Graph Theory In Protein Sequence Clustering And Tertiary Structural Matching

Rosni Abdullah, Nur'Aini Abdul Rashid and Fazilah Othman

School of Computer Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia

Abstract. The principle of graph theory which has been widely used in computer networks is now being adopted for work in protein clustering, protein structural matching, and protein folding and modeling. In this work, we present two case studies on the use of graph theory for protein clustering and tertiary structural matching. In protein clustering, we extended a clustering algorithm based on a maximal clique while in the protein tertiary structural matching we explored the bipartite graph matching algorithm. The results obtained in both the case studies will be presented.

Keywords: Graph Theory, Protein Clustering, Structure Matching, Bipartite Algorithm.
PACS: 07.05.Kf and 07.05.Rm.

INTRODUCTION

Graph theory is a domain in mathematics about the study of graph representation and manipulation of certain objects like a structure or string sequence. Graph can be referred to as a group of vertices (or nodes) and edges (or links). The edge carries a relationship between two adjacent vertices. The principle of graph theory which has been widely used in computer networks is now being adopted for work in bioinformatics, such as protein clustering, protein structural matching, biological data analysis and protein folding and modeling. In this paper, we present two case studies on the use of graph theory in graph based clustering and graph based structural matching.

RELATED WORK

In this section we present the related work for graph based clustering and graph based structural matching.

Graph Based Clustering

Graphs can be used to represent relations between protein sequences. The label of the node in the label weighted graph is the protein sequence tag and the weighted edge is the similarity values between two protein sequences. The Matsuda, Ishihara, and Hashimoto algorithm [7] introduced new graph structure called p-quasi complete

graph for describing a family of sequence with a confidence measure and used graph structure to generate a set of sub-graphs from the sequence. Their proposed method classifies the whole genome using similarity based algorithm. They then extended the single linkage clustering method for classifying the whole genome. Single linkage clustering method in graph is finding the maximal connected sub-graph.

ProtoMap [8] is a fully automated hierarchical clustering of protein sequences. Their work produced well defined groups of strongly connected protein families and sub-families. Kawaji et al. [9] uses graph partitioning method to cluster the protein sequences in related groups. They focus on approximate distantly related proteins without overlapping groups by iterative partitioning. This solution resolved the problem of high computational cost in method by [7]. The algorithm uses the normalized cut algorithm and Kernighan and Lin [10] heuristic to partition the graph. ProClust is an extended version of the algorithm proposed by Bolten et al. [11]. [11] developed a graph based protein clustering using transitive homology. The similarity values between two sequences are calculated using Smith-Waterman algorithm.

Sun Kim [12] proposed a graph clustering algorithm using two graph properties: bi-connected component and articulation point. Bi-connected components represent the protein family and articulation points represent the multi-domain protein. Cluster-C [13] uses proximity matrix values to build graph. The calculation of proximity matrix is the product of running the Smith-Waterman algorithm on all versus all protein sequence comparison.

Graph Based Structural Matching

Willet [1] attempted to extend his works in cheminformatics technique to bioinformatics. He claimed that both fields have large volumes of molecular data which require heavy computational processes. Willet has identified protein secondary structure elements (SSEs) as a linear structure represented by graph. The node in the graph will be the SSEs (α-helix or β-strand) and the edge will be the geometric relationships between pairs of SSEs. Each edge in the graph carry three-part of attributes containing (a) the angle between a pair of vectors that describe the SSEs; (b) the distance of closest approach between two vectors; and (c) the distance between midpoints.

Bruno [2] has represented carbohydrate structures in Complex Carbohydrate Structure Database (CCSD) by labeled graphs and the matching is performed using subgraph isomorphism algorithm. The vertices and edges of the labeled graph denote the residues and inter-residue linkages of the carbohydrate structure, respectively.

Adrian 2003 [3] has applied graph theory to SCWRL, an existing program for side-chain conformation prediction to solve the combinatorial problem. Side chain is represented as vertices in an undirected graph. The edge between residues only appears if it has rotamers with nonzero interaction energies. Once all the side chains have been represented in graphs, those without edges between each other will be partitioned into connected subgraphs and later divided into biconnected components. The combinatorial problem is reduced by finding the minimum energy of these small biconnected components. The results are combined to identify the global minimum energy conformation.

Another work focused on graph representation is by Huan 2004 [5]. They have created three versions of graph representation on protein structure and run each version in a frequent subgraph mining algorithm. From these representations, they want to identify which version can carry out the most discriminatory subgraph signatures of the protein structure. Donald [6] represents a protein structure as constraint network where the vertices are the protein atoms, the edges are the distance constraint between atoms and every angle between edges is stored.

CASE STUDY 1: GRAPH BASED PROTEIN SEQUENCE CLUSTERING ALGORITHM

Brief Introduction And Background

Graph-Based methods transform the clustering problem into graph partitioning problem using graph algorithm and heuristic methods. Normally undirected graph is used assuming the symmetric relations between objects. The choice of similarity measure depends on the problem domain. Finding a cluster in a set of protein sequences is analogous to finding cliques in a graph which involved partitioning the graph. Relations between protein sequences can be presented using a graph. Vertices in the graph are the protein sequences while edges represent a relation between two vertices. A clique is a subset of vertices, where all vertices in the subset are directly connected to each other.

Methods

The general overview of the graph based clustering algorithm follows closely to the one given by [12]. The general procedure for the clustering algorithm is divided into four main phases. They are as follows:

1. Pre-processing

Compute the similarity or distance values between two proteins. For a database with K sequences, the complexity of this phase is $O\left(\frac{K(K-1)}{2}\right)$. [12], [13] and [14] use Smith Waterman algorithm to calculate the similarity which result in $O\left(m \times n \times \frac{K(K-1)}{2}\right)$ complexity where K is the number of protein sequences in database and m and n are length of proteins. In our proposed algorithm we use the N-Gram Hirschberg (NGH) algorithm [15]. The complexity of our algorithm is $O\left(\frac{mn}{n} \times \frac{K(K-1)}{2}\right)$ where N is the size of the N-Gram.

2. Build Graph G_t

Build directed Graph G_t based on the similarity values calculated in step 1. The threshold value is determined by user. Threshold values are set by biologists to filter out "false positive", a condition where a protein sequence is incorrectly clustered into

the wrong group. A strict threshold will result in too many true sequences being discarded and a very low threshold will results in bringing in the "non-member" into the clusters. This is a very subjective issue and is to be resolved by the biologists.

3. Find all cliques in the graph C_G. This algorithm is the extension of large scale clustering algorithm based on extraction of maximal clique [13].
4. Post-Processing: Eliminate sub-clusters and Merge clusters.

Two families are merged into one bigger family if they share more than 80% similar sequences. Protein sequences that exist in too many families are deleted.

Results And Analysis

TABLE 1. Comparison of Clustering Results Among Different Algorithm.

Dataset / Algorithm	Performance Metrics	COG	PFAM1	NCBI	Swiss-Prot
NGHGC	JC	0.77	0.98	0.87	0.86
	RS	0.88	0.99	0.87	0.79
	P	0.84	0.99	0.92	0.87
	R	0.89	0.99	0.94	0.97
HGC	JC	0.91	0.98	0.97	0.98
	RS	0.91	0.98	0.97	0.98
	P	0.96	1.0	1.0	0.99
	R	0.94	0.98	0.97	0.99
BAG	JC	*n.a*	0.27	0.60	0.50
	RS	*n.a*	*n.a*	*n.a*	*n.a*
	P	*n.a*	1.0	1.0	1.0
	R	*n.a*	0.2	0.60	0.50
SEQOPTIC	JC	*n.a*	0.85	0.66	0.81
	RS	*n.a*	*n.a*	*n.a*	*n.a*
	P	*n.a*	0.98	0.82	0.99
	R	*n.a*	0.87	0.78	0.22
Blasclust	JC	*n.a*	0.04	0.11	0.06
	RS	*n.a*	*n.a*	*n.a*	*n.a*
	P	*n.a*	1.0	1.0	1.0
	R	*n.a*	0.04	0.11	0.06

Table 1 gives the experimental results of running of HGC and N-Gram Hirschberg Clustering (NGHGC) for all datasets. The results for BAG, SEQOPTICS and *blastclust* are taken from [14]. The values are the quality of clusters produced by each algorithm relative to the "true" clusters. Value 0 implies that the "derived" cluster is totally different from the "true" clusters while value 1 implies that the "derived" clusters are the same as the "true" clusters. Four performance metrics were used, namely Jaccard (JC), random statistics (RS), precision (P) and recall (R). The quality of clusters produced using the COG data set by the NGHGC algorithm is less than those by HGC measured by all the four metrics. However NGHGC algorithm is much faster when producing the distance matrix based on the experiments and results of [15]. This is the main advantage of NGHGC over HGC and this will become more obvious when dealing with bigger datasets.

CASE STUDY 2: GRAPH THEORY FOR PROTEIN TERTIARY STRUCTURAL MATCHING

Brief Introduction And Background

In this case study, graph theory is applied to both data representation and matching algorithm. For data representation, a reference frame is used to calculate a feature vector for matching. To find similarity between two structures, bipartite graph matching algorithm is applied. The reference frames extracted from protein tertiary structure are used to acquire the feature vectors for matching by using the bipartite graph matching algorithm. The reference frame is generated from backbone fragment specifically the sequential order of atom N-Cα-C. Each structure can be described as a graph where the atom and the connection between the others atoms are visualized as the nodes and the edges, respectively. Each node is labeled by a set of coordinates (x, y, z). Details of reference frame for this work is covered in [16].

Methodology

The matching is based on the backbone fragment of each structure. The backbone fragments (chain of 3 atoms, N-Cα-C) are extracted to form a reference frame. To find maximum matching, a maximum flow method and breadth-first search is applied to the problem. The bipartite graph consists of two partitions, A and B. Partition A is exclusively to represent structure A (query), and Partition B for structure B (target). Each vertex in Partition A represents a set of new coordinates calculated based on reference frames identified in the query. If the query structure has N identified reference frames, Partition A will have N vertices. The same applies to Partition B with the target. For initial matching, edges are created from each vertex from Partition A, to all vertices in Partition B. Here, N x N edges are formed. The edges' weight is the similarity between two sets of coordinates (each from partition A and B). From the constructed graph, maximum flow technique is applied to get as many weighted edges connecting vertices from both partitions. A maximum flow value is produced at the end. The query will be tested with target structures from the database. The target with highest maximum flow value is considered similar to the query.

Results

To benchmark the matching result, we refer to the dataset and result from work by Carlo et. al. [17] on comparison of protein secondary structures based on indexing technique. They compared the target structure 110M against 21,000 proteins in the database, and came out with top 25 hits. In our experiment, we downloaded first 10 structures from the top 25 hits (112M, 109M, 1MCY, 1JDO, 111M, 103M, 1MNO, 108M, 106M, 1MWC) and randomly add another 10 structures (113L, 112L, 111L, 110L, 109L, 108L, 107L, 104L, 103L, 102L) as input to our program, with the same target protein 110M. The expected matching result is 10 highest matching values

should go to the first 10 structures taken from the benchmark result in [17]. Our result shows that from the top ten hits, eight structures are correctly classified (Table 2).

TABLE 2. Top 10 Hits for Matching Result.

Protein	Matching Value	Label
109M	144.641465	true positive, TP
106M	140.205632	true positive, TP
111M	140.204562	true positive, TP
103M	139.608685	true positive, TP
108M	138.334305	true positive, TP
1JDO	134.549607	true positive, TP
1MCY	134.399218	true positive, TP
112M	126.016013	true positive, TP
103L	46.401593	true positive, FP
104L	42.408290	true positive, FP

By using confusion matrix evaluation, the result shows 80% of precision, sensitivity and accuracy as compared to [17].

SUMMARY AND FUTURE WORK

Both the graph based clustering and graph based matching involved compute intensive tasks. As way forward, we hope to explore parallel methods to speed up computation.

ACKNOWLEDGMENTS

Work reported here is pursued under the Short-term Grant by Universiti Sains Malaysia for "Tertiary Protein Structure Matching Based on Geometrical Features" [304/PKOMP/636038] and E-Science fund by Ministry of Science, Technology and Innovation, Malaysia (MOSTI) for "Parallel Sequence Alignment and Clustering Algorithms for Sequence Analyses of Fish Species" [305/PKOMP/613121].

REFERENCES

1. Willett, P., "Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures", Journal of Medical Chemistry, 2005, vol. 48, No. 13.
2. Bruno, I. J.; Kemp, N. M.; Artymiuk, P. J.; Willett, P., "Representation and Searching of Carbohydrate Structures Using Graph-Theoretic Techniques", Carbohydr. Res. 1997, 304, 61-67.
3. Adrian A. Canutescu, Andrew A. Shelenkov and Roland L. Dunbrack, Jr., "A Graph-Theory Algorithm for Rapid Protein Side-Chain Prediction", Protein Sci. 2003, 12: 2001-2014.
4. Bunke, H., "Graph Matching: Theoretical Foundations, Algorithms, and Applications", in Proc. Vision Interface 2000, Montreal, 2000, 82-88.
5. J. Huan, W. Wang, D. Bandyopadhyay, J. Snoeyink, J. Prins and A. Tropsha, "Mining Protein Family Specific Residue Packing Patterns From Protein Structure Graphs", Proc. RECOMB 2004, San Diego, California, USA, 2004, pp. 308-315.
6. Donald J. Jacobs, A.J. Rader, Leslie A. Kuhn, and M.F. Thorpe, "Protein Flexibility Predictions Using Graph Theory", in *PROTEINS: Structure , Function, and Genetics*, 44:150-165, 2001.

7. Matsuda, H., Ishihara, Y., Hashimoto, A., (1999), Calssifying Molecular Sequences Using Lingkage Graphs with their Pairwise Similarities, Theoretical Computing Science Vol (210), pp 305-325.
8. Yona, G., Linial, N & Linial, M., "ProtoMap: Automatic Classification of Protein Sequences a Hierarchy of Protein Families, and Local Maps of the Protein Space, *http://www.ls.huji.ac.il/michall/papers/Golan-proteins.pdf*, Vol. 37, pp 360-378, 1999.
9. Kawaji, H., Takenaka, Y., & Matsuda, H., (2004) Graph-Based Clustering for finding distant relationship in large set of protein sequences, Bioinformatics Vol 20 No 2, pp 243-250, Oxford University Press.
10. Kernighan and Lin, "An Efficient Heuristic Procedure for Partitioning Graphs", Bell System Technology, vol. 49, no. 2, pg. 291-308, 1970.
11. Bolten, E., Shliep, A., Schomburg, D. & Schrader, R. (2001), Clustering Protein Sequence-Structure Prediction by transitive Homology, Bioinformatics Vol 17, No 11, pp 932-941, Oxford University Press.
12. Kim, S. & Lee, J. (2006) BAG: A Graph Theoritic Sequence Clustering Algorithm. *International Journal of Data Mining and Bioinformatics, Vol 1 No 2*, 178-2006.
13. Mohzeni-Zadeh ,Brezellec and Risler , "Cluster C, Xifeng Yan, Philip S. Yu and Jiawei Han, "Substructure Similarity Search in Graph Databases", *Proc. of 2005 Int. Conf. on Management of Data (SIGMOD '05)*, pg 766 – 777, Baltimore, MD, 2005.
14. Chen, Y., Reilly, K.D., Sprague, A.P. & Guan, Z., "SEQOPTICS:A Protein Sequence Alignment Clustering Methods", Proceeding of the First International Multi-Symposiums on Computer and Computational Sciences(IMSCCS'06), Zhejian, China: IEEE Computer Society, 2006, (pp. 69-75).
15. Nur'Aini, Rosni and Abdullah Zawawi(2006) Nur'Aini Abdul Rashid, Rosni Abdullah, Abdullah Zawawi Haji Talib, Fast Dynamic Programming Based Sequence Alignment Algorithm, The 2nd International Conference on Distributed Framework for Multimedia Applications DFMA 2006, Pulau Pinang Malaysia, 15-17 May 2006,pp. 153-159.
16. F. Othman, R. Abdullah, J. Ali, Reference Frame for Protein Structure Matching, Konferensi Kebangsaan Biologi Bermatematik 2006 (KKBM06), Hotel Residence@UNITEN Bangi, Malaysia, 22nd – 23rd August 2006.
17. Carlo Ferrari, Concettine Guerra and Giuseppe Zanotti, "A Grid-aware Approach to Protein Structure Comparison", *Journal of Parallel and Distributed Computing*, 63 (2003), 728-737.