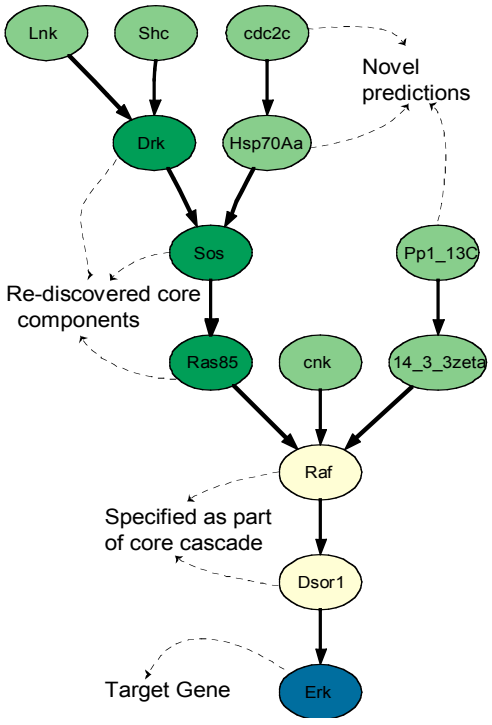
Influence Flow: Integrating Pathway-specific RNAi Data with Protein Interaction Networks

Rohit Singh[a] Bonnie Berger[a,b]
[a]*Computer Science and Artificial Intelligence Lab., MIT* [b]*Dept. of Mathematics, MIT (Corresp. author: bab@mit.edu)*

Understanding the detailed structure of signaling sub-systems is a major biological challenge. Often, the core cascade of the sub-system is well-understood and our goal is to ascertain the other genes/proteins involved and the corresponding network topology (e.g., the MAP Kinase signaling network). Towards this goal, we describe *influence flow*– a novel method for generating high-confidence hypotheses about a specific signaling network's topology. These hypotheses may then be used to direct further experiments.

Our method combines pathway-specific RNA interference (RNAi) data with genome-wide protein interaction networks. The RNAi data is generated from a functional genomic screen of a specific signaling pathway. These screens work as follows: a known end-effector gene of the pathway is chosen as the reporter gene (e.g., *Erk* in the MAPK pathway [1]). Every other gene in the genome is systematically knocked-down using RNAi and the effect on the reporter is measured. The experiment produces a list of genes (*hits*) that significantly influence the reporter and, for each hit, a score indicating the relative strength of its influence. The second input to our method is genome-wide protein-protein interaction (PPI) data (protein-DNA interactions can also be included). To minimize false negatives in PPI data, we use computational methods to predict new PPIs from other data sources and from PPI data in other species [2]. To mitigate the impact of false positives in the data, we can estimate confidence values for each edge in the PPI network and take these into account during our computations.



A part of the inferred network for the MAP Kinase subsystem

Given these inputs, we search for a *directed acyclic protein network* such that all its edges are consistent with the input PPI data, all its nodes and their relative placement is consistent with the RNAi data. Furthermore, we require that the output topology reflect the following biological intuition: for most proteins *not* in the core cascade, their influence on the end-effector protein is transmitted via the core cascade. Specifically, given an input PPI network N, an RNAi reporter gene T, the corresponding list of RNAi hits $L = \{i\}$ and their scores $S = \{s_i\}$, our desired network G must satisfy the following conditions A1-A4 and be optimal under condition A5:

A1. All the nodes in G are present as RNAi hits (i.e. in L).

A2. Each edge in G is directed. Also, each directed arc $a \rightarrow b$ in G is either in the core cascade or corresponds to an edge a — b in N.

A3. Every node in G has a directed path to the target gene T.

A4. Nodes closer to T should have higher RNAi scores. If G has an arc $a \rightarrow b$ that is not part of the core cascade, then $s_a \leq s_b$.

A5. For most nodes, the path(s) towards T should be routed through the core cascade, i.e., the last segment of the path(s) should only contain edges from the known core cascade.

The optimal network G must satisfy A1-A4 and have the maximum number of nodes that satisfy A5. We compute this solution by formulating an integer linear program (ILP), borrowing ideas from the multi-commodity network flow literature [3].

The constraints A1-A5 are quite simple; yet, the inferred influence flow network contains surprisingly plausible hypotheses. When supplied only a part of the known MAPK cascade (in fly), our method can successfully discover the other known components. The core MAPK cascade is *Drk* → *Sos* → *Ras85* → *Raf* → *Dsor1* → *Erk*. For test purposes, we specified to our method only a truncated cascade consisting of *Raf*, *Dsor1*, and *Erk*. Our method was able to retrieve all the remaining core nodes (*Drk, Sos, Ras85*). Furthermore, *Ras85* and *Sos* were two of the three nodes with the most flow in the ILP solution (in our method this quantity is a proxy for the node's importance in the solution). Experiments to test novel connections proposed by the influence network are in progress.

[1] A. Friedman, and N. Perrimon. *A functional RNAi screen for regulators of receptor tyrosine…* Nature, 444(7116), pp 230-4, 2006.
[2] R. Singh, J. Xu, and B. Berger. *Pairwise global alignment of protein interaction networks.* RECOMB 2007. To appear.
[3] M.O. Ball *et al. Handbook in Operations Res. and Management Sci.: Network Models*. Elsevier, 1995.