Using Similarity Flooding for Extracting Similar Parts of Proteins

Hassan Sayyadi [*], Sara Salehi [†], and Mohammad Ghodsi [‡]

Abstract—**Proteins are the main players in the game of life. Good understanding of their structures, functions, and behaviors leads to good understanding of drugs, diseases, and thus our health. So, much effort has been done to study and categorize proteins. Nowadays, tens of thousands of proteins have been found. Moreover, the problem of comparing the proteins is hard. Therefore efficient methods are needed to deal with this problem. In this paper, we used one important computational geometric method and one graph matching method: "Delaunay Tetrahedralization" and "Similarity Flooding" to propose a new idea to extract similar parts of proteins by combining both of these methods.**

Keywords: Protein Matching, Protein Comparison, Tetrahedralization, Similarity Flooding

1 Introduction

The number of known proteins is increasing every day; tens of thousands have been studied and categorized by now.To understand the functions and behaviors of a newly found protein, one should find well studied proteins with similar structure. In fact, the behavior of a protein is related to its sequence of amino acids and its 3D structure. So the comparison of proteins is a key technique not only in finding similarities in the structures of proteins but also to categorize them and to define families and super-families of the proteins. Like many comparison problems, this problem is hard because there is neither an exact definition of the likelihood of proteins structures nor an efficient algorithm exits for it. Although there exist optimal dynamic programming algorithms for comparing the sequences of amino-acids, the result is highly related to the definition of the relations of the sequences, which has not been uniquely defined [1]. The problem of comparing the 3D-structures of proteins becomes even harder. There is no efficient algorithm which guarantees the optimality of the answer. In fact, this problem is NP-hard. When the proteins become more complicated, the relationship models are more varied even than the models of sequence relatedness.

In this paper we will propose a model for protein matching or extracting similar parts of two given proteins. We focus on the computational geometric approach and the graph matching method used to model and compare the sequence and 3D-structure of proteins.

The remainder of this paper is organized as follows: We first have a glance at the related works. There are two major methods used in the literature: "Delaunay Tetrahedralization" and "Similarity Flooding". We will explain the required information in the next section as the background knowledge, and then propose a new idea in section 4 which can improve the current methods. We will then present experimental results of the implemented method which shows its effectiveness.

2 Related works

Delaunay Triangulation and Delaunay Tessellation are common computational geometric methods used in the bioinformatics. For example in [2] the Delaunay tessellation the α-carbons of the protein molecule is used to study the HIV-1 protease. This is because this model provides objective and robust definition of four nearest-neighbor amino acid residues as well as a four-body statistical potential function. The other usages are studied in the fields of packing analysis [2, 3], fold recognition [4], virtual mutagenesis [5], and structure comparison.

Authors of [6] consider the Delaunay Tetrahedralization determined by the alpha carbon positions of some particular protein. Starting at the amino-terminal residue, the edges of the Tetrahedralization that connect to a residue that has already been encountered are recorded as a relative residue difference. For example, if there is an edge between the 5th alpha carbon and the 3rd one, this edge is represented as 2. When the edge of a particu-

[*]Computer Engineering Department, Sharif University of Technology, Iran. Email: sayyadi@ce.sharif.edu

[†]Computer Engineering Department, Azad University Tehran-South branch, Iran. Email: sarasalehi@ace.tju.ir

[‡]Computer Engineering Department, Sharif University of Technology, Iran. IPM School of Computer Science, Tehran, Iran: This author's work has been partly supported by IPM School of CS (contract: CS1385-2-01). Email: ghodsi@sharif.edu

lar residue is exhausted a 0 is recorded to indicate a new residue. This linear representation will contain each edge in the Tetrahedralization exactly once. Furthermore, secondary structural components will be indicated by particular subsequences. Two one-dimensional representations are then compared by a dynamic programming scheme adapted from protein sequence analysis, thus reducing protein structural similarity to sequence similarity of the appropriate structure strings. In [7] the Euclidean metric for identifying natural nearest neighboring residues via the Delaunay tessellation in Cartesian space and the distance between residues in sequence space. In addition, authors of [8] find recurring amino-acid residue packing patterns, or spatial motifs, that are characteristic of protein structural families, by applying a novel frequent sub-graph mining algorithm to graph representations of protein three-dimensional structure. Graph nodes represent amino acids, and edges are chosen in one of three ways: first, using a threshold for contact distance between residues; second, using Delaunay tessellation; and third, using the recently developed almost-Delaunay edges.

3 Background Knowledge

3.1 Delaunay Tetrahedralization

Delaunay Tetrahedralization is a special type of Tetrahedralization which is defined based on the Voronoi diagram through the principle of duality. A Voronoi box is formed through the intersection of planes and is therefore a general irregular polyhedron. The facets of the Voronoi boxes correspond in the dual graph to the Delaunay edges which connect the points of P.

- *Delaunay Edge*: Let P be a finite set of points in a sub-domain Ω^n of the n-dimensional space R^n. Two points p_i and p_j are connected by a Delaunay edge e if and only if there exists a location $x \in \Omega^n$ which is equally close to p_i and p_j and closer to p_i, p_j than to any other $p_k \in P$. The location x is the center of an n-dimensional sphere which passes through the points p_i, p_j and which contains no other points p_k of P.

$$\begin{aligned} e_{Delaunay}(p_i, p_j) \quad &\Leftrightarrow \quad \exists x: \ x \in \Omega^n \\ &\wedge \quad |x - p_i| = |x - p_j| \\ &\wedge \quad \forall k \neq i, j: \ |x - p_i| < |x - p_k| \end{aligned}$$

Combining this criterion for the three edges of a triangle (Fig. 1) and furthermore for the four triangles of a tetrahedron leads to the following criterion for Delaunay tetrahedron.

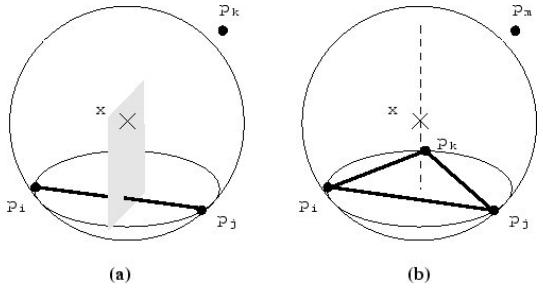

Figure 1: a) Delaunay Edge b) Delaunay Triangle criteria.

- *Delaunay Tetrahedron*: Let P be a finite set of points in a sub-domain Ω^n of the n-dimensional space R^n, where $n \geq 3$. Four non-coplanar points p_i, p_j, p_k and p_l form a Delaunay tetrahedron T if and only if there exists a location $x \in \Omega^n$ which is equally close to p_i, p_j, p_k and p_l and closer to p_i, p_j, p_k, p_l than to any other $p_m \in P$. The location x is the center of an n-dimensional sphere which passes through the points p_i, p_j, p_k, p_l and which contains no other points p_m of P. For $n = 3$ only one such sphere exists which is the circumsphere of T.

$$\begin{aligned} T_{Delaunay}(p_i, p_j, p_k, p_l) \Leftrightarrow \exists x: \ & x \in \Omega^n \\ \wedge \quad & |x - p_i| = |x - p_j| = |x - p_k| = |x - p_l| \\ \wedge \quad & \forall m \neq i, j, k, l \ : \ |x - p_i| < |x - p_m| \end{aligned}$$

A Delaunay tetrahedron must consist of Delaunay edges and Delaunay triangles. The edge and triangle criteria are implicit, because the existence of the n-dimensional sphere in *Delaunay Edge* criterion and in *Delaunay Triangle* criterion is guaranteed by the sphere in *Delaunay Tetrahedron* criterion.

3.2 Similarity Flooding

Matching or finding similar elements of two data schemas or two data instances plays a key role in data warehousing, e-business, or even biochemical applications. Authors of [9] present a matching algorithm named "Similarity Flooding" based on a fixpoint computation that is usable across different scenarios. As the example illustrating the Similarity Flooding Algorithm is shown in Fig. 2, the algorithm takes two graphs (schemas, catalogs, or other data structures) as input, and produces as output a mapping between corresponding nodes of the graphs.
As a first step, the schemas should be translated from their native format into graphs G_1 and G_2. Next the pair-wise connectivity graph (PCG) should be made that

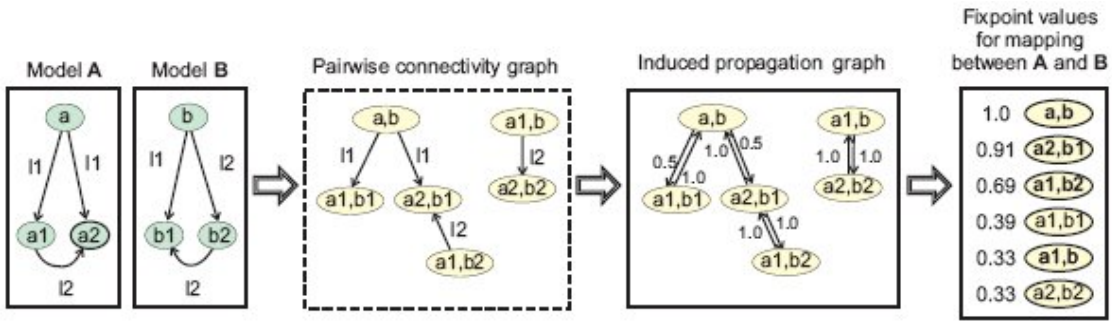

Figure 2: Example illustrating the Similarity Flooding Algorithm

is an auxiliary data structure derived from G_1 and G_2. If N_1 represent the set of all nodes in G_1 and respectively N_2, each node in the connectivity graph is an element from $N_1 \times N_2$ and is called "map-pair". Furthermore, edges in connectivity graph defined as follow:

$$\begin{aligned} ((x_1, y_1), P, (x_2, y_2)) \quad &\in \quad PCG(G_2, G_2) \\ &\Leftrightarrow \\ (x_1, P, x_2) \in G_1 \quad &and \quad (y_1, P, y_2) \in G_2 \end{aligned}$$

Each map-pair contains one node from each graph and the similarity score between them. The initial similarity for each map-pair is obtained using a simple string matcher that compares common prefixes and suffixes of literals in each node. Finally, computing the similarities relies on the intuition that elements of two distinct models are similar when their adjacent elements are similar. In other words, a part of the similarity of two elements propagates to their respective neighbors as follow:

$$\begin{aligned} \sigma^{k+1}(x\ , y) = \sigma^k(x, y) \\ + \sum_{(a_i,x)\in G_1,(b_i,y)\in G_2} \sigma^k(a_i, b_i).W((a_i, b_i), (x, y)) \\ + \sum_{(x,a_i)\in G_1,(y,b_i)\in G_2} \sigma^k(a_i, b_i).W((x, y), (a_i, b_i)) \end{aligned}$$

where $\sigma^k(x, y)$ shows the similarity between x and y after iteration k and $W((a_i, b_i), (x, y))$ is the propagation weight of the similarity between a_i and b_i to the similarity between x and y. The above computation is performed iteratively until the Euclidean length of the residual vector $\triangle(\sigma^n, \sigma^{n-1})$ becomes less than ϵ for some $n > 0$. If the computation does not converge, it will be terminated after some maximal number of iterations.

4 Proposed Method

In this section we present the proposed approach. Here we combine sequence similarity which is a simple extension of amino acid or nucleotide similarity and structural similarity which is the residues position similarity. Using both techniques leads to an efficient method for extracting similar parts of proteins.
The different phases of the proposed method may be represented as follow:

1. Protein Tetrahedralization
2. Creating pair-wise graph
3. Similarity propagation
4. Extracting similar components.

We will discuss each phase in the following subsections.

4.1 Protein Tetrahedralization

Each Protein is a sequence of residues in the 3D space in which each two consequent residues are connected by one edge called "chain-edge". Firstly, for each protein, we use Delaunay Tetrahedralization algorithm to convert the protein sequence to a tetrahedralized shape (Fig. 3(c)). Since all proteins have 3D shape, using Delaunay algorithm leads to create edges called "Tetrahedralization-edge" between atoms which are close to each other in space, regardless of their distance in protein sequence. This closeness has an extremely high influence on structural similarity which will be discussed in proposed method for extracting similar parts of proteins.
Inasmuch as Tetrahedralization algorithm creates convex shape, in order to have a much more similar shape to the real protein shape, we need to eliminate edges whose length are more than α for Tetrahedralization-edges and more than β for chain-edges. Obviously, the value of β is more than α, because of the importance of chain-edges in proteins comparison (Fig. 3(d)). We now construct a graph from the tetrahedralized shape. Each node in this graph contains one number which is the amino acid

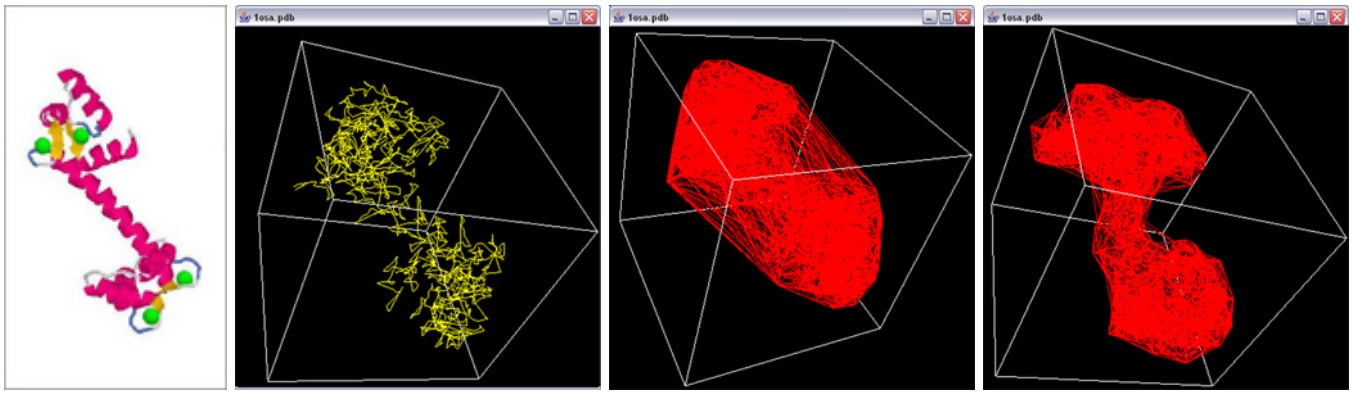

Figure 3: a) Protein b) Protein chain c) Tetrahedralized protein d) Tetrahedralized protein after removing worthless edges

number of corresponding atom in protein and coordinates (x, y, z) expressing the coordination of that atom. This graph has two different types of edges:

1. Edges which belong to the protein chain and also can be the part of the tetrahedrons named chain-edges.
2. Edges obtained from Tetrahedralization which do not belong to the protein chain named Tetrahedralization-edges.

Consequently, each protein is converted to one graph which not only contains the protein chain but also contains edges connecting atoms near each other in 3D space. These graphs are data structures for similarity flooding algorithm used in protein matching.

4.2 Creating Pair-wise Graph

Pair-wise Connectivity graph (PCG) arises from two protein's graphs P_1 and P_2 which were created through Tetrahedralization in the first phase. If N_1 and N_2 show the sets of all nodes in P_1 and P_2, each node in the pair-wise graph is an element from $N_1 \times N_2$. We call such nodes map-pairs. The edges of pair-wise graph are categorized to 3 parts depending on their map-pairs (see Fig. 4):

1. If a chain-edge exists between the first nodes of two map-pairs and there is a chain-edge between the second nodes of those map-pairs in their proteins, then we will connect these two map-pairs with an edge of type chain.

$$((x, y), CH, (x', y')) \in PCG(P_1, P_2) \Leftrightarrow$$
$$(x, CH, x') \in P_1 \text{ and } (y, CH, y') \in P_2$$

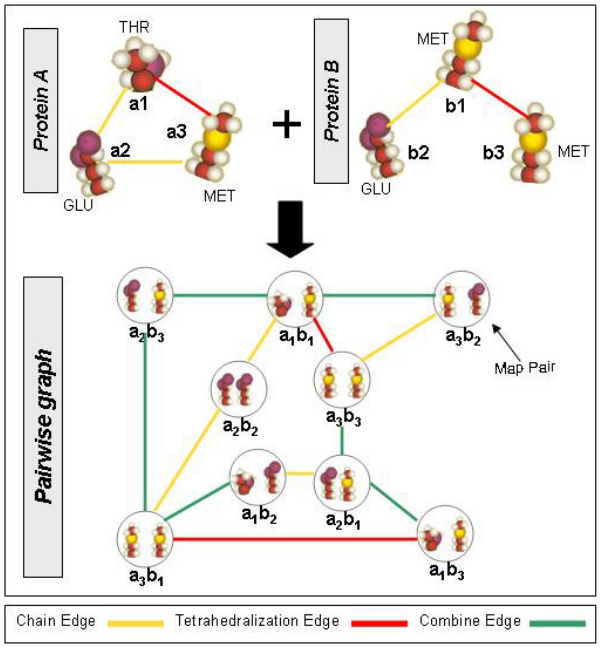

Figure 4: Pairwise connectivity graph for proteins

where CH represents the edge of type chain.

2. If a Tetrahedralization-edge exists between the first nodes of two map-pairs and there is a Tetrahedralization-edge between the second nodes of those map-pairs in their proteins, then we will connect these two map-pairs Tetrahedralization-edge.

$$((x, y), T, (x', y')) \in PCG(P_1, P_2) \Leftrightarrow$$
$$(x, T, x') \in P_1 \text{ and } (y, T, y') \in P_2$$

where T represents the edge of type Tetrahedraliza-

tion.

3. If a Tetrahedralization-edge exists between the first nodes of two map-pairs and there is a chain-edge between the second nodes of those map-pairs in their proteins or vice versa, then we will connect these two map-pairs with edge of type combine.

$$\begin{array}{rcl} ((x,y),C,(x',y')) & \in & PCG(P_1,P_2) \Leftrightarrow \\ (x,T,x') \in P_1 & and & (y,CH,y') \in P_2 \\ & or & \\ (x,CH,x') \in P_1 & and & (y,T,y') \in P_2 \end{array}$$

where C represents the edge of type combine.

We categorized these edges to three types, because the influence of their nodes in similarity propagation weights in the proposed method will differ from each other.

4.3 Similarity Propagation

In the created pair-wise graph, the primary similarity of each map-pair depends on the similarity between two nodes of that map-pair which are the atoms of the proteins. This similarity derives from amino acids scoring matrix. The two-dimensional matrix contains all possible pair-wise amino acid scores. Scoring matrices are also called substitution matrices because the scores represent relative rates of evolutionary substitutions. According to the similarity flooding algorithm the similarity of a map-pair increment based on the similarities of its neighbors in the pair-wise graph. Hence, the similarities of neighbors are affective in calculating final similarity between two atoms in each map-pair. It seems more rational for neighbors related to chain-edges to be much more affective in protein matching. Similarly, the weight of neighbors related to combine edges is more than those of Tetrahedralization-edges. Thus, over a number of iterations, the initial similarity of any two nodes propagates through the graphs. Similarity propagation in each iteration is computed as follows:

$$\begin{array}{rcl} Sim^{i+1}(x,y) & = & a * Sim^i(x,y) \\ & + & (1-a) * NeighborAffect(x,y) \end{array}$$

$$\begin{array}{rcl} NeighborAffect(x,y) & = & CHF * CHAffect^i/NF \\ & + & CF * CAffect^i/NF \\ & + & TF * TAffect^i/NF \end{array}$$

In the above equation, $Sim^i(x,y)$ defined as the similarity between x and y in each map-pair after i number of iteration(s), and a is the learning rate from the neighbors. Moreover,

- *CHAffect* is the average similarity of neighbors connecting with chain-edge to the respective map-pair and is calculated as bellow:

$$CHAffect^i(x,y) = \sum_{\substack{((x,y),CH,(x_i,y_i)) \\ \in PCG(P_1,P_2)}} \frac{Sim^i(x_i,y_i)}{CHSize}$$

  and CHF is the propagation weight of *CHAffect*.

- *CAffect* is the average similarity of neighbors connecting with combine edge to the respective map-pair and is defined as bellow:

$$CAffect^i(x,y) = \sum_{\substack{((x,y),C,(x_i,y_i)) \\ \in PCG(P_1,P_2)}} \frac{Sim^i(x_i,y_i)}{CSize}$$

  and CF is the propagation weight of *CAffect*.

- *TAffect* is the average similarity of neighbors connecting with Tetrahedralization-edge to the respective map-pair and is computed as bellow:

$$TAffect^i(x,y) = \sum_{\substack{((x,y),T,(x_i,y_i)) \\ \in PCG(P_1,P_2)}} \frac{Sim^i(x_i,y_i)}{TSize}$$

  and TF is the propagation weight of *TAffect*.

In the above formula, $CHSize$ ($CSize$ and $TSize$) is the number of edges of type chain (combine and Tetrahedralization) connected to the map-pair. The sum of the CHF,CF and TF must be equal 1 to restrict *NeighborAffect* between valid rang which will be discussed in experimental result. Furthermore, NF is normal factor applying to cases in which there is no neighbors related to edges of one type. For example, assume that there isn't any neighbor related to Tetrahedralization-edge, but there are edges of chain and combine types. Hence, we should set $NF = CHF + CF$ to normalize *NeighborAffect* into valid rang.

4.4 Extracting Similar Components

Due to the fact that the similarity degree of each map-pair in pair-wise graph expresses the matching degree of its atoms, we should extract similar components of two proteins by eliminating map-pairs and their related edges in pair-wise graph which have similarity degree less than γ. Consequently, the pair-wise graph transform to forest in which we have several connected components. Each connected component declares one similar part of two proteins, and each map-pair in connected components expresses matched atoms between two proteins. Similarly, each edge in connected components shows conforming

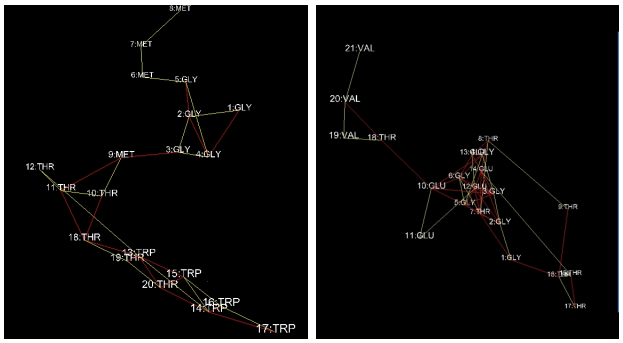

Figure 5: a) Protein number 1 after Tetrahedralization b) Protein number 2 after Tetrahedralization

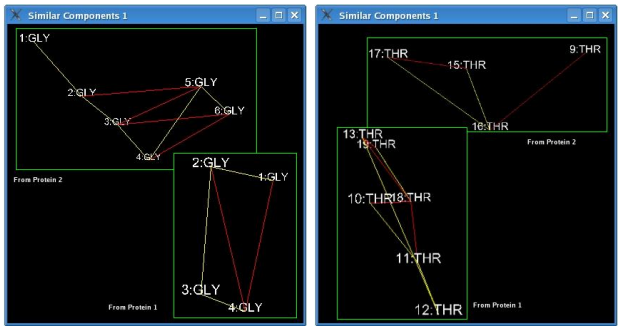

Figure 6: Extracted similar components

edges between two proteins. Hence, by extracting each protein's nodes and edges from the connected component, we obtain two connected sub-graphs, each of which belongs to one of two given proteins. Connected components with the number of nodes less than η are not valuable for the result of matching, therefore they should be removed. For example, assume that you have two pictures, and you want to match them. Obviously if one pixel of them is analogous, you can not assert that these two pictures are the same or you find valuable matching. Hence, the number of nodes of each connected components should be noticed and connected components containing less than η number of atoms should be eliminated.

5 Experimental Result

In this section we will explain our implementation of the proposed method. We read protein information from PDB files. Then we used Visad[1] package to do Delaunay Tetrahedralization on the input protein chain, this Tetrahedralization contributes our sequence chain of protein to convert to tetrahedralized shape which is needed in counting structural similarity in proposed protein matching algorithm. We can construct filters that apply $\alpha = 2$ for creating boundary for Tetrahedralization-edges length and $\beta = 5$ for removing worthless chain-edges. Hence, edges which are longer than these thresholds were removed. Fig. 5 shows the accuracy obtained after applying above filters in two proteins shape.
Two tetrahedralized shapes in the form of graph data structure were used to create pair-wise graph. After this creation, we used amino acid scoring matrices presented in Blast book[2] to assign primary similarity to each map-pair. Inasmuch as a part of the similarity of two nodes in each map-pair propagates to their respective neighbors, we propagated the similarities of map-pair via proposed equation in which CHAffect,CAffect and TAffect were defined as the affect of neighbors of the map-pairs. The final similarities were obtained after applying filters shown in bellow.

Parameter	a	CHF	CF	TF
Value	0.8	0.5	0.3	02

As you see in the above table, owing to the importance of neighbors in chain, we set the CHF factor more than two times of the TF factor, and correspondingly the CF factor more than the TF factor.
The similarity values between two amino acids in the scoring matrix are between -4 and +11, and these similarities between one amino acid with itself are between +4 and +11. After propagating, the similarities still remained between -4 and +11. Hence, to avoid removing map-pairs which contain two identical nodes or two different nodes with reasonable similitude degree, we set the threshold $\gamma = +3$. Therefore, map-pairs whose similarities are less than this threshold were eliminated. After this removal, components which included more than $\eta = 3$ atoms from each protein were extracted. Extracted similar components from small parts of two proteins which shown in Fig. 5 are represented in Fig. 6 (Owning to have clear figures, we choose very small part of proteins for this figures). In addition, due to the fact that will be explained in conclusion section, our experiments applied on some parts of proteins instead of complete set of atoms in each protein.

6 Conclusion and Future Works

In this paper we proposed a novel method used to extract similar parts of proteins based on computational geometry and graph matching methods. Our method used the Delaunay Tetrahedralization of the α-carbon atoms

[1] http://www.ssec.wisc.edu/ billh/visad.html
[2] http://safari.oreilly.com/0596002998/blast-CHP-4-SECT-3

in the protein molecules to add some edges in protein structure for short distance nodes in counting structural similarity. This method can build a robust model for protein matching but it needs some enhancements. Because of the large number of residues in each protein, the pair-wise graph created from two proteins will contain a large number of map-pairs and therefore large amount of memory is needed. Hence, we need some heuristics to reduce pair-wise graph nodes. For example, we can remove or do not create any map-pair with initial similarity less than a defined threshold. Furthermore, we need some test collections for optimizing proposed model parameters such as similarity propagation weight of each type in propagation graph and thresholds used in the model. Moreover, this model will be test on more real proteins with real size. In our experiments we use some parts of real proteins instead of complete structure of real protein because of mentioned problems.

Moreover, our proposed method can extracts similar parts of the two given protein precisely. Furthermore, it includes both structural and sequential similarity. Our model has great flexibility in all aspects and by changing different model parameters such as propagation weight of different edge type, we can change the influence of structural and sequence similarity.

References

[1] I. Eidhammer, I. Jonasses, and W. Taylor, "Structure comparison and structure patterns," 2000. [Online]. Available: citeseer.ist.psu.edu/eidhammer99structure.html

[2] J. Finney, "Random packing and the structure of simple liquids, the geometry of the random close packing," in *Proc R Soc*, 1970, p. 479493.

[3] A. Tropsha, C. Carter, S. Cammer, and I. Vaisman, "Simplicial neighborhood analysis of protein packing (snapp) a computational geometry approach to studying proteins," in *Methods Enzymol*, 2003, p. 509544.

[4] W. Z. S. Cho, I. Vaisman, and A. Tropsha, "A new approach to protein fold recognition based on delaunay tessellation of protein structure," in *Pacific Symposium on Biocomputing*, Singapre, 1997, pp. 487–496.

[5] C. Carter, B. LeFebvre, S. Cammer, A. Trosha, and M. Edgell, "Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations," in *J Mol Biol*, 2001, p. 625638.

[6] J. Roach, S. Sharma, M. Kapustina, and C. Carter, "Structure alignment via delaunay tetrahedralization," in *Proteins: Structure, Function, and Bioinformatics*, vol. 60, no. 1, April 2005, pp. 66–81.

[7] D. Bostick, M. Shen, and I. Vasiman, "A simple topological representation of protein structure: Implications for new, fast, and robust structural classification," in *Proteins: Structure, Function, and Bioinformatics*, vol. 56, no. 3, 2004, pp. 487–501.

[8] J. Hun, D. Bandyopadhyay, W. Wang, J. Snoeyink, J. Prins, and A. Trosha, "Comparing graph representations of protein structure for mining family-specific residue-based packing motifs," in *Journal of Computational Biology*, vol. 12, no. 6, 2005, pp. 657–671.

[9] S. M. H. Garcia-Molina and E. Rahm, "Similarity flooding: A versatile graph matching algorithm and its application to schema matching," in *Proc. 18th Intl. Conf. on Data Engineering(ICDE)*, San Jose CA, 2002.