

Local Connectivity Patterns Improve Missing Link Prediction in Biological Networks

Victor Missirian¹, Vladimir Filkov²

1 Introduction

Elucidating biological network interactions is an important problem in biology. While opening the possibilities for systemic understanding of living systems, at the same time biological networks obtained via large-scale experimental technologies are incomplete, noisy, and often biased. Technologies for identifying/validating individual links in networks exist but their systemic use is tedious and expensive in practice. When a portion of the network is already known computational methods can help to narrow down possibilities for the missing links by considering other data types, like gene expression, and network properties, like their architecture, especially since biological networks can be distinguished and characterized by local connectivity patterns, e.g. connected nodes share many neighbors, network motifs, etc.

Here we report preliminary results from our studies of missing-link prediction in incomplete protein-protein interaction networks and gene regulation networks based on the local connectivity of the networks. We compare three link-scoring mechanisms: expression correlation and two local connectivity measures, number of shared two-paths, and number of shared three paths, in terms of their efficacy of predicting TF-DNA and protein protein interactions in *Saccharomyces cerevisiae*. To test how well we can predict new from existing interactions, we split the real sets of known interactions into a revealed subset and a hidden subset. Given knowledge of only the revealed interactions, we test how well the three different scoring mechanisms can predict the hidden interactions. Our results show that by considering the local connectivity patterns increased the prediction precision for missing links ten fold compared to considering gene expression correlation.

2 Previous Work

The limitations of time series gene expression data has been addressed in the past. Filkov, Skiena, and Zhi [4] record that only approximately 20 percent of known regulatory interactions are associated with strong correlations time series gene expression data sets. Clauset, Moore, and Newman [2] present a hierarchical network model that they used to fit a revealed subset of the known interactions and to predict the remaining hidden interactions, for interaction sets from several domains. On each model network, they revealed a random subset of links, fit their hierarchical models to the revealed links, and then used the fit model to predict the remaining links in that network. We use the same random sampling strategy to evaluate the structural methods that we test. Clauset et. al. also test over many different structural algorithms to predict missing links in a network, including common neighbors, which we make use of in our experiments.

3 Data and Methods

We used a portion (alpha) of a well-known time-series gene expression data set by Spellman et al. [6], encompassing 6117 genes over 18 time points. We used two interaction data sets: the DIP-core [3] protein-protein interaction set and the Harbison et al. TF-DNA interaction set [5]. We filter out all interactions where either gene has no expression profile in the alpha experiment, and we subset the TF/DNA interaction set by pvalue 0.001. After processing, the DIP-core and Harbison TF-DNA interaction sets have 5887 links over 2654 nodes, and 10315 links over 3427 nodes, respectively.

To simulate the situation where a part of the network is known and missing links are to be predicted from it, we hide a varying number of links from the known interaction data sets, from 97.5% down to 50%. In these partially revealed PPI and TF-DNA networks, we score any pair of nodes that are not already connected using three different approaches: Pearson correlation between the nodes' expression profiles, number of two-paths connecting the nodes, number of three paths connecting the nodes, as illustrated in Figure 1. The two- and three-paths are made of only revealed links. We predict the top ranked links.

¹Department of Computer Science, University of California at Davis, CA, USA. E-mail: vmissirian@ucdavis.edu

²Department of Computer Science, University of California at Davis, CA, USA. E-mail: filkov@cs.ucdavis.edu

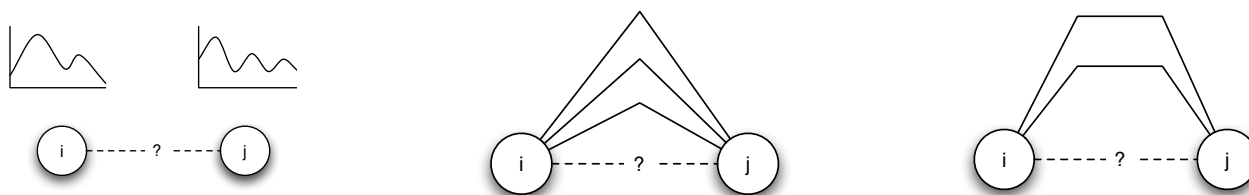


Figure 1: The three mechanisms for scoring missing links. Link (i, j) is scored based on, from left to right: the Pearson correlation of the gene expression profiles at its nodes, the number of distinct two paths connecting its nodes, and the number of distinct three paths connecting its nodes.

4 Results and Discussion

On each measure, for each revealed (i.e. not hidden) interaction set size, on each of 100 iterations, we sample a set of revealed interactions and use it to predict the set of hidden interactions. The number of gene pairs that we predict is equal to the number of hidden interactions, $N = total(1 - \%revealed)$. The results are given in Table 1. We included the "rand" measure as a control. For rand, we select the set of genes incident to any revealed interaction to create a subgraph, then randomly choose N pairs from this subgraph, and compute the expected number of these gene pairs that would be in the hidden interactions set.

Connectivity based ranking generally outperform the expression measure. Both two-paths and three-paths show dramatic improvement as the percentage of revealed interactions increases. In the future, we plan to come up with a comprehensive measure of combining the structural and expression-based measures, and we plan to include more types of local connectivity patterns.

% revealed	DIP-core				Harbison et al. TF/DNA			
	rand	ge	p_2	p_3	rand	ge	p_2	p_3
2.5	0.647	0.887	0.717	0.664	2.167	2.384	2.153	2.210
5	0.545	1.037	0.885	0.580	1.380	1.569	1.400	1.610
10	0.401	1.148	1.785	0.741	0.785	0.941	1.137	2.784
20	0.263	1.091	5.566	2.805	0.400	0.578	1.336	6.692
50	0.103	0.871	10.211	10.977	0.122	0.314	2.425	16.349

Table 1: Missing link prediction precision, (%), for the three different methods: expression correlation, ge , two-paths, p_2 , and three-paths, p_3 . Rand is the expected precision if links were predicted randomly.

References

- [1] Cho, R.J., et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.* 2:6573.
- [2] Clauset, A., Moore, C., and Newman, M.E.J. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* 453:98.
- [3] Deane, C.M., Salwinski, L., Xenarios, I., and Eisenberg, D. 2002. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics*, 1:349356.
- [4] Filkov, V., Skiena, S., and Zhi, Z. 2001. Analysis techniques for microarray time-series data. *RECOMB*, pp. 124-131.
- [5] Harbison, C.T., et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431:99-104.
- [6] Spellman, P.T., et al. 1998. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Mol. Biol. Cell*, 9:12:3273-3297.