Residue packing in proteins: Uniform distribution on a coarse-grained scale

Zerrin Bagci[a)]
Center for Computational Biology and Bioinformatics, and Department of Molecular Genetics and Biochemistry, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15213 and Chemical Engineering Department and Polymer Research Center, Bogazici University, Bebek 80815, Istanbul, Turkey

Robert L. Jernigan[b),c)]
Molecular Structure Section, Laboratory of Experimental and Computational Biology, MSC 5677, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-5677

Ivet Bahar[d)]
Center for Computational Biology and Bioinformatics, and Department of Molecular Genetics and Biochemistry, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15213

(Received 28 September 2001; accepted 12 November 2001)

The high packing density of residues in proteins ought to be manifested in some order; to date this packing order has not been thoroughly characterized. The packing regularity in proteins is important because the internal organization of proteins can have a dominant effect on functional dynamics, and it can aid in the design, simulation and evaluation of structures. Packing metrics could also inform us about normal sequence variability, an issue that, with the accumulating genome data, becomes increasingly important. Other studies, indicating a possible correlation between packing density, sequence conservation, and folding nucleation [O. B. Ptitsyn, J. Mol. Biol. **278**, 655 (1998)], have emphasized the importance of packing. Here, residue clusters from protein databank structures, each comprised of a central residue and all neighbors located within the first coordination shell, have been rigidly re-oriented and superimposed in a self-consistent optimization. About two-thirds of residues are found to follow approximately the relative orientation preferences of face-centered-cubic packing, when examined on a coarse-grained scale (one site per residue), while the remaining one-third occupy random positions. The observed regularity, which becomes more pronounced after optimal superimposition of core residues, appears to be the result of uniform sampling of the coordination space around each residue on a coarse-grained scale with hydrophobic clustering and volume exclusion, to achieve packing densities close to that of the universal closest packing of identical spheres. © *2002 American Institute of Physics.* [DOI: 10.1063/1.1432502]

I. INTRODUCTION: MACROMOLECULAR CONFORMATIONS AND INTERNAL PACKING IN PROTEINS

Many aspects of protein structures relate to internal packing considerations, including the design, simulation and evaluation of structures,[1] as well as sequence conservation and even folding nucleation.[2,3] Historically, there has been a strong focus on the conformations of protein backbones; however, because of competition between local and long-range interactions, it is not clear where the greatest regularity should appear. The regularity observed here is found in the *orientation angles* among close residues, irrespective of their sequential separation. The distances between residues depend on the specific amino acid pairs involved because of their different *sizes*. The angular positions could also be somewhat residue-specific, because of the different *shapes* of amino acid side chains. The present study sheds light on how amino acids of different sizes and shapes are compatible with dense, regular packing, when observed on a coarse-grained scale.

Backbone conformational isomers in small molecules and polymers are also regular insofar as only small ranges of torsion angles are allowed; the bond lengths, however, depend on the specific types of bonded atoms. The best-known rotational isomers are the *trans*, *gauche*$^+$, and *gauche*$^-$ states of hydrocarbons.[3–6] A similar type of conformational regularity is observed in the side chains of proteins.[7,8] Side chains can be disordered in protein crystals, but are usually constrained to sample only among the different rotational isomers. Distributions over the protein backbone torsion angles exhibit a greater breadth,[9] reflecting distortions arising from the competition among preferred regularities in the backbone conformation, hydrogen bonding, and the nonbonded or packing interactions. Even α-helices and β-strands can be viewed as selected rotational states caused by the drive to achieve high packing densities or optimize

[a)] Electronic mail: bagci@pitt.edu
[b)] Electronic mail: Robert_Jernigan@nih.gov
[c)] Present address: Laurence H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, Iowa 50011
[d)] Corresponding author. Electronic mail: bahar@pitt.edu

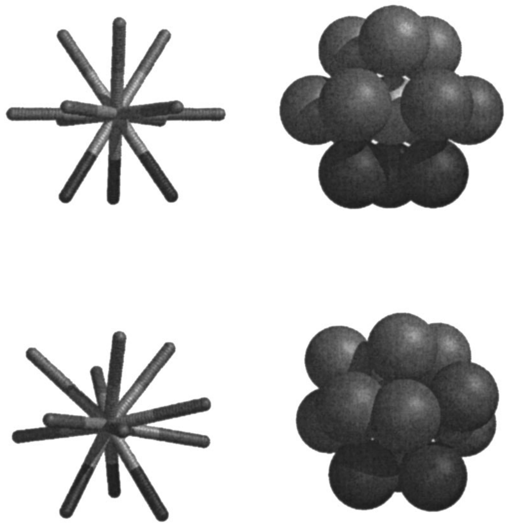

FIG. 1. Densest packing of identical spheres (top) in face centered cubic (fcc) geometry, shown on the right and the relative neighbor directions on the left side. The central sphere has six neighbors in the central plane, three above and three below. The positions of the upper spheres are staggered with respect to the positions of the lower ones. Observed residue packing in proteins (bottom) for residue clusters, with directions shown on the left and sphere packing on the right. The protein packing directions closely match the sphere packing directions shown at the top.

nonbonded interactions,[10] i.e., a structural manifestation of the hydrophobic effect. Recent studies suggest that protein α-helices or DNA double helices are simply optimal shapes, achieving the highest packing densities locally.[11,12] In addition to the many other proposed reasons for amino acids to be selected to make so many of the functional molecules in biology, one can even wonder if the peptide backbone might have been evolutionarily selected for its unusual malleability to meet, on a local basis, these global packing restraints. We will show that the local orientations of consecutive residues along the sequence do conform to the regular packing geometry preferred by nonbonded neighbors, which in turn closely matches the universal (closest packing) of identical spheres. The observed regularity in the occurrences of similar coordination angles, on the other hand, is the result of an optimal discretization of the coordination space around a central residue, because the coordination space is *uniformly* sampled by near neighbors when observed on a coarse-grained scale.

II. DIFFERENT VIEWS FOR RESIDUE COORDINATION IN PROTEINS

Several previous studies have attempted to characterize the coordination geometry of side chains.[13–15] However, atomic details can obscure the search for regularity, which may only be observable with a coarse-grained view, such as is frequently utilized for general descriptions of protein architecture. Coarse-grained views of protein packing have ranged from the extreme regularity typical of the closest packed (face-centered-cubic; fcc; Fig. 1, upper diagrams) lattice,[16] or the perfect complementarity resembling a jigsaw puzzle,[17] to a completely random arrangement devoid of complementarity and directionality similar to the arrangement of nuts and bolts in a jar.[18] We find that, for coarse-grained descriptions of proteins, in which a single interaction site represents each residue, taken here as the C^β atoms, a latticelike model is approximated, rather than either the jigsaw puzzle or the nuts-and-bolts model.

The commonly observed insensitivity of structures to single site mutations[19,20] has been attributed to a ductile reassociation of sidechains.[21] This ductility could originate either in disordered packing, as with the nuts-and-bolts model, or more likely in the preferential but sequence-independent packing characteristic of protein interiors, where each site can readily accommodate some range of residue substitutions. Our aim is to search for the occurrence of a regularity or internal order in folded proteins, other than those observed at the level of secondary structures. Tertiary structures will be observed here from a different perspective, in the absence of a model and observation frame that depend on the atomic details and geometric characteristic of the particular amino acids.

Lattice models have been widely exploited in theoretical and computational studies of protein structures on a coarse-grained scale, and among other representations, the face-centered-cubic (fcc) lattice has proven to be particularly useful in early threading studies.[22] In other studies, we and others have demonstrated that the fcc packing is *not* the only possible architecture with which one can fit well the coordination geometry of residues, but simply *one* of many possible descriptions.[23,24] These results were obtained using a *constrained fit* method.[24] There, database-extracted clusters—consisting of a central residue and the m surrounding residues located in the first coordination shell—were constrained by suitable rigid-body rotations to occupy angular positions as close as possible to the coordination directions of different target lattices. As expected, the quality of the match improves with the coordination number of the target lattice.[23,24] However, the more important basic question is instead, what is the actual geometry in protein structures?

III. METHOD

Here we use two different approaches to determine what are the real geometries. In the first method, shortly referred to as *optimal superimposition*, a Monte Carlo algorithm is utilized for superimposing residue clusters collected from known protein structures. We consider a statistical ensemble of proteins, mainly a representative set of nonhomologous structures deposited in the Protein Data Bank (PDB).[25] A residue cluster is composed of a central residue, and the set of all neighboring residues located within a first coordination shell. A radius of 6.8 Å is used for defining the first coordination volume, based on our earlier statistical analyses of database structures.[26–28]

The database-extracted clusters are represented, each, by a bundle of unit directional vectors that originate at the central residue and point to the coordinating residues. The number of directional vectors in a given bundle is equal to the coordination number of the central residue. It varies in the range $3 \leqslant m \leqslant 14$, depending on the location (surface or core) of the central residue. The individual bundles are then optimally superimposed onto each other by an iterative Monte Carlo scheme, in which a randomly selected bundle is subjected to an incremental rigid-body rotation, while all others are held constant. The root-mean-square deviation between the tips of the matching (closest) pairs of directional pairs, averaged over all pairs, is used as a criterion for accepting or rejecting each move.[24] Calculations show that the results converge after 10^6 moves, and results are reproducible when 1000 sets of bundles are independently analyzed. This set is large enough to ensure statistical convergence—as verified by repeating the calculations with different sets of clusters, and performing longer simulations—and yet small enough to be optimally superimposed within feasible computational time. The computational time for optimally superimposing 1000 bundles is about 50 h using an SGI O2 R5000 workstation.

The second approach is based on Voronoi tessellation (VT) methods that have been widely used for examining protein packing, volumes and surface area[17,29–38] starting from the original studies of Richards[30] and Finney.[31] An advantage of the VT methods, and the Delaunay tessellations that essentially contain the same information,[39] is that the coordination range need not be defined prior to calculations. Thus, biases that arise from the adoption of a fixed coordination volume around a given residue are avoided in the second approach. A major difficulty in this approach is, on the other hand, the assignment of the Voronoi polyhedra to surface residues, which may necessitate including or modeling solvent molecules. In the application of VT to single-site-per-residue models of proteins,[37,39] the space is divided into polyhedra enclosing the individual residues; the bisector planes perpendicular to inter-residue vectors define the faces of the Voronoï polyhedra, and the intersections of these planes form the edges. In this method, surface-exposed polyhedra can extend to infinity or be very elongated due to the lack of neighbors. Such complications are avoided by discarding the Delaunay tetrahedra whose circumscribed sphere radius exceeds the cutoff distance of 10 Å.[37,39] Thus a cutoff distance is adopted in these *modified* VT methods used in combination with Delaunay tetrahedra.

IV. RESULTS

About two-thirds of coordinating residues are superimposable along seven preferred directions.

The optimal superimposition of clusters leads to a relatively diffuse distribution of coordination angles.[40] Figure 2(A) displays the results for the superposition of 1000 bundles of directional unit vectors extracted from the PDB. The surface and the projected contour map on the lower plane represent the probability distribution of the orientations assumed by the directional vectors of the optimally superimposed bundles, expressed in terms of the polar (θ) and azimuthal (ϕ) angles with respect to a fixed frame. Parts (A) and (B) refer to the clusters including and excluding the first (bonded) neighbors along the chain sequences, respectively. Seven peaks are distinguishable in both cases. Except for the weakest, each peak exhibits occupancy near 10%, based on the fraction of residues located within 20° of solid angle about each directional vector. The sum of these seven probabilities is 0.63, so approximately two-thirds of residues are found to occupy these coordination states, and the remaining one-third occupies other positions in space. For random packing, the probability of occupancy of a coordination state defined by angular deviations up to 20° around a given central direction would be $\int_0^{\pi/9}\cos\theta\,\mathrm{d}\theta/\int_0^{\pi}\cos\theta\,\mathrm{d}\theta=0.03$; so the total probability of occupancy of seven such states becomes 0.21. This indicates that the preferred packing architecture is favored by a factor of three (0.63:0.21), over random packing.

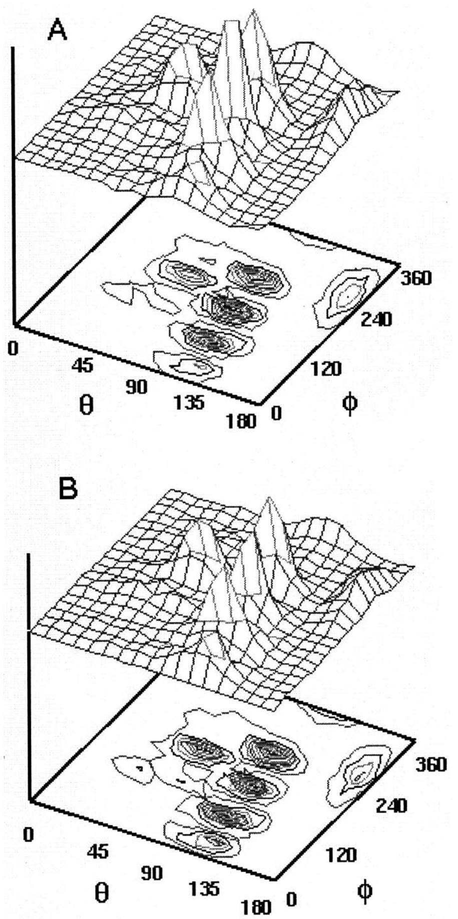

FIG. 2. Distribution of coordination angles (θ: polar; ϕ: azimuthal) obtained from optimal superimposition of clusters comprised of closely interacting residues including (A) all neighbors, and (B) nonbonded neighbors only around a central residue. An optimization algorithm coupled with Monte Carlo iterations, executed up to 3×10^6 steps is adopted to achieve optimal superpositions. At each step, a randomly chosen cluster is rotated randomly and the mean deviation from all other clusters is computed. If the mean deviation decreases with respect to the original state, the new rotation is accepted and vice versa. The mean deviation is the average distance between the tips of the closest unit vectors, evaluated for all pairs of directional vectors for all 10^6 pairs of clusters. Comparison of parts (A) and (B) demonstrates that the distribution is almost unaffected by including or excluding bonded neighbors. Comparison with Fig. 1 shows that significantly fewer coordination angles are included. Yet, the occupied sites are closely clustered, similarly to the dense packing in regular lattices, while about one-third of the coordination space is either unoccupied or sparsely populated.

Preferred directions are identical for bonded and nonbonded neighbors.

One could attribute the observed selection of particular coordination sites to the angular regularities imposed by the backbone. Calculations repeated for nonbonded neighbors alone show that this is not the case; an almost identical distribution is obtained [Fig. 2(B)]. This result corroborates pre-

TABLE I. Most probable coordination states observed upon optimal superimposition of residue clusters for proteins (Ref. 2).

Coordination number		Coordination states (°)												Ptot[†]
		1	2	3	4	5	6	7	8	9	10	11	12	
Surface	θ	40	45			95	90							0.40
$3\leqslant m\leqslant 4$	θ	30	170			50	110							(3.3)
All	θ	40	35	45	95	105	55	90*					120	0.63
$(3\leqslant m\leqslant 14)$	ϕ	10	200	285	350	50	115	180*					115	(3.0)
Core	θ	45	45	45	95	105	60	100	85	105	140			0.65
$m\leqslant 10$	ϕ	40	180	280	360	60	100	140	240	300	220			(2.2)
$m\geqslant 12$	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 lattice	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	

*For subset of specific aminoacids.

[†]Total probability for the full set of coordination states. Parenthetical numbers are the enhancements over random occupancies.

vious analyses suggesting that bonded and nonbonded neighbors need not be distinguished in order to describe satisfactorily inter-residue contact topology, and that they exhibit a similar extent of order. Furthermore, it suggests some of the reason for the success of dynamic models of proteins, where interactions between sequentially bonded residues are treated the same way as interactions between close nonsequential residues.[41]

The preferred directions cluster together leaving a fraction of the coordination space unoccupied, except for the core residues.

An interesting observation is that the most probable coordination sites are confined to a small subspace of the coordination space. As pointed out above, the fraction of residues that occupy this subspace is about two-thirds (or 0.63 when counting coordinating residues located within 20°; see Table I). The remaining one-third could refer to residues that are more loosely or randomly packed, probably being exposed to solvent. To understand the origin of this biased distribution of coordination sites, subsets of clusters composed of $m=10$ or more residues have been considered. These are ''core'' residues, based on observed coordination numbers in folded structures[25] (Berman *et al.*, 2001). The optimal superimposition of this subset of ''dense'' clusters yields the two ''global'' views of the angular distribution displayed in Fig. 3, parts (A) and (B). The figures display the most probable angular positions visited by the first neighbors around a reference residue located at the center of the coordination sphere. This distribution indicates that complete coverage of the coordination space is approached when the coordination geometry of core residues is considered. The ten most heavily populated sites, shown as dark patches on the sphere surface, correspond to the orientations listed in Table I for $m\geqslant 10$.[42,43]

The fraction of residues occupying the most probable ten coordination sites of core residues displayed in Figs. 3(A) and 3(B) is counted to be 0.65, allowing for 20^0 deviations about central directions. For a random arrangement of coordinating residues, the expected probability would be 0.30 for the ten regions. The observed probability indicates enhancement above random by a factor greater than two (0.65:0.30). We note that core residues cannot select their preferred coordination states as efficiently as other residues, due to more severe constraints.

Core residues' packing approximates the fcc geometry on a coarse-grained scale.

The azimuthal angle differences between the six neighboring sites 4–9 in Fig. 3(B) (see also Table I) are approximately 60°. An approximately hexagonal arrangement is consistent with $\Delta\phi=60°$ reported in our previous examination[21,44] of triplets of sidechains. These six sites can be viewed as comprising the middle layer in a closely packed arrangement (Fig. 1). Out of the remaining four sites, three, labeled 1–3, lie in the upper hemisphere ($\theta=45°$) and are

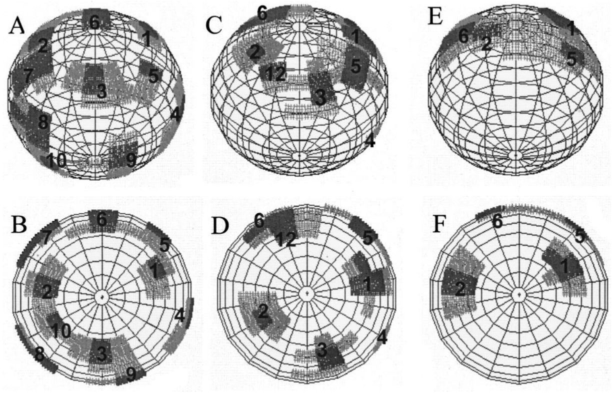

FIG. 3. Coordination sites obtained by repeating the optimal superimposition of three different subsets of clusters: Parts (A) and (B) refer to clusters composed of $m=10$ or more neighbors. The darkest patches indicate the most densely populated orientations; lighter patches are relatively less frequently occupied orientations. (C) and (D) display the results for all observed coordination numbers ($3\leqslant m\leqslant 14$). These are identical to those shown in Fig. 2, except for the rigid-body rotation of the coordination space, so as to facilitate the comparison with the other clusters. (E) and (F) show the results for the subset of clusters comprised of four or fewer neighbors. Essentially, the same sites are occupied in all cases, which could be associated with a partially filled, distorted fcc geometry. See Table I for identification of the labeled sites. All orientations are identified by the index numbers at the top of Table I.

separated by 120±20°. Interestingly, so far this arrangement closely approximates either hcp or fcc geometry. Finally, the last residue occupies a staggered position in the lower layer, thus conforming only to the fcc packing geometry. Hence overall, the optimal geometry in the core closely resembles fcc packing with two empty sites. We term this *an incomplete, distorted fcc packing*.

Additional calculations performed with even higher density clusters ($m \geqslant 12$) showed that the remaining two unoccupied sites are also filled in these most densely packed regions. See the results for $m \geqslant 12$ in Table I.

Increase in local packing density conforms to a gradual filling of fcc geometry directions.

As a further validation of the above-mentioned distorted, incomplete fcc geometry, we tested whether the seven optimal coordination directions found for all residues (Fig. 2) conform to the directional vectors of the core packing architecture. The coordination angles displayed in Fig. 2 refer to an arbitrary reference frame. The reference frame can instead be chosen so that the mean coordination directions of the superimposed bundles are oriented (insofar as possible) along those of the fcc lattice. The optimum rigid-body rotation of the complete set of superimposed bundles yields the coordination angles displayed in Figs. 3(C) and 3(D), which confirms our hypothesis of incomplete fcc coordination that is gradually filled as the local packing density increases. The sets of coordination directions, corresponding to ''all'' ($3 \leqslant m \leqslant 14$) and ''core'' ($m \geqslant 10$) residues, exhibit angular deviations between them below 30°. And finally, when cases having fewer neighbors ($m \leqslant 4$) alone are considered (mostly surface residues), we find four of these same sites to be occupied, approximately [Figs. 3(E) and 3(F)]. Table I summarizes the optimal coordination angles obtained for the various cases, along with the fraction of residues located in these sites (last column).

We conclude that the optimal internal angular architecture inside proteins can be well represented by fcc packing, the main difference being that not all sites are occupied, and some slight distortions in directional vectors are observed. This behavior emerged upon confinement to the subsets of densely packed ($m \geqslant 10$) clusters on a coarse-grained scale. It was essential to consider all neighbors within a first coordination shell, irrespective of their radial distance, and focus on their angular positions alone. Our analysis verifies that the same regular geometry holds even for surface residues, though more coordination sites are empty. It is interesting that coarse-grained models of polymers also have been shown to conform to the fcc lattice.[45,46]

Does internal packing in proteins follow the universal closest packing of identical spheres?

The fcc packing of spheres, although widely accepted to be the closest packing geometry for identical spheres, has only recently been rigorously proven.[47,48] The fact that protein interiors exhibit a tendency to assume this regular packing pattern—closely consistent with the observation[30,32] of packing densities of the order of 0.74—suggests that the same tendency is valid for protein interiors, when residues are examined at a coarse-grained scale. This type of nonspecific organization may be a manifestation of the hydrophobic drive to maximize the packing density.

It is worth pointing out that packing densities slightly higher than the fcc value have been reported for the interior of proteins,[33] with the likely explanation being that amino acids are not spheres, but have asymmetric shapes and internal degrees of freedom. Consequently, they can be reorganized or reconfigured to maximize the efficiency of packing, for example, to fill interstices, as in the jigsaw model. Size and shape differences, as in the nuts-and-bolts model, could also improve the efficiency of packing. These models could perhaps be reconciled with our observed regularities or uniformities provided that there are size and shape compensations in a given cluster, which become less discernible on a coarse-grained scale.

Delaunay tessellations yield comparable packing geometries if their coordination numbers are confined to ranges.

In the above calculations, we considered the neighbors located within a cutoff distance $r_c = 6.8$ Å from each central residue. This distance is indicated by extensive examinations of database structures to be the range of the first coordination shell around amino acids, in the presently adopted one-site-per-residue representation of folded proteins. One might wonder if the same results could be reproduced by other approaches of condensed matter physics, such as Voronoi tessellations (VT), in which the coordination range need not be defined prior to calculations.

In their modified VT method used in conjunction with Delaunay tetrahedra, Soyer *et al.*[37] found that the mean number of faces per Voronoï polyhedron is $\langle f \rangle = 13.97$ when each amino acid is represented by its geometric center, and that the mean number of edges per face is 5.14. This mean coordination number is higher than that observed in our analysis of clusters suggesting that a broader interaction distance is implicitly considered in the VT method. The reported[37] mean inter-residue distance (6.6 Å) is indeed comparable to the uppermost inter-residue distance (6.8 Å) included in the above-mentioned analysis.

For closer comparison with the optimal superimposition results, we use the *delaunay3.m* function of the Matlab 6.0 package to construct the Delaunay tetrahedra whose vertices coincide with the C^{β} atoms, with the centers of the spheres circumscribed by these tetrahedra defining the vertices of the VT cells. This method, applied to ~1200 clusters from the PDB yields significantly higher coordination numbers ($8 \leqslant m \leqslant 21$) than those ($3 \leqslant m \leqslant 14$) found with the cutoff distance of 6.8 Å. The mean coordination number is found to be 13.81, in close agreement with those of Soyer *et al.*[37] The mean inter-residue distance is 7.44 Å, this slightly higher value being attributed to the fact that we include all tetrahedra in our case, i.e., no upper cutoff value has been adopted for eliminating surface clusters with highly asymmetric shapes.

Next, we focus on the most probable coordination angles for VT cells. Given the high computational cost for the optimal superimposition of high coordination clusters, we focus on a subset of ~500 residue clusters, taken from the acetyl cholinesterase structure.[49] The mean coordination number is 14.22 for this subset, and the mean inter-residue distance is

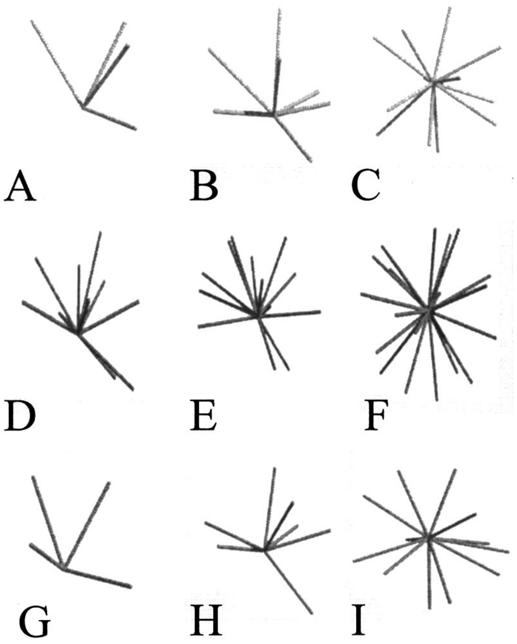

FIG. 4. Comparison of the coordination directions of residue clusters obtained with optimal superimposition [(A), (B), and (C)], Delaunay tesselation [(D), (E), and (F)], and coarse-graining of the Delaunay results [(G), (H), and (I)]. Panels (A), (B), and (C) show the results for representative surface $(m=4)$, intermediate $(m=7)$, and core $(m=12)$ clusters, using the orientations given in Table I. Panels (D), (E), and (F) refer to their counterparts in the Delaunay tessellation method, i.e., $m=10$, 14, and 19, respectively. Panels (G), (H), and (I) show the results after merging the coordination sites in (D), (E), and (F) according to their probabilistic weights so as to match the coordination numbers of the clusters displayed in (A), (B), and (C). There, a fairly close correspondence is seen between the first and last rows.

7.96 Å, slightly larger presumably due to the presence of a central cavity in the investigated protein. The results are summarized in Fig. 4. Parts (A)–(C) display the orientations of the clusters originally found Table I for the three cases of surface $(m\leqslant 4)$, all $(3\leqslant m\leqslant 14)$, and core $(m\geqslant 12)$ residues, in which $m=4$, 7 and 12 preferred directions could be discerned. The results for three comparable subsets obtained from Delaunay tessellation are displayed in parts (D)–(F). These are more heavily populated, in general. A common feature of the results in part (A)–(C) is that a substantial portion of the coordination space is unoccupied in the case of residues having low coordination numbers, and that this subspace is gradually filled as the number of neighbors increases. A careful examination also suggests that comparable orientations of coordination are selected in both sets, but the larger numbers of coordination vectors in the clusters (D)–(F) obscure this comparison. For a more transparent comparison, the clusters derived from Delaunay tetrahedra are mapped into simpler renditions (G)–(I), in which the sufficiently close and weakly populated sites are merged together in conformity with their statistical weights. This may be viewed as a coarse-grained renormalization, or smoothing out of the original distribution, to obtain fewer, but more probable, coordination directions. The mean angular deviation between the superimposed pairs of directional vectors turns out to be 6° for the panels (A) and (G) after this operation, 17° for the pairs of panels (B)–(H), and again 17° for the panels (C)–(I). This correspondence between the two sets demonstrates that the more detailed results from the tessellation method are compatible with those from the former analysis, provided that the distribution of coordination sites is viewed at a coarse-grained scale.

Is the apparent fit to fcc geometry simply the best discrete representation of random packing?

It is worth emphasizing that in the former optimal superimposition calculations about one-third of residues did not conform to *regular* packing geometry, but were diffusely, or randomly distributed in space. In a recent study of the distributions of free volumes in proteins using the Delaunay triangulation method, the free volume distributions in folded proteins are found to be liquidlike, or similar to glassy materials, although the packing densities are comparable to those of crystalline solids.[50] The interiors of proteins are concluded to be more like randomly packed spheres near their percolation threshold than like jigsaw puzzles. The coordination number of ~14 and the five-fold symmetry observed by Soyer *et al.*[37] were also shown to closely conform to the *random* packings of hard spheres. A tendency to pack as in an ideal icosahedral structure with dodecahedral cells was pointed out by Soyer *et al.* for the residues that are buried in the bulk of the protein,[37] which could be correlated with the present observations, given that the dodecahedral cells have 12 vertices and ensure the closest packing on a local scale

These observations lead us to consider more critically the origin of the observed regularities. Although the fcc-like coordination angles are enhanced by a factor of more than two over the random distributions, the aggregation of directional vectors along well-defined orientations could simply originate in well-packed bundles, rather than an actual preferred packing geometry.

In order to test this possibility we performed the following calculations: We consider four cells in a square lattice and assume that there is one neighbor in each of these four cells. The four coordination directions are thus selected such that each of the four neighboring cells are equally populated. This corresponds to the uniform case (see the following). For the random case, the four directions are randomly selected in space; they are not forced to point to different compartments. So, two cases are compared: (i) *uniform* (angular) distribution of coordinating residues in the neighborhood of a central residue so as to fill completely the space with an approximately constant density, and (ii) *random* distribution of coordinating residues. In the former case, the four directional vectors around the central site are assumed to occupy distinct quadrants (or lattice cells) in a 2-d space. This constraint may be viewed as a regularity imposed by an excluded volume effect. In the second case, no such restrictions apply. In both cases, the distribution of the coordination angles α is *uniform,* as shown in part (a) of Fig. 5, regardless of the presence or absence of volume exclusion. We generate 2000 bundles of four directional vectors, for each case, and these bundles are rigidly rotated so as to optimally superimpose their directional vectors. The resulting distributions of coordination angles for cases (i) and (ii) are displayed in the respective parts (b) and (c) of Fig. 5. The rms deviation

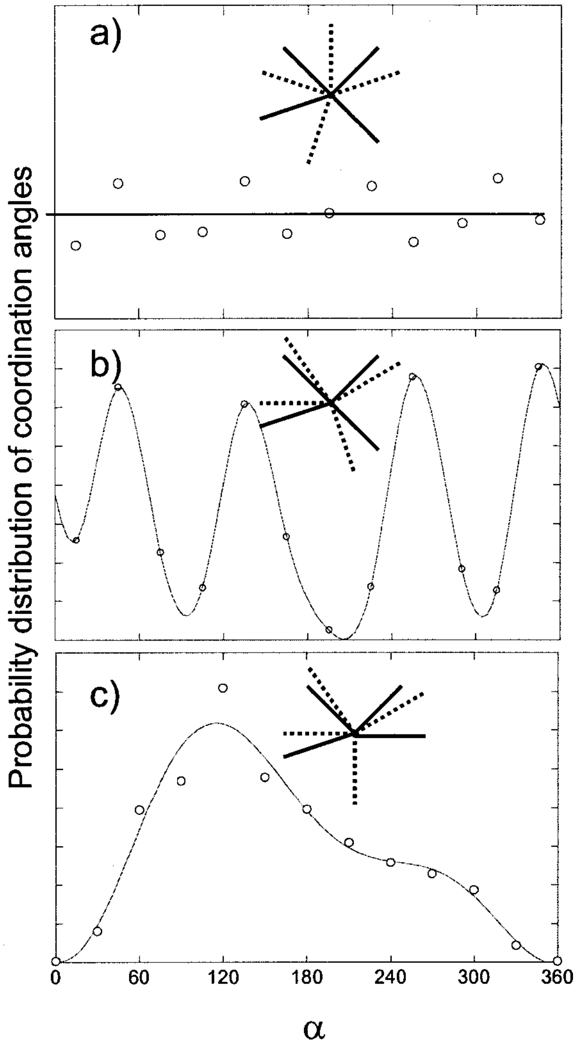

FIG. 5. Results from calculations performed for clusters of four directional vectors on a square lattice. Part (a) displays the original, uniform distribution of coordination directions, expressed in terms of the coordination angles α, with respect to a fixed frame. Parts (b) and (c) display the same distribution obtained after superimposition of 1,000 bundles of four directional vectors. The directional vectors are constrained to occupy distinct quadrants (a simple volume exclusion) in part (b), and are randomly oriented (no volume exclusion) in (c), which shows that the optimal superimposition of coordination sites that are uniformly sampling the coordination sites leads to four discrete, regular directions, whereas in the absence of a restriction to sparse uniformly, or in the absence of competition for space (or excluded volume), superimposition results in a Gaussian distribution.

between the clusters after 400 000 MC steps decreases from 0.48 to 0.28 for part (b) and decreases from 0.83 to 0.52 for part (c), for directional vectors of unit magnitude.

Although the original distribution of coordination angles is uniform, after optimal superimposition, the case with the excluded volume constraint leads to four discrete positions conforming to square lattice geometry, while we end up with a Gaussian distribution in the absence of any competition for space.

These results suggest that the observed preference of protein residues for fcc directions can likewise be a consequence of tight packing and excluded volume within the bundles in which the directional vectors have no actual net directional preferences apart from *uniformly* sampling (or parsing) the coordination space in the neighborhood of the central residue. In the other case of a totally *random* distribution of coordination angles, without excluded volume, a diffuse probability surface with a single peak is found (not shown) after optimally superimposing 1000 such bundles of coordination number $f=12$, as opposed to the surface with 12 peaks presently obtained for core residues.

V. SUMMARY

Just as semi-empirical potentials have a term included to reflect rotational isomers [typically $\cos(n\phi)$], empirical protein contact potentials could be constructed in a similar way to include the presently shown angular coordination with fcc packing geometry. By combining these with distance information, a generalized term can be obtained which will force collapse and compaction of the protein. Use of this distorted fcc lattice for simulations will require treatment of variable separations, reflecting the variable sizes and flexibilities of residues. Future extensions of these studies will likely offer insights into understanding sequence variability.

Notably the present considerations differ from the characterizations of atomic packing, which depend much more on the details of the structures and interatomic interactions.[34,50,51] In general, regularities in molecular materials appear to be manifested in orientations, not in distance distributions, i.e., bond lengths and packing distances may vary, but bond torsion angles are regular, and as shown here, the distribution of the coordination angles of residues can be discretized as sites conforming with the universal closest packing of identical spheres. A notable feature is that even regions with lower packing density choose among the same discrete sites, as if they are disposed to fill only the unoccupied sites if needed.

In conclusion, the present study provides evidence for the following generic properties of packing in proteins:

(i) For coarse-grained protein structures at the level of one point per residue, where only the space of the protein is considered, there is a relatively constant, or uniform, density of residues. Residues near the surface have a lower density only if the solvent-filled space immediately exterior were included. The high and fairly uniform packing density throughout the protein interior conforms closely to the conclusion reached in a recent analysis of 30 000 crystal structures on the atomic scale.[29]

(ii) In the case of the most densely packed interior regions, fcc packing emerges as the *optimal* solution for discretization of the uniform packing geometry. Thus, the view of universal closest packing of spheres is valid for the cores of proteins. Small distortions from perfect fcc geometry are observed, presumably imparted by size and shape differences among different types of amino acids. Yet, these perturbations are not strong enough to obscure the fact that the coordination directions tend to conform to the discrete sites on an "imperfect" fcc lattice (Fig. 1, lower diagrams), these sites being gradually filled, as more neighbors pack together.

(iii) The coarse graining at this level of one point per residue can be inferred to correspond to a homopolymeric chain where all residues are equivalent in their packing behavior, and pack approximately as spheres do. Filling discrete positions is likely a result of the hydrophobic effect that favors close packing among available residues even when a coordination shell is not fully populated, rather than the alternative of being dispersed more uniformly, but at lower density.

[1] B. I. Dahiyat and S. L. Mayo, Proc. Natl. Acad. Sci. U.S.A. **94**, 10172 (1997).
[2] O. B. Ptitsyn and K. L. Ting, J. Mol. Biol. **291**, 671 (1999).
[3] O. B. Ptitsyn, J. Mol. Biol. **278**, 655 (1998).
[4] A. Abe, R. L. Jernigan, and P. J. Flory, J. Am. Chem. Soc. **88**, 631 (1966).
[5] P. J. Flory, *Statistical Mechanics of Chain Molecules* (Interscience, New York, 1969).
[6] W. L. Mattice and U. W. Suter, *Conformational Theory of Large Molecules* (Wiley, New York, 1994).
[7] J. W. Ponders and F. M. Richards, J. Mol. Biol. **193**, 775 (1987).
[8] J. M. Thornton, *Protein Folding* (Freeman, New York, 1992), pp. 59–81.
[9] A. Fiser, R. K. Do, and A. Sali, Protein Sci. **9**, 1753 (2000).
[10] H. S. Chan and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **87**, 6388 (1990).
[11] A. Maritan, C. Michelletti, A. Trovato, and J. R. Banavar, Nature (London) **406**, 287 (2000).
[12] A. Stasiak and J. H. Maddocks, Nature (London) **406**, 251 (2000).
[13] J. Singh and J. M. Thornton, J. Mol. Biol. **211**, 595 (1990).
[14] G. Vriend and C. Sander, Proteins **11**, 52 (1991).
[15] J. Singh and J. M. Thornton, *Atlas of Protein Side-chain Interactions* (Oxford University Press, New York, 1992).
[16] G. Raghunathan and R. L. Jernigan, Protein Sci. **6**, 2072 (1997).
[17] F. M. Richards, Annu. Rev. Biophys. Bioeng. **6**, 151 (1977).
[18] S. Bromberg and K. A. Dill, Protein Sci. **3**, 997 (1994).
[19] W. A. Lim and R. T. Sauer, J. Mol. Biol. **219**, 359 (1991).
[20] B. W. Matthews, Adv. Protein Chem. **46**, 249 (1995).
[21] I. Bahar and R. L. Jernigan, Folding Des. **1**, 357 (1996).
[22] D. G. Covell and R. L. Jernigan, Biochemistry **29**, 3287 (1990).
[23] A. Godzik, A. Kolinski, and J. Skolnick, J. Comput. Chem. **14**, 1194 (1993).
[24] Z. Bagci, R. L. Jernigan, and I. Bahar, Polymer **43**, 451 (2002).
[25] H. M. Berman, J. Westbrook, Z. Freng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shinyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[26] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[27] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[28] I. Bahar and R. L. Jernigan, J. Mol. Biol. **266**, 195 (1997).
[29] J. Tsai, R. Taylor, C. Chothia, and M. Gerstein, J. Mol. Biol. **290**, 253 (1999).
[30] F. M. Richards, J. Mol. Biol. **82**, 1 (1974).
[31] J. L. Finney, J. Mol. Biol. **96**, 721 (1975).
[32] C. Chothia, Nature (London) **254**, 304 (1975).
[33] Y. Harpaz, M. Gerstein, and C. Chothia, Structure (London) **2**, 641 (1994).
[34] M. Gerstein, J. Tsai, and M. Levitt, J. Mol. Biol. **249**, 955 (1995).
[35] J. L. Finney, J. Mol. Biol. **119**, 415 (1978).
[36] C. W. David, Biopolymers **27**, 339 (1988).
[37] A. Soyer, J. Chomilier, J. P. Mornon, R. Jullien, and J. F. Sadoc, Phys. Rev. Lett. **85**, 3532 (2000).
[38] J. Liang, H. Edelsbrunner, P. Fu, P. V. Sudhakar, and S. Subramaniam, Proteins **33**, 1 (1998).
[39] H. Wako and T. Yamato, Protein Eng. **11**, 981 (1998).
[40] The polar and azimuthal angles can be expressed with reference to any coordinate system whose origin coincides with the central (superimposed) residues in the ensemble of clusters. For example, the seven peaks in Fig. 2 can be described in terms of different sets of (θ, ϕ) angles, depending on the choice of the orientation of the reference frame.
[41] I. Bahar, A. R. Atilgan, and B. Erman, Folding Des. **2**, 173 (1997).
[42] The coordination states for the case $3 \leqslant m \leqslant 14$ are identical to those displayed in Fig. 2, except for the reorientation of the reference frame, to capture the correlation between the residue coordination directions and the directional vectors of the fcc lattice.
[43] The last column in Table I refers to the total probability of the set of coordination states in each row, evaluated from the fraction of residues located within 20° deviation from the listed coordination centers. Numbers in parentheses represent their enhancement factor relative to random occupation.
[44] I. Bahar and R. L. Jernigan, *Perspectives in Structural Biology* (Indian Academy of Sciences, Hyderabad, 1999), pp. 209–225.
[45] R. F. Rapold and W. L. Mattice, Macromolecules **29**, 2457 (1996).
[46] P. Doruker and W. L. Mattice, Macromolecules **30**, 5520 (1997).
[47] B. Cipra, Science **281**, 1267 (1998).
[48] N. J. A. Sloane, Nature (London) **395**, 435 (1998).
[49] J. L. Sussman, M. Harel, F. Frolow, C. Oefner, A. Goldman, L. Toker, and I. Silman, Science **253**, 872 (1991).
[50] J. Liang and K. A. Dill, Biophys. J. **81**, 751 (2001).
[51] P. J. Fleming and F. M. Richards, J. Mol. Biol. **299**, 487 (2000).