PACKING REGULARITIES IN FOLDED PROTEINS

by

Elife Zerrin Bağcı

B.S. in CHE, Boğaziçi University, 1997

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

2001

ACKNOWLEDGMENTS

I would like to thank to Prof. İvet Bahar for her being the first person to appreciate me as a scientist candidate. This is meaningful to me because she is one of the rare persons who combines abilities of a good scientist, instructor, advisor and administrator with a conscientious character. I would also like to thank to Assoc. Prof. Türkan Haliloğlu for her kind help and nice character. I was privileged to be their student.

Many thanks to Prof. A. Rana Atılgan, Prof. Mehmet Çamurdan and Assoc. Prof. Pemra Doruker for their kindly attendance to examination jury.

Thanks to the members of Physics Department, especially to Prof. Cihan Saçlıoğlu. I learned much from them.

My thanks are also due to Hatice Köse, Sezen Gürdağ, Sinem Özyurt; my friends in Polymer Research Center, and in USA but not so far from me, Alper Acar and Betül Ünlüsü.

Many thanks to Boğaziçi University, in general, because I understood here that being hard-working and learned is not enough to be successful in life.

Special thanks to my parents and my sister Bilge for their love, motivation, and help during this study, and for the rest of my life.

This thesis is dedicated to all scientists.

ABSTRACT

PACKING REGULARITIES IN FOLDED PROTEINS

Using lattice models of proteins is a common method to reduce conformational space. Protein structures can be satisfactorily threaded onto several such lattices, the accuracy of the fit increasing with the coordination number of the lattice. Despite the suitability of various lattice geometries, the optimal packing geometry of residues in folded structures, or the generic preference for regular packing, if any, remains unclear. In this thesis, a degree of intrinsic regularity in residue packing is revealed upon optimal superimposition of clusters of residues from Protein Data Bank structures. This regularity can be identified as an incomplete distorted face-centered cubic packing, i.e. the closest packing of identical spheres, emerging when the tertiary structure is observed at a coarse-grained (single-site-per-residue) scale. It is apparently favored by the drive for maximizing packing density, and shows little variation with specific amino acid type. Both the extreme cases of solvent-exposed and completely buried residue neighborhoods approximate this generic packing, their only difference being in the number (and not the type) of coordination sites that are occupied (or left void for solvent occupancy). Interestingly, these sites are not staggered even for the solvent-exposed residues on the surface and it is concluded that all residues, even those at the protein surface are densely packed. The packing density is approximately uniform when the volume of solvent surrounding the residues is excluded.

ÖZET

KATLANMIŞ PROTEİNLERDE REZİDÜ YERLEŞME DÜZENİ

Latis modeller proteinlerin konformasyon uzayını azaltma amacıyla sıklıkla kullanılır. Protein yapıları çeşitli latislere latisin koordinasyon sayısı arttıkça artan bir doğrulukta örülebilir. Proteinlerin modellenmesi amacı ile çeşitli latis geometrileri uygun olduğu halde, katlanmış yapılardaki rezidülerin belirli bir yerleşme düzenini seçip seçmediği konusunda henüz bir bilgi bulunmamaktadır. Bu tezde, rezidülerin proteinlere özgü düzenli bir yerleşme eğilimi olduğu ortaya çıkarılmıştır. Bu sonuç Protein Bilgi Bankası'ndan alınan protein yapılarındaki rezidü topluluklarının optimal olarak üst üste çakıştırılmasıyla elde edilmiştir. Bu düzen, proteinlerin üçüncül yapılarının düşük çözünürlüklü uzayda (rezidü başına tek bir nokta) incelenmesi ile anlaşılmaktadır. Buna göre rezidüler tamamlanmamış ve deformasyona uğramış yüzey merkezli kübik bir yerleşmeyi tercih etmektedir. Yüzey merkezli kübik yerleşme aynı büyüklükteki kürelerin en yoğun yerleşmesidir. Bu çalışmada bulunan yerleşme düzeni, proteinlerin molekül içi yoğunluğunu arttırma eğiliminden kaynaklanmaktadır ve rezidü tipine göre küçük farklılıklar göstermektedir. Hem çözücüye temas eden, hem de tamamen proteinin içine gömülmüş olan rezidülerin – en uç örnekler – komşulukları da bu proteinlere özgü genel yerleşme düzenine yakın benzerlik göstermektedir. Rezidüler arasındaki farklılık, doldurulan (ya da çözücü için boş bırakılan) koordinasyon merkezlerinin sayıları şeklindedir. Çözücüye temas eden yüzey rezidüleri için bile aynı yerleşme düzeni geçerlidir. Sonuç olarak rezidülerin çevresindeki çözücüye ayrılan hacim ihmal edildiği takdirde, proteinin yüzeyi dahil olmak üzere tüm bölgelerinde aynı oranda muntazam ve yoğun bir yerleşme düzeni bulunduğu anlaşılmıştır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

m	Coordination number of the residue cluster
N	Number of residue clusters in a set
P	Probability of a coordination state
R, r	Radius of the sphere
z	Coordination number of the lattice
ε	Distance deviation
ϕ	Azimuthal angle
θ	Polar angle
Ala	Alanine
Arg	Arginine
Asn	Asparagine
Asp	Aspartic Acid
bcc	body-centered cubic
Cys	Cysteine
emb	Embedded lattice
fcc	Face-centered cubic
Gln	Glutamine
Glu	Glutamic Acid
Gly	Glycine
hcp	Hexagonal closest packed
His	Histidine
Ile	Isoleucine
Leu	Leucine
Lys	Lysine
Met	Methionine
opt	Optimal
PDB	Protein Data Bank
PF	Packing factor
Phe	Phenylalanine

Pro	Proline
RMS	Root-mean-square
sc	Simple cubic
Ser	Serine
Thr	Therionine
Trp	Tryptophan
Tyr	Tyrosine
Val	Valine

1. INTRODUCTION

Globular proteins have a unique three-dimensional structure under physiological conditions. This so-called native structures have significantly lower energy than all other conformations. Being compact crystals with a very high order at the atomic level, one is encouraged to examine their packing characteristics thoroughly (Richards and Lim, 1994). Residue packing has been suggested to play a selective role in determining protein structure in an "inverse folding" study by Ponder and Richards (1987). Two packing criteria were used: avoidance of steric overlap and complete filling of available space of rotamers of sidechains in determining the allowed sequences for a given tertiary structure. Templates specific to different structural classes obtained with that kind of analysis can be used to distinguish between these classes. An assumption of the study is that only a subset, the "interior" residues establish the basic architecture of the mainchain whereas the "exterior" residues are mainly involved in the energetics of the structure, not the geometry of the core parts of the mainchain. Another study by Munson *et al.* (1996) studied the effects of packing by redesigning hydrophobic core of a four-helix-bundle protein, ROP. The stability and structural properties of the protein were altered. In spite of computational studies and repacking experiments with mutations, the way in which residues are likely to be packed in folded proteins, or the preferential geometry towards which they tend to constrain the structure, is still an open issue, and so is even the existence of such inherent geometric packing preferences.

Controversial views have been advanced for the packing of sidechains in globular proteins. These differ in regularities and degrees of freedom. In one case, ideal packing conforming with the closest packed cubic geometry of identical spheres has been proposed (Raghunathan and Jernigan, 1997). In that study, the positional directions of the neighbors of a given residue were stated to coincide with the directions in face-centered cubic lattice, thus indicating an ideal architecture for folded proteins. In another case, residues were proposed to pack with a complementarity similar to a jigsaw puzzle (Richards, 1977). Opposing both of these views, Bromberg and Dill (1994) proposed that sidechain packing had a completely random arrangement free of directionality and complementarity. However they indicated that as maximum compactness is finally approached, the entropy

of sidechains was lost which supports the view of efficient packing. These suggestions are modeled in parts A, B and C of Figure 1.1, respectively.

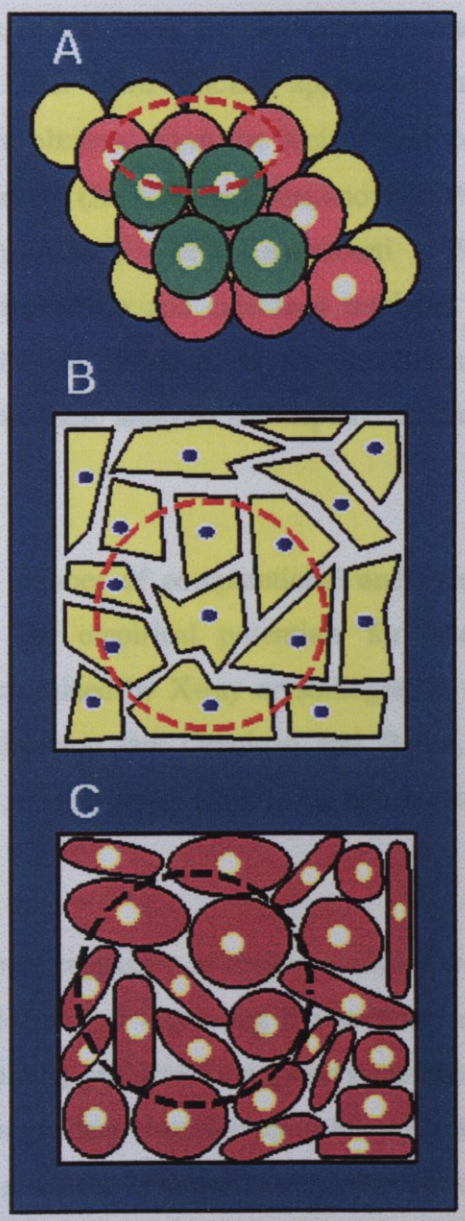

Figure 1.1. Different views of residue packing in folded proteins

The aim of the present work is to explore the existence of generic packing preferences and their role in determining the native conformation. Understanding how amino acids are packed, assessing the extent of randomness/regularity in their spatial rearrangements, and defining their coordination patterns are issues of crucial importance for designing proteins and their complexes. That kind of study concentrates on how the

protein backbone is folded to the native structure. Therefore, long range order should be analyzed although historically protein backbone preferences have been studied extensively. Studies on the regularities of the polypeptide backbone started with the pioneering work of Ramachandran *et al.* (1963). In their work, the allowed conformational states for secondary structures were identified. Later more efficient approaches in which pseudobonds and pseudotorsional angles were obtained from the loci of four consecutive C^{α} atoms have been employed by several groups (DeWitte and Shakhnovich, 1994; Oldfield and Hubbard, 1994; Bahar *et al.* 1997). Also Pal and Chakrabarti (1999) proposed a graphical representation for protein mainchain and sidechain torsional angles, which can aid in identifying backbone regularities. Whereas in the Ramachandran plots only a single residue is examined at a time, these analyses can capture regularities over several residues on the mainchain.

Until now, a large number of computational and theoretical studies have been directed at characterizing the empirical potentials for inter-residue interactions by exploiting the structures determined by X-ray crystallography or NMR spectroscopy and employing the inverse Boltzmann law (Jernigan and Bahar, 1996). Incorporating the preferred packing geometry, if any, could increase the discriminative power of knowledge-based potentials.

The existence of some regularity in residue packing would be of great utility for reducing the computational time and increasing the accuracy of conformational searches while solving the protein folding problem for a given sequence. Presently, the methods of bioinformatics can predict secondary structure up to 80 per cent accuracy (Petersen *et al.*, 2000), and a significant progress can be made in tertiary structure prediction by combining the tools for predicting secondary structure with those efficiently discriminating between alternative non-bonded interaction geometries.

The coordination patterns of residues that appear irrespective of amino acid type might be the reason for commonly observed insensitivity of structures to single site mutations. A study by Lim and Sauer (1991) revealed that even the majority of the combinations of three sites mutations in the core of lambda repressor did not change the general characteristics of its structure. What insensitivity to mutations indicates is the

absence of specificity not the irregularity in packing. A non-specific packing may relate to an ordered packing in which residues are packed similarly to hard spheres, or to disordered packing having the freedom to accommodate local changes without causing changes in the overall structure. Thus, whether packing is ordered or not cannot be inferred from tolerance or intolerance to mutations alone.

Proteins may have evolutionarily selected and conserved a regular architecture required for biological function and stability. The emergence of helical motifs in proteins is suggested by Maritan *et al.* (2000) by such generic packing preferences. A property of globular proteins which differentiates them from random coils is therefore related to evolution, compactness, induces secondary structure formation (Chan and Dill, 1990; Gregoret and Cohen, 1991; Hao *et al.* 1992; Hunt *et al.* 1994; Yee *et al.* 1994).

Although these studies indicate a relationship between regular backbone conformations and packing efficiency, no direct evidence of a regular non-bonded coordination associated with the drive for maximizing packing efficiency has yet been shown. Studies aimed at revealing sidechain coordination geometry have indicated a degree of non-randomness in residue packing. The orientation of neighboring sidechains with respect to a reference frame embedded on the mainchain were recently investigated for all kinds of residue pairs, and it was found that some coordination states are selected with probabilities about ten times higher than expected for random distributions (Bahar and Jernigan, 1996). Later a residue-specific backbone-dependent library for sidechain isomers had been proposed. Side chains could be packed onto known backbone structures utilizing their isomeric states (Keskin and Bahar, 1998). However, the geometries revealed in these analyses are based on reference frames embedded in the mainchain. The observed coordination states could be biased by the local backbone structure.

A more informative analysis of residue packing should be independent of any biased reference frame. The question to be answered is simple and direct: is there any regularity in residue packing, observed at a coarse-grained scale, single site per residue? Residues can be represented by their C^{α}- or C^{β}-atoms. Bonded and non-bonded neighbors need not be distinguished. A justification for this approximation comes from the Gaussian network model which successfully describes the fluctuation dynamics of proteins despite

the use of a single parameter harmonic potential for all (bonded and non-bonded) contacts (Bahar *et al.*, 1997a; Demirel *et al.*, 1998). Another support is provided by Covell and Jernigan (1990). In their study, five protein sequences were threaded onto restricted spaces, in which the native structures are represented by one lattice point per residue. Non-bonded interactions that exist in the native folded state appeared as the most energetically favored interactions, which indicates that non-bonded interactions are also important as bonded interactions.

Employing a direct study of packing architecture, Raghunathan and Jernigan (1997) have indicated that residue coordination in folded proteins is regular and conforms to the cubic closest geometry also named face-centered cubic (fcc) geometry. This extremely regular packing is in contrast to the nuts-and-bolts description (Bromberg and Dill, 1994) and do not agree with observations of preferential but relatively "ductile" association of sidechains (Bahar and Jernigan, 1996).

In this study, the contrasting views of packing are reconciled. The method employed for presenting the fcc geometry as the ideal architecture of folded proteins is tested. Clusters of residues (central residue and neighboring residues within a shell of 6.8 Å) in proteins are constrained to fit to other predefined geometries. Those fits will be referred 'constrained fits' because the databank residue clusters are constrained to occupy a priori defined lattice sites.

The drive for maximizing the packing efficiency stabilizes secondary structures (helices) – as been pointed out in different studies (Chan and Dill, 1990; Maritan *et al.*, 2000; Stasiak and Maddocks, 2000). In this study, regularities in tertiary packing are examined.

Another noteworthy feature is that residues are closely clustered in all regions, core or surface. Thus, the lower coordination number of surface residues does not imply a lower density packing, but simply the occupation of a smaller subspace of the coordination volume of closely packed residues, the remainder being apparently allocated to solvent molecules. The requirement of achieving a high inter-residue packing geometry at all regions – and the tendency for assuming fcc packing geometry – could be used as effective

constraints for reducing the conformational space in the search and/or engineering of tertiary structures of proteins.

The present study finally focuses on threading of protein structures onto five different lattices. In the second chapter, general information on proteins and crystals is presented. Followed in the third chapter, the results of the study and discussion are given. Finally in the fourth chapter, conclusions drawn from the results and recommendations for further investigation are presented.

2. GENERAL INFORMATION ON PROTEINS AND CRYSTALS

2.1. Proteins

In this section, general information on proteins and their folded structures will be given.

2.1.1. Proteins are Essential to Life with Their Diverse Structures

Virtually every property that characterizes a living organism is affected by proteins. Nucleic acids are also essential for life, encode genetic information – mostly specifications for the structures of proteins – and the expression of that information depends almost entirely on proteins.

Life forms make use of many chemical reactions to supply themselves continually with chemical energy and to use it efficiently, but by themselves these reactions could not occur fast enough under physiological conditions (aqueous solution, 37°C, pH 7, atmospheric pressure) to sustain life. The rates of these reactions are increased, by many orders of magnitude, in organisms by the presence of enzymes, which are also proteins. Proteins store and transport a variety of particles ranging from macromolecules to electrons. They guide the flow of electrons in the vital process of photosynthesis; as hormones, they transmit information between specific cells and organs in complex organisms; some proteins control the passage of molecules across the membranes that compartmentalize cells and organelles; proteins function in the immune systems of complex organisms to defend against intruders the best known among which are the antibodies; and proteins control gene expression by binding to specific sequences of nucleic acids, thereby turning genes on and off. Proteins are the crucial components of muscles and other systems for converting chemical energy into mechanical energy. They also are necessary for sight, hearing, and the other senses. And many proteins are simply structural, providing the filamentous architecture within cells and the materials that are used in hair, nails, tendons and bones of animals.

In spite of these diverse biological functions, proteins form a relatively homogeneous class of molecules. All are linear polymers, built of various combinations of the same 20 amino acids. They differ only in the sequence in which the amino acids are assembled into polymeric chains. The secret to their functional diversity lies partly in the chemical diversity of the amino acids but primarily in the diversity of the three-dimensional structures that these building blocks can form, simply by being linked in different sequences. The awesome functional properties of proteins can be understood only in terms of their relationship to the three-dimensional structures of proteins (Creighton, 1993).

2.1.2. Proteins are Linear Heteropolymers That Have Unusual Interactions with Water

Proteins are linear polymers that have structural aspects different from synthetic polymers. The apparent size of a polymer of given length depends markedly upon the chemical nature of both the polymer and the solvent in which it is dissolved. In a good solvent, a polymer chain is highly expanded because the interactions between the solvent and the units of the chain are preferred over the interactions between the chain units themselves. In a poor solvent these reactions are reversed and the chain contracts in an attempt to exclude contact with the solvent as far as possible. These contracted chains will usually aggregate and precipitate in further attempts to avoid solvent contact. These relations are clear both conceptually and experimentally in such homopolymers as polyethylene or polystyrene. In a more complex case of a heteropolymer with varying properties along the chain, one can expect differing stiffnesses considering availability of all conformations in different parts of the chain. The detailed character of the solvent now begins to play a much more selective role in the behavior of the solutions. A given solvent may appear to be good for one part of the polymer and poor for another. Thus, the swelling or shrinking of a polymer in a particular polymer/solvent pair will reflect the subtle compensation of strongly opposing forces. The apparent stiffness of the various regions of the polymer will vary with the solvent conditions.

From such a point of view, water is actually a poor solvent for the polypeptide chain under conditions where the native folded state is stable. Very little water is normally

found within the interiors of globular proteins. They are about as compact in that sense as they can be. These dense molecules are, however, frequently very soluble in water without any evidence of aggregation or precipitation. Such behavior would normally be taken as evidence of a good solvent. In polypeptides this curious behavior is related to the differing chemical properties of the individual amino acid residues. This ambivalent relationship between polymer and solvent for the polypeptide/water pair appears to be at the root of the unusual behavior of this system and of all of the biological functions that follow from it (Creighton, 1992).

2.1.3. The Polymeric Nature of Proteins

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 (Creighton, 1993; Branden and Tooze, 1999). These amino acids are connected by peptide bonds formed by a condensation reaction between the amino group of one amino acid and the carboxyl group of another as a water molecule is liberated. The repeated amide N, C^{α}, and carbonyl C atoms of each residue form the backbone of the polypeptide from which the various sidechains project (Branden and Tooze, 1999).

The most common and perhaps the most useful way of classifying these amino acids is according to the polarities of their sidechains (R groups). This is because proteins fold to their native conformations largely in response to the tendency to remove their hydrophobic sidechains from contact with water and to solvate their hydrophilic sidechains. According to this classification scheme, there are three major types of amino acids: those with non-polar R groups, those with uncharged polar R groups, and those with charged polar R groups. The amino acids alanine, valine, leucine, isoleucine, methionine, proline, phenylalanine and tryptophan are usually classified as non-polar amino acids. The amino acids serine, threonine, asparagine, glutamine, tyrosine, and cysteine are commonly classified as uncharged polar amino acids. Cysteine has the unique property among amino acids to form disulfide bonds that has great importance in protein structure: It can join separate polypeptide chains or cross-link two cyteines in the same chain. The amino acids lysine, arginine, histidine, aspartic acid, and glutamic acid are charged polar amino acids.

Glycine is unique because it has no sidechain. The 20 amino acids vary considerably in their physicochemical properties such as polarity, acidity, basicity, aromaticity, bulk, conformational flexibility, ability to cross-link, ability to hydrogen bond, and chemical reactivity. These several characteristics, many of which are interrelated, are largely responsible for proteins' great range of properties (Voet and Voet, 1995).

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 therefore chiral atoms which can exist in two mirror-image forms, called the L-isomer and the D-isomer. Only L-amino acids are present in proteins.

In the late 1930s, Linus Pauling and Robert Corey began X-ray crystallographic studies of the precise structure of amino acids and peptides. One of the important findings was that peptide unit is rigid and planar. There is no freedom of rotation about the bond between the carbonyl carbon atom and the nitrogen atom of the peptide unit because this link has partial double-bond character. In contrast, the link between the C^{α} atom and the carbonyl carbon atom is a pure single bond. The bond between the C^{α} atom and the peptide nitrogen atom also is a pure single bond. Consequently, there is a large degree of rotational freedom about these bonds on either side of the rigid peptide unit (Stryer, 1988). The combinations of the rotations about these two bonds are described as psi and phi angle combinations and shown in Ramachandran plots on which densely populated regions indicate conformations of α-helices and β-strands.

2.1.4. Four Levels of Protein Structure

Four levels of protein structure are defined. Primary structure is the amino acid sequence, or, in other words, the arrangement of amino acids along a linear polypeptide chain. Secondary structure occurs mainly as α-helices or β-strands. Certain amino acid sequences favor either α-helices or β-strands; others favor formation of loop regions. Secondary structure elements usually arrange themselves in simple motifs. Motifs are formed by packing sidechains from adjacent α-helices or β-strands close to each other. Tertiary structure is formed by packing secondary structures into one or several compact

globular domains. Proteins that contain more than one chain have a quaternary structure. The primary structure of a protein specifies its final tertiary structure.

The main driving force for folding water-soluble globular protein molecules is to pack hydrophobic sidechains into the interior of the molecule, thus creating a hydrophobic core and hydrophilic surface. The core is densely packed so the hydrophobic side chains with different shapes are packed like a three-dimensional jigsaw puzzle. The problem of constructing a hydrophobic core from the hydrophilic chain is solved by the formation of regular secondary structures such as α-helices or β-sheets that are characterized by their mainchain NH and CO groups participating in hydrogen bonds and packing these secondary structures within the interior of the protein molecule (Branden and Tooze, 1999).

2.1.5. Folded Proteins are Globular

Folded proteins are globular and highly compact with respect to random coil conformation and their such folded structures are specific to their amino acid sequence. To determine the relationship between the amino acid sequence and the folded structure is known as the folding problem. The problem is still a challenge for protein science. Characterization of the structure of a protein is a prerequisite for understanding its function. Therefore understanding the principles of folding and the relation between structure and function is of great importance for designing new proteins and drugs.

2.2. Crystals

2.2.1. Definition of Crystals

It is understood by the invention of optical microscope and X-ray diffraction that the regularity observed in crystals at the macroscopic level is due to an underlying regular pattern in the arrangement of atoms, ions or molecules. Crystals possess a periodicity that produces long-range order so that the local atomic arrangement is repeated at regular intervals in the three dimensions of space. The atoms of a small volume called the unit cell

are repeated at specific intervals. All unit cells in a crystal are identical (Van Vlack, 1989; Petrucci, 1985).

2.2.2. Crystal Lattices

The lattices that will be considered in the present study are simple cubic (sc), body-centered cubic (bcc), embedded (emb), fcc and hexagonal closest packed (hcp) lattices. The unit cells of sc, bcc and fcc lattices are presented in Figure 2.1.

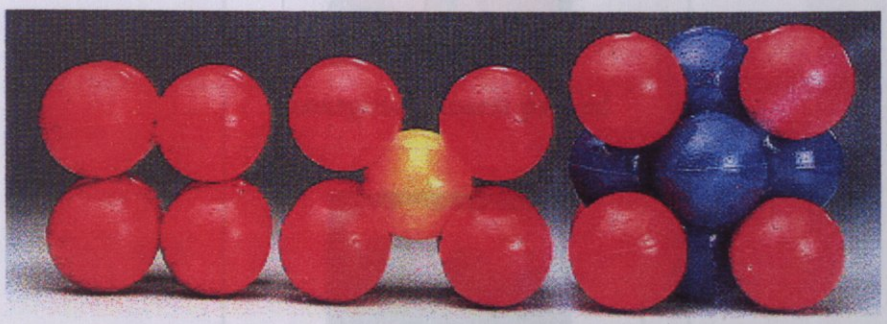

Figure 2.1. Space filling models of sc, bcc and fcc unit cells

In simple cubic lattice, a site is in contact with six other sites: four at the same plane, one at the upper plane and one at the lower plane. The number of contact sites are defined as coordination number. The coordination number in bcc lattice is eight: four contact sites at the upper plane and four at the lower plane. The coordination number in fcc lattice is twelve: six contact sites at the same plane, three at the upper plane, three at the lower plane. Fcc structure is also called cubic closest packed because it has the known but recently proved feature (Cipra, 1998; Sloane, 1998), of being the densest packing that can be achieved with identical size hard spheres. One more structure has the densest packing feature: hcp which also has a coordination number of twelve. These structures are illustrated in Figure 2.2. Both in the hexagonal close packed structure and cubic close packed structure the central site has six contacting sites at the same plane, three at the upper and three at the lower plane. The difference is that the upper and lower three sites in the cubic closest packed lattice are staggered (yellow and blue layers) but they are not in hexagonal closest packed lattice. Therefore the structure is repeated in two layers (red and yellow) in hexagonal closest packed lattice whereas it is repeated in three layers (red,

yellow and blue) in cubic closest packed lattice. To verify that fcc and cubic closest packed presentations are the same with different orientations Figure 2.3 (Petrucci, 1985) is presented. The embedded lattice has a coordination number of ten: six simple cubic contact sites and four bcc-like contact sites that have tetrahedral arrangement.

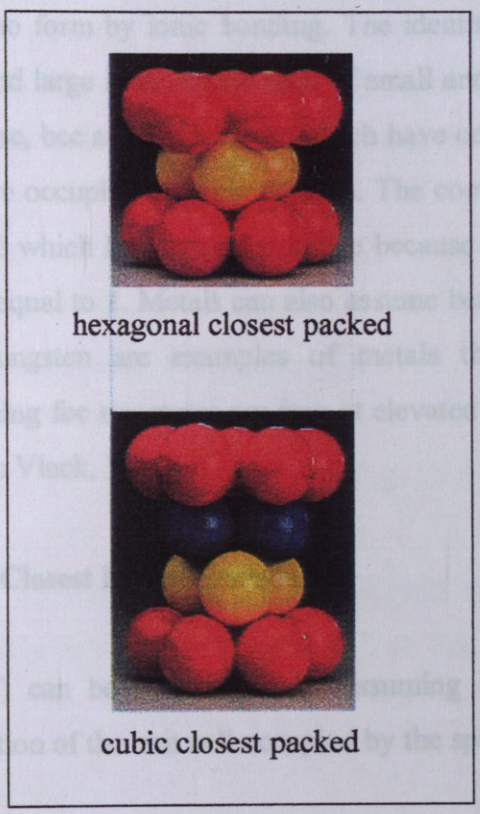

Figure 2.2. Closest packed structures

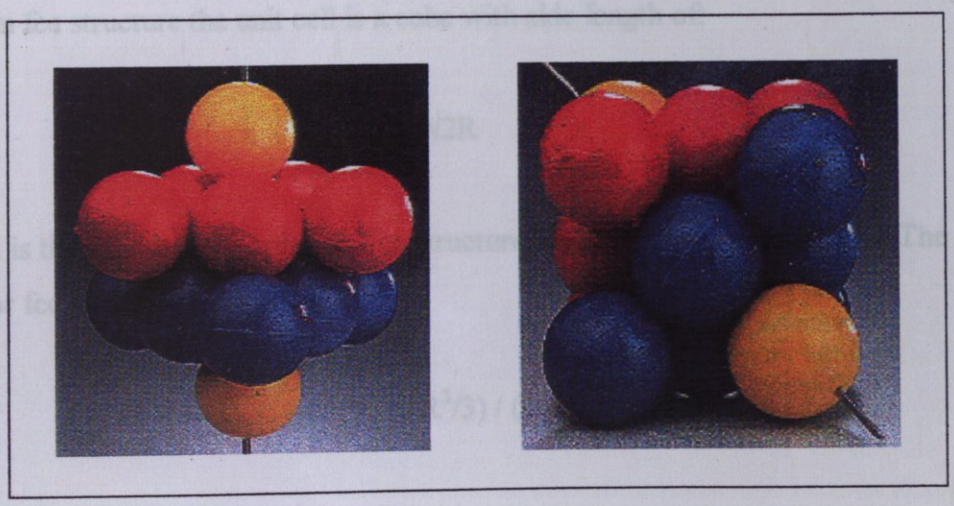

Figure 2.3. Two views of the fcc unit cell

2.2.3. Crystal Structures

Crystals form through covalent, ionic and metallic bonding. The familiar diamond lattice is a tetrahedral lattice formed by sp^3 hybrid covalent bonds and it has a coordination number of four. Lattices also form by ionic bonding. The identities of lattices depend on the relative sizes of small and large ions. As the ratio of small and large ion radii increases from 0.22 to 1, tetrahedral, sc, bcc and fcc lattices (which have coordination numbers of 4, 6, 8 and 12 respectively) are occupied by ionic crystals. The coordination number in pure metals can be as high as 12 which leads to a fcc lattice because the atoms have only one size thus have a radii ratio equal to 1. Metals can also assume bcc structures. Iron at room temperature, chromium, tungsten are examples of metals that have bcc structures. Examples for metals assuming fcc structures are iron at elevated temperatures, aluminum, copper, lead and silver (Van Vlack, 1989).

2.2.4. Packing Factors in Closest Packed Structures

Packing factor (PF) can be determined by assuming hard spheres model and calculating the volume fraction of the unit cell occupied by the spheres:

$$PF = \text{volume of spheres} / \text{volume of unit cell} \tag{2.1}$$

In fcc structure the unit cell is a cube with side length of:

$$a = 2\sqrt{2}R \tag{2.2}$$

where R is the radius of the spheres. Fcc structure has four atoms per unit cell. The packing factor for fcc structure is therefore:

$$PF = 4\ (4\pi R^3/3) / a^3 = 4\ (4\pi R^3/3) / (2\sqrt{2}R)^3 = 0.7405 \tag{2.3}$$

On the other hand, if we consider heterogeneous packing, the PF for equal composition of large and small spheres, the small spheres being contiguous to the large spheres in fcc arrangement yields:

$$PF = [4\,(4\pi r^3/3) + 4\,(4\pi R^3/3)] / (2r+2R)^3 \qquad (2.4)$$

The number of small spheres per unit cell is four (one full at the middle and 12 quarters at the sides of the unit cell). The length of the cube is

$$a = 2\sqrt{2}R = 2r+2R \qquad (2.5)$$

so that substitution for $r = (\sqrt{2}-1)R$ in Equation (2.4) gives

$$PF = [4\,(4\pi((\sqrt{2}-1)R)^3/3) + 4\,(4\pi R^3/3)] / (2(\sqrt{2}-1)R+2R)^3 = 0.7931 \qquad (2.6)$$

This result implies that the packing factor of 0.7405 for the closest packing of identical spheres is exceeded in heterogeneous packing (Van Vlack, 1989).

3. RESULTS AND DISCUSSION

3.1. Definition of Residue Clusters and Their Bundles of Directional Vectors

A total set of 28730 clusters of residues comprised of one central residue and m bonded and non-bonded neighbors within its first coordination shell (of radius 6.8 Å) have been collected from a non-homologous set of 150 databank structures for this study. Figure 3.1 illustrates a cluster formed by a central residue (Gly65) in myoglobin. Its Protein Data Bank (PDB) code is 1mbn (Watson, 1969). The cluster has m = 10 neighboring residues. The dashed lines represent the directions of the coordination vectors that connect the central residue to its neighbors. Residues are represented by their C^{α}-atoms in the case of Gly and C^{β}-atoms for all other amino acids. The present cluster of coordinating residues contains the residues 22, 25, 26, 29, 62, 64, 66, 67, 68, 69 of the myoglobin chain. A bundle of unit directional vectors pointing from the central residue towards the m coordinating residues characterizes each cluster.

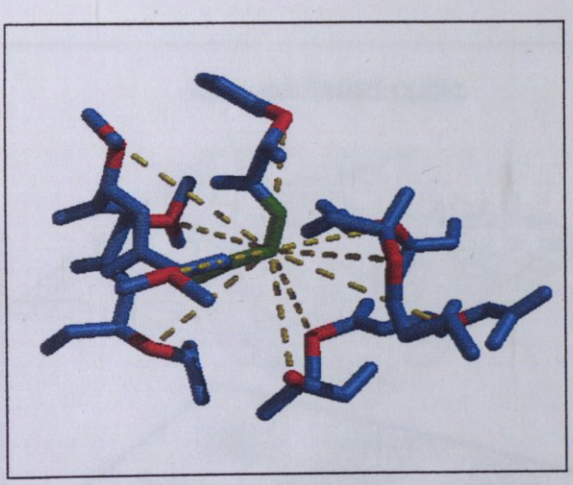

Figure 3.1. Residue cluster in myoglobin

3.2. Constrained Fit of Residue Clusters to Lattices

It has been proposed that folded proteins have an ideal architecture corresponding to fcc geometry (Raghunathan and Jernigan, 1997). To test the validity of this argument, a

constrained Monte Carlo algorithm is employed. 1000 clusters extracted from PDB protein structures are reoriented to fit to fcc unit directions. A cluster is chosen and rotated randomly as a rigid body. The rotation is accepted if the distance deviation between the m unit vectors of residue cluster and 12 unit vectors of fcc geometry is decreased, otherwise it is rejected. The deviation is called the constrained root-mean-square (RMS) deviation and calculated from:

$$<\varepsilon>_{cons} = \Sigma_k \varepsilon_k / N \qquad (3.1)$$

where $1 \leq k \leq N = 1000$ and ε_k is the distance deviation between the k^{th} cluster and target fcc directions.

The resulting probability surface and corresponding contours are shown in Figure 3.2. Twelve peaks emerge at the locations (shown by polar angle, θ, and azimuthal angle, ϕ) corresponding to fcc geometry (see Table 3.1). This shows that the directional vectors of residue clusters have been distributed to 12 coordination sites of fcc.

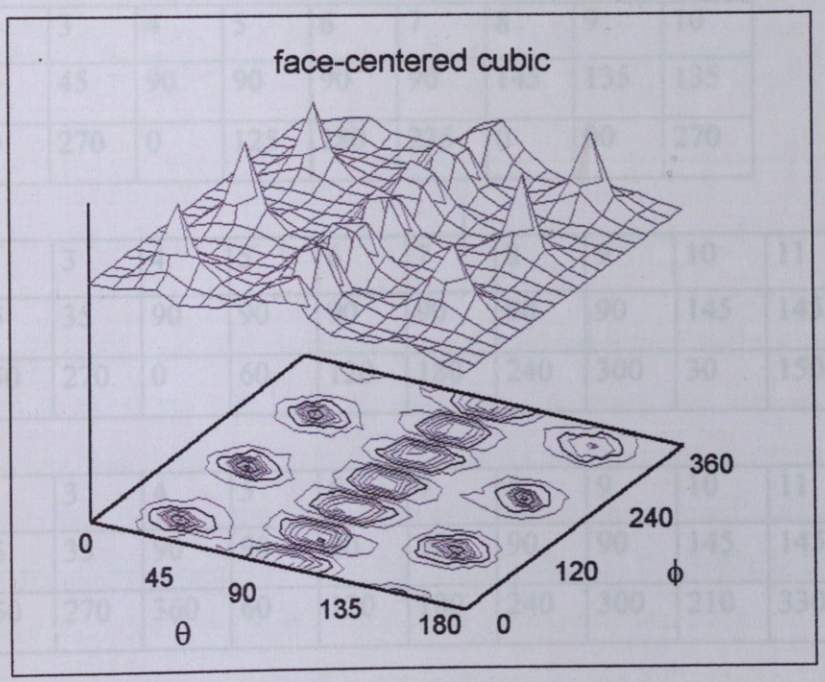

Figure 3.2. Fcc geometry obtained by constrained fit method

To test if the fcc geometry is the only geometry that can be obtained by this method, four other target lattices were also targeted. These lattices, described in the second chapter, are sc, bcc, emb and hcp lattices. Their coordination directions are given in Table 3.1. It is seen in Figure 3.3 that PDB structures can also be fit to these geometries using the constrained fit algorithm. Thus, fcc geometry cannot be viewed as the ideal architecture of folded proteins using the latter algorithm adopted by Raghunathan and Jernigan (1997).

Table 3.1. Coordination angles for different lattice geometries

sc	1	2	3	4	5	6
θ (°)	45	45	90	90	135	135
ϕ (°)	90	270	0	180	90	270

bcc	1	2	3	4	5	6	7	8
θ (°)	55	55	55	55	125	125	125	125
ϕ (°)	45	135	225	315	45	135	225	315

emb	1	2	3	4	5	6	7	8	9	10
θ (°)	35	45	45	90	90	90	90	145	135	135
ϕ (°)	0	90	270	0	125	180	235	0	90	270

hcp	1	2	3	4	5	6	7	8	9	10	11	12
θ (°)	35	35	35	90	90	90	90	90	90	145	145	145
ϕ	30	150	270	0	60	120	180	240	300	30	150	270

fcc	1	2	3	4	5	6	7	8	9	10	11	12
θ (°)	35	35	35	90	90	90	90	90	90	145	145	145
ϕ (°)	30	150	270	360	60	120	180	240	300	210	330	90

The constrained fit algorithm was executed up to $3x10^6$ iterations for the five fits. The RMS deviations decreased during the course of the executions as can be observed in

part (a) of Figure 3.4. Part (a) and part (b) of the Figure 3.4 display the RMS deviations up to $2x10^6$ iterations.

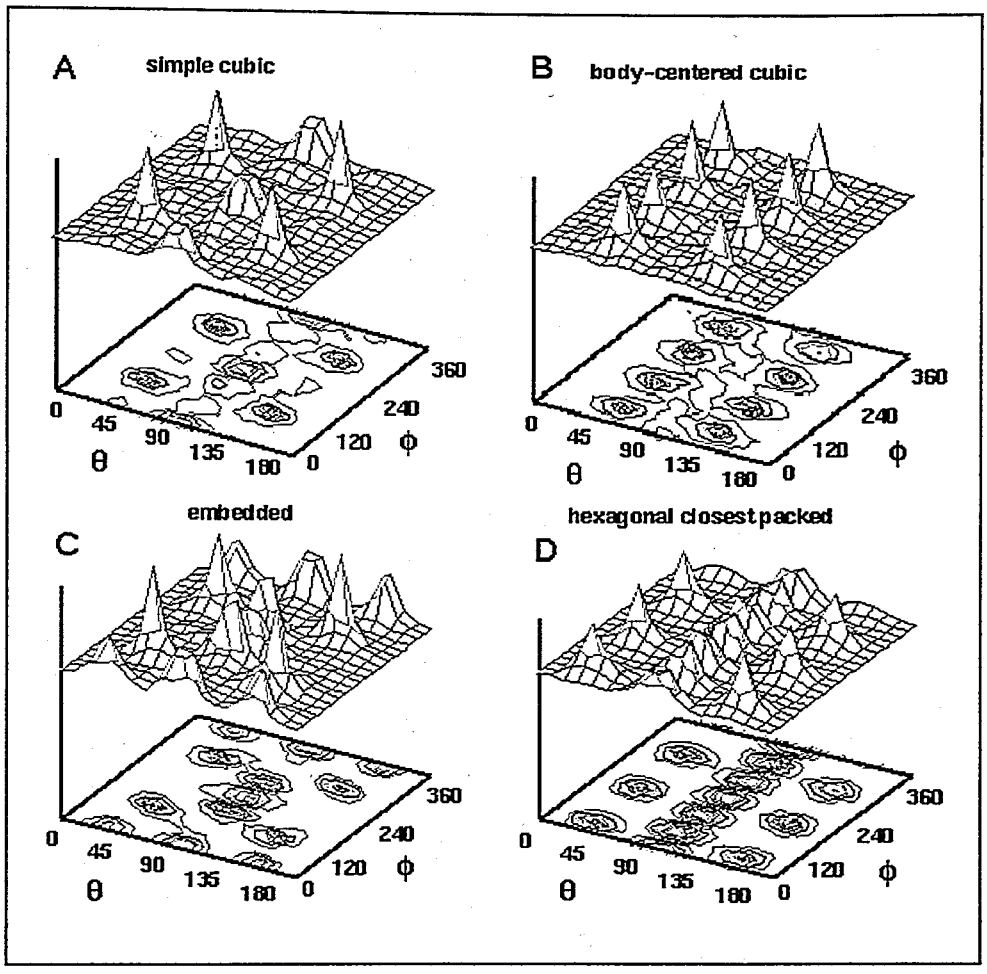

Figure 3.3. Sc, bcc, emb, hcp geometries obtained by constrained fit method

Constrained RMS deviations between the clusters and target lattices are found to decrease to 0.20, 0.21, 0.25, 0.30, 0.37 for the hcp, fcc, emb, bcc and sc unit cells, respectively, starting from approximately 0.65 for randomly oriented clusters. This does not however imply that clusters, themselves are optimally superimposed onto each other. The point is that the coordination number of the target lattice (z) is usually larger than that (m) of the cluster (coordination number of residue clusters is approximately 6.5 on the average) such that any of the z!/[m!(z-m)!] combinations of the z directions taken m at a time could be adopted for achieving the best fit. Different clusters therefore select different subsets amongst the z accessible choices, and this freedom results in a relatively poor (RMS deviation = 0.56 - 0.60 as shown in part (b) of Figure 3.4) superimposition between

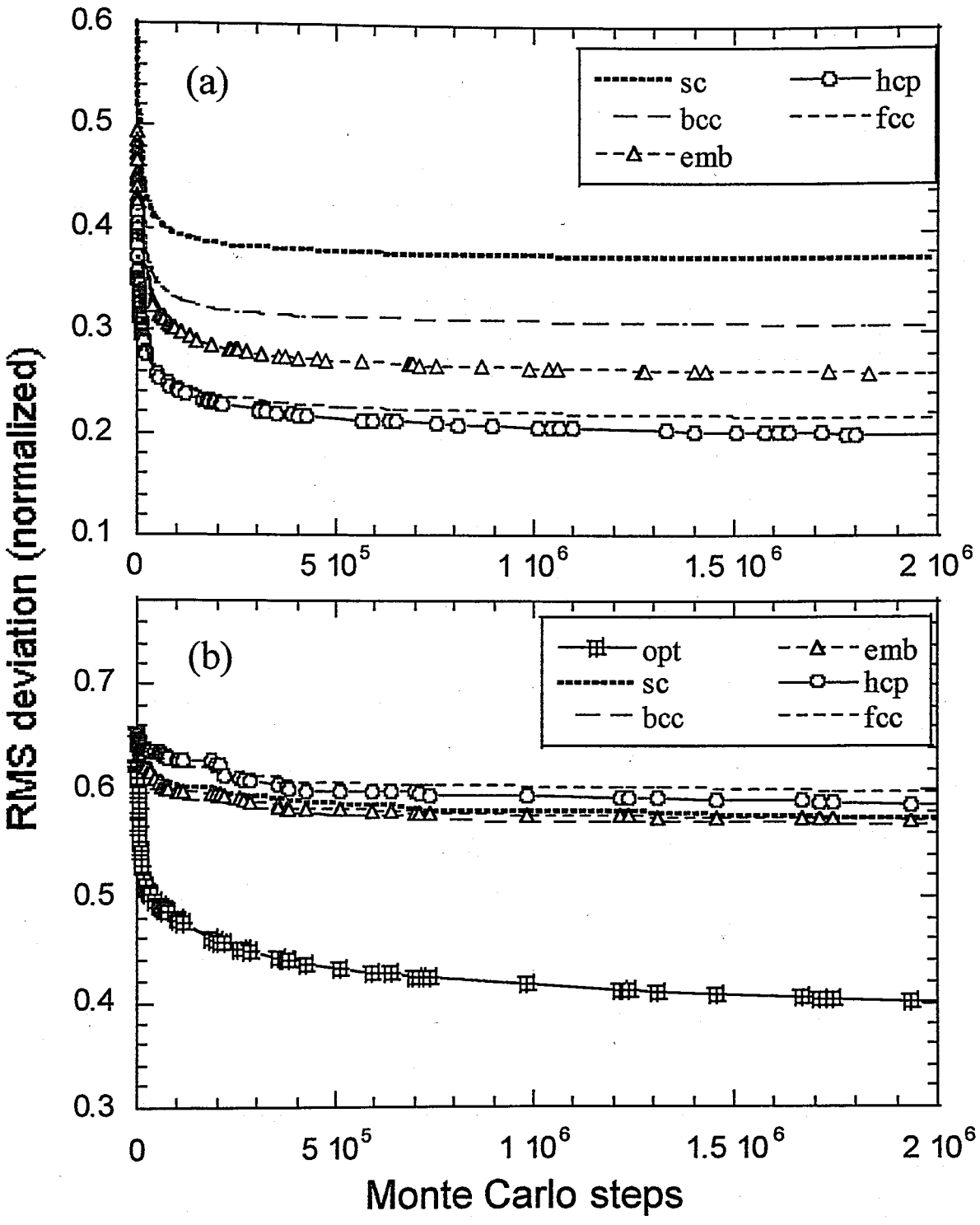

Figure 3.4. RMS deviation in Monte Carlo algorithms for (a) constrained fit to target lattices and (b) optimal superimposition of clusters

the clusters, themselves. As an example, two vector sets that have coordination number of m = 6 could be perfectly well fitted onto two different subsets of fcc target vectors but they would not be superimposed onto each other. When the clusters are optimally

superimposed onto each other, the RMS deviation between them decreases to 0.39, as shown in part (b) of Figure 3.4 with the abbreviation 'opt'. This kind of fit of residue clusters, called the optimal superimposition, is described in the next subsection.

3.3. Optimal Superimposition of Clusters. Generic Distribution of Amino Acids

The generic packing geometry of residue clusters should be free of any predefined geometries. This geometry is found in the present thesis from optimal superimposition of residue clusters, irrespective of amino acid type or coordination number. An unconstrained Monte Carlo algorithm is employed to this aim. The reason to use it is justified as follows: Although an exhaustive search would be possible for constrained fit, it is impossible from the point of view of execution time to perform an optimal superimposition because of the large number of degrees of freedom in rotating N = 1000 residue clusters. The method is employed by selecting a residue cluster and rotating it randomly as a rigid body. Then its mean deviation from all other residue clusters is calculated and compared with the preceeding value. If the deviation is decreased, the rotation is accepted, otherwise it is rejected. The deviation $\varepsilon_{i,j}$ between each pair of clusters (i, j) is calculated as

$$\varepsilon_{i,j} = \Sigma_k s_k / \min(m_i, m_j) \tag{3.2}$$

where s_k is the distance between the tips of the closest vectors selected from the two clusters and m_i is the coordination number of the i^{th} cluster. The term $\min(m_i, m_j)$ denotes the minimum of the two coordination numbers. For the superimposition of N clusters, all pairwise combinations of vector sets are taken. The pairwise mean deviations are calculated as

$$\langle\varepsilon\rangle = \Sigma_i \Sigma_j \varepsilon_{i,j} / [N(N-1)/2] \tag{3.3}$$

The denominator simply represents the total number of combination of clusters.

This algorithm has been executed up to $3\text{x}10^6$ Monte Carlo steps and the coordination geometry in part (a) of Figure 3.5 has been obtained. As a verification of the statistical accuracy of the results, longer runs (10^7) were executed. The coordination

geometry displayed in part (d) of Figure 3.5 was obtained. The distribution preserves the same features, except for a slight enhancement of the most densely occupied sites. The running time is about 50 hours (real time) for $3x10^6$ iterations during optimal superimposition of 1000 clusters, and grows exponentially with increasing number of residue clusters and coordination number of residue clusters.

The results of optimal superimposition are displayed in Figure 3.5. The number of peaks in part (a) is significantly lower than the coordination numbers of the lattices targeted in constrained fit calculations, except for the sc lattice. The peaks indeed reflect the average coordination number ($\approx$ 6.5) in the databank clusters. The most populated coordination directions are listed below.

	1	2	3	4	5	6	7
θ (°)	110	105	70	65	115	165	120
ϕ (°)	170	250	210	130	90	270	20
$P(\theta,\phi)$	0.15	0.10	0.08	0.02	0.10	0.13	0.05

The last row designates the respective probabilities of the individual coordination directions, directly found from the fraction of residues located within a 20° solid angle deviation with respect to the central directional vectors. The sum of these probabilities is 0.63, i.e. more than half of the residues occupy these coordination states, while the remainder selects any other suitable positions in space. The random probability of occupancy of a solid angle of 20° is 0.03, found from the ratio of its surface area to that of the entire coordination sphere. For random packing, the total probability of the above listed seven coordination states would therefore be 0.21. The difference between random (0.21) and observed (0.63) values discloses the existence of a preferred packing architecture, favored by a factor of 3 with respect to random packing.

One might describe this preference in coordination directions to the regularities of the backbone, or the dominant effect of the bonded neighbors, but this is not the case. The part (b) of Figure 3.5 displays the coordination geometry of bonded residues and part (c) displays non-bonded residues. So, the most probable coordination sites remain unchanged when bonded residues are excluded from the clusters. This results corroborates previous

analyses suggesting that bonded and non-bonded neighbors need not be necessarily distinguished for satisfactorily describing inter-residue contact topology.

Interestingly, the directional vectors are not uniformly distributed in space, but closely clustered to cover only a portion of the coordination sphere. The remaining empty (or sparsely occupied) regions can be anticipated as those allocated to solvent molecules.

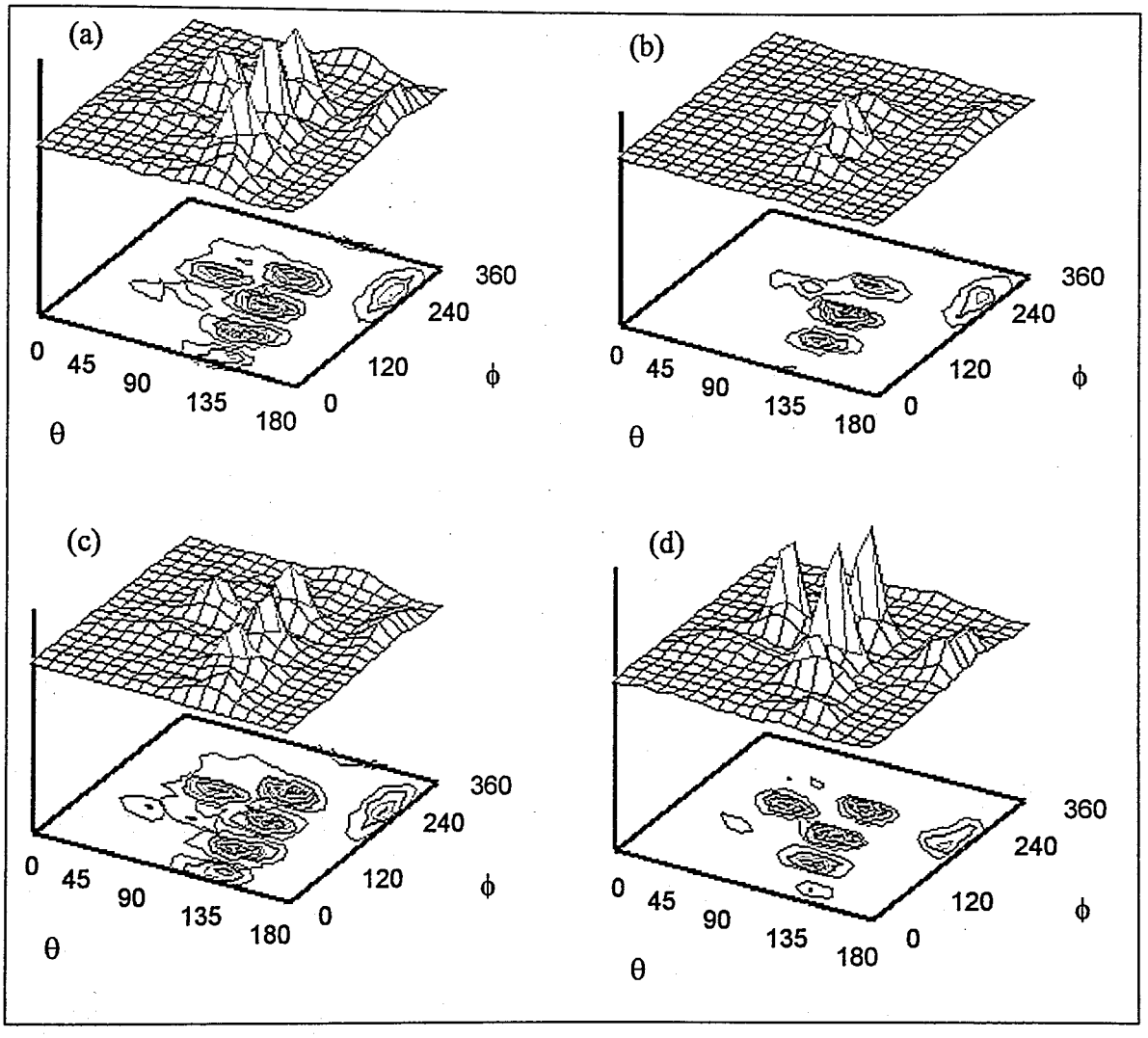

Figure 3.5. Coordination geometry obtained by optimal superimposition for (a) 'all' residues, (b) bonded residues, (c) non-bonded residues, (d) 'all' residues after a longer run

3.4. Optimal Superimposition for Specific Amino Acids at the Center of Clusters

Figure 3.6 displays the optimal superimposition results for each type of amino acid occupying the center of the examined clusters. The complete set of clusters has been considered to this aim. The clusters have been grouped into 20 subsets according to the identity of the central residue, and optimally superimposed within each group.

Recurrent patterns with slight variations can be detected for different amino acids in Figure 3.6. An additional eight coordination state emerges in some cases. Table 3.2 lists the resulting residue-specific coordination states. It is observed that (i) there is a rather weak residue specificity, the coordination directions being preserved with only small deviations in coordination angles, and (ii) not all coordination states are occupied in the neighborhood of all types of amino acids. Residues near a central amino acid usually select sites from amongst these eight most probable directions, depending on the type of amino acid.

The last two rows in Table 3.2 list the mean values for the directional vectors characterizing the most frequently occupied coordination sites. The first of these lists simply the seven sites already identified for all clusters in part (a) of Figure 3.5 irrespective of residue type, along with the eight site that is preferentially occupied in a number of specific residues. And the last row is another representation of the same set of directional vectors, expressed with respect to a different reference frame. The transformation to this new frame aims at stipulating the correspondence to the fcc geometry.

3.5. Coordination of Core Residues

From optimal superimposition results, it was deduced that some portion of the coordination sphere was left unoccupied – or more exactly weakly populated. It is anticipated that this feature is indicative of the solvent-exposed regions. To verify this conjecture, subsets of clusters composed of m = 10 or more residues have been considered. These are evidently densely packed clusters, and could be viewed as reflecting the behavior of core residues. Their optimal superimposition yielded the distribution of coordination angles displayed in part (a) of Figure 3.7. There are now more peaks, and

these are more or less evenly distributed in space. The directional vectors characterizing the centers of coordination are listed below.

	1	2	3	4	5	6	7	8	9	10
θ (°)	45	45	45	95	105	60	100	85	105	140
ϕ (°)	40	180	280	360	60	100	140	240	300	220

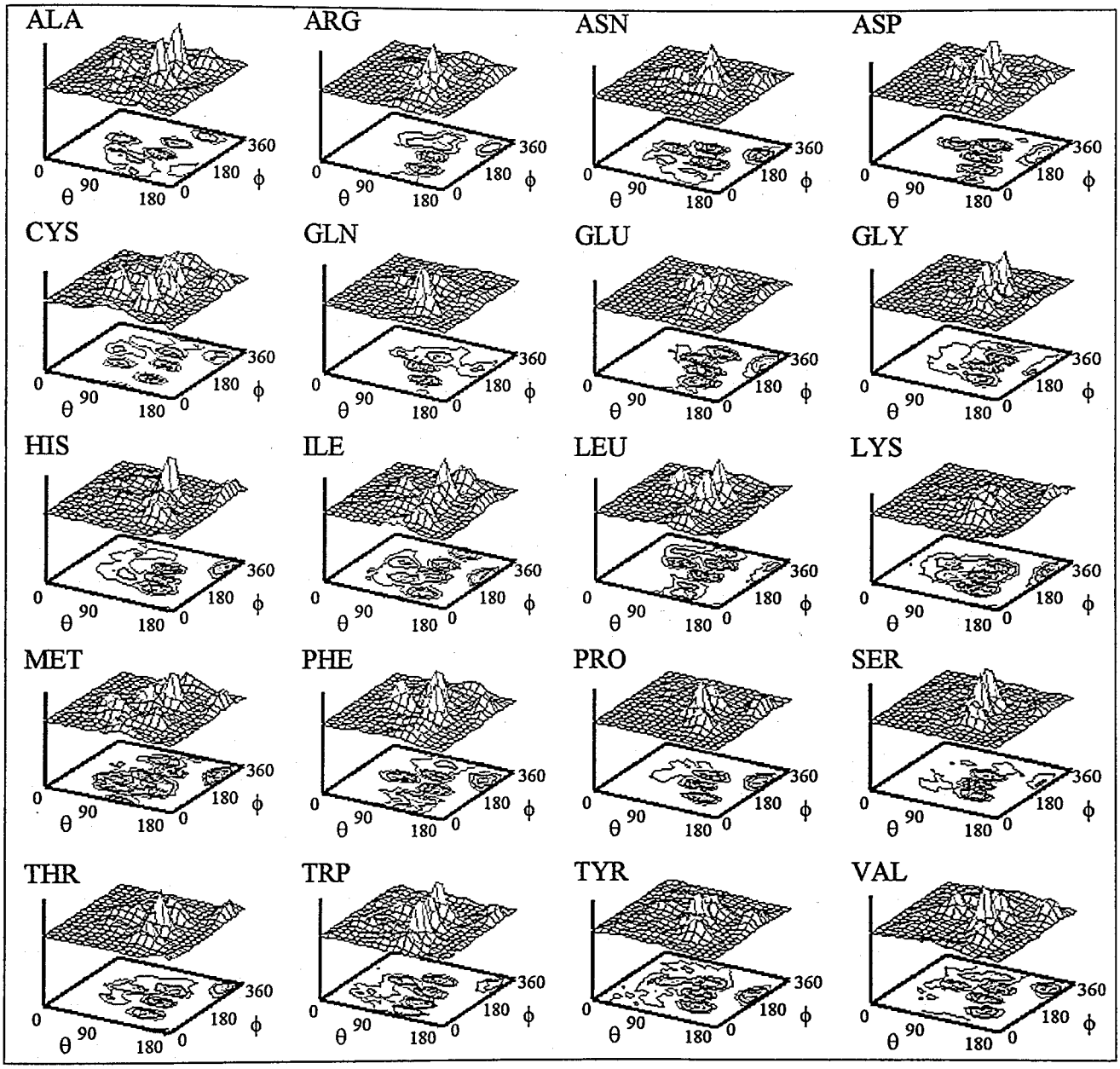

Figure 3.6. Coordination geometry around specific amino acids

Table 3.2. Spherical angles for the coordination directions of residues near specific amino acids

		1	2	3	4	5	6	7	8
ALA	θ, ϕ	115, 170	115, 250	50, 210	-	90, 110	135, 330	115, 30	
ARG	θ, ϕ	115, 170	105, 240	75, 230	-	125, 90	160, 300	115, 20	
ASN	θ, ϕ	120, 170	95, 250	65, 195	65, 130	120, 90	155, 260	-	
ASP	θ, ϕ	105, 170	100, 250	65, 210	-	120, 90	165, 260	130, 20	
CYS	θ, ϕ	120, 170	100, 255	45, 210	70, 130	120, 85	155, 300	95, 0	
GLN	θ, ϕ	95, 170	95, 240	55, 210	-	120, 90	160, 230	135, 20	
GLU	θ, ϕ	115, 170	105, 280	85, 225	-	125, 90	165, 270	-	
GLY	θ, ϕ	115, 170	115, 245	-	75, 130	125, 105	165, 220	-	100, 310
HIS	θ, ϕ	110, 160	110, 230	-	70, 140	125, 95	165, 310	-	
ILE	θ, ϕ	115, 170	105, 250	60, 210	95, 130	130, 70	165, 240	-	110, 330
LEU	θ, ϕ	115, 170	105, 250	65, 210	-	105, 90	165, 300	120, 40	95, 320
LYS	θ, ϕ	105, 170	105, 250	55, 210	-	120, 90	165, 290	-	
MET	θ, ϕ	95, 190	110, 245	-	60, 130	125, 110	165, 280	95, 10	85, 290
PHE	θ, ϕ	120, 170	100, 250	65, 190	-	120, 90	155, 270	120, 25	95, 320
PRO	θ, ϕ	105, 170	105, 230	-	-	125, 90	165, 270	-	
SER	θ, ϕ	105, 170	100, 230	-	60, 110	115, 110	165, 260	-	
THR	θ, ϕ	115, 170	110, 250	75, 210	60, 130	120, 90	165, 320	-	
TRP	θ, ϕ	115, 170	100, 250	60, 210	-	115, 110	165, 310	115, 45	
TYR	θ, ϕ	105, 170	100, 240	50, 195	-	120, 90	165, 230	-	
VAL	θ, ϕ	110, 170	110, 235	60, 210	-	125, 90	155, 290	-	90, 300
AVG	θ, ϕ	110, 170	105, 250	70, 210	65, 130	115, 90	165, 270	120, 20	100,310
$AVG^{(fcc)}$	θ, ϕ	40, 10	35, 200	45, 285	95, 350	105, 50	55, 115	120, 115	90, 180

The fraction of residues occupying these sites was calculated to be 0.65, again allowing for 20° deviations with respect to the central directional vectors. The individual probabilities vary all as 0.07 ± 0.01. In this case the random distribution would give a total probability of 0.30 for the ten regions. The observed total probability (0.65) thus indicates an enhancement in the selection of these directions by a factor about 2.

The distributions of coordination sites for three special cases, Ala, Cys and Gly, are presented in the respective parts (b)-(d) of Figure 3.7. Their corresponding coordination states are given in Table 3.3. Comparison of parts (a) and (b) shows that the geometry of clusters for 'all' types of central amino acid is well represented by that of alanine. Parts (c) and (d), on the other hand, exhibit some distinctive features. Glycine samples an additional eleventh state, consistent with its higher conformational freedom.

Table 3.3. Coordination states of amino acids in the core

		1	2	3	4	5	6	7	8	9	10	11
ALL	θ (°)	45	45	45	95	105	60	100	85	105	140	
	ϕ (°)	40	180	280	360	60	100	140	240	300	220	
ALA	θ (°)	40	50	50	95	115	65	105	85	105	145	
	ϕ (°)	20	170	280	340	60	100	160	220	280	220	
CYS	θ (°)	35	40	45	70	130	70	115	85	100	150	
	ϕ (°)	60	180	290	360	60	120	150	220	320	250	
GLY	θ (°)	30	40	50	90	85	65	105	80	105	145	130
	ϕ (°)	30	180	280	340	40	90	160	220	280	230	45

For illustrative purposes, the superimposition of two bundles of vectors corresponding to two residue clusters (m = 10) is depicted in Figure 3.8.

3.6. Comparison of Optimal Packing Architecture with Lattice Geometries

Let us now examine the generic coordination geometry of core residues ($m \geq 10$) in more detail. The azimuthal angle difference between the successive directional vectors 4-9 is approximately 60°. This approximately hexagonal organization conforms to the high propensity of the value $\Delta\phi = 60°$ noted in our previous examination (Jernigan and Bahar, 1999) of triplets of residues. The sites 4-9 identified here may be viewed as composing the

middle layer in a regular close-packed arrangement. Out of the remaining four sites, three (1-3) lie in the upper hemisphere ($\theta = 45°$) and are separated by $120 \pm 20°$. Interestingly, this arrangement closely approximates both of the hcp and fcc geometries. Finally, the tenth residue occupies a lower layer ($\theta = 140°$) position that can be compared to a staggered site of the fcc lattice. Thus, overall the optimal geometry in the core closely resembles that of an fcc packing with two empty sites. We may call that an incomplete, distorted fcc packing.

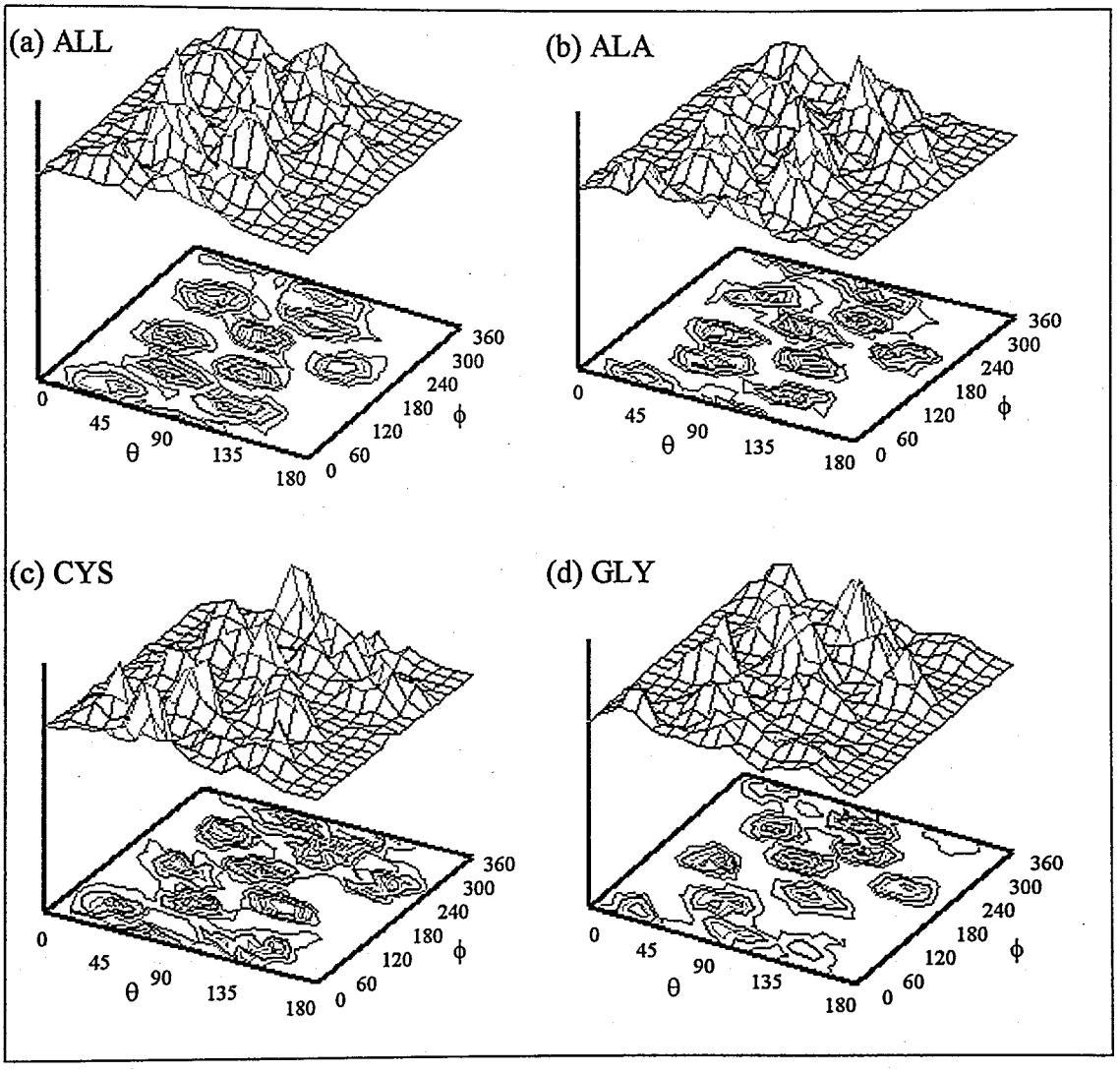

Figure 3.7. Coordination geometry around (a) 'all' amino acids, (b) only Ala, (c) only Cys, (d) only Gly, in the core regions of the examined proteins

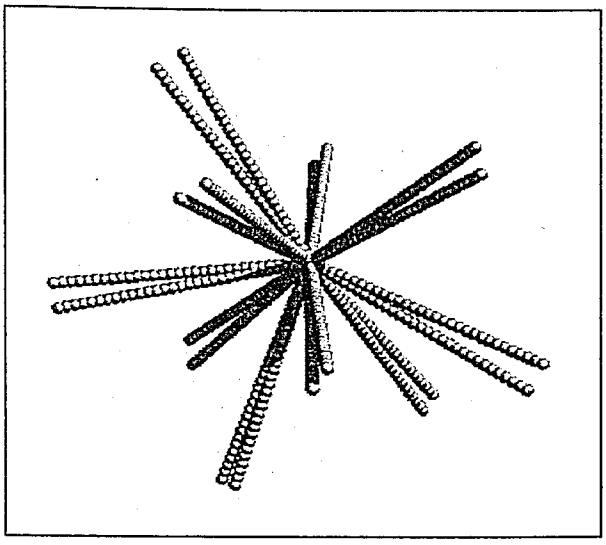

Figure 3.8. Two bundles of vectors corresponding to two residue clusters (m = 10) superimposed onto each other

The so-called distortions with respect to the fcc lattice can be rationalized by a closer examination. First, there are two unoccupied sites, because the subset of cluster under examination consist mainly of clusters of m = 10 coordinating residues. Additional calculations performed with higher density clusters ($m \geq 12$) indeed verified that the remaining two unoccupied sites become also filled upon optimal superimposition of such clusters which can be seen in Table 3.4. On the other hand, the preference for site 10 over the unoccupied sites 11 and 12 could be associated with the relatively large (100° instead of 60°) azimuthal angle difference between the two nearest hexagonal sites 7 and 8. The relatively small (40°) azimuthal angle difference between the site 6 and its nearest neighbors (5 and 7) on the other hand, is apparently accommodated by a polar angle distortion (60° instead of 90°).

As a further validation of the above identified distorted, incomplete fcc geometry, we checked if the seven optimal coordination directions found for all residues could also be associated with the directional vectors of fcc packing. This was indeed confirmed, as presented in the last row of Table 3.2, and reproduced in Table 3.4. The serial numbers of directional vectors are rearranged in Table 3.4, so as to match those of the fcc geometry. Interestingly, even the additional site observed for a number of specific residues (Table 3.2) precisely conforms to one of the directional vectors (7^{th}) of the fcc geometry.

Table 3.4. Coordination states of surface to core central residues

Coord. number		Coordination states (°)												
		1	2	3	4	5	6	7	8	9	10	11	12	Ptot
3≤m≤4	θ	40	45			95	90							0.40
	ϕ	30	170			50	110							(3.3)
3≤m≤14	θ	40	35	45	95	105	55	90					120	0.63
	ϕ	10	200	285	350	50	115	180					115	(3.0)
m≥10	θ	45	45	45	95	105	60	100	85	105	140			0.65
	ϕ	40	180	280	360	60	100	140	240	300	220			(2.2)
m≥12	θ	45	25	50	70	100	75	80	75	105	140	145	130	0.76
	ϕ	60	170	280	340	40	120	160	220	260	200	330	120	(2.1)
fcc	θ	35	35	35	90	90	90	90	90	90	145	145	145	–
	ϕ	30	150	270	360	60	120	180	240	300	210	330	90	

Finally, surface residues (subset of clusters that have coordination numbers of 4 or less) have been examined. Evidently, fewer peaks are observed in this case, but these can be easily allocated to four of the fcc geometry as can be seen in Table 3.4. The fraction of residues occupying these four sites near solvent-exposed residues is calculated to be 0.40, again allowing 20° solid angle deviations with respect to the centers of coordination sites. Random occupation would on the other hand yield a ratio of 0.12, i.e. the preference for these sites is enhanced by a factor of 3.3 over random distribution.

The observed twelve directional vectors identified for the most densely packed core residues ($m \geq 12$) exhibit an occupation ratio of 0.76, as opposed to the random ratio of 0.36. Therefore approximately three quarters of residues occupy these 'regular' coordination directions, while the remainder are 'disordered' when the highest coordination clusters are examined. A gradual enlargement in the fraction of residues that occupy the 'regular' coordination sites is seen as increasingly denser clusters are examined. This tendency is revealed in the last column of Table 3.4. But from the point of view of enhancement relative to random distribution, an opposite tendency is observed, i.e. residues in densely packed regions cannot effectively select from the allowed coordination sites, apparently due to the more severe constraints that are imposed by non-bonded interactions. The enhancement factors, simply found from the ratio of the actual occupancies to those expected for a random distribution, are given in parentheses in the last column.

For a visual comparison of the fcc geometry and the packing geometry of the densest residue clusters ($m \geq 12$), Figure 3.9 is displayed. In this figure, the fcc directional vectors and the corresponding residue cluster model (first row) are depicted, along with the 12 directional vectors given in the Table 3.4 for $m \geq 12$ (second row). The second column displays the corresponding space-filling models.

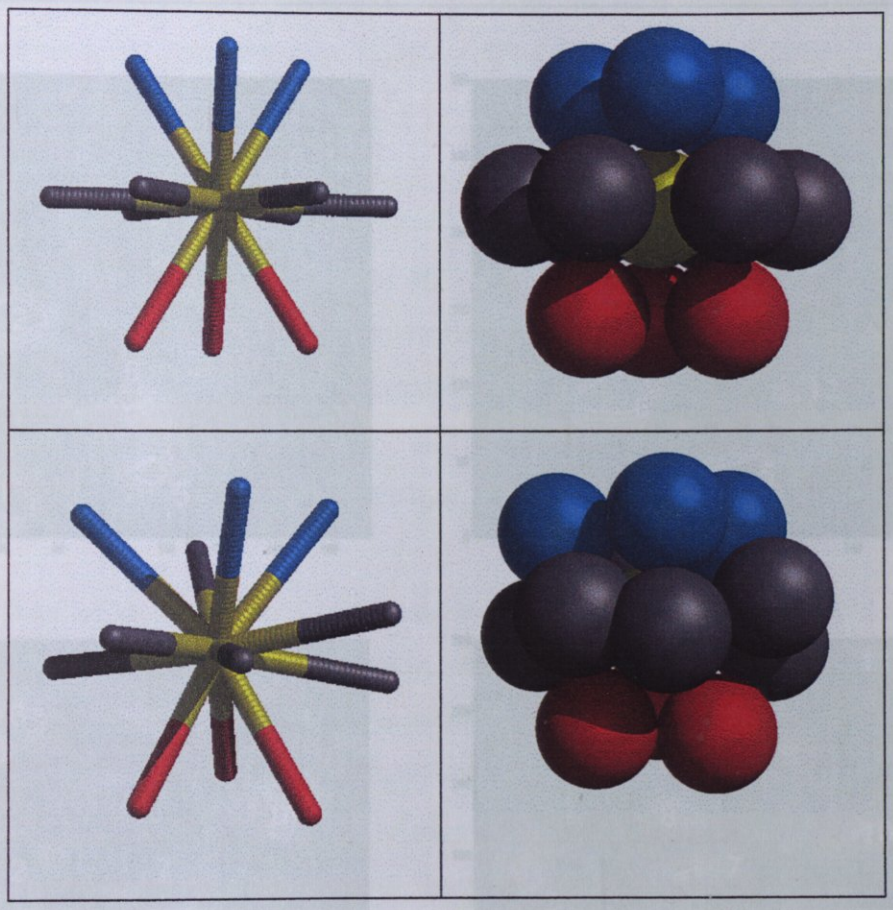

Figure 3.9. Directional vectors and the corresponding residue cluster models for the ideal fcc packing (top), and the densest core packing (bottom)

3.7. Generic Behavior: Incomplete, Distorted Fcc Geometry

Table 3.4 provides a summary of the correspondence between the coordination directions identified for all residues ($3 \leq m \leq 14$), core residues ($m \geq 10$), surface residues ($m \leq 4$) and the most densely packed clusters ($m \geq 12$), on the one hand, and the

coordination directions of the fcc geometry, on the other. The fraction of residues occupying these coordination directions, and the enhancement in the selection of these sites, relative to random distribution, are listed in the last column.

Figure 3.10 also provides a summary of the results obtained by optimal superimposition of clusters of different coordination numbers:

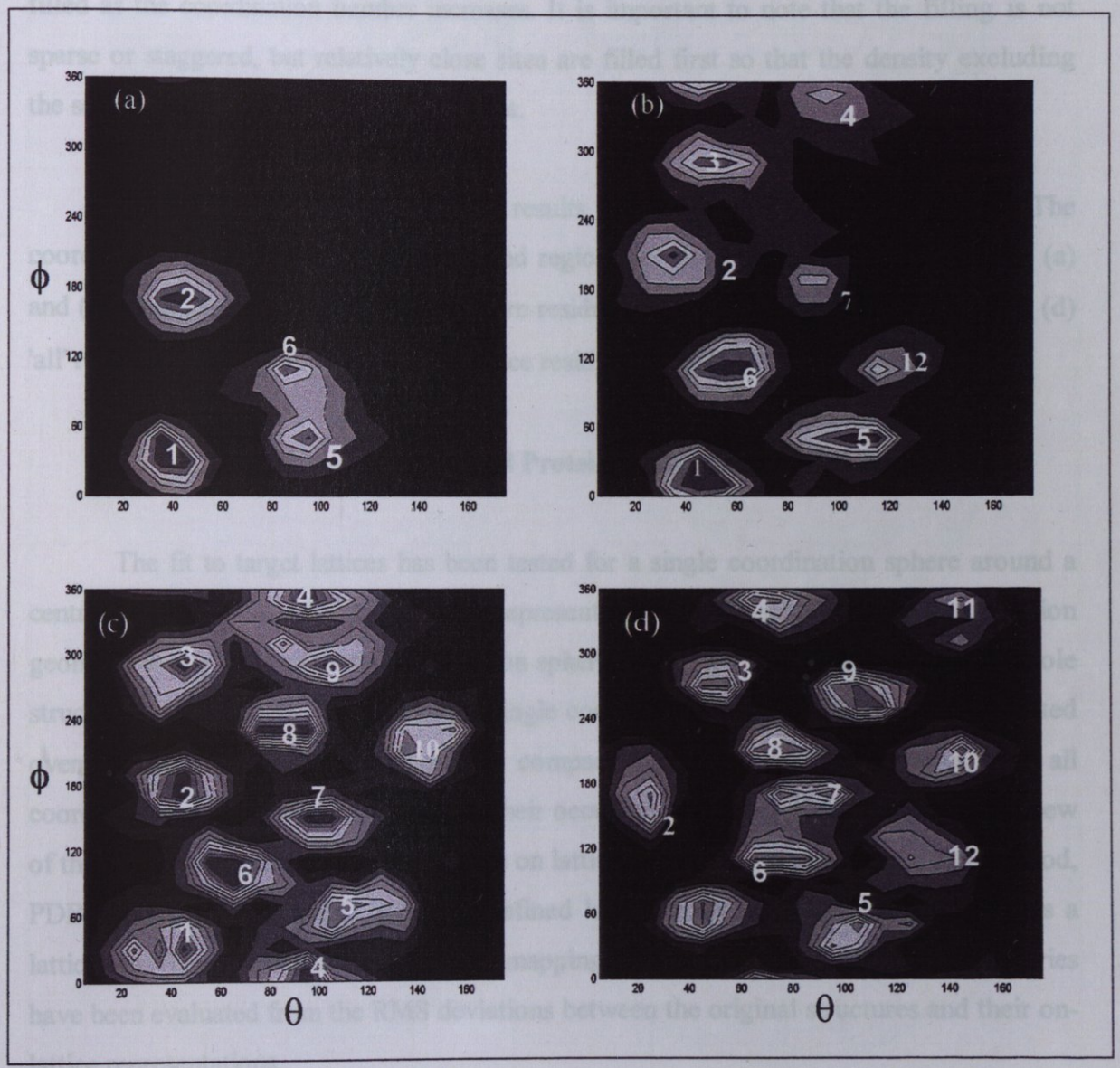

Figure 3.10. Coordination geometry around (a) surface residues ($m \leq 4$), (b) all residues ($3 \leq m \leq 14$), (c) core residues ($m \geq 10$), (d) densest core residues ($m \geq 12$)

In this figure, part (a) displays the most probable coordination sites for surface residues ($m \leq 4$). The sites are assigned numbers consistent with those in Table 3.4. Part (b) shows the results for all residues, irrespective of their coordination number, obtained by the rigid body rotation of the map in part (a) in Figure 3.5. Here, the additional site that has been observed for a subset of specific residues is included for visualization purposes (denoted with label 7). Parts (c) and (d) describe the core residues that have coordination numbers $m \geq 10$ and $m \geq 12$, respectively. As can be seen in this figure, sites are gradually filled as the coordination number increases. It is important to note that the filling is not sparse or staggered, but relatively close sites are filled first so that the density excluding the solvent space is approximately constant.

It is also possible to present these results with a different kind of illustration. The coordination sphere and its most populated regions are presented in Figure 3.11. Part (a) and (b) show the coordination sites for core residues ($m \geq 10$) with two views, (c) and (d) 'all' residues ($3 \leq m \leq 14$), (e) and (f) surface residues ($m \leq 4$).

3.8. Threading of Folded Proteins onto Several Lattices

The fit to target lattices has been tested for a single coordination sphere around a central residue. Although a given lattice representation can adequately fit the coordination geometry at the level of single coordination sphere, it may be less adequate when a whole structure is being fitted. This is because single coordination geometry may not be repeated over the entire space. Furthermore, in compact structures such as proteins, not all coordination sites are accessible, due to their occupancy by other chain segments. In view of these limitations, threading calculations on lattices have been performed. In this method, PDB structures are threaded onto a predefined lattice, such that each residue occupies a lattice site. The level of accuracy of the mappings to the five different lattice geometries have been evaluated from the RMS deviations between the original structures and their on-lattice representations.

Figure 3.12 illustrates the results for two example proteins. Part (a) displays the X-ray structure (gray) and the best fitting (fcc) lattice representation (black) for an α-protein, myoglobin (PDB code: 1mba). Part (b) displays the X-ray (gray) and lattice (black)

structures for a β-protein, plastocyanin (PDB code: 1plc). The respective RMS deviations between the X-ray and lattice structures are 2.04 Å and 2.29 Å.

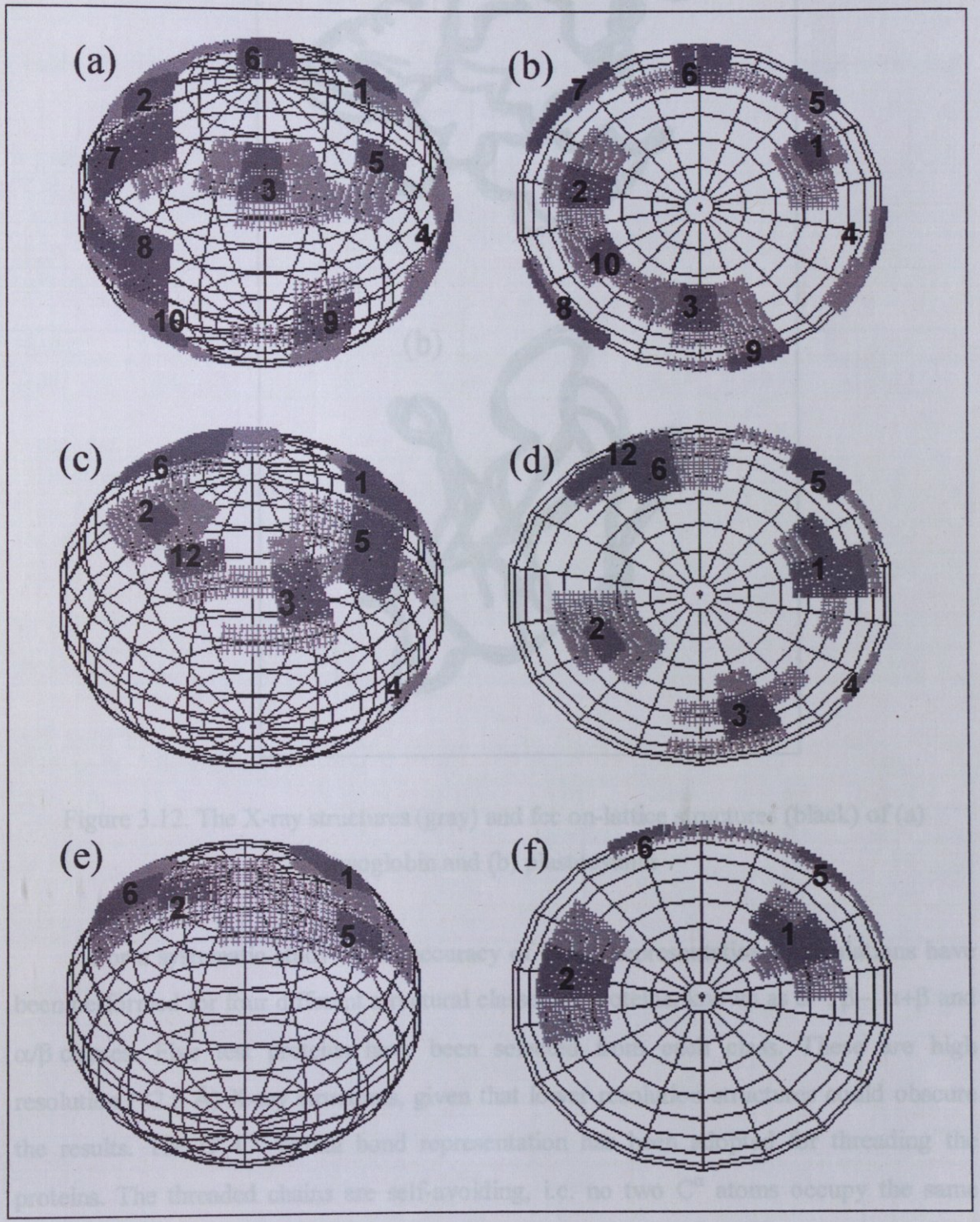

Figure 3.11. Coordination geometry around (a) and (b) core, (c) and (d) 'all', (e) and (f) surface residues

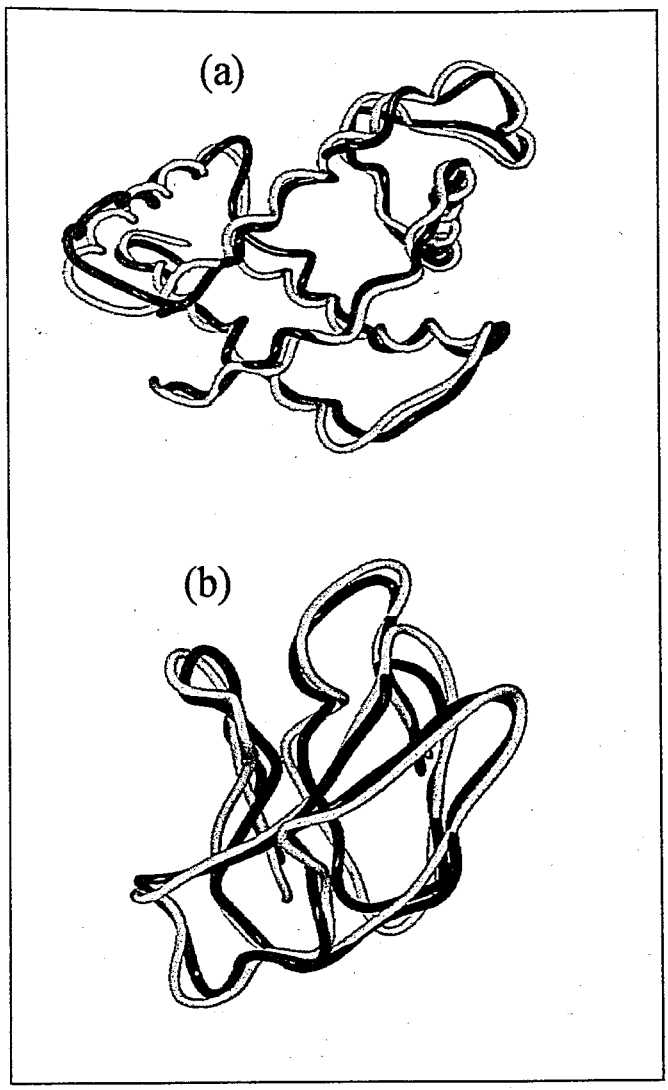

Figure 3.12. The X-ray structures (gray) and fcc on-lattice structures (black) of (a) myoglobin and (b) plastocyanin

For a systematic study of the accuracy of lattice representations, calculations have been performed for four different structural classes of proteins, known as α–, β–, α+β and α/β classes. Five test proteins have been selected from each class. These are high resolution (<2.5 Å) X-ray structures, given that lower resolution structures could obscure the results. The C^{α}-C^{α} virtual bond representation has been adopted for threading the proteins. The threaded chains are self-avoiding, i.e. no two C^{α} atoms occupy the same lattice site. The RMS deviations between the original positions of the α–carbons and their approximate on-lattice positions are listed in Table 3.5. The values given in parentheses are taken from the study of Godzik *et al.* (1993). The reported values are the deviations in Å,

the lattice edge being taken as 3.8 Å (i.e. the length of C^{α}-C^{α} virtual bonds) for the sc, bcc, fcc and hcp lattices. In the emb lattice, there are two lattice lengths, given by 1 and $3^{1/2}/2$, which were each multiplied by 3.8 Å.

Table 3.5. Threading results (RMS deviations in units of Å) for PDB structures belonging to four different structural classes

α-proteins

PDB code	Resolution (Å)	size	sc	bcc	emb	fcc	hcp
1AVHA	2.3	318	3.02	2.54	2.52	2.00	2.03
1LE4	2.5	139	2.81	2.25	2.21	1.96	1.93
1MBA	1.6	146	2.77 (2.7)	2.44 (2.4)	2.40	2.04 (1.9)	2.11
1MBC	1.5	153	2.82	2.44	2.37	1.97	2.06
2LH1	2.0	153	2.93	2.47	2.41	2.07	2.07

β-proteins

PDB code	Resolution (Å)	size	sc	bcc	emb	fcc	hcp
1HILA	2.0	217	5.07	3.51	2.51	2.28	2.48
1MAMH	2.45	217	5.01	3.46	2.79	2.31	2.54
1PLC	1.33	99	4.31 (4.3)	3.99 (2.9)	2.60	2.29 (3.3)	2.20
2AYH	1.6	214	4.35	4.05	2.83	2.62	2.27
8FABA	1.8	206	5.63	3.91	2.95	2.22	2.49

α+β proteins

PDB code	Resolution (Å)	size	sc	bcc	emb	fcc	hcp
1DNKA	2.3	250	3.48	4.02	2.68	2.47	2.44
1PPN	1.6	212	3.82	2.79	2.53	2.16	2.19
2AAK	2.4	150	3.39	3.73	2.44	2.46	1.96
2ACT	1.7	218	3.90	3.01	2.76	2.24	2.34
4BLMA	2.0	256	3.99	2.78	2.38	2.19	2.14

α/β-proteins

PDB code	Resolution (Å)	size	sc	bcc	emb	fcc	hcp
1DHR	2.3	236	3.43	3.23	2.78	2.37	2.16
1FX1	2.0	147	3.39	3.73	2.49	2.37	2.36
1OFV	1.7	169	3.60	3.07	2.42	2.15	2.25
2DRI	1.6	271	3.73	3.08	2.67	2.18	2.15
2YPIA	2.5	247	3.85 (3.3)	2.60 (2.8)	2.71	2.20 (2.0)	2.21

A summary of the results from threading calculations is presented in Table 3.6. Lattices having higher coordination numbers yield closer representations of the original structure. But these results are also associated with the details in the lattices and different structural classes of proteins. β–proteins and those belonging to α+β and α/β classes are harder, in general, to be threaded onto sc and bcc lattices compared to α-proteins. This deficiency can be partly attributed to the C^{α}-C^{α} virtual bond angle of 120° in β-strands, as opposed to its value of 90° in α–helices (Bahar *et al.*, 1997). The former can be readily accommodated by the fcc and hcp lattice cells; whereas the sc and bcc cells do not comply with 120° bond angles, hence the relatively high RMS deviations observed for β, α+β and α/β proteins in the sc and bcc cases (Table 3.6). Notice that the highest RMS deviations take place in the case of β-proteins threaded onto sc lattices, which can be understood from the fact that the sc lattice does not contain any coordination angle other than 90°.

Table 3.6. Average RMS deviations (Å) between databank structures and their lattice models

Class	sc	bcc	emb	fcc	hcp
α	2.87	2.43	2.38	2.00	2.04
β	4.87	3.78	2.74	2.34	2.40
α+β	3.72	3.27	2.56	2.30	2.21
α/β	3.60	3.14	2.61	2.25	2.23

4. CONCLUSION AND RECOMMENDATIONS

Packing architecture of residues in folded proteins can be modeled with a variety of regular geometries with approximately equal fidelity. Not surprisingly, the fcc lattice directional vectors were pointed out in a previous study (Raghunathan and Jernigan, 1997) to closely match the packing geometry, irrespective of amino acid type. However, in the present work, an alternative geometry with an equal coordination number (hcp) is shown to yield the same level of accuracy, when all clusters $3 \leq m \leq 14$ are considered.

Different coordination geometries were obtained, in the present thesis, for surface, core and completely buried residues, when the clusters are optimally superimposed irrespective of any predefined target representation. Comparison of the coordination maps with those of regular lattices reveals that the optimal architecture is a distorted, incomplete fcc packing. A cubic closest packed geometry may thus be viewed as a generic characteristic of the residue packing architecture in protein interiors.

Another important conclusion is that residues are packed closely and uniformly at all structural regions. Even for the clusters having relatively low coordination numbers, the coordination sites are closely clustered in space, i.e. the coordinating residues do not fill sparsely the coordinating sphere in the neighborhood of the central residues, but are closely grouped to occupy sites approximating those of a fcc packing. Therefore the uniform (high) densities of residues are maintained even at solvent-exposed regions, with the only difference that not all sites are occupied. Alternatively stated, the number of coordinating residues are different for surface and core residues. However, when only the space allocated to residues is considered, the density throughout the proteins is uniform. Interestingly, the same feature has been pointed out by Tsai *et al.* (1999) based on a different knowledge-based study. The compact – condensed matter - nature of proteins is also revealed by a Voronoi tesselation study (Soyer *et al.*, 2000).

The recently observed fcc coordination geometry does not preclude the complementarity in packing as suggested by Richards, 1977. It has been stated (Harpaz *et al.*, 1994) that packing density in proteins can even exceed that (0.7405) of fcc geometry

which has been shown (Cipra, 1998; Sloane, 1998) to be the closest packing geometry for identical spheres. It is possible to exceed this upper limit for packing of identical spheres by considering particles with respective radii in the ratio of 1.00:0.414, for example, which leads to a packing density of 0.7931 on a fcc lattice. This suggests that the size and shape heterogeneity of amino acids is well-suited for maximizing packing density. The intrinsic tendency of residues (when examined at the coarse-grained level of single-site-per-residue) to assume fcc packing architecture, can originate from the drive for maximizing packing density.

Despite the observed intrinsic regularity, the random positioning of about $1/3^{rd}$ of residues (on the average) could be selected to optimize the bonded and non-bonded interactions in a given irregular context, hence the adaptability of tertiary structures to single-site mutations. And the $2/3^{rd}$ of residues are packed in conformity with a well-defined architecture.

Core residues being more severely constrained cannot select the well-defined coordination states as efficiently as other residues. The seven coordination states identified for all residues in folded proteins are favored, on the average, by a factor of 3 compared to that implied by a random packing. The fraction of residues assuming these coordination states is indeed 0.63, as opposed to the value of 0.21 expected from random packing. Core residues, on the other hand, exhibit a relatively weaker freedom for selecting the well-defined coordination loci. The total frequency of occupancy of the twelve well-defined coordination loci is 0.76 for clusters having $m \geq 12$. This number exceeds the random probability (0.36) of the 12 coordination sites by a factor of about two, only.

In view of the present findings, the surface cannot be distinguished from the core by a lower packing density but rather by a flexibility (or ductility) evidenced by its high fraction (0.60) of residues occupying random positions. The solid-like core, on the other hand, can be understood in the view of the staggered closest packed distribution of residues. More than $2/3^{rd}$ of residues occupy these well-defined coordination sites. Solid-like versus liquid-like nature of the protein interior and exterior was suggested by Fraunfelder *et al.* (1979) and shown to be consistent with the Lindemann criterion by Karplus and coworkers (Zhou *et al.*, 1999). The Lindemann criterion is related to a

'disorder' parameter. The disorder parameter or the Lindemann parameter is the ratio of the rms atomic fluctuation to lattice constant *a* of a crystal. If this ratio reaches a certain value, fluctuations cannot increase without damaging or destroying the crystal lattice. Thus, this critical value is an indicator of the substance's being solid or liquid.

Packing architecture exhibits weak residue specificity, consistent with the faster divergence compared to structure. This is inferred from the similar coordination geometries observed for specific residues. This result conforms to the fact that during protein evolution, sequences diverge faster than structure. The same coordination architecture can thus accommodate different types of residues. Alternatively, mutations that drive structural changes could be associated with the differentiation of the coordination geometries around twenty amino acids. Behe *et al.* (1991) stated that packing does not determine the native fold. The weak specificity presently observed can indeed imply that packing does not determine the unique structure for a given sequence. On the other hand, knowledge of the generic packing architecture of residue clusters in folded structures, and its correlation with secondary structure, might provide important guidance in reducing the space for conformational search and in the computational predictions of 3D structure.

The regular packing geometry traced here at a coarse-grained scale is likely to be the result of an evolutionary imposed preference, for tolerating mutations while optimizing residue packing. The emergence of helical motifs was indeed shown (Michelletti *et al.*, 1999) to be the result of evolutionary pressure for selecting structures having a high degree of thermodynamic stability, which can accommodate amino acid sequences that fold reproducibly and rapidly. Helices satisfy such optimal packing constraints (Maritan *et al.*, 2000).

Examination of coordination geometry can be extended to other studies. It has been pointed that by employing bioinformatics tools, secondary structures of proteins can be predicted up to 80 per cent accuracy. In order to predict the folded structure, it would be useful to consider the packing of these secondary structures. This can be achieved by examining the coordination geometry around residues in α–helices only or β–strands only, as an extension of the present work.

The optimal superimposition method can be modified to examine folded structures of proteins in more detail. For example the identity of neighbors can be distinguished in the superimposition algorithm. The algorithm can be constrained to favor superimposition of similar neighbors. Thus, the coordination geometry will not only point the sites that are populated but also point which sites are populated by which type of neighbors. The identity of neighbors can be distinguished by their being hydrophobic or hydrophilic; small or large; α–helix, β-strand or coil forming.

REFERENCES

Bahar, I. and R. L. Jernigan, 1996, "Coordination Geometry of Nonbonded Residues in Globular Proteins" *Folding & Design,* Vol. 1, pp. 357-370.

Bahar, I., M. Kaplan and R. L. Jernigan, 1997 "Short-range Conformational Energies, Secondary Structure Propensities, and Recognition of Correct Sequence-Structure Matches", *Proteins:Structure, Function, and Genetics,* Vol. 29, pp. 292-308.

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

Behe, M. J., E. E. Lattman and G. D. Rose, 1991, "The Protein Folding Problem: The Native Fold Determines Packing, But does Packing Determine the Native Fold?", *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 88, pp. 4195-4199.

Branden, C. and J. Tooze, 1999, *Introduction to Protein Structure,* Garland Publishing, New York.

Bromberg, S. and K. A. Dill, 1994, "Side-chain Entropy and Packing in Proteins", *Protein Science*, Vol. 3, pp. 997-1009.

Chan, H. S. and K. A. Dill, 1990, "Origins of Structure in Globular Proteins", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 87, pp. 6388-6392.

Cipra, B., 1998, "Mathematics: Packing Challenge Mastered At Last", *Science,* Vol. 281, pp. 1267.

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

Creighton, T. E., 1993, *Proteins Structures and Molecular Properties*, W. H. Freeman and Company, New York.

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

Demirel, M. C., A. R. Atilgan, R. L. Jernigan, B. Erman, and I. Bahar, 1998, "Identification of Kinetically Hot Residues in Proteins", *Protein Science*, Vol. 7, pp. 2522-2532.

DeWitte, R. S. and E. I. Shaknovich, 1994, "Pseudodihedrals: Simplified Protein Backbone Representation with Knowledge-based Energy", *Protein Science,* Vol. 3, pp. 1570-1581.

Frauenfelder, H., G. A. Petsko and D. Tsernoglou, 1979, "Temperature-dependent X-ray diffraction as a Probe of Protein Structural Dynamics", *Nature*, Vol. 280, pp. 558-563.

Godzik, A., A. Kolinski and J. Skolnick, 1993, "Lattice Representations of Globular Proteins: How Good are They?", *Journal of Computational Chemistry*, Vol. 14, pp. 1194-1202.

Gregoret, L. M. and F. E. Cohen, 1991, "Effect of Packing Density on Chain Conformation", *Journal of Molecular Biology*, Vol. 219, pp. 109-122.

Hao, M. H., S. Rackovsky, A. Liwo, M. R. Pincus and H. A. Scheraga, 1992, "Effects of Compact Volume and Chain Stiffness on the Conformations of Native Proteins", *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 89, pp. 6614-6618.

Harpaz, Y., M. Gerstein and C. Chothia, 1994, "Volume Changes on Protein Folding", *Structure*, Vol. 2, pp. 641-649.

Hunt, N. G., L. M. Gregoret and F. E. Cohen, 1994, "The Origins of Protein Structure. Effects of Packing Density and Hydrogen Bonding studied by a Fast Conformational Search", *Journal of Molecular Biology,* Vol. 241, pp. 214-225.

Jernigan, R. L. and I. Bahar,, 1996, "Structure-derived Potentials and Protein Simulations", *Current Opinion in Structural Biology,* Vol. 6, pp. 195-209.

Jernigan, R. L. and I. Bahar, 1999, "Geometric Regularities Among Bonded and Non-bonded Residues in Proteins", in M. Vijayan, N. Yathindra and A. S. Kolaskar (Eds.), *Perspectives in Structural Biology (GN Ramachandran Volume)*, pp. 209-225, Indian Academy of Sciences, Universities Press, Hyderabad.

Keskin, O. and I. Bahar, 1998, "Packing of Sidechains in Low-resolution Models for Proteins", *Folding & Design*, Vol. 3, pp. 469-479.

Lim, W. A. and R. T. Sauer, 1991, "The Role of Internal Packing Interactions in Determining the Structure and Stability of a Protein", *Journal of Molecular Biology,* Vol. 219, pp. 359-376.

Maritan, A., C. Micheletti, A. Trovato and J. R. Banavar, 2000, "Optimal Shapes of Compact Strings", *Nature*, Vol. 406, pp. 287-290.

Micheletti, C., J. R. Banavar, A. Maritan and F. Seno, 1999, "Protein Structures and Optimal Folding from a Geometrical Variational Principle", *Physical Review Letters*, Vol. 82, pp. 3372-3375.

Munson, M., S. Balasubramanian, K. G. Fleming, A. D. Nagi, R. O'Brien and J. M. Sturtevant, 1996, "What Makes a Protein a Protein? Hydrophobic Core Designs That Specify Stability and Structural Properties", *Protein Science,* Vol. 5, pp. 1584-1593.

Oldfield, T. J. and R. E. Hubbard, 1994, "Analysis of Cα Geometry in Protein Structures", *Proteins: Structure, Function and Genetics,* Vol. 18, pp. 324-337.

Pal, D. and P. Chakrabarti, 1999, "Graphical Representation of the Salient Conformational Features of Protein Residues", *Protein Engineering,* Vol. 12, pp. 523-526.

Petersen, T. N., C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr and S. Brunak, 2000, "Prediction of Protein Secondary Structure at 80% Accuracy", *Proteins: Structure, Function, and Genetics*, Vol. 41, pp. 17-20.

Petrucci, R. H., 1985, *General Chemistry*, Macmillan Publishing Company, New York.

Ponder, J. W. and F. M. Richards, 1987, "Tertiary Templates for Proteins. Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes", *Journal of Molecular Biology*, Vol. 193, pp. 775-791.

Raghunathan, G. and R. L. Jernigan, 1997, "Ideal Architecture of Residue Packing and Its Observation in Protein Structures", *Protein Science*, Vol. 6, pp. 2072-2083.

Ramachandran, G. N., C. Ramakrishnan and V. Sasisekharan, 1963, "Stereochemistry of Polypeptide Chain Configurations", *Journal of Molecular Biology*, Vol. 7, pp. 95-99.

Richards, F. M., 1977, "Areas, Volumes, Packing, and Protein Structure", *Annual Reviews in Biophysics and Bioengineering*, Vol. 6, pp. 151-176.

Richards, F. M. and W. A. Lim, 1994, "An Analysis of Packing in the Protein Folding Problem", *Quarterly Reviews of Biophysics*, Vol. 26, pp. 423-498.

Sloane, N. J. A., 1998, "Kepler's Conjecture Confirmed", *Nature*, Vol. 395, pp. 435-436.

Soyer, A., J. Chomilier, J. P. Mornon, R. Jullien and J. F. Sadoc, 2000, "Voronoi Tessellation Reveals the Condensed Matter Character of Folded Proteins", *Physical Review Letters*, Vol. 85, pp. 3532-3535.

Stasiak, A. and J. H. Maddocks, 2000, "Mathematics: Best Packing in Proteins and DNA", *Nature,* Vol. 406, pp. 251-253.

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

Tsai, J., R. Taylor, C. Chothia and M. Gerstein, 1999, "The Packing Density in Proteins: Standard Radii and Volumes", *Journal of Molecular Biology,* Vol. 290, pp. 253-266.

Van Vlack, L. H., 1989, *Elements of Materials Science and Engineering*, Addison-Wesley Publishing Company, Inc., New York.

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

Watson, H. C., 1969, "The Stereochemistry of the Protein Myoglobin", *Progress in Stereochemistry,* Vol. 4, 299-304.

Yee, D. P., H. S. Chan, T. F. Havel and K. A. Dill, 1994, "Does Compactness Induce Secondary Structure in Proteins? A Study of Poly-alanine Chains Computed by Distance Geometry" *Journal of Molecular Biology,* Vol. 241, pp. 557-573.

Zhou, Y., D. Vitkup and M. Karplus, 1998, "Native Proteins are Surface-molten Solids: Application of the Lindemann Criterion for the Solid versus Liquid State", *Journal of Molecular Biology*, Vol. 285, pp. 1371-1375.