

# Annotating Genes with unknown Function by a Systems Approach

Frank Emmert-Streib<sup>1,2,\*</sup>, Lin Chen<sup>1</sup> and John D. Storey<sup>1,2</sup>

1. Department of Biostatistics, University of Washington, Seattle, WA 98195-5065, USA

2. Department of Genome Sciences, University of Washington, Seattle, WA 98195-5065, USA

\*E-mail: fes99@u.washington.edu

## Abstract

Making predictions about the molecular function unknown genes might be involved in is an important but difficult problem. Early methods based on expression data from microarray experiments used various clustering methods to annotate genes with respect to their co-appearance in clusters. Despite the simplicity of the hypothesized principle these methods were quite successful. We introduce a new method to annotate genes with unknown function based on a systems approach. It has been argued that using shortest paths that occur in gene or complex networks are helpful to study functional or communication relations among network nodes [1, 4, 5, 6]. Given a gene network representing causal gene-gene interactions we hypothesis that the ranking of similarity values between genes pairs is correlated to the frequency of co-occurrences of gene pairs found in shortest paths within the gene network. For a recently inferred transcriptional regulatory network of *Saccharomyces cerevisiae* [2] we demonstrate, first, that our underlying hypothesis is correct in a statistical sense by measuring the pairwise similarity values between genes in terms of the semantic similarity regarding the gene ontology entries of annotated genes [3] and, second, make predictions for the function of unannotated genes.

## References

- [1] Arita, M., The metabolic world of *Escherichia coli* is not small, PNAS, 100(6) (2004) 1543-1547.
- [2] Chen L, Emmert-Streib F, Storey JD (2007, submitted.) Harnessing naturally randomized transcription to infer regulatory relationships among genes .
- [3] Lord, P., Stevens, R, Brass, A. and Goble, C., Semantic similarity measures as tools for exploring the Gene Ontology. In Pacific Symposium on Biocomputing, 8 (2003) 601612.
- [4] Rahman, S. A., Schombur, D., Observing local and global properties of metabolic pathways: 'load point' and 'choke points' in the metabolic networks, Bioinformatics, 22(14) (2006) 1767-1774.
- [5] Trusina, A., Rosvall, M. and Sneppen, K., Communication Boundaries in Networks, Phys. Rev. Let., 94 (2005) 238701.
- [6] Zhou, X. Kao, M-C J., Wong, W. H., Transitive functional annotation by shortest-path analysis of gene expression data, PNAS, 99(20) (2002) 12783-12788.