*Protein Science* (1997), 6:2072–2083. Cambridge University Press. Printed in the USA.


# Ideal architecture of residue packing and its observation in protein structures

G. RAGHUNATHAN[1] AND R.L. JERNIGAN

Molecular Structure Section, Laboratory of Experimental and Computational Biology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, MSC 5677, 12 South Drive, Bethesda, Maryland 20892-5677

(RECEIVED May 7, 1997; ACCEPTED June 20, 1997)

**Abstract**

A simple model of sphere packing has been investigated as an ideal model for long-range interactions for the packing of non-bonded residues in protein structures. By superposing all residues, the geometry of packing around a central residue is investigated. It is found that all residues conform almost perfectly to this lattice model for sphere packing when a radius of 6.5 Å is used to define non-bonded (virtual) interacting residues. Side-chain positions with respect to sequential backbone segments are relatively regular as well. This lattice can readily be used in conformation simulations to reduce the conformational space.

**Keywords:** equivalent sphere packing model; ideal packing; long-range interactions; packing geometry, protein lattices; residue packing

A particular protein fold is stabilized by a delicate balance of forces originating in the different classes of interactions. In one classification scheme, there are local and non-local forces. Local forces involve those interactions among atoms or residues that are sequentially adjacent, and one manifestation of these forces in protein structures is, for example, the presence of $\alpha$-helices, where the conformations of individual amino acids are stabilized by local intra-chain hydrogen bonds. Significant progress has been made in characterizing these local interactions since the initial work on the permissibility of allowed conformations for peptides by Ramachandran et al. (1963). Among many other studies, we have recently related the empirical occurrences of conformations to secondary structure propensities (Bahar et al., 1997). On the other hand, non-local interactions are more difficult to treat because they originate substantially in the solvent interactions. Overall, these result in a segregation in which the charged amino acids are usually placed on the exterior and the non-polar amino acids most often are in the interior of the protein due to hydrophobic forces (Kauzmann, 1959). However, these hydrophobic interactions are not so specific, with the magnitude of their non-residue-specific component being approximately five times as large as their residue-type-specific component (Bahar & Jernigan, 1996a). This relatively uniform, strong interaction between hydrophobic residues raises the issue of whether this could also imply some regular, uniform positioning of residues in response. A regularity in packing might imply some regular directions among groups of close, non-bonded residues; however, previously hydrophobic interactions have been considered to be relatively non-specific (Miyazawa & Jernigan, 1985, 1996; Behe et al., 1991; Bahar & Jernigan, 1996a). But, as we will show below, non-specific interactions are not necessarily without directionality. One of the aims of the present paper is to present observations on the geometries of residue packing in globular proteins. We find that this packing is highly regular at the coarse-grained level of one point per residue and conforms to a specific type of regular lattice that closely approximates the ideal case of high density sphere packing. This is true despite the competition in the balance among the various forces, such as that arising from the constraints in sequence, which would prevent a completely independent placement of residues. This regularity in packing that we will demonstrate can, in a sense, be viewed as a manifestation of hydrophobic forces.

Residue packing in the protein interior has long been considered to be essential to the native-like character, stability, and function of proteins (Richards, 1974; Richards & Lim, 1993). If it were better understood, it could be useful for designing mutant proteins with altered properties. Among the many studies of mutant structures, an interesting one was reported by Eriksson et al. (1992) who obtained crystal structures of six "cavity-producing" mutants in the core of T4 lysozyme and reported their decreased stabilities. However, there are many cases where repacking is able to accommodate with ease small changes to the atom population. An example of this is the changes to staphylococcal nuclease, which has cav-

Reprint requests to: Robert L. Jernigan, MSC 5677, Room B-116, Building 12-B, National Institutes of Health, Bethesda, Maryland 20892-5677; e-mail: jernigan@lmmb.nci.nih.gov.

[1]Current address: Structural Bioinformatics, 10929 Technology Place, San Diego, California 92127.

ities in the wild type, where it was shown that cysteines can be substituted with unnatural amino acids having rotatable alkyl groups in the side chains (Wynn et al., 1996) without significant perturbation to the protein. In another surprising case (Baldwin & Matthews, 1994), mutations in the interior residues of T4 lysozyme caused large motions of the helices but only minor variations in the side-chain torsion angles.

In addition to stability and structural changes there is the issue of a protein's function. Lim et al. (1994) showed that larger substitutions can be accommodated in the core of lambda repressor. However, these mutations cause large shifts in $\alpha$-helix positions and a reduced DNA binding. Mutants of varying sizes have been designed to alter the stabilities of the four-helix bundle protein Rop (Munson et al., 1996). There is no clear basis for suggesting which mutants would lead to better packing without introducing significant large scale structural changes that could affect its function. Because of the wide variety of effects observed, of which only a few examples have been mentioned, a full comprehension of the range of packing effects may be difficult from experiment alone.

### *Model for ideal long-range interactions*

One model that has been developed and has proven to be useful to treat local interactions in flexible chains, not globular proteins, is the simple flexibility of equivalent freely jointed chains. In this case, a sequential segment of the real chain is chosen to encompass sufficient atoms to be defined as an ideal bond or freely jointed link so that it actually behaves effectively as if it were a freely jointed link. Here, in a similar way, it is possible to define a simple but useful model for long-range interactions chosen by enlarging the unit of consideration by increasing the number of atoms included until the ideal, model behavior is obtained. For treating long-range interactions, we define an equivalent volume that is large enough to include sufficient atoms so that the residues themselves become ideal in the geometries of their interactions. The ideal model pursued here for long-range interactions is one in which the interacting species behave in their packings as if they were spheres; this is presumed to be the simplest case for packing. The present lattice was previously employed for treating polyethylene melts; in that case, two monomers were sufficiently large when taken as a single unit to achieve this effective sphere-packing behavior (Rapold & Mattice, 1996; Cho & Mattice, 1997).

Here we focus on residue packing rather than atom packing and report the distinctly directional ways in which amino acids pack together. The most critical element for observing a highly regular packing has been to look at residue packing in a coarse-grained way rather than at the details of atom packing. For all interacting groups surrounding a central residue, we re-orient the packing unit to obtain the best overall coincidence of positions for all cases. As we will see, the residue packing is then highly regular and leads somewhat surprisingly to discrete preferred relative positions for non-bonded residues around any interacting residue, with geometries following the rules of high-density sphere packing. Thus, protein residues show overall an accord with this model and are packed in a highly regular, lattice-like way. Understanding this packing is extremely important for simulating and sampling protein folds and could be applied in various simple ways to substantially reduce the conformational choices. For protein conformation simulations, this discrete view of the packing of interacting nonbonded residues reduces in significant ways the size of the conformational space requiring either enumeration or sampling.

## Results and discussion

### *Amino acid "coordination"—Directional behavior*

First, we look to see how regular the distribution of the number of non-bonded neighbors is about a central residue for residues of different types. This analysis is performed on a set of 161 standard proteins selected as described in Methods, with the corresponding $C^{\alpha}$ atom positions chosen to represent all amino acids. Cumulative frequency distributions of the numbers of non-bonded residues approaching a central residue within 6.5 Å are shown in Figure 1. The curve including all residue types (Fig. 1, upper left) has a rather broad peak ranging from 4 to 6 residues but there is also a significant population with as many as 10 non-bonded neighbors. The peaks for most hydrophobic amino acids occur sharply near 6 non-bonded neighbors (Fig. 1, upper right). Smaller amino acids (Fig. 1, lower left) have somewhat fewer contacts. Charged amino acids (Fig. 1, lower right) have two distinct peaks, one at 4 and the other at 6. It is likely that the peak at 4 corresponds to a surface location on the protein exterior exposed to solvent and the peak at 6 corresponds to interior placement. The numbers of such contact values (Fig. 1) for all amino acid types and charged amino acids are slightly larger than those reported earlier (Miyazawa & Jernigan, 1996) because they used side-chain centers instead to represent amino acids. However, no change is observed in the distributions about hydrophobic amino acids, which are usually in the protein interior with a larger number of residue contacts. Hydrophilic amino acids, a majority of which are on the exterior and exposed to solvent, have in general fewer contacts. It is striking that all residues manifest only two peaks in these distributions at the same numbers of contacts. This already suggests some significant regularity in the packing behavior.

We previously showed a high specificity in the selection of ligand atoms and amino acids for various ions and in the directions of these interactions (Jernigan et al., 1994). Such specificities can be useful for protein and ligand design with altered stabilities and function. While this orientation specificity is likely to depend on the electronic state of the ion, part of the regularity could originate in simple packing. There have been several analyses (Singh & Thornton, 1992; Bahar & Jernigan, 1996b) of the relative orientation between spatially adjacent residues in terms of conventional internal coordinates, viz., virtual bond lengths, virtual angles, and virtual torsion angles connecting various non-bonded interaction sites, often taken as atoms. Usually these studies concern both sequentially connected and unconnected residues. Although such studies show preferred distributions of these atomic interactions, they do not show extreme regularities in geometry. Hence, we sought to explore instead the regularities in the overall directions of non-bonded residue interactions. Specifically, by considering residues at the coarse-grained level of one point per amino acid, we want to learn about the packing of amino acids and the extent of the regularity in the underlying directional behavior, akin to the directional specificities we observed for the distribution of ligands around ions (Jernigan et al., 1994). The exclusion of sequentially connected residues enhances their regularity.

Here, the neighbors surrounding any given amino acid are represented by vectors from the central amino acid. These 37,095 sets of vectors from the set of 161 diverse proteins are optimally superimposed (see Methods) and their spatial distribution over the two angles $\theta$ and $\phi$ of a spherical polar coordinate system are analyzed. These are shown in Figure 2 for two different choices of

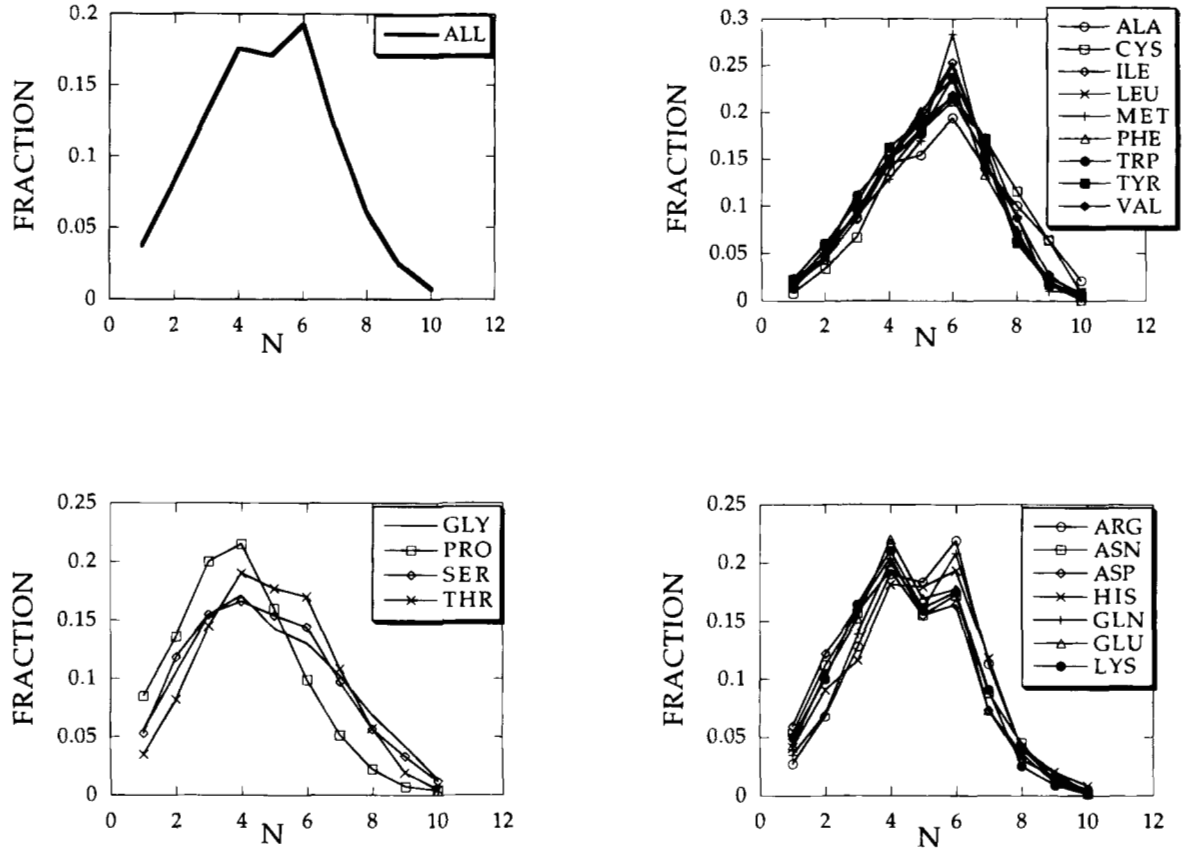


**Fig. 1.** Frequency distributions of the numbers of non-bonded contacts. Top left is the cumulative distribution for all 20 amino acid types. The top right panel comprises mostly hydrophobic amino acids. Amino acids in the bottom left panel are smaller in size. The bottom right is for charged and polar amino acids.

points. In two top parts of Figures 2, $C^{\alpha}$ atoms are used to represent amino acids and in lower parts of Figure 2, centers of each amino acid's side-chain atoms are used. Distributions of non-bonded residues (Fig. 2, two left parts) exhibit nine strong distinct peaks. There are six peaks at $\phi = 90°$. These are at regular intervals of 60° along $\theta$, at 0°, 60°, 120°, 180°, 240°, and 300°. There are three more peaks around $\phi = 35°$. These are near $\theta = 30°$, 150°, and 270°. Thus, these peaks are staggered with respect to the six peaks for $\phi = 90°$. The centers of the peaks are at identical positions for both the $C^{\alpha}$ and side-chain centers. However, the peaks are sharper for centers of side chains (Fig. 2, lower left) than those for $C^{\alpha}$ positions, as might be expected, since these are more specific interacting sites. The two figures on the right side (Fig. 2) include, in addition, the sequential neighbors. Figure 2, top right, for example includes non-bonded neighbors and three more points, the preceding and succeeding $C^{\alpha}$ atoms and the side chain of the central amino acid. Similarly, three centers are added in Figure 2, lower right. When these additional positions are used in the superposition, three added peaks appear near $\phi = 145°$. These occur at intervals of 120° with $\theta$ values near 90°, 210°, and 330°. These three peaks are thus staggered, both with respect to the six peaks around $\phi = 90°$ as well as the three peaks around $\phi = 35°$. Thus, all 12 peaks are staggered and well separated in space. Distributions of residues for all 20 different amino acid types display nearly identical patterns for their neighbors' positions (Fig. 3). The peaks for the polar amino acids are slightly sharper than those for the hydrophobic amino acids. Thus, at the coarse-grained level, the residues tend to behave extremely uniformly in their relative directions of approach regardless of the central residue type.

This reveals a general and extremely interesting feature of protein architecture that has important implications for protein folding and design. It illustrates clearly that residue packing can be treated in a general way by using these discrete points in space. Notably, the set of three peaks on the left and right sides of Figure 2 (two right-most parts) are staggered, both with respect to one another and also with respect to the six peaks along the $\phi = 90°$ line. Thus, all peaks are totally staggered. The staggered dispositions of these peaks are reminiscent of the staggered positions of high-density sphere packing. In the present case, the non-bonded and sequentially bonded neighbors tend to approach a central amino acid in highly specific directions, and we can attribute these specificities to the requirement for maximum spatial separation of residues. From Figure 2, it is apparent that non-bonded residues around a residue fill only one side of the three-dimensional space and the sequential neighbors complete the other part of the conformational space around a residue. This lattice can be used to reduce significantly the conformational space to be explored in protein folding studies compared to the continuum. It is interesting that the two peaks in the distribution of contacts (Fig. 1) for buried and surface would have exactly the same residue packing densities. For totally buried residues there are nine possible interaction sites so the peak

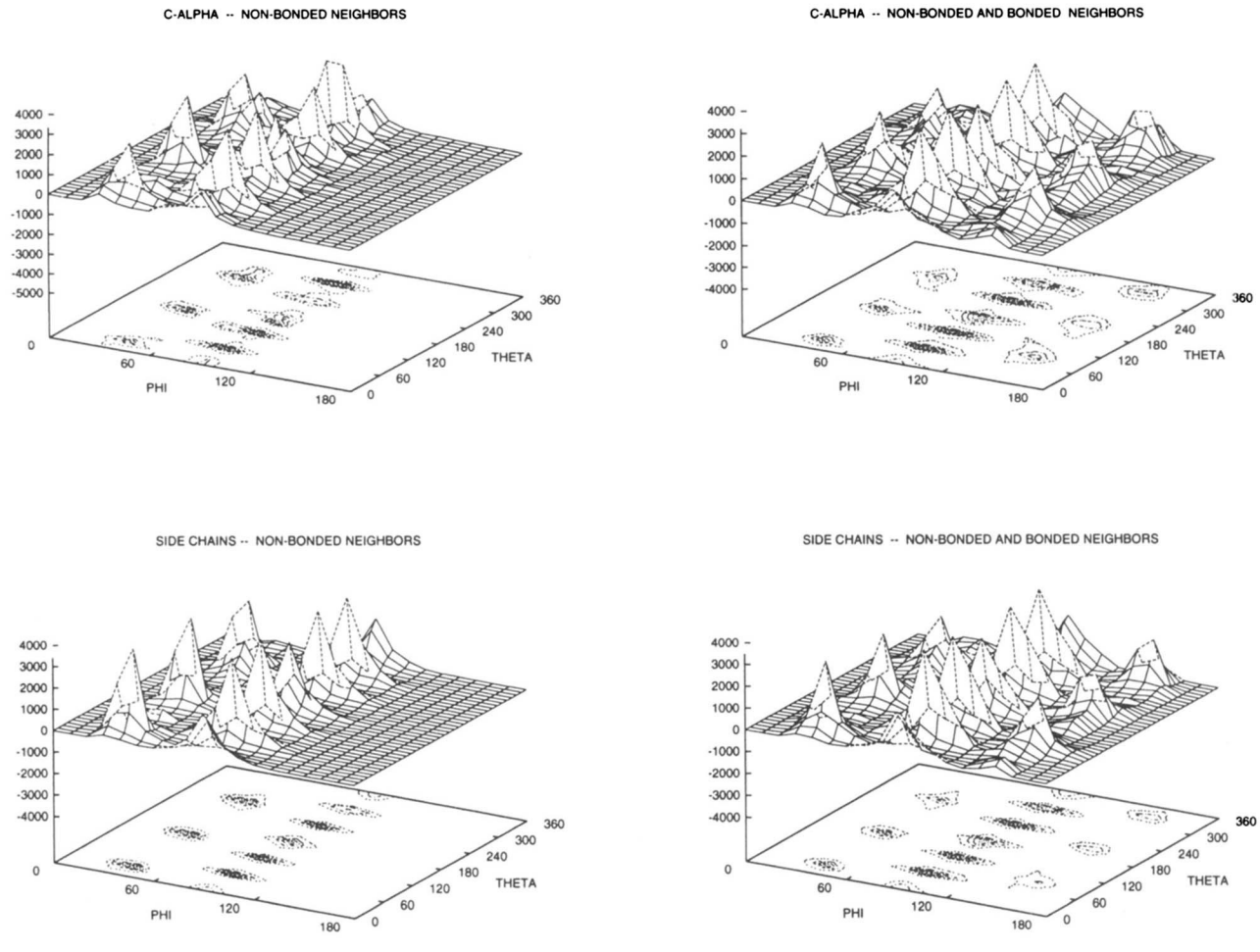


**Fig. 2.** The directions of neighbors surrounding a central amino acid are given in terms of the angles $\theta$ and $\phi$ in the polar coordinate system. Data in each of these four plots are comprised of a superposition of 37,095 sets of vectors from 161 diverse proteins. In the top two panels, $C^{\alpha}$ atoms are used to represent amino acids and, in the bottom two panels, centers of side chains are used. Directions of sequentially non-bonded residues (top left and bottom left) each have nine peaks. The six peaks around $\phi = 90°$ are all staggered at 60° intervals of $\theta$ and are at $\theta = 0°$, 60°, 120°, 180°, 240°, and 300°. The three peaks on the left around $\phi = 35°$ are at intervals of 120° at $\theta = 30°$, 150°, and 270°. All nine peaks are staggered. The peaks for side-chain positions (bottom) are sharper than those for $C^{\alpha}$ atoms (top). The figures on the top and bottom right include both bonded sequential neighbors and non-bonded neighbors. The top right figure contains, in addition, the side chain of the central amino acid. The position of the six peaks at the center of these figures and three peaks on the right are very similar to those of the figures on the left. The three new peaks on the right side of these figures are at $\phi = 145°$ and are around $\theta = 90°$, 210°, and 330°. All 12 peaks are staggered.

having six contacts corresponds to a residue density of exactly two-thirds. For surface residues, the simplest way to achieve this is to omit the three non-bonded sites above the hexagonal plane at $\phi = 90°$. This means that, for the other peak in the number distributions, four of the six states would be occupied or again a precise density of two-thirds. Thus, such an "ideal" protein would have a geodesic dome-like surface comprised of imperfect hexagonal plates.

*Knowledge-based lattice*

The present model of a protein is sufficiently coarse grained so that, to a good approximation, all residues pack identically as if they were of the same size and shape. It is this fuzzy regularity that permits us to view a protein as a homo-polymer placed on points of a regular lattice. Later we will also discuss how side chains can be placed upon this lattice. The order and symmetry in the spatial distribution of neighbors prompts us to look for underlying principles of protein structure and organization and to explore the feasibility of using this set of points for folding studies. In other words, we want to address the question: Can we arrive at a basic unit, which can represent the spatial arrangement of amino acids around a central amino acid? We realize that, indeed, we can fit the characteristic peaks in the observed distribution of neighbors around a central amino acid with discrete lattice points in three-dimensional space.

A model of the unit cell of this lattice is shown in Figure 4. It can be used directly for protein chain generation, since it conforms almost perfectly to the angular distributions of residues observed in Figures 2 and 3. Let us assume that each point in this lattice represents an amino acid. Each amino acid can have a maximum coordination of 12 sites. Of these, nine can be occupied by non-bonded neighbors and three are filled by the sequential neighbors and side chain of the central amino acid. This semi-regular solid unit is referred to as a cubo-octahedron and has 14 faces and 12

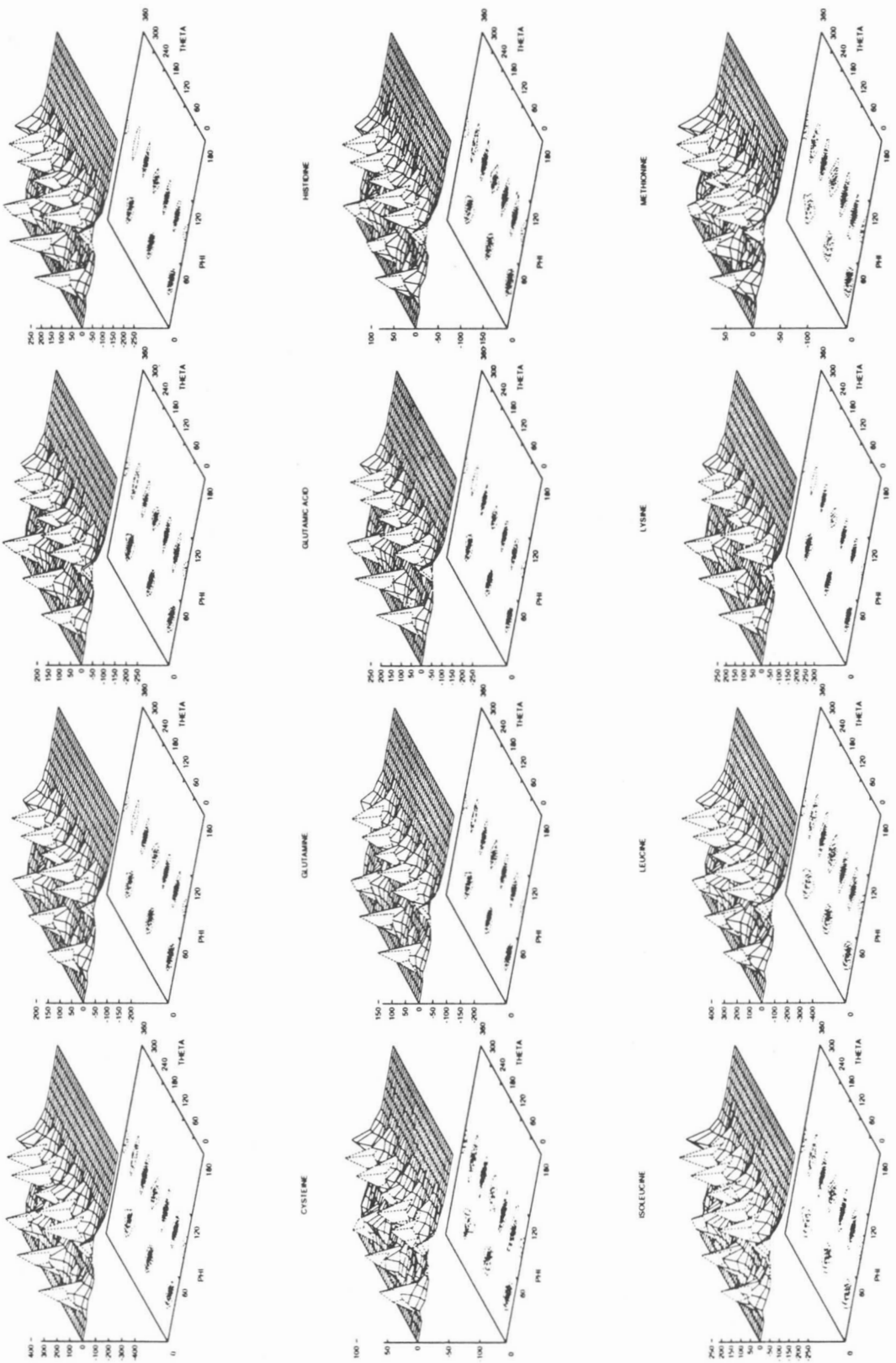


**Fig. 3.** *See caption on facing page.*

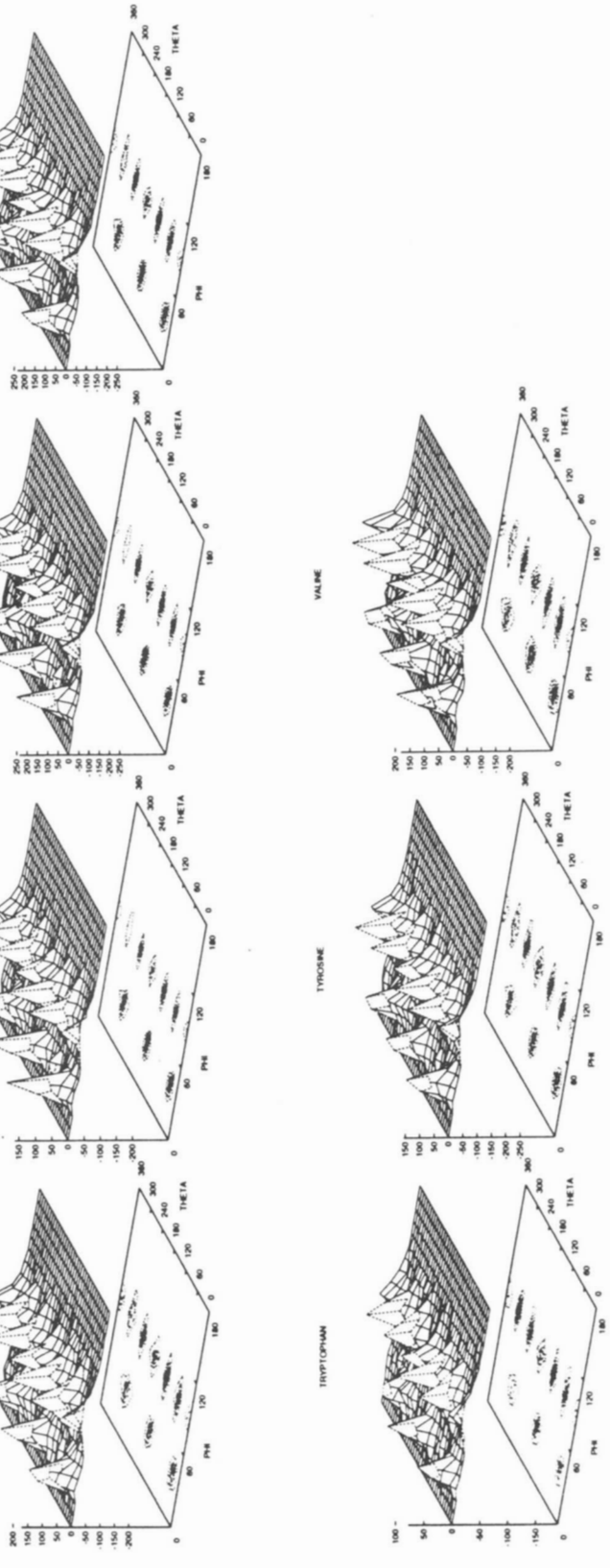


**Fig. 3.** Distribution of non-bonded neighbors for each of the 20 amino acid types. Side-chain centers are used as points to represent the amino acids. The plots are all quite similar and suggest a nearly identical regularity in the packing of amino acids at the coarse-grained level of one point per amino acid.

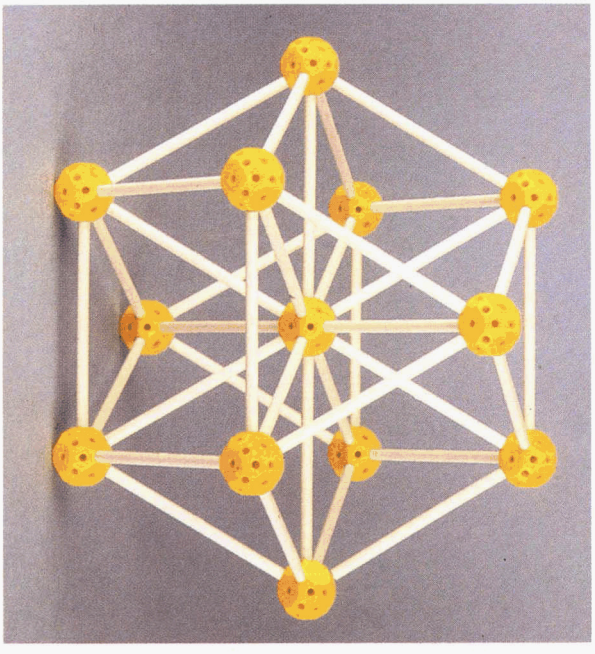

**Fig. 4.** A model illustrating the unit cell of the proposed lattice. Directions of the peaks in the distribution of neighbors around amino acids shown in Fig. 2 are used directly in constructing this model. Adjacent sticks connecting the six horizontal balls make an angle of 60° with the central ball. The top three balls are staggered with respect to the bottom three balls. The solid is made of alternating square and triangular faces and would be most ideal for uniformly filling space with densely packed spheres. Each triangle has three squares as neighbors and each square has four triangles as neighbors. There are six squares and eight triangles on its surface.

vertices. It is a unique convex polyhedron and contains only two kinds of regular polygons, namely triangles and squares. The 14 faces are composed of eight triangles and six squares (Wells, 1956). Each triangle is surrounded by three squares and each square by four triangles. There are six points in the horizontal plane, three above this plane and three below this plane. Adjacent vectors in the horizontal plane make an angle of 60° with the center. The top three vectors make angles of 120° with one another and likewise the bottom three vectors. The three points on the top and bottom are staggered with respect to one another and also staggered with respect to the six points in the horizontal plane. Thus, all 12 points are staggered and hence maximally separated from one another. The striking directional distribution of amino acids as fitted to these unique points can be the result of one of the best ways for uniform spheres to pack densely among themselves. This packing arrangement is different from hexagonal packing. In hexagonal packing, the three points in the top and bottom layers are oriented similarly so that each point in the top layer lies directly above a point in the bottom layer; whereas these two non-adjacent layers are staggered in the present case.

This lattice is a version of the face-centered cubic lattice that has connections permitted only between face-centered points and the 12 nearest corners. All connections therefore have the same length.

The 12 unit vectors (all requiring normalization by $\sqrt{2}$) are: in the *xy* plane, (1,1,0), (1,−1,0), (−1,1,0), (−1,−1,0); in the *yz* plane, (0,1,1), (0,1,−1), (0,−1,1), (0,−1,−1); and in the *xz* plane (1,0,1), (1,0,−1), (−1,0,1), (−1,0,−1).

When taken in connected pairs, these give the bond angles 0°, 60°, 90°, 120°, and 180°, and, in connected triples, the torsion angles 0°, 54.7°, 70.5°, 109.5°, 125.3°, and 180°. When expressed this way, the unit vectors are simpler to understand.

The set of points in Figure 4, when repeated, can form a periodic array of lattice points in space. An infinite number of layers can be grown from the center, radially outward. The number of points on the surface of any layer is given by the formula

$$N = 10 \cdot L^2 + 2$$

where $L$ is the layer number (Kappraff, 1991). Thus, the first layer has 12 points, the second 42, the third 92, and so on. Here we have constructed physical models to see how the lattice propagates in all directions in space and also to inspect visually tracings of protein chains prior to attempting any computer simulations. Figure 5 shows a view of the stacking of an array of triangular faces. These models not only are aesthetically appealing and illustrate the spatial propagation of the lattice but also offer one of the best ways to visually inspect protein chain tracings, side-chain stereochemistry in relation to backbone, secondary structure, and the other geometric details of protein structures.

### *Geometry of the lattice and possible conformations*

Let us assume for the moment, that each point in the lattice represents a $C^{\alpha}$ atom. Since the lattice grows radially outward, all bonds between adjacent $C^{\alpha}$ atoms have a length of 3.8 Å. This regular length differs from several versions of the cubic lattice that have different lengths between sequential neighbors. Bond angles of 60°, 90°, 120°, and 180° are possible in this lattice. For protein simulations, these angles translate into one to three distances (between a point and its third neighbor) of 3.8 Å , 5.4 Å, 6.7 Å, and

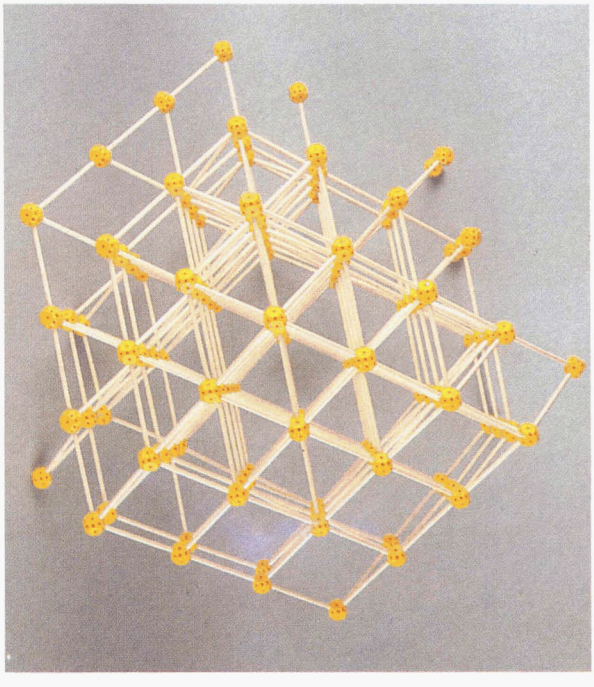

**Fig. 5.** A view of the lattice model illustrating the spatial propagation by highlighting the packing of triangular faces.

7.6 Å. Now, let us consider possible torsion angles on the lattice. Two adjacent unit cells are shown in Figure 6. There are four red balls, each making 60° with the vector formed by the two yellow balls linking the two units. There are four blue balls that are at 120° to this vector, two gray balls at 90° and one black ball at 180°.

A set of four sequentially connected points can be defined by two bond angles, which we denote as $\delta_1$ and $\delta_2$, and one dihedral angle, $\psi$. Let us consider the yellow ball at the far left of Figure 6 as the first atom, the middle yellow ball as the second atom, and the yellow ball at the far right as the third atom. The fourth atom can be placed in 11 possible positions. As we discussed before, the four red balls have a bond angle of 60°, the four blue balls have bond angles of 120°, the two gray balls have 90°, and the black ball at the extreme right has a bond angle of 180°. Although all sets of red, blue, and gray balls have the same bond angles, each of these same-color balls has a unique torsion angle with respect to the first three atoms. In a simulation, some of these points might be filled by non-bonded neighbors but the rest would be available for growing the protein chain. If, as is consistent with protein structures, we assume that we do not permit the acute bond angle of 60° or the straight 180° for chain generation, then there are 16 different possible conformations possible on this lattice. All pairs of conformational angles are listed in Table 1.

**Table 1.** *Possible conformations on the structure-based cubo-octahedral lattice*[a]

| State | $\delta_1$ | $\delta_2$ | $\psi$ | States available for the next move |
|---|---|---|---|---|
| 1 | 90° | 90° | 0° | 2,3,4,5,6 |
| 2 | 90° | 90° | 180° | 1,2,3,4,5,6 |
| 3 | 90° | 120° | 55° | 7,10,11,13,14,16 |
| 4 | 90° | 120° | 125° | 8,9,11,12,15,16 |
| 5 | 90° | 120° | −55° | 8,9,11,12,15,16 |
| 6 | 90° | 120° | −125° | 7,10,11,13,14,16 |
| 7 | 120° | 90° | 55° | 1,2,3,4,5,6 |
| 8 | 120° | 90° | 125° | 1,2,3,4,5,6 |
| 9 | 120° | 90° | −55° | 1,2,3,4,5,6 |
| 10 | 120° | 90° | −125° | 1,2,3,4,5,6 |
| 11 | 120° | 120° | 0° | 8,9,11,12,15,16 |
| 12 | 120° | 120° | 70° | 8,9,11,12,15,16 |
| 13 | 120° | 120° | 109° | 7,10,11,13,14,16 |
| 14 | 120° | 120° | −70° | 7,10,11,13,14,16 |
| 15 | 120° | 120° | −109° | 8,9,11,12,15,16 |
| 16 | 120° | 120° | 180° | 8,9,11,12,15,16 |

[a] $\delta_1$ and $\delta_2$ are the two bond angles and $\psi$ is the dihedral angle that serve to define the positions of any four connected lattice points. Bond angles for the acute angle of 60° and the straight angle of 180° are not included.

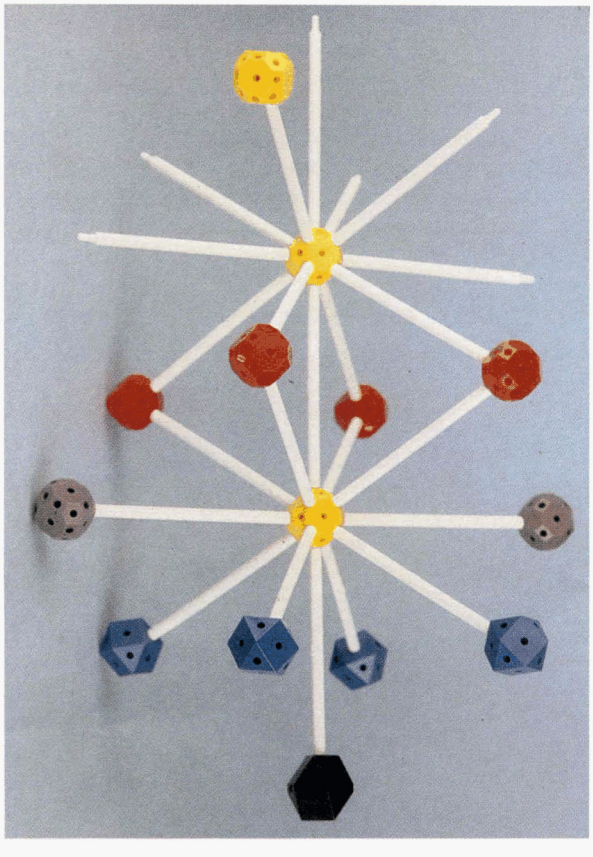

**Fig. 6.** An illustration of the bond angles and torsion angles possible in the lattice. With respect to the central bond (formed by the second and third yellow balls from the left), the four red balls are at 60°, the two gray balls are at 90°, the four blue balls at 120°, and the black ball on the right is at 180°. With respect to the yellow ball on the left, these balls have several torsion angles (see text).

The usual virtual bond angles connecting sequential $C^\alpha$s are used in conformational studies of polypeptide chains (Flory, 1969). These inter-$C^\alpha$ bond angles and torsion angles were used in describing secondary structures and protein chains (Srinivasan et al., 1975), in dipeptide conformational studies (Nishikawa et al., 1974), and in simulations to fold a simplified protein chain. Correlations of dihedral angles of successive residues have been reported (DeWitte & Shakhnovich, 1994). More recently, by using a much larger data set, the inter-dependencies of the bond angles and torsion angles for all amino acids have been analyzed in detail (Bahar et al., 1997). One dominant theme that emerges from these studies is the observance of two strong peaks in the plots of bond angles against torsion angles for four connected $C^\alpha$ atoms. A combination of bond angles in the neighborhood of 90° and a torsion angle near 55° characterizes the $\alpha$-helix region. The combination of 120° and 180° for bond angle and torsion angle characterizes the $\beta$-strand region. These combinations exist in the present lattice. In Table 1, conformations corresponding to $\{\delta_1 = 120°, \delta_2 = 120°, \psi = 180°\}$ represent a $\beta$-strand. $\alpha$-helix is represented by $\{\delta_1 = 90°, \delta_2 = 120°, \psi = 55°\}$ and $\{\delta_1 = 120°, \delta_2 = 90°, \psi = 55°\}$. One has to alternate between the latter two states for an $\alpha$-helix. Plastic models and fits of helices to this lattice give relatively small deviations for various $\alpha$-helices. The above combination of bond angles and a dihedral angle $\psi$ of 55° represents a right handed $\alpha$-helix. The same combination of bond angles with $\psi = -55°$ represents a left-handed $\alpha$-helix. Together with the remaining 13 conformations (Table 1), turns of various kinds can be fit well. Thus, various secondary structures such as helices, strands, and turns can be realistically described with this lattice. The 16 conformations constitute all possible configurations of four adjacent lattice points. However only six combinations of $\delta_2$ and $\psi$ are available for an additional point (see Table 1). In a simulation of a growing polypeptide chain, each new point would thus have a maximum choice

of six states possible, four with a bond angle of 120° and two with a bond angle of 90°.

*Fit of proteins to lattice*

Although this lattice provides excellent fits for close non-bonded residues, it is interesting to investigate how well this lattice can fit overall structures, even though that is not the main focus of this paper. There have been many works in which lattices of various types were used to fit protein structures (Covell & Jernigan, 1990; Kolinski & Skolnick, 1994; Park & Levitt, 1995; Reva et al., 1995, 1996).

We have fit several proteins with the present lattice as shown in Figure 7. The RMS deviation between the corresponding pairs of actual $C^{\alpha}$ atom positions and lattice points for myoglobin, an all-$\alpha$-helix protein, is 2.04 Å, which is near half the virtual bond length or lattice unit of 3.8 Å. The eight helices and the turns are described well. Lysozyme, a protein that contains a mixture of helices and strands, fits with an RMS deviation of 1.86 Å. Carboxypeptidase, a large protein with 306 amino acids, fits with an RMS deviation of 2.52 Å. The relatively good fits to this lattice suggest that the lattice captures the essential geometric features of both long-range/non-bonded and short-range/local interactions in protein structures. We feel that this lattice can be used to describe protein structures with high fidelity and hence should find wide use for protein simulations.

*Side-chain stereochemistry*

In most of the discussion so far, a lattice point has been taken to represent a $C^{\alpha}$ atom. Next, we want to learn whether it is possible also to include side chains on the lattice. Side chains of all amino acids have unique directional dispositions with respect to the directions of the preceding and succeeding residues. This feature is illustrated in Figure 8, which shows the directions of side-chain vectors of leucines, when the vectors $C^{\alpha}_{i-1}$–$C^{\alpha}_{i}$ and $C^{\alpha}_{i}$–$C^{\alpha}_{i+1}$ in the chosen proteins are superimposed for all of the residues in the set of proteins. Although the vectorial superposition highlights the specific direction of the side chain, a description in terms of the internal conformational angles would be useful for chain genera-

**Table 2.** *Lattice states of side chains, relative to backbone states. States given are those occupied by the $C^{\alpha}$ of the $i^{th}$ amino acids and the corresponding positions for side chains of the $(i-1)^{th}$ amino acid*[a]

| $C^{\alpha}$ state | Side-chain states |
|---|---|
| 1 | 2, 6 |
| 2 | 1, 3 |
| 3 | 5, 6 |
| 4 | 1, 5 |
| 5 | 2, 4 |
| 6 | 3, 4 |
| 7 | 9, 10 |
| 8 | 9, 11 |
| 9 | 8, 16 |
| 10 | 7, 8 |
| 11 | 15, 16 |
| 12 | 14, 15 |
| 13 | 11, 14 |
| 14 | 13, 16 |
| 15 | 12, 13 |
| 16 | 11, 12 |

[a]See Table 1 for the defining conformational angles corresponding to the above states.

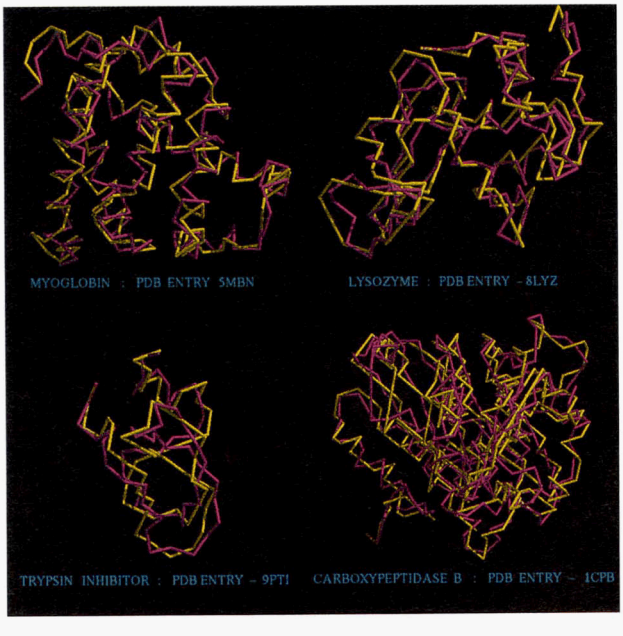


**Fig. 7.** Fits of some protein crystal structures to the present lattice. RMS deviation between the corresponding pairs of actual $C^{\alpha}$ atom positions and lattice points for myoglobin (top left), an all $\alpha$-helix protein, is 2.04 Å, which is just about half of the virtual bond length or lattice unit of 3.8 Å. The eight helices and the turns are described well. Lysozyme (top right), a protein that contains a mixture of helices and strands, fits with an RMS deviation of 1.86 Å. Trypsin inhibitor is shown bottom left. Carboxypeptidase (bottom right), a large protein with 306 amino acids, fits with an RMS deviation of 2.52 Å.

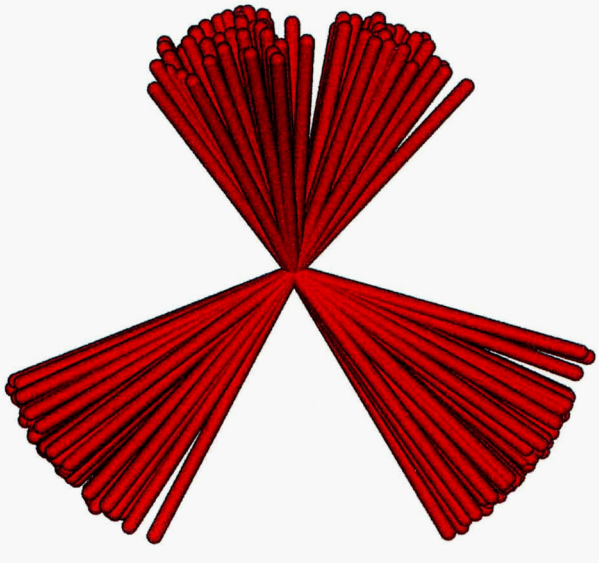

**Fig. 8.** Stereochemistry of an amino acid's side chains of 98 leucines with respect to their preceding and succeeding amino acids. Side-chain centers of the $i^{th}$ amino acid tend to be clustered in a specific direction (top of the figure), when vectors $C^{\alpha}_{i-1}$–$C^{\alpha}_{i}$ and $C^{\alpha}_{i}$–$C^{\alpha}_{i+1}$ are superimposed.

tion and for fixing side chains relative to the backbone. This observation is shown more quantitatively in Figure 9. The torsion angle for generating the side chain position of the $i^{th}$ amino acid $SC_i$, namely, $C^{\alpha}_{i-2}$–$C^{\alpha}_{i-1}$–$C^{\alpha}_{i}$–$SC_i$ is usually different by $-120°$ to $-190°$ from that of the backbone torsion angle, $C^{\alpha}_{i-2}$–$C^{\alpha}_{i-1}$–$C^{\alpha}_{i}$–$C^{\alpha}_{i+1}$. This small range of difference in torsion angle is observed for all 20 amino acids. However there are some variations among the 20 different amino acids. Correlated variations of virtual bond and torsion angles in the chosen set of proteins for all amino acid types as well as for some representative amino acid types are given in Figure 9.

Variations in the number and locations of the peaks in the observed distributions of the 20 amino acids correspond to the characteristic side-chain orientations and hence to the distribution over the different conformational minima possible about side-chain rotatable bonds. The locations and shapes of the observed peaks are very similar for Arg, Gln, Glu, Lys, and Met, and, as a representative, the observed region for Gln is shown in Figure 9, lower left part. For these five amino acids, the $\gamma$ atom is approximately tetrahedral and unbranched. A strong correlation between the side-chain torsion angles $\chi_1$ and $\chi_2$ has been reported from conformational energy calculations (Ponnuswamy & Sasisekharan, 1971a, 1971b) and from protein crystal structure analysis (Chandrasekaran & Ramachandran, 1970; Janin et al., 1978). Janin et al. (1978) reported quite similar characteristics in the distribution of the observed side-chain torsion angles for the above five amino acids. A significant amount of data had $\chi_1$ in the trans domain. Furthermore, in these amino acids, almost 90% of the observed data is distributed among five pairs of torsion angles $\chi_1$ and $\chi_2$, namely, $g^-t$, tt, $tg^-$, $g^+t$, and $g^+g^+$. Leucine, which differs from the above five amino acids in that it has a branched $C^{\gamma}$ atom, has a similar distribution. Compared to leucine, the breadth of the peak for valine, which is smaller by a -CH2 group, is smaller. Phe, Tyr, Trp, and His have two peaks each and the representative observed distribution for Phe is shown in Figure 9, lower right part. These amino acids have an aromatic ring in their side chains and a branched $C^{\gamma}$ atom. There is a pseudo-symmetric distribution of ring atoms about the $C^{\beta}$–$C^{\gamma}$ bond in Phe, Tyr, and His, which renders rotations related by 180° about $\chi_2$ to be equivalent. The occupied region is smaller for the smaller amino acids, such as Ser, Thr, Asn, and Asp. The location of the peak for proline differs from the other amino acids due to its unique ring structure. Thus, in contrast to the remarkable regularity for all 20 amino acids displayed in the coarse-grained level of non-bonded residue packing, there are still some larger significant variations in the short-range distributions.

Information on the relative orientations of a side chain and its preceding and succeeding neighbor backbone can be useful for placing side chains either on or off lattice while growing a protein chain. The individual side chains for the 20 amino acids still have characteristic peaks within a narrow region of conformational space. The proposed lattice also has appropriate geometry for incorporating the different amino acid side chains. Addition of side chains will add more detail and specificity to the generated protein chains. Since side chains occupy a definite volume, their inclusion will exclude a significant part of conformational space available to a growing peptide chain and hence reduce conformational searches considerably. This, in turn, would enhance the feasibility for generating chains in a restricted conformational space as chain lengths increase. Thus, inclusion of side chains may actually permit consideration of longer protein chains.

Lattice methods are ideal for generating and, especially, for enumerating large numbers of structures. Unlike other, conven-

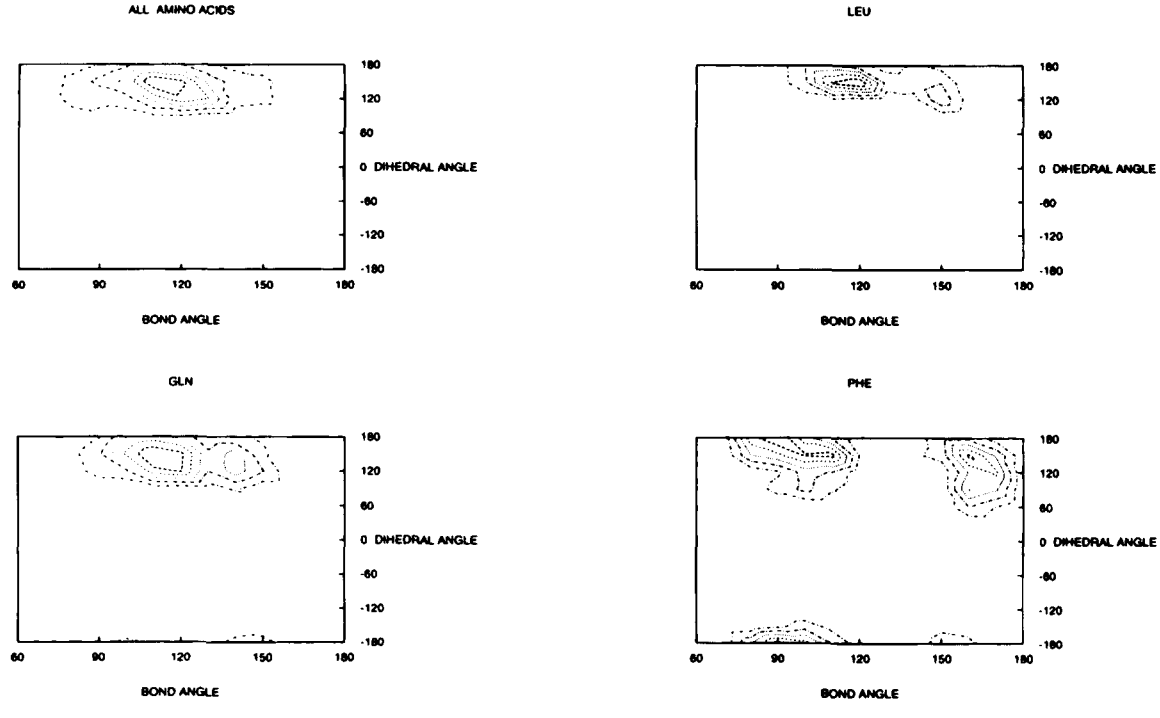


**Fig. 9.** Correlation between the virtual bond angles of the side chain and the torsion angle relating the side chain and the backbone of the succeeding amino acid for some amino acids. Bond angles on the abscissa refer to the angle $C^{\alpha}_{i-1}$–$C^{\alpha}_{i}$–$C^{\alpha}_{i+1}$. The torsion angle on the ordinate is the difference between the two torsion angles DA1–DA2, where DA1 is for $C^{\alpha}_{i-2}$–$C^{\alpha}_{i-1}$–$C^{\alpha}_{i}$–$C^{\alpha}_{i+1}$ and DA2 is for $C^{\alpha}_{i-2}$–$C^{\alpha}_{i-1}$–$C^{\alpha}_{i}$–$SC_i$, where $C^{\alpha}$ is the coordinate of the $C^{\alpha}$ atom and SC is the mean coordinate of the side-chain atoms.

tional methods such as molecular mechanics/dynamics and Monte Carlo methods, the final structures selected are not biased by starting conformations. Relatively large systems can be studied without the problems of local energy minima. Lattice calculations possess a number of advantages over other procedures: Integer based computations, simple criteria for self-avoidance, and a feature essential for putting in distance constraints, the certainty that rings can be closed. However, major advances are required in two areas to take full advantage of the potential of these methods. The first of these involves the development of potential functions to rank order and discriminate among structures. These methods usually involve residue-pair potentials, which give the likelihood that a residue comes within interacting distance of another residue compared to the mean force field operative between all groups in the protein. Potentials have been developed that include long-range and short-range interactions (Miyazawa & Jernigan, 1985, 1996; Bryant & Lawrence, 1993; Wodak & Rooman, 1993; Casari & Sippl, 1992; Nishikawa & Matsuo, 1993; DeWitte & Shakhnovich, 1994; Kocher et al., 1994; Godzik et al., 1995; Levitt et al., 1995; Sippl, 1995; Bahar & Jernigan, 1996a). The second and less developed area is the development of alternative lattices to describe protein architecture better. Most previous lattice simulations have used cubic lattices or several variants of the cubic lattice. There have also been some reports that have used other types of lattices, such as the tetrahedral lattice (Hinds & Levitt, 1992) and the knight's walk lattice (Skolnick & Kolinski, 1991). The lattice proposed here differs from these. It is knowledge-based in the sense that the set of points for this lattice has been chosen directly to coincide with the observed local distributions of non-bonded amino acid directions around a central amino acid. A large number of high-quality, diverse crystal structures have been used as the input data for this analysis. As we discussed above, the lattice captures features of the geometries of long-range, non-local interactions and also of short-range, local interactions. This lattice is based on observed distributions of directions of amino acid neighbors in space and reproduces the essential features of dense packing of uniform spheres. The observed distributions and the lattice points represent a best way to pack spheres in three dimensions. Hence, it should be ideal for representing the inner cores of proteins that are dominated by hydrophobic interactions. Further, it has relatively good combinations of conformational bond angles and torsion angles to describe local interactions of sequential neighbors, such as $\alpha$-helices, $\beta$-strands, and various turn types with high fidelity. Protein chains fit well to this lattice. The geometry of the lattice also permits the correct placement of side chains. We believe the proposed lattice has the essential characteristics to represent proteins realistically and hence that it is ideal for protein simulations. Lattices with higher coordination numbers could be used to fit structures better, but they would be less useful for chain enumerations because of the larger number of choices for placing successive units in growing chains.

## Methods

All non-bonded amino acids surrounding a central amino acid within a distance of 6.5 Å or less are considered to be its non-bonded neighbors. This distance is based on the occurrence of the first peak in the radial distribution of residues in the interior of proteins. Similar results were found for distance criteria of 6.0 Å and 7.0 Å (results not shown). Also, we have chosen not to look in detail at the distribution of distances. The directions of the non-bonded neighbors about a central residue, which is positioned at the origin, are represented as a set of unit vectors from the origin. Sets of these unit vectors corresponding to 37,095 amino acids in 161 proteins were superimposed on this template set as follows: The data set of unit vectors was rotated about the $x$, $y$, and $z$ axes in 12° intervals. All distances $d_{ij}$ among all $i$ unit vectors of the data set and the $j$ unit vectors of the template set were calculated. The best superposition was chosen for the orientation with the minimum sum of distances between the protein unit vectors and the nearest points of the target set. The best superposition between the pairs of vector sets does not depend on the sequential order of the vectors. Examination of the superpositions at various intermediate stages suggests strong and distinct clusters in specific directions. In the second stage, these standard directions were chosen and subsequently used for superposing the final sets of vectors as in the first stage.

The following criteria were used to select the set of high-quality, diverse protein structures from the protein data bank: No two proteins have a higher sequence identity than 25%; resolution of the structures is 2.0 Å or better; and crystallographic $R$-factors are below 20%. Application of these criteria led to the selection of 161 proteins.

Calculations were carried out for 37,095 sets of vectors of the 161 chosen proteins. Two separate calculations were carried out. In one of them, only the non-bonded (sequentially non-adjacent) residues were considered for generating the neighbor lists. Similar calculations were also performed including the sequential neighbors in the vector set. Calculations were performed on a CRAY YMP. It requires 13.7 s to superimpose two sets of non-bonded neighbors only (141 h for all the 37,095 vector sets in 161 proteins) and 26.5 s (to superimpose two sets containing both non-bonded and sequentially bonded neighbors (273 h for the set of proteins). For each of these two calculations, two additional sets of calculations were carried out. In one of them the $C^{\alpha}$s were used to represent amino acids and in the other the centers of side chains. In each of the above four sets of calculations (Fig. 2), analyses were carried out separately for each of the 20 types of amino acids.

## References


Bahar I, Jernigan R. 1996a. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J Mol Biol 266*:195–214.

Bahar I, Jernigan R. 1996b. Coordination geometry of non-bonded residues in globular proteins. *Folding & Design 1*:357–370.

Bahar I, Kaplan M, Jernigan R. 1997. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. *Proteins*. Forthcoming.

Baldwin E, Matthews B. 1994. Core-packing constraints, hydrophobicity, and protein design. *Curr Opin Biotechnol 5*:396–402.

Behe M, Lattman E, Rose G. 1991. The protein-folding problem: The native fold determines packing, but does packing determine the native fold? *Proc Natl Acad Sci USA 88*:4195–4199.

Bryant SH, Lawrence CE. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins 16*:92–112.

Casari G, Sippl MJ. 1992. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J Mol Biol 224*:725–732.

Chandrasekaran R, Ramachandran G. 1970. Studies on the conformation of amino acids. XI. Analysis of the observed side group conformations in proteins. *Int J Protein Res 2*:223–233.

Cho J, Mattice W. 1997. Estimation of long-range interactions in coarse-grained rotational isomeric state polyethylene chains on a high-coordination lattice. *Macromolecules 30*:637–644.

Covell DG, Jernigan RL. 1990. Conformations of folded proteins in restricted spaces. *Biochemistry 29*:3287–3294.

DeWitte R, Shakhnovich E. 1994. Pseudodihedrals: Simplified protein backbone representation with knowledge-based energy. *Protein Sci 3*:1570–1581.

Eriksson A, Baase W, Zhang X-J, Heinz D, Blaber M, Baldwin E, Matthews B. 1992. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science 255*:178–183.

Flory P. 1969. *Statistical mechanics of chain molecules*. New York: Interscience Publishers.

Godzik A, Kolinski A, Skolnick J. 1995. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci 4*:2107–2117.

Hinds D, Levitt M. 1992. A lattice model for protein structure prediction at low resolution. *Proc Natl Acad Sci USA 89*:2536–2540.

Janin J, Wodak S, Levitt M, Maigret B. 1978. Conformation of amino acid side-chains in proteins. *J Mol Biol 125*:357–386.

Jernigan R, Raghunathan G, Bahar I. 1994. Characterization of interactions and metal ion binding sites in proteins. *Curr Opin Struct Biol 4*:256–263.

Kappraff J. 1991. *Connections. The geometric bridge between art and science*. New York: McGraw-Hill, Inc.

Kauzmann W. 1959. Some factors in the interpretation of protein denaturation. *Adv Prot Chem 14*:1–63.

Kocher J-PA, Rooman MJ, Wodak S. 1994. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol 235*:1598–1613.

Kolinski A, Skolnick J. 1994. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins 18*:338–352.

Levitt M, Hirshberg M, Sharon R, Daggett V. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comp Phys Commun*. Forthcoming.

Lim W, Hodel A, Sauer R, Richards F. 1994. The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc Natl Acad Sci USA 91*:423–427.

Miyazawa S, Jernigan R. 1985. Estimation of effective inter-residue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.

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

Munson M, Balasubramanian S, Fleming K, Nagi A, O'Brien R, Sturtevant J, Regan L. 1996. What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. *Protein Sci 5*:1584–1593.

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

Nishikawa K, Moomany F, Scheraga H. 1974. Low-energy structures of two dipeptides and their relationship to bend conformations. *Macromolecules 7*:797–806.

Park BH, Levitt M. 1995. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol 249*:493–507.

Ponnuswamy P, Sasisekharan V. 1971a. Studies on the conformation of amino acids. IV. Conformations of serine, threonine, cysteine, and valine. *Int J Protein Res 3*:1–8.

Ponnuswamy P, Sasisekharan V. 1971b. Studies on the conformation of amino acids. V. Conformation of amino acids with delta-atoms. *Int J Protein Res 3*:9–18.

Ramachandran G, Ramakrishnan C, Sasisekharan V. 1963. Stereochemistry of polypeptide chain configurations. *J Mol Biol 7*:95–99.

Rapold R, Mattice W. 1996. Introduction of short- and long-range energies to simulate real chains on the 2nd lattice. *Macromolecules 29*:2457–2466.

Reva BA, Finkelstein AV, Rykunov DS, Olson AJ. 1996. Building self-avoiding lattice models of proteins using a self-consistent field optimization. *Proteins 26*:1–8.

Reva BA, Rykunov DS, Olson AJ, Finkelstein AV. 1995. Constructing lattice models of protein chains with side groups. *J Comput Biol 2*:527–535.

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

Richards F, Lim W. 1993. An analysis of packing in the protein folding problem. *Q Rev Biophys 26*:423–498.

Singh J, Thornton J. 1992. *Atlas of protein side-chain interactions*. Oxford: IRL Press.

Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr Opin Struct Biol 5*:229–235.

Skolnick J, Kolinski A. 1991. Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure, and dynamics. *J Mol Biol 221*:499–531.

Srinivasan R, Balasubramanian R, Rajan S. 1975. Some new methods and general results of analysis of protein crystallographic structural data. *J Mol Biol 98*:739–747.

Wells A. 1956. *The third dimension in chemistry*. Oxford: Clarendon Press.

Wodak S, Rooman M. 1993. Generating and testing protein folds. *Curr Opin Struct Biol 3*:247–259.

Wynn R, Harkins P, Richards F, Fox R. 1996. Mobile unnatural amino acid side chains in the core of staphylococcal nuclease. *Protein Sci 5*:1026–1031.