Predicting gene function by combining expression and interaction data

R.J.P. van Berlo and L.F.A. Wessels and S.D.C. Martes and M.J.T. Reinders
Delft University of Technology
Information and Communication Theory Group
Mekelweg 4, 2628 CD Delft, The Netherlands
r.j.p.vanBerlo@ewi.tudelft.nl

Abstract

In this study we combined the spurious protein interaction data from the Database of Interacting Proteins with the recently published gene expression data of S. cerevisiae *grown with limited nutrient limitations under different physical/chemical conditions (Tai et al. [2]) in order to predict protein interactions and protein functions with more confidence. Because proteins often have multiple functional annotations, we propose to employ a continuous metric (e.g. the cosine angle) for measuring functional similarity. We show that it is possible to extract multiple functional associations of a gene, but only by applying a strict Pearson correlation threshold on the gene expression data. Using this strategy, we were able to predict the function of six formally unclassified proteins. Additionally, we revealed six small networks of interacting proteins. These networks strongly match with existing biological knowledge. Furthermore, transcription factors could be assigned to four of these interaction networks by employing a recently published transcription database (Harbison et al. [1]).*

1. Introduction

Since microarray technology has reached maturity and since large-scale (physical) interaction data sets have been published, integrative bioinformatics approaches have appeared for identifying gene interactions and novel functions of genes. These studies indicate that integrative bioinformatics approaches pay off and are efficient methods for the elucidation of gene function and for the identification of protein interactions.

However, when the researchers mention that their methods are capable of predicting gene function, they mean that the methods are able to identify at least one functional (GO- or MIPS-) annotation of a gene at level three or four of specificity. However, genes often have multiple functional annotations. Here we proposed a new metric, the cosine angle, to capture functional similarity between two proteins, taking into account all the functional annotations. We further investigate under which conditions this metric reliably predict protein function.

We combined the nearly 15000 putative protein-protein interactions from the Database of Interacting Proteins [4], from which 2000 are interactions between a protein with unknown function and a protein with known function, with the recently published gene expression data by *Tai et al.* [2]. In doing so, our goal is to reliably predict the function of unknown proteins by integrating their expression and interaction data. The gene expression data set consists of triplet measurements of gene expression levels of *S. cerevisiae* grown with four different nutrient limitations in both aerobic and anaerobic chemostat cultures.

2. Methods

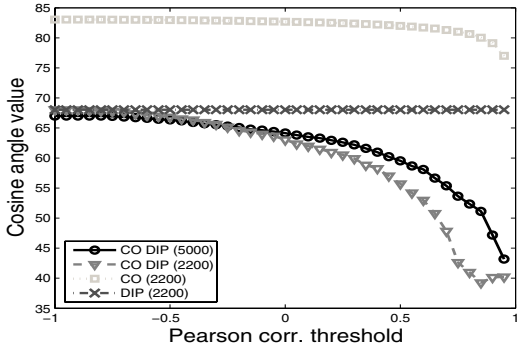
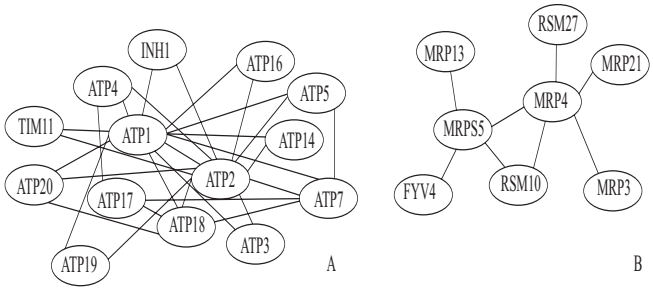


Figure 1. Mean cosine angle for different pairwise Pearson correlation thresholds $[-1 : .05 : .95]$. The values are presented for the case that either only DIP (DIP) or expression data (CO) is used $[N = 2200]$ and for the case that both data are combined (CO DIP) $[N = 2200, 5000]$. The number of selected features N was changed by using different coefficient of variation thresholds.



Interaction	*Uncl. Protein*	*Clas. Protein*	*Func. Desc. Clas. Protein*
1	YLR312C	MPD2	Protein disulfide isomerase-related protein
2	FYV4	MRPS5	Mitochondrial ribosomal protein, small subunit

C

Complex	*Motif*	*TF*	*Corr*
A	CCAAT	HAP2/3/4/5	$\geqslant 0.45$
B	ACCGA		

D

Figure 2. Examples of the different obtained results. A) and B) are two examples of the six found interaction networks. C) shows two examples of the assigned functions to formally unclassified proteins. Finally, D) displays over-represented pentanucleotides (motifs) detected by RSA tools [3] and probable transcription factors based on research of Harbison [1]. The last column gives the minimum (absolute) pairwise correlation coefficient between the transcription factor and the group of proteins in the particular complex. For complex B the motif did not coincide with a known transcrption factor in [1].

Proteins can have multiple functional annotations. Therefore, we propose to represent the function of a protein as a vector $(F_1^i, \ldots, F_m^i, \ldots, F_M^i)$. The length of the vector (M) equals the number of (MIPS) functional categories used for the proteins in *S. cerevisiae*. An element F_m^i equals one when the functional category m is assigned to protein i and zero otherwise. We have chosen to employ the cosine angle as metric for measuring the similarity between the functions (vectors) of two proteins. This metric enables us to take into account all the functional annotations of a protein and is defined as $\arccos\left(\frac{V^i \cdot V^j}{\|V^i\| \|V^j\|}\right)$.

3. Results

We employed four criteria to define pairs of proteins, which are subsequently assumed to have the same function. This allows the prediction of the function of an unknown protein paired with a known protein. The following criteria were employed: 1) the presence of only a DIP interaction; 2) a threshold on the Pearson correlation of the expression level of the two proteins; 3) both previous requirements; and 4) both requirements 1) and 2) and additionally requiring that the expression level of each of the two proteins exceeds a minimum coefficient of variation. By only considering pairs for which the MIPS annotation is known, the function prediction can be evaluated by comparing both annotations using the the cosine angle metric. With the fourth pairing strategy we obtained the highest prediction power for a strict threshold on the Pearson correlation: $(\rho_{thres} = 0.85)$, see Figure 1. Using this strategy for pairs in which one of the proteins has an unknown MIPS annotation, we were able to predict the function of six formally unclassified proteins (see Figure 2C).

The putative interactions between the established protein pairs were combined to identify networks of interacting proteins. We could identify six interaction networks. These networks strongly match with existing biological knowledge, e.g. Figure 2A-B.

Harbison et al. [1] made a regulatory map containing binding information about 102 TFs, i.e. for each TF the regulatory map indicates to which genes in the yeast genome the TF can bind (upstream) and thus possibly manipulate the expression levels. Using this map, we could identify TFs that were significantly overrepresented in four of the six interaction networks (see Figure 2D).

4. Discussion

The integration method leads to more reliable predictions about protein function and to more reliable protein interaction predictions. For example, all genes in complex A contain the pentanucleotide CCAAT. It is known that the heteromeric transcription factor complex consisting of the four subunits, $HAP2$, $HAP3$, $HAP4$ and $HAP5$, binds to this CCAAT motif. Each of these four genes was at least weakly correlated to this respiration complex (between 0.45 and 0.98). Interestingly, the protein HAP4 was most strongly related to this respiration complex $(0.71 - 0.98)$. This is the gene that does not bind DNA directly, but augments the binding of $HAP2$ and $HAP3$ and appears to be the most important protein for transcriptional regulation.

References

[1] C. T. Harbison et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, Sept. 2004.

[2] S. L. Tai et al. Two-dimensional tanscriptome analysis in chemostat cultures. *The J. of Biol. Chem.*, 280(1):437–447, Jan. 2005.

[3] J. van Helden et al. Extracting regulatory sites from the upstream region of yeast genes by comp. anal. of oligonucleotide frequencies. *J. of Mol. Biol.*, 281(5):827–842, Sept. 1998.

[4] I. Xenarios et al. Dip, the database of interacting proteins. *Nucleic Acids Research*, 30(1):303–305, Jan. 2002.