Improving the Accuracy of Functional Inferences from Links between Proteins Detected by Computational or Experimental Methods

Michael Thompson[1] **Edward Marcotte**[1,2] **Matteo Pellegrini**[1]
mjt@proteinpathways.com
Todd Yeates[1,2] **David Eisenberg**[1,2]

[1] Protein Pathways, Inc. 1145 Gayley Ave., #304, Los Angeles, CA 90024, USA
[2] Molecular Biology Institute, Box 951570, University of California, Los Angeles, Los Angeles, CA 90095-1570, USA

Keywords: functional genomics, bioinformatics, protein function, phylogenetic profiles, domain fusion

1 Introduction

Various methods have been developed for detecting functional couplings among proteins. These include recently developed computational techniques such as the protein phylogenetic profile method [1] and the Rosetta Stone method [2], as well as experimental techniques such as mRNA expression profiling [5]. Here, we present a general approach for quantitatively discriminating more informative functional links from less informative ones. We make the simple assumption that a pair of linked proteins which also share links to other proteins represents a more reliable connection than a pair of linked proteins which share few or no links to other proteins (Figure 1). This approach works on links established by the Rosetta Stone method, the phylogenetic profile method, and on links obtained from mRNA expression profiles.

2 Method and Results

We denote the link between proteins i and j as $l_{i,j}$ which can take on binary or real values depending on the underlying functional coupling detection method. For each protein, i, we can construct a vector $\vec{v}_i = (l_{i,1}, l_{i,2}, \ldots, l_{i,n})$ of couplings to all the other n proteins in the set (proteome).

A "link distance", $d_{i,j}$, can be computed between all pairs of proteins, i, j, based on their link vectors. For the cases where $l_{i,j}$ takes on binary or real values, this link-distance is the Hamming or Euclidean distance, respectively. For the purposes of this work, $d_{i,j}$ was normalized by the maximum link-distance observed in the dataset, d_{max}. Thus, all $d'_{i,j} = \frac{d_{i,j}}{d_{max}}$ lie in the range $[0, 1]$.

The link-distance, $d'_{i,j}$, can be treated as a parameter for optimization of functional inferences made between coupled proteins. The set of links generated by a given method can be filtered by choosing a threshold value, t, for the link-distance. Only pairs with $d'_{i,j} \leq t$ are included in the filtered subsets.

The validity of this approach was tested using the keyword-overlap statistic [3]. Keywords for proteins were downloaded from the genomic annotations provided by the Kyoto Encyclopedia of Genes and Genomes (KEGG) [4]. The average keyword overlap was calculated over the linked proteins in each filtered subset of the data. Then, an average was taken over keyword recovery for the proteins in

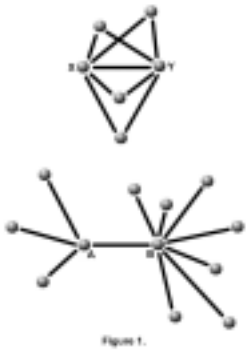

Figure 1: **Discriminating among pairs of functionally-coupled proteins.** In this abstraction, the linked proteins X and Y are also linked to a common set of partners. In contrast, A and B are linked to disparate sets of parters. In general, functional inferences made between pairs like X, Y are more accurate than those made between pairs like A, B.

the subset based on randomly chosen links for each protein. This randomization process was repeated 5000 times to produce an expectation for keyword recovery by random couplings. The ratio of the original average keyword overlap to this random expectation keyword overlap was taken as the signal to noise ratio presented in the results.

The approach was tested on 3 datasets. From the work of Rotstein *et al.*, sets of links from phylogenetic profiles and the rosetta stone method were obtained for *M. tuberculosis* [6]. The third set of links were derived by Marcotte *et al.* from mRNA expression data on *Saccharomyces cerevisiae* [3, 5].

In this work, the link vectors for each of the 3 datasets were binary. Filtered subsets of links were obtained by applying various thresholds, t, and the signal to noise ratio for keyword recovery was computed. These results are presented in Figure 2. While the number of proteins with links changes as a function of the filtering criterion t, little or no change in the number of *annotated* proteins with links were observed in the 3 datasets for $t > 0.3$. Thus, for clarity, Figure 2 displays results for $t \leq 0.3$.

We see from Figure 2 that a dramatic improvement in signal to noise for keyword recovery can be achieved for sets of links obtained by the Rosetta Stone and phylogenetic profile methods. For pairs of proteins with identical patterns of couplings ($d'_{i,j} = 0$), we found that the functional inferences are 40 and 32 times better than random for the Rosetta Stone method and the phylogenetic profile method, respectively. This is compared to 12 and 9 times better than random, respectively, for the raw pools of links generated by those two methods. The improvement for the links derived from mRNA expression profiles was no so large. The raw pool of links provides functional inferences that are 5 times better than random while the best subset of links gives functional inferences that are 9 times better than random. This is likely due to the relatively poor quality of the initial pool of links derived from the mRNA expression data, as noted by Marcotte, *et al.* [3].

References

[1] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O., Assigning protein functions by comparative genome analysis, *Proc Natl Acad Sci U S A.*, 96(8):4285–4288, 1999.

[2] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D., Detecting protein function and protein-protein interactions from genome sequences, *Science*, 285(5428):751–753, 1999.

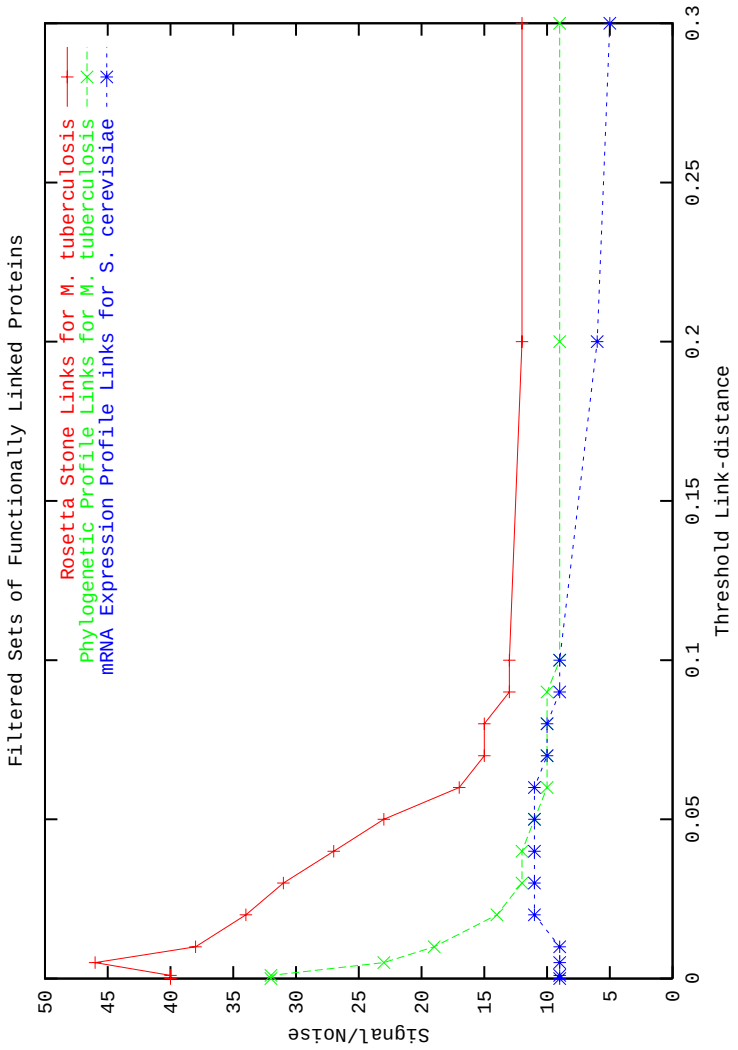

Figure 2: **The precision of functional inferences increases as the link-distance between pairs of coupled proteins decreases.** For 3 different sets of functional couplings, application of the link-distance as a filtering criterion results in discrimination of highly-informative functional inferences.

[3] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D., A combined algorithm for genome-wide prediction of protein function, *Nature*, 402(6757):83–86, 1999.

[4] URL: http://www.genome.ad.jp/kegg/.

[5] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci U S A.*, 95(25):14863–14868, 1998.

[6] Rotstein, S.H., Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D., Computational Assignment of Function to proteins of *Mycobacterium tuberculosis*. (Submitted), 1999.