Structural Alignment Using Graphs

N.VEDAVATHI [1], DHARMAIAH GURRAM[1]
1.Asst.Professor in Mathematics, K L University, Guntur (dist) A.P,India-522502

ABSTRACT

Protein structure or sequence alignment methods are widely used to discover similar regions between proteins and to assess the similarity by a score. Especially structural alignment methods, which are capable of capturing structural thus functional homologies, are useful tools for protein fold classification, protein structure modeling and structure based annotation. With rapidly growing experimental structure information , the need for fast and accurate structural alignment algorithms is apparent. In this paper we showed that graph theoretical properties such as connectivity, clustering coefficient, second connectivity, characteristic path length and centrality measures can be used effectively as the scoring function in structural alignment of proteins.

Index Terms*—Contact maps, graph theoretical properties, structural alignment.*

I. INTRODUCTION

Structure alignments of proteins may provide information about structural similarity of functional units (domains) and overall similarity of two known structures for classification and annotation purposes. Several structural properties of the proteins are used to obtain the optimum alignment of structures. In this work, we represent the protein structure as a graph and network properties of the graph are shown to represent similar regions between two distinct protein structures. We claim that network properties of the graphs can be used as a target function to find similarities between proteins. Each protein can be represented as graphs and then the structure alignment problem can be converted into inexact sub-graph matching problem where so many heuristic algorithms are already developed. In this paper, we used nine different graph theoretical properties and showed their applicability for structural alignment on two different data sets.

II. BACKGROUND AND RELATED WORK

Structural alignment methods try to obtain the optimum overlay of proteins based on their three dimensional coordinates. The resulting alignment is a superposition of amino acids where structurally similar regions are aligned with each other. The goodness of the fit is measured by the root mean square distance (RMSD) which calculates the mean distance between Cα atoms of corresponding amino acids [1]. There are different approaches for solving the structure alignment problem which can be classified into two categories, superposition and clustering methods [2]. Superposition methods translate and rotate one protein in three dimensional spaces to minimize the protein’s intermolecular distance to other protein. Clustering methods establish the amino acid clusters and compare the intra molecular amino acid to amino acid distances of one protein to another.

CE (Combinatorial Extension) is a widely used structure alignment method based on clusters of amino acids that uses inter residue distances [3]. Protein sequence is broken into and represented by a set of aligned fragment pairs (AFP). AFPs are of fixed size, it’s reported that 8 is the optimum size in terms of speed and accuracy. The alignment of two proteins A and B is defined as a path of AFPs in a similarity matrix S of size (nA-m) * (nB-m) where m is the AFP size and nA and nB are the lengths of proteins. An alignment may start from any AFP and after that consecutive AFPs are added in such an order that the next added AFP cannot contain any residue that was included in the previous AFP. Gaps are allowed but there is an upper limit to the length of a gap segment to reduce running times, the limit is 30. In the process of addition of new AFPs, not all the possibilities are explored; several heuristics are employed to reduce the search space.

CE uses three distance measures to evaluate similarity and AFP path extension alternatives. The first measure is the average of the sum of distances between residues of two different AFPs where each residue participates once. First measure is used to decide how well two AFPs combine, it is the path extension heuristic. The second measure is similar to the first one but all possible distances between non-neighbor residues are averaged for two different AFPs. Second measure evaluates the goodness of a single AFP, whether two protein fragments match well. The third measure is the root mean square distance calculated from superimposed structures and is used in the final steps to pick best alignments and optimization.

III. METHODOLOGY

Contact maps are widely used to represent the 3-Dimensional protein structures [4, 5]. A contact map shows which amino acids are in close vicinity of each other when the protein folds into its functional form. Contact maps can be represented as graphs where the residues correspond to the nodes and the contacts correspond to the links. There are many definitions of a contact in the literature. In this work, we used the definition given by Atilgan et. al [6]. If the distance between C_α atoms for the residues *i* and *j* is smaller than 6.8 A^o, then these residues are considered to be in contact [4, 5].

There are many network properties that can be used for graph operations. The first network property we used is the degree or the connectivity k which measures the number of neighbors of each residue in the protein. [7] The connectivity of a graph is a measure that shows its robustness as a network. The distribution of the degree frequency has a normal distribution in a protein and shows scale free property [8].

We have developed a new property to measure the compactness of the graph which we called as second connectivity (S(k)). If the structure is made up of small compact domains rather than one globular structure, it would have low second connectivity numbers. So this value can be used to determine the similar parts of the proteins that have such structural features. The second connectivity of a node is calculated by the sum of the contacts of all its neighbors. The third network property is the clustering coefficient so-called cliquishness which measures how well the neighbors are connected to each other. The clustering coefficient for each node is calculated as in (1);

$$C_n = \frac{2E_n}{k(k-1)} \qquad (1)$$

where E_n is the actual edges of the residue *n* and *k* is the degree. [7, 9]

In addition to these properties we used characteristic path length as a network property which was also used by Taylor *et. al.*[7] and Sinha *et. al.*[8]. Characteristic path length (L) is smaller in globular proteins and larger in fibrous proteins because of the variations in the shortest paths in the protein structures. Moreover, characteristic path length L_i for each residue is calculated by the average of the shortest paths from the residue *i* to all the other residues given as in (2);

$$L_i = \frac{1}{(N-1)} \sum_{j=1}^{N} \sigma_{ij} \qquad (2)$$

where σ_{ij} is the shortest path length between nodes *i* and *j* and *N* is the number of residues of a protein.[7]

Graph properties can only capture overall structural properties of the proteins but do not measure physiochemical interactions between the atoms that are in contact in the folded form. Therefore we employed weighted characteristic path lengths (wL) which have weights as contact potentials beside neighboring information. Contact potentials are statistical potentials that are calculated from experimentally known 3D structures of proteins which calculate frequencies of occurrences of all possible contacts and convert them into energy values so that frequently occurring contacts over random values would have positive contact scores. We used the contact potential matrix from Dill *et. al.* [10].

Several measures are used to discover the centrality of a node in a graph. Betweenness (Freeman [11]), Clossness (Sabidussi [12]), graph (Hage and Harary [13]), and stress (Shimbel [14]) centrality measures are the best known measures in the literature and their formulas are given respectively in equations (3), (4), (5), (6) [15, 16]. If the centrality measure of a node is high, this node has a central role (they are part of most frequently used pathways) in the structure of a protein and can be crucial in its folding [7].

$$C_B(i) = \sum_{s \neq i \neq t \in V}^{N} \frac{\sigma_{st}(i)}{\sigma_{st}} \qquad (3)$$

$$C_C(i) = \frac{1}{\sum_{t \in V} d_G(i,t)} \qquad (4)$$

$$C_G(i) = \frac{1}{\max_{t \in V} d_G(i,t)} \qquad (5)$$

$$C_S(i) = \sum_{s \neq i \neq t \in V}^{N} \sigma_{st}(i) \qquad (6)$$

where d_G is the degree matrix, σ_{st} is the shortest path matrix and $\sigma_{st}(i)$ is the matrix for the number of the paths between the nodes *s* and *t* pass through node *i*. In this paper to verify the applicability of the network properties in structural alignment problem, we calculated the difference between the network property values of the CE aligned residues of two protein structures then we checked to see whether such a difference value could be obtained randomly to show the statistical significance of the results.

First, a pair of structurally similar proteins is aligned with CE alignment algorithm [3]. For each protein, the values of different network properties are calculated. Then the Euclidean distances between network attributes of the aligned residues in the CE alignment are found. If the distance is close to zero, it would mean that the network values of the aligned regions are very similar. This would prove our claim that these properties can be used as a target function in structure alignment problem. We used two methods to check whether such a difference between network attributes could be obtained randomly.

In the first method; we kept the network values of the first protein the same and randomly shuffled the existing network values in the second protein. This way we make sure distribution of network values is kept the same but these values are assigned to different residues. Then we calculated the distance between aligned residues arising from random assignment of network attributes. This method is called as “shuffled method”. This procedure is repeated 1000 times. In the second method, we basically shifted the network values of the second protein randomly while keeping the values of the first protein then calculated the distances of CE aligned residues. This procedure is called “shifted method” and also repeated 1000 times. The reason for the second method lies in the fact that the network values may not be independent of each other and these values may be correlated for the neighboring residues. Random shuffling method would not capture the effect of such correlations. That is why we shifted the values randomly, thus keeping the local ordering of the values the same but these values would be assigned to different neighboring amino acids. Mean and standard deviations of the distances of CE aligned residues of each network value are calculated for 1000 random runs and these values are compared to actual distance values calculated based on CE alignments via their Z scores as in (7).

$$z = \frac{x-\mu}{\sigma} \qquad (7)$$

where *x,* is the “real distance” from the values in the order of CE alignment, μ is the averages and σ is the standard deviations.

IV. DATASETS

We used two different datasets. First data set was created by Capriotti *et. al.*[5]. There are 158 protein pairs in this dataset that are structurally similar. However, their sequence identity is less then 30% and the average sequence similarity is about 16%. Therefore, this dataset is being considered as a difficult set to find alignments between pairs using common alignment techniques. The second dataset is chosen from Astral 40 database [18]. In this data set, the sequence identities of each pair are less then 40% and their average is 17.8%. This dataset, moreover, is created based on SCOP[18] classification. Therefore, the protein pairs are built within the same sub-family in the dataset and this dataset consists of 3064 pairs.

V. RESULTS

The CE alignment of two proteins (12AS and 1PYS) from Caprioti data set is given in Figure 1. The network values for both proteins corresponding to part of the aligned regions are summarized in Table 1. These calculations are done for each pair in both data sets. The results for Caprioti data set were given in Table 2 and 3. The results for Astral 40 data set were given in Table 4 and 5.

The rows in the tables show the graph theoretical properties. *k*, *C*, and *S(k)* are degree, clustering coefficient and second connectivity respectively. *L* and *wL* are characteristic path length and its weighted form. *Cb*, *Cc*, *Cg* and *Cs* are the centrality measures which are betwenness, clossness, graph, and stress centrality respectively. X, µ and Z are the average scores of the randomly generated pairings and actual CE alignment distances. # shows how many pairs have higher z scores than 1.96 (the Z value used for to 95% significance testing). % shows percentage of the pairs in the data set that has significantly lower distances in the CE alignment than the randomly generated networks values.

In both data sets degree and the second connectivity had the lowest distances for CE alignments. Centrality measures were not as important as these attributes. Betweeness centrality measure was the most distinguishing centrality measure for both data sets. As expected, shifted method yielded lower percentages for both data sets than the shuffled method.

```
Structure  Alignment  Calculator,  version  1.02,
last modified: Jun 15, 2001.

CE Algorithm, version 1.00, 1998.

Chain 1: pdbdir/12AS.pdb:A (Size=330)
Chain 2: pdbdir/1PYS.pdb:A (Size=350)
Alignment length = 211 Rmsd = 3.45A Z-Score =
5.3  Gaps  =  125(59.2%)  CPU  =  15s  Sequence
identities = 14.2%

Chain 1:  9 QRQISFVKSHFSRQLEERLGLIEVQAPILSR
Chain 2:100 LHPITLMERELVEIFRAL-GYQAVEGPEVES
```

Fig. 1. A part of an example of the CE Alignment result between the chain A of 12AS and the chain A of 1PYS. Calculated values for each graph theoretical property for the bold part is in Table 1 as an example.

TABLE 1

CALCULATED NETWORK VALUES FOR BOTH PROTEINS

Res Num	21	22	23	24	25	26
12AS	**R**	**Q**	**L**	**E**	**E**	**R**
Res Num	112	113	114	115	116	117
1PYS	**E**	**I**	**F**	**R**	**A**	**L**
1st k	8	9	12	10	7	8
2nd k	8	10	9	9	7	6
1stcliq	0,64	0,58	0,44	0,53	0,76	0,61
2ndcliq	0,64	0,42	0,61	0,58	0,76	0,87
1st ss	H	H	H	H	H	H
2nd ss	H	H	H	H	H	H
1st sk	74	85	108	86	63	76
2nd sk	68	81	74	74	59	52
1st L	5,67	5,48	5,04	5,36	5,75	5,37
2nd L	5,41	5,16	5,17	5,21	5,31	5,32
1st wL	6,57	6,63	5,15	6,50	6,85	6,69
2nd wL	6,80	5,73	5,82	6,73	6,33	6,04
1st Cb	882,44	923,16	3633,08	1402,67	713,15	1180,16
2nd Cb	748,84	4088,64	994,19	941,19	676,65	618,22
1st Cc	0,0005	0,0006	0,0006	0,0006	0,0005	0,0006
2nd Cc	0,0007	0,0007	0,0007	0,0007	0,0007	0,0007
1st Cg	0,1111	0,1111	0,1111	0,1111	0,1111	0,1111
2nd Cg	0,1000	0,1000	0,0909	0,0909	0,0909	0,0909
1st Cs	4995,44	5483,92	9702,06	6124,84	1321,29	4057,23
2nd Cs	2196,42	9416,08	4633,18	5952,57	3238,14	2038,70

TABLE 2

THE RESULTS FROM RANDOMLY SHUFFLED METHOD (CAPRIOTTI DATASET)

	x	**μ**	**Z**	**#**	**%**
k	22,91	34,90	7,85	142	89,87
C	1,39	1,89	5,85	129	81,65
S(k)	271,89	439,56	9,17	142	89,87
L	13338,58	17855,23	6,24	132	83,54
wL	8,08	12,46	12,24	138	87,34
Cb	12,75	17,97	9,46	137	86,71
Cc	0,0082	0,0091	8,6922	137	86,71
Cg	0,3234	0,3849	6,8792	117	74,05
Cs	296164,26	334466,22	5,34	109	68,99

TABLE 3
THE RESULTS FROM SHIFTED METHOD (CAPRIOTTI DATASET)

	x	μ	Z	#	%
k	22,91	34,60	4,20	131	82,9
C	1,39	1,88	4,13	124	78,5
S(k)	271,89	435,11	3,88	129	81,6
L	13338,58	17798,15	4,67	121	76,6
wL	8,08	12,31	3,53	122	77,2
Cb	12,75	17,81	3,62	125	79,1
Cc	0,0082	0,0090	3,0510	115	72,8
Cg	0,3234	0,3826	2,3328	84	53,2
Cs	296164,26	333401,59	2,54	92	58,2

TABLE 4
THE RESULTS FROM RANDOMLY SHUFFLED METHOD (ASTRAL 40 DATASET)

	x	μ	Z	#	%
k	19,55	29,50	6,75	2708	88,38
C	1,22	1,67	5,29	2479	80,91
S(k)	223,35	349,74	7,36	2759	90,05
L	25477,08	30430,77	4,76	2083	67,98
wL	11,30	15,05	8,07	2498	81,53
Cb	15,72	19,89	6,80	2600	84,86
Cc	0,0077	0,0082	7,4336	2398	78,26
Cg	0,2877	0,3401	5,7695	2103	68,64
Cs	2949407,03	3035718,00	3,13	1796	58,62

TABLE 5
THE RESULTS FROM SHIFTED METHOD (ASTRAL 40 DATASET)

	x	μ	Z	#	%
k	19,55	29,22	3,64	2478	80,87
C	1,22	1,66	3,58	2331	76,08
S(k)	223,35	345,71	3,22	2379	77,64
L	25477,08	30362,23	2,71	1813	59,17
wL	11,30	14,90	2,33	1859	60,67
Cb	15,72	19,74	2,60	2117	69,09
Cc	0,0077	0,0082	2,1432	1741	56,82
Cg	0,2877	0,3378	1,5773	1346	43,93
Cs	2949407,03	3035201,37	1,96	1486	48,50

CONCLUSION

In this work we showed that network property values rather than actual distances as used in existing methods can be used for structural alignment. CE aligned pairs had very similar network attributes and this similarity was significant at 95 % significance level.

The results indicate the most similar network attribute between aligned pairs was the second connectivity which shows more similarity than connectivity itself; therefore using second connectivity will lead us to better alignments. Betweenness as a centrality measure has more information content than the other centrality measures. Using weights from statistical potentials in calculating characteristic path length improved the results. Here we showed that most of the network values are significantly similar for CE aligned regions of two proteins. Next, we can develop a function that would give high similarity score if the network values are close. One can use an optimization algorithm such as dynamic programming to find the highest alignment score corresponding to the optimum structural alignment.

This work is a first attempt to use a different function rather than actual atomic distances in structural alignment problem. Other methods are highly dependent on actual coordinates of the proteins accuracy of which may change with the experimental procedure that is used to obtain the structure. Our function is less dependent on actual coordinates and therefore more robust than existing methods.

ACKNOWLEDGMENT

The authors gratefully acknowledge, Prof. Rita Casadio and Prof. Anna Tramontano for helpful discussions and comments made during the course of our work.

REFERENCES

[1] Kabsch W. (1978) A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. Acta Crystallogr. A, 34, 827–828.

[2] Eidhammer I., Jonassen I., Taylor W.R. (2000) Structure Comparison and Structure Patterns. J Comp. Bio., 7, 685–716.

[3] Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 11(9) 739-747.

[4] Vendruscolo. M.. E. Kussel. and E. Domany: Recovery of Protein Structure from Contact Maps. *Structure Fold. Des.* 2 (1997) 295-306.

[5] Fariselli. P. and R. Casadio: A Neural Network Based predictor of Residue Contacts in Proteins. *Protein Eng.* 9 (1996) 941-948.

[6] A. R. Atilgan. P. Akan. C. Baysal: Small-World Communication of Residues and Significance for Protein Dynamics. *Biophys. J.* 86 (2004) 85-91

[7] Taylor T.. Vaisman I.I.: Graph theoretic properties of networks formed by the Delaunay tessellation of protein structures. *Phys. Rev. E. Stat. Nonlin. Soft. Matter Phys.* 73 (2006) 041925

[8] Ganesh Bagler and Somdatta Sinha Network properties of protein structures, Statistical Mechanics and its Applications,
Physica A, Volume 346, Issues 1-2, 1 February 2005, 27-33

[9] Vendruscolo. M.. N. V. Dokholyan. E. Paci. and M. Karplus: Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev.* **65** (2002) 061910

[10] Liang. J. and K.A. Dill: Are proteins Well-Packed? *Biophys. J.* 81 (2001) 751-766

[11] Freeman L. C. A set of measures of centrality based on betweenness. Sociometry, (1977), 40:35-41.

[12] Sabidussi, G. The Centrality index of graph, Physicometrica, (1966) 31:581-603.

[13] Hage P. and Harray F. Eccentricity and centrality in networks, Social Networks, (1995), 17:57-63.

[14] Shimbel A., Structural Parameters of communication networks, Bulletin of Mathematical Biophysics, (1953), 15:501-507.

[15] Ulrik Brandes (2001) A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology 25*(2):163-177. 2001.

[16] A Measure of Betweenness Centrality Based on Random Walks, M. E. J. Newman, DOI: cond-mat/0309045, arXiv, 2003-09-01.

[17] Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL compendium in 2004. *Nucleic Acids Research 32:*D189-D192 (2004).

[18] Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.