An Improved Algorithm for Protein Structural Comparison Based on Graph Theoretical Approach

Cheng-Hsien Hsu [a], Sheng-Lung Peng*[b], and Yu-Wei Tsay [b]

[a] Department of Management Information System, Takming University of Science and Technology, Taipei, 11451, Taiwan.
[b] Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien, 97401, Taiwan.

*Author for correspondence; e-mail: slpeng@mail.ndhu.edu.tw

ABSTRACT

It is known that Proteins play crucial roles in most biological processes. It is also known that the function of a protein is determined by its structure. Thus, knowledge of the structures of proteins provides us a way toward the understanding of life science. However, common experimental methods, e.g., X-ray crystallography and NMR spectroscopy, are labored and high cost. Therefore, many studies have been made for the protein structural similarity. In this paper, we propose an improved algorithm based on graph theoretic approach for this problem. At first, a protein is transferred into a labeled graph according to its secondary structures, chemical properties, and topological relations. Next, for two graphs, the maximum common edge subgraph is computed for measuring the structural similarity of the corresponding proteins. By performing a practical technique, a maximum common edge subgraph of two graphs can be found efficiently. Finally, a common substructure of the given proteins can be found by a backtracking from the maximum common edge subgraph. Experimental results show that our method outperforms the RMSD method, especially in the evolutionary relatedness among various strains. This graph-based approach provides a practical direction for measuring protein structural similarity.

Keywords: graph theory, protein similarity, bioinformatics, RMSD.

1. INTRODUCTION

Proteins consist of a sequence of amino acid residues assembled together into a polypeptide chain, called the primary structure. The local conformation of the polypeptide chain or the spatial relationship of amino acid residues that are close together in the primary sequence leads to a secondary structure, e.g., the α helix and β strand. Tertiary structures are the result of secondary structures packing on a more global level. Many proteins contain more than one polypeptide chain. Result of another level packing or a group of different proteins

packed together is a 3-dimensional conformation called quaternary structure. Basically, the conformation of a protein determines its biological function, performing their faculty in organisms, driven by numbers of molecular interactions. Knowledge of a protein structure can yield useful information about the functional property of specific protein. Measuring protein structural similarity attempts to establish an equivalence relation between polymer structures based on their shapes and 3-dimensional conformations. This procedure is a valuable tool for the evaluation of proteins with low sequence similarity, especially when evolutionary relations between proteins cannot be easily detected by alignment techniques.

In general, proteins with high sequence identity and high structure similarity tend to possess functional conserved [1]. In most cases, the relations of protein sequence and structure function homology are well-recognized from conserved regions. However, the bias of expected sequence and structure similarity relationship still remains unexplored [2]. Numerous amino acid sequences may yield various structures, as well as similar sequences sometimes yield dissimilar structures. In [3], the authors illustrated an asymmetric relation of protein sequence and structure. Thus, comparison of protein structures is believed that can reveal distant evolutionary relationships that would not be detected by sequence information alone.

The root mean square deviation value (RMSD) [4], gives a direction to measure average distance of given structural proteins. RMSD attempts to calculate the minimal value from two superimposed proteins by translation and rotation, indicating the divergence between two proteins. It is also an excellent measure for nearly identical structures. The smaller RMSD value of the two structures, the more similar conformation they have. In addition, there is a fairly general agreement that RMSD requires performing an exhaustive search against whole proteins. Although the RMSD protocol is most popularly implemented, it suffers from a few drawbacks; little attention has been given to this point.

First, the approach of RMSD ignores many significant global alignments that may result in misjudgments. Because all atoms are equally weighted during the RMSD calculation, it becomes biased to a local structure than to a global one. It is inability to perform overall numbers of atoms. Only selected backbone atoms between two residues are chosen for least squares fitting. Another drawback is that the protein environments of physical-chemical properties are completely ignored, such as ligand complexity, compound reaction, and molecule bonding [5]. Third, although an RMSD value seeks to reach a global minimum between two proteins, there are no standard criteria for RMSD value. Unrelated proteins may have a large RMSD, but two relevant proteins consisting of identical substructures may also have [6]. Once the shapes of two proteins turn into divergent, RMSD loses its effectiveness. This may lead to a confusing analysis in such a case.

In this paper, we propose an algorithm to measure the similarity of given proteins. Instead of finding the minimum common superstructure on two proteins, we transfer proteins to graphs for finding their maximum common substructure, attempting to overcome the inherent drawback of RMSD.

2. GRAPH-BASED APPROACH

Graph similarity measure is an important topic in various fields. If graphs are used for representing structured objects, then the problem of measuring object similarity turns into the problem of computing the similarity
of graphs [7]. Therefore, numerous attempts have been made on graph similarity to show its efficiency in recent years.

Let $G = (V', E)$ be a graph where $V'$ is the vertex set of $G$ and $E$ is the edge set of $G$. Usually, a vertex represents an object and an edge $(u, v)$ represents a relation between vertices $u$ and $v$. In graph-based approach, a vertex can be a residue or a secondary structure element (SSE). The use of graph theory for comparing two protein structures starts from [8]. In this approach, two main issues are considered. The first issue is how to transform a protein into a graph. The other issue is how to determine the measurement for similarity. Once we know how to do for these two issues, the steps of proposed algorithm are as follows.

1. Transform proteins into graphs.
2. Compute the similarities according to our determined measurement.
3. Infer possible results, e.g., classification, evolutionary tree, and so on, according to their similarity scores.

In the remaining of this section, we describe our algorithm in detail according to the outline above. Figure 1 shows an example for our algorithm. The detailed transformation is presented in the following subsections.

2.1 Graph Transformation

Most algorithms in graph-based approach transform a protein into a labeled graph no matter considering residue level or SSE level. In residue level, a vertex can be labeled by its corresponding amino acid while in SSE level a vertex can be labeled by its corresponding SSE, e.g., helix, sheet (strand), or coil. In this paper, we also consider an extra attribute, i.e., polar or non-polar. For the edge set, most algorithms consider the geometry edge, determined by the distance of two vertices. Sometimes, by considering more possible edge relationships, we can obtain a more accurate representation. In this paper, we define the forbidden edges and remove them from the graph defined by the geometry edges.

**Figure 1.** In the graph model, each secondary structure is regarded as a vertex, transforming a protein structure to an undirected simple graph.
2.1.1 Constructing P-graphs

For convenience, the protein graph constructed from a protein is called a P-graph. In the construction, each vertex in the P-graph is created according to the dictionary of protein secondary structures (DSSP). Each vertex is labeled by two attributes according to its properties. The first attribute is about its shape type, e.g., H-helix (containing G, H, and I), T-hydrogen turn (containing T, E, and B), or C-coiled (containing only C) [9]. The second attribute is about its chemical property. The twenty kinds of amino acids are classified into three classes. The most important chemical property we considered is about the structure stabilization. For example, cysteines can form covalent disulfide bonds to other cysteine residues, and prolines can form a cycle to the polypeptide backbone. Thus the first class is \{C, P\}. Depending on the polarity of a side chain, a polypeptide chain behaves according to its hydrophilic or hydrophobic character [10]. Thus, the second class is the polar set \{R, N, D, E, Q, H, K, S, T, Y\} and the final class is the non-polar set \{A, G, I, L, M, F, W, V\}. This attribute is determined by the majority of amino acids with the same property in the secondary structure. For each vertex, we store these two attributes in a table, called P-Table. We summarize the construction of vertex set as follows.

Vertex set: \( V = \{v_1, v_2, \ldots, v_n\} \)

- \( v_i \); the vertex corresponding to the \( i \)-th secondary structure of the input protein.
- **DSSP Label**: \( D[v] \) = 1 for helix, 2 for sheet, and 4 for coil.
- **Polarity Label**: \( P[v] \) = 1 for polar, 2 for non-polar, and 4 for Cys or Pro.

Note that the purpose for defining the values of \( D[v] \) and \( P[v] \) in the power of two is for arithmetic computing.

Now we consider the construction of edge set for a P-graph. In proteins, a secondary structure is described by patterns of hydrogen bonds between backbone amide and carboxyl groups. Thus, an edge is defined by a correlated hydrogen bond. In this case, a connection is established among secondary structures sequentially, called physical edge. As in Figure 2(a), four substructures \( B_1, B_2, B_3, \) and \( B_4 \) are in the P-graph \( G_p \). The edges \( b_1 = (B_1, B_2), b_4 = (B_2, B_3), \) and \( b_5 = (B_3, B_4) \) are physical edges.

![Figure 2](image-url)

**Figure 2.** (a) A P-graph \( G_p \) with \( V_p = \{B_1, B_2, B_3, B_4\} \) and \( E_p = \{b_1, b_2, b_3, b_4, b_5\} \). (b) The P-Table of \( G_p \). (c) The line graph \( L(G_p) \) of \( G_p \).
Proteins are conformed by various chemical bonding, e.g., in Figure 2(a), $B_1$ is spatially adjacent to $B_4$ but being deficient in their relation. Thus, another connection is defined to declare the relation between vertices, called geometric edge. Let $d$ be a given value. Then, we add a geometric edge between two vertices if they are not connected by a physical edge and their distance is no greater than $d$. For example, in Figure 2(a), edges $b_1 = (B_1, B_4)$ and $b_3 = (B_1, B_3)$ are geometric edges. In addition, in order to avoid the connection of vertices with the same attribute in the P-graph label, we define the forbidden edges. The purpose is to delete those edges with matching properties due to the repulsive interaction among different regions of polypeptide chains [11]. We summarize the construction of edge set as follows.

**Edge set:** $E = \{e_1, e_2, \ldots, e_m\}$

- **Physical Edge:** $(v_i, v_j) \in E$ if $j = i + 1$.
- **Geometric Edge:** $(v_i, v_j) \in E$ if the distance $dist(v_i, v_j) \leq d$ in 3D space for a given threshold $d$.
- **Forbidden Edge:** Remove the above defined edges $(v_i, v_j)$ if both $D[v_i] = D[v_j]$ and $P[v_i] = P[v_j]$.

Note that in this paper the threshold $d$ is the minimum distance among all the distances of physical edges.

### 2.1.2 Constructing Line Graphs

For an undirected graph $G$, the line graph of $G$, denoted as $L(G)$, is the graph obtained from $G$ by letting the vertex set of $L(G)$ be the edge set of $G$ and any two vertices of $L(G)$ are connected by an edge if their corresponding edges in $G$ share a common end vertex. As illustrated in Figure 2(c), each vertex in the line graph $L(G_\mu)$ is labeled with the corresponding edge in the original graph $G_\mu$. For example, the edge $b_i$ in $G_\mu$ is adjacent to edges $b_j$ and $b_k$ via vertex $B_i$, and is adjacent to edge $b_l$ through vertex $B_i$. Therefore, there are edges connecting vertex $b_i$ to vertices $b_j$, $b_k$, and $b_l$ in $L(G_\mu)$.

For constructing modular graph (defined in the next subsection), the chemical relation of $G_\mu$ is also annotated in each vertex in $L(G_\mu)$ from its P-Table. For example, in Figure 2(b), vertex $b_i$ in $L(G_\mu)$ can be viewed as a group of chemical substances obtained by $B_i$ and $B_j$ in $G_\mu$. Thus $b_i$ will be labeled as $(3, 5)$ which is computed according to $(2+1, 4+1)$. In such a manner, $b_i$ is composed with substructures helix and sheet, and one of them is polar component.

### 2.1.3 Constructing Modular Graphs

A modular graph is also named compatibility graph. Table 1.

<table>
<thead>
<tr>
<th>PID</th>
<th>Fold</th>
<th>Domain</th>
<th>Species</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2EYA</td>
<td>Crambin-like</td>
<td>Crambin</td>
<td>Abyssinian cabbage</td>
<td>Plant thionin</td>
</tr>
<tr>
<td>2FD7</td>
<td>Crambin-like</td>
<td>Crambin</td>
<td>Abyssinian cabbage</td>
<td>Plant thionin</td>
</tr>
<tr>
<td>1AB1</td>
<td>Crambin-like</td>
<td>Crambin</td>
<td>Abyssinian cabbage</td>
<td>Plant thionin</td>
</tr>
<tr>
<td>1CRN</td>
<td>Crambin-like</td>
<td>Crambin</td>
<td>Abyssinian cabbage</td>
<td>Plant thionin</td>
</tr>
<tr>
<td>1OKH</td>
<td>Crambin-like</td>
<td>Viscotoxin A3</td>
<td>European mistletoe</td>
<td>Plant thionin</td>
</tr>
<tr>
<td>1ED0</td>
<td>Crambin-like</td>
<td>Viscotoxin A3</td>
<td>European mistletoe</td>
<td>Plant thionin</td>
</tr>
<tr>
<td>1OLR</td>
<td>Crambin-like</td>
<td>Viscotoxin C1</td>
<td>European mistletoe</td>
<td>Plant thionin</td>
</tr>
<tr>
<td>1WUW</td>
<td>Crambin-like</td>
<td>Beta-Hordothionin</td>
<td>Barley</td>
<td>Plant thionin</td>
</tr>
<tr>
<td>1NBL</td>
<td>Crambin-like</td>
<td>Hellethionin D</td>
<td>Helleborus urpurascens</td>
<td>Plant thionin</td>
</tr>
<tr>
<td>1H5W</td>
<td>Upper collar protein gp10</td>
<td>Upper collar protein gp10</td>
<td>Bacteriophage phi-29</td>
<td>Phage Connector</td>
</tr>
</tbody>
</table>
The process of constructing modular graph from $L(G_A)$ and $L(G_B)$.

Table 2. Corresponding values of CATH codes for selected macromolecules are shown. C is for secondary structure content; A is for general spatial arrangement of secondary structures; T is for spatial arrangement and connectivity of secondary structures; H is for manual curation of evidence of evolutionary relationship; S indicates $\geq 35\%$ sequence similarity; O indicates $\geq 60\%$ sequence similarity; L indicates $\geq 95\%$ sequence similarity; I indicates 100% sequence similarity; and D means unique domains.

<table>
<thead>
<tr>
<th>PID</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>H</th>
<th>S</th>
<th>O</th>
<th>L</th>
<th>I</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>2EYA</td>
<td>3</td>
<td>30</td>
<td>1350</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>2FD7</td>
<td>3</td>
<td>30</td>
<td>1350</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>1AB1</td>
<td>3</td>
<td>30</td>
<td>1350</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1CRN</td>
<td>3</td>
<td>30</td>
<td>1350</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1OKH</td>
<td>3</td>
<td>30</td>
<td>1350</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1ED0</td>
<td>3</td>
<td>30</td>
<td>1350</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1ORL</td>
<td>3</td>
<td>30</td>
<td>1350</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1WUW</td>
<td>3</td>
<td>30</td>
<td>1350</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1NBL</td>
<td>3</td>
<td>30</td>
<td>1350</td>
<td>10</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1H5W</td>
<td>3</td>
<td>30</td>
<td>1350</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

A simple idea is that two graphs are more similar, the larger the common induced subgraph is. In this paper, we use the maximum common edge subgraph (MCES) for measurement. Another possibility is to use the minimum common supergraph [13].

To find a maximum common subgraph $G_{ab}$ of $G_A$ and $G_B$, a usual way is to construct the line graphs of $G_A$ and $G_B$ first. Then we combine these two line graphs into a modular graph $M_{ab}$.

Finally, the problem of finding the maximum common subgraph $G_{ab}$ of $G_A$ and $G_B$ can be reduced to the problem of finding the maximum clique of $M_{ab}$ [14, 15]. It is the reason why we introduce the constructions.
of line graphs and modular graphs. As the modular graph $M_{ia}$ has been constructed, an MCES finding between $G_a$ and $G_b$ is simply to locate the maximum clique in $M_{ia}$ and then backtracks to the original graphs $G_a$ and $G_b$. Note that the problem of finding a maximum clique for a graph is NP-hard [16]. However, it can be solved by a well known branch-and-bound algorithm [17]. Currently, it is doable for graphs with up to 15,000 vertices [18]. That is, this approach is suitable for proteins with up to 15,000 residues or SSE. Fortunately, in the SSE level, most proteins are below this scale.

Once the maximum common edge subgraph $G_{ab}$ was found from $G_a$ and $G_b$, a formula is needed to measure the similarity of proteins $A$ and $B$. In view of MCES, it is required to consider both vertices and edges in subgraph $G_{ab}$. Thus, the Jaccard similarity coefficient is adapted for our graph scoring function [19]. The Jaccard index is commonly a statistic used for comparing the similarity and diversity of sample sets. The concept of Jaccard index can be expressed as the following equation:

$$sim(G_a, G_b) = \frac{(|V(G_a)|+|E(G_a)|)^2}{(|V(G_a)|+|E(G_a)|+|V(G_b)|+|E(G_b)|)}$$

In this formula, the denominator is the multiplication of the size of the graph $G_a$ and the size of the graph $G_b$ and the numerator is a square of the size of the graph $G_{ab}$ [20]. In this respect, it quantifies the relation between $G_a$ and $G_b$ by their MCES $G_{ab}$ from 0 (dissimilar) to 1 (identical).

3. RESULTS

To validate the protein structure comparison of our graph-based approach, we tested some small proteins of plant crambin. Ten protein structures are selected as materials, namely, 2EYC, 2FD7, 1AB1, 1CRN, 1OKH, 1ED0, 1ORL, 1WUW, 1NBL, and 1H5W. As illustrated in Table 1 (from protein data bank (PDB) [21]), all the molecular descriptions are listed. For the further explanation of proposed results, it may be helpful to refer the protein database - CATH [22] (Class, Architecture, Topology, and Homologous) shown in Table 2. It is a hierarchical domain classification of protein structures in PDB classified according to a combination of automated and manual procedures.

Table 3 summarized the structural comparison from the viewpoint of RMSD to P-graph. In the following lines, we will discuss it in detail from two aspects. First, although RMSD aligns the structural

<table>
<thead>
<tr>
<th>PID</th>
<th>2FD7</th>
<th>1AB1</th>
<th>1CRN</th>
<th>1OKH</th>
<th>1ED0</th>
<th>1ORL</th>
<th>1WUW</th>
<th>1NBL</th>
<th>1H5W</th>
</tr>
</thead>
<tbody>
<tr>
<td>2EYA</td>
<td>1.2</td>
<td>0.92</td>
<td>1.1</td>
<td>0.75</td>
<td>1.1</td>
<td>0.33</td>
<td>1.4</td>
<td>0.30</td>
<td>1.6</td>
</tr>
<tr>
<td>2FD7</td>
<td>-</td>
<td>0.5</td>
<td>0.5</td>
<td>0.33</td>
<td>0.7</td>
<td>0.28</td>
<td>0.9</td>
<td>0.28</td>
<td>0.9</td>
</tr>
<tr>
<td>1AB1</td>
<td>-</td>
<td>-</td>
<td>0.4</td>
<td>0.44</td>
<td>0.6</td>
<td>0.40</td>
<td>1.0</td>
<td>0.40</td>
<td>1.0</td>
</tr>
<tr>
<td>1CRN</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.1</td>
<td>0.00</td>
<td>1.0</td>
<td>0.00</td>
<td>1.0</td>
<td>0.00</td>
</tr>
<tr>
<td>1OKH</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.9</td>
<td>1.00</td>
<td>0.9</td>
<td>1.00</td>
<td>0.9</td>
<td>1.00</td>
</tr>
<tr>
<td>1ED0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.2</td>
<td>1.00</td>
<td>0.9</td>
<td>0.83</td>
<td>1.5</td>
</tr>
<tr>
<td>1ORL</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.9</td>
<td>0.83</td>
<td>1.7</td>
<td>0.00</td>
</tr>
<tr>
<td>1WUW</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.7</td>
<td>0.00</td>
<td>1.7</td>
</tr>
</tbody>
</table>
similarity from one another, it is hard to realize the correlation within a group by RMSD. For example in Table 3, protein 2FD7 suggests that it should be similar to protein 1AB1 according to RMSD (0.5Å). However, according to the annotations and molecule descriptions in Table 1 and Table 2, 2FD7 is much more identical to 2EYA. In this case, by using our approach, it is able to correctly compare one protein and its superfamily. For 2FD7 to 2EYA, the proposed method offers a better solution (0.92) than RMSD. Here is another example which seems to support our concern. In respect of RMSD, 1OKH indicates the nearness with 2FD7, 1AB1, 1CRN, and 1WUW even 1OKH is a viscotoxin protein that is irrelevant to crambin protein or hordothionin.

Similar examples are abundant. Hence, only a bare general sketch can be given of. For the further explanation of Table 3, evolutionary tree is constructed for helping to illustrate the consequence of Figures 4 and 5 by PHYLIP [24] using neighbor-joining [25].

From the annotation of Table 1 and Table 2, it would be possible to account for comparative relationships as follows: 2EYA, 2FD7, 1AB1, and 1CRN (crambin protein) are believed as closely related, so does 1OKH, 1ED0, and 1ORL (viscotoxin protein). Besides, 1NBL and 1H5W would be relatively distant from crambin protein and viscotoxin protein. As mentioned above, it is uncertain to verify one relative relationship from RMSD. Comparing with Figures 4 and 5, it may be enough to explain the proposed hypothesis.

Second, the proposed method presents a convincing viewpoint, showing an inferred evolutionary relationship among various strains. There are several examples. According to P-graph matching, one may notice that 1OKH, 1ED0, and 1ORL are mutually shape-related. The same observation can also be applied to Figure 5, which is tending to unanimity with CATH code and PDB annotations. It is not saying that RMSD cannot be utilized in protein structural comparison; instead, the value of RMSD should be described together with other parameters, e.g., sequence identity, number of carbon alpha’s, backbone atoms, and so on. Nevertheless, our results show that our approach is more competitive with existing RMSD approach, although it relies on more information than RMSD. For completeness, we show a schematic view of selected crambin proteins.
Figure 5. A phylogenetic tree is inferred by proposed rule.

Figure 6. Protein structures with display style in schematic view.

In Figure 6. Since the structures of 1AB1 and 1CRN are very similar, we omit the structure of 1AB1.

In this study, attributes of edges are grading into three types: Type-1 edges join the initial relation from secondary structures; Type-2 edges determine the topological relation to improve an integrity in P-graph, and Type-3 edges are deleted due to the repulsive interaction. Adopting from several
properties on vertices and edges, protein comparison could be improved by incorporating with more specific parameters using the graph-based approach.

3.1 Program and Environment
We have implemented an MCES finding procedure described in the previous sections, and tested with small protein PDB datasets. The environment is running under 2GHz PC with 512MB of main memory in platform of Linux-2.6.11-1.1369. The implementation is temporarily written using Bash-3.00.16(1) and Octave-3.0.0. It may be worth pointing out that the construction of a P-graph is limited by the number of secondary structures. Take protein 1CXR for example. In PDB file, 1CXR contains only one helix structure. Therefore, it would be untrue to transfer a single node to its line graph or modular graph tested.

4. CONCLUSIONS
Graphs are a powerful and useful tool for various subfields of science and engineering. In many applications, the relations of objects can be modeled as graphs. Therefore, measuring the similarity of objects becomes determination the similarity of graphs.

In this paper, we propose a method to determine a structural similarity of two proteins based on graph theory. With the help of MCES finding, it is efficient to locate a relative similarity of protein structures. Comparing with general RMSD approach and its ability, our method shows an alternative conception and promotive advantage on its efficiency. This graph-based approach provides a practical direction for protein structural comparison.

In the future, we may expect that the concept of the graph-based approach can be applied to multi-domain protein comparison.

5. REFERENCES


