The goal of biclustering in a gene expression data matrix is to find a submatrix such that the genes in the submatrix show highly correlated activities across all conditions in the submatrix. A measure called Mean Squared Residue (MSR) is used to simultaneously evaluate the coherence of rows and columns within a submatrix. In this paper a new method for biclustering gene expression data is developed. In the first step high quality bicluster seeds are generated using K-Means clustering algorithm. Then more genes and conditions (node) are added to the bicluster. Before adding a node the MSR X of the bicluster is calculated. After adding the node again the MSR Y is calculated. The added node is deleted if Y minus X is greater than MSR difference threshold or if Y is greater than d (MSR threshold) which depends on the dataset. The MSR difference threshold is different for gene list and condition list and it depends on the dataset also. Proper values should be identified through experimentation in order to obtain biclusters of large size. Since it is very difficult to calculate the value of MSR difference threshold, in this algorithm an iterative search is used where MSR difference threshold is initialized with a small value and it is incremented after each iteration. A bicluster is obtained from Yeast dataset with a unique structural appearance. This proves that the newly introduced concept of MSR difference threshold will result in high quality biclusters. The results obtained on bench mark datasets prove that this algorithm is better than many of
the existing biclustering algorithms.

Reference

- Shyama Das and Sumam Mary Idicula "A Novel Approach in Greedy Search Algorithm for Biclustering Gene Expression Data" International Conference on Bioinformatics, Computational and Systems Biology (ICBCSB), WASET, 2009.
- SGD GO Termfinder [http://db.yeastgenome.org/cgi bin/GO/goTermFinder]
- Federico Divina and Jesus S. Aguilar-Ruize, "Biclustering of Expression Data with Evolutionary computation", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, pp. 590-602, 2006.

Index Terms

Computer Science Data Mining
Key words

Biclustering
gene expression data

K-Means clustering

Mean Squared Residue