BMC Bioinformatics

Poster presentation

A method for validation for clustering of phenotypic gene knockdown profiles using protein-protein interactions information

Nikolay Samusik, Yannis Kalaidzidis and Marino Zerial

Address: Max-Plank Institute of Cell Biology and Genetics, Dresden, Germany

from Fifth International Society for Computational Biology (ISCB) Student Council Symposium
Stockholm, Sweden 27 June 2009

Published: 19 October 2009

BMC Bioinformatics 2009, **10**(Suppl 13):P3 doi: 10.1186/1471-2105-10-S13-P3

This article is available from: http://www.biomedcentral.com/1471-2105/10/S13/P3

The functional characterization of the molecular components of biological systems is an outstanding task in the post-genomic era and is often addressed by gene knockdown screening which quantifies phenotypes in the form of multiparametric profiles. The identification of sets of genes producing similar phenotypic effects is fulfilled by the clustering of profiles. Here we address the question of how to quantify the biological significance of clusters of phenotypic profiles with respect to protein-protein interaction (PPI) data.

The dataset we used contains 20385 gene knockdown phenotypic profiles resulting from the genome-wide image-based multiparametric siRNA screen on endocytosis in HeLa cells, which has been carried out on in our lab. Each profile is produced from images of cells in a specific gene knockdown condition, and comprises a 40-dimensional vector of phenotypic parameters, which describe the distribution of endosomes labeled by fluorescent cargo. The profiles capture the morphological changes in the endocytic system resulting from different gene knockdowns. The dataset has been pre-filtered, discarding profiles showing weak effects.

Protein-protein interaction data was taken from EMBL STRING database [1]. We demonstrated that the interacting proteins are 17% more likely to produce a correlated phenotype in the siRNA screen than non-interacting proteins. Based on this observation, we developed a method for the cross-validation of the clustering of phenotypic profiles using PPI data. The relevance of a cluster with respect to PPI is measured by the ratio of the number of interactions between genes in the cluster to the total number of interactions of these genes, and by the average size of a fully connected PPI module in the cluster. The significance of these values for a cluster of a size N is assessed by bootstrapping, using sets of N nearest neighbors of random phenotypic profiles, and the p-value combines both measures.

We have applied this method to the results of clustering of the phenotypic profiles by means of a modified version of the "quick shift" [2] density-based clustering algorithm equipped with k-nearest neighbor (kNN) adaptive [3] von Mises-Fisher kernel for density estimate. The proposed measure shows significant p-value for the majority of the clusters. The p-value is sensitive to the artificial deterioration of cluster quality by addition of adjacent datapoints or removal of those on the border, and thus must allow optimization of clustering parameters.

To demonstrate that, we have created a method for the tree-cutting in hierarchical clustering which maximizes the significance of resulting clusters with respect to the measure we developed.

References

1. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P and von Mering C: **STRING 8 – a global view on proteins and their functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, **37 Database:** D412–416, Epub 2008 Oct 21.
2. Vedaldi A and Soatto S: **Quick Shift and Kernel Methods for Mode Seeking.** *Proceedings of the European Conference on Computer Vision* 2008, **5305:**705–718.
3. Georgescu B, Shimshoni I and Meer P: **Mean shift based clustering in high dimensions: a texture classification example.** *Proceedings of Ninth IEEE International Conference on Computer Vision* 2003, **1:**456–463.