154. Prediction of gene function from biological interactions

Joannis Apostolakis, Daniel Güttler, Florian Sohler and Ralf Zimmer[1]

Keywords: Protein function prediction, target identification, co-regulation protein interaction

1 Introduction

Large scale experiments on gene regulation, protein interaction and genetic interaction allow the derivation of relations between genes that may be used for identifying molecular factors relevant for a specific function or state of a cell. Of particular interest is the use of such experiments to unravel the molecular mechanisms underlying disease, in order to identify targets for therapy or diagnosis. One very common approach to this problem is to perform differential display experiments between known sample states, such as diseased and healthy tissue. However, the states of the samples are not well defined since they contain systematic bias that makes direct identification of possible targets very difficult. Further the comparison between healthy and diseased states limits the available data on samples of known state. Here, we generalize the problem of target identification as a particular instance of function prediction and use a "guilt by association" approach to identify the function of genes. While this approach in itself is not new, we extend it to different types of experimental data and combinations thereof. The presented methods have require knowledge of at least some relevant genes, on the other hand they allow the use of a significantly larger amount of experimental data within the same framework. As a first step, we investigate the possibility of gene function prediction on the yeast compendium data set, especially in combination with biological data from other sources, such as the scientific literature or other large scale experiments, e.g. protein-protein interaction scans.

2 Methods

Large systems of biological relations between genes can be described as simple graphs, whose vertices represent genes or proteins and edges represent specific relations between connected vertices. We use a simple and general approach for the prediction of protein function based on neighborhood in such graphs and show that reasonable predictions can be made for a large number of genes from the knowledge of the function classes their neighbors belong to. We analyze the quality of the prediction depending on the function being predicted, the function class size and the type of data used for the prediction (co-regulation, protein-protein interaction and co-occurrence in scientific abstracts). Co-regulation association is obtained by placing edges between each gene and the 10 closest genes in terms of correlation coefficient of expression [3]. For every gene, the number of neighboring genes that belong to each functional category is used to obtain a p-value for the prediction that the central gene also belongs to that functional class [3]. An alternative approach based on supervised learning with support vector machines is also tested.
Different possibilities for combining the information of different edge types are and we have started to automatize a procedure for scanning the available literature for information that may support classifications, that have been predicted with high reliability, and that are not found in the original classification hierarchy that was used for learning.

[1] Department for Practical Informatics and Bioinformatics, Ludwig Maximilians University Munich, Theresienstr 39, D-80333, email: apostola@bio.informatik.uni-muenchen.de

3 Data and Results

For the evaluation of the protein function prediction based on general biological associations we used the yeast compendium data (YCD) set by Hughes et al. [1], protein interaction data derived from yeast 2 hybrid experiments, mass spectrometric characterization of isolated complexes and the literature (von Mering et al. [2] and references therein). Finally we used a simple text mining approach to obtain co-occurrence of gene names in scientific abstracts taken from the Medline database. For the functional classes we used the hierarchical MIPS functional classification, as well as a simple non overlapping classification into 13 different functional groups [2]. These classifications depend mainly on the integration of the different genes in the functional organization of the cell and not so much on function type (e.g. reaction catalyzed).

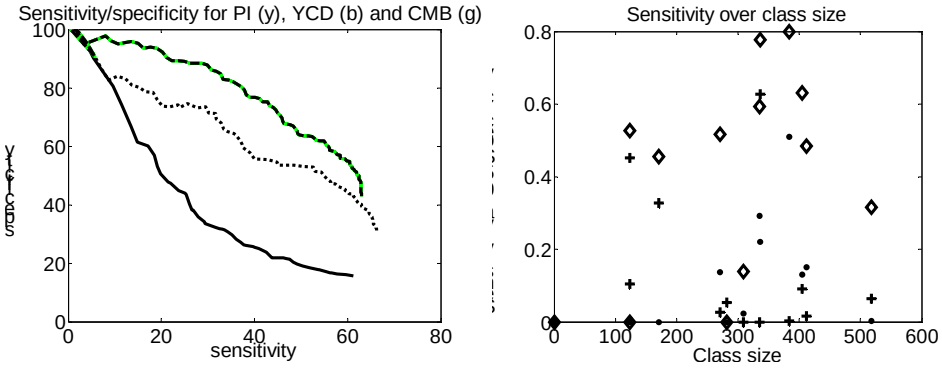

Figure 1: (a) Sensitivity/specificity plots for function prediction based on protein interaction (dotted), expression co-regulation (continuous) and a logical combination of the two (dashed). (b) Class size dependence of the sensitivity at 50% specificity (dots, plus signs and diamonds respectively)

In figure 1a a typical sensitivity/specificity plot for different types of data is shown, while figure 1b shows the sensitivity at a specificity of 50% in dependence of the class size. The results indicate that reasonable prediction sensitivities are obtained for high specificities. Predictivity depends strongly on the type of data used to obtain the association graph. Co-occurrence of gene names in scientific abstracts yields the best results for function prediction (see extended version). However the results obtained for large scale experimental data yield also satisfactory results. It is further shown that indirect connectivities in the graphs still allow reasonable prediction. Comparison of the quality of prediction for different functional classes shows that results strongly depend on the type of function. Furthermore, the prediction for some functional classes depends strongly on the type of data used. In the case of co-regulation data this indicates that different experiments are more or less suited for the identification of different functions, and that some functional classes may not be significantly regulated at all, as is often assumed to be the case for so-called housekeeping genes. The results indicate the power of the approach for a coarse screening of candidates for specific functions.

4 References

[1] Hughes T.R. et al. , 2002, Functional discovery via a compendium of expression profiles. *Cell* 7;102(1):109-26.

[2] von Mering et al., 2002, Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399-403

[3] Wu L.F., et al.., 2002, Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters *Nat. Genet*., 31(3):255-65.