Tissue tree gene biclustering

Avinash Das

March 25, 2013

1 Problem

Given the gene expression across multiple tissues, find gene clusters and most probable tree defining relations between tissues. Tissue tree implies the relation of the different tissues in an embryo, with embryo as root. Cluster of gene implies group of genes homogeneously all active or inactive in tissues of subtree. For eg. one of intended outcome would be genes those are homogeneously active genes in tissue subtree with heart as the root, i.e. these genes will be active in left ventricle, right ventricle and all tissues related to heart.

2 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 Generative process

Latent state variable Z: 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:

$$\begin{aligned} z_{ju} &= 0 \implies \textit{gene } j \textit{ is homogenously inactive in subtree with } u \textit{ as root.} \\ z_{ju} &= 1 \implies \textit{gene } j \textit{ is homogenously active with high expression in subtree with } u \textit{ as root.} \\ z_{ju} &= 2 \implies \textit{gene } j \textit{ have inhomogenously activity in subtree with } u \textit{ as root.} \end{aligned}$$

In addition, 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.

Prior over tissue tree π can be defined by Coalescent clustering.

Z can be then sampled from multinomial distribution with a Dirichlet prior.

$$\begin{aligned} \pi &\sim \text{coalescent}() \qquad (1) \\ Z|\pi, \phi &\sim \text{mult}(\phi) \\ \phi &\sim \text{Dir}(\alpha, \beta, \gamma) \end{aligned}$$

Generation of expression level: Given the state variable Z, gene expression is generated from a Bernoulli with parameter only dependent on state variable Z.

$$\begin{aligned} Y|Z = i &\sim \text{Bin}(p_i) \\ p_i &\sim \text{Beta}(\alpha_i, \beta_i) \forall i \in \{1, 2, 3\} \qquad (2) \end{aligned}$$

The beta priors can be chosen so to reflect our expected behaviour of expression level in different gene state. One example of prior that can be taken is shown in fig. 1.

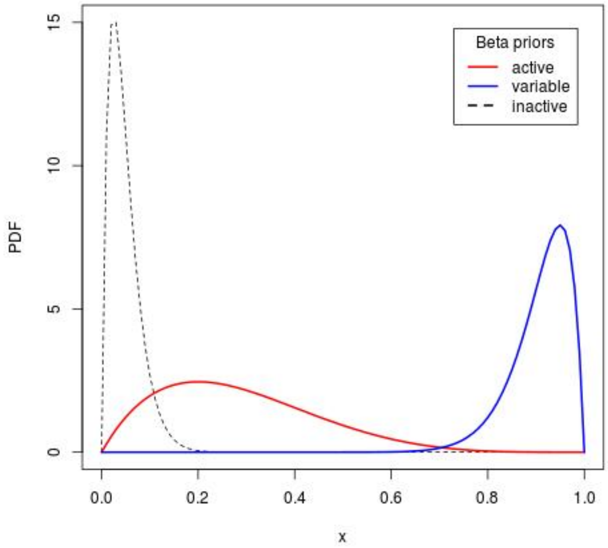

Figure 1: Beta priors for homogeneous active, homogeneous inactive and inhomogeneous gene state. The priors are consistent with the interaction matrix data; e.g. $E(p_2)$ = average that any gene is expressed in interaction matrix.

4 Expected outcomes

- We will obtain a tissue tree.

- At each node of the tree we will get cluster of genes which are active or inactive in the subtree.
- δ parameter in Kingman's coalescent will give the distance between the each tissue. For eg. we will get information if brain is closer to heart than kidney.
- Probably clustering model will be resistant to error due manual annotation.