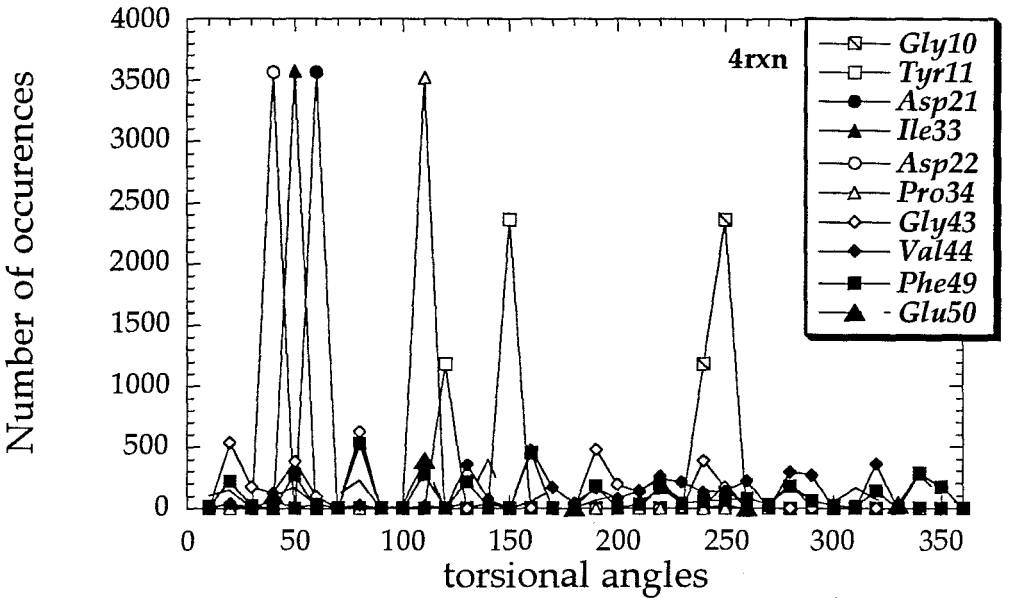


FIGURE 6.5. The distribution of rotational angle (ϕ_i) for ten flexible bonds of rubredoxin (4rxn).

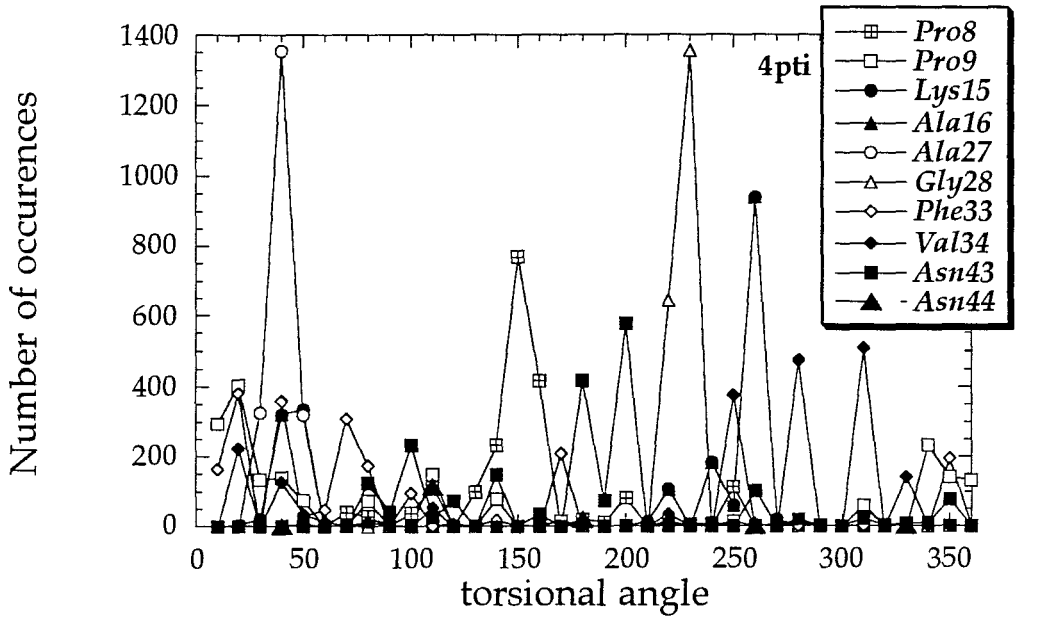


FIGURE 6.6. The distribution of rotational angle (ϕ_i) for ten flexible bonds of trypsin inhibitor (4pti).

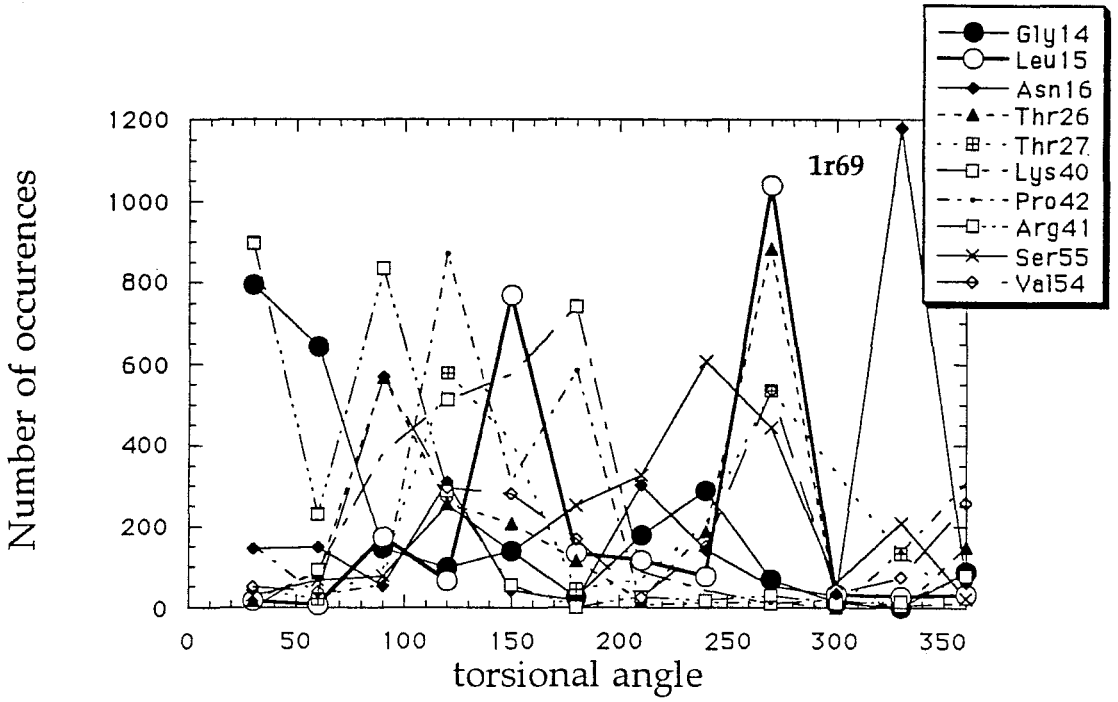


FIGURE 6.7 The distribution of rotational angle (ϕ_i) for ten flexible bonds of 434 repressor (1r69).

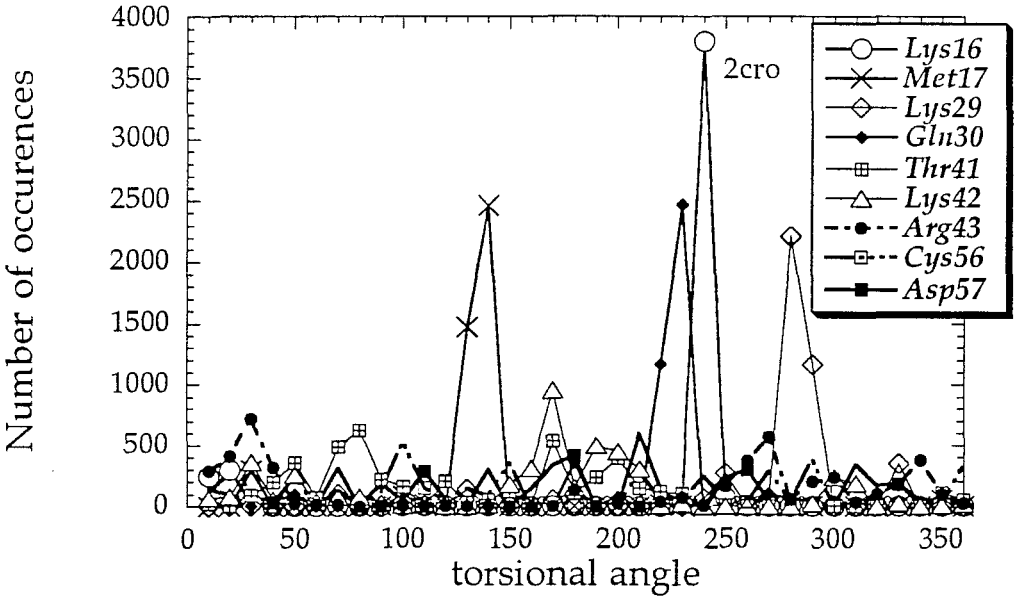


FIGURE 6.8 The distribution of rotational angle (ϕ_i) for ten flexible bonds of 434 cro protein (2cro).

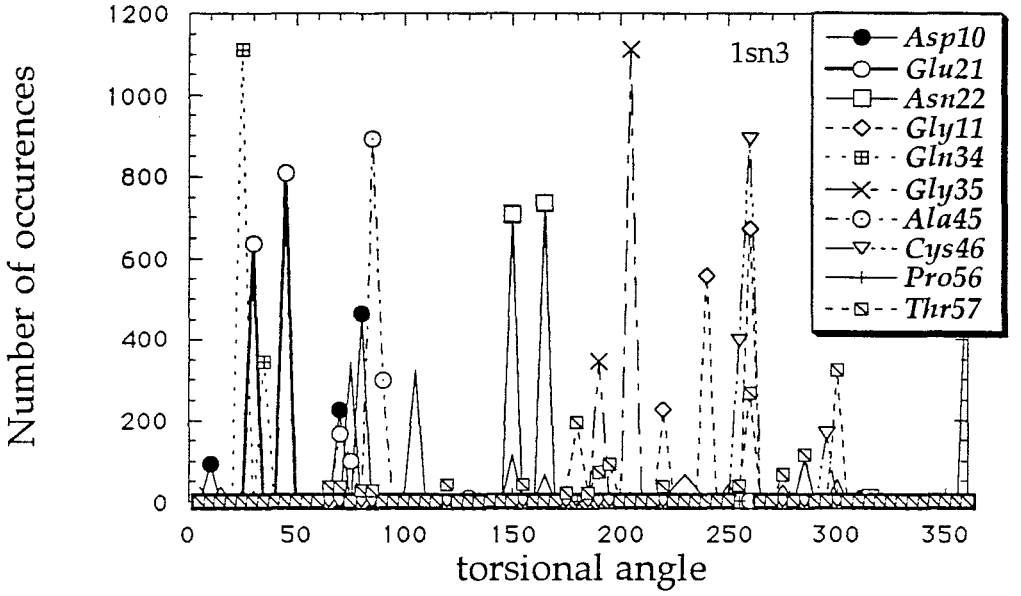


FIGURE 6.9 The distribution of rotational angle (ϕ_i) for ten flexible bonds of scorpion neurotoxin (1sn3).

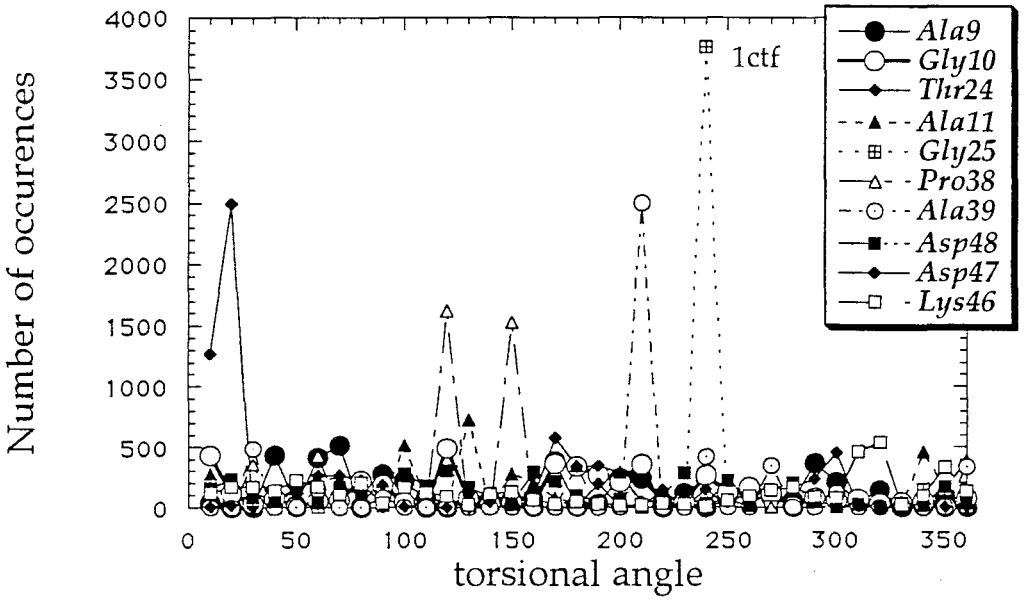


FIGURE 6.10 The distribution of rotational angle (ϕ_i) for ten flexible bonds of ribosomal protein (1ctf).

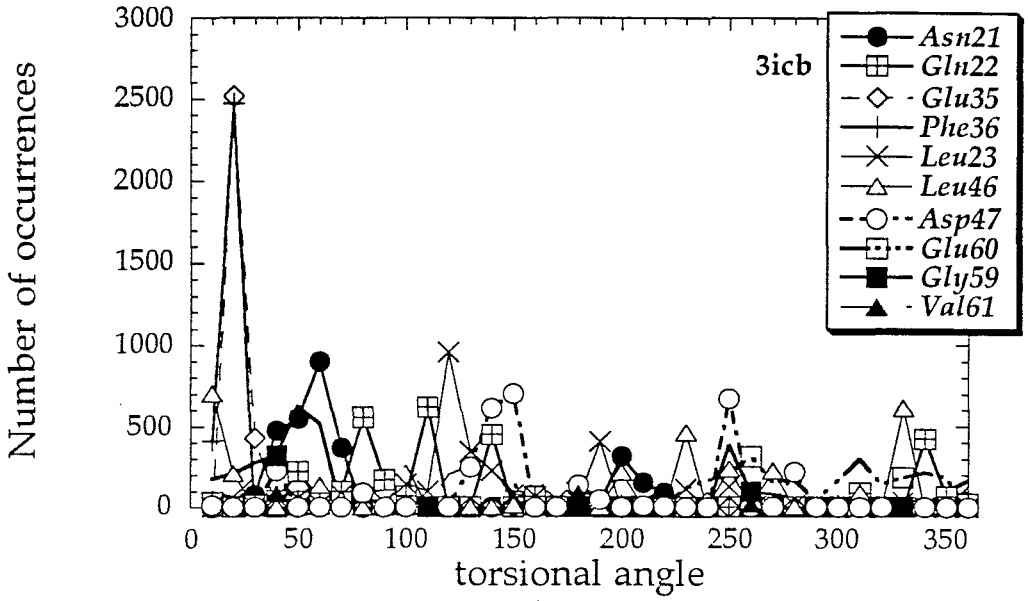


FIGURE 6.11. The distribution of rotational angle (ϕ_i) for ten flexible bonds of calcium binding protein (3icb).

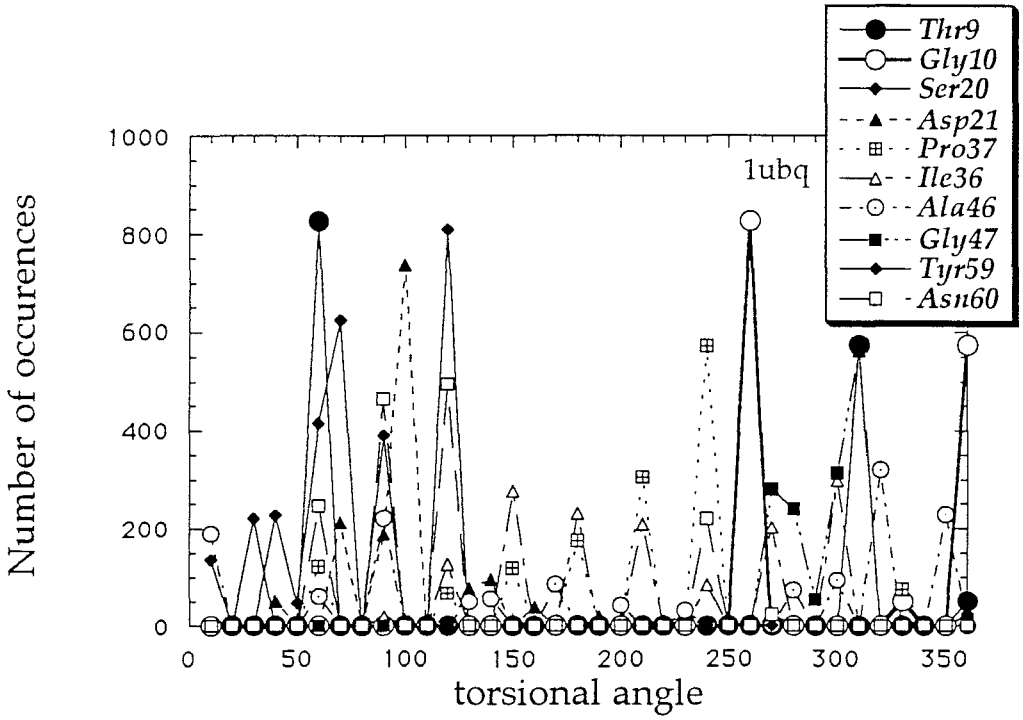
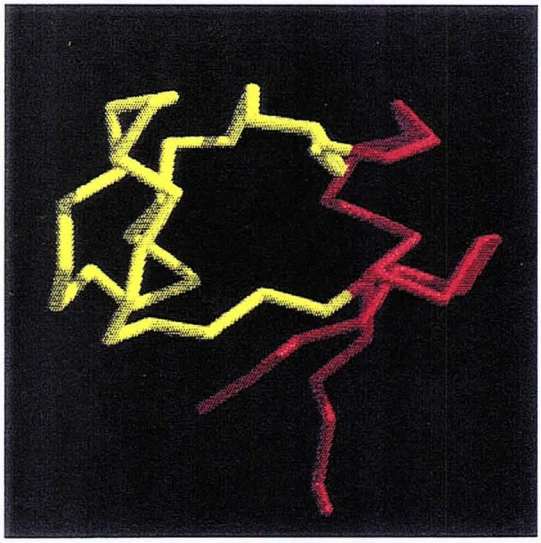
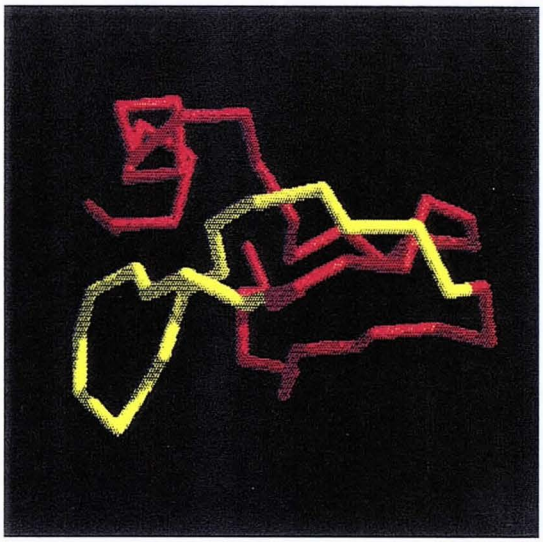


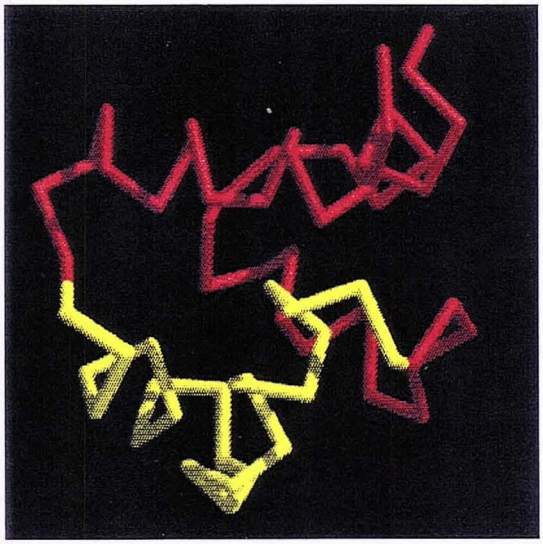
FIGURE 6.12. The distribution of rotational angle (ϕ_i) for ten flexible bonds of ubiquitin (1ubq)



a) 4rxn

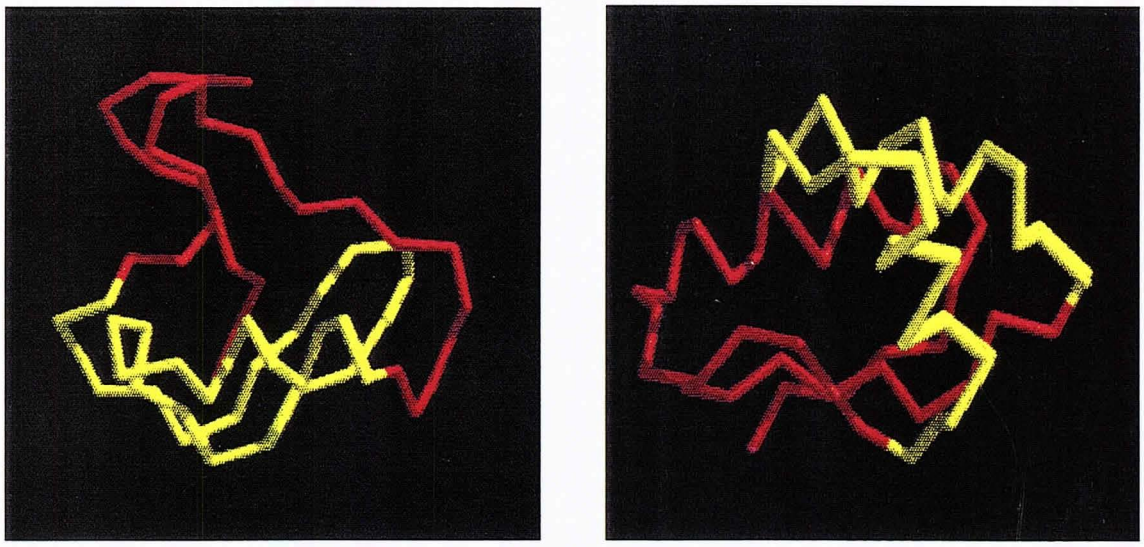


b) 4pti

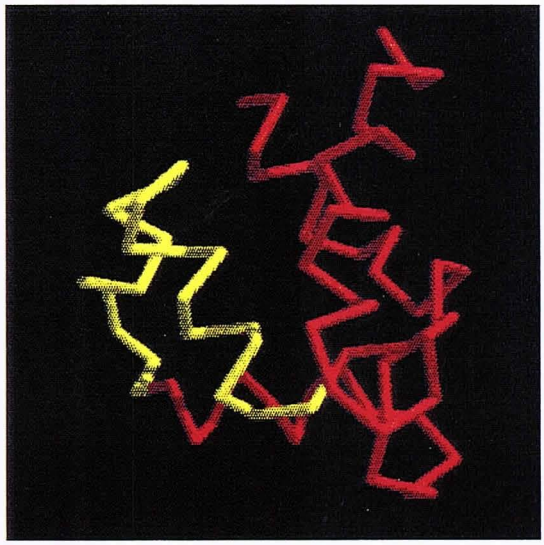


c) 2cro

FIGURE 6.13. The most stable regions of the proteins a) 4rxn, b) 4pti, c) 2cro. The yellow part is the superimposed structures of the stable regions with the native structure.



a) 1sn3 b) 1ctf



c) 3icb

FIGURE 6.14 The most stable regions of the proteins a) 1sn3, b) 1ctf, c) 3icb, The yellow part is the superimposed structures of the stable regions with the native structure.

TABLE 6.13. The number of conformations considered for the statistical analysis and the most stable regions of the proteins.

PDB Name	Number of Conformations	The Most Stable Region
4rxn	3570	Ile 12 - Cys 42
4pti	1997	Arg 17 - Thr 32
1r69	2486	------
2cro	4456	Thr 18 - Lys 40
1sn3	1455	Glu 23 - Thr 55
1ctf	4454	Asn 12 - Ala 37
3icb	1455	Ser 24 - Thr 45
1ubq	1451	-------

7. CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions

Computations for predicting the native three-dimensional structure of proteins using a low resolution model lead to the following conclusions:

(a) After generating a complete set of conformations by rotating virtual C^{α}–C^{α} bonds, it is seen that for 434 repressor (1r69), calcium binding protein (3icb), ribosomal protein (1ctf), ubiquitin (1ubq), the lowest energy conformations exhibit the lowest RMS deviations from the native structures at the same time. In the case of rubredoxin (4rxn), 434 cro protein (1r69) and scorpion neurotoxin (1sn3), the RMS deviation of the lowest energy conformations with respect to the native conformation remains lower than 2.0 Å, which is a satisfactory result in the view of the use of a coarse-grained model. The only protein which fails to be correctly recognized is trypsin inhibitor (4pti). The prediction of the tertiary structure of the latter is however, significantly improved upon determination of flexible residues by the Gaussian dynamics method.

(b) The usefulness of the model and energy parameters depend on how well they discriminate the native-like folds from the non native conformations. Park and Levitt [1] used contact, van der Waals, surface area and histogram energy functions to evaluate the energies of the conformations which were generated by four-state model. Most of the energy functions, especially non-distance-dependent ones, failed to discriminate the correct native folds. The presently used knowledge-based potentials [8] are evaluated from distance-dependent sidechain-sidechain interaction potentials extracted at 0.4 Å on the basis of low resolution model. The effects of the size of the proteins and the fraction of the hydrophobic residues are not considered. According to the results of the present study, it can be safely said that, these knowledge-based potentials are able to identify the native-like conformations.

(c) When the lowest energy conformations whose energies are evaluated according to total (E_{total}), long-range (E_{LR}) and non specific backbone-backbone energies E_{bb} are compared, it is seen that the effect of short-range interaction energy is negligible. The non specific backbone-backbone energies E_{bb} fail to predict the correct native folds even though they can give a general information about the proteins.

(d) The study is repeated with different flexible residues. Although the model and the energy parameters again succeed in predicting the near native folds (the correct tertiary structures), results for rubredoxin and trypsin inhibitor are observed to highly sensitive to the choice of flexible bonds. Locations of flexible residues are so important that improper choices cannot lead to correct native structures.

(e) A statistical analysis of the dihedral angle preferences in the set of all generated low energy conformations for each protein, yields an information on the protein regions with a strong tendency to assume the correct fold. These regions may be involved in early stages of the proteins, possessing a strong tendency to assume native-like rotational state, therefore act as folding nuclei.

(f) The secondary structural elements were kept as rigid blocks. The tertiary structures of the proteins were thus predicted by the determination of the optimal organization of the known secondary structural elements in space. This approach is based on a sequential folding mechanism. The results show that a sequential folding mechanism is applicable to the investigated proteins, if the secondary structural elements and flexible regions are accurately known.

(g) Exhaustive enumeration is performed by rotating virtual C^{α}–C^{α} bonds using large intervals such as 30°, and then the computations are repeated with smaller intervals (10°) in the neighborhood of the values obtained from the previous computations as a second step. In the second part of the study, the conformations are generated using smaller steps (10°). Lower RMS deviations (< 1.0 Å) are obtained with respect to the first part. It is better to use smaller steps sizes to capture the native folds. However, even proceeding at 30° intervals is suitable for locating the native-like conformation. This suggests that the energy minimum near the native fold is sufficiently deep and broad to be distinguished even with a relatively coarse-grained enumeration technique.

7.2. Recommendations

The results show that this method is successful for predicting the tertiary structure of small, single domain proteins. Calculations were performed for eight proteins. It would be complementary to increase the number of proteins used in calculations and check the validity of the method and the model.

Location of flexible regions is very important for obtaining the most native-like structure. Instead of dividing the proteins into three or four linear segments, containing two or three successive flexible residues, it would be better to decrease the number of segments and increase the number of flexible residues in these segments. For future work, it is recommended to repeat the calculations using such longer flexible regions.

In order to determine the most stable regions of the proteins, a statistical analysis of the torsional state preferences in the set of all generated low energy conformations for each protein was performed. For that purpose, the probability distribution curves for the rotational angles (ϕ_i) of the flexible bonds were obtained. A similar analysis can be repeated by considering only the top 50 lowest energy conformations, and evaluating the probability of each torsional state according to the energy of the conformations. It is highly likely that such an analysis will yield rotational angles in close agreement with native structure values, at least in some conserved regions of the proteins.

REFERENCES

1. Park, B. and M. Levitt, "Energy Functions that Discriminate X-ray and Near-native Folds from Well-constructed Decoys," *Journal of Molecular Biology*, Vol. 258, pp. 367-392, 1996.

2. Levitt, M. and M. Chothia, "Structural Patterns in Globular Proteins," *Nature*, Vol. 261, pp. 552-558, 1976.

3. Huang, E. S., S. Subbiah, and M. Levitt, "Using a Hydrophobic Contact Potential to Evaluate Native and Near-native Folds Generated by Molecular Dynamics Simulations," *Journal of Molecular Biology*, Vol. 257, pp. 716-725, 1995.

4. Dill, K. A., S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas and H. S. Chan,"Principles of Protein Folding: A Perspective from Simple Exact Models," *Protein Science*, Vol. 4, pp. 561-602, 1995.

5. Mairov, V. N. and G. M. Crippen, "Contact Potential that Recognizes the Correct Folding of Globular Proteins," *Journal of Molecular Biology*, Vol. 227, pp. 876-888, 1992.

6. Miyazawa, S. and R. L. Jernigan, "Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-chemical Approximation," *Macromolecules*, Vol. 18, pp. 534-552, 1985.

7. Monge, A., E. J. Lathrop, J. R. Gunn, P. S. Shenkin and R. A. Friesner, "Computer Modeling of Protein Folding: Conformational and Energetic Analysis of Reduced and Detailed Protein Models," *Journal of Molecular Biology*, Vol. 247, pp. 995-1012, 1995.

8. Bahar, I. and R. L. Jernigan, "Inter-residue Potentials in Globular Proteins and the Dominance of Highly Specific Hydrophilic Interactions at Close Separation," *Journal of Molecular Biology*, Vol. 266, pp. 195-214, 1997.

9. Covell, D. G., "Folding Protein α-carbon Chains into Compact Forms by Monte Carlo Methods," *Proteins,* Vol. 14, pp. 409-420, 1992.

10. Gunn, J. R., A. Monge, R. A. Friesner and C. H. Marshall, "Hierarchical Algorithm for Computer Modelling of Protein Tertiary Structure: Folding of Myoglobin to 6.2 Å Resolution," *Journal of Physical Chemistry,* Vol. 98, pp. 702-711, 1994.

11. Kolinski, A. and J. Skolnick, "Monte Carlo Simulation of Protein Folding: 1. Lattice Model and Interaction Scheme," *Proteins,* Vol. 18, pp. 338-352, 1994.

12. Vieth, M., A. Kolinski, C. L. Brooks and J. Skolnick, "Prediction of the Folding Pathways and Structure of the GCN4 Leucine Zipper," *Journal of Molecular Biology,* Vol. 237, pp. 361-367, 1994.

13. Wilson, C. and S. Doniach, "A Computer Model to Dynamically Simulate Protein Folding: Studies with Crambin," *Proteins,* Vol 6, pp. 193-209, 1989.

14. Bowie, J. U., R. Lüthy, and D. Eisenberg, "An Evolutionary Approach to Folding Small α-helical Proteins that Fold into a Known Three-Dimensional Structure," *Science,* Vol. 253, pp. 164-170, 1991.

15. Dandekar, T., and P. Argos, "Folding the Main-chain of Small Proteins with the Genetic Algorithm," *Journal of Molecular Biology,* Vol. 236, pp. 844-861, 1994.

16. Unger, R. and J. Moult, "Genetic Algorithms for Protein Folding Simulations," *Journal of Molecular Biology,* Vol. 231, pp. 75-81, 1993.

17. Sun, S., "Reduced Representation Model of Protein Structure Prediction: Statistical Potential and Genetic Algorithms," *Protein Science,* Vol. 2, pp. 762-785, 1993.

18. Chothia, C., "One Thousand Families for the Molecular Biologist," *Nature,* Vol. 357, pp. 543-544, 1992.

19. Orengo, C. A., D. T. Jones, and J. M. Thornton, "Protein Superfamilies and Domain Superfolds," *Nature*, Vol. 372, pp. 631-634, 1994.

20. Hendlich, M., P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari and M. J. Sippl, "Identification of Native Protein Folds Amongst a Large Number of Incorrect Models. The Calculation of Low Energy Conformations from Potentials of Mean Force," *Journal of Molecular Biology*, Vol. 216, pp. 167-180, 1990.

21. Sippl, M. J. and S. Weitckus, "Detection of Native-like Models for Amino acid Sequences of Unknown Three-dimensional Structure in Data Base of Known Protein Conformations," *Proteins*, Vol. 13, pp. 258-271, 1992.

22. Jones, D. T., W. R. Taylor, and J. M. Thornton, "A New Approach to Protein Fold Recognition," *Nature*, Vol. 358, pp. 86-89, 1992.

23. Godzik, A., A. Kolinski, and J. Skolnick, "Topology Fingerprint Approach to the Inverse Protein Folding Problem," *Journal of Molecular Biology*, Vol. 227, pp. 227-238, 1992.

24. Bryant, S. H. and C. E. Lawrence, "An Empirical Energy Function for Threading Protein Sequence Through the Folding Motif," *Proteins*, Vol. 16, pp. 92-112, 1993.

25. Covell, D. G. and R. L. Jernigan, "Conformations of Folded Proteins in Restricted Spaces," *Biochemistry*, Vol. 29, pp. 3287-3294, 1990.

26. Williams, R. L., J. Vila, G. Perrot and H. Scheraga, "Empirical Solvation Models in the Context of Conformational Energy Searches: Application to Bovine Pancreatic Trypsin Inhibitor," *Proteins*, Vol 114, pp. 110-119, 1992.

27. Hinds, D. A. and M. Levitt, "A Lattice Model for Protein Structure Prediction at Low Resolution," *Proceedings of National Academy of Sciences USA*, Vol. 89, pp. 2536-2540, 1992.

28. Hinds, D. A. and M. Levitt, "Exploring Conformational Space with a Simple Lattice Model for Protein Structure," *Journal of Molecular Biology,* Vol. 243, pp. 668-682, 1994.

29. Wang, Y., H. Zhang, W. Li and R. A. Scott, "Discriminating Compact Non-native Structures from the Native Structure of Globular Proteins," *Proceedings of National Academy of Sciences USA,* Vol. 92, pp. 709-713, 1995.

30. Bahar, I., A. R. Atilgan and B. Erman, "Direct Evaluation of Thermal Fluctuations in Proteins Using a Single Parameter Harmonic Potentials," *Folding & Design,* Vol. 2, pp. 173-181, 1997.

31. Ptitsyn, O. B., and A. A. Rashin, "A Model of Myoglobin Self-organization," *Biophysical Chemistry,* Vol. 3, pp. 1-20, 1975.

32. Srinivasan, R. and G. D. Rose, "LINUS: A Hierarchic Procedure to Predict the Fold of a Protein," *Proteins,* Vol. 22, pp. 81-99, 1995.

33. Taylor, W. G., *Patterns in Protein Sequence and Structure,* Spring-Verlag, New York, 1992.

34. Monge, A., R. A. Friesner, and B. Honig, "An Algorithm to Generate Low-Resolution Protein Tertiary Structures from Knowledge of Secondary Structure," *Proceedings of National Academy of Sciences USA,* Vol. 91, pp. 5027-5029, 1994.

35. Stryer, L., *Biochemistry,* W. H. Freeman and Company, New York, 1988.

36. Branden, C. and J. Tooze, *Introduction to Protein Structure,* Garland Publishing, New York, 1991.

37. Pauling, L., and R. B. Corey, "The Structure of Proteins: Two Hydrogen-Bonded Helical Configuration of The Polypeptide Chain," *Proceedings of National Academy of Sciences USA,* Vol. 37, pp. 205-211, 1951.

38. Voet, D. and J. G. Voet, *Biochemistry*, John Wiley and Sons, Inc., New York, 1995.

39. Chothia, C., "Conformation of Twisted β-pleated Sheets in Proteins," *Journal of Molecular Biology*, Vol. 75, pp. 295-302, 1973.

40. Thornton, J. M., *Protein Folding*, New York: W. H. Freeman and Company, 1992

41. Shortle, D., Y. Wang, J. R. Gillispie, and J. O. Wrabl, "Protein folding for Realists: A timeless Phenomenon," *Protein Science*, Vol. 5, pp. 991-1000, 1996.

42. Creighton, T. E. (editor), *Protein Folding*, W. H. Freeman and Company, New York, 1992.

43. Shakhnovich, E. I., "Theoretical Studies of Protein-folding Thermodynamics and Kinetics," *Current Opinion in Structural Biology*, Vol. 7, pp. 29-40, 1997.

44. Strang, G., *Linear Algebra and Its Applications*, New York: Harcourt, Brace, Jovanovich, 1988.

45. Tirion, M. M., "Large Amplitude Elastic Motions in Proteins from a Single Parameter, Atomic Analysis," *Physical Review Letters*, Vol. 77, pp. 1905-1908, 1996.

46. Bernstein, E. E., T. F. Koetzle, G. J. B. Williams, J. E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. J. Tasumi, "The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures," *Journal of Moleculae Biology*, Vol. 112, pp. 535, 1977.

47. Brant, D. A. and P. J. Flory, "The Configuration of Random Polypeptide Theory," *Journal of American Chemical Society*, Vol. 87, pp. 2791-2800, 1965.

48. Bahar, I., M. Kaplan and R. L. Jernigan, "Short-range Conformational Energies, Secondary Structure Propensities, and Recognition of Correct Sequence-Structure Matches," *Proteins,* in press.

49. Watenpaugh, K. D., L. C. Sieker, and L. C. Hensen, "Crystallagraphic Refinement of Rubredoxin at 1.2 Å Resolution," *Journal of Molecular Biology*, Vol. 138, pp. 615, 1980.

50. Marquart, M., J. Walter, J. Deisenhofer, W. Bode and R. Huber, "The Geometry of the Reactive Site and of the Peptide Groups in Tyripsinogen and its Complexes with Inhibitors," *Acta Crystallography, Section B*, Vol. 39, 1983.

51. Mondragon, A., S. Subbiah, C. Almo, M. Drottar and M. Harrison "Structure of the Amino-terminal of Domain of Phage 434 Repressor At 0 Å Resolution," *Journal of Molecular Biology*, Vol. 205, pp. 189, 1989.

52. Mondragon, A., C. Wolberger , and S. C. Harrison, "Structure of Phage of 434 Cro Protein at 5 Å Resolution," *Journal of Molecular Biology*, Vol. 205, pp. 179, 1989.

53. Bugg, C. E., "Structure of Variant-3 Scorpion Neurotoxin from Centruroides Sculpturatus Ewing, Refined at 1.8 Å Resolution," *Journal of Molecular Biology*, Vol. 493, pp. 493, 1983.

54. Leijonmarck, M. and A. Liljas, "Structure of the C-Terminal Domain of the Ribosomal Protein L7-L12 from Escherichia Coli at 1.7 Å," *Journal of Molecular Biology*, Vol. 195, pp. 555, 1987.

55. Szenbbenyi, D. M. E. and K. Moffat, "The Refined Structure of Vitamin D-Dependent Calcium-Binding Protein from Bovine Intestine Molecular Details, Ion Binding, and Implications for the Structures of the Calcium-Binding Proteins," *Journal of Biological Chemistry*, Vol. 261, pp. 8761, 1986.

56. Vijay, S., C. E. Bugg and W. J. Cook, "Structure of Ubiquitin Refined at 1.8 Å Resolution," *Journal of Molecular Biology*, Vol. 194, pp. 531, 1987.

57. Flory, P. J., *Statistical Machanics of Chain Molecules*, New York: Hanser Publishers, 1988.

58. Mattice, W. L. and U. W. Suter, *Conformational Theory of Large Molecules*, John Wiley & Sons, Inc., New York, 1994.

59. Kabasch, W., "A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors," *Acta Crystallography Section A*, Vol. 34, pp. 827-828, 1978.

60. Anfinsen, C. B., "Principles that Govern the Folding of Protein Chains," *Science,* Vol. 181, pp. 223-230, 1973.

61. Lund, O., J. Hansen, S. Brunak, and J. Bohr , "Relationship between Protein Strucutre and Geometrical Constraints," *Protein Sciencee,* Vol. 5, pp. 2217-2225, 1996.

62. Finkelstein, A. V., A. M. Gutin, and A. Y. Badretinov, "Perfect Temperature for Protein Structure and Folding," *Proteins,* Vol. 23, pp. 142-150, 1995.