Mechanistic Bayesian Networks
Modeling the Epithelial-Mesenchymal Transition Interactome

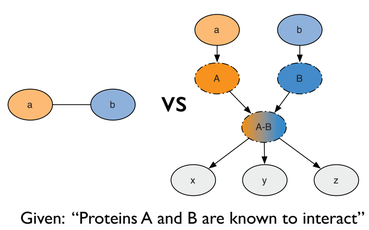
Abhik Shah, Venkat Keshamouni, Peter Woolf
University of Michigan, Ann Arbor, MI

Abstract
We present Mechanistic Bayesian Networks (MBN), a strategy for modeling biological systems using an integration of mechanistic knowledge with high-throughput experimental observations. We apply this method to elucidate the important interactions underlying the Epithelial-Mesenchymal transition. EMT is a process in which fully-differentiated epithelial cells undergo a change to a mesenchymal phenotype. It is a common process during development and implicated in early steps of cancer metastasis. We query the Michigan Molecular Interactions database (MiMI) and mirBase, a database of computationally predicted miRNA-target interactions, to create initial interaction networks. To reduce the visual complexity of the networks and to highlight causally central interactions, we refine the networks by applying mRNA and miRNA expression data.

We have a lot of existing knowledge about biological systems (in databases and literature) but it tends to be non-specific and often incorrect. An interaction observed in another tissue and under other conditions may not be valid in our system of interest. Experimental data, on the other hand, is specific to our system of interest but rarely provides data suitable for modeling specific mechanisms. By integrating these two sources of information, we should be able to learn models with detailed mechanisms that are specific ti our system of interest.

Mechanistic Bayesian Networks
MBN is a strategy for modeling biological systems with Bayesian networks while being faithful to known mechanisms. It consists of the following principles:

1. Don't conflate entities of different types (mRNA, protein) into one node.
2. Use hidden nodes to represent unobserved entities.
3. Don't accept prior knowledge indiscriminately.
4. Asses mechanisms by their downstream effect rather than upstream precursors.



Given: "Proteins A and B are known to interact"

We evaluate the interaction between proteins A and B, not by testing for correaltion between the mRNA profiles for A and B but by the interaction's downstream effects on other genes. This indicates the likelihood of the interaction and its causal centrality to the rest of the dataset. While we don't observe the protein-protein interaction in our dataset, we can observe its downstream effects.

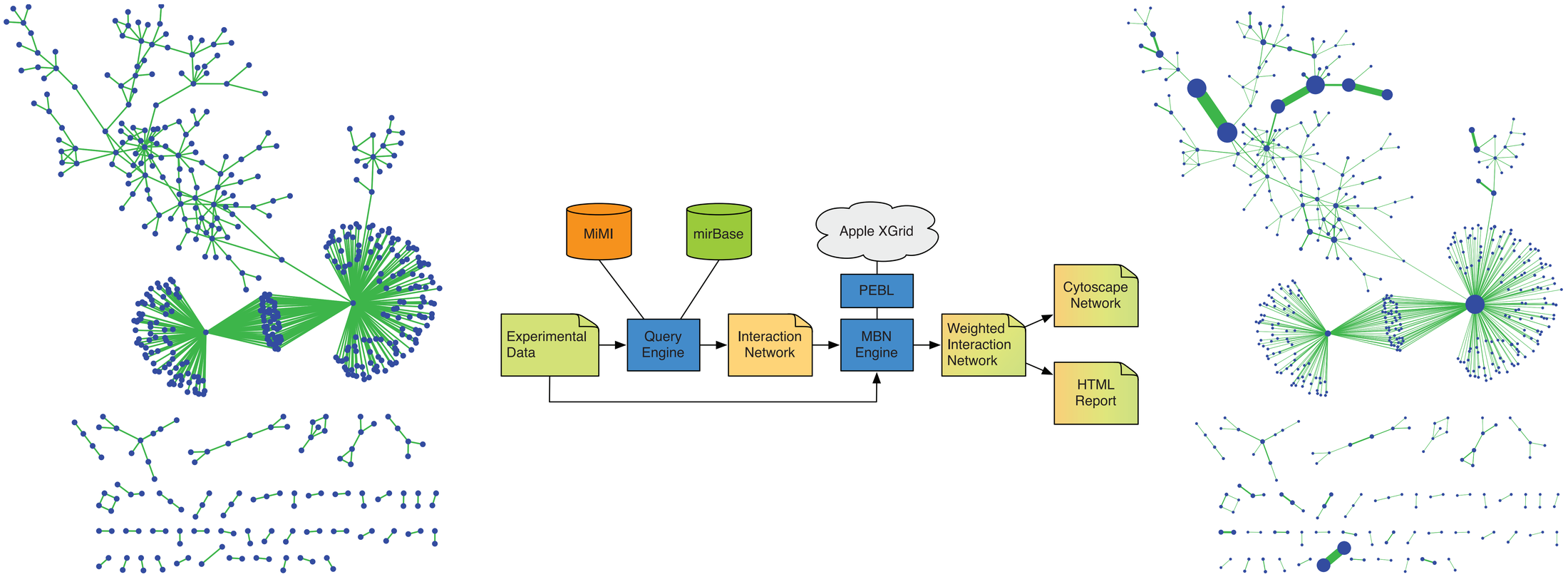
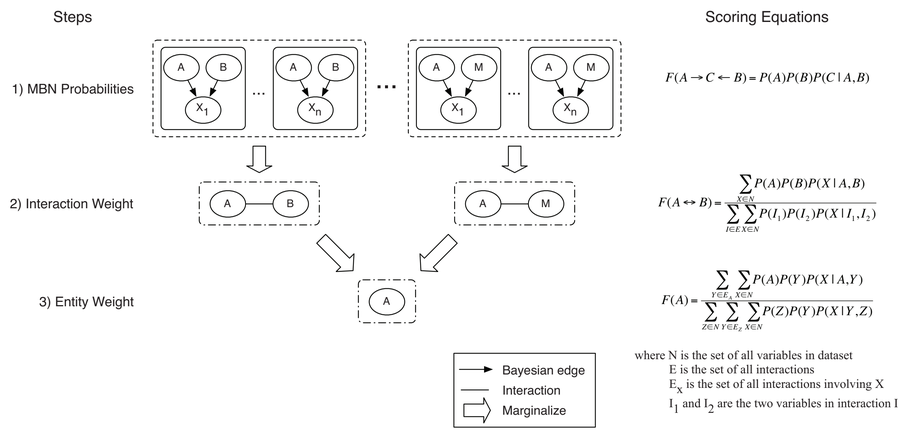


Fig 1. MBN analysis pipeline and protein-protein interaction networks, before and after using experimental data. For this analysis, the experimental data consisted of mRNA and miRNA expression data and we queried MiMI and mirBase for protein-protein and miRNA-target interactions respectively. The analysis was done using the Python Environment for Bayesian Learning (PEBL) and executed over an Apple XGrid using PEBL's distributed computing features. Note: miRNA-target and combined networks not shown.

MBN Scoring



Scoring Equations

$$F(A \rightarrow C \leftarrow B) = P(A)P(B)P(C \mid A,B)$$

$$F(A \leftrightarrow B) = \frac{\sum_{X \in N} P(A)P(B)P(X \mid A,B)}{\sum_{I_1 \in E_X} \sum_{X \in N} P(I_1)P(I_2)P(X \mid I_1,I_2)}$$

$$F(A) = \frac{\sum_{Y \in E_X} \sum_{X \in N} P(A)P(Y)P(X \mid A,Y)}{\sum_{Z \in N} \sum_{Y \in E_Z} \sum_{X \in N} P(Z)P(Y)P(X \mid Y,Z)}$$

where N is the set of all variables in dataset
E is the set of all interactions
E_X is the set of all interactions involving X
I_1 and I_2 are the two variables in interaction I

We use a novel criteria to evaluate a given putative interaction using experimental data. Rather than test for correlation between upstream precursors (like mRNA profiles for a protein-protein interactions), we look for downstream effects. To score an interaction between proteins A and B, for example, we model A and B as the parent nodes for all other variables in the dataset, use regular Bayesian scoring methods and then marginalize to determine the interaction weight. This score indicates both the likelihood of the interaction being valid and its causal centrality with respect to the rest of the dataset. We further marginalize the interaction weights to calculate the weight for each entity.

Results
The results of the analysis are exported as a set of Cytoscape networks and html pages that show the relationship underlying each interaction and allow for a drilldown exploration of the dataset.

We identified many interactions that appear to be central to the EMT process and are doing further validation. Many of the interactions involve cell adhesion and motility and some have been implicated in cancers. Interestingly, of the two network hubs, only one scores well with the experimental data and neither is involved in high-scoring interactions whereas some of the highest scoring interactions involve proteins with few interactions. This shows the utility of integrating experimental data with interaction networks -- network topology, while helpful, doesn't provide the full picture.

Python Environment for Bayesian Learning
This analysis was conducted using PEBL, our open-source and peer-reviewed software for Bayesian analysis. PEBL offers a set of features unmatched by other open or proprietary packages:

- Learns with observational and interventional data (eg. microarray after knockout)
- Handles missing values and hidden variables
- Provides greedy hill-climbing and simulated annealing learners
- Facilities for running on Apple's XGrid, IPython clusters and Amazon's Elastic Cloud Computing platform (EC2).
- Written in python but core routines use optimized matrix operations and custom C extension modules

Available at http://pebl-project.googlecode.com

Acknowledgements
This work was supported by National Institutes of Health: Grant #U54 DA021519