BMC Systems Biology

Poster presentation

Identification of phenotypes in patient microarrays

J Keith Vass*, Gabriela Kalna and Des J Higham

Address: The Beatson Institute for Cancer Research, Dept of Mathematics, University of Strathclyde, Glasgow, UK

Email: J Keith Vass* - k.vass@beatson.gla.ac.uk

* Corresponding author

from BioSysBio 2007: Systems Biology, Bioinformatics and Synthetic Biology
Manchester, UK. 11–13 January 2007

Published: 8 May 2007

BMC Systems Biology 2007, **1**(Suppl 1):P75 doi:10.1186/1752-0509-1-S1-P75

This abstract is available from: http://www.biomedcentral.com/1752-0509/1?issue=S1

Background

The focus of patient studies on sample or gene classification oversimplifies, discarding the diversity in any group and loses much information as "noise". Natural variations represent an enormously rich source of expression-perturbations allowing identification of a network with over 10^7 gene co-expression pairs in leukaemia samples. We classify gene-expression as up, down or unchanged, allowing the construction of 3 types of co-expression matrices – up-together, down-together and up-down. These matrices are analysed by singular value decompositon (SVD) and successive vectors used to cluster genes, revealing highly connected subgraphs. These clusters can be used as bait to interrogate other network types – revealing potential negative effectors or consequential down-regulated genes.

Results

We demonstrate that many co-expression relationships show highly consistent behaviour: genes that are frequently down-regulated together are also often up together; genes showing a relationship of: gene-1 up -> gene-2 down often show gene-1 down <- gene-2 up.

In an acute myeloid leukemia dataset (Valk et al 2005) we identify 2 easily biologically recognizable clusters: 1) DNA copying/cell-division and 2) Erythroid Band 3 multiprotein complex with known controlling factors. We discuss the use of our analysis to disect the biology of control of both clusters and assess the possibility that they are functionally linked.

Materials and methods

Classification is by the Z-score method of Quackenbush, 2001, where each sample is compared to the mean of all samples in the dataset. The classified data are initially stored in a 3 value matrix (-1, 0, 1) from which 2 matrices are derived (0,1) for the up-classifications and for down. From these we calculate the co-expression patterns by matrix dot-products. To filter out relationships that were likely to be due to chance, given the density or number of 1's for each gene, Monte Carlo simulation methods estimate the distribution of scores for randomized vectors of all possible densities, by permuting the order of each and then recording the number of times 1's occur for both vectors at each position. The test was repeated 1000 times for every pair of vectors and the scores which included 99.5% (calculated by the R-package25 function quantile) of the tests was used as a cut-off. We estimated false discovery rate by randomizing the order of each gene, then constructing the matrices – this give around 8% of the edges found with un-shuffled data. Two matrix types (up-together and down-tegether) are symmetric but the up-down matrix is square and non-symmetric. The nodes (gene probesets) are re-ordered by using successive eigen or SVD vectors with clusters members requiring a maximum of one-unlinked edge to be included.

Conclusion

Avoiding prior knowledge, using very large sample numbers and classifying expression data allows us to calculate very large networks of up to about 10^7 edges with almost 20,000 genes. We have succeeded in clustering smaller

matrices of about 9,000 genes and identified groups of genes immediately recognizable by biologists. We have demonstrated that these "phenotypes" can be used to further identify candidates involved in regulating or being regulated by the identified group of genes. This approach provides a framework for the logical analysis of transcriptional effects and is generally applicable to other types of large scale data.