Statistical discovery of signaling pathways from an ensemble of weakly informative data sources

Edoardo M. Airoldi, Florian Markowetz*, David M. Blei & Olga G. Troyanskaya*
Computer Science Department & Lewis-Sigler Institute for Integrative Genomics, Princeton University

Brief introduction. Signaling pathways are complex biological mechanisms. For instance, consider the Mitogen-activated protein (MAP) kinase pathway, available on Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2006). It involves 55 genes/proteins, it sub-divides into four interconnected modules (pheromone response, high osmolarity glycerol response, hypotonic shock response, and starvation response), and their circuitry is currently instantiated by 104 directed interactions with various meanings (e.g. activation, phosphorylation, binding, and inhibition)

Historically, the inner workings of each pathway have been put together piece by piece, with the aid of experimental evidence. Searching PubMed, we found more than 400 articles that specifically mention the MAP kinase pathway, in the 90s—before the boom of high-throughput technologies such as microarrays. Evidence in support of the role played by a gene or protein in the MAP kinase pathway, or in support of a specific interactions between a pair of genes or proteins, has been screened by a group of experts who ultimately hand-curated the pathway and its inner mechanism, as we currently understand it.

High-throughput technologies held the promise of supporting the automated recovery of pathway-specific signal. Their intrinsic noise, however, has hindered progress in the community. Successful efforts to date have been able to *confirm* aspects of known biology—at great cost and labor. Discovery of novel mechanisms guided by computational predictions has arguably not happened yet, at any reasonable scale.

The problem lies as much in the methods as it lies in the data (see Table 1). Many past efforts aimed at inferring pathways from microarrays without paying much attention to the other kinds of data available to the community, or neglected to consider the empirical evidence that biologists would find convincing enough (in support of candidate signaling pathway relations) to encourage testing in the lab.

	p-values (KS test)			
Data set	ρ	partial ρ	shrunk ρ	max R^2
Brem et al. (2005)	0.005892	0.5690550	0.1395886	0.3422675
Roberts et al. (2000)	0.530624	0.5180326	0.9481925	0.4210444
Hughes et al. (2000)	0.269332	0.6280315	0.2999146	0.3929368
Gasch et al. (2000)	0.720287	0.4224010	0.6875346	0.3815168

Table 1: Strength of the statistical signal about the pheromone response (MAP-K) pathway in four microarray studies. P-values correspond to a Kolmogorov-Smirnov test that suggests how different the distributions of various statistics are, between those genes that interacting in the MAP kinase pathway and those that don't. We evaluated correlation coefficients ρ, partial correlation coefficients and shrunken partial correlation coefficients (Schäfer and Strimmer, 2005), as well as the maximum R^2 (over arrays in each data set) of a relational regression model—attributes of neighboring genes are used as regressors, in turn.

Our approach. We propose a probabilistic model that integrates high-throughput interactions and multivariate attribute data with low-throughput knock-out experiments, and existing hand-curated biological

knowledge-bases. The data sources we set out to integrate covers:

1. Data on gene/protein relationships give us direct information on the *edges* in the pathway diagram. In this study, we integrate microarray gene expression data, transcription factor location data and protein-protein interaction data.
2. Data on gene/protein characteristics contain descriptions of genes/proteins and tell us how they are built, or other characteristics and features. This information is about the *nodes* in the pathway diagram and only indirectly about the edges. Here, we consider two types of node information: the Gene Ontology (Ashburner et al., 2000) annotation of proteins and their domain architecture (Finn et al., 2006; Letunic et al., 2006).
3. Functional data shows how the pathway reacts to experimental stimulations and perturbations. It is not information on single edges or nodes in the pathway, but tells us about the information flow along *paths* in the graph. We integrate into our model phenotypic profiles of gene knockout experiments.

A pathway is a graph G on a node set $\mathcal{V}$ of size n. Our goal is to infer an edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ and edge attributes from data. We define a model by a set of variables for edge existence, X_{ij} $(i, j \in \mathcal{V})$, and variables for edge attributes, like direction or strength of regulation.

Models for edge attributes. For microarray gene expression data, we base our model on the gene-by-gene (partial) correlation matrix. The distribution of (partial) correlations is known analytically for $X_{ij} = 0$ (Hotelling, 1953), and for $X_{ij} = 1$ we use a uniform distribution (Schäfer and Strimmer, 2005). TF-DNA binding data is given as p-values for a TF binding to DNA. Under the null distribution ($X_e = 0$) the p-values are known to be uniformly distributed, while we expect to see a small p-value if there is an edge ($X_e = 1$), which we model by Beta$(a, 1)$ (Pounds and Morris, 2003). Protein-Protein interaction data is binary. To define the likelihood we use error rates described in the literature (Chiang et al., 2007).

Models for node attributes. In relational regression, attribute values at neighboring nodes serve as regressors for attribute values of each gene in turn. For attribute k, for instance, $X_{ik} = \alpha_k + \sum_{j \in N_i} \beta_k X_{jk} + \epsilon$, where the set of nodes $\{j \in N_i\}$ is computed given G. (This setting can be generalized to other linear models.) In our model, G is only partially observable and gets updated at each iteration of the MCMC sampler.

Models for functional data. When modeling functional data, knock-outs are considered IID probes to the biological system that are quantitatively described by genome-wide transcriptional responses. The model is a simple mixture, $\pi\ f_0(x) + (1 - \pi)\ f_1(x)$, that captures whether each gene is responding or not to the stimulus given by the knock-out experiment. This model induces hard constraints on the pathway G that is primary objective of the inference process. If the log-odds of a response are favorable then the pathway G must include a path between the knock-out and the responding genes, otherwise a path must not exist.

Integrating prior knowledge. Prior knowledge about a pathway is given by a collection of edge-specific random variables that encode the prior probability of the edge being present. Specifically, such a prior probability is encoded by a uniform distribution on the interval $(0, 1)$ when we have no information about a relation, whereas it is encoded by a Beta$(a, 1)$ when we do have information about the presence of a relation.

Related work. The main competing approach to ours is the *Physical Network Model* by Yeang et al. (2004). They use PPI, TF-DNA and KO data and combine them in a likelihood function which is maximized by a max-product algorithm. In contrast to their approach, (1.) we formulate a model that is more suitable for large-scale applications, (2.) we include node information and thus are the first to integrate all three types of data, (3.) we use available prior information from pathway databases.

References

M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, May 2000.

T Chiang, D Scholtens, D Sarkar, R Gentleman, and W Huber. Coverage and error models of protein-protein interaction data by directed graph analysis. *Genome Biol.*, 8(9):R186, 2007.

R.D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer, and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Research*, 34:D247–D251, 2006.

Harold Hotelling. New light on the correlation coefficient and its transforms. *J. R. Statist. Soc. B*, 15: 193–232, 1953.

M. Kanehisa, S. Goto, M. Hattori, Aoki-Kinoshita, Itoh K.F., Kawashima M., Katayama S., M. T., Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, 34:D354–357, 2006.

I Letunic, RR Copley, B Pils, S Pinkert, J Schultz, and P Bork. Smart 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, 34:D257–60, 2006.

S Pounds and SW Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10): 1236–42, 2003.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4 (1):art.32, 2005.

Chen-Hsiang Yeang, Trey Ideker, and Tommi Jaakkola. Physical network models. *Journal of Computational Biology*, 11(2):243 – 262, 2004.