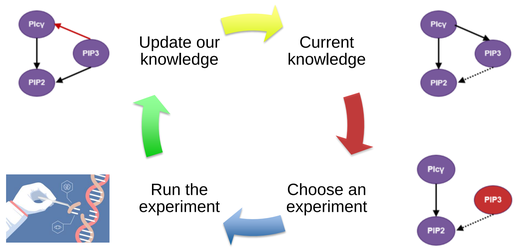
Discovery of Gene Regulatory Networks Using Adaptively Selected Perturbation Experiments

Thesis Advisor: Dr. Jeffrey W. Miller
Collaborators: Mair Lab, Dept. of Genetics and Complex Diseases

Michele Zemplenyi, PhD Candidate
Harvard University, Dept. of Biostatistics
mzemplenyi@g.harvard.edu

Objective

To build an adaptive learning algorithm for inferring gene networks by iterating between experimentation and analysis.



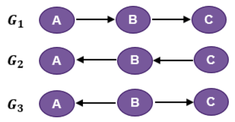
Background

Representing the network

- Causal Bayesian network (directed acyclic graph)
- Nodes: random variables representing gene expression values
- Edges: causal, regulatory relationships

Markov equivalence classes

- Networks that represent the same set of dependencies and conditional independencies, e.g.



Benefit of perturbation data

- Observational data alone cannot distinguish among networks in the same equivalence class.
- Perturbing node A affects descendants of A, allowing us to distinguish $\{G_1\}$ from $\{G_2, G_3\}$.

Entropy as a measure of uncertainty

- Three possible edge relationships for two nodes: $A \rightarrow B$, $A \leftarrow B$, and $A \perp B$.

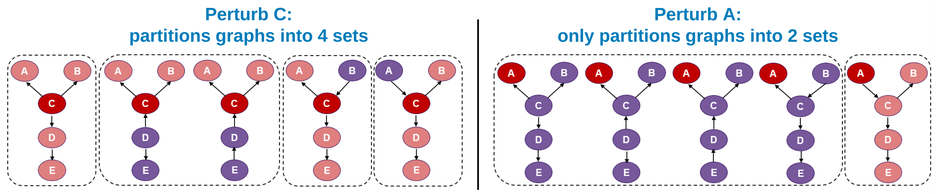
- Tong and Koller (2001) define edge entropy as:

$$H(A \leftrightarrow B) = -P(A \rightarrow B)\ log\ P(A \rightarrow B) - P(A \leftarrow B)\ log\ P(A \leftarrow B) - P(A \perp B)\ log\ P(A \perp B)$$

- The larger this entropy, the less certain we are about the relationship between A and B.

Selection Criterion

Motivation: perturb the node whose descendant sets are maximally different.



- Sampling variability leads to uncertainty in the descendant set partitions.
- Capture uncertainty of whether node B is a descendant of A, $d(A)$, via:

$$H\big(B \in d(A)\big) = -P\big(B \in d(A)\big) * log\ P\big(B \in d(A)\big) - \big([1 - P\big(B \in d(A)\big)] * log[1 - P\big(B \in d(A)\big)]\big)$$

$P\big(B \in d(A)\big)$ **is calculated empirically from graphs sampled from the posterior on graphs G given data D:**

$$P(B \in d(A)|D) = \sum_{[B \in d(A)] \in G} P(G|D)$$

Descendant entropy: captures the overall uncertainty in the descendants of A.

$H(A) = \sum_j H\big(j \in d(A)\big)$ **for all nodes j in $G \backslash A$**

- Greatest information gain achieved by perturbing the node with the largest descendant entropy.

- "Adaptive descendant" method uses descendant entropy as the criterion; substitute $d(A)$ with $c(A)$ to compare with the "adaptive child" method.

Adaptive Learning Algorithm

1. Sample n_{obs} instances from the ground truth Bayesian network.

2. Use MCMC to obtain posterior samples from $P(G \mid D)$.
   - Metropolis-Hastings proposals with an edge addition, deletion, or reversal.

3. Calculate $P\big(B \in d(A)\big)$ and $H\big(B \in d(A)\big)$ empirically from the sampled graphs for all node pairs.

4. Check the stop criterion. Continue performing experiments until either:
   - minimum desired entropy is achieved
   - maximum number of allowed experiments is reached

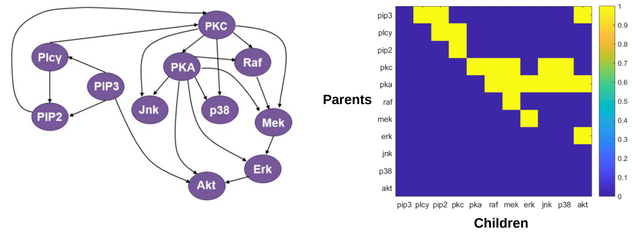
5. Select $arg\ max_{\{A \in T\}} H(A)$ as the next node to perturb.
   - T: set of nodes eligible for interventions.

6. Generate n_{pert} instances from the Bayesian network.

7. Combine the new data with the existing data and return to step 2.

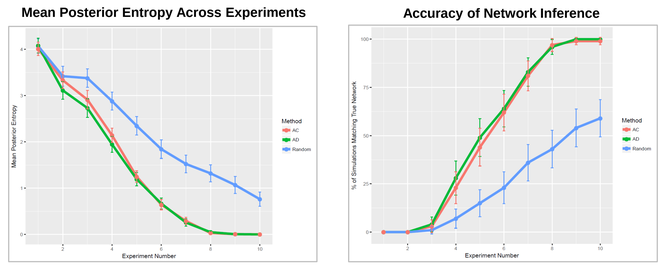
Results

Protein-signaling network used to simulate gene expression data, adapted from Sachs et al. (2005):

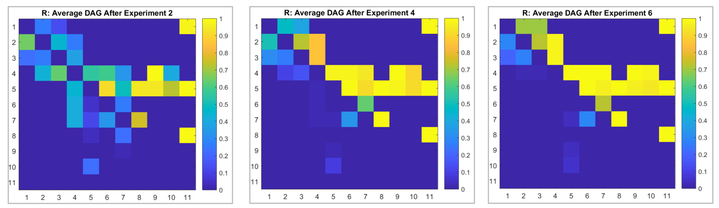


The adaptive descendant (AD) and adaptive child (AC) methods outperform random (R) node selection, as evidenced by:

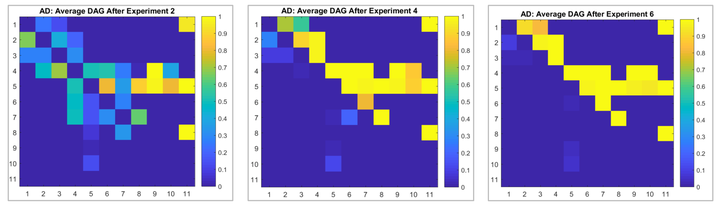
(1) faster rates of entropy reduction, and
(2) greater accuracy in inferring the true causal network.



Random Node Selection



Adaptive Descendant Node Selection



References

1. Tong, S., Koller, D. Active learning for structure in Bayesian networks. *IJCAI* 2001, p863-869.
2. Eaton, D., Murphy, K. Bayesian structure learning using dynamic programming and MCMC. *UAI* 2007, p101-108.
3. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, Nolan, G. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 2005. 308:523-528.
4. Li, G., Leong, TY. Active learning for causal Bayesian network structure with non-symmetrical entropy. *PAKDD* 2009, p290-301.