Assessment of Protein-Graph Remodeling via Conformational Graph Entropy

Sheng-Lung Peng[*] and Yu-Wei Tsay

Department of Computer Science and Information Engineering
National Dong Hwa University, Hualien 974, Taiwan
slpeng@mail.ndhu.edu.tw

Abstract. 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 this extended graph entropy is applied, it helps a conformational on protein graph modeling. Besides, it also contributes to the protein structural comparison indirectly if a protein graph is solid.

Keywords: Protein structural similarity, Protein graph, Graph entropy.

1 Introduction

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 for representing of structured objects, then the problem of measuring object similarity turns into the problem of computing the similarity of graphs [1]. Protein remodeling is another field where multiple domains within structures are considerably complicated.

For decades, many studies have been devoted on defining topological relations and notation on protein structures. A schematic description is essentially expected to describe its topology. Mathematical formulation of structural patterns helps to facilitate the composition in a polypeptide chain. A schematic description has the advantage of simplicity, which makes it possible to implement in an alternative way as graph-theoretic approach [2]. By selectively neglecting protein structural features, it has a potential to detect further homologous relationships based on various geometric methods and motivations.

2 Preliminaries

The structure of a protein can be regarded as a conformation with various local elements (helix, sheet) and forces (Van der Waal's forces, hydrogen bonds), folding into its specific characteristic and functional structure. With the help of *graph transformation*, folded polypeptide chains are represented as a graph by several

[*] Corresponding author.

D.-S. Huang et al. (Eds.): ICIC 2013, CCIS 375, pp. 43–48, 2013.

mapping rules. Proteins contain complex relationships in its polymer; reaction of residues, interaction of covalent, bonding of peptides, packing of hydrophobic, are essential part in structure determination. The intention is to transform a protein structure into a graph.

2.1 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 [3]. In Table 1, we outline some categories of protein graph approach to a set of graphs, representing each specific graph rewriting and graph measuring skills. And also, it will be useful to begin with the summarized common research into the following matters:

Geometric Relation: It has been shown that the conformation of a protein structure is determined geometrically by using various constraints [4]. The most common method for protein modeling is to reserve its topological relation on graphs. From the aspect 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 polypeptide folding.

Chemical Relation: Comparing with geometric relationship, in chemical properties, it describes a more complicated description on protein graph model; owing to various chemical properties of amino acids, it includes electrostatic charge, hydrophobicity, bonding type, size and specific functional groups [5]. By giving values to edges and vertices in graph, each labeled component corresponds to a type of chemical relation.

Table 1. Recent studies for constructing protein graphs

Studies	Vertex set	Edge set
[6]	C_α atoms	labeled edges
[7]	DSSP[1]	attributed edges
[8]	side chains	defined by interacted energy
[9]	residues	defined by geometrical constraints
[10]	SSE[2]	labeled edges

[1] Dictionary of protein secondary structures. [2] Secondary structure elements.

2.2 Entropy

Entropy defines quantitatively an equilibria property within a system and it implies the principle of disorder from the second law of thermodynamics [11]. 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 [12]. Similarly, this concept is also included in life. As we known, life is composed by many cells, tissues, and organs from one of the vital element -- protein. Since proteins are biochemical compounds, consisting of one or more polypeptide chains, the arrangement of protein polymers is assumed to be in a compact state, according to its backbone dihedral angles and side chain rotamers. This is so called *conformational entropy*. In general, a protein graph model should also obey the second law of thermodynamics.

For an n-object system G, asume that each object i is associated a probability P_i. Then the entropy of the system G is defined as follows [13].

$$I(G) = -\sum_{i=1}^{n} P_i \log P_i \tag{1}$$

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

3 Our Method

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

Usually, the entropy of a graph is defined by its degree sequence. In this case, every regular graph with the same order has the same entropy. For example, both of C_4 and K_4, the complete graph with four vertices, have the same entropy, namely, 2. However, C_4 and K_4 are different. Thus, this definition is not enough to distinguish these two graphs. It motivates this research.

For a given graph G = (V, E) and two vertices u and v belonging to V, let $d(u,v)$ denote the length of the shortest path between u and v. Let $S_k(u) = \{v \mid d(u,v) = k\}$. In graph theory, $S_k(u)$ is called the *k-distance neighborhood* of u and is also called the *k-sphere* of u [14]. Let the function $f(u) = \sum(|S_i(u)|/(n-i+1))$ and $f(V)=\sum f(u)$. Assume that $V=\{v_1, v_2, ..., v_n\}$. We define Q_i for each v_i as follows.

$$Q_i = \frac{f(v_i)}{f(V) - S_1(v_i) + 1} \tag{2}$$

Note that in Formula 1, $\sum P_i$=1. However, in Formula 2, $\sum Q_i$ is not necessary equal to 1. Thus, we call $I(G)$ the *extended entropy* of graph G by replacing P_i with Q_i in Formula 1. By this extension, we obtain that $I(C_4)$=2.122 but $I(K_4)$=1.245.

4 Results

In this experiment, we validate the remodeling function of P-graph by using extended graph entropy to verify stability of a given P-graph. For the P-graph construction, please refer [7]. Thus, we only concern the connectivity impact to protein structural similarity. Various types of MHC are chosen as the material for the verification of proposed method, namely, **1HDM**, **1K5N**, **2CRY**, **1VCA**, **2Q3Z**, and **1ZXQ**. 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). Due to a great diversity of microbes in the environment, MHC genes widely vary their peptides through several mechanisms [15].

4.1 P-Graph Entropy Comparison

Let $G = (V, E)$ be the P-graph after remodeling from the construction proposed by [7], vertices of V in G are created according to the dictionary of protein secondary structures (DSSP). Under this metric, a protein secondary structure is represented by a single letter code, *e.g.*, H-helix (containing **G**, **H**, and **I**), T-hydrogen turn (containing **T**, **E**, and **B**), and C-coiled (containing only **C**). For controlling one variable in this experiment, let the edge set E in G be changed from a specific range.

A preliminary comparison of MHC proteins are shown in Table 2. In the table, **PID** is the protein identification number in PDB [16]. Since MHC proteins are composed by multiple polypeptide chains, there are multimeric **Domain**. Besides, the **Dens** means the density in the graph; it defines as $2|E|/(|V|(|V|-1))$ 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, if **AVG** = 10, then +20% increases the criteria length from 10 to 12, which increases the number of edges in E. When the number of edges in G is raised, certainly, the density is also increased.

Table 2. The selected proteins with corresponding extended entropies

PID	Domain	-40%	-20%	AVG	+20%	+40%
1HDM	B	NA	3.252	3.319	3.531	3.660
Dens		0.381	0.476	0.524	0.667	0.762
1K5N	A	4.029	4.456	4.777	5.001	5.243
Dens		0.345	0.491	0.582	0.636	0.782
2CRY	A	NA	NA	4.029	NA	NA
Dens		0.667	0.667	0.667	1.000	1.000
1VCA	A	NA	3.253	3.412	3.412	3.249
Dens		0.333	0.476	0.571	0.571	0.857
2Q3Z	A	5.418	5.904	6.610	7.565	9.715
Dens		0.221	0.308	0.413	0.551	0.750
1ZXQ	A	NA	3.480	3.630	3.866	4.176
Dens		0.321	0.429	0.500	0.607	0.786

It is interesting to observe the relation between $|E|$ and $I(G)$ in the following. First, when the density in G raises, it appears that the graph G goes from sparse to dense. However, its extended entropy does not totally decrease with its density. It seems a little anomalous in this appearance. Second, the edge set in protein remodeling issue can be possibly determined from its entropy. By the definition, the P-graph G should be a connected graph, *i.e.*, once the G becomes a spanning tree, the conformation can be decided from its entropy. For instance of **1VCA**, its P-graph is not a connected graph when density is 0.333 (-40%) but its entropy is 3.253 when the density rises to 0.476 (-20%). There is considerable validity to this concept though it should be verified by further proof and experiment. Third, it seems that E is 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 criteria. This is an essential fact to stress. It may be worth pointing out that the construction of P-graph is limited by V. Taking protein **1CXR** as an example, in PDB file, **1CXR** contains only one helix structure. Therefore, it would be unsuitable to transfer it into a one-vertex graph.

4.2 P-Graph Entropy Verification

For the purpose of validity according to the previous assumptions, a method for protein structural comparison is adapted to measure its similarity. *Graph spectra* gives an alternative solution for 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. Please refer [17] for the detail. In Table 3, we list the results of protein structure remodeling matters. The field **Old** shows a remodeling based on specific value of edge length, and **New** indicates the edges in *G* are determined by extended entropy. Values in each column display a local and global comparison of their graph spectra. As the structural alignment method, the smaller of the value, the more similar of their similarity. If our entropy suggests a better result in protein comparison, then we simply mark **Better** denoted as “+” ; otherwise, it is marked as “=” (unchanged) or “-” (worse). In summary, the extended entropy determines a better conformational graph from protein structure remodeling.

Table 3. A comparison of protein structure remodelings by proposed method

PID		1K5N	2CRY	1VCA	2Q3Z	1ZXQ
1HDM	**Old**	2.45|7.93	0.00|23.4	2.00|15.7	4.24|24.0	2.45|13.7
	New	1.73|7.75	0.00|21.1	2.23|13.9	2.83|23.7	2.24|12.1
	Better	+	+	+	+	+
1K5N	**Old**		1.00|26.6	2.23|19.6	3.32|26.9	2.65|18.0
	New		0.00|23.7	2.45|17.4	2.23|21.1	2.53|16.0
	Better		+	+	+	+
2CRY	**Old**			1.00|14.9	0.00|12.3	0.00|17.1
	New			1.00|12.9	0.00|34.1	0.00|14.9
	Better			+	=	+
1VCA	**Old**				1.41|17.7	1.00|5.39
	New				1.00|29.7	1.00|4.47
	Better				=	+
2Q3Z	**Old**					2.24|19.4
	New					1.41|28.6
	Better					=

4.3 Program and Environment

The environment is running under 2 Ghz PC with 512 MB of main memory with Linux-2.6.11-1.1369. The implementation is temporarily written using Bash-3.00.16(1) and Octave-3.0.0.

5 Conclusion

In this paper, we propose a benchmark to determine graph stability for protein structure remodeling based on graph entropy. With the help of this extended entropy validation, it concludes a conformational confirmation on protein structural comparison. This graph-based approach offers a practical concept to support protein structural alignment. In the future, a labeled protein remodeling is also expected to be verified by this extended entropy formula.

Acknowledgement. This work is supported in part by the National Science Council, Taiwan, under the Grant NSC 101-2221-E-259-004

References

1. Bunke, H.: Graph Matching: Theoretical Foundations, Algorithms, and Applications. In: Proc. Vision Interface 2000, pp. 82–88 (2000)
2. Gilbert, D., Westhead, D.R., Nagano, N., Thornton, J.M.: Motif-based Searching in TOPS Protein Topology Databases. Bioinformatics 15, 317–326 (1999)
3. Vishveshwara, S., Brinda, K.V., Kannan, N.: Protein Structure: Insights from Graph Theory. Journal of The Comp. Chem. 1, 187–211 (2002)
4. Lund, O., Hansen, J., Brunak, S., Bohr, J.: Relationship between Protein Structure and Geometrical Constraints. Protein Science: A Publication of the Protein Society 5, 2217–2225 (1996)
5. Nelson, D.L., Cox, M.M.: Lehninger Principles of Biochemistry, 4th edn. Freeman (2004)
6. 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. Comput. Biol. 12, 657–671 (2005)
7. Hsu, C.-H., Peng, S.-L., Tsay, Y.-W.: An Improved Algorithm for Protein Structural Comparison based on Graph Theoretical Approach. Chiang Mai Journal of Science 38, 71–81 (2011)
8. Canutescu, A.A., Shelenkov, A.A., Dunbrack, R.L.: A Graph-theory Algorithm for Rapid Protein Side-chain Prediction. Protein Sci. 12, 2001–2014 (2003)
9. Samudrala, R., Moult, J.: A Graph-theoretic Algorithm for Comparative Modeling of Protein Structure. J. Mol. Biol. 279, 279–287 (1998)
10. Borgwardt, K.M., Ong, C.S., Schonauer, S., Vishwanathan, S.V.N., Smola, A.J., Kriegel, H.-P.: Protein Function Prediction via Graph Kernels. Bioinformatics 21, i47–i56 (2005)
11. Shannon, C.E.: Prediction and Entropy of Printed English. Bell Systems Technical Journal 30, 50–64 (1951)
12. Chang, R.: Physical Chemistry for the Biosciences. University Science (2005)
13. Simonyi, G.: Graph Entropy: a Survey. Combinatorial Optimization 20, 399–441 (1995)
14. Dehmer, M., Emmert-Streib, F.: Structural Information Content of Networks: Graph Entropy based on Local Vertex Functionals. Computational Biology and Chemistry 32, 131–138 (2008)
15. Pamer, E., Cresswell, P.: Mechanisms of MHC Class I – Restricted Antigen Processing. Annual Review of Immunology 16, 323–358 (1998)
16. Berman, H.M., Westbrook, J., Feng, Z., et al.: The Protein Data Bank, Nucl. Nucl. Acids Res. 28, 235–242 (2000)
17. Peng, S.-L., Tsay, Y.-W.: On the Usage of Graph Spectra in Protein Structural Similarity. Journal of Computers 23, 95–102 (2012)