# Predicting Cancer Regulatory Networks: A Variational Bayesian Approach

Kevin Tee, Michael Liang

Department of Computer Science
University of California, Berkeley

{kevintee,liangmichael}@berkeley.edu

**For many human diseases, such as cancer, gene regulatory networks can provide insight into disease etiology and pathogenesis. Recent technological advances in the measurement of genome-wide gene expression allow for many computational inferences of such networks. In this method, we use a Bayesian approach to constructing these networks, comparing our results to MERLIN. Both methods were applied to publicly available breast tumor microarray data from the TCGA database. Additionally, we validate our results with comparison to ChIP-seq data and Gene Ontology analysis, suggesting advantages and disadvantages to our method.**

## Introduction

Cancer is a disease of the genome that is associated with accumulation of mutational events that lead to disregulation of the cellular system. Genes involved in cancer development and progression are classified as oncogenes, tumor suppressor genes and genomic stability genes [1]. These

genes have a key role in the regulation of the cell-cycle, proliferation and cell differentiation, and in the regulation of apoptosis [2]. Mutations in over 100 genes are known to drive tumorgenesis, affecting a broad classes of proteins such as transcription factors, chromatin remodelers, growth factors (e.g.,EGFR), growth factor receptors (e.g.,HER2), signal transducers, regulators of apoptosis and DNA repair genes [2]. Within any given tumor there are between 2-8 mutated driver genes modulating the activity of critical molecular pathways [3]. Pilot studies by TCGA and others demonstrated that patients harbor genomic alterations or aberrant expression in different genes and these genes often participate in a common pathway [4,5], indicating that pathway-level genomic perturbations are key features in the underlying cancer biology. Gene regulatory network (GRN) in these cancer pathways has been recognized as potentially important prognosis markers in risk of metastasis or ultimately the target for personalized treatments [6].

High-throughput technologies have produced enormous amount of genome-wide gene expression data. Many data driven mathematical and computational approaches have been developed for probing the molecular interactions of cancer pathways. The molecular interactions between the genes and their products are complex and dynamic. Due to the high dimensionality of gene expression data and insufficient sample measures, the GRN with high prediction accuracy may not reveal the true biological relations. Further, different inference methods may yield different GRNs that are difficult to validate. Thus, the translational potential of gene expression profiling in cancer diagnosis, prognosis and in the development of personalized medicine may not be fulfilled due to the lack of consensus among the inferred GRNs and poor understanding of the underlying mechanisms.

In this paper, we aim to compare two methods using ChIP-seq data as gold standard and validate the clustering of genes into modules using Gene Ontology. The first method is modular

regulatory network learning with per gene information (MERLIN), which assumes a conditional Gaussian distribution for the conditional relationships between regulators and genes and is based on a probabilistic graphical model representation of a regulatory network. The second method is based on stochastic block model, a generative Bayesian approach that learns clusters as latent variables. Both methods will be applied to the publicly available breast tumor microarray data from the TCGA database to infer GRN. We will investigate the differences in the regulatory networks inferred from these two methods based on the established cancer pathway databases and Gene-Ontology analysis.

# Methodology

We investigated 6 cancers: BRCA, COAD, KIRC, LUSC, OV, and UCEC, with the microarray gene expression platform obtained from the TCGA data portal. A corresponding data matrix with rows representing genes and columns representing patient samples was then generated for each cancer data set. These data values were Lowess-normalized, $\log 2$ transformed ratio values comparing expression in the respective patient samples to measurements of the Stratagene Universal Human Reference. There were a total of 8499 genes in these experiments. We direct the reader to the appendix for derivations.

## Covariance Estimation

Typically, in gene co-expression networks, the pairwise co-expression measure is determined through some metric of correlation (Pearson, Spearman, etc.). In our approach, we estimate the covariance matrix using an inverse Wishart prior. The sample covariance is estimated as

follows, where $n$ is the number of patient samples:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(X_i - E[X])(X_i - E[X])^T$$

Note that this is an unbiased estimate of the covariance matrix. However, since the number of patient samples is less than the number of genes (the dimension), the sample covariance is not necessarily positive semi definite. This leads to problems when applying an inverse Wishart prior, which requires the covariance matrix to be PSD. To combat this issue, we add a diagonal matrix to the covariance matrix, where $\lambda$ is defined as the absolute value of the largest non-positive eigenvalue.

$$\Sigma = \Sigma + \lambda I$$

The inverse Wishart distribution is defined as

$$X \sim \mathcal{W}(\mathbf{\Psi}^{-1}, \nu) = \frac{|\mathbf{\Psi}|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p(\frac{\nu}{2})}\,|\mathbf{X}|^{-\frac{\nu+p+1}{2}}\,e^{-\frac{1}{2}tr(\mathbf{\Psi}\mathbf{X}^{-1})}$$

which is conjugate to the normal distribution with known mean and unknown covariance. Therefore, it leads to a relatively simple posterior update:

$$X|\Sigma \sim \mathcal{W}(\mathbf{\Psi} + \sum_{i=1}^{n}(X_i - E[X])(X_i - E[X])^T, n + \nu)$$

For choice of hyperparameters, $\nu$ is set to the number of genes to be weakly informative, and $\mathbf{\Psi}$ is set to the identity.

## Building Clusters with Stochastic Block Models

Given the covariance matrix, a threshold is set where the absolute value of every element above that threshold is set to 1, otherwise it is set to 0. We define this new matrix to be the binary matrix, a square matrix with dimension equal to the number of genes.

To learn clusters, we use a method called Stochastic Block Model (SBM), a generative model which associates each vertex of a network to a latent variable $Z_i$ drawn from a multinomial distribution such that $Z_{iq} = 1$ if vertex $i$ belongs to class $q$. There are a total of $Q$ classes, corresponding to the number of clusters in our model. Since $p(Z_i|\alpha)$ is a multinomial distribution, we place a Dirichlet prior on $\alpha$. Therefore,

$$\alpha \sim Dir(n^0)$$

$$Z_i \sim Multinomial(\alpha)$$

where $n^0 = (n_1^0, ..., n_Q^0)$, and $n_q^0, \forall q$, a non-informative Jeffreys prior. Note that the $0$ corresponds to the initial condition of $n$, which will be updated with more iterations.

The edges are drawn from a Bernoulli distribution, with a Beta prior to model the connectivity matrix, $\pi$, a $Q$ x $Q$ matrix of connection probabilities between clusters (classes). Therefore,

$$\pi \sim Beta(\eta^0, \xi^0)$$

$$X_{ij}|(Z_{iq}Z_{jl} = 1) \sim Bernoulli(\pi_{ql})$$

where $\eta_{ql}^0 = \xi_{ql}^0 = 1/2, \forall q, l$, again non informative hyperparameters for the Beta distribution.

Considering an undirected graph without self loops, we arrive at the following conditional probabilities:

$$P(Z|\alpha) = \prod_{i=1}^{N} \prod_{q=1}^{Q} \alpha_q^{Z_{iq}}$$

and

$$\begin{aligned} P(X|Z, \pi) &= \prod_{i<j} P(X_{ij}|Z_i, Z_j, \pi) \\ &= \prod_{i<j} \prod_{q,l} Bernoulli(X_{ij}|\pi_{ql})^{Z_{iq}Z_{jl}} \\ &= \prod_{i<j} \prod_{q,l} (\pi_{ql}^{X_{ij}}(1-\pi_{ql})^{1-X_{ij}})^{Z_{iq}Z_{jl}} \end{aligned}$$

In the case of a directed graph, the condition is $i \neq j$ as opposed to $i < j$.

## Estimation Using Variational Bayes EM

The distribution $p(Z, \alpha, \pi|X)$ is intractable, so we propose an approximation to this distribution using variational inference. The marginal log-likelihood can be decomposed into two terms, the second of which being the KL divergence.

$$\ln p(X) = \mathcal{L}(q(Z, \alpha, \pi) + KL(q(Z, \alpha, \pi)||p(Z, \alpha, \pi, X))$$

where

$$\mathcal{L}(q(Z, \alpha, \pi) = \sum_{Z} \int \int q(Z, \alpha, \pi) \ln(\frac{P(X, Z, \alpha, \pi)}{q(Z, \alpha, \pi)}) d\alpha d\pi$$

and

$$KL(q(Z, \alpha, \pi)||p(Z, \alpha, \pi, X)) = -\sum_{Z} \int \int q(Z, \alpha, \pi) \ln(\frac{P(Z, \alpha, \pi|X)}{q(Z, \alpha, \pi)}) d\alpha d\pi$$

In mean field variational inference, we assume that $q(Z, \alpha, \pi)$ can be factorized such that

$$q(Z, \alpha, \pi) = q(\alpha)q(\pi)q(Z) = q(\alpha)q(\pi) \prod_{i=1}^{n} q(Z_i)$$

Using a variational Bayesian EM algorithm, we define the E-step to be the optimization of each $q(Z_i)$ and the M-step to be the approximations of $q(\alpha)$ and $q(\pi)$.

# Results

# Discussion

## References

[1] Vogelstein, B. and Kinzler, K.W. Cancer genes and the pathways they control. Nat Med 10, 789-99 (2004).

[2] Croce, C.M. Oncogenes and cancer. N Engl J Med 358, 502-11 (2008).

[3] Vogelstein, B. et al. Cancer genome landscapes. Science 339, 1546-58 (2013).

[4] Parsons, D.W. et al. An integrated genomic analysis of human glioblastoma multiforme. Science 321, 1807-12 (2008).

[5] Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 1061-8 (2008).

[6] Drier, Y., Sheffer, M. and Domany, E. Pathway-based personalized analysis of cancer. Proc Natl

Acad Sci U S A 110, 6388-93 (2013).