Journal of Bioinformatics and Computational Biology
Vol. 6, No. 6 (2008) 1067–1087

ICP Imperial College Press
www.icpress.co.uk

MINING OVERREPRESENTED 3D PATTERNS OF SECONDARY STRUCTURES IN PROTEINS

MATTEO COMIN*,† and CONCETTINA GUERRA*,‡,§

*Department of Information Engineering
University of Padova, 35131 Padova, Italy
†ciompin@dei.unipd.it
‡guerra@dei.unipd.it

§College of Computing, Georgia Institute of Technology
Atlanta, GA 30332-0280, USA

GIUSEPPE ZANOTTI

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

Received 8 November 2007
Revised 5 March 2008
Accepted 9 April 2008

We consider the problem of finding overrepresented 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 overrepresented angular patterns. We propose a variant of the Apriori method that obtains overrepresented 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 nonredundant proteins. Supplementary material is available at http://www.dei.unipd.it/nciompin/papers/CGZ-jbcb-suppl.pdf/.

Keywords: Protein's structures; secondary structures; computational geometry; overrepresented motifs.

1. Introduction

The problem of finding recurrent three-dimensional (3D) patterns in proteomic data is of biological interest, and therefore has been studied in different contexts and with various techniques.[1,2] Although information on the fold of a protein is already totally contained in its amino acid sequence, calculation of the minimal energy among all possible conformations is an overwhelming task even for the

fastest computer. Developing methods for discovering statistical rules that govern structures may help answer basic biological and biochemical questions. Several statistical models have been introduced to evaluate the statistical significance of arrangements of points or substructures in 3D space.[3–5] Structure prediction methods[6,7] may be aided by the statistical analysis presented here. The characterization and significance of global constraints on secondary structure elements (SSEs) can be of help in the definition of scoring functions to assess the quality in predicted models and correct packing defects. Furthermore, such knowledge could be used to guide the engineering of stable protein modules: the detection of overrepresented and underrepresented arrangements may be taken as an indication of physical possibility or impossibility of artificial structures.[8] In protein structure comparison, hashing techniques based on SSEs are often used for the efficient retrieval of similarity information for a target structure against all structures of the Protein Data Bank (PDB) or a subset of representative structures.[9–11] The statistical properties of geometric invariants can assist in defining hashing strategies. For this reason, a great deal of effort has been spent over the years to disclose hidden rules about the organization of SSEs.[12,13]

A simplified description of 3D protein structure considers it as an arrangement of SSEs. The possible ways SSEs aggregate in space are somewhat limited: all protein structures determined up to now can be grouped in a relatively limited number of different folds. Moreover, it is well known that interacting SSEs show marked preferences in their relative 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.[14] Packing angle preferences between interacting α-helices have also been studied extensively and general rules extracted.[15–23] The angle of interacting SSEs is measured between linear segments associated to secondary structures, such as the axis of an α-helix or the endpoints of a β-strand. Interaction between α-helices has been defined according to different criteria. The most common criterion is that the length of the line of closest approach perpendicular to both segments must be within a predefined threshold. If the helices are assumed as infinite in length, this distance is always defined. In Walther *et al.*'s paper,[22] the definition is extended to the more realistic case where helices are of finite length. A simpler criterion for interaction, also used in this paper, defines a pair of helical structures as interacting if the distance between their midpoints is less than a given threshold.

In Bowie's paper,[15] it was observed that the expected uniform random distribution of angles is actually biased toward angles near 90°. When this geometric bias was taken into account, the observed peaks in the interacting 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.

The global interaction of multiple SSEs in space is much less analyzed and understood.[24] In the past, we have conducted a statistical analysis on

the arrangements of triplets of SSEs, including noninteracting α-helices and β-strands.[25,26] We found that the distribution of triplets of angle cosines is far from being random, with a marked preference for certain angle combinations. Specifically, we showed that planar configurations of α-helices and β-strands appear more frequently than expected for uniformly distributed vectors, and alignment is strongly preferred compared to that expected for uniformly distributed vector triplets.

The present study extends the previous analysis to quartets of SSEs. We measure the distribution of the six angles formed by quartets of secondary structure segments that are not necessarily directly interacting, and compare such distribution with that which would be expected for a randomly distributed set of quartets of vectors in order to determine significant deviations. We present results on the distribution of secondary structures within a selected set of 300 nonredundant proteins. The results are consistent with prior studies on triplets in that planar configurations of unit vectors of α-helices and β-strands appear more frequently than expected for uniformly distributed vectors. For overrepresented patterns of angles, the relation $\gamma + \delta = \phi$ characterizing the angles γ, δ, and ϕ formed by three coplanar vectors in 3D space holds generally for more than a triplet of segments which is a subset of the overrepresented quartet. This implies that, more often than expected, the SSEs are located on parallel planes.

One of the findings of our analysis is that overrepresented angular patterns of quartets are often compatible with known motifs of interacting helices or strands, although such motifs are a very small fraction of the quartets that contribute to the frequency of the patterns. In other words, the same angular patterns are shared by interacting SSEs as well as by quartets comprising interacting and noninteracting SSEs. Furthermore, we show that in the top overrepresented pattern among all configurations of four helices or four strands, at least one pair of SSEs is in contact but generally more than one. Finally, we observe that overrepresented patterns of angles tend to form clusters in the six-dimensional (6D) space corresponding to six angles, and determine such clusters.

Another contribution of the paper is the development of an efficient methodology that computes the overrepresented patterns of quartets by assembling frequent patterns of triplets of segments. The combinatorial explosion makes the problem of determining quartets remarkably more computationally demanding than that of triplets. We tested the results of our method by a comparison with the exhaustive enumeration of quartets. While the exhaustive enumeration of all quartets of SSEs for even a limited set of 300 representative proteins takes more than 3 days on a standard PC, our method requires less than 20 minutes. This significant reduction in computation time is achieved by exploiting an idea similar to the Apriori algorithm often used in data mining. Apriori constructs frequent patterns of k elements (or itemsets of size k) by joining frequent patterns of size $k-1$. In our case, frequent patterns of six angles are constructed by joining frequent patterns of three angles. Careful attention has to be paid to the fact that the closure property of the frequency that is assumed in Apriori does not hold for all triplets of angles of a

sextuple. Our construction can apply to sets with more than four SSEs for which the enumeration could be prohibitive.

The paper is organized as follows. We formulate the problem in Sec. 2 and describe the Apriori-based method to generate overrepresented patterns in Sec. 3. In Sec. 4, we ask the question of whether two sets of unit vectors on a sphere that have the same angular pattern are in fact geometrically similar. It is easy to show that there are basically few distinct configurations which cannot be superimposed corresponding to a given pattern of angles. A discussion of the overrepresented patterns is in Sec. 5.

2. Problem Description

Given a dataset of protein structures, we address the problem of finding overrepresented arrangements of SSEs in terms of geometrical properties. Most papers in the literature study the distribution of geometrical properties, in particular angles, of pairs of interacting SSEs.[19,21] The scope of this paper is twofold: first, to identify recurrent configurations of SSEs by means of their geometrical properties; and second, to pose as the basis to efficiently extract recurrent configurations composed by an arbitrary number of SSEs. Here, we focus on overrepresented configurations of multiple SSEs not necessarily interacting and analyze the distribution of angles of such configurations. We present a study of the distribution of angular patterns of SSEs computed without considering any notion of distance or distinguishing between interacting and noninteracting elements.

Our task is to design a framework to extract overrepresented arrangements of k SSEs by combining the results obtained with arrangements of $k-1$ SSEs. We discuss in detail how to efficiently obtain overrepresented arrangements of four SSEs by using the distribution of triplets of SSEs rather than by an exhaustive generation. This specific case will pose as 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 all SSEs ordered according to the backbone chain. A line segment is associated to each SSE. For a β-strand, the line segment is the best-fit segment of the set of atoms of the strand; while 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 3D space. Thus, a protein is a list of m unit vectors $(s_1, s_2, \ldots, s_m)$.

The arrangement of SSEs is described in terms of the angles formed by all pairs of corresponding vectors. Let α_{hk} be the angle of s_h and s_k, $0^\circ \leq \alpha_{hk} \leq 180^\circ$. 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 six 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 Fig. 1. It is easy to show that, in the general case, the six angles are not completely independent. More precisely,

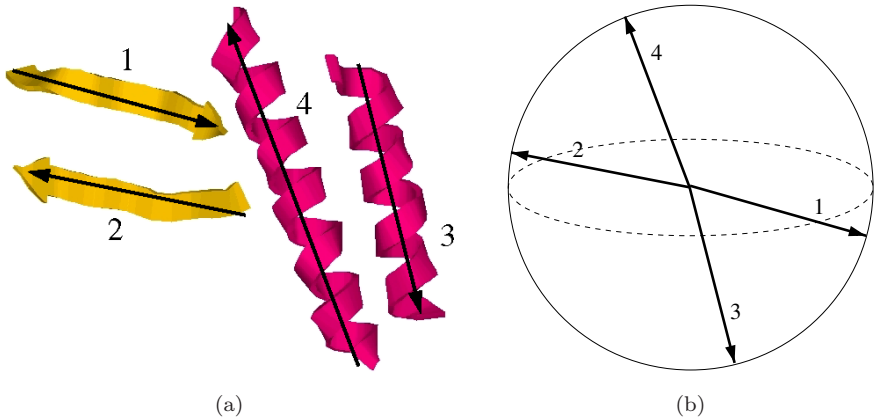


Fig. 1. (a) An example of vector representation for a quartet of SSEs. (b) The unit vectors translated to the origin (into the unit sphere).

given five 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, which will be considered in Sec. 4, is whether it is possible to superimpose, by a rigid transformation, two quartets forming the same angles.

For the purpose of our analysis, the angular values are discretized into uniform intervals, with every interval represented by an integer. More precisely, the range 0°–180° is divided into 10 intervals, and an angle α represented by the integer i such that $i * 18° \leq \alpha < (i + 1) * 18°$. We experimented with several partition criteria and chose the number of intervals equal to 10. This appears a reasonable choice if we consider the approximations introduced in calculating the best-fit segments for strands and helices; furthermore, this ensures a reasonable number of items per interval. A quartet of SSEs is represented by six integer values, each within the range $[0, 10]$. In the following, we refer to the discretized angles simply as "angles".

3. Discovery of Overrepresented Patterns

Our approach to find overrepresented angular patterns is similar to the Apriori algorithm used for data mining applications. Our algorithm finds overrepresented arrangements of quartets of segments from overrepresented triplets of segments. It does so by joining overrepresented triplets of angles to obtain overrepresented 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 or closure property, which states that all nonempty subsets of a frequent set must also be frequent. In other words, if an itemset cannot pass the test of being frequent, then all of 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 which are subsets of Q are not frequent. This is the case for the triplet of angles $T = (\alpha_{13}, \alpha_{23}, \alpha_{24})$, which cannot be formed (in the general case) by a triplet of segments that is a subset of an element of S because the three angles involve all four segments of a single element of S. However, there are four triplets of angle subsets of a frequent sextuple Q that must be frequent: $(\alpha_{12}, \alpha_{13}, \alpha_{23})$, $(\alpha_{23}, \alpha_{24}, \alpha_{34})$, $(\alpha_{13}, \alpha_{14}, \alpha_{34})$, and $(\alpha_{12}, \alpha_{14}, \alpha_{24})$. Indeed, the four triplets are obtained 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 start by giving an overview of our approach, and then describe each step in detail. The first two steps use hashing techniques to compute and store all triplets of angles into hash tables. Geometric hashing techniques were originally devised in the area of computer vision,[27] mostly for object recognition applications, and were later applied to several matching problems arising in computational biology.[9,11,28] The next two steps 3 and 4 test the deviation of the obtained angular distribution from that of random vectors. Finally, steps 5 and 6 are based on a variant of the Apriori method to determine and evaluate the distribution of quartets.

Procedure: *Pattern Discovery*

(1) Initialization: From the given protein dataset, 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 a hash table, indexed by the triplets of angles, that stores all triplets of segments. Derive a 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 overrepresented.
(5) Join step: Construct candidate sextuples of angles from triplets of C.
(6) Verification step: Prune candidate sextuples to find the overrepresented ones.

3.1. *Building the hash table*

We build a four-dimensional (4D) 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, while 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 bin size for the histograms. Each cell in 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 Comin *et al.*[9]

Procedure: *Insert Protein*
Given a 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:

(1) Compute the angles $(\alpha_{uv}, \alpha_{vz}, \alpha_{uz})$ and determine the triplet type.
(2) Access the cell of the hash table at the location indexed by triplet type and by the quantized values of $(\alpha_{uv}, \alpha_{vz}, \alpha_{uz})$.
(3) Append to the list of records at that cell a new record which contains
    - the name of protein P; and
    - the identifier of each SSE of the triplet.

The above procedure is repeated for all proteins in the dataset. 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 bias due to the features considered, we compute the null distribution of such properties.

Note that, for each protein, all triplets of SSEs are inserted into the hash table without any distinction between interacting and noninteracting elements. This table represents the distribution of angles of triplets of SSEs for our dataset of real proteins. Here, we want to construct the distribution of angles of randomly generated triplets of segments. This new distribution cannot be obtained by simply generating three angles at random, since the angles in a triplet are highly constrained and therefore their distribution is far from uniform. Instead, we construct feasible triplets of random segments and compute their angles.

The random generation of a triplet of angles consists of 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 z-coordinate. A versor is now uniquely determined by two parameters: its coordinate $z \in [0, 1]$, and its azimuth $\beta \in [0, 2\pi]$.

Given a dataset of n real proteins, we generate n sets of random vectors, each corresponding to a real protein and containing the same number of SSEs of such protein. Then, for each of the n sets, we compute the angles of all triplets of random vectors and update the hash table accordingly.

We have already observed that the triangular inequality holds for any order of the 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 the 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) is $Count(\alpha, \beta, \gamma)/V_c(\alpha, \beta, \gamma)$, where $Count(\alpha, \beta, \gamma)$ is the number of elements in that cell.

The generation of the n sets 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 overrepresented angles, we experimented with different selection policies. To preserve a reasonable number of candidates, we select configurations of angles that have a frequency larger than the mean in the corresponding cell of random data.

3.3. *Join and verification steps*

The operation *join* merges four frequent triplets — $(\alpha_{12}, \alpha_{13}, \alpha_{23})$, $(\alpha_{23}, \alpha_{24}, \alpha_{34})$, $(\alpha_{13}, \alpha_{14}, \alpha_{34})$, 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 triplet, 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 triplet coincide with the first two segments of the second triplet. 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 $(\alpha_{23}, \alpha_{24}, \alpha_{34})$, 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

In this section, we discuss why two sets of vectors that share the same angular pattern may not be necessarily superimposed by a 3D rigid transformation. We recall that we are considering only angles between SSEs, but not distances between them. We want to stress that the notion of superimposition is a much stronger condition. For example, two sets of segments with the same angular pattern that are one the mirror of the other, from a physical prospective, can serve the same function, although they cannot be rigidly superimposed. To this end, here we enumerate the number of possible arrangements of vectors that generate the same angular pattern. This is of interest since geometric superposition of proteins, and specifically of SSEs, is often used in protein structure alignment methods to determine structural similarity.[29]

We define as "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) with 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 nonparallel vectors, forming a given triplet of angles (α, β, γ). For example (see Fig. 2), consider four vectors forming a regular pyramid with vertex in O; 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 nonequivalent since they are one the mirror of the other.

The following simple procedure constructs two nonequivalent triplets with the same angles. All vectors forming a given angle δ with a given vector v are rays of

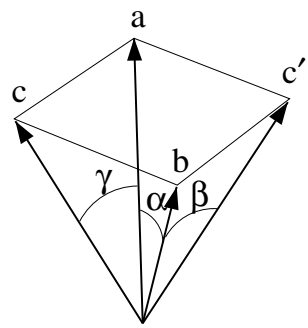


Fig. 2. An example of two triplets, (a, b, c) and (a, b, c'), with the same pairwise angles, one the mirror of the other.

the cone with vertex in O and forming angle δ 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 nonparallel 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 nonequivalent arrangements doubles. Consider a sextuple of angles $(\alpha_{12}, \alpha_{13}, \alpha_{23}, \alpha_{24}, \alpha_{34}, \alpha_{14})$. To construct all nonequivalent 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 the 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''').

In conclusion, given a sextuple of angles, there can be up to four different arrangements of segments that generate this angular pattern. On the other hand, the superimposition is a very constricting condition and two arrangements that may be considered similar in practice might not be under this restrictive requirement. Nevertheless, the number of possible arrangements remains particularly small. In our experiments, we observed that for a given angular pattern, although all four configurations of segments might exist, in practice there is a strong bias towards one or two specific arrangements. Again, this reduces the search space when performing motif search and it can be used to filter out infrequent and unfeasible motifs.

5. Results and Discussion

We selected a set of 300 nonredundant proteins from different families, and computed the set of all triplets of SSEs and their associated linear segments (see supplementary material for a complete list). 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.[30] 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. We recall that, at this stage, all triplets of SSEs are considered without any distinction between interacting and noninteracting elements.

To better appreciate the distributions of real and random triplets, we report both histograms in Figs. 3(a) and 3(b). By construction, the two histograms contain the same number of triplets and are computed using sets of segments with the same composition; consequently, they can be directly compared. The hash table contains 520 nonempty 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 frequent is shown in Fig. 3(c).

5.1. *Analyzing overrepresented patterns of angles*

The pattern discovery process finds a set of overrepresented arrangements of four SSEs. Each arrangement is described by six ordered angles, where an angle corresponds to a specific pair of SSEs that 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.

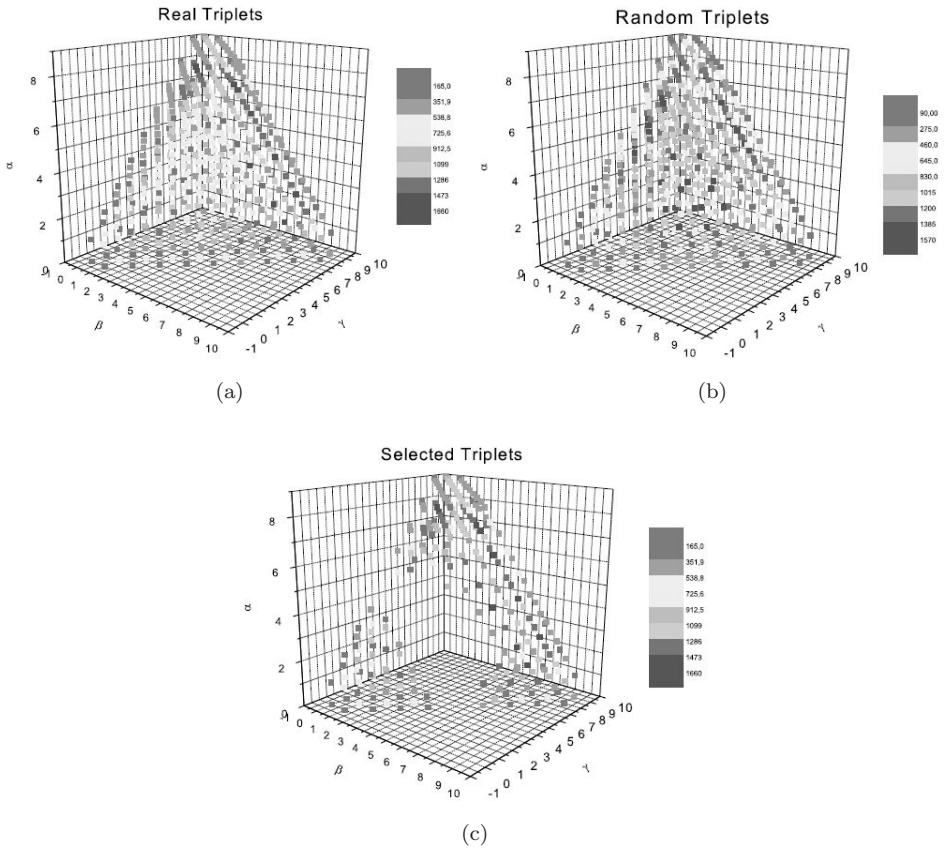


Fig. 3. 3D histograms of the distributions of angles between triplets of SSEs. Each axis represents an angle, and the frequency of each triplet follows the color coding.

By merging patterns, the discovery procedure selects a set of 785 overrepresented 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 overrepresented pattern is composed by the discretized angles $(1, 2, 3, 7, 8, 9)$, corresponding to angles in the ranges $(18°–36°, 36°–54°, 54°–72°, 126°–144°, 144°–162°, 162°–180°)$, and has a frequency of 6,439; the top second pattern 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. The overall distribution of patterns of angles, ranked by their frequency, is illustrated in Fig. 4. 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, using the same

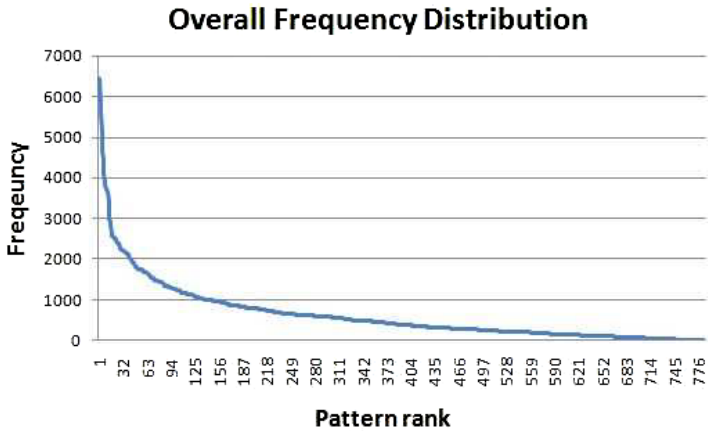


Fig. 4. Overall frequency distribution of patterns of angles.

program without filtering out the infrequent triplets, the exhaustive generation of all possible quartets of SSEs takes more than 3 days.

We observed that overrepresented patterns of angles tend to form clusters in the 6D space corresponding to six angles. Thus, we further analyzed the set of overrepresented patterns by clustering them using as distance the Euclidean distance between angular patterns in 6D space. We experimented with different clustering algorithms and different numbers of clusters and, based on the measure of silhouette,[31] we selected the k-means algorithm with three clusters. Clusters 1 and 3 contain, respectively, the first and second most frequent patterns. 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 Table 1.

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° to 72°. 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 orientation (cluster 1) or with two SSEs in the same orientation and the other two in the opposite orientation (cluster 3). The smaller number of elements in cluster 2 reflects the tendency of SSEs to form antiparallel configurations.

In all clusters, the angles vary from 0° to 72° and from 126° to 180°, while values between 80° and 100° are completely absent. For example, in parallel and antiparallel β-sheets, each β-strand typically forms a small angle with the two nearby strands. The same is true for interacting α-helices, which pack forming small angles; furthermore, they are hardly found perpendicular to each other.[22,23]

We observed that the top pattern $(1, 2, 7, 8, 8, 9)$ is compatible with a known motif of interacting strands; however, the quartets of interacting strands contribute less than 450 to the total frequency of 5,780. Thus, although the number

Table 1. The top ten most frequent patterns for the three clusters.

α_0	α_1	α_2	α_3	α_4	α_5	Frequency
			(a) Cluster 1			
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
			(b) Cluster 2			
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
			(c) Cluster 3			
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

of interacting quartets of SSEs is very small compared to all possible quartets, still their angular patterns appear among the top overrepresented arrangements.

We now compare our results with the set of all exhaustive quartets. To this end, we provide here a list of the top ten patterns ranked according to their frequency as obtained by an exhaustive enumeration [see Table 2(a)]. For the same top patterns, we also list the frequency of random quartets obtained with a procedure similar to the one described in Sec. 3.2 for triplets. As can be seen from the table, the top angular patterns do not deviate too much from the random distribution, and thus do not appear more frequently than expected by chance. On the other hand, if we rank the exhaustive quartets by the difference of their frequency from that of random quartets, we obtain a different outcome [see Table 2(b)]. The top patterns in this new ranking are, in almost all cases, the same patterns that are detected as

Table 2. The top angular patterns of the exhaustive enumeration of quartets ranked by (a) their frequency and (b) the difference of their frequency from that of random quartets.

α_0	α_1	α_2	α_3	α_4	α_5	Freq.	Rand. Freq.	Diff.
					(a)			
3	4	5	6	7	8	16119	17224	−1105
2	3	5	6	7	8	13447	12467	+979
2	4	5	6	7	8	12912	13274	−362
3	3	5	6	7	8	11054	10525	+528
2	4	5	6	7	7	10236	12575	−2339
3	3	4	6	7	8	9795	9613	+181
1	4	5	6	7	8	9503	8582	+921
2	3	4	6	7	8	9496	9089	+406
3	4	4	6	7	8	9355	10361	−1006
2	3	4	5	6	8	8627	10043	−1416
					(b)			
1	2	3	7	8	9	7439	3267	+4171
1	2	7	8	8	9	5861	2087	+3773
1	2	3	7	8	8	6517	3258	+3258
1	3	6	7	8	9	5939	2807	+3131
1	2	3	6	8	9	5861	2813	+3047
1	1	2	7	8	9	4658	1831	+2826
1	3	5	6	8	9	6146	3373	+2772
2	3	6	7	8	9	5764	3022	+2741
1	2	6	7	8	9	5076	2378	+2697
1	3	4	6	8	9	5915	3364	+2550

overrepresented by our method. In Fig. 5, the distribution of the 785 overrepresented patterns is shown and compared with that of all 2,262 quartets ranked by their deviation from random quartets. We find that among the top 100 quartets, 81 are discovered as overrepresented by our algorithm. In the figure, the curve of elements in common grows very fast until it saturates, thus confirming that our results are in accordance with those of a more computing-intensive procedure.

5.2. *Searching for motifs*

We have seen that overrepresented patterns tend to be arranged into specific spatial conformations that can be described in terms of groups of parallel and antiparallel SSEs. Moreover, we have observed that the overall frequency of angular patterns drops dramatically after the first top patterns, and that this tendency is preserved even within the different clusters. We now deepen this investigation by analyzing only homogeneous configurations, i.e. those containing four strands or four helices. A similar study can be carried out for any combination of strands and helices, but due to lack of space we report only two cases here. It is interesting to notice that even in this restricted scenario, we obtain similar results for the clusters, but with a preference for antiparallel pairs, corresponding to the top-ranked pattern of angles $(1, 2, 7, 8, 8, 9)$.

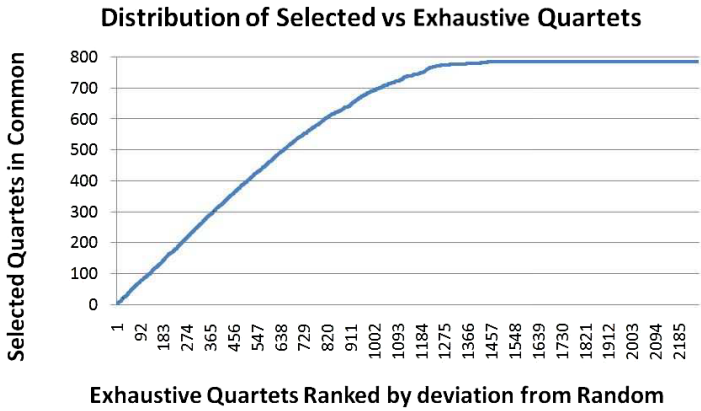


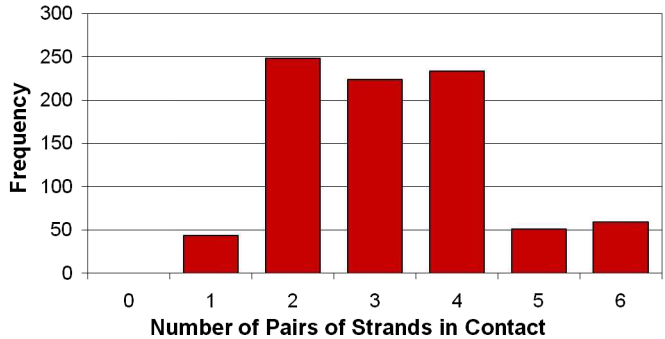
Fig. 5. Distribution of overrepresented quartets over all quartets ranked by deviation from random quartets.

In general, motifs are arrangements of SSEs that contribute to a given fold. In most cases, the SSEs involved in a motif are also close in space, implying that their pairwise distance must be limited. The overrepresented patterns considered so far have included the SSEs of the selected set of proteins, regardless of their distances; moreover, the overrepresented patterns are extracted by using only angular information. We now investigate whether such patterns are also close in space.

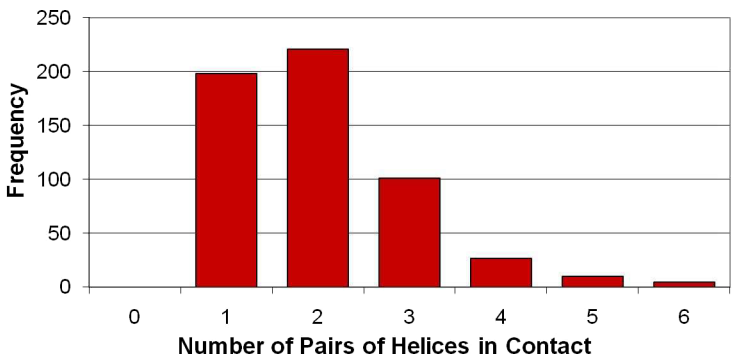
We consider only homogeneous patterns of SSEs from the top-ranked configuration $(1, 2, 7, 8, 8, 9)$, and we define two SSEs to be in contact if the distance between the midpoints of their associated vectors is less than a given threshold (18 Å in our analysis). Notice that, following our definition, in a motif not all pairs of SSEs must be in contact; it is enough that every SSE is in contact with at least another SSE. In our case of four SSEs, a motif should have at least three pairs of SSEs in contact. Figure 6 shows the number of SSEs in contact for the quartets of segments with the top angular pattern.

It is interesting to observe 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.[21] Nevertheless, more than 65% of the elements have at least two SSEs in contact. Better results are obtained when considering only strands; in this case, 65% of the elements have at least three SSEs in contact. Again, we can observe that purely angular patterns which are overrepresented are in most cases close in space.

To better appreciate the proximity of these overrepresented configurations, in Fig. 7 we present different examples of four strands with angles $(1, 2, 7, 8, 8, 9)$. In all of 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 include SSEs from the same β-sheet [Fig. 7(c)] as well as from different β-sheets [Figs. 7(a) and 7(b)]. The fact that



(a)

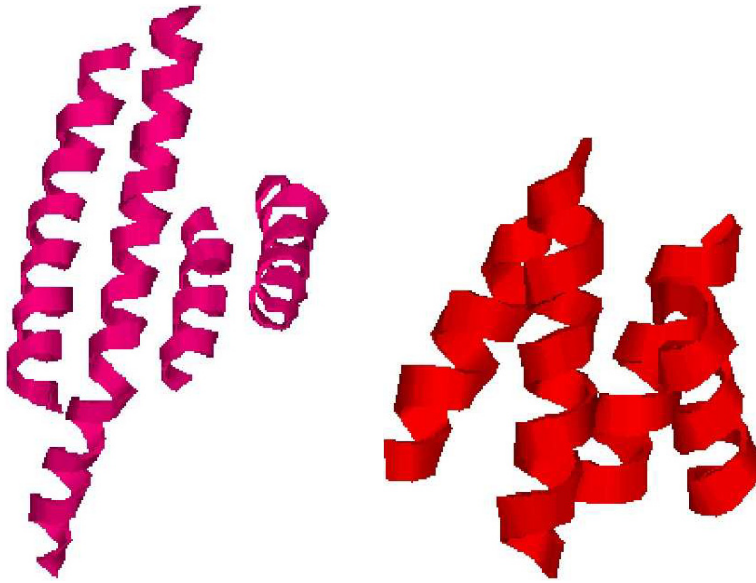


(b)

Fig. 6. Frequency of pairs of SSEs in contact: (a) quartets of strands, and (b) quartets of helices.

(a) (b) (c)

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



(a) (b)

Fig. 8. Examples of patterns of angles composed by helices only: (a) pattern $(0, 1, 1, 2, 2, 3)$, protein 3mdd, SSE: 0-3-13-15; (b) pattern $(1, 2, 7, 8, 8, 9)$, protein 1aa7, SSE: 0-3-7-8.

most, but not all, 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.

A similar observation holds for quartets formed by all helices. In Fig. 8, the two most frequent angular patterns generated by four helices are shown. In Fig. 8(a), the four helices are such that all pairwise angles are smaller than 72° (parallel configuration); whereas in Fig. 8(b), an antiparallel configuration is represented.

As we observed in Sec. 4 and illustrated in Fig. 7, SSEs belonging to the same quartet do not necessarily correspond to structures that can be superimposed by rotation and translation. Nevertheless, they may perform the same functional role. For this reason, although a limited number of configurations generate the same angular pattern, we keep all of these arrangements grouped together in the same pattern.

6. Conclusions

We have proposed an efficient algorithm to extract overrepresented quartets of SSEs that avoids the exhaustive generation of patterns. We have shown that careful analysis of the angular bias of random vectors is essential in the determination of overrepresented 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 overrepresented patterns could be used to guide the engineering of stable protein modules or to predict their 3D structures. Other applications can be designed by replacing the null distribution with that of a specific family of proteins. The fact that most of the elements in this class are also close in space consolidates the idea that particular arrangements of angles are more likely than others, and that these configurations are precisely the ones which biologists should be interested in.

Acknowledgments

We would like to thank the reviewers for their useful comments. This research was partially sponsored by the Ateneo Project of University of Padova "Computational assessment of protein–protein interaction networks: target prediction and validation as guide for modern system biology".

References

1. Efimov AV, Structural trees for protein superfamilies, *Proteins* **28**(2):241–2260, 1997.
2. Persson B, Bioinformatics in protein analysis, *EXS* **88**:215–231, 2000.
3. Barker JA, Thornton JM, An algorithm for constraint-based structural template matching: Application to 3D templates with statistical analysis, *Bioinformatics* **19**(13):1644–1649, 2003.
4. Chen BY, Fofanov VY, Kristensen DM, Kimmel M, Lichtarge O, Kavraki LE, Algorithms for structural comparison and statistical analysis of 3D protein motifs, *Pac Symp Biocomput*, pp. 334–345, 2005.
5. Stark A, Sunyaev S, Russell RB, A model for statistical significance of local similarities in structure, *J Mol Biol* **326**:1307–1316, 2003.
6. Samudrala R, Levitt M, A comprehensive analysis of 40 blind protein structure predictions, *BMC Struct Biol* **2**:3–10, 2002.
7. Zhang Y, Skolnick J, The protein structure prediction problem could be solved using the current PDB library, *Proc Natl Acad Sci USA* **102**(4):1029–1034, 2005.
8. Laskowski RA, MacArthur MW, Moss DS, Thornton JM, PROCHECK: A program to check the stereochemical quality of protein structures, *J Appl Crystallogr* **26**:283–291, 1993.
9. 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.
10. Dror O, Benyamini H, Nussinov R, Wolfson H, MASS: Multiple structural alignment by secondary structures, *Bioinformatics* **19**(Suppl 1):95–104, 2003.
11. Dror O, Benyamini H, Nussinov R, Wolfson H, Multiple structural alignment by secondary structures: Algorithm and applications, *Protein Sci* **12**:2492–2507, 2003.
12. Brenner SA, Predicting the conformation of proteins from sequences. Progress and future progress, *J Mol Recognit* **8**(1–2):9–28, 1995.
13. Eisenhaber F, Persson B, 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.
14. Chothia C, Levitt M, Richardson D, Structure of proteins: Packing of α-helices and pleated sheets, *Proc Natl Acad Sci USA* **74**:4130–4134, 1977.

15. Bowie JU, Helix packing angle preferences, *Nat Struct Biol*, **4**:915–917, 1997.
16. Chothia C, Levitt M, Richardson D, Helix to helix packing in proteins, *J Mol Biol* **145**:215–250, 1981.
17. 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* **132**:275–288, 1979.
18. Efimov AV, Complementary packing of alpha-helices in proteins, *FEBS Lett* **463** (1–2):3–6, 1999.
19. Lee S, Chirikjian GS, Interhelical angle and distance preferences in globular proteins, *Biophys J* **86**(2):1105–1117, 2004.
20. Nakamura H, Roles of electrostatic interaction in proteins, *Q Rev Biophys* **29**(1):1–90, 1996.
21. Reddy B, Blundell T, Packing of secondary structural elements in proteins. Analysis and prediction of inter-helix distances, *J Mol Biol* **233**:464–479, 2003.
22. Walther D, Eisenhaber F, Argos P, Principles of helix-helix packing in proteins: The helical lattice superposition model, *J Mol Biol* **255**:536–553, 1996.
23. Walther D, Springer C, Cohen FE, Helix-helix packing angle preferences for finite helix axes, *Proteins* **33**(4):457–459, 1998.
24. Hespenheide BM, Kuhn LA, Discovery of a significant, nontopological preference for antiparallel alignment of helices with parallel regions in sheets, *Protein Sci* **12**(5):1119–1125, 2003.
25. Guerra C, Lonardi S, Zanotti G, 3D matching of proteins based on secondary structures, *Proc IEEE Symposium on 3DPVT*, Padova, Italy, pp. 812–821, 2002.
26. Platt DE, Guerra C, Zanotti G, Rigoutsos I, Global secondary structure packing angle bias in proteins, *Proteins* **53**(2):252–261, 2003.
27. Wolfson HJ, Rigoutsos I, Geometric hashing: An overview, *IEEE Comput Sci Eng* **4**(4):10–21, 1997.
28. Shulman A, Nussinov R, Wolfson HJ, Recognition of functional sites in protein structures, *J Mol Biol* **339**:607–633, 2004.
29. Yona G, Kedem K, The URMS-RMS hybrid algorithm for fast and sensitive local protein structure alignment, *J Comput Biol* **12**:12–32, 2005.
30. Gerstein M, A resolution-sensitive procedure for comparing protein surfaces and its application to the comparison of antigen-combining sites, *Acta Cryst* **48**:271–276, 1992.
31. Kaufman L, Rousseeuw PJ, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1990.

Matteo Comin received his Laurea degree in Computer Science in 2003 from the University of Padova, Italy. In 2003, he was a visiting student at Purdue University, USA; and from 2004 to 2006, a research intern at the IBM T.J. Watson Research Center. In 2007, he received a Ph.D. in Computer Science from the University of Padova. Currently, he is an Assistant Professor at the same university. His interests are in the design of algorithms and applications for pattern discovery, data compression, geometric pattern matching, structural proteomics, and grid computing.

Concettina Guerra works in the areas of computational biology and computer vision. Her recent interests fall in the domains of protein classification, recognition, and docking. Formerly an Associate Professor at the University of Rome, Italy, she joined the Department of Information Engineering of the University of Padova, where she became a Professor in the Faculty of Engineering. She has visited extensively with US institutions, including Rensselaer Polytechnic and Carnegie Mellon University, and has been on the Computer Science faculty of Purdue University for over a decade. Now, she is a Professor at the College of Computing at the Georgia Institute of Technology.

Giuseppe Zanotti is a Full Professor of General and Inorganic Chemistry at the University of Padova, Italy. His fields of interest are biophysics and structural biology. He has deposited more than 75 crystal structures of proteins at the Protein Data Bank (http://www.rcsb.org). He is also interested in the theoretical aspects of the phase problem in crystallography, and in the analysis of conformational aspects of the structure of globular proteins.