Mining Over-Represented 3D Patterns of Secondary Structures in Proteins

Matteo Comin*
*Department of Information Engineering
University of Padova, Italy
Padova 35131, Italy
ciompin@dei.unipd.it

Concettina Guerra*†
†College of Computing
Georgia Institute of Technology
Atlanta, GA 30332-0280 USA
guerra@dei.unipd.it

Giuseppe Zanotti‡
‡Department of Chemistry and VIMM
University of Padova
Padova 35131, Italy
giuseppe.zanotti@unipd.it

ABSTRACT

We consider the problem of finding over-represented arrangements of Secondary Structure Elements (SSEs) in a given dataset of representative protein structures. While most papers in the literature study the distribution of geometrical properties, in particular angles and distances, between pairs of interacting SSEs, in this paper we focus on the distribution of angles of all quartets of SSEs and on the extraction of over-represented angular patterns. We propose a variant of the Apriori method that obtains over-represented arrangements of quartets of SSEs by combining arrangements of triplets of SSEs. This specific case will pose the basis for a natural extension of the problem to any given number of SSEs. We analyze the results of our method on a dataset of 300 non redundant proteins.

1. INTRODUCTION

The problem of finding recurrent three-dimensional patterns in proteomic data is of biological interest and therefore has been studied in different contexts and with various techniques [6, 16]. In fact, although the information on the fold of a protein is already totally contained in its amino acid sequence, the calculation of the minimal energy among all the possible conformations is a task which is overwhelming even for the fastest computer. For this reason, a great deal of efforts has been spent over the years in order to disclose hidden rules about the organization of secondary structure elements [2, 8].

A simplified description of the three-dimensional protein structure is that of considering it as an arrangement of SSEs. The possible ways SSEs aggregate in space is someway limited: all protein structures, till now determined, can be grouped in a relatively limited number of different folds. Moreover, it is well known that interacting SSEs show marked preferences in their reciprocal orientation. For example, interacting β-strands are very often organized in sheets, where each strand is disposed in a roughly parallel or antiparallel orientation with respect to the neighboring ones [3]. Preferences between interacting α-helices have been also studied extensively and general rules extracted [4, 7, 15]. Nevertheless, it has been shown that the expected uniform random distribution of angles is actually biased toward angles near 90^o[1]. When this geometric bias was taken into account, the observed peaks in the helix-helix angle distribution were significantly attenuated: correcting for statistical bias, the true preference for particular packing angles in soluble proteins is not as strong as previously thought.

Moreover, the relative arrangement of non-interacting SSEs in space is less obvious [11]. In order to analyze their global disposition, in the past we have conducted a statistical analysis on the occurrences of triplets of SSEs [10, 17]. We found that the distribution is far from being random, with a marked preference for specific angle combinations. This knowledge could be used to guide the engineering of stable protein modules or to predict the three-dimensional structure [13].

The present study extends the previous analysis, taking into account quartets of SSEs. It presents an analysis of the distribution of secondary structures within a selected set of non redundant proteins. It constructs frequent patterns of k elements (or itemsets of size k) by joining frequent patterns of size $k-1$.

2. PROBLEM DESCRIPTION

Given a data-set of proteins structures, we address the problem of finding over-represented arrangements of SSEs in terms of geometrical properties. Most papers in the literature study the distribution of geometrical properties, in particular angles, between pairs of interacting SSEs [14, 18]. Here we focus on over-represented configurations consisting of more than two SSEs and analyze the distribution of angles of such configurations. Our task is to design a framework to extract over-represented arrangements of k SSEs, by combining the results obtained with arrangements of $k-1$ SSEs. We discuss in details how to obtain over-represented arrangements of four SSEs by using the distribution of triplets of SSEs instead of generating all quartets of SSEs from the data set. This specific case will pose the basis for a natural extension of the problem to any given number of SSEs.

Each protein structure of the dataset is given with the list of SSEs ordered according to the backbone chain. A line segment is associated to each SSE. For a β-strand the segment

BIOKDD'07, August 12, 2007, San Jose, California, USA.

is the best fit segment of the set of atoms of the strand, for an α-helix it is the best fit axis. For the purpose of our analysis, a line segment is assumed to be a unit vector applied in the origin of a reference system in three-dimensional space. Thus a protein is a list of m unit vectors $(s_1, \cdots, s_m)$.

An arrangement of SSEs is described in terms of the angles formed by all pairs of corresponding vectors. Let α_{hk} be the dihedral angle of s_h and s_k, $0^o \leq \alpha_{hk} \leq 180^o$. A triplet of SSEs (s_{i1}, s_{i2}, s_{i3}), with $i1 < i2 < i3$, is described by three angles α_{12}, α_{13} and α_{23} satisfying the triangle inequality. A quartet of SSEs $S = (s_{i1}, s_{i2}, s_{i3}, s_{i4})$, with $i1 < i2 < i3 < i4$, gives rise to 6 dihedral angles $Q = (\alpha_{12}, \alpha_{13}, \alpha_{23}, \alpha_{24}, \alpha_{34}, \alpha_{14})$. A schematic representation of the unit vectors derived from a quartet of SSEs can be found in Figure 1. It is easy to show that, in the general case, the six angles are not completely independent. More precisely, given 5 of the α_{hk} angles, the sixth angle can take only one of two possible values. The derivation of such values is omitted for lack of space. Furthermore, when three out of four segments are mutually orthogonal then one of the angles formed by the fourth segment with the three segments is uniquely determined by the other two angles. Another important question, that will be considered in section 4, is whether it is possible to superimpose, by a rigid transformation, two quartets forming the same angles.

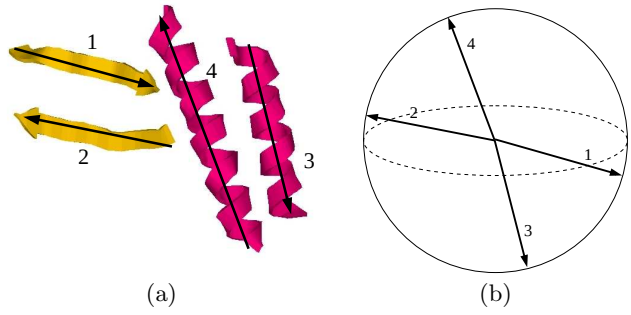


Figure 1: (a) An example of vector discretization for a quartet of SSEs. (b) The unit vectors translated to the origin (into the unit sphere).

The angular values are discretized into uniform intervals, with every interval represented by an integer. More precisely, in our work the range $0^o - 180^o$ is divided into 10 intervals, and an angle α represented by the integer i such that $i * 18^o \leq \alpha < (i+1) * 18^o$. Thus a quartet of SSEs is represented by 6 integer values each in the range [0,10]. In the following we refer to the discretized angles simply as angles.

3. DISCOVERY OF OVER-REPRESENTED PATTERNS

Our approach is similar to the Apriori algorithm used for data mining applications. Apriori finds frequent associations of attributes of k elements (or itemsets of size k) by joining frequent associations of itemsets of size $k-1$. Similarly, our algorithm finds over-represented arrangements of quartets of segments from over-represented triplets of segments; it does so by joining over-represented triplets of angles to obtain over-represented sextuplets of angles.

However, our approach differs substantially from Apriori in the way the patterns are joined together to obtain patterns of larger size. At the basis of the Apriori mining algorithm is the anti-monotone property that states that all non empty subsets of a frequent set must also be frequent. In other words, if an itemset cannot pass the test of being frequent, then all its supersets will fail the same test.

The anti-monotone property does not hold for the angles formed by sets of segments. Consider a frequent sextuple of angles $Q = (\alpha_{12}, \alpha_{13}, \alpha_{23}, \alpha_{24}, \alpha_{34}, \alpha_{14})$ and all quartets S of segments with angles Q. Even though Q is frequent, it is possible that triplets that are subsets of Q are not frequent. This is the case of the triplet of angles $T = (\alpha_{13}, \alpha_{23}, \alpha_{24})$ that cannot be formed (in the general case) by a triplet of segments which is a subset of an element of S, because the three angles involve all 4 segments of a single element of S. However, there are four triplets of angles subsets of a frequent sextuple Q that must be frequent. These are $(\alpha_{12}, \alpha_{13}, \alpha_{23})$ and $(\alpha_{23}, \alpha_{24}, \alpha_{34})$, $(\alpha_{13}, \alpha_{14}, \alpha_{34})$ and $(\alpha_{12}, \alpha_{14}, \alpha_{24})$. Indeed, the four triplets are formed by the four different ways of choosing three segments out of four. Frequent triplets of angles are extracted by comparing the observed frequencies of triplets of angles with those of randomly distributed vectors.

We now describe our mining procedure. We start by giving an overview of our approach, and then describe each step in detail.

PROCEDURE: *Pattern Discovery*

1. Initialization: From the given protein data set generate the set A of all ordered triplets of angles associated to ordered triplets of SSEs, sorted according to the order along the backbone.

2. Build an hash table indexed by the triplets of angles that stores all triplets of segments.

   Derive the 3D histogram of the distribution of the triplets of A from the hash table. The histogram has $b = 10$ bins along each axis, for a total of b^3 bins or cells.

3. Build the distribution of triplets of angles of random unit vectors and derive the corresponding 3D histogram.

4. Based on the deviation between the histogram of observed triplets of angles and that of random triplets, determine the subset $C \subset A$ of triplets that are over-represented.

5. Join step: construct candidate sextuples of angles from triplets of C.

6. Verification step: prune candidate sextuples to find the over-represented ones.

3.1 Building the Hash Table

We build a four-dimensional hash table with the following index structure: for a given triplet of vectors, three indexes are given by the quantized values of the angles of the triplet, the fourth index depends on the composition of the triplet in terms of the number and position of helices and strands. This index, called *triplet type*, is used when a separate analysis is requested for helices and strands. The size of the cells of the table is the same as the binsize for the histograms.

Each cell of the table contains a list of records, one for every triplet that hashed into it. The following procedure inserts protein P into the hash table and is a variant of the one described in [5].

PROCEDURE: *Insert Protein*
Given protein P, all triplets of secondary structures of P are examined and for each triplet (p_u, p_v, p_z) with $u < v < z$ the following steps are executed:

i. Compute the angles $(\alpha_{uv}, \alpha_{vz}, \alpha_{uz})$ and determine *triplet type*.

ii. Access the cell of the hash table at the location indexed by *triplet type* and by the quantized values of $(\alpha_{uv}, \alpha_{vz}, \alpha_{uz})$.

iii. Append to the list of records at that cell a new record that contains:

- the name of protein P.
- the identifier of each secondary structure element of the triplet.

The above procedure is repeated for all proteins in the data set. The construction of the table is computationally intensive. However, the number of proteins of the dataset to be inserted is relatively small.

3.2 Generating Random Triplets

The selection of the frequent triplets is the crucial point of the overall procedure: a wrong selection can produce a meaningless starting point that can lead to unreliable results. Thus this step must be carefully designed. We observe that the distribution of geometric properties of triplets strongly depends on the features considered. To avoid the bias due to the features considered, we compute the null distribution of such properties.

The random generation of a triplet of angles is decomposed into the generation of three versors. A versor is a vector of unit length that we assume to be in the semisphere identified by a positive value of the z coordinate. A versor is now uniquely determined by two parameters: its coordinate $z \in [0, 1]$, and its Azimuth $\beta \in [0, 2\pi]$. We have already observed that the triangular inequality holds for any three angles α, β, γ of a triplet of segments; it translates into the following three constraints: $\alpha + \beta \geq \gamma$, $\alpha + \gamma \geq \beta$, $\beta + \gamma \geq \alpha$. This implies that not all cells of the hash table can be populated by triplets of segments; in other words, there are cells that will remain empty. Furthermore, some cells can only be partially populated. Thus when deciding which cells correspond to most frequent triplets of angles, we have to take into account the above consideration and normalize by the volume of the region of the cell that can in fact be populated. This region is determined by considering that the above three constraints correspond to the equations of the three boundary planes $\alpha + \beta = \gamma$, $\alpha + \gamma = \beta$, $\beta + \gamma = \alpha$ delimiting the populated area in 3D space. By intersecting each cell of the 3D array with the three boundary planes we find out which region, if any, has to be excluded and consequently compute the volume V_c of the populated region. Thus the frequency of a cell (α, β, γ) will be: $Count(\alpha, \beta, \gamma)/V_c(\alpha, \beta, \gamma)$.

Given a data set of n real proteins to analyze, we generate the distribution of angles of n sets of random vectors, each corresponding to a protein of the dataset and containing the same number of SSEs of such protein.

The generation of the ensemble of random vectors is repeated several times and, at the end, each cell of the hash table has the average of the values of the cell over all random generations. This results in a 3D histogram representing all triplets of angles, where each triplet has attached a mean and a variance. For the selection of over-represented angles we experimented with different selection policies. To preserve a reasonable number of candidates we select the configurations of angles that have a frequency above the mean.

3.3 Join and Verification Steps

The operation *join* merges four frequent triplets $(\alpha_{12}, \alpha_{13}, \alpha_{23})$ and $(\alpha_{23}, \alpha_{24}, \alpha_{34})$, $(\alpha_{13}, \alpha_{14}, \alpha_{34})$ and $(\alpha_{12}, \alpha_{14}, \alpha_{24})$ into the candidate sextuple $(\alpha_{12}, \alpha_{13}, \alpha_{23}, \alpha_{24}, \alpha_{34}, \alpha_{14})$. The four triplets to be merged are such that the last angle of the first triplet is the same as the first angle of the second; the second element of the first triplet is the same as the first element of the third triplet, and so on. Recall that all angles are discretized. Furthermore, note that two triplets may coincide.

Once a candidate sextuple has been identified in step 5, the verification procedure checks that there is in fact a statistically significant number of quartets of vectors with that sextuple of angles. This number will provide the actual frequency of the sextuple of angles. The verification step is needed because some triplets of segments contributing to the count of frequent triplets of angles cannot be joined into quartets of segments. For instance, the two triplets might be from different proteins. Two triplets of segments (s_1, s_2, s_3) and (t_1, t_2, t_3) associated to SSEs of the same protein and forming angles $(\alpha_{12}, \alpha_{13}, \alpha_{23})$ and $(\alpha_{23}, \alpha_{24}, \alpha_{34})$, respectively, can be joined into a quartet of segments with angles $(\alpha_{12}, \alpha_{13}, \alpha_{23}, \alpha_{24}, \alpha_{34}, \alpha_{14})$ if $(s_2 = t_1$ and $s_3 = t_2)$, i.e. the last two segments of the first triples coincide with the first two of the second triples. Two such triplets of segments are called "consistent" and they contribute one to the frequency count of the associated sextuple.

To efficiently search for consistent triplets, we use the hash table built in step 2 containing the triplets of segments of all proteins. The frequency or count of a candidate sextuple $(\alpha_{12}, \alpha_{13}, \alpha_{23}, \alpha_{24}, \alpha_{34}, \alpha_{14})$ is determined as follows. Access the hash table at the cells $E1$ and $E2$ indexed by $(\alpha_{12}, \alpha_{13}, \alpha_{23})$ and by $(\alpha_{23}, \alpha_{24}, \alpha_{34})$ respectively. For each triplet (s_1, s_2, s_3) in $E1$ with associated protein name P search in $E2$ for all triplets (s_2, s_3, t), with any arbitrary t, of the same protein P. For each such triplet increment the count if the last angle α_{14} is compatible with the candidate sextuple under examination.

4. SPATIAL ARRANGEMENTS OF VECTORS WITH THE SAME ANGULAR PATTERN

It is interesting to determine whether two sets of vectors with the same angular pattern can be superimposed by a 3D rigid transformation, or whether the spatial conformations of the two sets of vectors differ in their 3D shape. Protein structure comparison algorithms that align SSEs also use a shape similarity measure based on the rigid superposition of the structures [21].

We define equivalent two sets of vectors that can be superimposed by a rigid transformation. We first look at the case

of triplets of vectors (a, b, c) and their angles (α, β, γ). We recall that the unit vectors are applied into the origin O of a coordinate system without considering the actual location of the SSE in 3D space. It is easy to see that there are two distinct triplets of vectors (a, b, c) and (a, b, c'), where c and c' are non parallel vectors, forming a given triplet of angles (α, β, γ). For example (see Figure 2), consider four vectors forming a regular pyramid with vertex in 0; label two opposite vectors of the pyramid a and b and the other two c and c'. The two triplets of vectors (a, b, c) and (a, b, c') have the same angles but are non equivalent since they are one the mirror of the other.

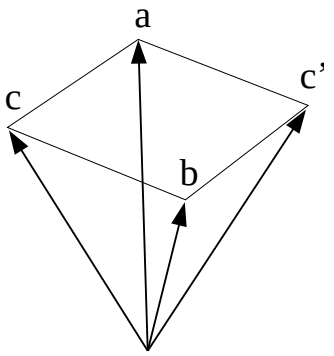


Figure 2: An example of two triplets, (a, b, c) and (a, b, c'), with the same pairwise angles, one the mirror of the other.

Perhaps more convincing is the following proof. All vectors forming a given angle δ with a given vector v are rays of the cone with vertex in O and forming δ angle with v. Given two vectors a and b forming angle α, a third vector forming angles β and γ with a and b, respectively, is at the intersection of two cones. Two cones intersect at either one or two lines. In the first case, the only possible triplet consists of vectors lying on the same plane ($\alpha + \beta = \gamma$); in the latter there are two non parallel vectors c and c' corresponding to two distinct triplets.

In conclusion, a triplet of angles (α, β, γ) corresponds to two spatial arrangements of unit vectors (a, b, c) and (a, b, c') that are one the mirror of the other; equivalently, there exists a transformation with determinant -1 mapping one triplet of vectors into the other. Loosely speaking, although two triplets of vectors cannot be superimposed by a rotation (with determinant 1), they correspond to a similar configuration in terms of angles.

If we extend this argument to quartets of vectors, the number of non equivalent arrangements doubles. Consider a sextuple of angles $(\alpha_{12}, \alpha_{13}, \alpha_{23}, \alpha_{24}, \alpha_{34}, \alpha_{14})$. To construct all non equivalent quartets of vectors corresponding to it, we follow a build-up approach. From the first three angles $(\alpha_{12}, \alpha_{13}, \alpha_{23})$ we construct either one triplet of vectors (a, b, c) or two (a, b, c) and (a, b, c'). Then, we derive the last vector d. There are four possible cases:

1. If $\alpha_{12} + \alpha_{23} = \alpha_{13}$ and $\alpha_{23} + \alpha_{34} = \alpha_{24}$, then there is a single triplet (a, b, c) and a single triplet (b, c, d). Thus, there exists a unique arrangement of four vectors.

2. If $\alpha_{12} + \alpha_{23} = \alpha_{13}$ but $\alpha_{23} + \alpha_{34} < \alpha_{24}$, then two distinct arrangements are possible, (a, b, c, d) and (a, b, c, d').

3. Otherwise, if $\alpha_{23} = \alpha_{34}$ then four different arrangements are possible, with three distinct vectors as last component of the quartet: (a, b, c, d), (a, b, c, d'), (a, b, c', d') and (a, b, c', d'').

4. In all other cases, the following four arrangements are possible: (a, b, c, d), (a, b, c, d'),(a, b, c', d'') and (a, b, c', d''').

5. RESULTS AND DISCUSSION

We selected a set of 300 non-redundant proteins from different families and computed the set of all triplets of SSEs and their associated linear segments. To include only significant SSEs, we required helices to have at least seven residues, corresponding to two complete turns of a regular helix. Strands were required to have at least three residues for proper fitting of a vector to the C_α coordinates. Secondary structures are represented by the best-fit line segments. A Singular-Value Decomposition (SVD) routine is used to associate a segment to each α-helix and β-strand [9]. Using this dataset we constructed the hash table of triplets of angles and compared it with the random distribution to determine the cells that deviate significantly from the corresponding cells for the random data. The hash table contains 520 non empty cells (containing a total of 398,853 triplets of vectors), of which 242 were selected as frequent (corresponding to 189,270 triplets). The histogram of the triplets of angles selected as frequents is shown in Figure 3.

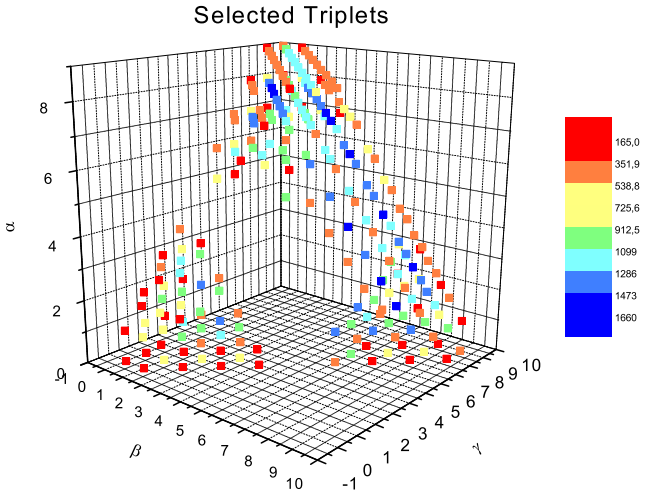


Figure 3: 3D histogram of the distribution of selected angles. Each axis represents an angle and the frequency of each triplet follows the color coding.

5.1 Analyzing Over-represented Patterns of Angles

The pattern discovery process finds a set of over-represented arrangements of four SSEs. Each arrangement is described by six ordered angles, i.e. an angle corresponds to a specific pair of SSEs which is identified by the sequential order of SSEs along the primary structure. Thus two arrangements forming the same six angles, but in a different order, correspond to two different patterns, even though they can be considered geometrically equivalent. We address this issue

by merging together patterns composed by the same angles and ignoring the relative order of angles.

By merging patterns, the discovery procedure selects a set of 785 over-represented patterns, formed by 485,021 quartets of segments, out of 2,262 patterns and more than 3,000,000 quartets obtained by the exhaustive search. The top pattern is composed by the discretized angles $(1, 2, 3, 7, 8, 9)$, corresponding to angles in the ranges $(18^o - 36^o, 36^o - 54^o, 54^o - 72^o, 126^o - 144^o, 144^o - 162^o, 162^o - 180^o)$, and has a frequency of 6,439, the top second has similar angles, (1,2,7,8, 8,9), and a smaller frequency of 5,780. The frequency count drops dramatically after the first few patterns. It is interesting to notice that the top 11 angular patterns (out of 785) cover about 10% of the quartets; coverage of the quartets of about 20% is obtained by 29 patterns and that of 50% by 122 patterns.

The overall discovery procedure is relatively fast; it takes approximately 20 minutes on a standard PC (AMD Athon 2.6 GHz). On the same machine, the exhaustive generation of all possible quartets of SSEs takes more than 3 days.

We observed that over-represented patterns of angles tend to form clusters in the six-dimensional space corresponding to six angles. Thus, we further analyzed the set of over-represented patterns by clustering them using as distance the Euclidean distance between angular patterns in six-dimensional space.

We experimented with different clustering algorithms and different numbers of clusters and, based on the measure of silhouette [12], we selected the k-means algorithm with 3 clusters. Clusters 1 and 3 contain, respectively, the first and second most frequent pattern. Cluster 2 contains the configuration of angles $(0, 1, 1, 2, 2, 3)$ that appears at position 16 in the overall ranking of patterns. The top patterns for each cluster are shown in Figure 4. In Figure 5 the cluster separation is highlighted by plotting the distribution of distances between the centroids of each cluster and the elements of all 3 clusters.

In all clusters the angles vary from 0^o to 72^o and from 126^o to 180^o, while values between 80^o and 100^o are completely absent. This is not surprising because the distribution is biased by the presence of many interacting SSEs. For example, in parallel and anti-parallel β-sheets, each β-strand typically forms a small angle with the two nearby strands. The same is true for interacting α-helices, that pack forming small angles; furthermore, they are hardly found perpendicular to each other [19, 20]. Cluster 2 is the smallest one, with 32,988 elements; it contains SSEs characterized by the same orientation: in fact, the angles between all pairs of SSEs are in the range 0^o to 72^o. The other two clusters are more densely populated; cluster 1 has 221,879 elements and cluster 3 has 230,154 elements. In these two clusters the SSEs are arranged with three SSEs with the same orientation and the other one with the opposite (cluster 1) or with two SSEs in the same orientation and the other two in the opposite orientation. The smaller number of elements in cluster 2 reflects the tendency of SSEs that are close in space to form anti-parallel configurations.

If we restrict the analysis to homogenous configurations, i.e. those containing four strands or four helices, we obtain similar results for the clusters, but with a preference for anti-parallel pairs, corresponding to the top ranked pattern of angles $(1, 2, 7, 8, 8, 9)$.

The over-represented patterns considered so far have included the SSEs of the selected set of proteins, regardless of their distances. We now consider homogenous patterns of SSEs that are close in space; we define two SSEs to be in contact if the distance between the mid-points of their associated vectors is less than a given threshold (18 in our analysis). Figure 6 shows the number of pairs of vectors in contact for the top configuration. It is interesting to notice that in all cases at least one pair of vectors is in contact, and very often three or more vectors are in contact. Notice that the use of the same threshold penalizes helices, because of their bigger steric hindrance [18]. Nevertheless, more than 65% of the elements have at least two SSEs in contact. To better appreciate the proximity of these over-represented configurations, in Figure 7 we show different examples of four strands, with angles $(1, 2, 7, 8, 8, 9)$. In all these examples the four strands are in contact. Although they display different arrangements, their pairwise angles are similar, thus they fall into the same cell of the hash table. These patterns of angles are obtained with SSEs from the same β-sheet (Figure 7(c)), as well as from different β-sheets (Figure 7(a) and (b)). The fact that most, but not all,

α_0	α_1	α_2	α_3	α_4	α_5	Frequency
1	2	3	7	8	9	6,439
1	2	3	7	8	8	5,586
1	1	2	7	8	9	4,657
1	2	3	6	8	9	4,085
1	2	3	7	7	8	3,728
1	1	2	6	7	8	3,648
1	2	2	7	8	9	3,401
1	2	3	6	7	9	2,958
1	1	2	8	8	9	2,833
1	1	2	7	8	8	2,494

Cluster 1

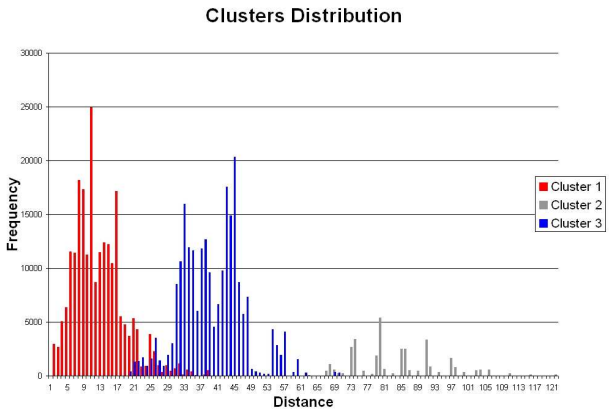
α_0	α_1	α_2	α_3	α_4	α_5	Frequency
0	1	1	2	2	3	2,623
1	1	1	2	2	3	2,162
0	1	1	1	2	2	2,123
0	1	1	2	3	3	1,667
0	1	1	2	2	2	1,445
0	1	1	1	1	2	1,311
0	1	1	1	2	3	1,246
0	1	2	2	3	3	1,178
1	1	1	2	3	3	1,039
1	1	2	2	2	3	1,010

Cluster 2

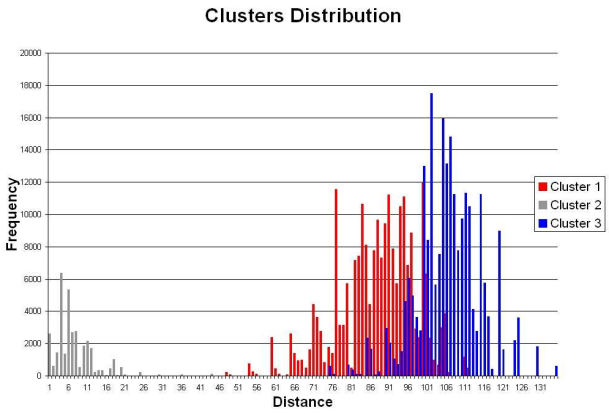
α_0	α_1	α_2	α_3	α_4	α_5	Frequency
1	2	7	8	8	9	5,780
1	3	6	7	8	9	5,100
1	2	6	7	8	9	4,437
2	3	6	7	8	9	3,884
1	3	7	7	8	8	3,831
1	2	7	7	8	9	3,637
1	1	7	8	8	9	2,916
1	3	6	7	8	8	2,572
1	3	7	7	8	9	2,544
0	3	7	7	8	8	2,525

Cluster 3

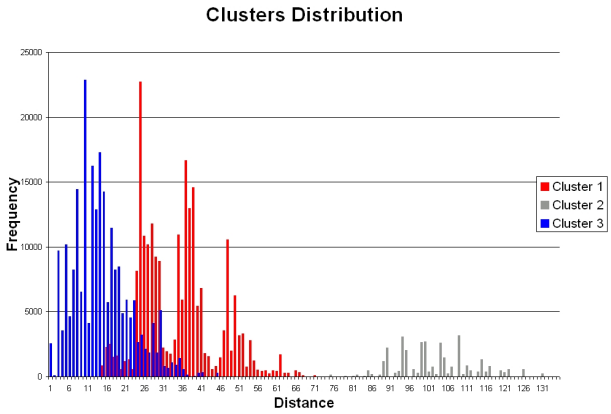
Figure 4: The ten top frequent patterns for the three clusters.



(a)
Distribution of distances from the centroid of Cluster 1.

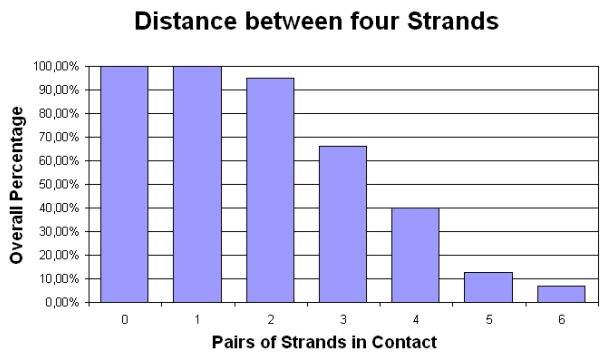


(b)
Distribution of distances from the centroid of Cluster 2.

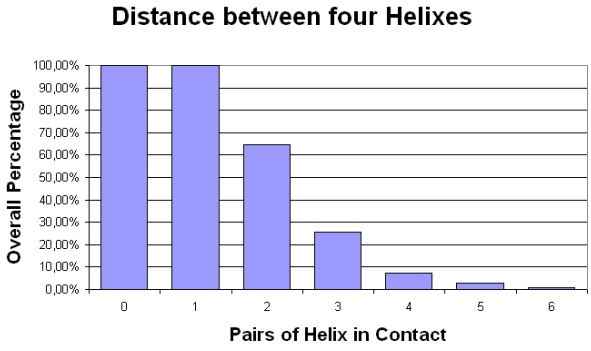


(c)
Distribution of distances from the centroid Cluster 3.

Figure 5: Distance distributions between centroids of clusters.



(a) Number of pairs in contact in quartets of strands.



(b) Number of pairs in contact in quartets of helixes.

Figure 6: Number of pairs of segments in contact.

(a) (b) (c)

Figure 7: Three examples of the pattern of angles (1,2,7,8,8,9) composed by all strands: (a) Protein 1hpl, SSE: 16-17-18-20; (b) Protein 1acc, SSE: 0-1-2-3; (c) Protein 1aor, SSE: 4-6-8-12.

SSEs are close in space consolidates the idea that arrangements of angles are influenced by atomic interactions, either directly or through other SSEs that do not explicitly belong to the quartet. Finally, as illustrated in Figure 7, secondary structure elements belonging to the same quartet do not necessarily correspond to similar structures, i.e. structures that can be superimposed by rotation and translation. For this reason it is impossible to associate a three-dimensional motif, or a group of motifs, to the most frequent quartets described above. The biological significance of the distributions observed needs a deepener investigation.

6. CONCLUSIONS

We have proposed an efficient algorithm to extract over-represented quartets of SSEs, that avoids the exhaustive generation of patterns. We have shown that a careful analysis of the angular bias of random vectors is essential in the determination of over-represented arrangements of secondary structures. This study provides a generalized framework that can be easily extended to patterns composed by more than four SSEs. The knowledge of over-represented patterns could be used to guide the engineering of stable protein modules or to predict their three-dimensional structures. Other applications can be designed by replacing the null distribution with that of a specific family of proteins.

7. REFERENCES

[1] Bowie J.U. Helix packing angle preferences. *Natural Structural Biology*, 4:915-917, 1997.

[2] Brenner, S.A. Predicting the conformation of proteins from sequences. Progress and future progress. *J. Mol. Recognit.*, 8(1-2):9-28, 1995.

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

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

[5] Comin M., Guerra C., Zanotti G. PROuST: a comparison method of three-dimensional structures of proteins using indexing techniques. *J. Comput. Biol.*, 11:1061-1072, 2004.

[6] Efimov, A.V. Structural trees for protein superfamilies. *Proteins*, 28(2):241-60, 1997.

[7] Efimov, A.V. Complementary packing of alpha-helices in proteins. *FEBS Lett.* , 463(1-2):3-6, 1999.

[8] Eisenhaber, F., Persson, B. and Argos, P. Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. *Crit. Rev. Biochem. Mol. Biol.*, 30(1):1-94, 1995.

[9] Gerstein, M. A Resolution-Sensitive Procedure for Comparing Protein Surfaces and its Application to the Comparison of Antigen-Combining Sites. *Acta Cryst.* , A48, 271-276, 1992.

[10] Guerra C., Lonardi S., and Zanotti G. 3D Matching of Proteins based on Secondary Structures. *Proc. IEEE Symposium on 3DPVT*, Padova, pages 812-821, 2002.

[11] Hespenheide BM, Kuhn LA. Discovery of a significant, nontopological preference for antiparallel alignment of helices with parallel regions in sheets. *1: Protein Sci.*12(5):1119-1125, 2003.

[12] Kaufman, L., Rousseeuw, P. J. (1990). Finding Groups in Data - An Introduction to Cluster Analysis. *Wiley Series in Probability and Mathematical Statistics*, 1990.

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

[14] Lee S, Chirikjian GS. Interhelical angle and distance preferences in globular proteins. *Biophys J.*, 86(2):1105-1117, 2004.

[15] Nakamura, H. Roles of electrostatic interaction in proteins. *Q. Rev. Biophys.*, 29(1):1-90, 1996.

[16] Persson, B. Bioinformatics in protein analysis. *EXS.*, 88:215-31, 2000.

[17] Platt D.E., Guerra C., Zanotti G., Rigoutsos I. Global secondary structure packing angle bias in proteins. *Proteins: Structure, Function, and Genetics*, 53(2):252-261, 2003.

[18] Reddy, B., and Blundell, T. Packing of secondary structural elements in proteins. Analysis and prediction of inter-helix distances. *J. Mol. Biol.*, 233:464-479, 2003.

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

[20] Walther D, Springer C, Cohen FE. Helix-helix packing angle preferences for finite helix axes. *Proteins*, 33(4):457-9, 1998.

[21] Yona G. and Kedem K. The URMS-RMS hybrid algorithm for fast and sensitive local protein structure alignment. *Journal of Computational Biology*, 12:12-32, 2005.