

A Rich Probabilistic Model to Predict Yeast Gene Function

Lan V. Zhang¹, Oliver D. King², Zeba Wunderlich³ and Frederick P. Roth⁴

Keywords: integrated probabilistic model, probabilistic decision tree, gene function, protein interaction, protein domain, and gene ontology.

1 Introduction.

Prediction of gene function is an important problem in the post-genome era. Traditionally, functions of unknown genes are inferred from two types of methods: one using the “guilt-by-association” principle (*e.g.* [1]), and the other using features of the gene of interest (*e.g.* [2]). Both types of methods have shown certain success in the task. Here we aim to combine the two principles using one rich probabilistic model.

2 Methods.

To predict whether a gene of interest possesses one particular function, we exploit knowledge of its other characteristics, the characteristics of its neighbors in an integrated biological network, as well as the type(s) of interactions or relationships that link them. We model genes and the interactions/relationships between them as an integrated network. Each node represents a gene or its protein product, and is associated with a “node vector” encoding characteristics of the gene/protein. Such node characteristics can include functional annotation, protein domains, sequence motifs, *etc.* Each node pair is assigned an “edge vector” encoding the interactions/relationships between the two adjoining nodes. Edge characteristics can include physical interaction, genetic interaction, sequence homology, transcriptional regulation, expression correlation, *etc.* We formulate the question as to estimate, for any given gene and function of interest, the conditional probability of the gene having that function given the characteristics of all other nodes and edges.

There have been several previous approaches to predict gene function by integrating multiple datasets [3-6]. Compared with previous methods, our model benefits from the fact that we do not assume conditional independence between different gene characteristics given the interactions, nor do we assume conditional independence between different edge characteristics. The model was trained using probabilistic decision trees [7], and evaluated using cross-validation.

3 Results.

¹ Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School. Email: lan_zhang@student.hms.harvard.edu

² Whitehead Institute for Biomedical Research. Email: oking@wi.mit.edu

³ Biophysics Program, Harvard Medical School. Email: wunderl@fas.harvard.edu

⁴ Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School. Email: fritz_roth@hms.harvard.edu

We applied our method to 13 functional categories derived from the MIPS database [8]. We used 4950 Pfam [9] protein domains as additional node labels, and three different interaction types — physical interaction [8], genetic interaction [8], and correlated expression [10]. The results were compared with a previous study [4]. Predictions were also made for a larger and more specific collection of Gene Ontology [11] function terms in yeast.

References

- [1] Schwikowski, B., Uetz, P. and Fields, S., 2000. A network of protein-protein interactions in yeast. *Nat Biotechnol*, **18**(12): p. 1257-61.
- [2] King, O. D., Foulger, R. E., Dwight, S. S., White, J. V. and Roth, F. P., 2003. Predicting gene function from patterns of annotation. *Genome Res.*, **13**(5): p. 896-904.
- [3] Deng, M., Zhang, K., Mehta, S., Chen, T. and Sun, F., 2003. Prediction of protein function using protein-protein interaction data. *J Comput Biol*, **10**(6): p. 947-60.
- [4] Deng, M., Chen, T. and Sun, F., 2003, Integrated probabilistic model for functional prediction of proteins, in The ACM-SIGACT 7th Annual International Conference on Computational Molecular Biology (RECOMB03): Berlin, Germany. p. 95-103.
- [5] Deng, M., Tu, Z., Sun, F. and Chen, T., 2004. Mapping Gene Ontology to proteins based on protein-protein interaction data. *Bioinformatics*, **20**(6): p. 895-902.
- [6] Letovsky, S. and Kasif, S., 2003. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, **19 Suppl 1**: p. i197-204.
- [7] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., 1984, Classification and regression trees. Wadsworth statistics/probability series. Belmont, Calif.: Wadsworth International Group. x, 358.
- [8] Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. and Weil, B., 2002. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, **30**(1): p. 31-4.
- [9] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., *et al.*, 2004. The Pfam protein families database. *Nucleic Acids Res*, **32**(Database issue): p. D138-41.
- [10] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*, **9**(12): p. 3273-97.
- [11] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.*, 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**(1): p. 25-29.