Tissue tree gene biclustering and gene dynamics

Avinash Das

August 25, 2013

1 Abstract

Problem of classification of genes is one of fundamental problem in field of genetics. We propose a hierarchical Bayesian method to perform gene classification. We develop a message passing based EM algorithm for inference. Final we demonstrate how the model can tease apart gene dynamics on tissue tree.

2 Introduction

2.1 Classification of genes

One of the classical problem in the area of computational biology is to classify genes. There are more than 40 thousand known genes in human genome. Function of many genes are poorly understood. They function in different manner in different context. Many of the these genes control the expression of other genes and sometime regulate its own expression. This add a layer of complexity to analyze and define classification of genes.

The task of gene classification is also important. Many of the genome wide studies like genome-wide association studies or differential gene expresssion analysis find set of genes. With better classification, we can analyze such experiment Gene annotation is one example of several existing gene classification methods. It define tree based annotation of genes. It also give relationship of genes with pathways.

Clustering is other popular method for gene classification. Clustering of genes tries to find set of genes which are co-expressed. Co-expressed are sometime defined as two or more genes are expressed together across all samples. This is limiting definition for classification. Genes are known to have different function in different context. Therefore, it is arguably more useful to define clusters as set of genes which are co-expressed in subset of the samples. Biclustering is approach to find significant submatrices in matrix. Biclustering of genes therefore will fish out co-expressed genes in subset of samples.

The main limitation of clustering or biclustering approach is, they are local search algorithm and hence may miss out important genes clusters. Another

assumption that samples usually have flat relation. Many times samples corresponds to tissues or disease sample or samples from different individuals. All such cases trees is better representation of relationships.

Another very useful classification of gene classification is: a) housekeeping genes: genes expressed in all tissues. b) tissue specific genes: genes which are expressed in few tissues. The later category are gene which considered as gene signature of tissue. The main limitation of these classification they are very simplistic.

We propose a novel method of gene classification that combines these two approaches of gene classification. We try to use inherent tree relationship between sample to impute better cluster. Further, we define homogeneous set of gene at each internal node of inferred tree.

2.2 In-situ hybridization data

The data is collected from http://www.eurexpress.org/, composed of expression of 5600 genes across 811 tissues. The genes are manually annotated and have discreet value as strong, medium, weak or no expression. For initial analysis we can treat the data as binary i.e. expressed or not expressed. Later it can be generalized to take four discreet values.

3 Methods

3.1 Kingman's coalescent

To infer tree from observed data, we used agglomerative clustering using Kingman's coalescent [Teh et al., 2008]. It defines exponential distribution as prior over trees. It calculates posterior of the trees given the observation. The inference is performed by message passing algorithm, by passing message upward from leaf to trees. One of the advantage of method is the distribution is exchangeable.

Let π be the tree defining genaology of n individuals. The tree π can be defined by $n-1$ coalescent events. The ith coalescing event is defined by ρ_{l_i} and ρ_{r_i}, the left and right subtree that are coalescing at waiting δ_i after $i-1$th coalescing event. The prior over tree was defined as:

$$p(\pi) = \prod_{i}^{n-1} \exp\left(-\binom{n-i+1}{2}\delta_i\right) \tag{1}$$

[Teh et al., 2008] proved that joint probability of observation and tree can

be given by:

$$p(x,\pi) \propto \prod_{i}^{n-1} exp\left(-\binom{n-i+1}{2}\delta_i\right)\tilde{Z}_i(X|\theta_i) \quad (2)$$

$$\textit{where, } \tilde{Z}_i(X|\theta_i) = \int\int p(a)k_{-\infty t_i}(x,y)M_i(y)dyda \quad (3)$$

$$M_i(y) \propto \prod_{b=l,r}\int k_{t_i t_{b_i}}(y,y_b)M_i(y_b)dy_b \quad (4)$$

$M_i(y)$ is message propagated upward to subtree θ_i from its both children. This can be calculated iteratively by propagating messages from leaf to root node. $\tilde{Z}_i(X|\theta_i)$ can be viewed as local likelihood. Inference is based on equation 4. At each step i, a duration is δ_i is sampled and then a pair ρ_{l_i}, ρ_{r_i} is chosen from the proposal distribution.

3.2 Latent gene state variable Z

Given we have reasonable accurate tissue tree representing the relationship between tissues, the problem of gene biclustering reduces to inferring the state of genes at each internal node of the inferred tree. Given a tree $\pi(V,E)$ and genes G, we define latent state variable $Z = \{z_{ju} \in \{0,1,2\} : \forall j \in G, u \in V\}$ as:

$z_{ju} = 0 \implies$ *gene* j *is homogenously inactive in subtree with* u *as root.*

$z_{ju} = 1 \implies$ *gene* j *is homogenously active with high expression in subtree with* u *as root.*

$z_{ju} = 2 \implies$ *gene* j *have hetrogenous activity in subtree with* u *as root.*

Once a node z_{ju} takes a value either 0 or 1 corresponding to states homogeneously active or in-active all its children node in subtree stays in same homogeneous state.

3.3 Generative process I

The problem of biclustering of the genes reduces to simultaneously inferring 1) tree of tissue and 2) gene state variable at each internal node of the tree. We define a generative process to generate gene expression data given the tree using Z as the internal latent variable:

$$\begin{aligned} \pi &\sim \text{kingsman}() \\ Z|\pi,\phi &\sim \text{mult}(\phi) \\ \phi &\sim \text{Dir}(\alpha,\beta,\gamma) \\ Y|Z=i &\sim \text{Bin}(p_i) \\ p_i &\sim \text{Beta}(shape1_i, shape2_i) \end{aligned} \quad (5)$$

The basic idea of generative model is very similar to Kingman's agglomerative clustering. The main difference is instead of using observed data (gene

expression) to infer tree, the generative process uses the state variable Z at each internal nodes. Therefore, gene expression is thought to be generated at leafs given the tree and gene state variable.

The message passing algorithm in [Teh et al., 2008] uses transition kernel $k_{t_i t_{bi}}(y, y_b)$ to define transition from one state to other. If Z is used to infer tree, transitions cannot be independent for both children. That is transition of left children to parent cannot be independent of state of right children. For instance: once a node become homogeneous all it descendant should be homogeneous. This kind of complicated dependencies demand a three dimensional transition kernel. Such kernel does not arrest a closed form solution of messages from equation 4.

3.4 Generative process II

To solve the problem described in previous section, we propose to split the tree inference with the inference of latent gene state Z. We infer tree directly from the observed gene expression from agglomerative clustering [Teh et al., 2008]. At each of the internal state we also infer gene expression Y at each internal nodes of tree by passing message downward and combing it with upward message.

Given the tree π and inferred gene expression at each internal node. We define a new generative process:

$$Y|Z = i, \pi \sim \text{Beta}(shape1_i, shape2_i) \tag{6}$$

The Y at internal nodes are probability therefore it arrest a $\mathtt{Beta}$ distribution. We can also assume given the tree structure and expression at internal nodes, gene expression becomes independent. This implies we can infer Z independently for each genes.

3.5 Inference

We used message passing algorithm to compute posterior probabilities of latent variable at internal nodes. Parameters from generative process defined from equation can be inferred by EM algorithm. The graphical model induced by the generative process contains hidden states Z and observations at each node. This similar to HMM just instead of Markov chain, induced graph is a tree. Therefore EM algorithm is similar to Baum-Welch algorithm [Rabiner and Juang, 1986].

The upward message α at node t can be calculated starting from leafs and propagating upward. The observed variables influencing Y_t are split into two subsets: a) $e^-_{Y_t}$ composed of observed variables emitted by descendants of node t (including t) and b) $e^+_{Y_t}$ composed of observed variable emitted by non-descendants of node t (excluding t) [Starr and Shi, 2004]. l and r are respectively left and right child of node t. p and s are respectively parent and

sibling of node t. $\oplus$ is outer product of two vector.

$$\begin{aligned}\alpha_t(i) &= Pr(e^-_{Y_t}|Z_t = i) \\ &= \sum_{i,j} \alpha_l(i)\alpha_r(j)\Gamma_{ij;k}Pr_k(Y_t)\end{aligned}$$

Where, $\Gamma_{ij;k}$ is transition matrix from $(Z_l = i, Z_r = j)$ to $Z_t = k$. $Pr_k(Y_t) = Pr(Y_t|Z_t = k)$ is the emission probability. This can be written in following matrix form. It is important to keep these matrices information so that algorithm can be implemented in higher language program like R or matlab without sacrificing much on speed.

$$\begin{aligned}\boldsymbol{\alpha_t} &= \boldsymbol{\delta P}(x_i) \quad \forall t \in \textit{leaf} \\ \boldsymbol{\alpha_t} &= (\boldsymbol{\alpha_l} \oplus \boldsymbol{\alpha_r})\boldsymbol{\Gamma Pr}(Y_t) \quad \forall t \not\in \textit{leaf}\end{aligned} \tag{7}$$

In the similar manner, we can derive the iterative formula for downward messages.

$$\begin{aligned}\beta_t(i) &= Pr(e^+_{Y_t}|Z_t = i) \\ &= \sum_{j,k} \alpha_s(j)\Gamma_{ij;k}\beta_p(k)Pr_i(Y_t)\end{aligned}$$

Equivalently ,

$$\boldsymbol{\beta_t} = \boldsymbol{\alpha_s}\Gamma_{ij;k}\boldsymbol{\beta_p}\mathbf{Pr}(\mathbf{Y_t}) \tag{8}$$

The complete likelihood of observed data is expressed in terms of α and β as: $L_T = \alpha_t\beta'_t = Pr(\mathbf{Y}^{(T)})$, for each t.

The unknown parameters transition and emission probabilities can be inferred by EM algorithm:

$$\begin{aligned}\textit{E step:}\quad \hat{u}_j(t) &= Pr(Z_t = j/y^{(T)}) = \alpha_t(j)\beta_t(j)/L_T \\ \hat{v}_{jk;i}(t) &= Pr(Z_l = j, Z_r = k, Z_t = i|\mathbf{y}^{(T)})\end{aligned}$$

$$\begin{aligned}\textit{M step:}\quad \gamma_{jk;i} &= \tilde{f}_{jk;i} \\ f_{jk;i} &= \sum \hat{v}_{jk;i}(t)\end{aligned}$$

Where, $\tilde{f}$ is normalized f. The shape parameter of emission probability can be found using the Newton's method that involves digamma and trigamma function.

4 Results

4.1 Convergence

The EM algorithm converge rapidly. Typically it takes less than 50 iteration to converge (change in likelihood $<$ 1e-6).

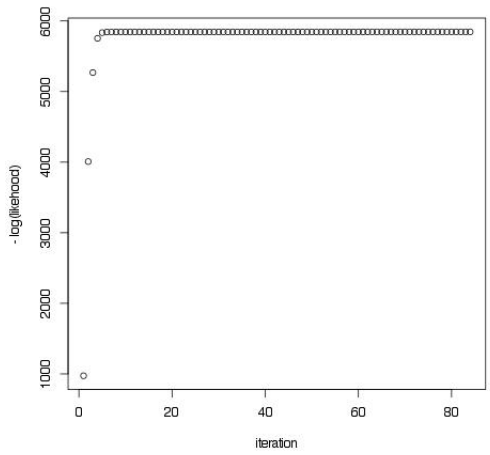

Figure 1: Rapid convergence of EM algorithm

We expect once as a node in the tissue tree is homogeneous active state all its descendant must be in active state. Therefore $\gamma_{11;1} \approx 1$. We expect similar behaviour for homogeneous inactive state. This seems to be the case for the transition matrix.

The emission probability of three states are consistent with definition of latent state variable of gene Z.

4.2 Simulation result

We generated a simulated data set with the proposed generative process. Then, by keeping uniform distribution as transistion matrix and random initialization parameter. We learned parameter by EM algorithm. The fig shows original transition matrix and inferred transition matrix.

4.3 Gene dynamics

The proposed method can be run on Tissue tree individually based on the independence assumption. The state variable of gene gives its dynamics in the tissue tree. Fig. shows the dynamics of one gene HMISC which knwn to e cardio vascular related. The agglomerative clustering nicely cluster together cluster corresponding to heart specific tissue. The swictching event is orange circle represent where transition probability convert to homogenous active stat from hetrogenous state.

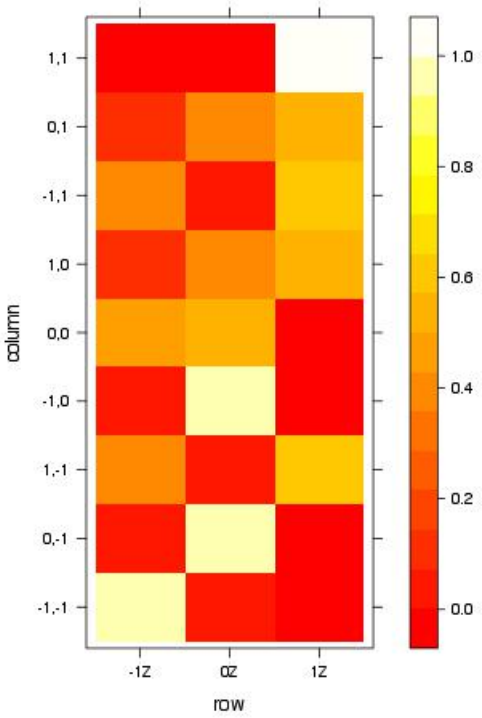

Figure 2: Transistion matrix of 3 state HMM.

References

[Rabiner and Juang, 1986] Rabiner, L. and Juang, B. (1986). An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16.

[Starr and Shi, 2004] Starr, C. and Shi, P. (2004). *An introduction to bayesian belief networks and their applications to land operations.* DSTO Systems Sciences Laboratory.

[Teh et al., 2008] Teh, Y. W., Iii, H. D., and Roy, D. (2008). Bayesian agglomerative clustering with coalescents. In *In Advances in Neural Information Processing Systems*. Citeseer.

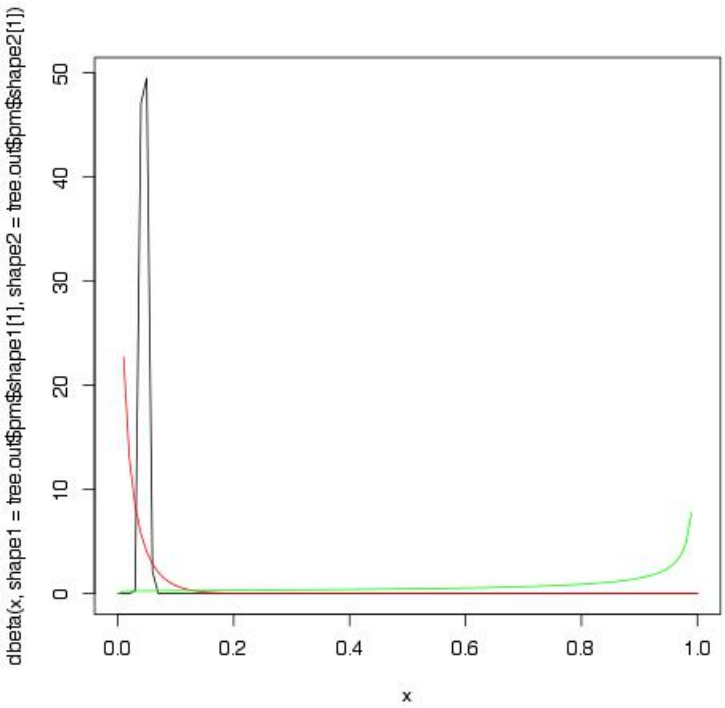

Figure 3: Emission probability from the homogeneously inactive(black), heterogeneous (red) and homogeneously active (green)

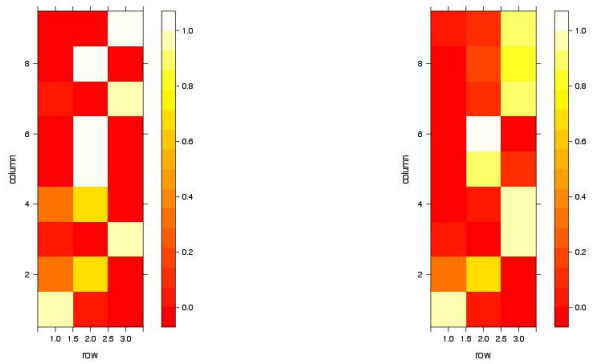

Figure 4: Transistion matrix of simulated data and estimated transition transition matrix

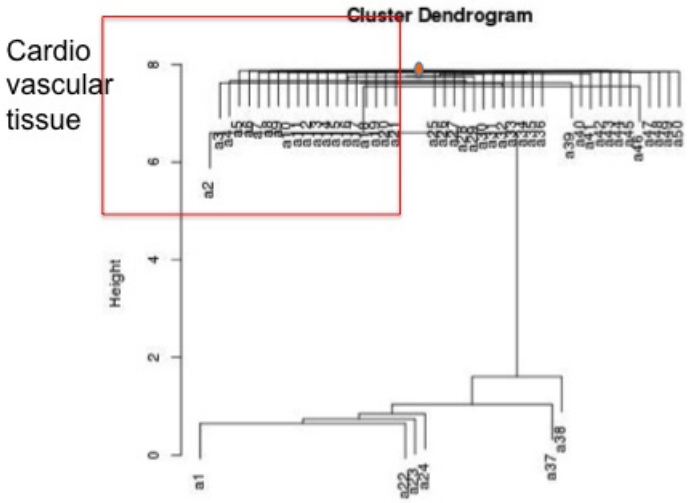

Figure 5: Emission probability from the homogeneously inactive(black), heterogeneous (red) and homogeneously active (green)