PROTEIN SIMILARITY/DISSIMILARITY USING CONTACT MAP AND MATCHING INDEX

R.MAGESWARI, K. SRINIVASA RAO, K. SIVAKUMAR

Abstract: Three dimensional structure of a protein determines its functional properties. Two proteins that are similar in three dimensional structures are likely to have similar functions. It is proposed to use graph theoretical concepts to analyze protein similarity using contact maps to predict the functional similarity using its structural similarity as a tool. The contact map of the protein is considered as a graph, where the carbon alpha atoms of all residues are taken as vertices and the edges of the graph depend on the distance between the carbon alpha atoms. In this paper, we find the matching indices based on the number of common neighbors shared by nodes. We observed high positive correlation between the matching indices of similar proteins 2WPY and 2WPZ and negative correlation between dissimilar proteins 2WPY and 1FT1.

Keywords: Contact map, Correlation coefficient, Matching index.

Introduction: There are twenty amino acids which make up the large biological molecule called protein. There are thousands of different proteins present in a cell and the nature of each type of protein is determined by a different gene. Proteins make up the cellular structure of organisms and proteins structure determines its biological function. In particular, structural similarity between proteins is a very good predictor of functional similarity. Understanding the structure of protein is significantly useful in understanding its function. Therefore, it necessitates analyzing the protein structure. A total of 1,00,000 protein structures were identified by the experimental methods such as X-ray crystallography, NMR and Electron Microscopy and are available in the Protein Data Bank (PDB). But, experimental methods are more expensive, time consuming and the process is also tedious. So, mathematical methods are introduced in the analysis of proteins.Similarity and dissimilarity of protein sequences is found using the Euclidean distance between the [1] spatial median of each protein. Structural similarity is predicted from the [2] structural descriptors derived from the geometric properties of secondary structure elements. Similarity is measured by the [3] compression ratio of the concatenated image using the image and audio compression – based approach.

Among the different mathematical methods for protein structure identification, graph theoretic methods are very easy to handle, less expensive and more effective. Contact map plays an important role in understanding proteins' structure. Using contact map of the protein, specific folds [4] in the protein are identified. Packing density & surface residues [5] are measured using the contact map of the protein and amino acids are sorted corresponding to the hydrophobicity properties.

Similarity is measured between protein structures from the matching pairs of secondary structure elements by the interaction of each pair between their axial line segments using a [6] fast bipartite graph matching algorithm. Similarity is measured by comparing the centroids of all secondary structures using [7] largest common sub graph detection algorithm.

Structural similarity of the two proteins is measured from the [8] maximum common edge sub graph of the labeled graph of the protein according to its secondary structures, chemical properties & topological relations. [9]Homology modeling based on clique finding, identification of side-chain clusters in protein structures upon graph spectrum are introduced to solve protein structure identification problems. [10] Content Based Image Retrieval (CBIR) and image registration techniques are used to measure the similarity between contact maps. Similarity of two protein structures is computed by measuring their [11] contact map overlap, which measures the similarity between two proteins (in lattice model) based on the pair wise distances of the C_α – atoms of each protein. Structural similarity between two proteins is measured by the [12] maximum clique within the graph using the maximum clique algorithm PLS (Phased Local Search).Similarity between two structures is measured using the[13] bipartite graph matching algorithm in which the reference frames extracted from protein tertiary structure are used to find the feature vectors for matching. Similarity is computed by [14] aligning the proteins to maximize the number of shared contacts in their corresponding contact maps.

In this paper, using graph theoretical concept, we extract the information of a protein by its contact map without having the 3D coordinates and measure the similarity & dissimilarity of the protein structure using matching index [15].

Preliminaries:

Definition: Let G = <V, E> be a graph in which V denotes the set of vertices, E denotes the set of edges and |V (G)| is the number of vertices of G. Two vertices u and v of G are called adjacent or neighbors if u-v is an edge of G. The degree of a vertex v is denoted by d(v) and is defined as the number of neighbors of v.

Contact Maps: The distance between two residues (d_{ij}) may be defined by the distance between two carbon alpha (C_α) atoms or between two carbon beta (C_β) atoms or it may be the minimum distance between any pair of atoms belonging to the side chain or to the backbone of two residues. A contact between two given atoms (or residues) exists when a certain distance is below a given threshold.

Let P be a protein with n atoms which are labeled

1,2,3,….n. We define the contact map of the protein as a matrix
$T = (t_{ij})_{1\le i,j\le n}$,
Where $t_{ij} = 1$ if $i \neq j$ & $d_{ij} \le 6$Å
= 0 otherwise.

A protein can be considered as a graph G = <V,E> for which each vertex $v_k \in V$ represents a residue of the protein and each $v_i - v_j \in E$ represents a contact between two residues v_i and v_j. On the other hand, there is an edge $v_i - v_j \in E$ if $t_{ij} = 1$.

The minimum Euclidean distance between two consecutive residues will be assumed to be 3Å and the maximum Euclidean distance between two consecutive residues is assumed to be 9Å. By changing D_{cutoff} threshold between 3 - 9Å for two C_α atoms, different contact maps can be viewed.

Matching Index:

The matching index M_{ij} measures the similarity of two nodes and is based on the number of common neighbors shared by nodes i and j. It is calculated as

$$M_{ij} = \frac{\sum common\ \_neighbirs}{\sum total\ \_number\ \_of\ \_neighbors}$$

Note that two vertices that are functionally similar need not always have to be connected. Also, complete graphs with same number of vertices are similar. Therefore, Choosing D_{cutoff} values so that to get a contact map, which is not a complete graph is important.

The Protein Data Bank (PDB) is a freely accessible database of 3D protein structures. Begun in 1971 with seven structures, it has now nearly 91,588 structures, with the yearly number of structures added to the database increasing each year.

It is observed that D_{cutoff} = 8Å yields the contact maps as a complete graph for the two similar proteins considered here. Hence, hereafter , we proceed with D_{cutoff} = 6Å. Due to computational reasons only five residues from the three proteins having PDB ID 2WPY, 2WPZ and 1FT1 are taken to construct the contact map for our similarity search. Protein sequences and tertiary structures (only C_α) of the three proteins with PDB ID 2WPY, 2WPZ and 1FT1 are extracted from the database PDB.

Figure 1: Protein 3D structure of (a) 2WPY, (b) 2WPZ and (c) 1FT1

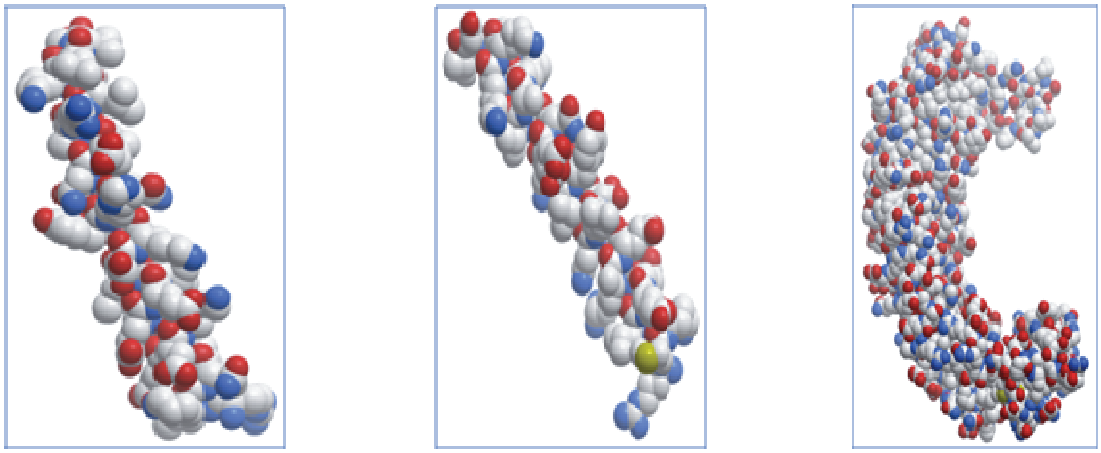


Figure 2: Ball and stick representation of carbon-alpha (C_α) of (a),(b) and (c)

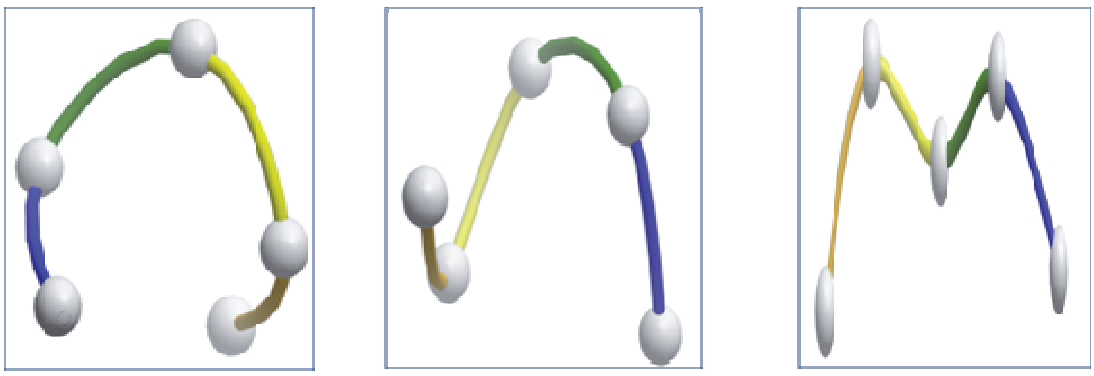
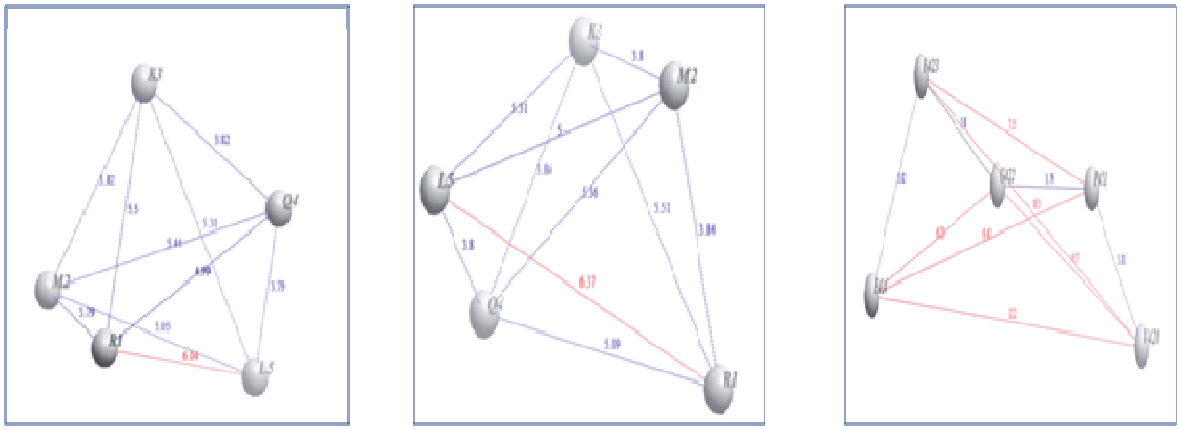


Figure 3: Distance Map (C_α atoms) of (a) , (b) and (c).



(a) (b) (c)

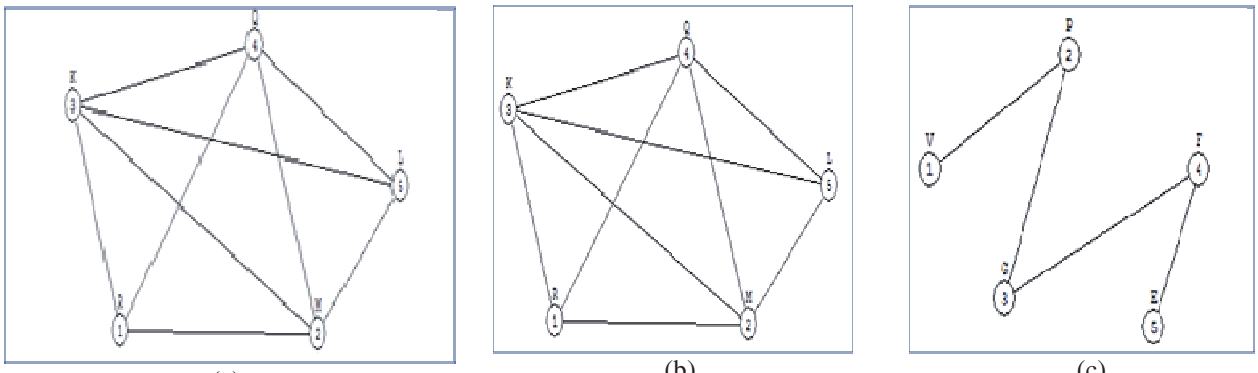
Contact Maps and Graphs:

The contact maps of (a),(b), and (c) are given below by using the definition stated above and the graphs are obtained from the contact maps as shown in figure 4.

$$\mathbf{T_a} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix} \qquad \mathbf{T_b} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix} \qquad \mathbf{T_c} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Contact map of (a) Contact map of (b) Contact map of (c)

Figure 4: Graph of the Contact Maps of (a), (b) and (c)



(a) (b) (c)

Consider the contact map of 2WPY. Here, vertex 1 is connected to the vertices 2,3 and 4. Vertex 2 is connected to the vertices 2, 3, 4 and 5. Number of common neighbors for the vertices 1 & 2 is 2. Total number of neighbors for the vertices 1 & 2 is 3.

Therefore, for the vertices 1 & 2, matching index is calculated as $M_{12} = 2/3 = .67$ and the remaining matching indices are calculated in the similar way and tabulated

below.

S.No	Matching index M_{ij}	2WPY	2WPZ	1FT1
1	M_{12}	0.67	0.67	0
2	M_{13}	0.67	0.67	0.5
3	M_{14}	0.67	0.67	0
4	M_{15}	1.0	1.0	0
5	M_{23}	1.0	1.0	0
6	M_{24}	1.0	1.0	0.5
7	M_{25}	0.67	0.67	0
8	M_{34}	1.0	1.0	0
9	M_{35}	0.67	0.67	0.5
10	M_{45}	0.67	0.67	0

Conclusion: We find the matching index of the vertices 1 & 5 for 2WPY and 2WPZ is 1. It shows even though they are not connected, they are functionally similar. The matching index of the vertices 1 and 2 for 1FT1 is 0. It shows even though they are connected by direct link, they are not functionally similar. We find the correlation between the matching indices of 2WPY and 2WPZ as 1 and it shows that the two proteins are exactly similar. The correlation between 2WPY and 1FT1 is -0.0890 and it shows the two proteins are dissimilar. We introduced a method to find the structural similarity/dissimilarity of proteins using its contact map and graph theoretical concepts.

References:

1. Abo-Elkhier, Mervat M, "Similarity /dissimilarity analysis of protein sequences using the spatial median as a descriptor" Journal of Biophysical Chemistry, 2012, Vol. 3, pp:142-148.
2. Pooja Jain and Jonathan D.Hirst, "Study of Protein Structural Descriptors Towards Similarity & Classification" , Computational Biophysics to Systems Biology, 2007, Vol. 36, pp:165-167.
3. Morihiro Hayashida and Tatsuya Akutsu, "Image compression – Based approach to measuring the similarity of protein structures", WSPC- Proceedings October 2007 pp:17.
4. Pankaj Barah & Som data Sinha, "Analysis of Protein folds using protein contact networks", Pramana Journal of physics, August 2008, Vol. 71, pp: 2.
5. Mahnaz Habibi, Changiz Eslahchi, Mehdi Sadeghi , Hamid Pezashk, "The interpretation of protein structures based on graph theory and contact map", Open Access Bioinformatics 2010, pp: 127-137.
6. William R.Taylor, "Protein structure comparison using Bipartite Graph Matching and its application to protein structure classification", *Molecular & Cellular Proteomics*, 2002, 1(4) pp. 334-339.
7. Stoicho Stoichev, Debrinka Petrova, "Protein Structure Models for Determining Protein Structure Similarity", Comp. Sys. Tech'06.
8. Cheng – Hsien Hsu, Sheng-Lung Peng and Yu-Wei Tsay, "An improved Algorithm for Protein Structural Comparison Based on Graph Theoretical Approach", Journal of Science, 2011, pp: 71-81.
9. Yan Yan, Shenggui Zhang, Fang-Xiang Wu, "Applications of graph theory in protein structure identification", Proteome Science 2011, 9.
10. Fernando Fernandes JR, Carlos Eduardo Lopes, Raquel Melo, Marcelo Santoro, Rodrigo Carceroni, Wagner Meira JR, Arnaldo Araujo, "An Image – Matching Approach to Protein Similarity Analysis", Computer Graphics and Image Processing, 2004. Proceedings pp: 17-24.
11. Pankaj k.Agarwal, Nabic H.Mustafa and Yusu Wang, "Fast Molecular Shape Matching Using Contact Maps", Journal of Computational Biology, Vol. 14, 2007, pp: 131-143.
12. Wayne Pullan, "Protein Structure Alignment Using Maximum Cliques and Local Search", M.A Orgun and J.Thornton(Eds): AI 2007, 4830, pp: 776-780.
13. Rosni Abdullah, Nu' Aihi Abdul Rashid & Fazilah Othman, "Graph Theory in Protein Sequence Clustering and Tertiary Structural Matching" , AIP Conference Proceedings, Vol. 971, pp:19.
14. B.Carr, W.Hart, E.Burke J.Smith, "Alignment of protein Structures with a Memetic Evolutionary Algorithm", Proceedings of the Genetic and Evolutionary Computation Conference, 2002.
15. Pavlopoulos et al ,"Using Graph theory to analyze biological networks" , Bio Data mining, 2011, Vol.4, pp: 10.

***

R.Mageswari/Assistant Professor/ Department of Mathematics/
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya (SCSVMV University)/
Kanchipuram - 631 561/Tamilnadu, INDIA/mageswari78@gmail.com
K. Srinivasa Rao/Associate Professor/ Department of Mathematics/
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya (SCSVMV University)/
Kanchipuram - 631 561, Tamilnadu, INDIA/raokonda@yahoo.com
K. Sivakumar /Associate Professor/ Department of Chemistry/
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University (SCSVMV University)/
Kanchipuram - 631 561/ Tamilnadu/ INDIA/shivamk25@yahoo.co.in