

TWAIN: a new tool for parallel gene finding

Steven Salzberg
salzberg@tigr.org

*Senior Director of Bioinformatics,
The Institute for Genomic Research,
9712 Medical Center Dr.,
Rockville, MD 20878*

*Research Professor,
Department of Computer Science,
Johns Hopkins University,
Baltimore, MD 21218.*

One of the most exciting recent developments in genomics is the availability of complete genome sequences from two or more closely related species. These closely related organisms can offer numerous valuable insights into functional differences, evolution, gene regulation, and other aspects of biology. For bioinformatics, these sequences offer a new and powerful opportunity: the chance to use sequence conservation to find new genes, create more accurate gene models, and find regulatory elements. TWAIN - our new parallel gene finder - continues the series of computational tools that we have recently developed in order to achieve better gene models as well as better alignments of two similar genomes. Using the probabilistic framework of generalized pair HMMs, the goal of our efforts is to develop a working two-organism gene finder that will be highly accurate for many different pairs of species while eliminating some of the restrictions imposed by other similar tools.

Joint work with: Mihaela Pertea and William Majoros

Untangling graphs: structure-based denoising of protein-protein interaction networks.

Brendan Frey and Quad Morris

*{frey, quaid}@psi.toronto.edu
Department of Electrical and Computer Engineering,
University of Toronto,
Toronto, Ontario M5S 3G4, Canada*

Many types of real-world networks (e.g., the WWW and Internet, metabolic networks, protein-protein interaction networks) share a very similar connectivity structure: most nodes have a small degree (i.e., are connected to very few other nodes) whereas a few nodes have a very large degree. This regularity is important because it can be used to denoise networks constructed using noisy edge-appearance data.

In this talk, I will describe a probabilistic generative model of edge-appearance data that can be used to untangle graphs hidden in a set of noisy edge detections. Exact inference in this model is generally intractable; however, I will describe a sum-product algorithm for efficient approximate inference that works when the structure priors are degree-based. These priors can model a rich class of distributions over graphs.

I will also describe an application of the new model to denoising protein-protein interaction networks. Though current experimental methods for detecting interactions are noisy, the connectivity structure of the network of false detections is very different from that of the true interaction network. I will describe how to use a small set of trusted interactions to learn structure priors that can be used to distinguish true and false detections in a much larger set of untrusted interactions.

Inferring molecular interaction networks

Tommi Jaakkola

tommi@ai.mit.edu

Associate Professor of Computer Science

Sloan Research Fellow

MIT Artificial Intelligence Laboratory

200 Technology Square, Cambridge, MA 02139

The available high-throughput data sources pertaining to transcriptional regulation such as expression arrays, protein-DNA binding data, and gene deletions are effective and limited in different ways. The integration of such complementary constraints is crucial for accurate reconstruction. I will discuss a new framework for inferring models of transcriptional regulation from heterogeneous data source. The models, which we call physical models, are based on verifiable molecular properties and expressed in terms of annotated molecular interaction graphs. The effects of perturbations such as gene deletions in these models are associated with (restricted) molecular cascades, paths in the annotated graph. By tying measurements to the properties of the molecular networks in a causal manner we can appropriately constrain the set of possible underlying molecular interpretations from diverse data sources. The soft constraints from the data sources can be expressed as potentials in a factor graph, where each configuration of variables specify an annotated molecular interaction graph. The reconstruction problem is reduced to an inference problem in the factor graph and can be solved approximately with local propagation algorithms such as the max-product. I will discuss our results in the context of the yeast pheromone response pathway as well as genomic scale analysis. I will also illustrate various extensions of the basic approach such as automated experiment design currently in use, along with modeling coordinated regulation.

Joint work with: Chen-Hsiang Yeang

Predicting protein function from protein/protein interaction data

Stanley Letovsky

sletovsky@aol.com

Bioinformatics Program and Department of Biomedical Engineering,

Boston University,

44 Cummington St., Boston, MA 02215, USA

The development of experimental methods for genome scale analysis of molecular interaction networks has made possible new approaches to inferring protein function. This talk will describe a method of assigning functions based on a probabilistic analysis of graph neighborhoods in a protein-protein interaction network. The method exploits the fact that graph neighbors are more likely to share functions than nodes which are not neighbors. A binomial model of local neighbor function labeling probability is combined with a Markov random field propagation algorithm to assign function probabilities for proteins in the network. The method has been applied to a protein-protein interaction dataset for the yeast *Saccharomyces cerevisiae* using the Gene Ontology (GO) terms as function labels. The method reconstructed known GO term assignments with high precision, and produced putative GO assignments to 320 proteins that currently lack GO annotation, which represents about 10% of the unlabeled proteins in *S. cerevisiae*.