

PROCEEDINGS

Open Access

# Adjusting protein graphs based on graph entropy

Sheng-Lung Peng\*, Yu-Wei Tsay

From 2013 International Conference on Intelligent Computing (ICIC 2013)  
Nanning, China. 28-31 July 2013

## Abstract

Measuring protein structural similarity attempts to establish a relationship of equivalence between polymer structures based on their conformations. In several recent studies, researchers have explored protein-graph remodeling, instead of looking a minimum superimposition for pairwise proteins. When graphs are used to represent structured objects, the problem of measuring object similarity become one of computing the similarity between graphs. Graph theory provides an alternative perspective as well as efficiency. Once a protein graph has been created, its structural stability must be verified. Therefore, a criterion is needed to determine if a protein graph can be used for structural comparison. In this paper, we propose a measurement for protein graph remodeling based on graph entropy. We extend the concept of graph entropy to determine whether a graph is suitable for representing a protein. The experimental results suggest that when applied, graph entropy helps a conformational on protein graph modeling. Furthermore, it indirectly contributes to protein structural comparison if a protein graph is solid.

## Background

Graph theory is now widely used in information theory, combinatorial optimization, structural biology, chemical molecule, and many other fields. Graph similarity measuring is a practical approach in various fields. When graphs are used to represent of structured objects, the problem of measuring similarities between objects becomes one of computing similarities between graphs [1]. Protein remodeling is another field wherein multiple-domains within structures are considerably complicated.

It is believed that proteins are important molecules for living organisms. In fact, they are essential parts of organisms and participate in almost every process within cells. A protein contains at least one linear chain of amino acid residues called a *polypeptide*. By various synthesis, *e.g.*, biosynthesis and chemical synthesis, a polypeptide is folded into a unique 3-dimensional structure. Usually, the structure of a protein determines its biological function performed in organisms. Knowledge of a protein structure can help us understand biological functions and evolution. Measuring protein similarities according to 3-dimensional structures of proteins

provides a valuable tool for evaluating proteins with low sequence similarities when evolutionary relations among proteins cannot be detected by sequence alignment techniques. To perform a structural comparison of molecules, accurate information of two superimposed protein structures must be obtained. However, optimizing these two quantities simultaneously is difficult. Unlike the sequence alignment problem, the structural alignment problem has not even been classified as solvable.

For decades, studies have attempted to define topological relations and notations on protein structures, a schematic description is essentially expected to describe its topology. Mathematical formulations of structural patterns can facilitate the composition in a polypeptide chain. A schematic description has the advantage of simplicity, making the implementation of graph theory as an alternative approach possible [2]. By selectively ignoring protein structural features, it has the potential to detect further homologous relationships based on various geometric methods and motivations.

The structure of a protein can be regarded as a conformation with various local elements (*e.g.*, helices, sheets) and forces (*e.g.*, Van der Waal's forces, hydrogen bonds), folding into its specific characteristic and functional

\* Correspondence: [slpeng@mail.ndhu.edu.tw](mailto:slpeng@mail.ndhu.edu.tw)  
Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien 974, Taiwan

structure. With the help of graph transformation, folded polypeptide chains can be represented as a graph using several mapping rules. Proteins contain complex relationships in its polymer: residual reactions, covalent interactions, peptide bonding, and hydrophobic packing are essential parts in structural determination. The intention is to transform a protein structure into a graph. Formally, a *graph transforming system* consists of a set of graph rewrite rules:  $L \rightarrow R$ , which  $L$  is called pattern graph and  $R$  is called replacement graph [3]. It is the key operation in graph transformation.

### Protein Remodeling

As mentioned to the protein remodeling, a study reviewed in detail of protein graph (abbreviated as P-graph) description can be found in [4]. Usually, the vertex set of a P-graph can be defined by  $C_\alpha$  atoms, residues, side chains, DSSP (the dictionary of protein secondary structures), and SSE (secondary structure elements). For the edge set, it is usually defined by the distance of two vertices with some labels, e.g., chemical properties. Figure 1 shows an overview of protein graph remodeling. Table 1 shown an outline of some categories of the protein graph approach to a set of graphs, representing each specific graph rewriting and graph measuring skills. Therefore, it is useful to begin with the summarized common research into the following matters: *geometric relation* and *chemical relation*.

Proteins have been represented in various ways using different levels of detail. The conformation of protein structure has been shown to be determined geometrically by various constraints [5]. Therefore, the most common method for protein modeling is to reserve its topological relationship in graphs. From the perspective of graph theory, a simplified representation of protein structure aims attention at connectivity patterns. It helps to go into details on interacted relation within a

**Table 1 Recent studies for constructing protein graphs.**

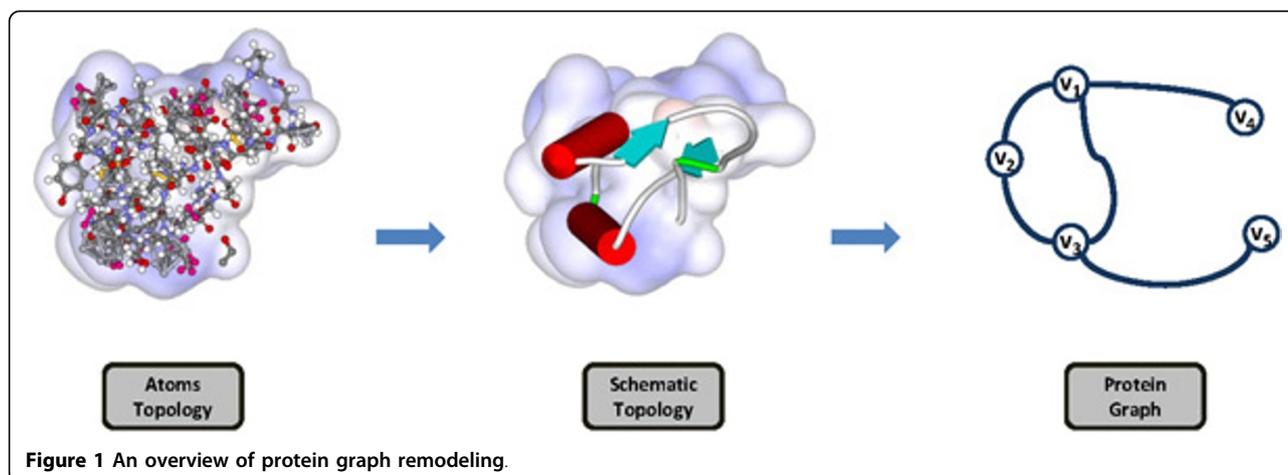
Ref.	Vertex Set	Edge Set
[20]	$C_\alpha$ atoms	labeled edges
[21]	DSSP	attributed edges
[22]	side chains	defined by interacted energy
[23]	residues	defined by geometrical constraints
[24]	SSE	labeled edges

polypeptide folding. In brief, the geometric-based protein modeling is to refine its edges (relations) among vertices (objects), adapting the information from inter-object distances for all pairs of objects.

Comparing with geometric relationship, chemical properties provides a more complicated description in the protein graph model. Amino acid contain various chemical properties, including electrostatic charge, hydrophobicity, bonding type, size and specific functional groups [6]. By giving values to edges and nodes in graph, each different labeled component that varies between the various types of chemical relation.

### Entropy

*Entropy* defines a quantitative equilibrium property within a system and it implies the principle of disorder from the second law of thermodynamics [7]. It is particularly important in describing how energy is applied and transferred in an isolated system. The higher the disorder, the greater the entropy of the system [8]. Similarly, this concept is presented in life. As we known, life is composed of many cells, tissues, and organs from the vital element of protein. Since proteins are biochemical compounds, consisting of one or more polypeptide chains, the arrangement of protein polymers are assumed to be in a compact state, according to its backbone dihedral angles and side chain rotamers. This is called *conformational entropy*. There is considerable evidence to prove that the same



**Figure 1** An overview of protein graph remodeling.

observation can be applied to a protein graph model. In such a case, a graph model should also follow the second law of thermodynamics.

For an  $n$ -object system  $G$ , assume that each object  $i$  is associated with a probability  $p_i$ . Then the entropy of the system  $G$  is defined as in Formula 1 [9].

$$I(G) = \sum_{i=1}^n -p_i \times \log_2 p_i \quad (1)$$

In graph theory, the entropy of a graph is usually defined by its degree sequence. For example, we consider the cycle with 4 vertices, *i.e.*,  $C_4$ . The degree sequence is (2, 2, 2, 2). Thus, the  $p_i$  for each vertex  $v_i$  is  $\frac{2}{8} = 0.25$ . By definition,  $I(C_4) = -4 \times 0.25 \times \log_2(0.25) = 2$ .

### Methods

In this section, we extend the concept of graph entropy to measuring protein graphs. To demonstrate the calculation of graph entropy exemplarily, peptide chains of MHC (Major Histocompatibility Complex) are selected as the materials for examining the utilities of graph entropy.

#### Graph entropy

For a given graph  $G = (V, E)$  and two vertices  $u$  and  $v$  in  $V$ , let  $d(u, v)$  denote the length of the shortest path between  $u$  and  $v$ . Let  $N_k(u) = \{v \mid d(u, v) = k\}$ . In graph theory,  $N_k(u)$  is called the  $k$ -distance neighborhood of  $u$  and is also called the  $k$ -sphere of  $u$  [10]. By counting  $k$ -distance neighbors of  $v_i$ , it gives a good account of nodes mutual connectivity in  $G$ . We define the following formula.

$$f(u) = \sum_{i=1}^k \frac{|N_i(u)|}{n - i + 1} \quad (2)$$

In Formula 2,  $k$  is the longest length for  $u$  to reach to a vertex, (*i.e.*,  $N_k(u) \neq \emptyset$  but  $N_{k+1}(u) = \emptyset$ ). The idea of our formula makes that every other vertex  $v$  contributes an impact to the current vertex  $u$ . In particular, the closer distance between  $v$  and  $u$ , the greater the impact of  $v$ . For simplicity, we let  $f(V) = \sum_{v \in V} f(v)$ . Assume that  $V = \{v_1, v_2, \dots, v\}$ . We define  $q_i$  for each  $v_i$  as follows.

$$q_i = \frac{f(v_i)}{f(V)} \quad (3)$$

Finally, our modified entropy formula for a graph  $G = (V, E)$  is as follows.

$$I'(G) = \sum_{i=1}^n -q_i \times \log_2 q_i \quad (4)$$

For convenience, we consider the graph depicted in Figure 1 as an example which is a P-graph based on small proteins of the plant crambin. This graph is an unlabeled graph corresponding to the protein. The following equations are easy to obtain:

$$\begin{aligned} f(v_1) &= f(v_3) = \frac{3}{5} + \frac{1}{4} \\ f(v_2) &= \frac{2}{5} + \frac{2}{4} \\ f(v_3) &= f(v_5) = \frac{1}{5} + \frac{2}{4} + \frac{1}{3} \end{aligned}$$

So the entropy of graph depicted in Figure 1 is:

$$\begin{aligned} I'(G) &= -q(v_1) \log_2 q(v_1) - q(v_2) \log_2 q(v_2) \\ &\quad - q(v_3) \log_2 q(v_3) - \dots - q(v_5) \log_2 q(v_5) \\ &= -2q(v_1) \log_2(q(v_1)) - q(v_2) \log_2(q(v_2)) \\ &\quad - 2q(v_4) \log_2(q(v_4)) \\ &= -2 \times 0.3188 \times \log_2(0.3188) - 0.2455 \times \log_2(0.2455) \\ &\quad - 2 \times 0.2214 \times \log_2(0.2214) \\ &= 2.4835 \end{aligned} \quad (5)$$

Let us consider the four graphs depicted in Figure 2. They are  $C_4$ , a cycle of four vertices,  $K_4$ , a clique of four vertices,  $P_4$ , a path of four vertices, and  $S_4$ , a star of four vertices. For  $C_4$  and  $K_4$ , since the four vertices are in the same situation, they have the same probability. Thus  $I(C_4) = I'(C_4) = I(K_4) = I'(K_4) = \log_2(4) = 2$ . However, for  $P_4$  and  $S_4$ , we have  $I(P_4) = 1.918$  and  $I(S_4) = 1.793$ . By Formula 4, we have  $I'(P_4) = 1.894$  and  $I'(S_4) = 1.995$ . The densities of  $P_4$  and  $S_4$  are the same, (*i.e.*, 0.5). However, the diameter of  $P_4$  is greater than that of  $S_4$ . According to traditional graph entropy,  $I(S_4) < I(P_4) < I(K_4) = I(C_4)$ . However, in our formula,  $I'(P_4) < I'(S_4) < I'(K_4) = I'(C_4)$ . Intuitively,  $S_4$  is more compact than  $P_4$ . Thus, our formula

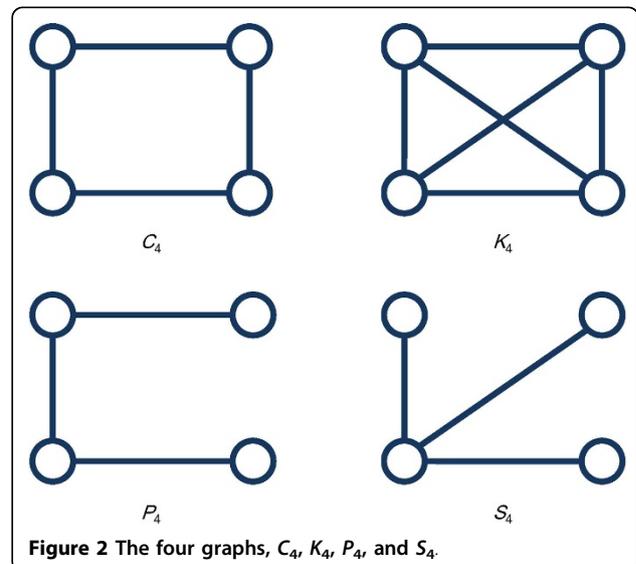


Figure 2 The four graphs,  $C_4$ ,  $K_4$ ,  $P_4$ , and  $S_4$ .

makes a better decision. Note that in graph entropy, the higher entropy of a graph indicates that the graph structure is more stable.

**Edge adjustment**

By the definition of  $I'(G)$ , its value is not increased monotonously if the density of  $G$  is increased. Thus, we have the following cases to determine how to adjust the graph. Assume that  $G$  is the current graph and  $I'(G) = x$ . Let  $I'(G - e) = y$  and  $I'(G + e) = z$  where  $G - e$  means that we remove the longest edge from  $G$  and  $G + e$  means that we add a shortest non-edge to  $G$ .

- **Case 1:**  $y = 0$  It means that after this edge is removed,  $G$  is no longer a connected graph.
- **Case 2:**  $z > x > y$  It means that by adding a new edge,  $G$  will become more stable.
- **Case 3:**  $x > z > y$  It means that  $G$  is stable enough.
- **Case 4:**  $y > x > z$  It means that by removing an old edge,  $G$  will become more stable.

As illustrated in Figure 3, it shows when edges are added or removed from a graph, their entropy values will be changed. A set of connected 5-node graphs is shown in Figure 4.

**Graph spectra**

Given two graphs  $G_A = (V_A, E_A)$  and  $G_B = (V_B, E_B)$ , the graph matching problem is to find a one-one mapping  $f : V_A \rightarrow V_B$  such that if  $(u, v) \in E_A$ , then the possibility of  $(f(u), f(v)) \in E_B$  is as higher as possible. Therefore, numerous attempts have been made on graph similarity to show its efficiency in recent years. In [11], it revealed that the problem of graph matching may be divided into different types depended on their levels. According to the graph scoring, it qualitatively measures a mutual

dependence of two objects [12]. Generally, a value of similarity ranges between 0 and 1, from dissimilar to identical.

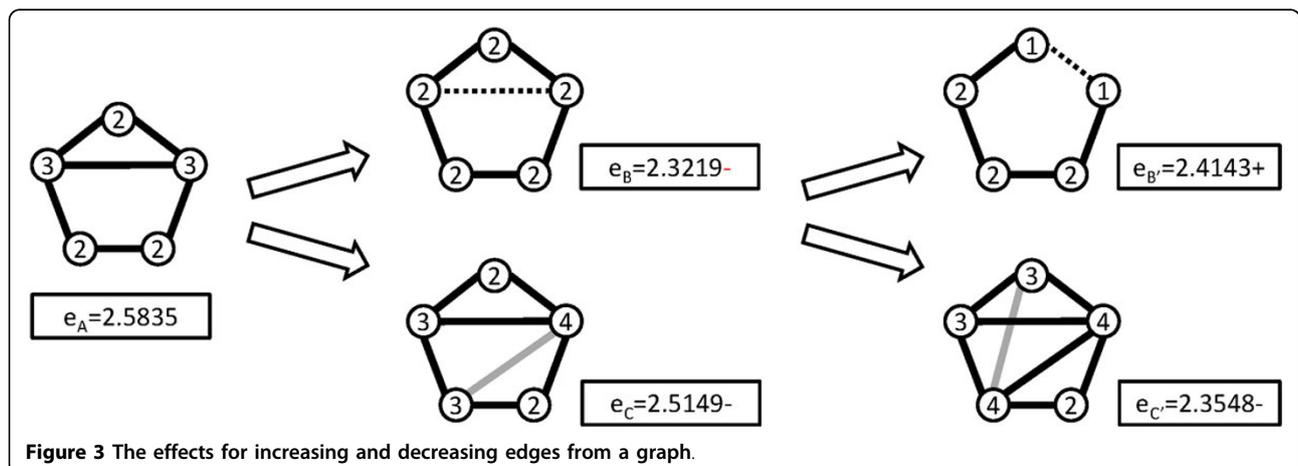
Occasionally, topologies of graphs are complicated; therefore, one practical way is to symbolize it as matrix, turning graph into numbers and vectors. Since it is hard to determine graph isomorphism, *graph spectra* gives an alternative solution for graph matching. By definition, a spectrum of a finite graph  $G$  is the spectrum of its adjacency matrix  $A_G$  and diagonal degree matrix  $D_G$ , whose entries  $a_{ij}$  and  $d_{ij}$  can be written in Formula (6) and Formula (7), respectively. That is, its connected neighbors of eigenvalues together with their multiplicities [13].

$$a_{ij} = \begin{cases} 1, & \text{if } (i, j) \in E, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

$$d_{ij} = \begin{cases} \text{deg}(v_i), & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases} \tag{7}$$

The *Laplacian spectrum* of  $G$  is the matrix,  $L_G = D_G - A_G$ , indicating a topological properties and connectedness of the graph. In brief, a graph spectra of  $G$  can be regarded as a set of *eigenvectors*-  $\lambda$  [14]. Apparently, comparing with the binary relation of graph  $G$ , the spectrum of  $G$  tends to improve its information on adjacent relation. We give some examples to describe a graph spectrum transformation. Let  $X$  be the resulting graph by removing one edge from  $K_4$ . Let  $Y$  be the graph depicted in Figure 1.

$$L_X = \begin{pmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{8}$$



**Figure 3** The effects for increasing and decreasing edges from a graph.

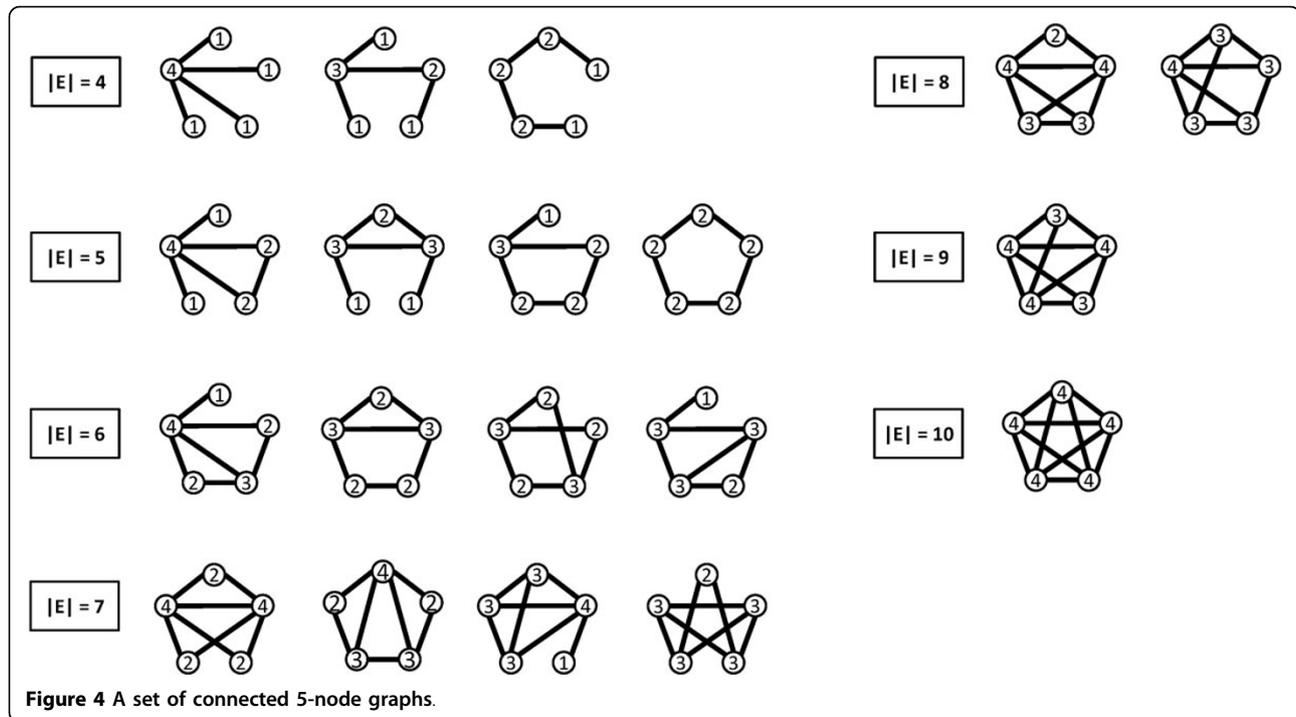


Figure 4 A set of connected 5-node graphs.

$$L_Y = \begin{pmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 3 & 0 & -1 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{pmatrix} \quad (9)$$

Obviously,  $|V_X| = 4 < |V_Y| = 5$ . Therefore, when spectra are different sizes, the smaller one may be padded with zero values to equalize the size of  $G_X$  and  $G_Y$ . By definition, the spectra  $\lambda_X$  and  $\lambda_Y$  can be obtained, i.e.,  $\lambda_X = [4, 4, 2, 0, 0]$  and  $\lambda_Y = [4, 4, 1, 1, 0]$ . The similarity between  $G_X$  and  $G_Y$  can be simply measured by the *Euclidean distance* of  $\lambda_X$  and  $\lambda_Y$ . In this case, the similarity of  $G_X$  and  $G_Y$  is 1.414.

## Results and discussion

In this experiment, we validated the remodeling function of the P-graph by using extended graph entropy to verify the stability of a given P-graph. For the P-graph construction, please refer to [15]. Thus, we were interested in only the impact of connectivity on protein structural similarities. Various types of MHC were chosen as the material to verify the verification of proposed method: **1HDM**, **1K5N**, **2ENG**, **1VCA**, **1ZXQ**, **1UXW**, **1A2Y**, **3ARD**, **2Q3Z**, and **2CRY**. MHC, as an immune system in most vertebrates, encodes for a small complex cell surface protein. It is also known for HLA (Human Leukocyte Antigen), one of the most intensively studied genes in human [16]. Due to a great diversity of microbes in the environment, MHC genes vary widely its peptide through several mechanisms

[17]; this is also the major reason why MHC proteins were selected as materials for this studies.

## P-graph comparison

Let  $G = (V, E)$  be the P-graph after remodeling from the construction proposed by [15]. Vertices of  $V$  in  $G$  are created according to the DSSP. Under this metric, a protein secondary structure is represented by a single letter code, H-helix (containing **G**, **H**, and **I**), T-hydrogen turn (containing **T**, **E**, and **B**), or C-coiled (containing only **C**). For controlling one variable in this experiment, let the edge set  $E$  of  $G$  be changed from a specific range.

A comparison of MHC proteins is shown in Table 2. In the table, **PID** is the protein identification number in PDB [18]. Since MHC proteins are composed of multiple polypeptide chains, they are multimeric **Domain**. Furthermore, **Dens** means the density in the graph. It is defined as  $\frac{2|E|}{|V|(|V|-1)}$  ranging from 0 to 1. **AVG** indicates the average distance within DSSP vertices. If the distance of  $v_i$  and  $v_j$  is no greater than **AVG**, then there is an edge between them. In the table,  $+ke$  ( $-ke$ ) means that we add (remove) the  $k$  shortest (longest) possible edges. For example,  $+1e$  means that we add the edge with the shortest length that is greater than **AVG**. In the table, **NaC** indicates that the resulting graph is not a connected graph.

The relationship between  $|E|$  and  $I'(G)$  is as follows. First, when the density in  $G$  increases, the graph  $G$  appears to go from sparse to dense. However, its extended entropy

**Table 2 A selected proteins with corresponding extended entropies**

PID	- 3e	- 2e	- 1e	AVG	+ 1e	+ 2e	+ 3e
<b>1HDM</b>	3.343	3.396	3.563	3.319	3.705	3.845	3.765
Dens	0.357	0.393	0.464	0.524	0.535	0.607	0.643
<b>1K5N</b>	4.305	5.545	5.564	4.537	4.614	4.732	3.787
Dens	0.436	0.457	0.475	0.509	0.527	0.564	0.571
<b>2ENG</b>	4.000	4.091	4.144	4.212	4.294	4.344	4.480
Dens	0.422	0.444	0.467	0.489	0.511	0.533	0.578
<b>1VCA</b>	3.106	3.171	3.254	3.221	3.249	3.467	3.493
Dens	0.381	0.429	0.476	0.524	0.571	0.619	0.667
<b>1ZXQ</b>	3.494	3.551	3.641	3.709	3.774	3.712	3.907
Dens	0.429	0.464	0.500	0.535	0.571	0.607	0.643
<b>1UXW</b>	5.562	5.563	5.646	5.764	5.855	5.950	6.079
Dens	0.456	0.463	0.478	0.500	0.515	0.529	0.551
<b>1A2Y</b>	NaC	NaC	2.414	2.507	2.581	2.512	2.510
Dens	-	-	0.400	0.500	0.600	0.700	0.800
<b>3ARD</b>	4.460	4.641	4.698	4.756	4.801	4.860	4.932
Dens	0.424	0.470	0.485	0.500	0.515	0.530	0.554
<b>2Q3Z</b>	6.611	6.730	6.775	6.834	6.885	6.996	7.302
Dens	0.474	0.486	0.493	0.503	0.511	0.525	0.547
<b>2CRY</b>	NaC	NaC	NaC	NaN	NaN	NaN	NaN
Dens	-	-	-	0.667	1.000	1.000	1.000

does not increase completely with its density. It seems a little anomalous in this appearance. Second, the edge set in protein remodeling issue can be determined from its extended entropy. By definition, the P-graph  $G$  should be a connected graph. Once the  $G$  becomes a disconnected graph, we cannot decide its entropy. For example, **1A2Y** is not a connected graph when the density is lower than 0.400. Third,  $E$  appears to be considerably related to  $V$  in graph entropy. Consider the P-graph **2CRY** as another example. If a protein remodeling function adapts a specific value on the basis of its geometrical edge, then it might be an error to assume a fixed value as a criterion. This is an

essential fact to stress. It is worth pointing out that the construction of a P-graph is limited by  $V$ .

#### P-graph verification

To validate the previous assumptions, a method for protein structural comparison is adapted to measure its similarity. Graph spectra gives an alternative solution to graph matching. It is a set of relational parameters, consisting of a characteristic polynomial and eigenvectors of its adjacency matrix or Laplacian matrix. Graph spectra quantitatively provide graph information, e.g., structure, topology, connectivity [19]. In Table 3 we list the results of protein

**Table 3 A comparison of protein structure remodelings.**

PID		1K5N	2CRY	1VCA	2Q3Z	1ZXQ	1A21	2ENG	1UXW	1A2Y	3ARD
<b>1HDM</b>	Old	7.93	23.36	15.68	24.01	13.74	6.54	12.57	7.92	5.75	8.27
	New	7.75	21.12	13.87	23.67	12.11	5.64	11.03	7.25	5.41	7.79
	Result	+	+	+	+	+	+	+	+	+	+
<b>1K5N</b>	Old	.	26.58	19.55	26.91	18.02	14.65	17.69	20.44	18.41	25.72
	New	.	23.70	17.39	21.13	15.99	12.83	15.84	17.17	16.94	23.64
	Result	.	+	+	+	+	+	+	+	+	+
<b>2CRY</b>	Old	.	.	14.87	12.33	17.13	14.39	15.62	19.33	6.81	18.42
	New	.	.	12.91	34.10	14.92	12.45	17.54	19.35	5.17	19.63
	Result	.	.	+	-	+	+	-	=	+	-
<b>1VCA</b>	Old	.	.	.	17.71	5.39	4.83	7.75	11.42	5.45	12.80
	New	.	.	.	29.68	4.47	3.21	6.82	10.07	4.83	11.67
	Result	.	.	.	-	+	+	+	+	+	+

**Table 3 A comparison of protein structure remodelings. (Continued)**

<b>2Q3Z</b>	Old	New	.	.	.	.	27.57	26.31	29.30	28.35	23.46	21.11	25.45	24.52	38.30	36.72	24.14	23.00
	Result	.	.	.	.	.	+		+		+		+		+		+	
<b>1ZXQ</b>	Old	New	.	.	.	.	.	.	3.98	3.41	7.96	7.49	10.52	9.67	6.21	6.53	9.14	8.87
	Result	.	.	.	.	.	.	.	+		+		+		-		+	
<b>1A21</b>	Old	.	.	.	.	.	.	.	.	.	6.24		12.76		7.37		12.85	
	New	.	.	.	.	.	.	.	.	.	5.41		11.38		6.91		11.06	
	Result	.	.	.	.	.	.	.	.	.	+		+		+		+	
<b>2ENG</b>	Old	.	.	.	.	.	.	.	.	.	.	.	4.65		11.42		14.19	
	New	.	.	.	.	.	.	.	.	.	.	.	4.17		10.39		13.82	
	Result	.	.	.	.	.	.	.	.	.	.	.	+		+		+	
<b>1UXW</b>	Old	.	.	.	.	.	.	.	.	.	.	.	.	.	16.24		5.71	
	New	.	.	.	.	.	.	.	.	.	.	.	.	.	15.41		3.45	
	Result	.	.	.	.	.	.	.	.	.	.	.	.	.	+		+	
<b>1A2Y</b>	Old	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	12.24	
	New	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	11.41	
	Result	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+	

structure remodeling matters. The field **Old** shows a remodeling based on the specific value of edge length, and **New** indicates that the edges in  $G$  are adjusted by extended entropy. The value in each entry is the distance of the two spectra. If our method obtains a better result in the comparison, then we simply mark “+” to denote a better result; otherwise, it is marked “=” (not bad) or “-” (worse). Table 4 shows the CATH codes for the selected macromolecules. In summary, the extended entropy determines a better conformational graph from protein structure remodeling.

#### Program and environment

The procedure for computing the extended entropy for a P-graph was implemented and has been tested with the MHC PDB dataset. The environment was running under 2 Ghz PC with 512 MB of main memory with

Linux-2.6.11-1.1369. The implementation was written using Bash-3.00.16(1) and Octave-3.0.0.

#### Conclusion

In this paper, we proposed a measurement to determine graph stability for protein structure remodeling based on graph entropy. Our modified entropy validation shows a positive result for protein structural comparison. This graph-based approach offers a practical concept to support protein structural alignment.

#### Competing interests

The authors declare that they have no competing interests.

#### Acknowledgements

This work is supported in part by the National Science Council, Taiwan, under the Grant No. NSC 101-2221-E-259-004.

#### Declarations

Publication charges for this article have been funded by the authors. This article has been published as part of *BMC Bioinformatics* Volume 15 Supplement 15, 2014: Proceedings of the 2013 International Conference on Intelligent Computing (ICIC 2013). The full contents of the supplement are available online at <http://www.biomedcentral.com/bmcbioinformatics/supplements/15/S15>.

Published: 3 December 2014

#### References

1. Bunke H: **Graph Matching: Theoretical Foundations, Algorithms, and Applications.** *Proc Vision Interface 2000* 2000, 21.
2. Gilbert D, Westhead DR, Nagano N, Thornton JM: **Motif-based searching in TOPS protein topology databases.** *Bioinformatics* 1999, **15**(4):317-326.
3. Ehrig H, Engels G, Kreowski H: **Handbook of Graph Grammars and Computing by Graph Transformation: Applications, Languages and Tools.** World Scientific Publishing Company; 1997.
4. Vishveshwara S, Brinda K, Kannan N: **Protein Structure: Insights from Graph Theory.** *Journal of the Comp Chem* 2002, **1**:187-211.
5. Lund O, Hansen J, Brunak S, Bohr J: **Relationship between Protein Structure and Geometrical Constraints.** *Protein Science: a Publication of the Protein Society* 1996, **5**(11):2217-2225.

**Table 4 CATH codes for the selected macromolecules.**

PID	Domain	C	A	T	H	S	O	L	I	D
1HDM	A2	2	60	40	10	152	1	1	1	1
	B2	2	60	40	10	137	1	2	1	1
1K5N	A2	2	60	40	9	1	1	1	1	1
2ENG	A2	2	40	40	10	1	1	1	1	1
1VCA	A1	2	60	40	10	135	1	1	1	1
	A2	2	60	40	10	62	1	1	1	1
1ZXQ	A1	2	40	40	10	123	1	1	1	1
	A2	2	40	40	10	121	2	1	1	1
1UXW	A2	2	60	40	10	9	1	1	1	1
	B1	2	60	40	10	3	1	1	1	1
1A2Y	A1	2	60	40	10	8	2	2	1	1
	B1	2	60	40	10	36	2	1	1	1
3ARD	C1	2	60	40	10	18	2	3	2	1
	D1	2	60	40	10	21	4	12	1	1

6. Nelson DL, Cox MM: **Lehninger Principles of Biochemistry**. Freeman; 4 2004.
7. Shannon C: **Prediction and Entropy of Printed English**. *Bell Systems Technical Journal* 1951, **30**:50-64.
8. Chang R: **Physical Chemistry for the Biosciences**. University Science; 2005.
9. Simonyi G: **Graph Entropy: a Survey**. *Combinatorial Optimization* 1995, **20**:399-441.
10. Dehmer M, Emmert-Streib F: **Structural Information Content of Networks: Graph Entropy based on Local Vertex Functionals**. *Computational Biology and Chemistry* 2008, **32**(2):131-138.
11. Zager LA, Verghese GC: **Graph similarity scoring and matching**. *Applied Mathematics Letters* 2008, **21**:86-94.
12. Rand WM: **Objective Criteria for the Evaluation of Clustering Methods**. *J Amer Statistical Assoc* 1971, **66**(336):846-850.
13. Brouwer AE, Haemers WH: **The Gewirtz graph: an exercise in the theory of graph spectra**. *Eur J Comb* 1993, **14**:397-407.
14. Cvetkovic DM, Doob M, Sachs H, Cvetkovi&ccacute; M, Horstlistprice : **Spectra of Graphs: Theory and Applications**, 3rd Revised and Enlarged Edition. Vch Verlagsgesellschaft Mbh; 1998.
15. Hsu CH, Peng SL, Tsay YW: **An Improved Algorithm for Protein Structural Comparison based on Graph Theoretical Approach**. *Chiang Mai Journal of Science* 2011, **38**(2):71-81.
16. Goodsell DS: **The Machinery of Life**. Springer, 2nd ed. edition; 2009.
17. Pamer E, Cresswell P: **Mechanisms of MHC class I-restricted antigen processing**. *Annual Review of Immunology* 1998, **16**(10):323-358.
18. Berman HM, Westbrook J, Feng Z, et al: **The Protein Data Bank**. *Nucl Acids Res* 2000, **28**:235-242.
19. Peng SL, Tsay YW: **On the Usage of Graph Spectra in Protein Structural Similarity**. *Journal of Computers* 2012, **23**:95-102.
20. Huan J, Bandyopadhyay D, Wang W, Snoeyink J, Prins J, Tropsha A: **Comparing Graph Representations of Protein Structure for Mining Family-specific Residue-based Packing Motifs**. *J Computational Biology* 2005, **12**(6):657-671.
21. Peng SL, Tsay YW: **Measuring Protein Structural Similarity by Maximum Common Edge Subgraphs**. *Advanced Intelligent Computing Theories and Applications* 2010, 100-107, LNCS, 6216.
22. Canutescu A, Shelenkov A, Dunbrack R: **A Graph-theory Algorithm for Rapid Protein Side-chain Prediction**. *Protein Science* 2003, **12**(9):2001-2014.
23. Samudrala R, Moulton J: **A Graph-theoretic Algorithm for Comparative Modeling of Protein Structure**. *J Mol Biology* 1998, **279**:279-287.
24. Borgwardt K, Ong C, Schönauer S, Vishwanathan SN, Smola A, Kriegel HP: **Protein Function Prediction via Graph Kernels**. *Bioinformatics* 2005, **21**(suppl 1):i47-i56.

doi:10.1186/1471-2105-15-S15-S6

**Cite this article as:** Peng and Tsay: **Adjusting protein graphs based on graph entropy**. *BMC Bioinformatics* 2014 **15**(Suppl 15):S6.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

