MINING SUPER-SECONDARY STRUCTURE MOTIFS FROM 3D PROTEIN STRUCTURES: A SEQUENCE ORDER INDEPENDENT APPROACH

ZEYAR AUNG[1] JINYAN LI[2]
azeyar@i2r.a-star.edu.sg jyli@ntu.edu.sg

[1]*Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613*
[2]*School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798*

Super-Secondary structure elements (super-SSEs) are the structurally conserved ensembles of secondary structure elements (SSEs) within a protein. They are of great biological interest. In this work, we present a method to formally represent and mine the sequence order independent super-SSE motifs that occur repeatedly in large data sets of protein structures. We represent a protein structure as a graph, and mine the common cliques from a set of protein graphs in order to find the motifs. We mine two categories of super-SSE motifs: the generic motifs that occur frequently across the entire database of protein structures, and the fold-preferential motifs that are concentrated in particular protein fold types. From the experimental data set of 600 proteins belonging to 15 large SCOP Folds, we have discovered 21 generic motifs and 75 fold-preferential motifs that are both statistically significant and biologically relevant. A number of the discovered motifs (both generic and fold-preferential) resemble the well-known super-SSE motifs in the literature such as beta hairpins, Greek keys, zinc fingers, etc. Some of the discovered motifs are of novel shapes that have not been documented yet. Our method is time-efficient where it can discover all the motifs across the 600 proteins in less than 14 minutes on a stand-alone PC. The discovered motifs are reported in our project webpage:
http://www1.i2r.a-star.edu.sg/~azeyar/SuperSSE/

Keywords: 3D Protein Structure, Super-secondary Structure, Structural Motifs Mining.

1. Introduction

Proteins are the workhorses in the cells of living organisms. A protein is made up of a sequence of amino acid (AA) residues which folds into a particular 3-dimensional (3D) structure by the various forces of nature. A 3D protein structure consists of frequent and structurally conserved elements called secondary structure elements (SSEs). Alpha helices and beta strands are the two common types of SSEs. There are in turn some ensembles of SSEs that are frequent and structurally conserved. They usually serve as the structural and/or functional units within a protein, and are called super-secondary structure elements (super-SSEs) [4].

Biologists are very interested in super-SSEs because they are usually associated with basic structural configurations and/or basic biological functions of the proteins.

Some of the well-known super-SSE types are helix-loop-helix, beta ribbon, beta hairpin, beta-alpha-beta, zinc finger, EF hand, Greek key, etc. Researchers have studied super-SSEs extensively for more than three decades [12, 21, 23–26].

A super-SSE motif is a particular type of structurally similar super-SSEs that occur frequently across a given set of protein structures. In this paper, we propose a method to (1) formally represent the sequence order independent (sequentially disconnected) super-SSEs with respect to their structural conformations, and (2) mine the motifs of those super-SSEs in a given set (either the entire database or a particular fold type) of protein structures.

Conventionally, a super-SSE is defined as a set of sequentially connected (i.e. sequence order preserved) SSEs that are neighbored to each other in 3D space. However, there exists a number of biologically significant structural motifs being composed of SSEs that are spatially proximate yet sequentially not connected [5, 9, 25]. Such a motif can be termed a *sequence order independent* motif.

In this work, we generalize the definition of a super-SSE by relaxing the sequence order constraint with a view to covering the sequence order independent motifs. For example, while the conventional definition covers only the sequence order preserved motif A–B–C as shown in Fig. 1(a), our definition can also deal with the sequence order independent motif A'–B'–C' as shown in Fig. 1(b).

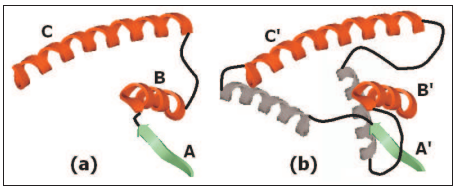

Fig. 1. (a) A conventional (sequence order preserved) beta-alpha-alpha super-SSE. (b) A sequence order independent super-SSE with the same spatial configuration.

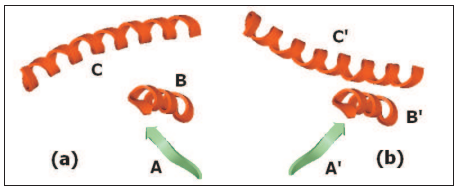

Fig. 2. Two beta-alpha-alpha super-SSEs with different structural configurations.

In our proposed method, we represent a protein structure as a labeled graph with each node being an SSE, and each edge being the relationship between two close enough SSEs. A clique (a fully connected sub-graph) within a graph corresponds to a super-SSE. We develop an algorithm to mine the frequent clique types (super-SSE motifs) in a given protein structure data set. From a experimental data set of 600 proteins, we can discover a number of generic and fold-preferential motifs that are both statistically significant and biologically relevant within a short time.

2. Motivations

2.1. *Need for a Formal Representation Scheme*

Traditionally, super-SSEs (both sequence order preserved and sequence order independent) are described less formally with the names such as helix-loop-helix, alpha-

beta-alpha, etc. Using this verbal description, we have only a very limited ability to identify and quantify the super-SSEs systematically. For example, we may be able to distinguish a beta-alpha-alpha from an alpha-beta-beta. But we may not be able to differentiate between two beta-alpha-alpha super-SSEs having different structural configurations as shown in Fig. 2. The ability to distinguish or classify such kinds of SSEs into different types is highly desirable for the biologists, since it will enable them to study super-SSEs in a more subtle manner [12].

Some methods such as [12, 24] try to identify the different types of super-SSEs by characterizing the loops between the constituent SSEs of a super-SSE. But, this approach is also limited because it requires the sequence order constraint, and its applicability is confined to the super-SSEs with only two elements.

Thus, there is a need for a formal representation scheme which enables the identification and quantitative manipulation (comparison, clustering, etc.) of super-SSEs in a more general manner (i.e. applicable to all kinds of super-SSEs regardless of their sequence order and the number of SSEs they contain). In this work, we try to address this formalization issue by representing proteins and super-SSEs as labeled graphs and labeled cliques respectively.

2.2. *Need for a Large-Scale Motif Mining Method*

Structural motif mining is an active area of research in structural bioinformatics. Different methods use different description of structural motifs, and try to mine the frequent motifs from a set of protein structures. Trilogy [3] explores the sequence–structure motifs made up of AA residue triplets; SPratt2 [17] mines the conserved residues within a fixed-size bounding sphere; MotifMiner [6] mines the frequent atom-sets; and Huan *et al.* [15, 16] mines the frequent sub-graphs/cliques of AA residues, etc.

In this study, we will focus on the discovery of structural motifs in terms of the super-SSEs. A number of methods, such as Koch *et al.* [19], MASS [9], PROTEP [1], and Szustakowski *et al.* [25], have been proposed to detect both sequence order preserved and sequence order independent super-SSE motifs. All of these methods adopt the *comparison-based* motif discovery approach [11] in which each method employs one of the many multiple structural alignment algorithms to generate the motifs. Unfortunately, such a comparison-based approach is only suitable for the discovery of motifs from small data sets with just tens of protein structures. In terms of its scalability, it is not suited for motif discovery from larger data sets with hundreds or thousands of proteins for the following reasons.

- Usually, a motif does not occur in all proteins in the data set, but only in a subset of it. Since we do not know *a priori* the motifs nor the subsets of proteins in which these motifs occur, we need to explore all the possible combinations of proteins in the data set. In order to retrieve the complete set of motifs from a given set of N proteins, a naive approach will take an exponential time, whist an intelligent approach, such as the one described

in Koch *et al.* [19], will still take an $O(N^3)$ time.
- If a greedy strategy is adopted to reduce the time cost, some pivot proteins can be selected to serve as seeds for multiple alignments. But, this cannot always guarantee a complete answer in the event where a motif does not occur in any of the selected pivots.
- Although a single run of an expensive comparison-based motif discovery algorithm (which may take several days to several weeks) may still be affordable, one will need multiple runs of the algorithm with different sets of parameters in order to secure the desired results. Such multiple runs are prohibitively expensive to be carried out in reality.

With a view to overcoming the abovementioned problems, we adopt the *pattern-mining* approach — also known as the *pattern-driven* approach [11] — for the large-scale discovery of super-SSE motifs. The pattern-mining strategy has been used for discovering sequence, structure, and sequence–structure motifs of various kinds [3, 6, 15–17]. However, to our best knowledge, it has not been used for the discovery of super-SSE motifs before.

Since we represent a protein structure as a graph, we need to apply pattern mining algorithms for graphs so as to discover our desired super-SSE motifs from the graphs. This has been technically infeasible until the recent emergence of the algorithms for the large-scale mining of graph databases for sub-graphs [28], quasi-cliques [22], and cliques [27].

In this work, we utilize one of these latest technologies, namely CLAN [27], to mine the frequent cliques representing the super-SSE motifs. CLAN is known to be a complete clique mining algorithm where it enumerates all the frequent closed cliques from a given database of graphs. It is also an efficient tool that can manage large graph databases with fast response times.

Recently, Huan *et al.* [15, 16] has used graph representation and mining to find the motifs of AA residue nodes. However, it should be noted that their objective is substantially different from ours in which they try to mine the small residue-based packing motifs rather than the relatively large super-SSE motifs as in our case.

3. Methods

3.1. *Formal Representation Of Super-SSE Motifs*

In this section we will describe how we formalize the representation of a protein and that of a super-SSE motif.

3.1.1. *SSE as a Vector*

We use the STRIDE algorithm [13] to identify the SSEs in protein structures. Since SSEs are relatively straight in structure, we can approximate each SSE with a vector (line segment) in 3D space [9, 21]. Fig. 3(b) shows the vector representation of SSEs.

3.1.2. *Protein Structure as a Graph*

We present a protein structure as a graph with its nodes being the SSE vectors, and edges being the relationships between these SSE vectors. Graph representation of protein structures has also been used previously in a number of protein structure comparison and analysis methods [1, 19, 21].

For a protein with n number of SSEs, we have a graph of n nodes. A pair of nodes in the graph is connected by an edge if the *distance of closest approach* [31] between the corresponding SSE vector pair is less than the distance threshold dt. The constituent SSEs in a super-SSE must be close enough to each other, i.e. less than dt, in order to act effectively as a structural/functional unit. Since we do not put an edge between any pair of nodes whose SSE vectors are farther than dt, those two SSEs can never become parts of a single super-SSE. We use $dt = 16$Å as the default value. The graph representation of a protein structure is depicted in Fig. 4.

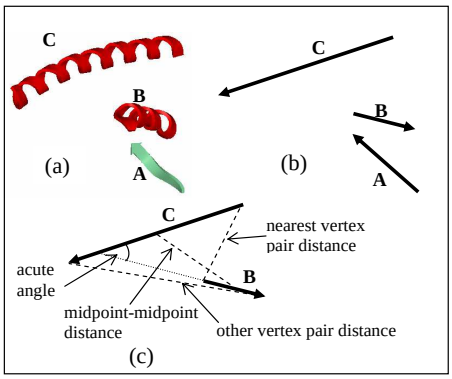

Fig. 3. (a) Original SSEs (b) Vector representation of SSEs and (c) Various types of relationships between SSEs.

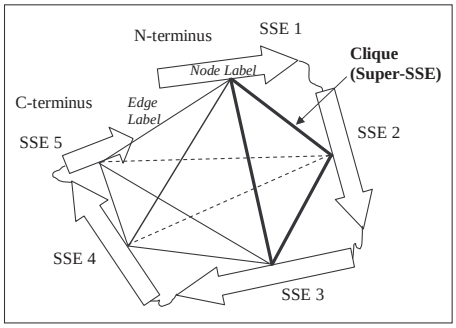

Fig. 4. Graph representation of a protein structure with 5 SSEs. A dotted line denotes a non-existing edge between two node because their SSEs are farther than the distance threshold dt.

Labels are assigned to all nodes and edges. Each node label corresponds to the attributes of the SSE it represents. We use two attributes: (1) type (alpha-helix or beta-strand) and (2) length (in terms of the number of AAs) of the SSE for a node label. Each edge label corresponds to the attributes of the relationship of the two SSEs it connects. We use four attributes: (1) acute angle, (2) nearest vertex-pair distance, (3) other vertex-pair distance, and (4) midpoint-midpoint distance between the two SSE vectors for an edge label. Fig. 3(c) demonstrates the four edge label attributes. Our graph representation scheme is sequence order independent in that the node and the edge labels do not carry any information regarding sequence positions or sequential connectivity of the SSEs.

For each label (either for node or edge), we quantize each attribute, concatenate the binary values for all attributes, and convert the concatenated bit string into a single integer value. The number of bins for each attribute is empirically determined.

3.1.3. *Super-SSE Motif as a Clique*

In a graph for a protein structure as described above, each *clique* (a sub-graph where every node is connected by an edge with every other node) can be viewed as a super-SSE. According to our definition, every constituent SSE within a super-SSE must be close enough (i.e. connected by an edge) to every other constituent SSE. Thus, any other kind of non-clique induced sub-graph does not qualify as a super-SSE.

Two given cliques (representing two super-SSEs) can be considered as structurally similar and thus belonging to the same type if they are isomorphic, i.e. all of their corresponding node and edge labels are matched. (Partially-matched cliques are not guaranteed to similar to each other despite their matching portions.) If the instances of a particular super-SSE type occurs *frequently* in a given set of protein, this super-SSE type can be defined as a *motif*.

3.2. *Mining Super-SSE Motifs*

A frequent clique is a clique that occurs in at least *st* graphs in a given set of protein structure graphs, where *st* is a user-defined support threshold. A frequent clique corresponds to a super-SSE motif.

We find the frequent cliques from the given set of graphs using a general-purpose frequent clique mining algorithm called CLAN [27]. CLAN reports the frequent cliques only in terms of their node labels. (Hereafter, we will name such a clique as a *node-frequent* clique.) In other words, the set of node-frequent cliques reported by CLAN is a superset of the set of actual frequent cliques with both their node and edge labels taken into account.

Thus, we have to test whether a clique reported by CLAN is actually frequent or not. Since CLAN reports only the node labels of the node-frequent cliques and their respective support values, we have to find the actual instances of these node-frequent cliques in all the protein graphs in the data set. We use the VF2 [7] sub-graph isomorphism algorithm to find these instances.

After finding all the instances for a node-frequent clique in all protein graphs, we find the frequent instance(s) (both in terms of their node and edge labels) that occur in at least *st* protein graphs in the given data set, and report them as the desired super-SSE motifs. (Note that, for one node-frequent clique, there may be more than one distinct frequent clique because of the different edge labels. On the other hand, for some node-frequent cliques, there may be no actual frequent clique at all. It was observed that the number of the actual frequent cliques is only about 10% of the original node-frequent cliques.)

We try to find two categories of super-SSE motifs: (1) the motifs that occur frequently across the entire database — termed the *generic motifs*, and (2) the motifs that occur concentratively in particular protein fold types (SCOP Folds in our case) — termed the *fold-preferential motifs*.

3.2.1. *Generic Motifs*

First, we find the generic motifs each of which occurs in at least st_g proteins across the whole given database of protein structures, where st_g is a user-defined support threshold. After we have discovered the generic motifs by the procedure described above, we need to assess their statistical significance. For that, we calculate the estimated p-values of them using the model described by He and Singh [14].

According to this model, we can represent a generic motif ω as a feature vector of the occurrences of the basic elements it contains.

$$\omega = \{y_1, y_2, \ldots, y_t\} \tag{1}$$

where t is the number of unique basic elements in the database, and $y_i\,(1 \leq i \leq t)$ is the number of occurrences of the i-th basic element in the motif ω.

Here, we treat each distinct *combined label*, which is a concatenated string of the label of an edge plus the labels of nodes connected by the edge, as our basic element. We can calculate the probability of ω occurring at random in a protein graph in the database as:

$$\hat{P}(\omega) = \prod_{i=1}^{t} P(Y_i \geq y_i) \tag{2}$$

where $P(Y_i \geq y_i)$ is the probability that the i-th basic element (combined label) occurs at least y_i times in a random vector. This is calculated based on the background distribution of the basic elements in the database. Finally, the p-value of the generic motif ω (termed *generic p-value*) is calculated as:

$$PV_g(\omega) = \sum_{\mu=T}^{N} bino(\mu, N, \hat{P}(\omega)) \tag{3}$$

where N is the number of graphs (proteins) in the database; T is the support, i.e, the number of proteins in which the motif ω occurs ($T \geq st_g$); and $bino(\mathbf{.}, \mathbf{.}, \mathbf{.})$ is the binomial distribution function. If the generic p-value is less than or equal to 0.05, the motif is considered statistically significant.

3.2.2. *Fold-preferential Motifs*

Second, we mine the fold-preferential motifs that occur more frequently in a certain protein fold type rather than in the other protein fold types. In particular, we find the motifs that are concentrated in certain SCOP Folds. (SCOP [30] is a protein structure classification system. A Fold in SCOP consists of a set of proteins that are generally similar to each other in terms of their 3D structures.) We define a particular motif as fold-preferential only if the motif occurs in at least twice the number of proteins in its most frequent SCOP Fold than in its second-most frequent SCOP Fold.

We find the fold-preferential motifs each of which occurs in at least st_f proteins in its most frequent SCOP Fold, where st_f is a user-defined support threshold. Then,

we calculate the statistical significance of the fold-preferential motifs in terms of another type of p-value named *fold-preferential p-value*.

We can calculate the fold-preferential p-value of a particular motif ω to occur by chance in a particular SCOP Fold by using a hypergeometric distribution [15]:

$$PV_f(\omega) = 1 - \sum_{i=0}^{K-1} \frac{\binom{F}{i}\binom{N-F}{T-i}}{\binom{N}{T}} \quad (4)$$

where N is the number of proteins in the entire database; T is the total number of proteins in which the motif occurs in the entire database; (For each motif for a SCOP Fold, we also have to enumerate its other instances outside its own Fold in the rest of the database by using the VF2 algorithm again.) F is the size of the SCOP Fold in which the motif most frequently occurs; and K is the number of proteins in which the motif occurs in this Fold ($K = T \cap F$). Again, if the fold-preferential p-value is less than or equal to 0.05, the motif is regarded as statistically significant.

4. Results and Discussions

We use the same database of 600 proteins as previously used in [2]. The list of the 600 proteins is given in the project webpage. This is a subset of the SCOP database [30] with less than 40% sequence homology. The PDB-style co-ordinates for these proteins are obtained from the ASTRAL database [29].

The database of 600 proteins is composed of 15 large SCOP Folds each having 40 member proteins. (If a Fold contains more than 40 members, we randomly select 40 from it.) The SCOP designations for these 15 Folds and their descriptions are given in Table 1.

First, we mine the generic super-SSE motifs that occur frequently across the whole database of 600 proteins with the support threshold of $st_g = 3\%$, and assign the generic p-values to the motifs. Then, we find the fold-preferential super-SSE motifs for each of the 15 SCOP Folds with the support threshold $st_f = 10\%$, and assign both the fold-preferential p-values and the generic p-values to the motifs.

We conducted our experiments on a single PC with Pentium D 3.2GHz processor and 2GB main memory running Windows XP. The time statistics show that the proposed method is efficient. The total running time using the default parameters (dt = 16Å, $st_g = 3\%$, and $st_f = 10\%$) is only 805 sec (13 min 25 sec) in which 178 sec is for constructing the protein structure graphs, 274 sec is for mining of the generic motifs, and 353 sec for mining the fold-preferential motifs.

The effects of varying the three important parameters dt, st_g and st_f are discussed in the project webpage.

4.1. *Generic Motifs*

We have discovered a total of 22 generic motifs among which 21 are statistically significant in terms of their generic p-values. All of these 21 generic motifs are 3-SSE

Table 1. Number of significant fold-preferential motifs discovered in the SCOP Folds ($dt = 16$Å, $st_f = 10\%$).

SCOP Fold	Description	#3-SSE motifs	#4-SSE motifs	Total
a.4	DNA/RNA-binding 3-helical bundle	0	0	0
a.39	EF Hand-like	0	0	0
a.118	alpha-alpha superhelix	0	0	0
b.1	Immunoglobulin-like beta-sandwich	1	0	1
b.40	OB-fold	0	0	0
c.1	TIM beta/alpha-barrel	8	0	8
c.2	NAD(P)-binding Rossmann-fold domains	20	3	23
c.3	FAD/NAD(P)-binding domain	13	4	17
c.23	Flavodoxin-like	0	0	0
c.37	P-loop containing nucleoside triphosphate hydrolases	1	0	1
c.47	Thioredoxin fold	3	0	3
c.55	Ribonuclease H-like motif	0	0	0
c.69	alpha/beta-Hydrolases	20	1	21
d.15	beta-Grasp (ubiquitin-like)	1	0	1
d.58	Ferredoxin-like	0	0	0
Total		67	8	75

motifs. (There are a vast number of 2-SSE motifs. In this work, we simply ignore them because they are considered less significant. On the other hand, we have not detected any frequent motif with the size larger than 3 SSEs.) We rank the motifs by their generic p-values. The distribution of the motifs' generic p-values is shown in Fig. 5.

The highest-ranked generic motif has the lowest p-value of 5.75×10^{-22}. Its random probability $\hat{P}(\omega)$ is 0.0272, and it occurs in 67 proteins across 7 distinct SCOP Folds. It resembles a version of a well-known conventional super-SSE motif called three-stranded beta hairpin [8, 10] with all beta strands approximately parallel to each other as shown in Fig. 7. (Higher-resolution images for Fig. 7–10 can be viewed in the project webpage.)

We have also discovered a number of other biologically relevant motifs that look

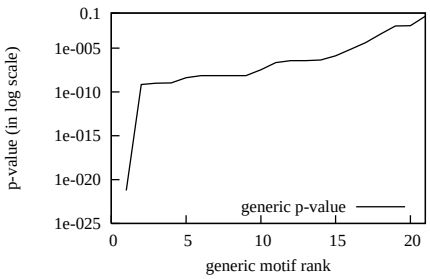

Fig. 5. P-values of generic motifs ($dt = 16$Å, $st_g = 3\%$).

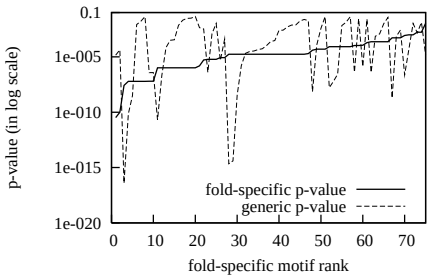

Fig. 6. P-values of fold-preferential motifs ($dt = 16$Å, $st_f = 10\%$).

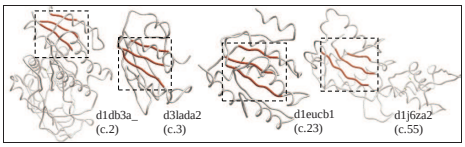

Fig. 7. Some instances of the rank #1 generic motif: a 3-SSE motif resembling a three-stranded beta hairpin with all parallel beta strands.

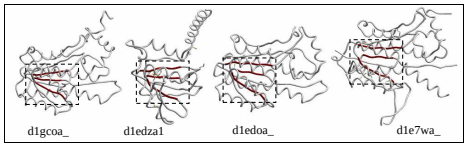

Fig. 8. Some instances of the rank #1 fold-preferential motif in SCOP Fold `c.2`: a 3-SSE motif resembling a three-stranded beta hairpin with two parallel and one angled beta strands.

like the well-known conventional super-SSE motifs such as different versions of beta hairpins, beta-alpha-beta, zinc fingers, etc. The complete list of 21 generic motifs and their occurrences is reported in the project webpage.

4.2. *Fold-preferential Motifs*

We have found a total of 110 fold-preferential motifs among which 75 are statistically significant in terms of both their fold-preferential and generic p-values. Among these 75 significant motifs, 67 are the 3-SSE and 8 are the 4-SSE motifs. 9 of the fold-preferential motifs overlap with the generic motifs.

The motifs are found in 8 out of the 15 SCOP Folds investigated. The number of motifs found for each Fold is given in Table 1. We rank the motifs by their fold-preferential p-values. The distributions of the p-values of both kinds for those 75 motifs are shown in Fig. 6.

The highest-ranked motif has the lowest fold-preferential p-value of 3.22×10^{-11}, and the genetic p-value of 1.40×10^{-5}. It is preferential to SCOP Fold `c.2`. It occurs in 10 proteins in `c.2`, but only in 2 proteins in the rest of the database. It is also similar to a version of the three-stranded beta hairpin motif [8, 10] with two parallel and one angled beta strands as shown in Fig. 8.

We have found a 4-SSE motif as our third-ranked motif. It is a beta-beta-beta-alpha motif (Fig. 9) which resembles the sequence order preserved version described in [18]. It has the fold-preferential p-value of 2.56×10^{-8}, and the genetic p-value of 3.80×10^{-17}. It is preferential to SCOP Fold `c.3`. It exists in 7 proteins in `c.3`, but only in 1 protein in the rest of the database.

We have also discovered a number of other biologically relevant motifs as our fold-preferential motifs. We report the full list of the 75 fold-preferential motifs in the project webpage.

It is observed that we have achieved our objective of formalization and specification of the super-SSE motifs as discussed in Section 2.1. Different versions of the motifs with the same verbal description can be classified based on their structural configurations. For example, we are able to distinguish the two different versions of the 3-SSE motifs resembling the three-stranded beta hairpin [8, 10] as shown in Figures 7 and 8. It has been previously observed that super-SSEs or SSE packings

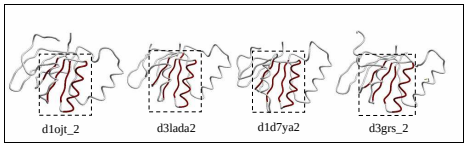

Fig. 9. Some instances of the rank #3 fold-preferential motif in SCOP Fold `c.3`: a 4-SSE motif resembling a beta-beta-beta-alpha motif.

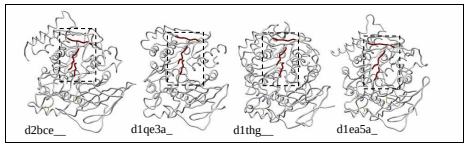

Fig. 10. Instances of a 3-SSE T-shape fold-preferential motif in SCOP Fold `c.69` (fold-preferential rank #44).

with the same SSE components but with the different configurations correspond to different biological functions [12, 20]. As such, it can be conjectured that those different versions of motifs may have different functions. However, an in-depth biological analysis will be required to verify this.

In addition, we have also discovered some new types of super-SSE motifs (both generic and fold-preferential) whose shapes have not been documented in the literature yet. For example, we have discovered a 3-SSE T-shape motif preferential to SCOP Fold `c.69` as shown in Fig. 10. Biologists can further investigate their detailed structural and functional properties, and possibly explore their potential usability in biomedical applications such as drug target finding.

5. Conclusion

In this paper, we have proposed a method to formalize the representation of sequence order independent super-SSEs, and mine the frequent super-SSE motifs from a large data set of protein structures. We have shown that our method is both effective and efficient. It can discover the generic and fold-preferential motifs that are statistically significant and biologically interesting within a short time. Biologists can further explore our discovered motifs to find out the potential usability of them in biomedical applications.

References

[1] Artymiuk, P. J., Spriggs, R. V., and Willett P., Graph theoretic methods for the analysis of structural relationships in biological macromolecules, *J. Am. Soc. Info. Sci. Tech.*, 56:518–528, 2005.

[2] Aung, Z. and Tan, K. L., Automatic 3D protein structure classification without structural alignment, *J. Comp. Biol.*, 12:1221–1241, 2005.

[3] Bradley, P., Kim, P. S., and Berger B., TRILOGY: discovery of sequence–structure patterns across diverse proteins, *Proc. Natl Acad. Sci., USA*, 99:8500–8505, 2002.

[4] Branden, C. and Tooze, J., *Introduction to Protein Structure*, Garland Publishing, 2nd edition, 1999.

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

[6] Coatney, M. and Parthasarathy, S., MotifMiner: efficient discovery of common substructures in biochemical molecules, *Knowl. & Infom. Sys.*, 7:202–223, 2005.

[7] Cordella, L. P., Foggia, P., Sansone, C., and Vento, M., An improved algorithm for matching large graphs, *Proc. IAPR GbRPR'01*, 149–159, 2001.

[8] Das, C., Raghothama, S., and Balaram, P., A designed three stranded beta-sheet peptide as a multiple beta-hairpin model, *J. Am. Chem. Soc.*, 120:5812–5813, 1998.
[9] Dror, O., Benyamini, H., Nussinov, R., and Wolfson, H., MASS: multiple structural alignment by secondary structures, *Bioinformatics*, 19(Suppl. 1):i95–i104, 2003.
[10] Efimov, A. V., Super-secondary structures involving triple-strand beta-sheets, *FEBS Lett.*, 334:253–256, 1993.
[11] Eidhammer, I., Jonassen, I., and Taylor, W. R., Protein structure comparison and structure patterns, *J. Comp. Biol.*, 7:685–716, 2000.
[12] Fernandez-Fuentes, N., Oliva, B., and Fiser, A., A supersecondary structure library and search algorithm for modeling loops in protein structures, *Nucleic Acids Res.*, 34:2085–2097, 2006.
[13] Frishman, D. and Argos, P., Knowledge-based secondary structure assignment, *Prot. Struct. Funct. Genet.*, 23:566–579, 1995.
[14] He, H. and Singh, A. K., GraphRank: statistical modeling and mining of significant subgraphs in the feature space, *Proc. ICDM'06*, 885–890, 2006.
[15] Huan, J., Bandyopadhyay, D., Prins, J., Snoeyink, J., Tropsha, A., and Wang, W., Distance-based identification of spatial motifs in proteins using constrained frequent subgraph mining, *Proc. CSB'06*, 227–238, 2006.
[16] Huan, J., Bandyopadhyay, D., Wang, W., Snoeyink, J., Prins, J., and Tropsha, A., Comparing graph representations of protein structure for mining family-specific residue-based packing motifs, *J. Comp. Biol.*, 12:657–671, 2005.
[17] Jonassen, I., Eidhammer, I., Conklin, D., and Taylor, W. R., Structure motif discovery and mining the PDB, *Bioinformatics*, 18:362–367, 2002.
[18] Kagawa, W., Kurumizaka, H., Ishitani, R., Fukai, S., Nureki, O., Shibata, T., and Yokoyama, S., Crystal structure of the homologous-pairing domain from the human Rad52 recombinase in the undecameric form, *Mol. Cell*, 10:359–371, 2002.
[19] Koch, I., Lengauer, T., and Wanke, E., An algorithm for finding maximal common subtopologies in a set of protein structures, *J. Comp. Biol.*, 3:289–306, 1996.
[20] Kurochkina, N. and Privalov, G., Heterogeneity of packing: structural approach, *Protein Sci.*, 7:897–905, 1998.
[21] Mitchell, E. M., Artymiuk, P. J., Rice, D. W., and Willett, P., Use of techniques derived from graph theory to compare secondary structure motifs in proteins, *J. Mol. Biol.*, 212:151–166, 1989.
[22] Pei, J., Jiang, D., and Zhang, A., On mining cross-graph quasi-cliques, *Proc. SIGKDD'05*, 228–238, 2005.
[23] Rao, S. T. and Rossman, M. G., Comparison of super-secondary structures in proteins, *J. Mol. Biol.*, 76:241–256, 1973.
[24] Sun, Z. and Blundell, T., The pattern of common supersecondary structure (motifs) in protein database, *Proc. HICSS'95*, 312–318, 1995.
[25] Szustakowski, J. D., Kasif, S., and Weng, Z., Less is more: towards an optimal universal description of protein folds, *Bioinformatics*, 21(Suppl. 2):ii66–ii71, 2005.
[26] Taylor, W. R. and Thornton, J. M., Prediction of super-secondary structure in proteins, *Nature*, 301:540–542, 1983.
[27] Wang, J., Zeng, Z., and Zhou, L., CLAN: an algorithm for mining closed cliques from large dense graph databases, *Proc. ICDE'06*, 73, 2006.
[28] Yan, X. and Han, J., CloseGraph: mining closed frequent graph patterns, *Proc. SIGKDD'03*, 286–295, 2003.
[29] `http://astral.berkeley.edu/`
[30] `http://scop.mrc-lmb.cam.ac.uk/scop/`
[31] `http://softsurfer.com/Archive/algorithm_0106/algorithm_0106.htm`