Mining the human interactome through gene expression time series analysis

Ferrè F (1), Clote P (1), Ausiello G (2), Via A (2), Cesareni G (3), Helmer-Citterich M (2)

(1) Biology Dept., Boston College, MA
(2) Center for Molecular Bioinformatics, Dept. of Biology, University of Rome Tor Vergata, Rome
(3) Dept. of Biology, University of Rome Tor Vergata, Rome

Motivation

Gene expression as measured by DNA microarray experiments offers at a glance an overall view of cellular processes, thus revealing complex relationships among genes. Sampling gene expression levels at different points in time produces a dynamic landscape of expression changes. This offers unique perspectives for the clustering of genes with similar temporal expression and for the reconstruction of functional modules. Time series microarray experiments are ideal for the study of dynamic processes, such as cell cycle, development and changes in gene expression in response to different levels of a new condition, like a drug, and can be used to highlight differences and similarities between different tissues, or between normal and pathological conditions. On the other hand, protein interaction networks can provide information about the skeleton of physical interactions underlying the relationships among the involved genes. Our goal is the study of human interaction networks and their temporal evolution in different tissues and conditions by means of the clustering of interacting proteins with similar temporal expression. The integration of gene expression time series data with interaction networks is a powerful tool for the discovery of gene modules and the analysis of how these modules are perturbed in pathological conditions, thus creating the basis for a molecular pathology of diseases.

Methods

While algorithms for the analysis of static microarrays can in principle be applied to dynamic microarray data, the study of gene expression time series raises specific problems and requires specifically developed computational tools in order to fully exploit the data information content. Dynamic time warping (DTW) is an algorithmic technique for the computation of the smallest distance and optimal alignment between two numerical sequences. DTW is evidently a very flexible tool for time series data comparison [1] since it identifies similarity between sequences where there is a shift in the time axis, and it accommodates sequences of different length. The quality of an alignment is estimated by the time warping distance (TWD), which is an alternative and possibly better measure than the commonly used distance measures (Pearson correlation coefficient, Euclidean distance, L1 distance). We used TWD as distance measure for our implementation of the Cluster Affinity Search Technique (CAST) algorithm [2], which is of particular interest for the analysis of noisy data since it explicitly incorporates an error model, thus it is particularly appropriate for biological applications. Unlike most clustering algorithms, CAST makes no assumptions about the number of clusters, their size or structure, which are discovered from the data. Arguably, the most relevant property of DTW is that it can identify and align sequences, which have approximately the same overall component shapes, but these shapes do not line up along the x-axis. DTW warps the time axis of one or both sequences to achieve a better alignment. This time shift may be indication of a functional relationship where the expression of one gene activates one or more other genes. Clustered genes can identify a gene module where similar expression through time is an indication of functional interaction. Genomic data helps the pathway reconstruction, for example through the identification of similar transcription factors binding sites. Sub-cellular localization (known or predicted) is used to further filter the clusters.

Results

Human interactome modules, obtained by the clustering of public human time series data [3-5], are benchmarked using curated protein-protein interactions from the MINT database [6] and gene ontology (GO) annotations, and a confidence index is derived accordingly. Cluster of genes similarly expressed through time are identified, highlighting gene modules which are associated with specific functions (for example cell cycle progression regulation), or which are affected in pathological conditions.

Contact email: citterich@uniroma2.it