Bayesian Inference of the Phenotype-Genotype Network Topology

R.S. Hageman, M.S. Leduc, B. Paigen, R. Korstanje, G.A. Churchill

The Jackson Laboratory, Bar Harbor, Maine 04609 USA

Abstract

Understanding the complex network of interactions between genotypes, genes and clinical phenotypes is critical to unraveling the pathology of disease. Computational methods to construct causal regulatory systems for mammalian systems can be used to generate hypothesis about these interactions, and predict the effects of perturbations (genetic and environmental) to network components. Expression Quantitative Trait Loci (eQTL) data consists of genotypes at markers across the genome and phenotypes (both clinical phenotypes and gene expression). EQTL data from segregating populations is an ideal system for inferring causality.

We propose a Bayesian framework for the inference of the genotype-phenotype network topology for segregating populations using eQTL data. Structural priors are implemented to penalize the graph density and can be used to integrate prior biological knowledge about the relationship between variables. An efficient Markov Chain Monte Carlo method is proposed to search across the model space. The result is not a single network, but an ensemble of highly probable networks that provides a measure of confidence in the estimated connections. We have applied our method to eQTL data from the kidneys of an MRLxSM intercross mouse population, and predicted a feedback loop in the canonical Renin-Angiotensin System (RAAS) pathway.

Bayesian Networks

The data is a mixture of continuous (phenotypes) and discrete (genotypes) random variables:

$$D = \left\{ \underbrace{X_1, X_2, \ldots, X_n}_{\text{phenotypes}}, \ldots, \underbrace{Q_1, Q_2, \ldots, Q_m}_{\text{genotypes}} \right\}.$$

The graph G is a **Directed Acyclic Graph** (DAG) that obeys the *Markov Property*.

The posterior density:

$$P(G \mid D) \propto P(D \mid G) P(G),$$

where $P(G)$ is the prior on the graph structure.

The likelihood density:

$$P(D \mid G) \propto P(D \mid \theta, G) P(\theta \mid G) d\theta,$$

where $P(\theta \mid G)$ is the prior on the parameters.

We define a **local family** as a continuous child with mixed parents (Figure), and describe their relationships with the hierarchical regression model:

$$y_i \sim N\left(\mu + \sum_{j=1}^{s}\left(\sum_{g=1}^{3} \gamma_j Q_{j,g[i]}\right) + \sum_{l=1}^{t} \phi X_l, \sigma_y^2\right),$$

$$\gamma_j \sim N(0, \sigma_\gamma^2),$$

$$\phi_l \sim N(0, \sigma_\phi^2).$$

For simplicity, we let $\beta = \{\mu, \gamma, \phi\}$, and rewrite the model in the form:

$$y_i = \beta_0 + \ldots + \beta_{s-1} Q_{A,i} + \beta_s Q_{B,i} + \beta_{s+1} Q_{H,i} + \ldots + \beta_{t-1} X_i' + \varepsilon_i.$$

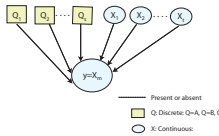

Figure 1: An example of a local family consisting of a continuous child $y=X_m$ with mixed parents $\pi_G(y) = \{Q_1, Q_2, \ldots Q_s, \ldots X_1, X_2, X_3, \ldots X_t\}$.

Structural and Parameter Priors

Structural Priors:
Restrict the model space and facilitate the MCMC search procedure.
Defining B as a *Gaussian belief network*, the prior is an energy function embedded in a Gibbs distribution:

$$P(G) \alpha \exp(-\tau \cdot e(G)) \qquad \text{where} \qquad e(G) = \sum_{i,j=1}^{N} \left| B_{i,j} - G_{i,j} \right|.$$

Parameter Priors:
Conditionally conjugate priors on the parameters yield a closed-form likelihood.
For a local model G the joint distribution of the parameters is

$$P(\beta, \sigma^2) \sim P(\beta \mid \sigma^2) P(\sigma^2)$$

where

$$P(\beta \mid \sigma^2) \sim N(\mu_{pr}, c^2 \Sigma_{pr})$$

$$P(\sigma^2) \sim IG(a, b).$$

MCMC Proposals

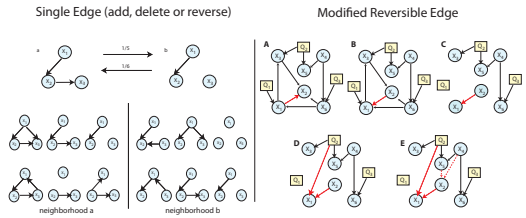

Figure 2: We utilize a modified Metropolis-Hastings algorithm with two types of proposals, *Single Edge Proposals* [1] and Reversible-Edge proposals [2]. In Single edge proposals, an edge is either removed, added or reversed (known to be slow in convergence with poor mixing) (**Left**). Our modified reversible edge proposal reverses an edge between two continuous nodes, orphans them, and then re-samples for the new parent sets (**Right**).

[1] Madigan D, York J (1995) Bayesian graphical models for discrete data. International Statistics Review 63: 215-232.
[2] Grzegorczyk M, Husmeier D (2008) Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. Machine Learning 71: 265-305.

Renal RAAS Pathway: MRLxSM Intercross Kidneys

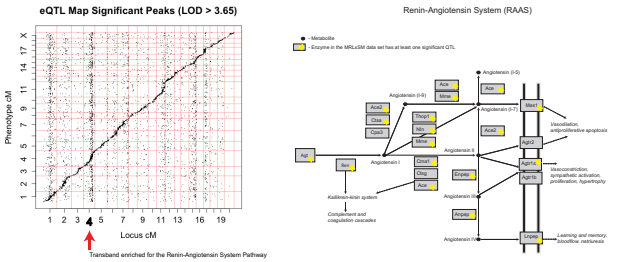

Figure 3: (**Left**) An eQTL map from single-trait analysis of kidney gene expression from an MRLxSM intercross. The trans-band on Chr 4 was enriched for genes in the Renin-Angiotensin system (RAAS) pathway (7/18 genes, P <0.001). Several other members had QTL elsewhere in the genome (7/18). These 14 transcripts and the associated significant QTL were selected as nodes in our network (**Right**).

Results

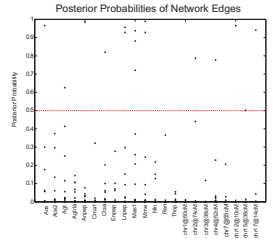

Figure 4: (**Left**) Posterior probabilities estimated by BMA for each network node. Each point is an entry in the consensus matrix which represents the probability of a connection associated with the given node. Connections with probabilities exceeding 0.5 serve as edges in the final weighted network.

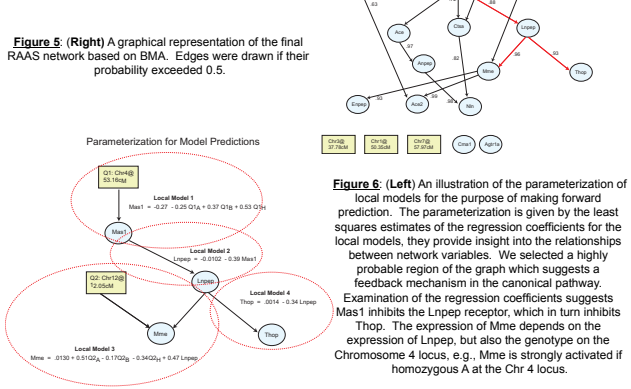

Figure 5: (**Right**) A graphical representation of the final RAAS network based on BMA. Edges were drawn if their probability exceeded 0.5.

Figure 6: (**Left**) An illustration of the parameterization of local models for the purpose of making forward prediction. The parameterization is given by the least squares estimates of the regression coefficients for the local models, they provide insight into the relationships between network variables. We selected a highly probable region of the graph which suggests a feedback mechanism in the canonical pathway. Examination of the regression coefficients suggests Mas1 inhibits the Lnpep receptor, which in turn inhibits Thop. The expression of Mme depends on the expression of Lnpep, but also the genotype on the Chromosome 4 locus, e.g., Mme is strongly activated if homozygous A at the Chr 4 locus.

Conclusions

- We described a novel Bayesian approach to the joint inference of the genotype-phenotype map from eQTL data. Local families of continuous children (phenotypes) and mixed parents (genotypes and phenotypes) are modeled with hierarchical-regression models. Constraints on the network are imposed through a structural prior, and search procedure designed for efficient graph sampling in mixed variable domains is proposed.

- We have applied our method to eQTL data from the kidneys of an MRL×SM intercross to infer causal relationships between members of the renal Renin-Angiotensin System (RAAS). Our analysis revealed a potential feedback loop in the canonical pathway. Further testing is required to validate these hypotheses.

- Our method provides a framework for extracting the most probable gene-gene interactions. Information of this type can be used to predict the effects of genetic interventions, e.g., drugs which attempt to modify genes in order to alter downstream phenotypes.

Acknowledgements

This project was funded with grants from the NIGMS GM076468 (GAC, RK), the NHLBI NSRA fellowship 1F32 HL095240-01 (RSH), AHA post-doctoral fellowship (ML), HL077796 and HL081162 (BP).