Details of our Motif Clustering Procedure

1 Bayesian Hierarchical Clustering Model

The data for each discovered motif is a count matrix $N_i$ which can have different widths and number of counts compared to other TF motifs. Our clustering will be based on a motif matrices with a fixed width $w$, so we assume each of these $n$ raw motif matrices should contain a submatrix $Y_{i}$ of dimension $w \times 4$ that will be considered the "central motif" upon which the clustering will be based. We include components for both within-motif and between-motif variability of the nucleotide counts $Y_{ijk}$ where $i$ indexes the motif, $j$ indexes the $w$ columns within each motif, and $k$ indexes the four possible nucleotides within each column.

Within-motif level: Count matrices $p(Y_i | \Theta_i) = \prod_{j=1}^{w} p(Y_{ij} | \theta_{ij})$

$$Y_{ij} = (y_{ija}, \ldots, y_{ijt}) \sim \text{Multinomial}(n_i, \theta_{ij} = (\theta_{ija}, \ldots, \theta_{ijt}))$$

Between-motif level: Frequency matrices $p(\Theta_i)$

$$\Theta_i = (\theta_{i1}, \ldots, \theta_{iw}) \sim F(\cdot)$$

where $F(\cdot)$ is an unknown distribution with $w$ dimensions for the columns $\times 4$ dimensions for the nucleotides (constrained to sum to one). This unknown distribution $F(\cdot)$ represents the common structure between the different motifs in the dataset. A common prior for an unknown distribution $F(\cdot)$ is a Dirichlet process $D(\gamma)$ with characteristic smooth measure $\gamma$. Using a Dirichlet process prior enables similar motifs to be clustered together into groups with identical frequency matrices. Ferguson (1974) states that if $\Theta_1, \ldots, \Theta_K$ are $n$ observations, taking on $K$ distinct values $\tilde{\Theta}_1, \ldots, \tilde{\Theta}_L$ drawn from $F(\cdot)$ with a Dirichlet process prior $D(\gamma)$, then the posterior distribution of $F(\cdot)$ has a point mass at each observation, which allows for the clustering of similar observations. If we were to draw an additional $(m+1)$-th observation $\Theta_{m+1}$ from this distribution $D(\gamma^*)$, that new observation would either come from our prior, or would take on a value exactly equal to one of the current $\Theta_i$'s, say $\Theta_k$, in which case $\Theta_k$ and $\Theta_{m+1}$ are defined as being in the same cluster. The conditional distribution $p(\Theta_{m+1} | \Theta_{-m+1})$ of one current observation $\Theta_k$, given all other observations $\Theta_{-i}$, also has $K$ point masses at each of the $\tilde{\Theta}_{-i}$'s that represent the unique values within $\Theta_{-i}$. Any observations $\Theta_m$ and $\Theta_n$ that have the same value are defined as being in the same cluster.

This conditional distribution allows us to implement our model via a Gibbs sampling algorithm, which is a Markov Chain Monte Carlo strategy for simulating unknown parameters (or sets of parameters) one at a time by conditioning on the current values of all the other parameters. For our motif clustering model, the Gibbs sampler could intuitively be based on $p(\Theta_i | \Theta_{-i})$. However, since our $\Theta_i$'s are actually unknown, a more efficient
clustering procedure involves drawing values of the clustering indicators directly, without dealing with drawing a frequency matrix $\Theta_i$ for each motif $i$ at each iteration. We denote our clustering indicators as $z_i$ where $z_i = l$ if $\Theta_i$ takes on the same value as $\tilde{\Theta}_l$ (and hence is in the $l$-th cluster) or $z_i = 0$ if $\theta_i$ is drawn from the prior distribution (and hence forms a new cluster). We would like to sample directly from the conditional distribution of these clustering indicators, i.e., we want to sample from $p(z_i|z_{-i}, Y)$ where we use the notation $z_{-i}$ to mean all the clustering indicators $z$ except the $i$-th one. With our model outlined above, the conditional probability of forming a new cluster with the $i$-th motif is

$$p(z_i = 0|z_{-i}, Y) \propto \frac{1}{n} \prod_{j=1}^w \frac{\prod_k \Gamma(Y_{ijk} + \alpha) \Gamma(4\alpha)}{\Gamma(\sum_k Y_{ijk} + 4\alpha) \Gamma(\alpha)^4}$$

whereas the probability that the $i$-th motif joining the $l$-th existing cluster that has a total count matrix of $\tilde{Y}_l$ is

$$p(z_i = l|z_{-i}, Y) \propto \frac{n_l}{n} \prod_{j=1}^w \frac{\prod_k \Gamma(Y_{ijk} + \tilde{Y}_{ijk} + \alpha) \Gamma(\sum_k \tilde{Y}_{ijk} + 4\alpha)}{\Gamma(\sum_k Y_{ijk} + \tilde{Y}_{ijk} + 4\alpha) \prod_k \Gamma(\tilde{Y}_{ijk} + \alpha)}$$

It should be noted that the usual form of the Dirichlet process prior has an additional parameter $a$ which is often referred to as the “prior weight” (see Liu (1996) for details). This parameter influences the tendency of the algorithm to form larger or smaller clusters by giving more or less weight to the probability $p(z_i = 0|z_{-i}, Y)$ given above. In our situation, we prefer larger clusters of potentially co-regulated genes, and so we use a small prior weight of $a = 1$. We account for spurious relationships caused by this decision by filtering our clusters as described in Section 5 below.

A complete iteration of our Gibbs sampling algorithm results in a complete sample $z$ of our clustering indicators, which also represents a complete partition of our motif matrices into clusters. The clustering algorithm was run until “convergence” i.e., until we were reasonably certain that both algorithms had moved to a good area of the posterior distribution. This convergence was evaluated by running multiple chains from very different starting configurations of our clustering indicators $z$ (e.g., all motifs in their own cluster, all motifs in one cluster, random clusterings, etc.) and examining the between-chain similarity of individual clusters, as well as statistics for the entire set of clusters, such as the number of clusters and the average cluster size.

## 2 Motif Alignment

An additional component of our model addresses the fact that we do not necessarily know which “central motif” $Y_i$ of length $w$ to use within the raw alignment matrix of length $n_i > w$ for motif $i$. We use the notation $a_i = j$ to mean that we are using the columns $j, j+1, j+w-1$ of our raw motif matrix $N_i$ as our central motif $Y_i$. For example, if
our clustering algorithm is based on a fixed width of \( w = 6 \) and our \( i \)-th raw motif matrix \( N_i \) has 8 positions, than we have three possible choices for our central motif: \( a_i = 1 \) (\( Y_i = \) columns 1 to 6 of \( N_i \)), \( a_i = 2 \) (\( Y_i = \) columns 2 to 7 of \( N_i \)), or \( a_i = 3 \) (\( Y_i = \) columns 3 to 8 of \( N_i \)).

Our hierarchical clustering model assumes that the \( Y_i \) for each motif is known (and thus \( a_i \) is known), so we need an additional step where, for each raw data matrix, the best location of the central motif \( a_i \) is drawn conditional the other motifs \( Y_{-i} \) and clustering indicators \( z_{-i} \) for the other motifs. Let \( Y_i^{a_i} \) denote the central motif \( Y_i \) that corresponds to the choice of a particular \( a_i \), then the posterior probability \( p(a_i) = p(a_i|z_{-i}, Y_{-i}) \) of \( a_i \) is

\[
p(a_i) \propto \frac{1}{n} \left[ \frac{\Gamma(4\alpha)}{\Gamma(\alpha^4)} \right]^w \prod_k \frac{\Gamma(Y_{ijk}^a + \alpha)}{\Gamma(\sum_k Y_{ijk}^a + 4\alpha)} + \sum_{c=1}^C \frac{n_c}{n} \prod_j^w \frac{\Gamma(Y_{ijk}^a + \tilde{Y}_{ck} + \alpha)}{\Gamma(\sum_k Y_{ijk}^a + Y_{ck} + 4\alpha)} \frac{\Gamma(\sum_k \tilde{Y}_{ck} + 4\alpha)}{\prod_k \Gamma(\tilde{Y}_{ck} + \alpha)}
\]

This alignment procedure is performed every tenth iteration of the Gibbs sampler described in the previous section.

### 3 Extension to Variable Width

We also include an optional extension of our model to allow the motif width within each cluster \( w_c \) (\( c = 1, \ldots, C \) where \( C \) is the current number of clusters) to be an unknown variable. Each cluster width \( w_c \) is modeled as being independent with prior distribution \( w_c \sim \text{Poisson}(\lambda) \) where \( \lambda \) is the expected \( \text{a priori} \) width of the motif in each cluster, which we assume is fixed and specified. We let \( B_{ck} \) be the “background” counts of nucleotide \( k \) over all columns of the raw matrix \( bN_i \) that are not included in the central motif matrix \( Y_{ic} \), which now must be taken into account by our model since each motif width is allowed to vary. These background columns represent the edges of the raw motif matrices that we do not believe are well conserved. We assume the background counts \( B_i = (B_{i1}, \ldots, B_{i4}) \) are a multinomial realization from an underlying vector of background nucleotide frequencies \( \theta_0 = (\theta_{01}, \ldots, \theta_{04}) \). With these added distributions, the posterior probability for the width of a particular cluster \( c \), conditional on the current members of that cluster \( z_c \) and their central motif alignments \( a_c \), is

\[
p(w_c|z_c, a_c) \propto \frac{\prod_j^{w_c} \Gamma(\tilde{Y}_{ck} + \alpha)}{\Gamma(\sum_k \tilde{Y}_{ck} + 4\alpha)} \cdot \frac{\Gamma(4\alpha)}{\Gamma(\tilde{Y}_{ck} + \alpha^4)} \cdot \prod_k \theta_{0ck}^{B_{ck}} \cdot \frac{\lambda^{w_c} e^{-\lambda}}{\Gamma(w_c)}
\]

where \( \tilde{Y}_{ck} \) and \( B_{ck} \) are, respectively, the central motif and background nucleotide counts in cluster \( c \). We implement this added component of our model by an additional step in our Gibbs sampling algorithm that, for each cluster \( c \), samples a new value of \( w_c \) from the conditional posterior distribution \( p(w_c|z_c, a_c) \).
4 Posterior Value of a Partition

The best set of clusters for a given motif dataset is the partition $\hat{z}$ with the highest posterior value. The posterior value $p(z|Y)$ of $z$ is calculated as the product of the likelihood value $p(Y|z)$ and the prior value $p(z|\alpha)$. If our partition $z$ has $L$ clusters, each with $n_l$ members and count matrix $\tilde{Y}_l$ (the sum of all $w \times 4$ count matrices in cluster $l$), then the posterior value in our fixed-width model is

$$p(z|Y) \propto \prod_{l=1}^{L} \prod_{j=1}^{w} \frac{\prod_{k} \Gamma(\tilde{Y}_{jk} + \alpha)}{\prod_{k} \Gamma(\sum_{k} \tilde{Y}_{jk} + 4\alpha)} \times \frac{L}{\prod_{i=1}^{n}(n_l - 1)!}$$

The posterior value of $z$ for our variable-width model is similar, but has additional terms for the background nucleotide counts $B$ and variable widths $w_c$.

5 Strength of Clusters

We measure the strength of each cluster by calculating the logarithm of the Bayes factor (Kass and Raftery, 1995) for the current cluster $l$, with members $z = (z_1, z_2, \ldots, z_{n_l})$, versus each member of the cluster forming its own cluster. For a cluster of $m$ motifs $(Y_1, \ldots, Y_m)$ and clustering indicators $z = (z_1, \ldots, z_m)$,

$$\text{Strength(Cluster } l) = \log \left[ \frac{P(z \text{ all same } | Y)}{P(z \text{ all different } | Y)} \right] = \log \left[ \frac{\prod_{j=1}^{w} \prod_{k} \Gamma(\tilde{Y}_{jk} + \alpha)}{\prod_{j=1}^{w} \prod_{k} \Gamma(\sum_{k} \tilde{Y}_{jk} + 4\alpha)} \times (m - 1)! \right]$$

where $\tilde{Y}$ again denotes the count matrix for the entire cluster together. We also calculated, for each motif in our best clusters, the probability of the motif belonging to its assigned cluster (as opposed to any of the other existing clusters). Any motifs with less than 75% probability of belonging to their assigned clusters were removed.

References

