BMC Bioinformatics

Meeting abstract

LSI based framework to predict gene regulatory information

Sujoy Roy[1,2], Lijing Xu[1,2] and Ramin Homayouni*[1,2]

Address: [1]Department of Biology, University of Memphis, Memphis, TN 38152, USA and [2]Bioinformatics Program, University of Memphis, Memphis, TN 38152, USA

Email: Ramin Homayouni* - rhomayon@memphis.edu

* Corresponding author

This abstract is available from: http://www.biomedcentral.com/1471-2105/10/S7/A5

Background

Latent Semantic Indexing (LSI), a vector space model for information retrieval, has shown promise in predicting functional relationships between genes using textual information in MEDLINE abstracts. The underlying principle is that genes may be represented as document vectors in a multi-dimensional hyperspace, and the conceptual relationship between any two genes is determined by the cosine of the angle between their vectors [1]. In this study, we sought to extend this concept for identification of putative transcription factors (TFs) that regulate a group of co-regulated genes. We hypothesized that co-expressed genes identified by microarray experiments are functionally related and that at least some of these genes have previously been linked explicitly or implicitly to TFs in the literature. A transcriptional module is then defined as a set of genes clustered together in LSI space with closely related TFs (Figure 1). We devised a framework using these assumptions to identify transcriptional modules from microarray and promoter motif data (Figure 2). The framework requires as input, co-expressed genes from a microarray dataset and a set of TFs that have consensus motifs in the promoter regions of the co-expressed genes. Usually the set of such motif-derived TFs is large and makes the identification of the critical ones difficult. The framework first identifies functionally related clusters of co-expressed genes based on their latent relationships from literature, and then adds to each cluster TFs that are closely associated with the genes in the cluster. The putative transcriptional modules are ranked based on the degree of relative literature coherence amongst the entities in them.

Results and discussion

The LSI-based algorithm allows prediction of TFs based on latent (implicit) relationships in the literature. A preliminary evaluation of our method using previously published knock-out experiments revealed that it has reasonable recall and precision. A more rigorous evaluation of the method will require several additional TF knock-out microarray experiments. This work provides proof of principle that the combination of motif analysis
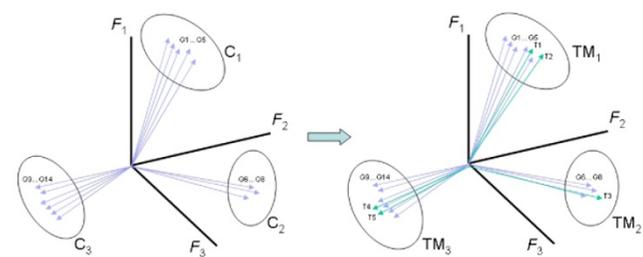

Figure 1
Gene clusters and transcriptional modules in LSI space. Gene document vectors (blue) are clustered together in LSI space based on the closeness of the angle between them. Next, transcription factor vectors (green) are added to the clusters if their cosine value is close to the average cosine of the gene cluster.
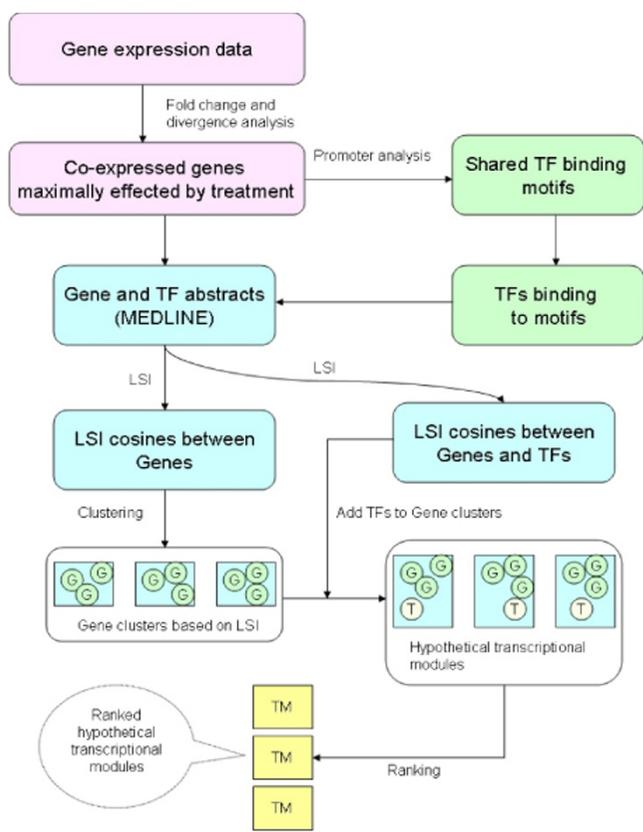
Figure 2
Overview of Framework.

and LSI may be used to identify putative transcriptional modules from microarray data.

References
1. Homayouni R, Heinrich K, Wei L, Berry MW: **Gene clustering by latent semantic indexing of MEDLINE abstracts.** *Bioinformatics* 2005, **21(1):**104-115.