PROTEINS: Structure, Function, and Genetics 53:252–261 (2003)

Global Secondary Structure Packing Angle Bias in Proteins

Daniel E. Platt,[1] Concettina Guerra,[2] Giuseppe Zanotti,[3] and Isidore Rigoutsos[1]
[1]*Bioinformatics and Pattern Discovery, IBM T. J. Watson Research Center, Yorktown Heights, New York*
[2]*Dept. Information Engineering, University of Padova, Padova, Italy*
[3]*Dip. Chimica Organica, Universitá di Padova, Padova, Italy*

ABSTRACT **While studies of secondary structure interactions have focused on local interacting features, there is a need for a more global characterization of packing-induced aligned packing of secondary structures. This study presents an analysis of the distribution of globally sampled secondary structures within selected subunits of a selected set of multimeric proteins. Comparisons are made between the distribution of the cosines of angles between triplets of linear segments associated to secondary structures and a theoretically obtained distribution for triplets of random uniformly distributed unit vectors. We show that, among all pairs of helix or strand segments, planar configurations appear more frequently than expected for uniformly distributed vectors, and alignment is strongly preferred compared to that expected for uniformly distributed vector triplets. Among all secondary structure triplets, pairs of angle cosines between helix strand segments deviate from uniformity corresponding to alignment and anti-alignment. Furthermore, among all helix or strand segments, including non-interacting secondary structures, the distribution of a single angle cosine indicates a strong preference for alignment and anti-alignment. Selection for interactive triplets shows results consistent with prior studies. Lastly, angle pairs are not statistically independent, indicating that alignment between two helix or strand segments is more likely if another helix or strand is aligned with either of the first two helices or strands. Selection for interactive segment triplets shows results consistent with prior studies. Proteins 2003; 53:252–261.** © 2003 Wiley-Liss, Inc.

INTRODUCTION

Previous studies[1–12] have definitively shown that "interacting" helices demonstrate preferred packing angles, and have considered a wide range of models to understand the mechanism of that preference. Packing angle is measured between linear segments associated to secondary structures, such as the axis of an α-helix or the endpoints of a β-sheet strand. Interaction between α-helices has been modeled in a number of studies in some detail. For the purposes of statistical study, several criteria have been proposed to select candidates for interacting secondary structures. Bowie[9] applied a criterion that the distance of closest approach between segments must be within a certain cutoff value. Walther et al.[10] reinforced the notion that the points on each segment of minimum distance between a pair of secondary structure segments must occur within the bounds of the secondary structure "elements," defined here as α-helices or β-sheet strands. At the nearest points on each axis, the line segment joining those points of closest approach is perpendicular to both segments. Therefore, the helices could be unrolled to lie flat on each other at these special points, increasing the possibility of interactions. In this study, while the focus is on secondary structures that are *not* interacting, we have considered the rolling criterion by itself, as well as a maximum distance cutoff to select "interactive" secondary structures (structures that are candidates for interaction). Since we measure triplets of secondary structure elements, the criterion that all three must be in direct contact leaves very few sample points. Relaxation of the distance cutoff constraint includes alignments between helices that may be in barrels, but not directly interacting.

However, while many previous studies have focused on the interactions between individual secondary structure elements, a study of the global interaction between multiple secondary structure elements remains relatively undeveloped. Bowie's study[9] introduced a correction to the way that packing angle distributions were normalized essentially to account for the distribution of area on a sphere. He tested this notion by comparing his geometrically normalized distributions for noninteracting secondary structures with the distribution obtained for uniformly distributed unit vectors, with the assumption that equality would validate the geometric normalization of the packing angle distributions.

This study considers the problem of detecting global bias on packing. Obviously, interactions will constrain the packing of secondary structure elements. However, it is legitimate and useful in a number of circumstances, such as the construction of scoring functions for protein structure prediction, or for the characterization of hash keys based on packing angles, to ask whether such interactions, as well as more global characteristics of protein structure, impose nonlocal packing constraints on secondary structures throughout the protein, and to seek ways to charac-

Grant sponsor: Italian Ministry of University and Research.

*Correspondence to: Daniel E. Platt, Bioinformatics and Pattern Discovery, IBM T. J. Watson Research Center, Yorktown Heights, New York. E-mail: watplatt@us.ibm.com

Received 10 May 2002; Accepted 17 January 2003

terize some of those global packing effects. This study considers the question by measuring the distribution of packing angles of secondary structure segments that are not necessarily directly interacting, and then testing for deviations from that which would be expected of a randomly distributed set of vectors. Therefore, both near-neighbor interactive secondary structures are considered as well as noninteracting secondary structure elements. Further, the relationship of multiple secondary structure segments is presented. The joint distributions of angles between triplets of secondary structure segments are measured, tested for deviation from that expected for random vectors, and tests for statistical dependence or interaction between the angles are performed. This addresses the question of whether it is more likely that one axis is aligned with a second axis if a third axis is aligned with that second axis, or alternatively, than would be expected for isolated pairs of segments.

The following behavior was observed. First, among all triplets of secondary structures, planar configurations of the triangles formed by three unit vectors are preferred more than expected for random vectors implying that triplets of secondary structures tend to lie on parallel planes. Moreover, among all such triplets the greatest concentration corresponds to alignment or anti-alignment, with anti-alignment being preferred. Second, there is an interaction between the angles, suggesting that two segments are more likely to be aligned if one of the segments has an alignment with a third axis. Third, angular preferences induced by detailed structural interactions are not visible to the distribution computed from all interactive and noninteracting triplets selected from within individual polymeric protein subunits due to the large number of noninteracting triplets compared to interactive triplets. The structural angular preferences only become visible when distributions for angles between locally interactive helices are selected. No structural preferences seem to appear for strand-strand interactions.

METHODS AND DATA

In this study, proteins were extracted from the PDB.[13] The proteins extracted from the PDB were selected as representatives of the SCOP[14] fold classes from the SCOP classification. The reason this was done was to select an unbiased set of proteins, or more accurately, a set of proteins with controlled bias that sample the universe of folds. This was compared with the results from Fischer classes,[15] with no significant differences in results. Since transmembrane proteins are underrepresented in the PDB due to the difficulty of crystallization required for X-ray crystallographic structure determination, there is a bias in the constructed sample selected in this study. The secondary structures were identified from the annotations in the PDB. PDB entries showing alternate coordinate sets were excluded. Single subunits were selected from each protein. This process yielded 331 subunits. A singular-value decomposition (SVD) alignment routine[16,17,18] was used to align a standard helix with known axis to sampled α-helices, using only C_α atom correspondences, and to each strand of the β-sheets.[19] Helices with fewer than 4 amino acid bases were discarded. A subunit so selected is shown in Figure 1, with its decomposition into segments shown in Figure 2.

Ordered triplets of secondary structure elements were selected sequentially, ranked according to the basis number of the start of the secondary structure element. No triplets were selected from triplets coming from more than one chain or subunit. The axis unit vectors $\hat{u}_1$, $\hat{u}_2$, and $\hat{u}_3$ were computed for the triplet. Cosines of the angles between triplets of the fitted axes were computed.

The distribution of cosine triplets was compared to the distribution expected for randomly distributed triplets of vectors. Distributions were computed for the following cases: first, all triplets including interactive and noninteractive axes, yielding 1,237,597 triplets; and second, interactive structure axes only, yielding 12,645 triplets. Distributions were computed for segments that were also broken down by secondary structure classifications. Triplets constructed from all-helix units were labeled "HHH." Triplets constructed from helix-helix-strand were labeled "HHS," and so on.

It is important to note that the interaction constraint imposed on the selection of triplets produces a bias. First, only triplets where each member is interactive with both of the others is considered. Application of a distance cutoff excludes α-helix barrels. Second, marginal distributions will only be counted from pairs interactive with some third secondary structure element. Any pair that interacts with larger numbers of third structure elements will be weighted more heavily than those that interact with smaller numbers. Thus, lone helix pairs will not be noticed, whereas barrels and sheets will show preference. This method selects direct contact helices in barrels by virtue of shared constraints, but also includes counts for angle preferences between barrel members that have no direct interaction.

Previous studies have definitively shown that interacting helices demonstrate preferred packing angle characteristics.[1–10] "Interacting" has been taken to mean that the point of minimum distance between secondary structure axes occurs within the bounds of the secondary structure elements.[5,10] The following describes the criteria for selecting interactive secondary structure candidates.

Given points $\vec{r}_1$ and $\vec{r}_2$ along the axes $\hat{u}_1$ and $\hat{u}_2$ of secondary structures 1 and 2, each with a terminal at $\vec{R}_1$, and $\vec{R}_2$, the points may be labeled in terms of the distances t_1 and t_2 along the axis by

$$\vec{r}_1 = \vec{R}_1 + \hat{u}_1 t_1,$$

$$\vec{r}_2 = \vec{R}_2 + \hat{u}_2 t_2.$$

The points of closest approach of the lines passing through $\vec{R}_1$ along $\hat{u}_1$ and through $\vec{R}_2$ along $\hat{u}_2$ can be shown to be connected by that line that is simultaneously perpendicular to both lines 1 and 2. Therefore, unrolling the helices allows them to lie flat on each other at these special contact points, permitting alignment between surface features of the secondary structures.

Imposition of a distance cutoff constraint, following Bowie[9] and Walther et al.,[10] would require triplets of

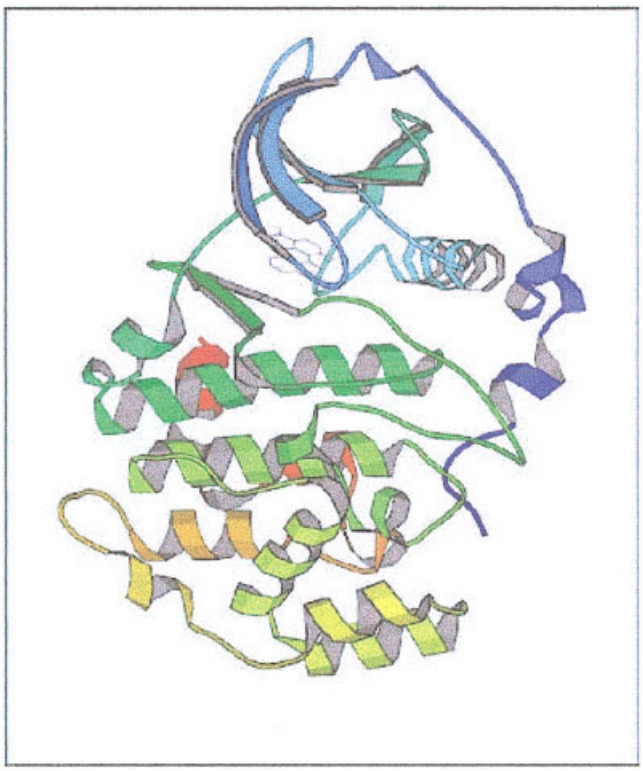

Fig. 1. The catalytic domain of protein kinase CK2 (PDB code 1F0Q). Arrows, beta strands; ribbon, alpha helices.

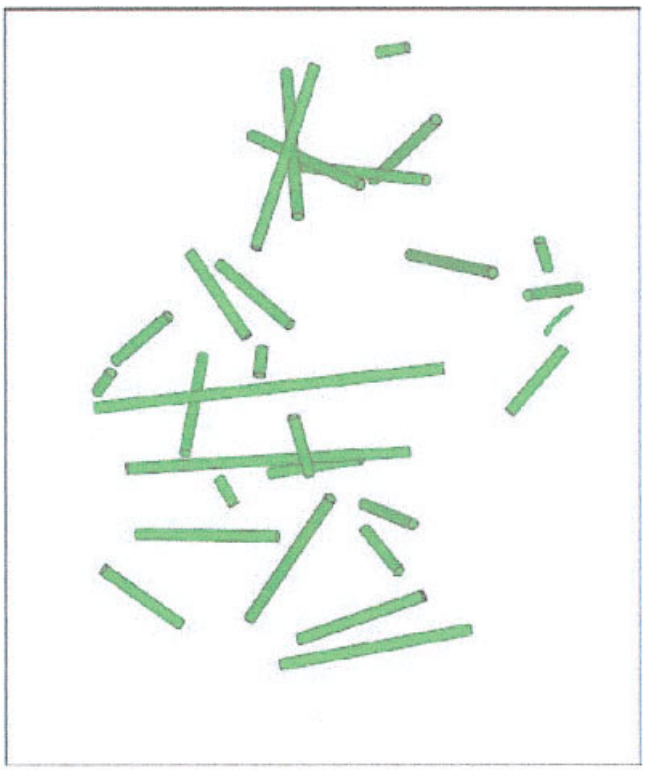

Fig. 2. Same as Figure 1, except that now helices and strands have been substituted by cylinders that approximate the segments used to describe secondary structural elements. Turns and coils are not considered.

secondary structure elements that only include cases where all members are in direct contact with each other. This left very few points, which would have corresponded to a very special condition. Particularly, this excludes barrels of helices *and* all β sheets. This distance was, therefore, not used to exclude any candidates in the consideration of any pair of secondary structure axes.

If the lengths of the secondary structure axes are L_1 and L_2, then the secondary structures are interactive if $0 \leq t_1 \leq L_1$ and $0 \leq t_2 \leq L_2$. Otherwise, helices will not "roll" on each other.

The next problem under consideration is to determine whether the secondary structure elements are arranged in space in a randomly uniform fashion, or whether the protein structures impose constraints in some regular way that deviates from what would be demonstrated by a uniformly distributed set of vectors.

The first step is to consider how such a uniformly distributed set of vectors would be oriented with respect to each other. The second step will then be to compare the measured distributions to the set of vectors.

It has been recognized that the distribution of angles between vectors distributed on a sphere produces a geometrically induced bias in the determination of modal packing angle preferences, which must be corrected by introducing a geometrical normalization factor.[9] This choice to normalize the distribution by dividing by sin θ made it difficult to recognize deviations in the alignment preferences for packing angles from the distribution expected for uniformly distributed unit vectors. An alternative treatment of the problem is to recognize that angles are not necessarily the most useful variables to express distributions of packing angles between secondary structure elements.

Consider a set of randomly distributed unit vectors $\hat{u}_1$, $\hat{u}_2$, and $\hat{u}_3$. The direction cosines between these unit vectors were defined

$$\Upsilon_1 = \hat{u}_2 \cdot \hat{u}_3,$$

$$\Upsilon_2 = \hat{u}_3 \cdot \hat{u}_1,$$

$$\Upsilon_3 = \hat{u}_1 \cdot \hat{u}_2,$$

These values can range from $-1 \leq \Upsilon_j \leq 1$. The distribution for the random vectors is expected to have the form

$$f(\Upsilon_1, \Upsilon_2, \Upsilon_3)d\Upsilon_1 d\Upsilon_2 d\Upsilon_3 = \frac{1}{(4\pi)^3} \oint d\Omega_1 \oint d\Omega_2 \oint d\Omega_3 \delta(\Upsilon_1 - \hat{u}_2 \cdot \hat{u}_3) d\Upsilon_1 \delta(\Upsilon_2 - \hat{u}_3 \cdot \hat{u}_1) d\Upsilon_2 \delta(\Upsilon_3 - \hat{u}_1 \cdot \hat{u}_2) d\Upsilon_3, \quad (1)$$

where the $\delta(x)$ are Dirac δ functions, and $d\Omega_i = d\Omega(\hat{u}_i)$. This may be evaluated to yield

$$f(\Upsilon_1, \Upsilon_2, \Upsilon_3)d\Upsilon_1 d\Upsilon_2 d\Upsilon_3 = \frac{1}{4\pi} \frac{d\Upsilon_1 d\Upsilon_2 d\Upsilon_3}{\sqrt{1 + 2\Upsilon_1\Upsilon_2\Upsilon_3 - \Upsilon_1^2 - \Upsilon_2^2 - \Upsilon_3^2}} \Theta(1 - \Upsilon_1^2)\Theta(1 - \Upsilon_2^2)\Theta(1 - \Upsilon_3^2)\Theta(1 - \Upsilon_1^2 - \Upsilon_2^2 - \Upsilon_3^2 + 2\Upsilon_1\Upsilon_2\Upsilon_3), \quad (2)$$

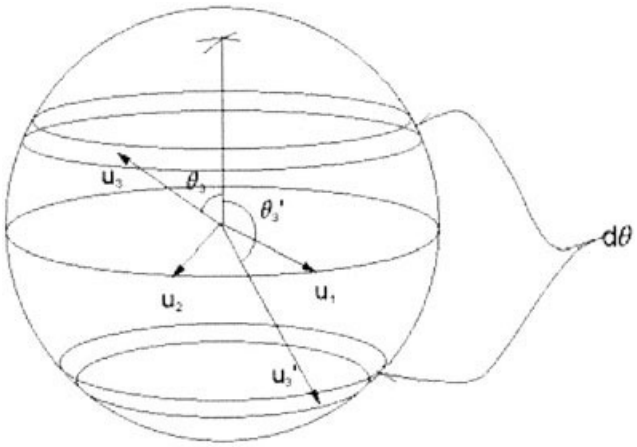

Fig. 3. Bias for angles between triplets of vectors uniformly distributed over a unit sphere.

where the $\Theta(x)$ are Heaviside step functions: $\Theta(x) = 1$ if $x \geq 0$, and $\Theta(x) = 0$ if $x < 0$. This evaluation is most simply carried out by recognizing that the orientation of the θ and ϕ coordinates of the vectors $\hat{u}(\Omega)$ pointing at solid angle $d\Omega = \sin\theta d\theta d\phi$ on the sphere are arbitrary. The first vector, $\hat{u}_1$ may be pointed at the $\theta_1 = 0$ pole. The second vector $\hat{u}_2$ may be pointed along the x-z plane, so that $\phi_2 = 0$.

If $\Upsilon_1 = \cos a$, $\Upsilon_2 = \cos b$, and $\Upsilon_3 = \cos c$, then the angles a, b, and c correspond to the sides of a spherical triangle. The distribution density $f(\Upsilon_1, \Upsilon_2, \Upsilon_3)$ diverges where $1 + 2\Upsilon_1\Upsilon_2\Upsilon_3 - \Upsilon_1^2 - \Upsilon_2^2 - \Upsilon_3^2 = 0$, which corresponds to the cases where $\cos c = \cos(a \pm b)$. This then corresponds to a situation where all the sides of the triangle lie in the same plane on the same great circle. Thus, even when the vectors are all uniformly distributed, planar arrangements are expected much more frequently than non-planar arrangements. This can also be understood by considering Figure 3. In this figure, two of the vectors, $\hat{u}_1$ and $\hat{u}_2$, form a plane, which intersects the sphere in a great circle. This circle may be taken to be the equator of a convenient coordinate system. The cosine of the angle between the pole of that equator and the third vector should then be uniformly distributed. There is a bias in the selection of the angle θ_3 or θ_3', measured from the polar axis, proportional to the sine of the angle. Thus, there is a preference for planar triangles in general, and for alignment in particular (any pair selected from the three vectors forms a plane with a preference for the third vector to lie close to that plane). The condition that $1 + 2\Upsilon_1\Upsilon_2\Upsilon_3 - \Upsilon_1^2 - \Upsilon_2^2 - \Upsilon_3^2 \geq 0$, imposed by the Θ function, represents the triangle inequality of spherical trigonometry.

The marginal distribution $f(\Upsilon_1, \Upsilon_2, \cdot)$, obtained by integrating over Υ_3 is a simple uniform distribution

$$f(\Upsilon_1, \Upsilon_2, \cdot)d\Upsilon_1 d\Upsilon_2 = \frac{1}{4}\Theta(1-\Upsilon_1^2)\Theta(1-\Upsilon_2^2)d\Upsilon_1 d\Upsilon_2. \tag{3}$$

Next, consider the problem of histogramming the distribution, and determining the probability that such a distribution would be observed given that the angles reflect uniformly distributed unit vectors. Three-dimensional histograms with L bins on each axis, yielding $B = L^3$ bins for the full distribution $f(\Upsilon_1, \Upsilon_2, \Upsilon_3)$, and $B = L^2$ for the marginal distribution $f(\Upsilon_1, \Upsilon_2, \cdot)$, and $B = L$ for the marginal distribution $f(\Upsilon_{1,\cdot,\cdot})$, were constructed for comparison with the distribution expected for uniformly random $\hat{u}_1$, $\hat{u}_2$, and $\hat{u}_3$. The total number of angles distributed over the bins was N, with n_j falling into each bin j. This type of measurement is expected to satisfy a multinomial distribution, with probability[20]

$$P(\{n_j\}) = \frac{N!}{\Pi_j n_j!}\prod_j p_j^{n_j}, \tag{4}$$

where $\Sigma_j n_j = N$, and $\Sigma_j p_j = 1$. There is an expectation value, a variance for each n_j and there are covariances between n_i and n_j of the form

$$E(n_j) = Np_j \tag{5}$$

$$\mathrm{var}(n_j) = Np_j(1-p_j), \tag{6}$$

$$\mathrm{cov}(n_i, n_j) = Np_i\delta_{ij} - Np_ip_j. \tag{7}$$

Also, and most usefully, the weighted sum of squared differences between the measured distribution and the expected values,

$$\mathscr{C}^2 = \sum_j \frac{(n_j - Np_j)^2}{Np_j} \tag{8}$$

is approximately χ^2_{B-1} distributed.

The selection of a bin index j from Υ_i is achieved by scaling Υ_i from 0 to 1 by noting that $-1 \leq \Upsilon_i \leq 1$, with the transformation $(\Upsilon_i - (-1))/(1 - (-1)) = (\Upsilon_i + 1)/2$. Then, for L bins numbered $j = 0, 1, \ldots, L-1$, this becomes $j = [L(\Upsilon_i + 1)/2]$, where $[x]$ is the largest integer less than or equal to x. If j is selected equal to L, it is placed in the $L-1$ bin. A value of Υ_i that corresponds to bin j is $(\Upsilon_i + 1)/2 = (j + 0.5)/L$, selecting the Υ_i from the center as being most representative of the bin, or $\Upsilon_i = 2(j + 0.5)/L - 1$. The size of a bin is $\Delta\Upsilon_i = \Delta[2(j + 0.5)/L - 1] = 2/L$. So, the probability of landing in bin j in the marginal distributions is $\Delta\Upsilon_1 \times 1/2 = 1/L$, which is what is expected.

The full distribution requires a little more care. The probability of landing in a bin (j_1, j_2, j_3) may be approximated by

$$p_{j_1 j_2 j_3} \approx \frac{1}{4\pi}\Delta\Upsilon^3 \frac{1}{\sqrt{1 + 2\Upsilon_1^\dagger\Upsilon_2^\dagger\Upsilon_3^\dagger - \Upsilon_1^{\dagger 2} - \Upsilon_2^{\dagger 2} - \Upsilon_3^{\dagger 2}}}\Theta(1 - \Upsilon_1^{\dagger 2} - \Upsilon_2^{\dagger 2} - \Upsilon_3^{\dagger 2} + 2\Upsilon_1^\dagger\Upsilon_2^\dagger\Upsilon_3^\dagger)\Theta(1-\Upsilon_1^{\dagger 2})\Theta(1-\Upsilon_2^{\dagger 2})\Theta(1-\Upsilon_3^{\dagger 2}), \tag{9}$$

where

$$\Upsilon_i^\dagger = 2\frac{j_i + 0.5}{L} - 1. \tag{10}$$

This approximation fails for $(\Upsilon_1^\dagger, \Upsilon_2^\dagger, \Upsilon_3^\dagger)$ in proximity to the divergent edges where $(1 - \Upsilon_1^{\dagger 2} - \Upsilon_2^{\dagger 2} - \Upsilon_3^{\dagger 2} + 2\Upsilon_1^\dagger\Upsilon_2^\dagger\Upsilon_3^\dagger = 0$. If $(\Upsilon_1^\dagger, \Upsilon_2^\dagger, \Upsilon_3^\dagger)$ happens to land outside of the permitted region, $f(\Upsilon_1^\dagger, \Upsilon_2^\dagger, \Upsilon_3^\dagger)$ returns a false 0. If it is near the boundary, then $f(\Upsilon_1^\dagger, \Upsilon_2^\dagger, \Upsilon_3^\dagger)\Delta\Upsilon^3$ is a poor approximation of p_j. This was handled in this study by computing the problematical contributions to $\mathscr{C}^2$ indirectly. The method chosen was to collect all the bins where p_j was not trusted into one set of bins S. Then define $N_S = \Sigma_{j \in S} n_j$, $p_S = \Sigma_{j \in S} p_j$. Then

$$\sum_{n_j, j \in S} p(\{n_j\}) = \sum_{n_j, j \in S} \frac{N!}{\Pi_j n_j!} \prod_j p_j^{n_j}$$

$$= \left(\sum_{n_j, j \in S} \frac{N_S!}{\Pi_{j \in S} n_j!} \prod_{j \in S} p_j^{n_j} \right) \frac{N!}{N_S! \Pi_{j \notin S} n_j!} \prod_{j \notin S} p_j^{n_j}$$

$$= \frac{N!}{N_S! \Pi_{j \notin S} n_j!} P_S^{N_S} \prod_{j \notin S} p_j^{n_j},$$

where $N_S + \Sigma_{j \notin S} n_j = N$ and $P_S = \Sigma_{j \in S} p_j$, is easily recognized as another multinomial distribution. Since $\Sigma_{j \in S} n_j$ may be directly computed from the occasions where sampled $(\Upsilon_1, \Upsilon_2, \Upsilon_3)$ land in those bins, and $\Sigma_{j \in S} p_j + \Sigma_{j \notin S} p_j = 1$, it is possible to obtain $\Sigma_{j \in S} p_j = 1 - \Sigma_{j \notin S} p_j$ from the trusted p_j, $j \notin S$. In this study, sites neighboring and next-to-neighboring poorly trusted points were put in set S.

A χ^2 test was performed comparing the measured distribution of secondary structure axes with the random vector distribution $f(\Upsilon_1, \Upsilon_2, \Upsilon_3)$ for each bin. Boundary bins, which are not approximated very well by a simple $f(\Upsilon_1^\dagger, \Upsilon_2^\dagger, \Upsilon_3^\dagger)\Delta\Upsilon^3$, were lumped into a single bin, and computed from the rest of the bins by applying conservation $\Sigma_j n_j = N$ and $\int d^3\Upsilon f(\Upsilon_1, \Upsilon_2, \Upsilon_3) = \Sigma_j p_j = 1$, shown above to preserve the multinomial character of the distribution of p's. A separate test was performed on the lumped boundary bins since a preference for alignment will be accumulated in those bins.

χ^2 tests were also performed on the marginals $f(\Upsilon_1, \Upsilon_2, \cdot)$, and $f(\Upsilon_1, \cdot, \cdot)$. If the vectors are random, the contribution of Υ_1 and Υ_2 to the joint distribution are independent and uniform. Even if the observed secondary structure axes are not uniformly distributed, a statistical interaction in the joint distribution between two of the angles would indicate that the probability of an alignment of one axis with another axis would be enhanced or diminished by the alignment of some third axis with the first axis. This would indicate a more global nonlocal interaction forced by the packing in the protein.

RESULTS

The measured distributions of the various combinations of secondary structure elements are presented. First, general characteristics of the distributions are reviewed. Second, a more detailed analysis of the interactive secondary structures are considered. Last, the global secondary structures are presented.

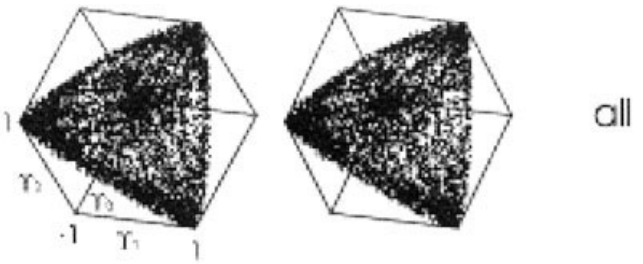

Fig. 4. Global secondary structure packing. A dot corresponds to 10 triplets falling in the same 3D table bin. Dots are displayed uniformly distributed within a bin. Only the bins with a number of triplets above average are displayed.

The triplets were sampled from "interactive" or "near," and "global" secondary structure elements. Global elements included all triplets. Interacting triplets were composed of the subset of all triplets such that the nearest points on each structure segment line were contained within the segments, as described earlier. The data were sampled in order of residue, so that each triplet member increased in rank of residue number. Triplets were selected by secondary structure class, either with no discrimination in class (labeled "All"), or by class type ("HHH" or "SSS" for pure helices or strands).

General Characteristics of the Distributions

Figure 4 shows a stereogram representing the joint distribution of angles density $f(\Upsilon_1, \Upsilon_2, \Upsilon_3)$, where the Υ_j axes were labeled x, y, z. Figure 4 shows both global and interactive ("near") triplets with no class discrimination. It is possible to discern that the boundaries are curved, reflecting the constraint, quadratic in each cosine, $1 + 2\Upsilon_1\Upsilon_2\Upsilon_3 - \Upsilon_1^2 - \Upsilon_2^2 - \Upsilon_3^2 \geq 0$, corresponding to the triangle inequality for spherical triangles. Further, the central region in the global triplets with no class discrimination is seen to be relatively devoid of points, showing some bias for selection of angles near the boundaries, which correspond to "flat" spherical triangles. The imposition of interactions, with no class selection, shows more structure in selection preferences.

Figures 5 and 6 correspond to the two-angle marginal joint distributions, where Υ_3 was integrated. The plots should show higher densities than the full three variate distributions. Further, it is much easier to compare the results with the distribution expected for uniformly random vectors, which would be uniform.

Characteristics of Interacting Secondary Structure Alignments

Figures for interactive subselection by class in the full three-variate joint distribution were not presented. The reason for this may be seen by noting that the number of triplets selected in each subset was too small a sampling number to show effectively on a $20 \times 20 \times 20$ grid.

Figures 5 through 10 correspond to the two-angle marginal joint distributions, where Υ_3 was integrated. The plots show higher sampling densities per cell than the full three-variate distributions, producing more a statistically

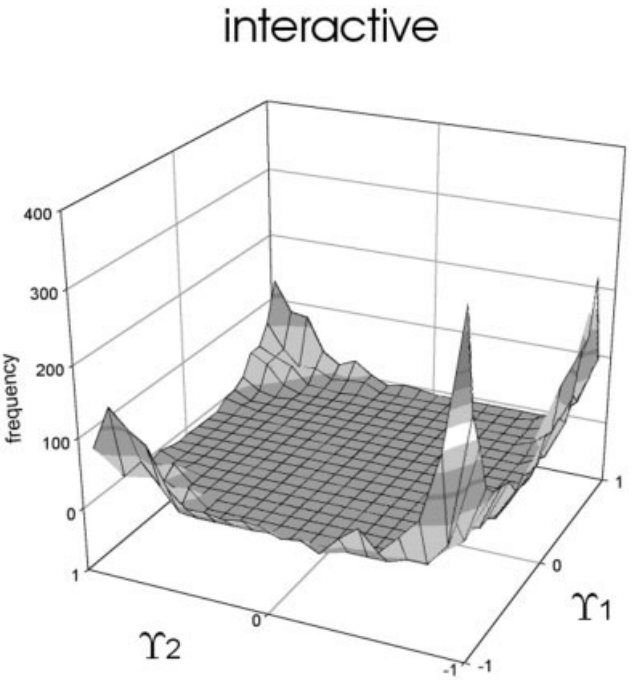

Fig. 5. Two-angle marginal joint distribution, where Υ_3 was integrated. Υ_1 corresponds to the cosine of the angle between the second and third triplet member, and Υ_2 corresponds to the angle between the third and the first triplet member. Interactive triplets of all classes HHH, HHS, . . . , and SSS are included in the plot.

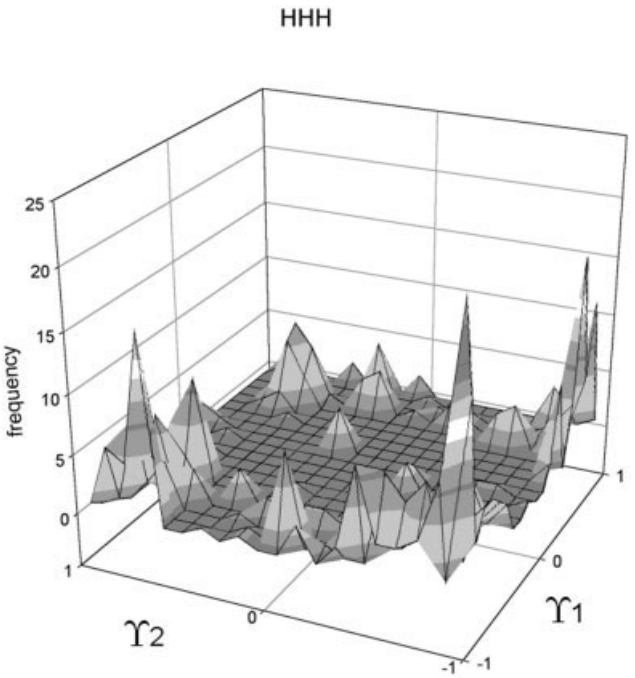

Fig. 6. Two-angle marginal joint distribution, where Υ_3 was integrated. Υ_1 corresponds to the cosine of the angle between the second and third triplet member, and Υ_2 corresponds to the cosine of the angle between the third and the first triplet member. Triplets of the HHH class only are included. The α helices of each triplet must be Interactive.

reliable measurement of the distributions. Further, it is much easier to compare the resulting distributions with that expected for uniformly distributed random vectors, since the marginal distributions for such vectors are expected to be uniform.

Characteristics of Interactive Secondary Structure Alignments

Figures 5 through 7 show the interactive data. The interpretation of joint angles are that Υ_1 corresponds to the angle between the second and third triplet member, and Υ_2 corresponds to the angle between the third and the first triplet member.

The number of points sampled by interactive "HHH" was too small to significantly resolve detailed structure. Even so, it can be seen that there are numbers of hits away from the axis corresponding to interaction angles described in other studies.[5,8,9] It has been noted earlier that this sampling does not refer to directly interacting helices. Directly interacting helix triplets are rare. Instead, helices that are mutually involved were identified by sharing perpendiculars connecting nearest points, as would be expected within helix barrels. While there is some chance that this could pick up stray helices, this appears to be a simple criterion for quickly identifying helices that are involved in larger mutual structures. It is observed that there are preferred packing angles between multiple helices induced by geometrical constraints in helix barrels.

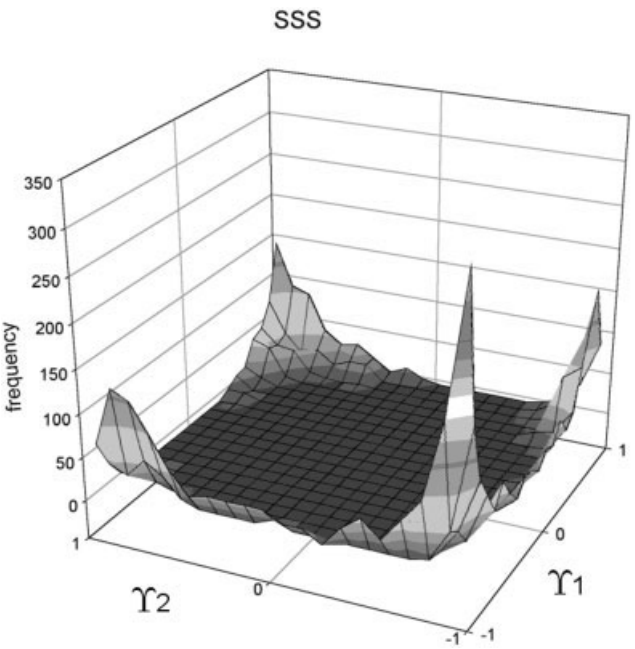

Fig. 7. Two-angle marginal joint distribution, where Υ_3 was integrated. Υ_1 corresponds to the cosine of the angle between the second and third triplet member, and Υ_2 corresponds to the angle between the third and the first triplet member. Only triplets of the SSS class are included. The β strands of each triplet must be Interactive.

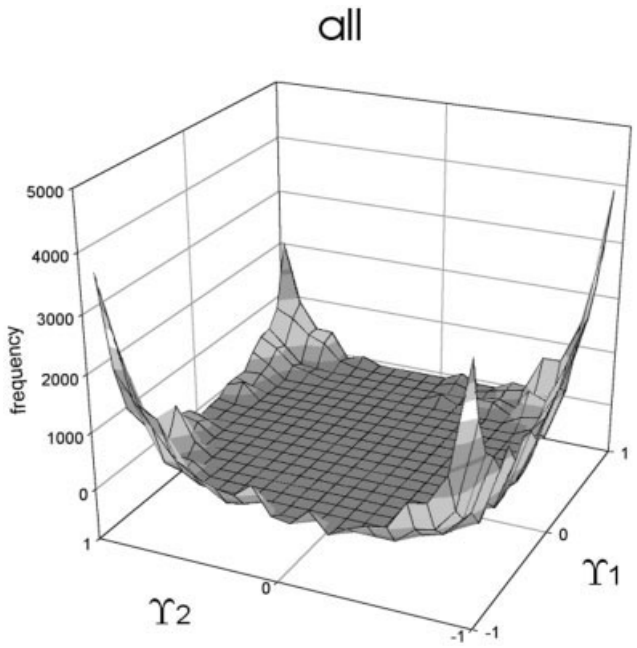

Fig. 8. Two-angle marginal joint distribution, where Υ_3 was integrated. Υ_1 corresponds to the cosine of the angle between the second and third triplet member, and Υ_2 corresponds to the cosine of the angle between the third and the first triplet member. All triplets of all classes HHH, HSS, . . . , and SHH are included. A total of 1,237,597 triplets were sampled.

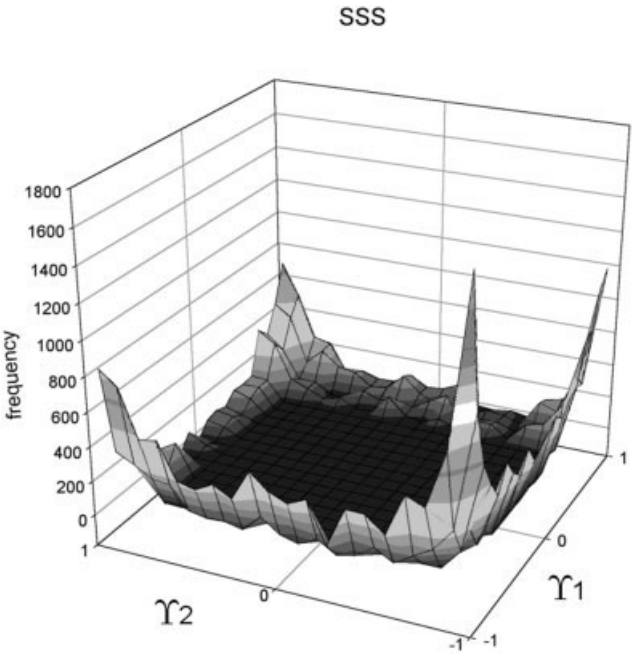

Fig. 10. Two-angle marginal joint distribution, where Υ_3 was integrated. Υ_1 corresponds to the cosine of the angle between the second and third triplet member, and Υ_2 corresponds to the cosine of the angle between the third and the first triplet member. All triplets of the SSS class are included. A total of 236,691 triplets were sampled.

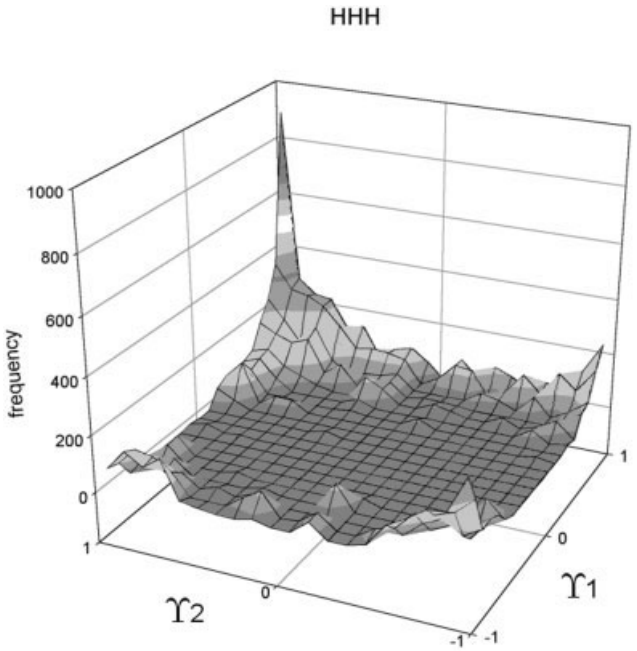

Fig. 9. Two-angle marginal joint distribution, where Υ_3 was integrated. Υ_1 corresponds to the cosine of the angle between the second and third triplet member, and Υ_2 corresponds to the cosine of the angle between the third and the first triplet member. All triplets of the HHH class are included. A total of 174,240 triplets were sampled.

Some peaks do correspond to the two-helix preferred packing angles reported elsewhere.

The case for interactive helices are listed under "HHH" in Table I. The single-angle distribution is nonuniform. The marginal two-angle joint distribution is nonuniform, and the angles interact. There is *not* a significant deviation of the three-angle distribution from that expected for random uniformly distributed vectors. This is more likely due to the small number of data points that qualified than due to a true equivalence (unable to reject the null hypothesis). This is partly supported by the test on the singular bins, which showed a significant accumulation of aligned axes. However, in this case, there *was* an indication of preferred angles, as seen in the figures, in the single-angle marginal distribution, reflecting preference of aligned structural features.

The case for interactive β-sheet strands, listed as "SSS" in Table I, shows many of the same features listed for α-helices. The single-angle distribution is nonuniform, the marginal two-angle joint distribution is nonuniform, and the angles interact. There is *not* a significant deviation of the three-angle distribution from that expected for random uniformly distributed vectors. This is more likely due to the small number of data points that qualified than due to a true equivalence (unable to reject the null hypothesis). This is partly supported by the test on the singular bins, which showed a significant accumulation of aligned axes. However, in this case, there was *not* an indication of preferred angles in the single-angle marginal distribution, reflecting preference of aligned structural features.

TABLE I. Sampling of Interactive Secondary Structure Alignments

Structure class	Triplets sampled	3D (χ^2_{1347})	Boundary (z)	2D (χ^2_{399})	1D (χ^2_{19})	Interaction (χ^2_{399})
All	12645	1636	36.9	135,619	18,357	2972.6
HHH	877	851.6	5.14	9,495.82	1,264.74	762.72
HHS	217	308.5	4.45	2,843.83	292.074	853.82
HSH	265	216.4	5.34	3,250.57	342.415	752.71
HSS	842	260.3	10.13	11,207.3	1,029.45	3,337.6
SHH	247	679.4	3.57	3,347.66	345.506	1,087.2
SHS	627	702.2	7.49	7,985.70	1,174.70	1,030.1
SSH	986	379.1	9.99	12,361.02	3,177.00	9,825.1
SSS	8584	1192	32.3	98,506.3	1,5991.4	2,690.7

TABLE II. Sampling of Global Secondary Structure Alignments

Structure class	Triplets sampled	3D (χ^2_{1347})	Boundary (z)	2D (χ^2_{399})	1D (χ^2_{19})	Interaction (χ^2_{399})
All	1,237,597	8,478.2	74.7	11,198,201	1,254,686	1,583,614
HHH	174,240	2,975.6	22.4	1,576,326.7	178,056.2	3,234.968
HHS	120,897	2,758.6	18.1	1,096,351.5	1232,837.4	11,393.02
HSH	133,669	2,843.9	25.4	1,218,183.0	141,552.7	2,593.925
HSS	155,873	3,317.7	27.0	1,418,761.9	159,842.9	8,296.430
SHH	136,988	3,321.2	16.0	1,243,243.9	144,068.3	11,355.34
SHS	137,537	3,701.6	17.0	1,249,919.4	143,228.4	4,127.662
SSH	141,702	3,818.5	37.3	1,294,968.6	155,794.0	12,150.64
SSS	236,691	5,063.5	43.7	2,167,446.6	252,862.7	4,007.338

Characteristics of Global Secondary Structure Alignments

Figures 8 through 10 show the two-angle marginal distribution for global triplets with secondary structure class subselection. The "HHH" set shows a very pronounced preference for parallel and anti-parallel alignment that was simply not clearly evident in the interactive subset selection case. There is a similar preference for parallel and anti-parallel alignment in the "SSS" set, but the "HHH" shows a much broader distribution of angles, with some indication from those packing angles besides parallel or anti-parallel that reflect direct interactions not excluded from the global set. However, the distribution shows a much larger preference for simple alignment than is evident from the interactive data set.

The global secondary-structure class-independent sampled triplets, shown in Figure 10, shows a similar preference for parallel and anti-parallel alignment as characteristic of the "SSS" subclass as observed in the interactive case. However, while the total number for nonsubset selection is 1,237,597, the numbers in the "SSS" set are now 236,691, which is a small fraction. Comparison of "All" in Figure 5 with "All-Interactive" in Figure 10 shows that while the fraction of triplets dominated by strand-strand alignments is smaller, the actual preference for parallel or anti-parallel alignment is actually greater.

Statistical Tests

The next case considered is the statistical tests on the distribution of all helix or strand axes within subunits, listed under "All" in Table II. The marginal distribution, composed of a cosine of the single angle between the second and third axes selected within triplets, would be expected to be uniform for random uniformly distributed vectors. This was not observed. The deviation of the sampled single-angle marginal distribution from that expected for randomly oriented vectors had a $\chi^2_{19} = 1.255 \times 10^7$; the probability of observing such a deviation by chance in a random sampling process is essentially zero. There is a strong preference for alignment or anti-alignment, with anti-alignment being preferred in comparison to the expected uniform distribution. There is no indication of structural preferences induced by interactions between secondary structure elements that had been studied elsewhere,[2–8] since this set seems to be dominated by strand-strand interactions.

The marginal joint distribution of the angles between second and third, and first and third axes out of the selected triplets would also be uniform if the vectors were random uniformly distributed vectors. However, just as the marginal distribution for the first angle showed a nonuniform distribution, the joint distribution is also nonuniform.

The full distribution of the cosines of all three angles sampled by the axes of secondary structure elements was compared with the distribution of random uniformly distributed vectors. Those bins that are poorly estimated by a simple product $f(\Upsilon_1^\dagger, \Upsilon_2^\dagger, \Upsilon_3^\dagger)\Delta\Upsilon^3$ where f becomes singular, were lumped into a single multinomial bin as outlined in previously. The distribution for global nonspecific secondary structure class selection deviates significantly from that expected for uniform randomly distributed vectors, with a $\chi^2_{1347} = 8478.2$. The lumped bins at the singular regions of the distribution correspond to alignment between the axes. The z-score indicating a preference for alignment over that expected for randomly distributed

vectors is $z = 74.7223$, with the probability of observing a score larger than this being essentially zero.

Tests were also performed for combinations of helix and strand data, listed as helix-helix-helix (HHH), helix-helix-strand (HHS), etc., which have been indexed in Tables I and II. In both tables, column labels are as follows: "Class" refers to the types of secondary structure axes; "Triplets Sampled" is the number of triplets in the class; "3D χ^2_{1347}" is the χ^2 score measuring the deviation of the actually observed sample from the peak value expected for unit vectors uniformly distributed on a sphere excluding boundary-sampled points, with an expectation value of 1,347 for vectors uniformly distributed on a sphere; "Boundary z" is the Gaussian z score measuring the deviation in number of points in the 3D sampled distribution that occurred in the boundary regions (if unit vectors uniformly distributed on a sphere described the distribution, this score would have an expectation value of 1); "2D χ^2_{399}" is the deviation of the distribution from the marginal of the 3D distribution, which for vectors uniformly distributed on a sphere would be uniform, with an expectation value 399; "1D χ^2_{19} refers to the marginal over 2 angle cosines, which would be uniform for vectors uniformly distributed on a sphere, with an expectation value of 19; "Interaction χ^2_{399}" is a standard contingency χ^2 test measuring the deviation of the measured sample from that expected from the marginals if there were no interaction between any two of the three angle cosines with an expectation value of 399, which is independent of the model of vectors uniformly distributed on a sphere. Table I shows the test results comparing the distribution measured for those axes selected as being "interactive" against those for unit vectors uniformly distributed on a sphere. Table II shows the test results comparing the distribution measured for all axes against those for unit vectors uniformly distributed on a sphere.

Statistical Dependence of Packing Variables

If the vectors were uniformly distributed, the above joint distribution would also be uniform, and the cosines would be independent. Even though the angles are not uniform, it is meaningful to ask if they are independent. Such an independence would indicate that the second selected axis would align with the third axis with the same probability regardless if the first axis aligned with the third axis or not, since constraints between the first and second axis had been integrated out in the marginal distribution. Even though the full three-angle distribution for random uniformly distributed vectors is nonuniform, indicating that aligned first and third, and aligned second and third vectors implies alignment between first and second vectors, the angles in the marginal joint distribution with the angular constraint between first and second vectors integrated out, are still independent. In considering the global set with no secondary class subset selection, a standard χ^2 contingency test of the independence of the first two angles in the marginal distribution shows a $\chi^2_{399} = 1.584 \times 10^7$; the probability of observing a deviation larger than this by chance is essentially zero. Similar results are observed whether the triplets sampled were interactive or global, and whether secondary structure class subtype selection was performed, as seen in Tables I and II, under the heading "Interaction." This implies that the axis of one secondary structural element is more likely to be aligned with a second one if that second axis is aligned with a third than if it is not aligned with another axis, even accounting for the fact that there are more aligned vs. unaligned axes. This implies that there is a highly significant statistical interaction, reflecting a packing induced nonlocal interaction.

CONCLUSIONS

It has been established that the packing-induced distributions of secondary structure elements are important to the prediction of protein folding.[2–6] particularly in attempting to evaluate the effects of individual mutations or modifications between similar proteins.[7] However, more global characterizations that might be reduced to scoring functions, are also valuable tools in starting the process of predicting the folding of a protein.[21,22] Further, the application of such angles to geometric hashing is natural since such applications are greatly improved by taking advantage of an understanding of the distribution of has keys. Such considerations originally motivated the study presented here.

The results outlined here suggest a global characterization of the tendency for secondary structure elements to align in proteins, regardless of whether they directly interact or not. Such a tendency could be due to some physical mechanism, or it could be due to sampling bias within the PDB[13] or SCOP[14] databases as applied in this study, even though the purpose of selecting representative PDB entries from the SCOP database was to use it to provide a spanning set of archetypes in order to remove or rationally control overrepresentation bias from the PDB.

The definition of interactive secondary structure elements adapted here is selected to include elements constrained by direct interaction and by mutual geometry, such as α-helix barrels or β-sheet strands. More stringent selection rules that impose a distance cutoff criterion would be relatively uninteresting in the consideration of a joint distribution of multiple packing angles. The application of this criterion for interactive secondary structures recovers many of the features of helix-pair packing, as well as geometrically induced constraints between pairs that are not directly interacting. Preferred packing between neighbors as reported elsewhere is still resolved among the peaks in the packing angle histograms.

For the case where the interactive constraint is relaxed, it is observed that there is a strong, though not absolute, tendency for secondary structure elements to anti-align or align, and that the tendency increases if more than one pair of such elements are aligned, indicating that there may be a globally induced nonlocal interaction encouraging the alignment. This remains true even though local interactions between helices and strands seem to show a broader distribution among nonparallel local orientations. Details of local interactions seems to be suppressed in consideration of the global distributions. There is a much

more pronounced tendency for simple parallel vs. antiparallel packing between all secondary structure types for global as opposed to "near," or interactive, packings. While tests seem to indicate that the χ^2 test for uniformity shows less significance for interactive triplets as opposed to global triplets in comparing Table I with Table II, two factors can contribute: first, there can be a more uniform distribution in the interactive set, and second, the uncertainty of the test may be increased by reducing the number of triplets sampled. Both factors are present in the comparison of interactive vs. global secondary structure triplet selection.

We, therefore, conclude that there is a strong global preference for parallel packing between secondary structures within the selected data set, including helix-helix interactions. The packing angles between secondary structures are not distributed independently. The probability of parallel packing between any pair of secondary structure axes is enhanced by a packing of either of the pair of axes with a third secondary structure axis.

The above conclusions suggest that there may be a mechanism selecting conformations in which secondary structures prefer parallel or antiparallel packings within the selected dataset. One such mechanism could be the fact that cylinders and sheets do not pack as well if they are not parallel as they will if they are. This produces a cost in surface to volume, and empty pockets. Such pockets tend to be expensive, costing between 2–8 kcal/mol,[23] but may provide benefits worth the cost in terms of permitting inter-domain movement[24] or providing a means of stabilizing with prosthetic groups.[23] Simple thermodynamic arguments supporting preferences for some packing geometries[11] show that exclusions of waters and other interactions[12] select particular geometries over others. It has been further shown that exclusion of waters and other interactions considered by Fernández et al. may be expected to produce 3-point cooperativity.[25] This leads to the same type of statistical interaction shown by the measurements reported in this study. In the balance between such benefits and their costs, the preference would be for alignment. Another group of proteins preferring aligned secondary structure elements, particularly helices, are transmembrane. However, few transmembrane structures have been determined by X-ray crystallography.

ACKNOWLEDGMENTS

Support for Guerra and Zanotti was provided in part by the Italian Ministry of University and Research under the FIRB Project "Bioinformatic for Genomics and Proteomics." Additional support for Guerra was provided by the Italian Ministry of University and Research under the FIRB Project "Enabling Platforms for High Performance Computational Grids Oriented to Scalable Virtual Organizations."

REFERENCES

1. Crick FHC. The packing of α-helices: simple coiled coils. Acta Crystallogr 1953;6:689–697.
2. Chothia C, Levitt M, Richardson D. Structure of proteins: packing of α-helices and pleated sheets. Proc Natl Acad Sci USA 1977;74: 4130–4134.
3. Richmond TJ, Richards FM. Packing of α-helices: geometric constraints and contact area. J Mol Biol 1978;119:537–555.
4. Cohen FE, Richmond TJ, Richards FM. Protein folding: evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. J Mol Biol 1979;132:275–288.
5. Chothia C, Levitt M, Richardson D. Helix to helix packing in proteins. J Mol Biol 1981;145:215–250.
6. Cohen FE, Kuntz ID. Prediction of the three-dimensional structure of human growth hormone. Proteins 1987;2:162–166.
7. Reddy BVB, Blundell TL. Packing of α-helices: geometric constraints and contact area. J Mol Biol 1993;119:537–555.
8. Walther D, Eisenhaber F, Argos P. Principles of helix-helix packing in proteins: the helical lattice superposition model. J Mol Biol 1996;255:536–553.
9. Bowie JU. Helix packing angle preferences. Nature Struct Biol 1997;4:915–917.
10. Walther D, Springer FC, Cohen FE. Helix-helix packing angle preferences for finite helix axes. Proteins 1998;33:457–459.
11. Finkelstein AV, Pittsyn OB. Why do globular proteins fit the limited set of folding patterns? Prog Biophys Mol Biol 1987;50:171–190.
12. Murzin AG, Finkelstein AV. General architecture of the α-helical globule. J Mol Biol 1988;204:249–269.
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank Nucleic Acids Res 2000;28:235–242; http://www.rcsb.org/pdb/.
14. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
15. Fischer D, Tsai C, Nussinov R, Wolfson H. A 3D sequence-independent representation of the protein data bank. Protein Eng 1995;8:981–997.
16. Arun KS, Huang TS, Blostein SD. Least-squares fitting of two 3-D point sets. IEEE Pattern Anal Mach Intell 1987;9:698–700.
17. Goryn D. On the estimation of rigid body rotation from noisy data. IEEE Pattern Anal Mach Intell 1995;17:1219–1220.
18. Kanatani K. Analysis of 3-D rotation fitting. IEEE Pattern Anal Mach Intell 1994;16:543–549.
19. Gerstein M. A resolution-sensitive procedure for comparing antigen-combining sites. Acta Cryst 1992;A48:271–276.
20. Freund JE. Mathematical statistics, 5th ed. Upper Saddle River, NJ: Prentice Hall, 1992.
21. Silverman BD. Hydrophobic moments of protein structures: spatially profiling the distribution. Proc Natl Acad Sci 2001;98:4996–5001.
22. Novotny J, Rashin AA, Bruccoleri RE. Criteria that discriminate between native proteins and incorrectly folded models. Proteins 1988;4:19–30.
23. Erikson AE, Basse WA, Wozniak JA, Matthews BW. A cavity-containing mutant of T4 lysozyme is stabilized by burried benzene. Nature 1992;355:371–373.
24. Hubbard SJ, Argos P. A functional role for protein cavities in domain:domain motions. J Mol Biol 1996;261:289–300.
25. Fernández A, Colubri A, Berry RS. Three-body correlations in protein folding: the origin of coopravitiy, Physica A 2002;307:235–259.