GSD Validation

Johnny Li

June 2019

1 Background

Gene regulatory network (GRN): a set of molecular regulators that facilitate chromatin interactions which results in coordinated gene expression. Molecular regulators include complexes of protein and RNA as well as the genomic regions that they may bind to. We make the assumption that a sample's transcriptional profile can be modeled as a linear combination of effects of GRNs.

We believe that chromatin capture data can improve our understanding of gene regulatory networks (GRNs), and we've developed a graph-based matrix decomposition method to integrate HiC and transcriptomic data. Classical decomposition approaches are not guaranteed to find components that correspond well with GRNs. Furthermore, they do not directly provide mechanistic insight into enhancer-promoter interactions, though they could be used to infer these mechanisms from HiC as an additional layer of analysis.

Our hypothesis is that GSD can outperform these baseline methods for 1) identification of genes and their relative transcriptional activity within a GRN and 2) identification of enhancers and their gene targets within a GRN. Note that 1 implies improved clustering of samples based on activities of GRNs in each sample. We will test this hypothesis on synthetic data.

Synthetic data will be generated by first creating a model of a GRN, and then taking linear combinations of transcriptional activities associated with distinct GRNs. We assume that promoters in a GRN are regulated by one or more enhancers by physical interactions. A program can be activating or repressive, and all regulatory interactions within a GRN will be in the same direction.

2 Formalism

Let $\vec{\boldsymbol{l}} = (l_1, \ldots, l_M)$ be a list of M non-overlapping genomic loci (e.g. 5kb bins) and let there be K GRNs that form the basis of regulatory activity. For GRN $k \in \{1, \ldots, K\}$, if k is an activating GRN, let $\vec{\boldsymbol{e}}^k = \{e_i^k | e_i^k \geq 0\}_{i \in \boldsymbol{l}}$ be enhancer activities for all regions in $\vec{\boldsymbol{l}}$, and let $\vec{\boldsymbol{p}}^k = \{p_i^k | p_i^k \geq 0\}_{i \in \boldsymbol{l}}$ be promoter activities. Enhancer and promoter activities for repressive GRNs are defined similarly, but with $e_i^k, p_i^k \leq 0$. A value of 0 indicates no enhancer/promoter activity, so

enhancers and promoters regions of GRN k are $\text{supp}(\vec{\boldsymbol{e}}^k)$ and $\text{supp}(\vec{\boldsymbol{p}}^k)$, respectively. Note that a genomic region may be both an enhancer and promoter, i.e. $\emptyset \subseteq \text{supp}(\vec{\boldsymbol{e}}^k) \cap \text{supp}(\vec{\boldsymbol{p}}^k) \subseteq \vec{\boldsymbol{l}}$.

Let $W^k(i,j) \geq 0$ be the weighted effect of an enhancer $i \in \text{supp}(\vec{\boldsymbol{e}}^k)$ on a promoter $j \in \text{supp}(\vec{\boldsymbol{p}}^k)$. Again, a value of 0 indicates no effect of an enhancer on a promoter, so $\text{supp}(W^k)$ is the set of true physical enhancer-promoter (EP) interactions that play a regulatory role in GRN k. Note that W^k is restricted to be non-negative when GRN k is an activating or repressing GRN. In this model, we assume that promoter activity for GRN k is a function of enhancer activities and their weighted effects, i.e. $\vec{\boldsymbol{p}}^k = f(\vec{\boldsymbol{e}}^k, W^k)$. Thus, $\vec{\boldsymbol{e}}^k$ and W^k are free variables for GRN k.

Finally, for N samples, we will simulate a data matrix of RNA-seq data $\mathbf{X} = \{\vec{\boldsymbol{x}}_i\}_N \in \mathbb{R}^{N,M}$. The transcriptional activity $\vec{\boldsymbol{x}}_i$ for sample i is a linear combination of promoter activities over K GRNs: $\vec{\boldsymbol{x}}_i = \vec{\boldsymbol{z}}_i \cdot \mathbf{P}$, where $\mathbf{P} = \{\vec{\boldsymbol{p}}^1, \ldots, \vec{\boldsymbol{p}}^K\} \in \mathbb{R}^{K,M}$ and $\vec{\boldsymbol{z}}_i = \{z_{ik}\}_{k=1\ldots K} \in \mathbb{R}$ is a vector of scores for each GRN. Then the full data matrix can be generated as

$$\mathbf{X} = \mathbf{Z} \cdot \mathbf{P}. \tag{1}$$

2.1 *In vitro* Data

We will use the following data types to ground our synthetic model:

- ChIP-seq peaks will be used to define all possible enhancer regions, $D_{\text{enh}} \subseteq \vec{\boldsymbol{l}}$.
- Gencode annotations will be used to define all possible promoter regions, $D_{\text{prom}} \subseteq \vec{\boldsymbol{l}}$.
- HiC will be used to define true EP interactions and their weights. HiC data D_{hic} will be encoded as a 2D matrix with rows and columns corresponding to loci $\vec{\boldsymbol{l}}$ and weights given by HiC interaction frequency.

2.2 $\vec{\boldsymbol{e}}^k$

First, we will sample enhancers $\text{supp}(\vec{\boldsymbol{e}}^k)$ from all possible enhancer regions D_{enh}. Then activities $e_i^k : i \in \text{supp}(\vec{\boldsymbol{e}}^k)$ may be uniformly set, or sampled from a Gaussian distribution.

2.3 W^k

First, we will sample true EP interactions $\text{supp}(W^k)$, and the corresponding weights will be set using interaction frequency scores from HiC data. We propose several different EP interaction sampling procedures:

1. For an enhancer $i \in \text{supp}(\vec{\boldsymbol{e}}^k)$, get interaction scores with any possible promoter region in D_{prom}. Sample true interactions weighted by interaction scores.

2. Cluster HiC graph to find small subgraphs of highly interacting regions. For an enhancer $i \in \text{supp}(\vec{\boldsymbol{e}}^k)$ that belongs to cluster c_i, identify promoter regions D_{prom} within the cluster. True EP interactions will be any enhancer $i \in \text{supp}(\vec{\boldsymbol{e}}^k)$ and promoter within the same cluster.

Option 2 lends itself to the model of gene regulation in which an enhancer regulates its target promoters by bringing them close in 3D space resulting in clusters of highly interacting loci. However, given the stochastic nature of interaction events, it is conceivable that within a GRN, an enhancer regulates promoters a and b, but a and b are not highly interacting. Option 1 may be a better model for this behavior.

2.4 $\vec{\boldsymbol{p}}^k$

A promoter's activity will be the weighted sum of its enhancer's activities:

$$p_i^k = \sum_{j \in \text{supp}(\vec{\boldsymbol{e}}^k)} e_j^k \cdot W^k(i,j) \tag{2}$$

3 Performance measures

We will provide GSD and all benchmark algorithms with the data in section 2.1: $D_{\text{enh}}, D_{\text{prom}}, D_{\text{hic}}$. We will evaluate performance by ability to identify

- $\text{supp}(\vec{\boldsymbol{e}}^k)$
- $\text{supp}(\vec{\boldsymbol{p}}^k)$, $\vec{\boldsymbol{p}}^k$
- $\text{supp}(W^k)$