Inferential Analysis for Dynamic Pathway Topology in Spatial-Temporal Transcriptome Dataset

March 13, 2017

Abstract

We aim to study the pathway structures in the genomic networks in Brainspan data. We apply the combinatorial inference to the spatial-temporal dataset in order to discover the topological features in the pathways, especially both the inner- and intra-connectivity property among pathways. Our goal is to find the times period when different pathways become linked and related to the disease.

1 Introduction

Spatiotemporal expression profile is believed to be informative in disentangling regulatory roles of genes during development. The brainspan dataset including RNA-seq and microarray expression profile, is comprehensive in both time span and regions. There have been a wide range of analyses of microarray dataset published by Kang *et al*, like functional-PCA and autism gene discovery. However, data mining towards pathway level is limited compared with others.

Pathways are important building blocks for realization and regulation of biological functions. Besides gene-gene interaction networks or protein-protein interaction networks, pathways are also intertwined to different level in respective developmental stages. It is intriguing to unveil the communication between networks, which is also crucial for better understandings of the importance of spatiotemporal factors.

An interesting scientific discovery would be different pathways, e.g., immunization and dendrites, may work together to influence the disease at certain period of time between infancy and adult. To depict the cross-talk behaviors, interactions in both gene and sub-pathway level, which are located in certain signaling pathways, will be studied cross 8 time windows and 16 regions.

2 Statistical Model

2.1 Mutual information

To screen the interactions between pathways, certain metric to measure interactions in pathway level is necessary. The SVD-based methods to describe correlation between gene sets have been widely adopted in modular network research. However, with continuous change of gene expression and different patterns, it is hard to summarize representative modules through spatiotemporal space, which makes it inflexible to deal with gene sets with known labels.

To simplify the description challenge, we first propose to treat data matrix $\boldsymbol{X} \in \mathbb{R}^d$ blockwisely. Namely, $\boldsymbol{X} \in \mathbb{R}^d$ consists of several multivariate normal random variables $\boldsymbol{X_i} \in \mathbb{R}^{|\boldsymbol{p}_i|}$, $i \in \{1...k\}$ and $\sum_i |\boldsymbol{p}_i| = d$.

$$\boldsymbol{X} = \left[\boldsymbol{X}_1, ..., \boldsymbol{X}_k\right]$$

For $k = 2$, the mutual information for random variables X_1 and X_2 can be calculated based on their joint distribution $f(\boldsymbol{X})$, where $X \sim N_d(0, \boldsymbol{\Sigma})$. A blockwise representation of $\boldsymbol{\Sigma}$ is:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

Then,

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix}$$

The marginal distribution of X_1 and X_2 are still normal with zero mean and covariance matrices $\{\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}\}$ respectively. The mutual information between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is defined as,

$$MI(\boldsymbol{X}_1, \boldsymbol{X}_2) = D_{KL}(f(\boldsymbol{X}_1, \boldsymbol{X}_2), f(\boldsymbol{X}_1)f(\boldsymbol{X}_2)) = \int f(\boldsymbol{X}_1, \boldsymbol{X}_2) log \frac{f(\boldsymbol{X}_1, \boldsymbol{X}_2)}{f(\boldsymbol{X}_1)f(\boldsymbol{X}_2)}$$

After the integral,

$$MI(X_1, X_2) = \frac{1}{2}(log(\frac{|\boldsymbol{\Sigma}_{11}||\boldsymbol{\Sigma}_{22}|}{|\boldsymbol{\Sigma}|}) + tr(\begin{bmatrix} 0 & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & 0 \end{bmatrix} \boldsymbol{\Sigma}))$$

Because we do not have $\boldsymbol{\Sigma}$, we use sample covariance $\widehat{\boldsymbol{\Sigma}}$ to estimate $\boldsymbol{\Sigma}$. Here $|\boldsymbol{p}_1| + |\boldsymbol{p}_2| = d$, what if $|\boldsymbol{p}_1| + |\boldsymbol{p}_2| > d$. A potential extension is to use conditional mutual information to represent the dependence.

Let $\boldsymbol{p}_B = \boldsymbol{p}_1 \cap \boldsymbol{p}_2$ and $\boldsymbol{p}_A = \boldsymbol{p}_B^C$, then the conditional joint distribution of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ will be

$$\boldsymbol{X}|x_i, i \in \boldsymbol{p}_B \sim N(\boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}(\boldsymbol{X}_B - \boldsymbol{\mu}_B), \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA})$$

The dependence between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is defined as $MI(\boldsymbol{X}_1, \boldsymbol{X}_2|x_i, i \in \boldsymbol{p}_B)$. To make each pair comparable, the dependence should be normalized, hence $\Pr(MI(\boldsymbol{X}_1, \boldsymbol{X}_2|x_i, i \in \boldsymbol{p}_B) \geq Obs)$ is calculated by permutation and the dependence is in $[0, 1]$ scale (lazy normalization!). Some other normalization methods reminiscent of pearson correlation coefficient(Strehl and Ghosh, 2002) are still under considerations if we want to incorporate MI in further framework.

2.2 Gaussian graphical model

We apply dynamic Gaussian graphical model to the dataset. Let $\boldsymbol{X} \in \mathbb{R}^d$ be the expression levels of d genes in the communities where our pathway of interest belongs to. It is a d-dimensional random vector and let $Z \in [0, 1]$ be the time covariate random variable. The time-varying Gaussian graphical model assumes that

$$\boldsymbol{X} \mid Z = z \sim N_d(\boldsymbol{0}, \boldsymbol{\Sigma}(z)), \tag{1}$$

where $\boldsymbol{\Sigma}(z)$ is the covariance matrix of $\boldsymbol{X}$ given $Z = z$. Let $\boldsymbol{\Theta}(z) = (\boldsymbol{\Sigma}(z))^{-1}$ be the inverse covariance matrix at $Z = z$. It is well-known that $\boldsymbol{\Theta}(z)$ encodes the conditional dependence relationships among the random variables at time $Z = z$, i.e., $\mathbf{\Theta}_{jk}(z) = 0$ if and only if the jth and kth variables are conditionally independent given all of the other variables.

2.2.1 Inferential Problems

We aim to explore the following inferential problems.
• **Confidence Set of Pathways Connected Giant.** Let V be the vertex set containing all genes in the pathway and $G(z)$ is the graph. Let $S(z)$ be the node set of the largest connected component in $G(z)$. We aim to construct a confidence set $\widehat{S}_\alpha(z)$ such that

$$\lim_{n\to\infty} \mathbb{P}\Big(S(z) \subseteq \widehat{S}_\alpha(z), \text{ for all } z \in [z_1, z_2]\Big) \geq 1 - \alpha.$$

• **Connectivity between Pathways.** Let V_1 and V_2 be two vertex sets belongs to two different pathways and $G_1(z)$ and $G_2(z)$ are the corresponding graphs. The pathway inter-connectivity test in the given time period $[z_1, z_2]$ is

$$\begin{aligned} &H_0: \text{ There exists } z_0 \in [z_1, z_2] \text{ such that there is no edge connecting } G_1(z_0) \text{ and } G_2(z_0);\\ &H_1: \text{ } G_1(z) \text{ and } G_2(z) \text{ is connected to each other for all } z \in [z_1, z_2]. \end{aligned}$$

Furthermore, we can explore the intensity level of the cross-walk. In specific, let $E_{12}(z)$ be the edge set connecting $G_1(z)$ and $G_2(z)$. The intensity level of the cross-walk is the cardinality $|E_{12}(z)|$. We aim to test

$$\begin{aligned} &H_0: \text{ } |E_{12}(z)| \leq N \text{ for all } z \in [z_1, z_2];\\ &H_1: \text{ There exists } z^* \text{ such that } |E_{12}(z^*)| > N. \end{aligned}$$

• **Change Point Detection** Let V_1 and V_2 be two vertex sets belongs to two different pathways and $G_1(z)$ and $G_2(z)$ are the corresponding graphs. We aim to test whether the connectivity is changed in $[z_1, z_2]$, i.e., for the inter-pathway,

$$\begin{aligned} &H_0: \text{ } G(z) \text{ is disconnected for all } z \in [z_1, z_2];\\ &H_1: \text{ There exists } z^* \text{ such that } G(z) \text{ is disconnected for } z \in [z_1, z^*] \text{ and connected for } z \in [z^*, z_1]; \end{aligned}$$

and for the intra-pathways.

$$\begin{aligned} &H_0: \text{ There is no edge connecting } G_1(z) \text{ and } G_2(z) \text{ for all } z \in [z_1, z_2];\\ &H_1: \text{ There exists } z^* \text{ such that } G_1(z) \text{ and } G_2(z) \text{ is disconnected for } z \in [z_1, z^*] \text{ and connected for } z \in [z^*, z_1]. \end{aligned}$$

Also, an important question will be estimating the change point z^*.

We may also estimate the change point of intensity level. For example, we test

$$\begin{aligned} &H_0: \text{ } |E_{12}(z)| \text{ is constant for all } z \in [z_1, z_2];\\ &H_1: \text{ There exists } z^* \text{ such that the value of } |E_{12}(z)| \text{ is changed at } z^*. \end{aligned}$$

And also we want to estimate the value of z^*.

2.2.2 Inferential Methods

We begin with inferring a single edge. Let $\boldsymbol{X}$ be the random vectors containing genes in two pathway. Let $(\boldsymbol{X}_1, Z_1), \ldots, (\boldsymbol{X}_n, Z_n)$ be n independent realizations of the pair of random variables $(\boldsymbol{X}, Z)$. For simplicity, in the following of our paper, we let $[z_1, z_2] = [0, 1]$. Let $K : \mathbb{R} \to \mathbb{R}$ be a symmetric kernel function. To obtain an estimate for $\boldsymbol{\Sigma}(z)$, we use the kernel smoothed covariance estimator

$$\widehat{\boldsymbol{\Sigma}}(z) = \frac{\sum_{i\in[n]} K_h(Z_i - z)\boldsymbol{X}_i\boldsymbol{X}_i^T}{\sum_{i\in[n]} K_h(Z_i - z)}, \tag{2}$$

where $K_h(Z_i - z) = K\left((Z_i - z)/h\right)/h$ and $h > 0$ is the bandwidth parameter. The choice of kernel is not essential as long as it satisfies some regularity conditions. Let $\widehat{\boldsymbol{\Theta}}(z) = (\widehat{\boldsymbol{\Sigma}}(z))^{-1}$ or other estimator like glasso and CLIME. Mirror-reflect method is used to address the boundary effects. if the dataset is high dimensional. We consider the debiased estimator

$$\widehat{\boldsymbol{\Theta}}^d_{jk}(z) = \widehat{\boldsymbol{\Theta}}_{jk}(z) - \frac{\left(\widehat{\boldsymbol{\Theta}}_j(z)\right)^T \left[\widehat{\boldsymbol{\Sigma}}(z)\widehat{\boldsymbol{\Theta}}_k(z) - \mathbf{e}_k\right]}{\left(\widehat{\boldsymbol{\Theta}}_j(z)\right)^T \widehat{\boldsymbol{\Sigma}}_j(z)}, \tag{3}$$

To this end, we construct a test statistic based on the de-biased estimator $\widehat{\boldsymbol{\Theta}}^d_{jk}(z)$ in (3). Let

$$T_E = \sup_{z\in[0,1]} \max_{(j,k)\in E(z)} \sqrt{nh} \cdot \left|\widehat{\boldsymbol{\Theta}}^d_{jk}(z) - \boldsymbol{\Theta}_{jk}(z)\right| \cdot \left(\frac{1}{n}\sum_{i\in[n]} K_h(Z_i - z)\right), \tag{4}$$

where the subscript E indicates that the maximum is taken over the edges in the set $E(z)$ for each z. Since the test statistic (4) involves taking the supreme over z and the maximum over all edges in $E(z)$, it may be difficult to evaluate its distribution. To approximate the distribution of the test statistic T_E, we generalizes the Gaussian multiplier bootstrap. We construct the bootstrap statistic as

$$T_E^B = \sup_{z\in[0,1]} \max_{(j,k)\in E(z)} \sqrt{nh} \cdot \left|\frac{1}{n}\sum_{i\in[n]} \left(\widehat{\boldsymbol{\Theta}}_j(z)\right)^T K_h(Z_i - z)\left(\boldsymbol{X}_i\boldsymbol{X}_i^T\widehat{\boldsymbol{\Theta}}_k(z) - \mathbf{e}_k\right)\xi_i\right|, \tag{5}$$

where $\xi_1, \ldots, \xi_n \overset{\text{i.i.d.}}{\sim} N(0, 1)$. We denote the conditional $(1 - \alpha)$-quantile of T_E^B given $\{(Z_i, \boldsymbol{X}_i)\}_{i\in[n]}$ as

$$c(1 - \alpha, E) = \inf\left\{t \in \mathbb{R} \mid P\left(T_E^B \le t \mid \{(Z_i, \boldsymbol{X}_i)\}_{i\in[n]}\right) \ge 1 - \alpha\right\}. \tag{6}$$

The quantity $c(1 - \alpha, E)$ can be calculated numerically using Monte-Carlo.

Now we are ready to test the problems above. Given a graph $G_0 = (V, E_0)$, we denote all the connected subgraphs of G_0 as $\{G_{0\ell} = (V_{0\ell}, E_{0\ell})\}_{\ell=1}^{k'}$. We define the edge set

$$\mathcal{C}(E_0, \text{Conn}) = \left\{(u, v) \in E_0^c \,\middle|\, u \in V_{0\ell}, v \in V_{0\ell'}, \ell \neq \ell'\right\}. \tag{7}$$

Now we discuss the connectivity test between pathways. We aim to test the intensity level of

```
Algorithm 1 Confidence Set of Pathways Connected Giant.
Input: $\widehat{\boldsymbol{\Theta}}^d(z)$ for $z \in [0,1]$.
Initialize: $t = 0$; $E_0(z) = \emptyset$ and connected node sets $\mathcal{V}(z) = \{\{1\}, \dots, \{d\}\}$ for $z \in [0,1]$ .
Repeat:
    1. Compute the critical edge set $\mathcal{C}(E_{t-1}(z), \text{Conn})$ for $z \in [0,1]$ and the conditional quantile $c(1-\alpha, \mathcal{C}(E_{t-1}, \text{Conn})) = \inf\Big\{t \in \mathbb{R} \mid P(T^B_{\mathcal{C}(E_{t-1}, \text{Conn})} \le t \mid \{(Z_i, \boldsymbol{X}_i)\}_{i\in[n]}) \ge 1-\alpha\Big\}$, where $T^B_{\mathcal{C}(E_{t-1}, \text{Conn})}$ is the bootstrap statistic defined in (5) with the maximum taken over the edge set $\mathcal{C}(E_{t-1}(z), \text{Conn})$.
    2. Construct the rejected edge set
       $$\mathcal{R}(z) = \left\{ e \in \mathcal{C}(E_{t-1}(z), \text{Conn}) \mid \sqrt{nh} \cdot |\widehat{\boldsymbol{\Theta}}^d_e(z)| \cdot \sum_{i\in[n]} K_h(Z_i - z)/n > c(1-\alpha, \mathcal{C}(E_{t-1}, \text{Conn})) \right\}.$$
    3. Update the rejected edge set $E_t(z) \leftarrow E_{t-1}(z) \cup \mathcal{R}(z)$ for $z \in [0,1]$.
    4. $t \leftarrow t+1$.
Until: $E_t(z) = E_{t-1}(z)$ for $z \in [0,1]$.
Output: $\widehat{S}_\alpha(z)$ be the vertex set of largest connected subgraph of $E_t(z)$.
```

the cross-walk is the cardinality $|E_{12}(z)|$

$$\begin{aligned} &H_0: \ |E_{12}(z)| \le N \text{ for all } z \in [z_1, z_2]; \\ &H_1: \ \text{There exists } z^* \text{ such that } |E_{12}(z^*)| > N. \end{aligned}$$

We can see that if $N = 0$, the test above is a test for whether there exists a cross-walk.

Algorithm 2 Confidence Set of Pathways Connected Giant (Detailed)

```
Input: $\widehat{\mathbf{\Theta}}^d(z)$ for $z \in [0,1]$.
We discretize the continuous interval $[0,1]$ into $z_1, \dots, z_m$
Initialize $t = 0$; $E_0(z) = \emptyset$ and connected node sets $\mathcal{V}(z) = \{\{1\}, \dots, \{d\}\}$ for $z \in [0,1]$ .
repeat
  $t \leftarrow t + 1$;
  for $z \in \{z_1, \dots, z_m\}$ do
    Find the critical edge set $\mathcal{C}_t(z) = \big\{(u,v) \in E^c_{t-1} \,\big|\, u \in S, v \in T, S \neq T$ and $S, T \in \mathcal{V}(z)\big\}$.
    Update the rejected set: $\mathcal{R}_t(z) = \{e \in \mathcal{C}_t(z) \mid \sqrt{nh} \cdot |\widehat{\mathbf{\Theta}}^d_e(z)| \cdot \sum_{i \in [n]} K_h(Z_i - z)/n > c(\alpha, \mathcal{C}_t(z))\}$;

    $E_t(z) \leftarrow E_{t-1}(z) \cup \mathcal{R}_t(z)$;
    Update the connected node sets $\mathcal{V}(z)$:
    for $(u,v) \in \mathcal{R}_t(z)$ do
      if $u, v$ belong to different node sets $S, T$ in $\mathcal{V}(z)$, i.e., $u \in S, v \in T$ and $S \neq T$ then
        $\mathcal{V}(z) \leftarrow (\mathcal{V}(z) \backslash \{S, T\}) \cup \{S \cup T\}$
      end if
    end for
  end for
until $\mathcal{R}_t(z) = \emptyset$ for all $z \in \{z_1, \dots, z_m\}$.
Output: $\widehat{S}_\alpha(z)$ be the vertex set of largest connected subgraph of $E_t(z)$.
```

Algorithm 3 Cross-walk Test

```
Input: $\widehat{\mathbf{\Theta}}^d(z)$ for $z \in [0,1]$.
Initialize: $t = 1$; $E_0(z) = V_1 \times V_2$ for $z \in [0,1]$.
Repeat:

  1. Compute the conditional quantile

     $$c(1-\alpha, E_{t-1}) = \inf\Big\{t \in \mathbb{R} \mid P(T^B_{E_{t-1}} \le t \mid \{(Z_i, \boldsymbol{X}_i)\}_{i \in [n]}) \ge 1 - \alpha\Big\}.$$

  2. Construct the rejected edge set

     $$\mathcal{R}(z) = \Big\{e \in E_{t-1} \mid \sqrt{nh} \cdot |\widehat{\mathbf{\Theta}}^d_e(z)| \cdot \sum_{i \in [n]} K_h(Z_i - z)/n > c(1-\alpha, E_{t-1})\Big\}.$$

  3. Update the rejected edge set $E_t(z) \leftarrow E_{t-1}(z) \backslash \mathcal{R}(z)$ for $z \in [0,1]$.

  4. $t \leftarrow t + 1$.

Until: $E_t(z) = E_{t-1}(z)$ for $z \in [0,1]$ or $|E_t(z)| > N$ for some $z \in [0,1]$.
Output: Reject $H_0$ if $|E_t(z)| > N$ for some $z \in [0,1]$.
```

3 Results

Notch and EGFR signaling pathways have been reported to be connected for the regulation of neural stem cell number and self-renewal(Aguirre, Rubio and Gallo, 2010). We are interested in how genes and sub signaling pathways interact with each other in developmental stages to adulthood over brain regions. Table 1 is the sub-pathways of Notch and EGFR signaling, in which only sub-pathways at the end of pathway hierarchy are selected. There are 103 genes of Notch signaling in the dataset, 343 for EGFR and 17 in both.

	Description	ID	Pathway	Genes
1	Constitutive Signaling by AKT1 E17K in Cancer	5674400	EGFR	24
2	NOTCH1 Intracellular Domain Regulates Transcription	2122947	Notch	43
3	Constitutive Signaling by NOTCH1 PEST Domain Mutants	2644606	Notch	52
4	Constitutive Signaling by NOTCH1 HD+PEST Domain Mutants	2894862	Notch	52
5	Pre-NOTCH Transcription and Translation	1912408	Notch	29
6	Regulation of RAS by GAPs	5658442	EGFR	61
7	Signaling by RAS mutants	6802949	EGFR	48
8	Activated NOTCH1 Transmits Signal to the Nucleus	2122948	Notch	26
9	MAP2K and MAPK activation	5674135	EGFR	38
10	Signaling by moderate kinase activity BRAF mutants	6802946	EGFR	38
11	Signaling by high-kinase activity BRAF mutants	6802948	EGFR	34
12	Paradoxical activation of RAF signaling by kinase inactive BRAF	6802955	EGFR	38
13	EGFR downregulation	182971	EGFR	23
14	SHC1 events in ERBB2 signaling	1250196	EGFR	21
15	Constitutive Signaling by Aberrant PI3K in Cancer	2219530	EGFR	62
16	PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling	6811558	EGFR	84
17	FRS-mediated FGFR1 signaling	5654693	EGFR	23
18	SHC-mediated cascade:FGFR2	5654699	EGFR	25
19	FRS-mediated FGFR2 signaling	5654700	EGFR	24
20	FRS-mediated FGFR4 signaling	5654712	EGFR	21
21	SHC-mediated cascade:FGFR4	5654719	EGFR	22
22	PI-3K cascade:FGFR1	5654689	EGFR	21
23	PI-3K cascade:FGFR2	5654695	EGFR	22
24	SHC-mediated cascade:FGFR1	5654688	EGFR	21
25	RAF activation	5673000	EGFR	24

Table 1: Sub-pathways of Notch and EGFR signaling