ELSEVIER

Polymer 43 (2002) 451–459

polymer

www.elsevier.com/locate/polymer

Residue coordination in proteins conforms to the closest packing of spheres ☆

Zerrin Bagci[a], Robert L. Jernigan[c], Ivet Bahar[a,b,*]

[a]*Chemical Engineering Department and Polymer Research Center, Bogazici University, Istanbul, Turkey*
[b]*Center of Computational Biology & Bioinformatics, and Department of Molecular Genetics & Biochemistry, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213 USA*
[c]*Molecular Structure Section, Laboratory of Experimental and Computational Biology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892-5677, USA*

Abstract

Coarse-grained protein structures have the unusual property of manifesting a greater regularity than is evident when all atoms are considered. Here, we follow proteins at the level of one point per residue. We confirm that lattices with large coordination numbers provide better fits to protein structures. But, underlying these protein structures, there is an intrinsic geometry that closely resembles the face-centered-cubic (fcc) lattice, in so far as the coordination angles observed in clusters of near neighboring residues are concerned. While the fcc lattice has 12 neighbors, the coordination number about any given residue in a protein is usually smaller; however, the neighbors are not distributed in a uniform, less dense way, but rather in a clustered dense way, occupying positions closely approximating those of a distorted fcc packing. This packing geometry is a direct manifestation of the hydrophobic effect. Surprisingly, specific residues are clustered with similar angular geometry, whether on the interior or on the exterior of a protein. © 2001 Published by Elsevier Science Ltd.

Keywords: Packing in proteins; Face-centered-cubic lattice; Uniform packing density

1. Introduction

Proteins are compact linear molecules with some regularities in their polypeptide backbone. Foremost among these are two regular motifs commonly observed: α-helices and β-strands. Studies on the regularities of the polypeptide backbone began with the pioneering work of Ramachandran and coworkers [1]. The occurrence of α-helices and β-strands in native proteins could be attributed to the accessibility of the corresponding rotational angles seen in the Ramachandran maps. Yet, the calculated energies of a single residue for the rotational angles do not completely account for the high frequency of occurrence of these secondary structures in folded proteins. These secondary structures are indeed stabilized significantly by interactions other than those between the nearest neighbor bonds, which are usually considered on Ramachandran plots. They are stabilized by so-called higher order interactions. For example, α-helices are stabilized by the hydrogen bonds between the C=O and N–H groups of the respective residues i and $i + 4$, while sheet formation stabilizes the β-strands by the association of polar groups of segments even further along the sequence. Thus, *non-bonded* interactions do play a major role in stabilizing secondary structures.

The importance of non-bonded interactions in proteins goes far beyond the stabilization of secondary structures. Proteins are unique in that each sequence of amino acids selects one or sometimes a few three-dimensional structure and the non-bonded interactions are accepted to be the major determinant of the different folds taken by different amino acid sequences.

Thus, in contrast to the statistical analysis of polymers in which significant information on conformational behavior can be extracted from the rotational isomeric state distribution of individual (or pairwise dependent) bonds, analyzing proteins' conformational preferences necessitates a thorough understanding of the preferences for the non-bonded interactions of amino acids. Not surprisingly, a large number of computational and theoretical studies have been directed at characterizing the empirical potentials for

☆ This paper was originally submitted to *Computational and Theoretical Polymer Science* and received on 14 December 2000; received in revised form on 28 February 2001; accepted on 28 February 2001. Following the incorporation of *Computational and Theoretical Polymer Science* into *Polymer*, this paper was consequently accepted for publication in *Polymer*.

* Corresponding author. Address: School of Medicine, University of Pittsburgh, Kaufmann Building, Suite 601, 3471 Fifth Avenue, Pittsburgh, PA 15213, USA. Tel.: +1-412-648-6671/+90-212-2631540, ext. 2003; fax: +1-412-648-6676/+90-212-2575032.

E-mail addresses: bahar@pitt.edu (I. Bahar), bahar@prc.bme.boun.edu.tr (I. Bahar), bahar@indigo.bme.boun.edu.tr (I. Bahar).

0032-3862/02/$ - see front matter © 2001 Published by Elsevier Science Ltd.
PII: S0032-3861(01)00427-X

inter-residue interactions [2]. The Protein structure Data (PDB) [3] where the atomic coordinates of all proteins determined by X-ray crystallography or NMR spectroscopy are deposited has been exploited as an important source for extracting information on inter-residue potentials. The essential approach is to use the so-called inverse Boltzmann law, i.e. calculate effective potentials, or free energies, from the observed frequencies of amino acid pairs, assuming that the examined PDB structures contain a complete equilibrated set of residue pairs.

Although significant efforts have been devoted to modeling inter-residue interaction potentials as a function of their spatial separation, much less attention has been given to the orientational preferences of amino acids or to the coordination directions of amino acids. Examinations of such angular preferences revealed [4–9], on the other hand, some non-randomness, which could signal the occurrence of some generic preferences for the packing of residues in proteins.

Clearly, the existence of some regularity in residue packing would be of great utility for reducing the computational time and increasing the accuracy of conformational searches. Presently, methods of bioinformatics can predict secondary structure with up to 80% accuracy [10], and significant progress can be made in tertiary structure prediction by combining the tools for predicting secondary structure with those efficiently discriminating between alternative non-bonded interaction geometries.

A common approach for reducing the conformational space of macromolecules is to adopt coarse-grained lattice models. Two major computational advantages of lattice models are to discretize the conformational space, and to be amenable to integer algorithms. Lattice models not only provide an efficient means of generating conformations and perturbing them to investigate the dynamic characteristics, but also provide insights about the global behavior of molecular systems that could not be explored in atomic detail [11].

In the present work, the possible regularity in the association of non-bonded residues in folded proteins will be investigated. The representation of the coordination ofresidues in terms of different lattice geometries will be explored. It will be shown that a diversity of lattice geometries, including simple lattices, such as simple cubic (sc), can be adopted for a satisfactory description of the structure at a coarse-grained level. Yet, the optimal association of non-bonded residues, or the preferred coordination directions, will be shown to approximate the face-centered-cubic (fcc) geometry.

The fcc geometry has been only recently shown [12,13], rigorously, to be the closest packing geometry for identical spheres [14,15], i.e. after approximately 400 years since the original conjecture of Kepler. Interestingly, the same geometry, termed second nearest neighbor diamond (2nnd) lattice, has been recently proposed for a coarse-grained representation of polymer chains whose backbone bond angles are approximately tetrahedral [16–18]. This lattice may be viewed as consisting of directional vectors that connect every other site on a tetrahedral lattice, and it has indeed been utilized for locating every other backbone atom in polyethylene-like chains. Monte Carlo (MC) simulations of polyethylene melts performed on the 2nnd lattice and reverse mapping of the equilibrium structures to full atomic continuous space yielded several properties, including the cohesive energy densities, that were in close agreement with experimental data [19].

In the present coarse-grained study, the virtual bond model is adopted for modeling the backbone. The geometric features of this model conform to those of an fcc lattice, hence the special suitability of the fcc directions for describing the polypeptide backbone. However, the tendency of residues to be coordinated in conformity with the fcc geometry is stronger than the preference that would be imparted by the geometric suitability for the backbone alone. Furthermore, this tendency is discerned even when bonded neighbors are excluded. The fact that the fcc packing is the closest packing geometry for identical spheres brings into consideration the possible drive for maximizing packing density as a possible (origin for the appearance of the fcc geometry. Besides, even the formation of helices in proteins and DNA has been recently shown to be a consequence of optimal packing requirement [20–21]. Clearly, residue side chains differ in their size and shape. But at a coarse-grained scale where such differences are ignored, an intrinsic bias towards a universal geometry that provides the most efficient packing for identical spheres can be discerned, which suggests that the drive for optimizing packing density may be stronger than previously thought. It can be concluded that close packing, though indirectly, is a result of hydrophobic effects. The reason is that hydrophobic interactions are generally non-specific; they do not distinguish between different interactions geometries, and thus could conform to any geometry that would maximize the packing density. On the other hand, the interactions between polar (or hydrophilic) groups require specific association of functional groups, along specific directions, and do not necessarily conform to a uniform, symmetric packing geometry. The fact that a regular packing is approximated signals the importance of hydrophobic effects.

2. Model and method

A single-site-per-residue model is adopted for representing protein structures. The C^{β}-atoms of all amino acids are used to this aim, except for glycine residues for which the C^{α}-atom is utilized in the absence of a side chain. Clusters of residues are collected from 150 non-homologous PDB structures. Each cluster consists of a central residue, and its nearest (bonded and non-bonded) neighbors within a first coordination shell of 6.8 Å [22]. The coordination number of residues varies in the range $3 \leq m \leq 14$. A total of 28,730 clusters has been collected, and organized in 20 subsets, depending on the identity of the central amino acid.

Table 1
Coordination angles for different lattice geometries

sc												
	1	2	3	4	5	6						
θ (°)	45	45	90	90	135	135						
ϕ (°)	90	270	0	180	90	270						
bcc												
	1	2	3	4	5	6	7	8				
θ (°)	55	55	55	55	125	125	125	125				
ϕ (°)	45	135	225	315	45	135	225	315				
emb												
	1	2	3	4	5	6	7	8	9	10		
θ (°)	35	45	45	90	90	90	90	145	135	135		
ϕ (°)	0	90	270	0	125	180	235	0	90	270		
hcp												
	1	2	3	4	5	6	7	8	9	10	11	12
θ (°)	35	35	35	90	90	90	90	90	90	145	145	145
ϕ (°)	30	150	270	0	60	120	180	240	300	30	150	270
fcc												
	1	2	3	4	5	6	7	8	9	10	11	12
θ (°)	35	35	35	90	90	90	90	90	90	145	145	145
ϕ (°)	30	150	270	360	60	120	180	240	300	210	330	90

Here, we concentrate on the similarities between the coordination angles, rather than on the coordination distances. Each cluster is thus reduced to a bundle of unit directional vectors representing the angular position of residues within the first coordination sphere. The objective is to identify the regularities, if any, in the coordination directions of residues.

Five lattice representations have been considered in order to assess the regular geometry that best suits the databank clusters. These are simple cubic (sc), body-centered-cubic (bcc), face-centered-cubic (fcc), hexagonal close packed (hcp) and a hybrid lattice formed by embedding a tetrahedral lattice into a simple cubic lattice, shortly designated as 'emb'. The corresponding coordination angles are 6, 8, 12, 12, and 10, respectively. The directional vectors associated with these lattices, expressed in terms of two spherical angles (θ: polar, ϕ: azimuthal), are listed in Table 1.

The clusters of directional vectors extracted from the PDB are compared in each case with the directional vectors characteristic of the different lattices. Best fit of the clusters on the directional vectors of the target lattice requires rigidly rotating the clusters so as to minimize the deviations between closest pairs of directional vectors. An iterative MC algorithm is conveniently used to this aim. Each step consists of an incremental rotation of the overall cluster, which is accepted if the deviation between the two sets of directional vectors is reduced, and rejected otherwise. For a cluster composed of m directional vectors, being superimposed on a target lattice of coordination number z, there are $z!/[m!(z-m)!]$ possible combinations for pairing the directional vectors. Each combination has been considered for determining the optimal set of superimposed pairs. The quality of the fit to the target lattice is assessed from the mean distance between the tips of the closest pairs of directional vectors.

In addition to the above procedure used for testing the suitability of target lattices, the occurrence of *off-lattice coordination states* has been explored. To this aim, the clusters were superimposed disregarding any predefined regular architecture. This so-called *optimal superimposition* scheme is significantly more expensive than the fit to a target lattice: the application to a set of $N = 1000$ clusters, for example, requires about 50 h (real time). At each step, a randomly selected cluster is rotated, and its root-mean-square (RMS) deviation with respect to the remaining $N - 1$ clusters is tested. The move is accepted if the RMS deviation is decreased. Results are found to converge by performing runs for a total of 3×10^6 steps. Usually, a set of 1000 clusters is verified to be large enough to obtain statistically reliable results, while being small enough to be superimposed within a reasonable computational time.

3. Results

3.1. Fit to target lattices

As a first step, we tested the suitability of the fcc geometry, which was recently pointed out to closely fit the coordination architecture of residues [23]. To this aim, the quality of the fit achieved by forcing the PDB clusters of directional vectors to match the directional vectors of the fcc lattice has been examined. The results displayed in Fig. 1 demonstrate that the angular positions of residues observed at the scale of a single site per residue can be

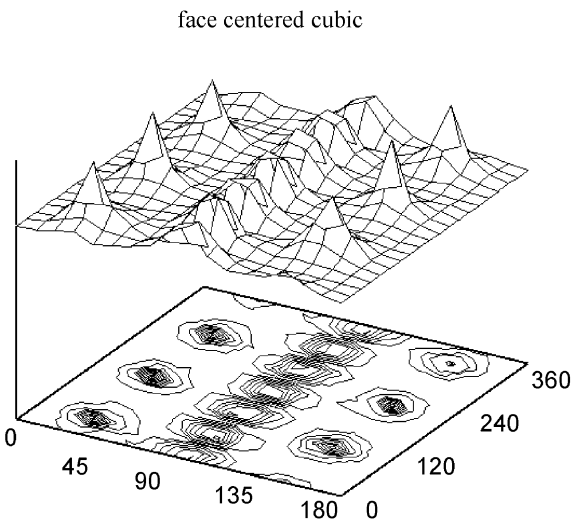

Fig. 1. Distribution of coordination angles for a representative set of 1000 residue bundles from the PDB, which have been suitably rotated to fit the fcc lattice geometry. The axes refer to the polar (θ) and azimuthal (ϕ) angles (in degrees) and the surface represents the probability of observing a first neighbor along a given direction.

construed — by suitable rigid body rotation of the overall clusters — to fit those of the fcc geometry. The surface in Fig. 1 represents the probability distribution of the spherical angles assumed by the directional vectors of the PDB clusters at the end of these constrained fit algorithms. The peaks exactly coincide with the fcc lattice coordination angles.

Interestingly, calculations repeated for the other four target lattices yielded results lending support to the applicability of other lattices, as well. See the distributions presented in Fig. 2 for the sc, bcc, hcp and emb lattice geometries tested here. In all cases, peaks are obtainable at the angular positions of the directional vectors associated with the different lattices.

3.2. Quality of the fit to target lattices

Figs. 1 and 2 suggest that different regular geometries can be adopted for modeling the packing architecture of residues in proteins. A quantitative measure of their level of accuracy is the RMS deviation between the actual clusters' directional vectors and those of the target lattices. In Fig. 3(a), the decrease in the RMS deviations as a function of the number of MC steps is displayed for all five types of target lattices. The RMS deviations decrease to 0.20, 0.21, 0.25, 0.30, 0.37 at 3×10^6 MC steps, for the hcp, fcc, emb, bcc, sc cells, respectively. The curves are displayed up to 2×10^6 MC steps. These are normalized values, i.e. values are expressed relative to lattice edges of unit size. The RMS deviations decrease with the increasing coordination

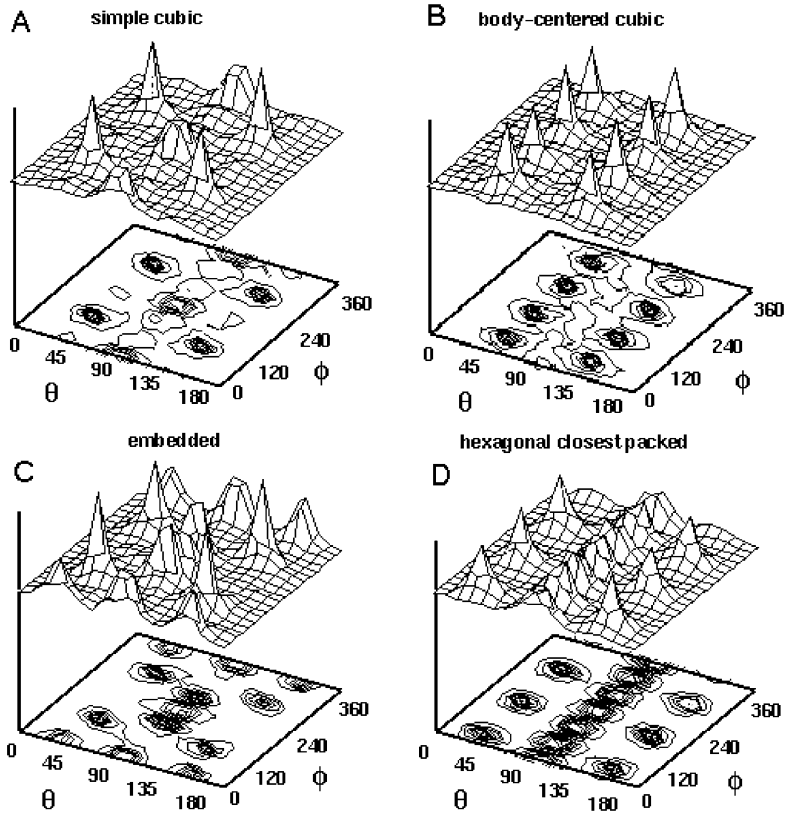

Fig. 2. Same as Fig. 1, for four alternative lattice geometries. The 3-D view of the probability distribution is displayed for (A) simple-cubic (sc), (B) body-centered-cubic (bcc), (C) tetrahedral embedded in simple cubic (emb), and (D) hexagonal close packed (hcp). See Table 1 for the coordination angles characterizing these lattice geometries. These coincide with the peaks in the respective distributions.

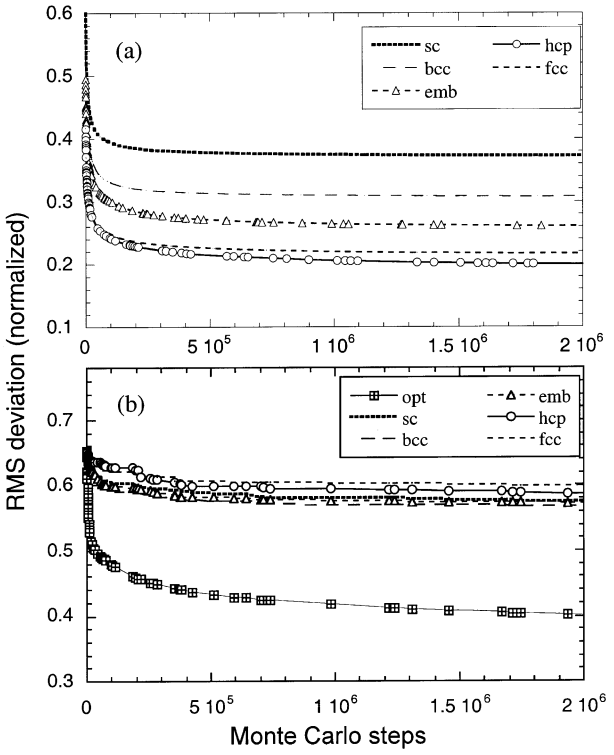

Fig. 3. RMS deviations between the PDB clusters and the target lattices as a function of MC steps are shown in (a). (b) depicts the accompanying decrease in the RMS deviation between the clusters themselves. The lowermost curve, labeled as 'opt' refer to the results from optimal superimposition of the clusters, where no target lattice is imposed.

numbers of the target lattice. This is expected since with larger z values, there are more choices for allocating the m vectors of the clusters, and a broader coverage of the coordination space.

Improvement in the suitability of the lattice with increase in its coordination number is indeed the common observation [24,25].

It is important to notice that in the present analysis, the clusters are forced to maximally match the directional vectors of the target sets, rather than being optimally superimposed on themselves. If, on the other hand, the deviations between the clusters of directional vectors (that have been rotated to fit the target lattices) are examined, significantly higher RMS deviations are observed. The reason is the allocation of the clusters' individual vectors to any possible choice of the target lattice directions. For example, two different clusters of $m = 6$ directional vectors can occupy two exclusive sets of coordination directions among the 12 accessible directions of the fcc lattice; therefore, the two clusters would not be superimposed, although they perfectly match the lattice geometry. The RMS deviations between the clusters that fit the target lattices are presented in Fig. 3(b). These are observed to decrease to 0.58 ± 0.02, starting from original values of about 0.65. The lowermost curve, on the other hand, represents the result from the optimal superimposition of the clusters themselves (see below), not imposing any target lattice assignment. A significantly lower RMS deviation (0.39) is achieved in this latter case.

3.3. Threading results

The fit to target lattices has been tested above for a single coordination sphere around a central residue. Although a given lattice representation can adequately fit the coordination geometry at the level of a single coordination sphere, it may become less adequate when an overall structure is being fitted because of the geometric constraints at the juxtaposition of successive lattice sites in space. Furthermore, in compact structures such as proteins, not all coordination sites are accessible, due to their occupancy by other chain segments. In view of these limitations, we performed threading calculations following the method originally proposed by Covell and Jernigan [26]. In this method, PDB structures are threaded onto a predefined lattice, such that each residue occupied a lattice site. The level of accuracy of the mappings to the five different lattice geometries will be evaluated from the RMS deviations between the original structures and their on-lattice representations.

Fig. 4 illustrates the results for two example proteins. Fig. 4(A) displays the PDB structure (lighter) and the best fitting (fcc) lattice representation (darker) for an α-protein, myoglobin (PDB code: 1mba [27]). Fig. 4(B) displays the X-ray and lattice structures for a β-protein, plastocyanin (PDB code: 1plc [28]). The respective RMS deviations between the X-ray and lattice structures are 2.04 and 2.29 Å, respectively, in Fig. 4(A) and (B).

For a systematic study of the accuracy of lattice representations, calculations have been performed for four different structural classes of proteins, known as α-, β-, α + β,- and α/β-classes [29]. Five test proteins have been selected from each class. These are high resolution (<2.5 Å) X-ray structures, given that lower resolution structure, could obscure the results. The C^{α}–C^{α} virtual bond representation has been adopted for threading the proteins. The threaded chains are self-avoiding, i.e. no two C^{α} atoms occupy the same lattice site. The RMS deviations between the original positions of the α-carbons and their approximate on-lattice positions are listed in Table 2. The reported values are the deviations in Å, the lattice edge being taken as 3.8 Å (i.e. the length of C^{α}–C^{α} virtual bonds) for the sc, bcc, fcc and hcp lattices. In the emb lattice, there are two lattice lengths, given by 1 and $3^{1/2}/2$, of which each was multiplied by 3.8 Å.

A summary of the results from threading calculations is presented in Table 3. Lattices having higher coordination numbers are again observed to yield closer representations of the original structure. However, there are important differences between the levels of accuracy achieved for different classes of proteins. β-proteins and those belonging to α + β- and α/β-classes are harder, in general, to fit onto sc, and bcc lattices compared to α-proteins. This deficiency

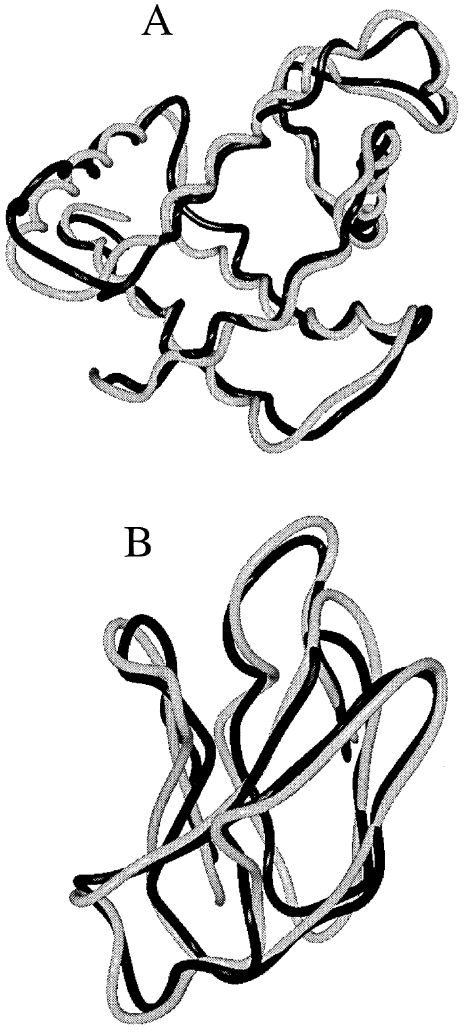

Fig. 4. Threading results for two example proteins: (A) myoglobin and (B) plastocyanin. The backbone is displayed in each case. The lighter diagram is the X-ray structure, and the darker is its on-lattice representation, after threading the C^{α}-atoms onto an fcc lattice. The ribbon diagram using the Midas package is displayed in both cases for visual clarity.

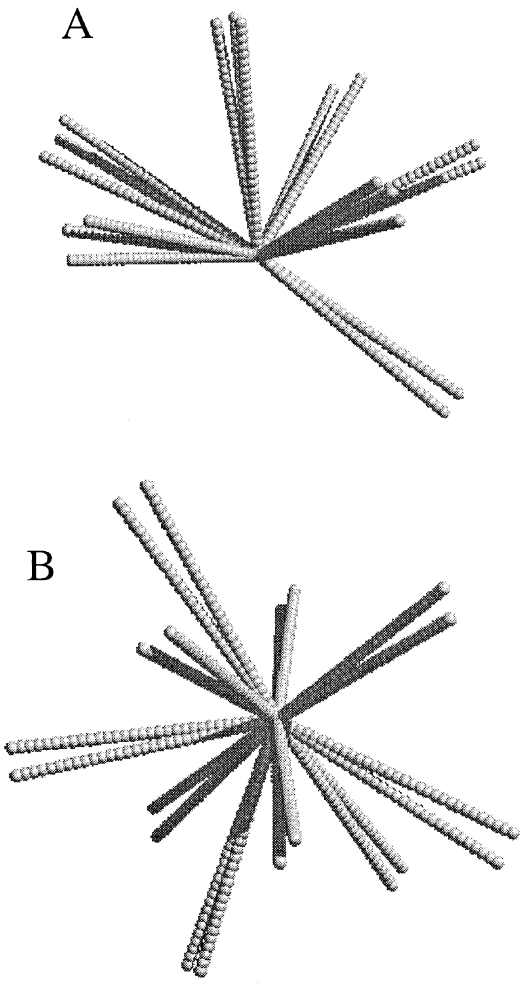

Fig. 5. Schematic illustration of the superimposition of clusters of directional vectors: (A) for three clusters of coordination numbers 6–7, and (B) two clusters of coordination number $m = 10$.

can be partly attributed to the C^{α}–C^{α} virtual bond angle of 120° in β-strands, as opposed to its value of 90° in α-helices [30]. The former can be readily accommodated by the fcc and hcp lattice cells (see Table 1); whereas, the sc, and bcc cells do not comply with 120° bond angles, hence the relatively high RMS deviations observed for β-, α + β- and α/β-proteins in the sc, and bcc cases (Table 3). Notice that the highest RMS deviations take place in the case of β-proteins threaded onto sc lattices, which can be understood from the fact that the sc lattice does not contain any coordination angle other than 90°.

3.4. Optimal superimposition of the clusters

Identification of the most probable coordination geometry of residues requires optimal superimposition of clusters onto themselves. Fig. 5(A) illustrates the superposition of three clusters as an example. Note that the clusters in the figure differ in their coordination numbers; two of them have seven neighbors, and the third has four. Fig. 5(B) illustrates the superimposition of two clusters of coordination number 10 each.

The clusters shown in Fig. 5 actually display the most highly populated coordination directions obtained upon optimal superimposition of cluster, for all types of clusters (Fig. 5(A)), irrespective of amino acid identity and coordination number, and clusters having coordination number $m \geq 10$, (Fig. 5(B)) usually located in the core of proteins. The former will be referred to as the generic coordination of all residues, and the latter as the generic coordination directions of 'core' residues.

A quantitative distribution of the generic distributions of coordination directions for 'all' and 'core' residues is given in the respective maps of Fig. 6(A) and (B). The maps represent the probability distribution of the spherical coordination angles θ and ϕ for the two cases. As mentioned above, sets of 1000 clusters with required coordination numbers, randomly selected form the PDB, yield

Table 2
Threading results (RMS deviations in units of Å) for PDB structures belonging to four different structural classes (the results in Ref. [24] for the same analysis are shown in parentheses)

PDB code	Resolution (Å)	Size	sc	bcc	emb	fcc	hcp
α-proteins							
1AVHA	2.3	318	3.02	2.54	2.52	2.00	2.03
1LE4	2.5	139	2.81	2.25	2.21	1.96	1.93
1MBA	1.6	146	2.77 (2.7)	2.44 (2.4)	2.40	2.04 (1.9)	2.11
1MBC	1.5	153	2.82	2.44	2.37	1.97	2.06
2LH1	2.0	153	2.93	2.47	2.41	2.07	2.07
β-proteins							
1HILA	2.0	217	5.07	3.51	2.51	2.28	2.48
1MAMH	2.45	217	5.01	3.46	2.79	2.31	2.54
1PLC	1.33	99	4.31 (4.3)	3.99 (2.9)	2.60	2.29 (3.3)	2.20
2AYH	1.6	214	4.35	4.05	2.83	2.62	2.27
8FABA	1.8	206	5.63	3.91	2.95	2.22	2.49
α + β proteins							
1DNKA	2.3	250	3.48	4.02	2.68	2.47	2.44
1PPN	1.6	212	3.82	2.79	2.53	2.16	2.19
2AAK	2.4	150	3.39	3.73	2.44	2.46	1.96
2ACT	1.7	218	3.90	3.01	2.76	2.24	2.34
4BLMA	2.0	256	3.99	2.78	2.38	2.19	2.14
α/β-proteins							
1DHR	2.3	236	3.43	3.23	2.78	2.37	2.16
1FX1	2.0	147	3.39	3.73	2.49	2.37	2.36
1OFV	1.7	169	3.60	3.07	2.42	2.15	2.25
2DRI	1.6	271	3.73	3.08	2.67	2.18	2.15
2YPIA	2.5	247	3.85 (3.3)	2.60 (2.8)	2.71	2.20 (2.0)	2.21

reproducible distributions. The lowermost curve in Fig. 3(b) shows the significant decrease in the RMS deviation obtained by MC iterations for 'all' clusters.

The centers of coordination obtained for the two groups of clusters are listed below:

Coordination state		1	2	3	4	5	6	7	8	9	10	11	12
$(3 \le m \le 14)$	θ	40	35	45	95	105	55	90[a]					120
	ϕ	10	200	285	350	50	115	180[a]					115
$m \ge 10$	θ	45	45	45	95	105	60	100	85	105	140		
	ϕ	30	170	270	350	50	90	130	230	290	210		

[a] This coordination state emerged from the examination of the packing geometry of specific residues (unpublished data).

On the other hand, repeating the calculations for the densest clusters ($m \ge 12$), the following coordination angles are found:

$m \ge 12$	θ	45	25	50	70	100	75	80	75	105	140	145	130
	ϕ	60	170	280	340	40	120	160	220	260	200	330	120

which can be compared to the directional angles of the fcc lattice listed in the last row of Table 1.

Examination of the above sets of optimal directional angles for the clusters of different coordination numbers reveals that the coordination angles approximate those of the fcc geometry. In the former case, only seven sites are occupied, whereas in the latter case, all sites are occupied. As a further test of the validity of an fcc geometry that is gradually filled with increasing coordination numbers, we investigated the coordination angles of surface residues ($3 \le m \le 4$). The values

$(3 \le m \le 4)$	θ	40	45	95	90	
	ϕ	30	170	50	110	

were found as the most highly populated coordination angles. It might be thought that the 12 degrees of freedom of an fcc lattice could simply afford sufficient flexibility to realize a good fit. However, the four sites mentioned above for surface residues are also arranged in a condensed way, in

Table 3
Average RMS deviations (Å) between databank structures and their models threaded onto different lattices, for four different classes of proteins

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

conformity with close sites on an fcc lattice. In other words, not only do they approximate four of the lattice directions of the fcc packing, but also they occupy nearby fcc directions, instead of being distributed sparsely over space. Likewise, the sites identified for all residues ($3 < m < 14$) are also clustered on a restricted coordination subspace. These observations indicate a behavior substantially different from merely fitting the structure to the fcc coordination geometry.

4. Discussion

The fit of every other methylene group in polyethylene-like chains to an fcc lattice is understandable in view of the fact that the successive bonds exactly conform to a tetrahedral geometry. The fcc lattice can indeed be viewed as a 'second level' of tetrahedral lattice. The 12 coordination vectors of the fcc lattice simply connect a given central site to its second neighbors along a tetrahedral lattice. Likewise, it has been possible to extend the same approach to the second generation of mapping/reverse mapping [31]. In this case, every fourth carbon atom along the chain is utilized, and about 1/12 of the lattice sites are occupied.

In the case of the polypeptide backbone, on the other hand, the accord with the fcc lattice can be explained by the geometric features of the virtual bond representation. First, all virtual bonds have fixed (3.8 Å) length, conforming to a fixed lattice dimension. Secondly, the angle between two successive lattice sites is either around 90° (α-helices) or 120° (β-strands), in conformity with the lattice geometry [30], and the most probable torsional angles can also be approximated by the choices accessible on an fcc lattice [26]. Whereas in the Ramachandran plots, only a single residue is examined at a time, other analyses capture regularities over several residues on the main chain. Pal and Chakrabarti [32] proposed a graphical representation for protein main chain and side chain torsional angles, which can aid in identifying backbone regularities. Also efficient analyses have been proposed by several groups [30,33,34] in which virtual C^{α}–C^{α} bonds and pseudodihedrals characterize backbone structure. These studies emphasize the regularities induced by the physicochemical nature of the polypeptide backbone.

However, what is unexpected is to see is that the fcc geometry is also the best fitting regular geometry approximating the coordination architecture of residues in folded proteins. The origin of this generic preference is probably fundamentally different from that of the polypeptide backbone. In this case, the non-bonded interactions are regular to some extent. The fact that this regularity coincides with the one that maximizes the packing density of identical spheres suggests that the tendency to assume an fcc geometry originates in the drive to optimize packing density.

It is worth recalling that the presently emphasized similarity to an fcc geometry resides in the coordination angles. The coordination distances, on the other hand, were variable. These would vary in the range $3.8 \leq r \leq 6.8$ Å, the lower bound being equal to the length of virtual bonds, and the upper bound being the cutoff distance that includes the first neighbors around a central amino acid, when viewed at the level of a single site per residue. It is clear that amino acids differ in their size and shape, but the present coarse-grained view neglects these differences, and focuses on similarities in coordination directions.

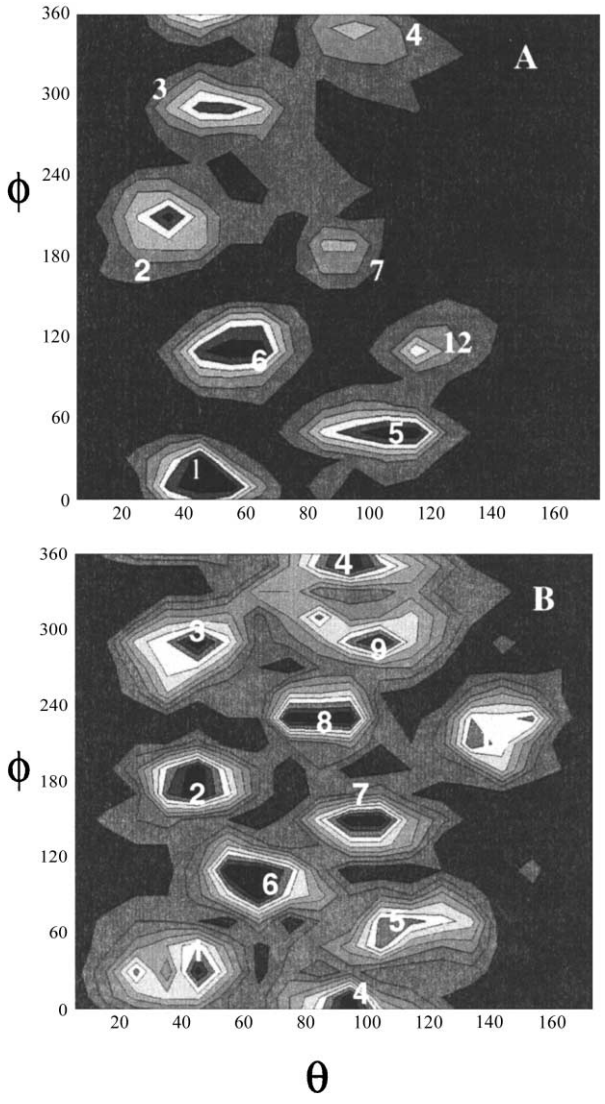

Fig. 6. Results for optimal superimposition of clusters, for two groups of clusters differing in their coordination numbers: (A) randomly collected 1000 clusters have been analyzed to extract the generic coordination angles for all residues, and (B) 300 clusters having high coordination number ($m \geq 10$) have been used to extract the optimal coordination geometry of core residues.

Significant variations were recently reported at the level of atomic packing in proteins, which were attributed to a complex combination of protein size, secondary structure composition and amino acid composition, and such differences have been proposed to serve as a measure of structural

homology studies [35]. Such differences are not discernible, however, at the presently adopted coarse-grained scale. In order terms, even the surface residues whose (residue) coordination number ($m = 3$–4) is significantly lower than that of core residues ($m \geq 10$) are closely packed in conformity with the accessible sites on the same fcc geometry. Our results conform to the remarks of Chothia and coworkers [36] on the occurrence of standard radii and volumes for residues in folded proteins.

One might wonder if the apparent regular geometry could be relevant to the crystallographic symmetry group or refinement methods in determining residue coordinates. However, such methods should not necessarily introduce a bias towards the fcc coordination directions, in particular, because the structures as a whole are fitted to different regular geometries, rather than the coordinates of the individual residues.

The emergence of helical motifs in proteins was shown [37] to be the result of evolutionary pressure for selecting structures having a high degree of thermodynamic stability, which can accommodate amino acid sequences that fold reproducibly and rapidly. Helices satisfy such optimal packing constraints [20]. The regular packing geometry that can be traced here at a coarse-grained scale can tolerate substitutions while optimizing residue packing. It remains to be examined that the effect of evolution on packing geometry by further studies of the coordination geometry in evolutionarily related proteins. For example, evolutionarily distant and close proteins can be directly examined to see the effects of evolution. Such systematic analyses should now be feasible with the increased availability of human or other species' genomic data.

As a further study, the extent of the lattice-like regularity might be expected to be more pronounced with increasing contents of secondary structures in specific proteins. We have not performed a direct analysis of coordination geometry in the neighborhood of residues belonging to α-helices or β-sheets. This could be a future extension of the present analysis. Whereas our threading calculations show that the fcc lattice can fit equally well different types of structural classes (Table 3), the improvement with increasing contents of secondary structure, which is a more detailed study, could be explored further.

Acknowledgements

Partial support by Bogazici University Research Funds (Project no. 00HA503) and useful discussions with P. Doruker and B. Ozkan are gratefully acknowledged.

References

[1] Ramachandran G, Ramakrishnan C, Sasisekharan V. J Mol Biol 1963;7:95–99.
[2] Jernigan RL, Bahar I. Curr Opin Struct Biol 1996;6:195–209.
[3] Bernstein FC, Koetzle TF, Williams GJB, Meyer EFJ, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. J Mol Biol 1977;112:535–42.
[4] Singh J, Thornton JM. J Mol Biol 1990;211:595–615.
[5] Vriend G, Sander C. J Appl Crystallogr 1993;26:47–60.
[6] Bahar I, Jernigan RL. Fold Des 1996;1:357–70.
[7] Singh RK, Tropsha A, Vaisman II. J Comp Biol 1996;3:213–22.
[8] Mitchell JBO, Laskowski RA, Thornton JM. Proteins 1997;29:370–80.
[9] Keskin O, Bahar I. Fold Des 1998;3:469–79.
[10] Petersen TN, Lundegaard C, Nielsen M, Bohr H, Bohr J, Brunak S, Gippert GP, Lund O. Proteins 2000;41:17–20.
[11] Dill KA, Bromberg S, Yue KZ, Fiebig KM, Yee DP, Thomas PD, Chan HS. Protein Sci 1995;4:561–602.
[12] Hales TC. Discr Comp Geom 1997;17:1–51.
[13] Hales TC. Discr Comp Geom 1997;18:135–49.
[14] Cipra B. Science 1998;281:1267.
[15] Sloane NJA. Nature 1998;395:435–6.
[16] Rapold RF, Mattice WL. J Chem Soc Faraday Trans 1995;91:2435–41.
[17] Rapold RF, Mattice WL. Macromolecules 1996;29:2457–66.
[18] Doruker P, Rapold RF, Mattice WL. J Chem Phys 1996;104:8742–9.
[19] Doruker P, Mattice WL. Macromolecules 1997;30:5520–6.
[20] Maritan A, Michelleti C, Trovato A, Banavar JR. Nature 2000;406:287–90.
[21] Stasiak A, Maddocks JH. Nature 2000;406:251–3.
[22] Bahar I, Jernigan RL. J Mol Biol 1997;266:195–214.
[23] Raghunathan G, Jernigan RL. Protein Sci 1997;6:2072–83.
[24] Godzik A, Kolinski A, Skolnick J. J Comp Chem 1993;14:1194–202.
[25] Park BH, Levitt M. J Mol Biol 1995;249:493–507.
[26] Covell DG, Jernigan RL. Biochemistry 1990;29:3287–94.
[27] Bolognesi M, Onesti S, Gatti G, Coda A, Ascenzi P, Brunori M. J Mol Biol 1989;205:529–44.
[28] Guss JM, Bartunik HD, Freeman HC. Acta Crystallogr 1992;48:790–811.
[29] Branden C, Tooze J. Introduction to protein structure. 2nd ed. New York: Garland Publishing, 1999.
[30] Bahar I, Kaplan M, Jernigan RL. Proteins 1997;29:292–308.
[31] Doruker P, Mattice WL. Macromol Theory Simul 1999;8:463–78.
[32] Pal D, Chakrabarti P. Protein Engng 1999;12:523–6.
[33] Oldfield TJ, Hubbard RE. Proteins 1994;18:324–37.
[34] DeWitte RS, Shakhnovich EI. Protein Sci 1994;3:1570–81.
[35] Fleming PJ, Richards FM. J Mol Biol 2000;299:487–98.
[36] Tsai J, Taylor R, Chothia C, Gerstein M. J Mol Biol 1999;290:253–66.
[37] Micheletti C, Banavar JR, Maritan A, Seno F. Phys Rev Lett 1999;82:3372–5.