

A New Approach to Clustering Biological Data Using Message Passing

Huimin Geng¹, Dhundy Bastola², and Hesham Ali¹

¹ *Department of Computer Science, University of Nebraska at Omaha*

² *Department of Pathology and Microbiology, University of Nebraska Medical Center
{hgeng, hali}@mail.unomaha.edu, dbastola@unmc.edu*

Abstract

Clustering algorithms are widely used in bioinformatics to classify data, as in the analysis of gene expression and in the building of phylogenetic trees. Biological data often describe parallel and spontaneous processes. To capture these features, we propose a new clustering algorithm that employs the concept of message passing. Message Passing Clustering (MPC) allows data objects to communicate with each other and produces clusters in parallel, thereby making the clustering process intrinsic. We have proved that MPC shares similarity with Hierarchical Clustering (HC) but offers significantly improved performance because it takes into account both local and global structure. We analyzed 35 sets of simulated dynamic gene expression data, achieving a 95% hit rate in which 639 genes out of total 674 genes were correctly clustered. We have also applied MPC to a real data set to build a phylogenetic tree from aligned mycobacterium sequences. The results show higher classification accuracies as compared to traditional clustering methods such as HC.

1. Introduction

Clustering algorithms are frequently and successfully used to organize multivariate data into groups with similar patterns, having been applied to problems ranging from the analysis of gene expression to the building of phylogenetic trees. Among all available clustering methods used by biologists, HC is perhaps one of the most popular. It arranges the data into a tree structure which is easily viewed and understood, and the hierarchical structure provides potentially useful information about the relationships between clusters. However, HC puts a priority on global distance (similarity) without honoring local structures. Further, in each clustering process, only the two clusters which have the highest similarity are merged and clusters cannot be generated in parallel.

To overcome these problems, we propose a new clustering algorithm which simulates spontaneous and intrinsic biological processes. Inspired by a real world situation in which initially unknown people can form groups by exchanging messages, MPC allows data objects to communicate with each other, and hence improves the performance.

2. Methods

2.1. Algorithm

MESSAGEPASSINGCLUSTERING

*for every cluster C_i // i from 1 to n
 $C_i.FROM=C_i.TO=0$;*

while (threshold not reached)

for every cluster C_i

*Send a message to C_j which is closest to C_i ;
 $C_i.TO=j$; $C_j.FROM=i$;*

for every cluster C_i

if ($C_i.FROM=C_i.TO=j$)

*Merge C_i and C_j to form a new cluster;
Update the distance table;*

2.1. Properties

While it may appear that MPC and HC produce similar outputs, we show that more often than not, the output clusters of the two approaches are different in favor of MPC methods. In particular, if the centroid is used as the distance criterion in MPC, the results obtained by MPC and HC are often different. Also if single, complete and average linkage distance criteria are used, the final clustering dendrograms from MPC and HC are equivalent, but intermediate clusters are often different. In the following lemmas, we used the term *core* to denote a cluster formed by merging two clusters.

Lemma 1: The distance between a pair of objects in a core is always smaller than that between the object inside the core and the object outside the core.

Lemma 2: The number of cores produced at each step is from 1 to $n/2$, where n is the number of clusters at the previous step.

Theorem: For centroid criterion, the clustering solutions using MPC and HC may be different. For single, complete, and average linkage, the final clustering dendrograms from MPC and HC are equivalent, but intermediate clusters are different. If the number of final clusters is specified in advance, MPC and HC may yield different solutions.

3. Results

We test the validity of the proposed approach in two ways: simulated data (microarray expression data) was used to show that the proposed method has high accuracy and stability; real data (aligned mycobacterium sequences) were used to look at how well the results generated by the proposed method agree with the real phylogenies and to show the superiority of the proposed method to existing techniques, such as HC and neighbor joining (NJ) methods.

An on-line simulator, eXPATGen [2], was used in the general evaluation of the proposed method. Employing the user-defined inputs to the simulator, dynamic mRNA profiles similar to those produced from microarray experiments were generated. Then the proposed method was used in the analysis of the data and the results were compared with the initial input.

We tested 35 data sets using the proposed method. For each data set, 10 to 100 genes with the serial of time points dimension ranging from 20-40 were included. A 95% hit rate was achieved, in which 639 genes out of a total of 674 genes were correctly clustered.

We also successfully applied the proposed method to constructing phylogenetic trees. The similarity matrix that we used was obtained from 34 strains of nine species of *Mycobacterium*. Figure 1 demonstrates the results of using the NJ program in the PHYLIP package (Felsenstein, 1989). It is seen that only five out of a total of nine groups of species show biologically relevant clusters. The phylogenetic tree obtained by the MPC method (Figure 2) shows reasonable relation among the *Mycobacterium* species. It generates exactly nine, representing multiple strains from nine different species. The analysis of the same data set using the HC method shows different topology and incorrect positioning of the same strains of *Mycobacterium* species, because HC is a greedy algorithm which considers only the global structure but does not honor local structure.

4. Conclusion

We proposed a new clustering algorithm which employs the concept of message passing. By taking advantage of the communication among data elements and by taking into account both local and global structure, MPC can describe parallel and spontaneous biological processes more precisely, and hence it can produce more accurate clustering solutions. Using the proposed approach, a 95% hit rate was achieved for the simulated gene expression data, and biologically relevant topologies for different *Mycobacterium* species were reflected in the phylogenetic trees obtained from the real data set.

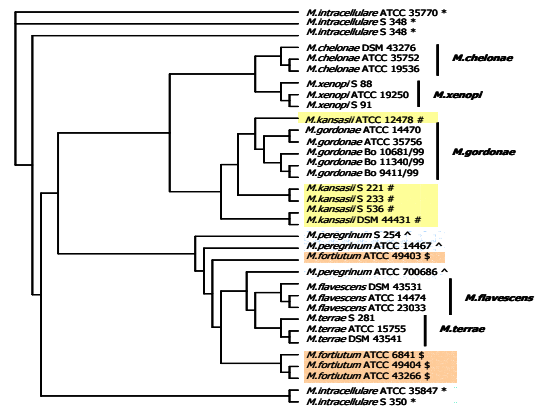


Figure 1. Phylogenetic tree constructed by NJ method

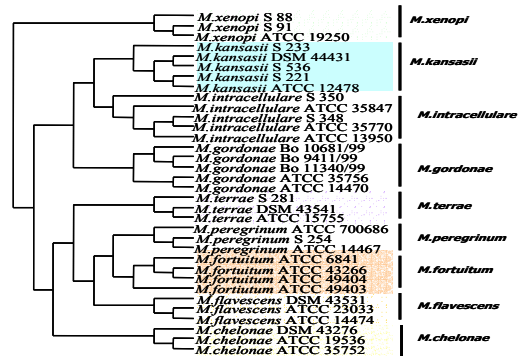


Figure 2. Phylogenetic tree constructed by MPC method

5. References

- [1] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns", *Proc Natl Acad Sci, U S A*, 1998 Dec. Vol. 95, pp. 14863-14868.
- [2] D.J. Michaud, A.G. Marsh, and P.S. Dhurjati, "eXPATGen: generating dynamic expression patterns for the systematic evaluation of analytical methods", *Bioinformatics*, 2003, Vol. 19 no. 9, pp. 1140-1146.