BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 6(10)

A graph-based clustering method applied to protein sequences

Pooja Mishra[1]* & Paras Nath Pandey[2]

[1]Center of Bioinformatics, University of Allahabad, Allahabad, India, [2]Department of Mathematics, University of Allahabad, Allahabad, India; Pooja Mishra E-mail: pooja.mishra0806@gmail.com; Phone: +91-9452377426; *Corresponding author

Received July 09, 2011; Accepted July 12, 2011; Published August 02, 2011

Abstract:
The number of amino acid sequences is increasing very rapidly in the protein databases like Swiss-Prot, Uniprot, PIR and others, but the structure of only some amino acid sequences are found in the Protein Data Bank. Thus, an important problem in genomics is automatically clustering homologous protein sequences when only sequence information is available. Here, we use graph theoretic techniques for clustering amino acid sequences. A similarity graph is defined and clusters in that graph correspond to connected subgraphs. Cluster analysis seeks grouping of amino acid sequences into subsets based on distance or similarity score between pairs of sequences. Our goal is to find disjoint subsets, called *clusters*, such that two criteria are satisfied: *homogeneity*: sequences in the same cluster are highly similar to each other; and *separation*: sequences in different clusters have low similarity to each other. We tested our method on several subsets of SCOP (Structural Classification of proteins) database, a gold standard for protein structure classification. The results show that for a given set of proteins the number of clusters we obtained is close to the superfamilies in that set; there are fewer singeltons; and the method correctly groups most remote homologs.

Keywords: Clustering, protein sequences, graph-theoretic approach.

Background:
Clustering refers to a procedure that assigns data objects to a set of disjoint classes, called *clusters*, so that objects within a class have similarity to each other in some sense. Unsupervised clustering means that clustering does not rely on predefined classes and training examples. Thus, clustering is some sort of pattern recognition. Cluster analysis consists of mathematical tools for recognizing natural and meaningful clusters within a set of samples. The importance of these tools is that they can divide similar data without any prior knowledge. That is why this field is also called unsupervised clustering. The existing clustering approaches such as k-means, fuzzy k-means etc., require the specification of initial cluster seeds, i.e. a priori knowledge of the number of natural clusters is essential and may be estimated by several potential algorithms or given randomly. Graph theory clustering methods resolve this problem, because they do not need a priori knowledge of the number of clusters. The most widely used graph clustering approaches are Markov clustering process (MCP) **[1]** and the cFinder algorithm **[2]**. The MCP approach forms clusters in the dataset using random walks in the full weighted graph that represents the similarities among the objects to be clustered. In computer science, graph theory has been widely used in many areas such as in chip and circuit design, reliability of communication networks, transportation planning, etc. However, it has been applied recently in biology. The aim of the current paper is to implement a graph theoretic approach for clustering of proteins. A graph G is an ordered pair G = (*V*, *E*), where $V = \{v_i, i = 1,\ldots, n\}$ is a set of points (nodes) and *E* is a set of edges denoted by e_{ij} or (v_i, v_j) connecting the points v_i and v_j. If the order of points v_i and v_j is not meaningful, the graph is called undirected; otherwise it is called directed. A weighted graph is a graph G in which each edge e has been assigned a real number say, *w(e)*, called the *weight (length)* of *e*. If no real number is associated with the edges the graph is said to be unweighted. If the number of elements in the vertex set V and edge set of a graph G are v and e respectively, then the incidence matrix denoted by M(G) is a v×e matrix and is defined by $M = [a_{ij}]$, the matrix element $a_{ij} = 1$, if jth edge e_j is incident on ith vertex v_i, and $a_{ij} = 0$, otherwise. The adjency matrix of a labeled graph G denoted by A (G) is a v×v matrix defined by $a_{ij} = 1$, when v_i is adjacent to v_j, otherwise $a_{ij} = 0$. We have already defined undirected graphs above; graphs that are not directed can be represented by a symmetric matrix, whereas directed graphs can be represented by using an asymmetric incidence matrix. Matrix representation of a graph is very convenient for the evaluation of any algorithm in computer processing. The graph-theoretic algorithms represent the problem data through an undirected graph. Each node (the protein sequences) is associated to a sample in the feature space, while each edge represents the distance between nodes connected under a suitably defined relationship. A cluster is thus defined as a connected sub-graph, obtained according to criteria peculiar of each specific algorithm. Algorithms based on this definition are capable of detecting clusters of various shapes and sizes, at least for the case in which they are well separated **[3]**. Moreover, isolated samples should form singleton clusters and then can be easily discarded as noise. Usually graph-based clustering algorithms do not require the setting of the number of clusters, but need however some parameters to be provided by the user. The algorithm applied in this paper overcomes this limitation, proving to be an effective solution in some real applications where a completely unsupervised method is desirable. This clustering approach is based on the algorithm described by Zahn **[4]**. At the first stage construct a Minimum Spanning Tree (MST) of the graph representing the samples. After that, identifies inconsistent edges and removes them from the MST. The remaining connected components are then the clusters in the graph G. An edge is inconsistent if the distance associated to it is greater than a predefined threshold. In order to determine the optimal value of this threshold, we used a novel method based on the use of the Fuzzy C-Means algorithm **[5]**.

Methodology:

Dataset:

The investigations were performed on the sequences taken from SCOP's **[6]** superfamily grouping. Proteins in the same superfamily are believed to be evolutionary related, and for this reason we chose such superfamily groupings as the correct groupings. The dataset taken consists of 500 sequences belonging to 6 super-families, namely globin-like (85 proteins), EF-hand (83), cuperdoxins(78), (Trans) glycosidases (81), Thioredoxin-like (79), Membrane all-alpha (94). This set was extracted from Astral-95 (http://astral.berkeley.edu/), so the maximum pairwise identity was 95%.

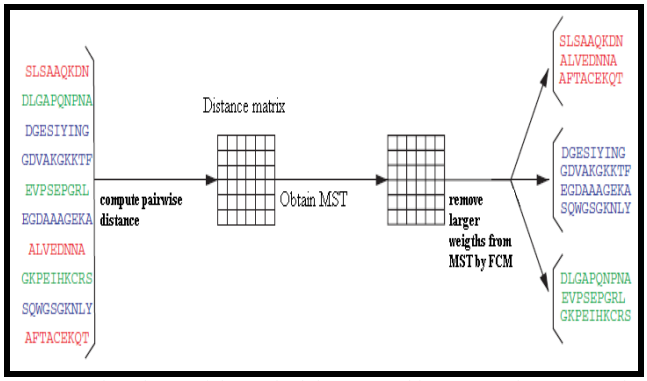

Figure 1: The scheme of the method that we used in our experiments. Proteins of the same color are evolutionary related.

Graph-Based Clustering Method

The clustering method applied here is based on graph theoretical cluster analysis. Firstly the complete graph is constructed, where each node is associated to a single protein to be clustered. The distances of a protein from all other remaining proteins in the dataset is calculated by NW algorithm **[7]**, and stored in N×N matrix, where N is the number of proteins in the dataset. The weight of each edge is the distance between the connected protein nodes, the diagonal will contain only zero, as the distance of a protein with itself is zero. Then, the Minimum Spanning Tree (MST) is computed for this graph. By removing all the edges with weights greater than a threshold δ, we arrive at a forest containing a certain number of subtrees (clusters). In this way, the method automatically groups protein nodes into clusters. As stated in **[8]**, the subtrees are independent of the particular MST, i.e algorithm chosen for deriving MST. In this paper, we have applied the Prims's algorithm. The optimal value of δ is determined by reformulating the problem as the one of partitioning the whole set of edges into two clusters, according to their weights. The cluster of the edges of the MST with small weights will contain edges to be preserved, while the edges belonging to the other cluster will be removed from the MST. This problem is solved by employing the *Fuzzy C-Means* (FCM) clustering algorithm **[9]**. (More details given in **Supplementary material**)

Result and Discussion:

We have considered the problem of clustering proteins according to their evolutionary relatedness and we are particularly interested in those cases in which some related proteins have very low sequence similarity. As a characterization of evolutionary relatedness, we used SCOP's superfamily grouping. SCOP is organized in a hierarchical manner at four main levels: class, fold, superfamily and family. At the superfamily level homology relationships may not be apparent from sequence considerations alone since proteins in the same superfamily can display varying degrees of sequence similarity. Therefore, at superfamily level, SCOP provides an excellent benchmark for testing how algorithms perform in cases, in which some related proteins have very low sequence similarity. Distance measure between two sequences is computed by the N-W alignment algorithm and PAM50 **[10]** mutation probability matrix. The distances is calculated between every pair of protein sequences in the dataset and stored in a square matrix of N×N, where N is the number proteins in a dataset to be clustered. The algorithm applied is summarized in **Figure 1**. We have tested the algorithm on different set of superfamilies, starting from 2, 3, 4, 5, 6 i.e. increasing the complexity of dataset to judge the efficiency of algorithm. This is shown in **Table 1 (see Supplementary material)**. From the results of the table, we conclude that the efficiency of the algorithm is satisfactory even when the number of superfamilies is increased in the datasets. **Table 1** shows that the algorithm predicts actual number of clusters in case of 2, 3, 4 and 5 set of superfamilies dataset. In the case of 6 superfamily dataset the predicted cluster is 8. The reason for this could be that there are some sequences which come in the twilight zone of the two or more groups and the algorithm can cluster the sequence in any one of that group. The use of accuracy rate to assess clustering performance is standard in any algorithm, but sometimes this measure can be misleading since it does not discriminate between positive and negative cases. That is, the accuracy rate is the sum of the correctly clustered cases. Another useful way to measure performance is using 'sensitivity' and 'specificity', for clustering a protein of unknown class, depending on the class predicted by the system and on the actual class of the protein. These measures are frequently used in two-class problems, but can be readily adapted for multiclass problems. Sensitivity (Se) and the specificity (Sp) can be defined as given in **Supplementary material.** Sometimes sensitivity and specificity are called true positive rate and true negative rate, respectively. Sensitivity measures the ability of the classifier system to correctly assign a protein to its real class. On the other hand, specificity measures the ability of the system to reject a given protein as belonging to a class to which it does not belong. The clustering algorithm is better than other existing algorithms in the sense that it does not require any priori information about the clusters, i.e. it is completely unsupervised. The other advantage of this algorithm is that it does not require the training of the algorithm; we can apply the algorithm directly to dataset and can obtain the clusters.

Conclusion:

We have applied existing graph theoretic techniques in the protein world and explored a new dimension for proteins. The algorithm has a low polynomial computational complexity and it is also efficient in practice. The graph-based clustering algorithm is applied to a cluster detection problem in a dataset of proteins where group of different proteins are merged together, and to detect to which group or class a particular protein belongs, with the condition that we are given only the primary sequence of that particular protein. The graph-based clustering algorithms are different from other clustering algorithms in that it does not require the user to set any parameter or threshold. This approach is can thus be used for classification of unknown proteins based on their similarity.

References:

[1] http://igitur-archive.library.uu.nl/dissertations/1895620/full.pdf
[2] Balázs Adamcsek1 *et al*. *Bioinformatics* 2006 **22**: 1021 [PMID: 16473872]
[3] http://www.springerlink.com/content/71770120u87l7827/
[4] Zahn C. *IEEE Transactions on Computers*. 1971 C-**20**: 68
[5] http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html
[6] Alexey G *et al*. *J Mol Biol.* 1995 **247**: 536 [PMID: 9016544]
[7] Needleman SB & Wunsch CD. *J Mol Biol.* 1970 **48**: 443 [PMID: 5420325]
[8] http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1671676
[9] http://www.sciencedirect.com/science/article/pii/0167865589900263
[10] Wheeler D. *Curr Protoc Bioinformatics*. 2002 3: Unit 3.5. [PMID: 18792939]

Edited by P Kangueane
Citation: Mishra & Pandey. Bioinformation 6(10): 372-374 (2011)

BIOINFORMATION open access

Supplementary material:

FCM is based on the minimization of the objective function.

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m (x_i - c_j)^2, \qquad 1 \le m < \infty$$

where m is a real number, x_i is the i-th measured data (weight of the i-th edge of the MST), c_j is the center of the cluster, u_{ij} is the degree of membership of x_i to the cluster j, C is the number of clusters and N is the number of objects to be clustered. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{x_i - C_j}{x_i - C_k} \right)^{\frac{2}{m-1}}} \quad \textbf{and} \quad C_j = \frac{\sum_{i=1} u_{ij}^m . x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

This iteration will stop when:

$$\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \Delta$$

where Δ is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m. At the end of the procedure, each edge x_i has been assigned to the cluster r such that:

$$r = \arg\max_j u_{ij}$$

At this point, all the edges of the MST are separated into two clusters. Then, we remove from the MST all the edges belonging to the cluster s whose center exhibits the largest value, i.e.:

$$s = \arg\max_j c_j$$

The applied clustering method is summarized as follows:
Construct a complete weighted undirected graph ***G;***
Obtain the ***MST*** of ***G;***
Remove from ***MST*** edges with larger weights by using the ***FCM*** algorithm.
The detected clusters are then the remaining subtrees of the ***MST***.

Sensitivity and Specificity:

$$Se = \frac{(TP \times 100)}{(TP + FN)}$$

$$Sp = \frac{(TN \times 100)}{(TP + FN)}$$

Table 1: Efficiency of clustering algorithm shown on different datasets

S. No.	Number of groups in the dataset	Total number of sequences in the dataset	Nunmber of cluster detected by Graph-Theoretic approach	$Se = \frac{(TP \times 100)}{(TP + FN)}$	$Sp = \frac{(TN \times 100)}{(TP + FN)}$
1	2	168	2	79.3	85.3
2	3	246	3	73.8	86.4
3	4	327	4	74.4	79.1
4	5	421	5	77.5	78.6
5	6	500	8	78.1	77.4