Analysis of Secondary Structure Elements of Proteins Using Indexing Techniques

Concettina Guerra
Dip. Elettronica e Informatica
Univ. of Padova, Italy
email: guerra@dei.unipd.it

Stefano Lonardi
Computer Science and Engineering
Univ. of California Riverside, USA
email: stelo@cs.ucr.edu

Giuseppe Zanotti
Dip. Chimica Organica e Centro Studi Biopolimeri,
Univ. of Padova, Italy
email: zanotti@chor.unipd.it

Abstract

In this paper we present a method for protein structure comparison that is based on indexing. Unlike most methods using indexing, ours does not use invariants of the C_α atoms of the proteins, rather it relies on geometric properties of the secondary structures. Given a set of protein structures, we compute the angles and distances of all the triplets of linear segments associated to the secondary structures of the proteins and use them to build hash tables. The hash tables can be used for fast retrieval of hypotheses of matches of a query protein against the database. We present and analyze the tables obtained for two separate sets of proteins that are representatives of all the folds in the PDB. The tables show an interesting distribution of the triplets of elements, especially if one takes into account that the elements of a triplet are generally not close in space. The majority of the elements are found to belong to two planar regions in the three-dimensional tables. The planar regions can be characterized as those whose corresponding triplets of structures lie on almost parallel planes.

Keywords: indexing, geometric hashing, pattern recognition, secondary-structure elements, globular proteins.

1. Introduction

The exponentially growing number of three-dimensional protein structures available through the Protein Data Bank (PDB) [1] has been paralleled by the development of several automatic programs for pattern recognition and detection of distant similarities among functionally unrelated proteins [3, 7, 10, 11, 13]. In fact, it has long been recognized that globular proteins, despite their huge number, can be grouped in a quite limited number of basic folds [18, 22].

In this paper we consider the problem of matching a query protein against a database of existing proteins for the automatic assignment of a newly determined structure to one of the existing protein families. This is a problem that has recently received a lot of attention in the biological literature [19, 20]. We propose an approach to solve this problem based on indexing using geometric invariants of the secondary structures of the proteins. We consider all triplets of secondary structures and their associated best-fit linear segments. Geometric properties of the three linear segments are computed and used to index a three-dimensional hash table; after the table has been built for a given set of proteins, each entry contains information about all triplets of geometrically similar structures in the given set. The table can then be used for different types of protein comparisons, for instance, for matching a query protein against the database, or for fast retrieval of common substructures among proteins. Moreover, the table can be used for a statistical analysis of the arrangement of the secondary structures.

Indexing techniques present obvious advantages over other search techniques when large databases are involved. The matching phase of the indexing is not heavily dependent on the size of the database since it does not require to match each protein structure separately. Drawbacks of indexing are excessive memory requirements for the hash table entries, sensitivity to quantization parameters, and the possibility of false positive matches. Geometric hashing and indexing for protein structural comparison were first proposed in [7], where invariant properties of quadruples

0-7695-1521-5/02 $17.00 © 2002 IEEE

of C_α atoms were used in the hashing scheme. Here we use higher-level properties of secondary structural elements which allow us to reduce the memory requirements and the computation time. Obviously, the result of the matching based on the secondary structures can only provide a coarse estimate for the matching at the atomic level.

One of the contributions of the paper is in the analysis of the arrangements of triplets of secondary structures. Spatial relationships among the elements of secondary structure have not been systematically analyzed, except for packing of helices or strands in contact [5, 6, 24]. We analyze the distribution of the hash tables for two separate representative sets of proteins from the PDB each consisting of approximately 300 folds. A more complete and detailed analysis of angle bias in secondary structure packing can be found in [21], where a comparison is made between the distribution of 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.

This paper is organized as follows. We first review existing approaches to matching 3D structures using indexing techniques and then present our indexing scheme and describe the hash table construction and its use for protein structural comparison. We focus on the analysis of the distribution of entries in the tables for two datasets of approximately 300 proteins with structurally different folds.

2. Previous Work

We briefly review some indexing methods based on geometric invariants and on the use of hash tables. The technique of *geometric hashing* was originally developed within the field of Pattern Recognition and Computer Vision to solve the model-based object recognition problem [14]. Given a database of object models, it consists of representing each model by storing redundant transformation-invariant information about it in a hash table. This table is compiled off-line. At recognition time, similar invariants are extracted from the sensory data and hashed into the table to find possible instances of the models in the scene. Geometric hashing has generally been applied to point sets either 2D or 3D under rigid transformations or the more general affine transformations. For matching 3D point sets, quadruples of points are used to define reference frames or bases in which the coordinates of all other points remain invariant. Models are stored into the table by considering all possible combinations of quadruples of points as bases and using the invariant coordinates of the remaining points to index the table. At recognition time, if the correct quadruple of points is chosen from the image points, the candidate matches are efficiently retrieved. Geometric hashing suffers from sensitivity to noise and excessive memory requirements. Several heuristics have been proposed to solve either one or both: coarse quantization of the hash bin, the selection of few relevant bases, the detection of ”seed matches” and the clustering of them.

In object recognition, most of the research on indexing has focused on which type of invariant to use. Ad-hoc high-level shape features, may be selected for applications to specific classes of objects. One important way to overcome the limitations of geometric hashing is to use more complex global invariants to index the hash table. In [4], it is shown that higher-dimensional index spaces lead to a drastic reduction in computation time. They tend however to produce many false positive matches and therefore require a careful verification phase.

Geometric hashing and indexing have been applied to solve various instances of the protein structure comparison problem, from the complete comparison of large sets of proteins, to the fast retrieval of patterns or motifs from the PDB, to the pairwise comparison of proteins allowing hinge bending [7], [8], [12], [23]. More recently, applications have included multiple structural alignment, that is the determination of the largest substructure common to all molecules of a given ensemble [17]. Unlike other existing protein structure comparison methods, they generally use only 3D information in the form of coordinates of the C_α atoms of the proteins and are sequence independent.

3. Index structure

We consider the problem of matching a query protein against a database of protein structures and use the secondary structures as primitives of the matching. Geometric properties of triplets of secondary structures serve to build and query the hash table. The secondary structure elements are approximated by linear segments and their angles and distances are computed. The three dihedral angles associated to a triplet of secondary structures are used to index a 3D hash table. Each table entry stores information about all triplets of segments of the proteins in the database that hashed into it. The distances between the segments of a triplet are also included in the table to be used in the matching phase to filter hypotheses of correspondences. The construction of the table is computation intensive; it requires O(n^3) time, for n secondary structures. However, once it is built it allows fast retrieval of candidate matches between the query protein and the proteins stored in the database.

3.1. Building the hash tables

All triplets of segments associated to α-helices and β-strands are considered in the analysis independently from their position in the polypeptide chain. The definition of helices and strands in a protein is taken from the PDB, without performing any specific control. The segment associated to a secondary structure belongs to the best-fit line through the $C\alpha$ atoms of the helix or the strand and has approximately the same length as the structural element. The details of the computation of the segments are omitted. A segment representing a helix usually corresponds quite well to the axis of the helix itself, except for very long helices that are often bent. For β-strands, the representation through a linear segment is generally more approximate, especially for long and bent strands. In our experiments all segments are considered oriented, the orientation being determined by the polypeptide chain, from the N to the C-terminus.

We build four separate three-dimensional hash tables. Each table is indexed by the three angles formed by three segments associated to three secondary structures. $table\_0$ corresponds to the triplets of β-strands only, $table\_1$ corresponds to the triplets consisting of two β-strands and one α-helix, $table\_2$ correspond to the triplets of one β-strand and two α-helices and, finally, $table\_3$ corresponds to the triplets of α-helices only.

Let (s_i, s_j, s_k) be a triplet of segments, where s corresponds either to an α-helix or a β-strand. Let α_{sr} be the dihedral angle formed by two segments s and r. The dihedral angle between two segments is the angle formed by the two planes perpendicular to the straight lines containing the segments themselves and therefore is defined in the range [0, 180]. In previous papers, the interval of angular values could be extended to [-180,180] because only packing helices and strands were considered [6, 24]. However, the absolute values of the angles are coincident in the two systems.

The triplet of segments (s_i, s_j, s_k) is then described by the three angles $(\alpha_{ij}, \alpha_{jk}, \alpha_{ki})$. The three angles, quantized into uniform intervals, give an index for accessing a cell of one of the four tables where the triplet of segments is stored. For the triplet of angles we use the canonical representation (α, β, γ), with $\alpha \leq \beta \leq \gamma$, so that only one copy is entered into a table for each triplet of segments. As a consequence, the distribution of the triplets (α, β, γ) is constrained by the three following inequalities:

$$\alpha \leq \beta$$
$$\beta \leq \gamma$$
$$\gamma \leq \alpha + \beta$$

The last inequality follows from the fact that α, β and γ can be mapped to the angles formed by three edges incident to one vertex of a tetrahedron. Note that there is no explicit information in our tables about the order of the segments along the polypeptide chain. A cell of the look-up table with index (α, β, γ) contains a list of elements each associated to three segments forming (α, β, γ) angles. Each element consists of the following:

- the 4-digit identifier of the protein which contains the three structures corresponding to the segments
- the starting and ending residue number of each of the three secondary structure elements
- the distances between the segments.

The distance is measured as the distance between the middle points of the two segments.

For each cell a counter of the number of entries in the cell is also kept. Due to the relatively large approximation introduced in the simplified segment representation, the angles are discretized into intervals of 10 degrees. Each table is then a three-dimensional matrix consisting of 18x18x18 elements cells. For instance, the cell with coordinates (0,0,0) contains all the triplets of segments with three dihedral angles each in the range [0, 10]. All triplets of segments in a protein, irrespective of their relative distance, are included in the tables, thus most of the triplets are relative to structural elements spatially distant.

3.2. Recognition

Once the four tables have been built, they can be queried to find similarities in the arrangment of the secondary structures of a query protein with the proteins stored in the database. A protein P is matched against the database of proteins by the following procedure:

- For each triplet (s_i, s_j, s_k) of secondary structures of P, compute the three angles (α, β, γ) and the three distances of the associated segments. In the corresponding hash table, look at the cell indexed by (α, β, γ) and tally a vote for each entry in the cell with similar distance values. Repeat the same for the adjacent neighboring cells.
- Formulate and rank hypotheses of matching by determining the proteins with the highest number of votes.
- Verify hypotheses, thus eliminating false positives.

Each entry in the cell contains triples of segments with almost identical dihedral angles, but they may have very different distance values. Obviously, similarity among angles does not imply spatial similarity, that is the possibility of aligning the two sub-structures. The only possible candidates for matching a triplet of segments of the query protein are the entries that have similar distance values to that triplet. Thus a vote is tallied only to those entries.

Because of the quantization, the most similar stored triplets may not lie in exactly the same hash cell as the query triplet, but in one of the neighboring cells. That is why our method examines also the adjacent cells to accumulate consensus for matching hypotheses.

Hyphotheses of matches of the query protein with the stored proteins are ranked according to the number of votes they accumulate. At this stage, there is no check for the consistency of the multiple triplet associations that provide consensus for the same hypothesis. In other words, it could, for instance, happen that two triplets of the query protein consisting of distinct elements are hashed into two triplets sharing some elements. False positive matches may thus arise and a verification phase is needed. However, in practice, simply counting the number of votes provides a reasonable matching score.

The verification of the hypothesized matches may be performed by a pairwise comparison between the proteins, either at the level of secondary structures [9] or by extending the matching to residue level.

Once compiled, the table can be used for different types of comparisons, for instance for all-to-all structure comparison.

4. Distribution of triplets of elements

The Protein Data Bank contains about 12,000 protein models, but its high level of redundancy makes any statistical analysis performed on the entire database highly biased. In order to avoid this problem, our analysis has been undertaken on two different and limited sets of structures, each of them containing at least one representative of all the folds in the database.

One of the representative sets is taken from Fischer et al. [8] and contains 268 structures. Of them, approximately 200 were used in the analysis mostly because the remaining ones did not include the specification of the secondary structures in the PDB. The other set is obtained using the unique folds deposited every year at the PDB as detected by SCOP [19]. Approximately 300 proteins of this set were analyzed.

In this section, we show the distribution of the angles of the triplets of segments associated to the secondary structures for the selected proteins. The number of entries in each of the four tables for the two representative sets is shown in table 1. Column two and four of the table give the number of entries of the complete tables, column three and five those of the tables using a cut-off distance (in brackets in the table).

The hash tables are displayed only for the set of proteins obtained by SCOP [19]. The results obtained for the other set are very similar and are not presented here. $table\_0$, corresponding to all the triplets of three β-strands, is shown in figure 1, where the three axes x, y and z represent α, β, and γ, respectively. The elements of a cell of the table are graphically represented by dots. For simplicity, a dot in the picture represents 10 elements (triplets) in the corresponding hash table cell. Dots within the same cell are shown uniformly distributed in the cell. As already noticed, for geometrical reasons not the entire three-dimensional table is populated: the constraints $\alpha \leq \beta \leq \gamma$ and $\gamma \leq \alpha + \beta$ imply that only a wedge of the table may contain non-empty cells. The inspection of all the tables shows a concentration of elements in certain regions of the wedge. Figures 2-4 display the regions of the four tables that are more populated, i.e. the regions of all the cells that contain a number of elements greater than 3σ, where σ is the standard deviation. It can be observed that these cells identify two planar regions in the $(\alpha, \beta, \gamma$) space defined by the equations:

$$\alpha + \beta = \gamma$$

and:

$$\alpha + \beta = 360 - \gamma \qquad ,$$

if $\alpha + \beta > 180$.

The above relationship is in some way unexpected. We are in fact considering elements of secondary structure that in general are not in contact, i.e. that do not directly interact. Nevertheless, since they belong to the same protein, their non-covalent interaction is mediated by a sequence of other strands or helices. We have also to consider that in general the straight lines that contain the segments do not intersect and are not coplanar. The previous relationship holds in fact for the case of three coplanar straight lines. Consider now three non-coplanar segments. It can be shown (see Appendix) that the relation $\alpha + \beta = \gamma$ holds if and only there exist three parallel planes each containing one of the three segments. The planes are constructed as follows. As it is well known, a pair of segments in space defines two parallel planes through either segment parallel to the other segment. Thus, the three pairs of segments in a triplet define six planes that are pairwise parallel. If $\alpha + \beta = \gamma$, then the six planes reduce to three parallel planes each through one segment. Due to the discretization of the $(\alpha, \beta, \gamma$) space, these planes are almost parallel for the majority of the triplets in the tables; in other words the tetrahedron formed by the three unit vectors associated to the three segments and through the origin of the coordinates system is almost ”flat”.

This property is expected for frequent motifs, such as the β-α-β motif [2], that is found in almost every protein structure that has a parallel β sheet. In a β-α-β motif, two adiacent parallel β strands are packed against an α helix and the helical axis is parallel to the plane of the two β strands. Moreover, often in a sheet most of the β strands lie on the same plane, and therefore satisfy the above relation. These cases however, generally involve packing elements or elements close in space. In our tables, close elements represent a small fraction of the overall entries. To the best of our knowledge, this property was not observed before in a systematic way for the elements of secondary structure not close in space.

5. Distribution of triplets of elements close in space

The analysis of the distribution of angles in triplets of elements spatially close, i.e. helices or strands linked by non-covalent interactions, was performed using a subset of elements of the previous tables.

The idea of including in the analysis only triplets of secondary structural elements that interact with each other, i.e. β-strands connected by hydrogen bond interactions or α-helices with at least some atoms at a distance close to the sum of their van der Waals radii, is impracticable once the hash table has been built. In fact, the original data base has been reduced to a collections of segments and the original atomic details are missing. Since the information on the distance between the middle points of each pair of segments is present in the database, we have applied a cutoff based on such distance, assuming that a distance between segments shorter than a given value should in general correspond to secondary structural elements directly interacting. Different cutoff values were chosen for β-strands and α-helices, since the mean distance between strands close in space is definitely shorter than that between helices. A compromise was used for mixed sets, thus cutoff values were 10, 15, 20 and 20 Angstrom for $table\_0$, $table\_1$, $table\_2$ and $table\_3$, respectively. This approximation should include most of the triplets we want to consider, along with some unwanted. It must be pointed out that tests performed using quite different cutoff distances (i.e., 15 Angstrom for $table\_0$ or $table\_3$, instead of 10 and 20 Angstrom, respectively) do not alter the general behavior of the table.

The results are displayed in Fig. 6- 7 and summarized in Table 2. Table 2 gives the local maxima of the cell counters in each of the 4 tables for the second set of proteins. The second column of the table gives the three indexes(angular ranges) of the cell corresponding to the maximum, the third column the value of the maximum in σ units.

The close structures represent a small fraction of the overall triplets: for this reason, each dot in the figure represents one triplet, instead of 10 triplets as in Fig. 1-5. The cells that include a number of dots greater than 2 σ are grouped in a quite small region: for $table\_0$ and $table\_1$ this area is quite narrow The maximum in the table corresponds to the value. (30, 150, 150). Somewhat different is the situation for $table\_3$, representing groups of three α-helices, which presents a less defined distribution. Nevertheless, it is significant that one of the three local maxima is coincident with that of the other sets. The broadening for $table\_3$ can be partially explained by the use of a cutoff distance larger than that considered for the other table: perhaps a relevant number of secondary structure elements not interacting to each other are included in the statistics. Moreover, it has been previously observed that angles between pairs of interacting helices [24] present a quite broadened distribution that necessarily influences the distribution of triplets.

The value of maximum of the distribution, indicates that the secondary structure elements close in space tend to pack in a parallel-parallel-antiparallel fashion: the axes of two of them form a small angle (30) and are oriented in the same direction, while the third axis makes a similar angle with the other two, but its direction is reversed. Besides, the majority of the elements close in space do not lie on parallel planes, indicating that this is an overall property of non-interacting elements. Moreover, if we neglect the direction of the axes, the maximum of the distribution of angles between triplets becomes roughly (30, 30, 20), i.e. most elements of secondary structure close in space tend to pack forming similar angles.

6. Conclusions

Regularities in the arrangements of secondary structural elements have been often observed in the past, but in general the analysis has been limited to a subset of proteins and to elements directly interacting. In this paper we have considered all combinations of triplets of elements and in all cases the same rules seem to apply. In particular:

- When the elements are not close in space, i.e. their atoms are not directly interacting, the three elements tend to lie on parallel planes;

- When elements are close in space, i.e. some of their atoms are close enough to give non-covalent interactions, they show a preference to pack with dihedral angles in well defined ranges, as previously observed by other authors for packing pairs of elements. In particular, they are often oriented with two of their axes parallel and the other anti-parallel.

The two previous statements do not represent ”rules” and are in fact not strictly obeyed: they arise from a statistical analysis of experimental data and indicate a general trend more than a well defined behavior.

7. Appendix

Consider three segments a, b, and c and let α, β and γ be the angles formed by a and b, b and c, and a and c, respectively. We assume that: $\alpha \leq \beta \leq \gamma$. We show that the relation $\alpha + \beta = \gamma$ holds if and only if there exists three parallel planes each containing one of the three segments. Furthermore, if two of the three segments are coplanar then two of the three parallel planes are coincident; if all three segments lie on the same plane, then the three parallel planes are all coincident with the one containing the three segments.

Proceedings of the First International Symposium on 3D Data Processing Visualization and Transmission (3DPVT'02)
0-7695-1521-5/02 $17.00 © 2002 IEEE

Given two segments consider the two parallel planes through either segment parallel to the other segment. Recall that the equation of the plane through the line r with cosine directors (λ, μ, ν) and parallel to the line t with cosine directors (λ', μ', ν') is:

$$(x - x')h + (y - y')k + (z - z')t = 0 \qquad (1)$$

where:

$$h = \mu\nu' - \mu'\nu \; ; k = \nu\lambda' - \nu'\lambda; t = \lambda\mu' - \mu\lambda';$$

and $(x', y'z')$ are the coordinates of a point of r.
For the three pairs of segments of a triplet, the above equations define six planes that are pairwise parallel. We show that the six planes are parallel and, in fact, reduce to three planes iff $\alpha + \beta = \gamma$.
For the proof, we consider the three unit vectors a_u, b_u, c_u associated to the three segments. Without loss of generality, we choose a coordinate system with the x axis on b_u, and the xy plane containing a_u. Thus a_u, b_u have cosine directors:

$$(\cos\alpha, \sin\alpha, 0), (1, 0, 0)$$

respectively. The third unit vector c_u forming an angle of β with b_u has cosine directors:

$$(\cos\beta, \mu, \nu)$$

that satisfy the constraint:

$$\cos^2\beta + \mu^2 + \nu^2 = 1$$

Since c_u forms an angle γ with a_u we have:

$$cos\gamma = cos\alpha\cos\beta + \mu\sin\alpha;$$

Thus:

$$\mu = (cos\gamma - cos\alpha\cos\beta)/\sin\alpha$$

$$\nu = \pm\sqrt{1 - \cos^2\beta - (cos\gamma - cos\alpha\cos\beta)^2/\sin^2\alpha}$$

The two planes p_1 and p_2 defined by the equation 1 through either a_u or b_u and parallel to the other unit vector have equations $z + d = 0$, where the value of d for p_1 (p_2) is determined by imposing that the plane contains a point of a_u (b_u). Let n be the normal to both planes; n has cosine directors $(0, 0, 1)$.
Similarly, the planes p_3 and p_4 defined by the equation 1 through either b_u or c_u and parallel to the other unit vector have equations:

$$Ax + By + Cz + D = 0$$

where:

$$A = 0$$

$$B = \mp\sqrt{1 - \cos^2\beta - (\cos\gamma - \cos\alpha\cos\beta)^2/\sin^2\alpha}$$

$$C = (\cos\gamma - \cos\alpha\cos\beta)/\sin\alpha$$

The fourth coefficient D is different for the two planes and is computed by imposing that the plane passes through a point of either segment. The normal m to both planes p_3 and p_4 has cosine directors: $(A/k, \pm B/k, C/k)$, where $k = \pm\sqrt{A^2 + B^2 + C^2} = \pm\sin\beta$. The angle formed by the two normals n and m is given by:

$$\cos nm = \pm(\cos\gamma - \cos\alpha\cos\beta)/\sin\alpha\sin\beta$$

The two vectors n and m are parallel if and only if:

$$\cos nm = \pm 1$$

that is:

$$\cos\gamma = \cos\alpha\cos\beta \pm \sin\alpha\sin\beta.$$

or, equivalently,:

$$\gamma = \alpha \pm \beta \qquad .$$

Since in the hash table for all triplets of angles we have: $\alpha \leq \beta \leq \gamma$, it follows that the four planes p_1, p_2, p_3 and p_4 are parallel iff $\gamma = \alpha + \beta$. If we consider the remaining two planes p_5 and p_6 defined by equations 1 relative to a_u and c_u, along the same lines we can obtain the same result. However, this is automatically derived by the following consideration. The three normals defined as above and through the origin of the reference system form a tetrahedron. If two normals are parallel, then necessarily the third one is parallel to the other two. This completes the proof.

References

[1] Abola, E. E., Sussman, J. L., Prilusky, ,J. and Manning, N. O. (1997). Protein Data Bank archives of three-dimensional macromolecular structures. *Methods Enzymol.*, 277, 556-571.

[2] Branden, C., Tooze, J., (1999). *Introduction to protein structure* Garland.

[3] Brown, N.P., Orengo, C.A., Taylor, W.R. (1996). A protein structure comparison methodology. *Comp. Chem.*, 20, 359-380.

[4] Califano, A., R. Mohan. (1992). Multidimensional indexing for recognizing visual shapes. *IEEE Trans. on*

Pattern Analysis and machine Intelligence, 16, 4, pp. 373-392.

[5] Chothia, C., Levitt, M. and Richardson, D. (1977). Structure of proteins: packing of α-helices and pleated sheets. *Proc. Natl. Acad. Sci.*, USA 74, 4130-4134.

[6] Chothia, C., Levitt, M. and Richardson, D. (1981). Helix to helix packing in proteins. *J. Mol. Biol.*, 145, 215-250.

[7] Fischer, D., Bachar, O., Nussinov, R., and Wolfson, H. (1992). An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J. Biomol. Struct. Dyn.* 9, 769-789.

[8] Fischer, D., Tsai, C.J., Nussinov, R., and Wolfson, H. (1995). A 3D sequence-independent representation of the protein data bank. *Protein Engineering 8*, 981-997.

[9] Guerra, C., Pascucci, V. (1998). Segment matching for protein secondary structure comparison. *RECOMB99*, (abstract), Lyon, 1999.

[10] Grindley, H.M., Artymiuk, P.J., Rice, D.W., and Willett, P. (1993). Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.*, 229, 707-721.

[11] Holm, L. and Sander, C. (1996). Mapping the protein universe. *Science*, 273, 595-602.

[12] Holm, L. and Sander, C. (1996). 3-D Lookup: Fast protein structure database searches at 90% reliability. *ISMB*.

[13] W. Kabsch, C. Sander, (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, vol. 22, pp. 2577-2637.

[14] Y. Lamdan, J. T. Schwartz, H. J. Wolfson, “Affine invariant model-based object recognition”, *IEEE Trans. on Robotics and Automation*, pp. 578-589, 1990.

[15] Laskowski R.A., MacArthur M.W., Moss D.S., Thornton J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26, 283-291.

[16] Laskowski R.A., MacArthur M.W., Moss D.S., Thornton J.M. (1992). Stereochemical quality of protein structure coordinates. *Proteins*, 12, 345-364.

[17] Leibowitz, N., Fligelman, Z.Y., Nussinov, R., Wolfson, H.J. (1999). Multiple structural alignment and core detction for geometric hashing. *Proc. ISMB99*, Heidelberg, Germany, 169-177.

[18] Lesk, A. M. (1991).*Protein architecture: a practical approach.* Oxford Univ. Press, Oxford.

[19] Murzin, A.G., Brenner, S.E.,Hubbard, T., Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal Mol. Biol.*, 247, 536-540.

[20] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997). CATH- A Hierarchic Classification of Protein Domain Structures. *Structure*, 5, 1093-1108

[21] D. Platt, C. Guerra, G. Zanotti, I. Rigoutsos (2002). Global secondary structure packing angle bias in proteins. (submitted.)

[22] Richardson, J. S. (1981). Anatomy and taxonomy of protein structures. *Adv. Protein Chem.* 34, 167-183.

[23] Verbitsky, G. Nussinov, R., Wolfson H.J. (1998). Structural comparisons allowing hinge bendings, swiveling motions. *Proteins*, 34, 232-254.

[24] Walther, D., Eisenhaber, F. and Argos, P. (1996). Principles of helix-helix packing in proteins: the helical lattice superposition model. *J. Mol. Biol.*, 255, 536-553.

	set 1 (Fischer et al)		set 2 (SCOP)	
	tot.	n (d)	tot.	n(d)
$table\_0$: 3 β-strands	56,084	597 (10)	115,343	1,234 (10)
$table\_1$: 2 β-strands - 1 α-helix	112,483	3,269 (15)	233,119	6,621 (15)
$table\_2$: 1 β-strand- 2 α-helices	105,912	4,934 (20)	239,578	10,866 (20)
$table\_3$: 3 α-helices	39,526	1,511 (20)	119,013	4,278 (20)

Table 1. Number of triplets contained in each hash table for the two data sets considered.

	angular coordinates	maximum
$table\_0$: 3 β-strands	40, 150, 150	63 (6σ)
$table\_1$: 2 β-strands - 1 α-helix	10, 150, 160	135 (7σ)
	30 150 160	132 (7σ)
$table\_2$: 1 β-strand- 2 α-helices	30, 150, 160	117 (7σ)
$table\_3$: 3 α-helices	30, 40, 70	34 (7σ)
	40, 140, 160	30 (7σ)
	90, 120, 140	34 (7σ)

Table 2. Local maxima of the cell counters in each of the 4 tables for the second set of proteins.

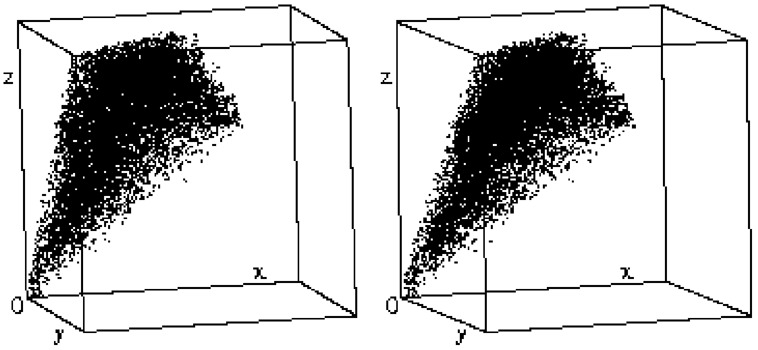

Figure 1. Stereo views of $table\_0$ (all triplets of three β-strands) Along the x, y and z axes are the dihedral angles α, β, and γ. Each dot represents ten triplets in a cell.

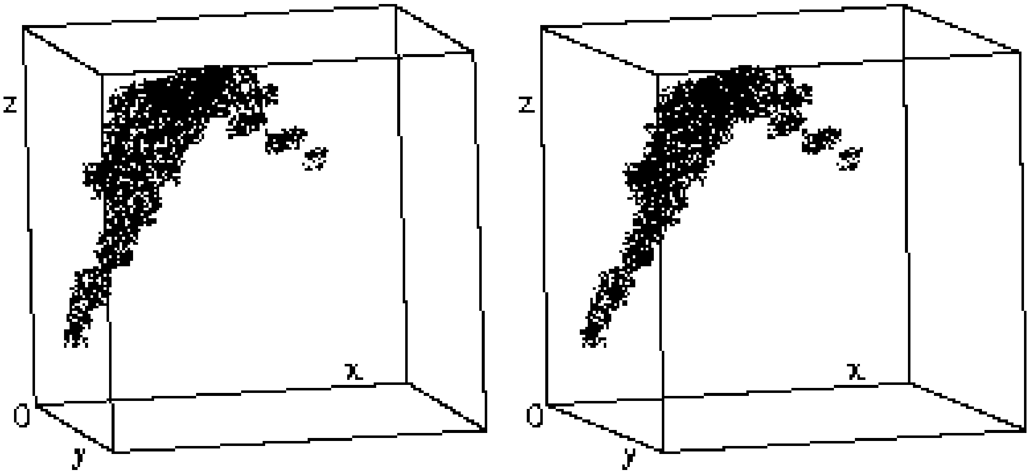

Figure 2. $table\_0$ **(triplets of three β-strands)**

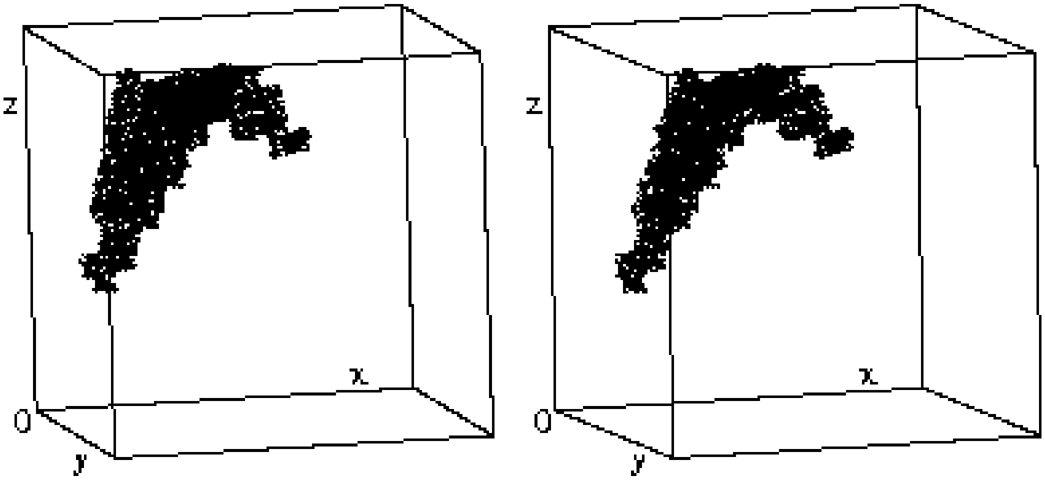

Figure 3. $table\_1$ **(triplets of two β-strands and one α-helix)**

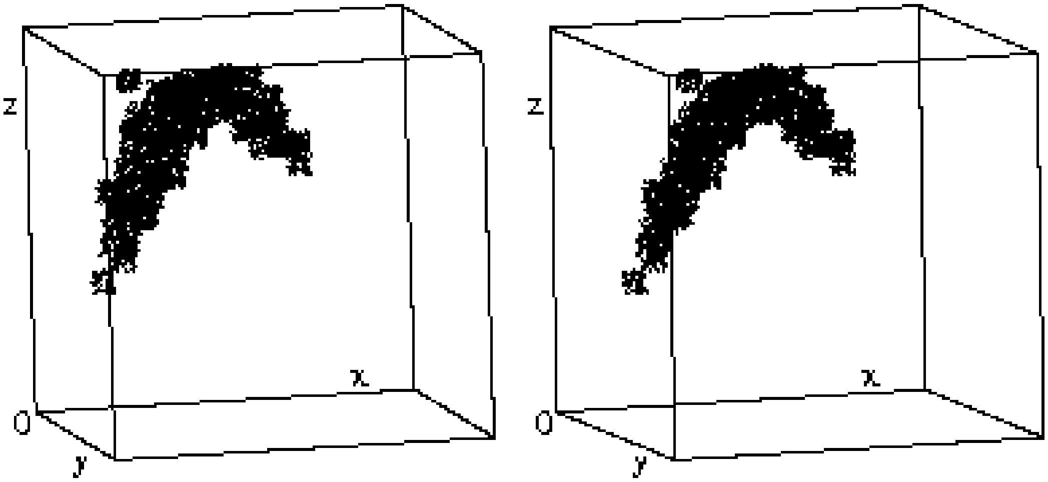

Figure 4. $table\_2$ **(triplets of one β-strand and two α-helices)**

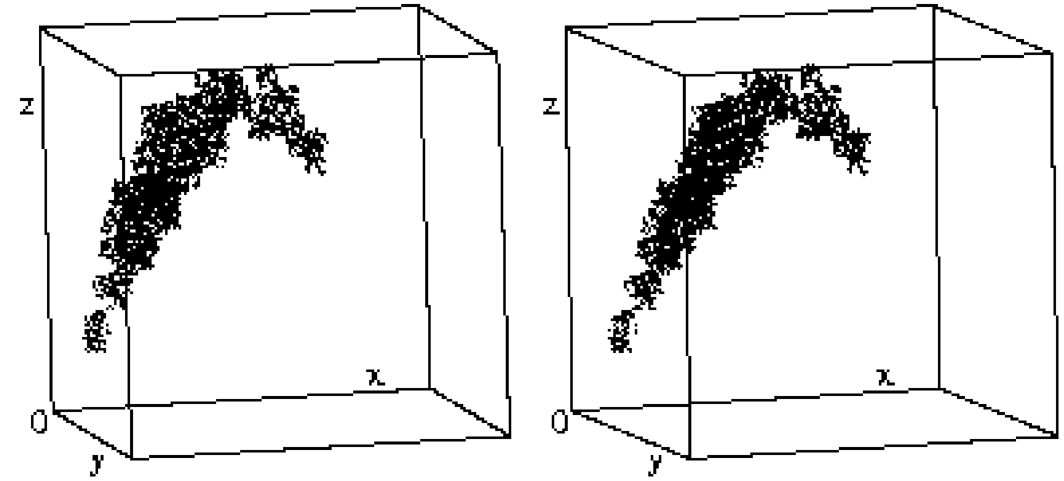

Figure 5. $table\_3$ **(triplets of three α-helices)**

Figures 2-5 show stereo views of the 4 hash tables.

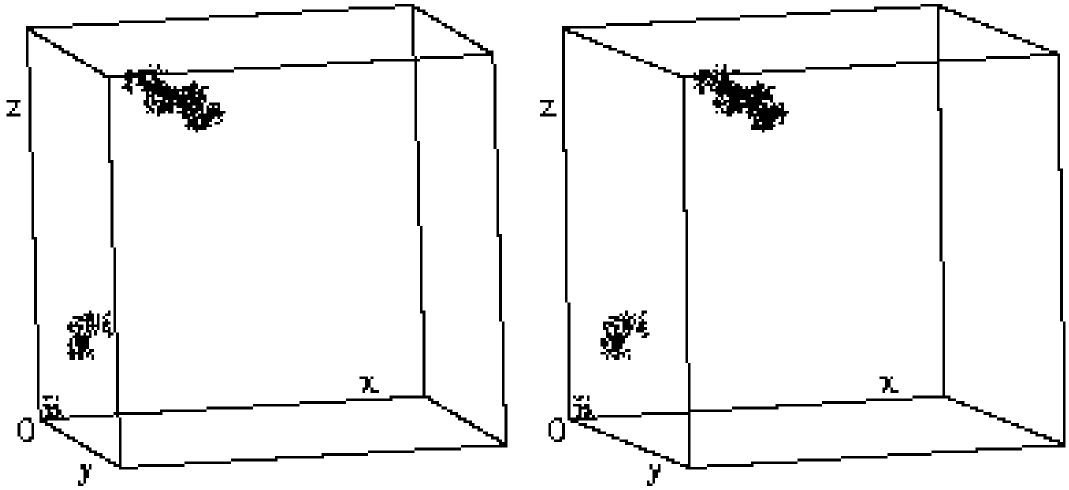

Figure 6. $table\_0$ **(triplets of three β-strand)**

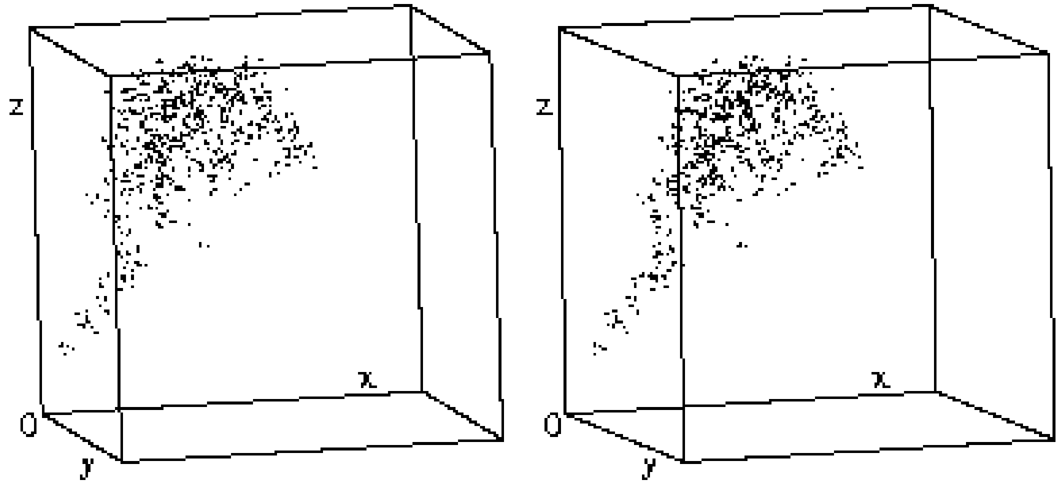

Figure 7. $table\_3$ **(triplets of three α-helices)**

Figures 6-7 display the angular distribution of the triplets of elements close in space.