CENTIPEDE 2.0

ANIL RAJ AND HEEJUNG SHIM

1. Overview

CENTIPEDE aims to infer motif sites bound by transcription factors based on the DNase I cleavage patterns measured from DNase-Seq assays. The model relies on two assumptions: (1) sites bound by transcription factors have higher DNase I sensitivity than unbound sites, and (2) each transcription factor has a characteristic DNase I cleavage profile at bound sites.

Given a putative binding site, CENTIPEDE models the number of reads mapped to each base pair as a mixture of two distributions, where the mixing proportions capture the probability of the factor being bound. Specifically, conditional on being bound, the total number of reads are modelled as drawn from a negative binomial distribution and the read profile, conditional on the total number of reads, are modelled as drawn from a multinomial distribution.

$$p(X_n, T_n^X | Z_n = 1) = p(X_n | Z_n = 1, T_n^X) p(T_n^X | Z_n = 1) \tag{1.1}$$
$$= \text{mult}(\pi; T_n^X) \text{NegBin}(\alpha, \tau) \tag{1.2}$$
$$p(X_n, T_n^X | Z_n = 0) = p(X_n | Z_n = 0, T_n^X) p(T_n^X | Z_n = 0) \tag{1.3}$$
$$= \text{mult}(\pi_o; T_n^X) \text{NegBin}(\alpha_o, \tau_o), \tag{1.4}$$

where $X_n \in \mathbb{N}_0^L$, $T_n^X = \sum_l X_{nl}$, $\pi \in \mathbb{S}^L$, $\mathbb{S}^L$ is the standard L-simplex, $\pi_o = \frac{1}{L}\mathbf{1}^L$, $\mathbf{1}^L$ is the L-dimensional vector of ones, $\alpha, \alpha_o \in \mathbb{R}^+$, and $\tau, \tau_o \in [0, 1]$. L is the length of the site around the binding motif.

A key limitation of this generative model is that it does not appropriately model the underlying spatially structured DNase I cleavage pattern, and the associated correlation between DNase I cutting rates at two different positions around the motif, induced by binding of a particular transcription factor. One way to model spatially structured signals would be to place a logistic-normal prior on the multinomial parameter π; the covariance matrix in the logistic-normal distribution can then be estimated from the data. However, the non-conjugacy of this prior makes exact inference intractable, leading us to use approximate techniques like variational inference to compute the posterior probabilities of the latent variables [1].

Alternately, we propose to replace the multinomial part of the CENTIPEDE model with a hierarchical multi-scale poisson model. Here, we describe the model and derive maximum likelihood estimators for parameters. See Kolaczyk (1999) for a detailed discussion of this inhomogenous multi-scale poisson process model.

Date: May 29, 2013.

[1] **This approach will be derived and discussed later**

2. Poisson-Binomial Model

Specifically, keeping Kolaczyk's notation for the parameters, let Y_{jk} be defined in terms of the data X_l as

$$Y_{J-1,k} = X_{2k} + X_{2k+1} \tag{2.1}$$

$$Y_{j,k} = Y_{j+1,2k} + Y_{j+1,2k+1}, \qquad j \in \{0, \ldots, J-2\}, \tag{2.2}$$

where $j \in \{0, \ldots, J-1\}$, $J = \log_2 L$, and, $k \in \{0, \ldots, 2^j - 1\}$. Note that, the range of the k index depends on the value of the j index. Conditional on the total number of reads at a site, the likelihood function of parameter $R = (R_{jk})$ factorizes as follows:

$$p(X_n | Z_n = 1, T_n^X, R) = \prod_{j,k} p(Y_{njk} | Z_n = 1, T_{njk}, R_{jk}), \tag{2.3}$$

where

$$R_{jk} = \gamma_{jk}\delta\left(\frac{1}{2}\right) + (1 - \gamma_{jk})\delta\left(B_{jk}\right) \tag{2.4}$$

$$p(\gamma_{jk}) = \text{Ber}(\pi_j), \tag{2.5}$$

$$\tag{2.6}$$

and, $\delta(a)$ is the Dirac delta function centered at a. For each scale j, the product is only over even values of the index k. For ease of notation, we introduce a variable $T_{njk} = Y_{n,j-1,k/2}$, the total number of reads at a coarser resolution from which reads at the finer resolution are drawn. Therefore, at a given scale and location, Y_{njk} is effectively drawn from a mixture of a binomial distribution, with parameter $\frac{1}{2}$ and a binomial distribution with parameter B_{jk}. If R_{jk} were treated as a fixed parameter instead of a random variable, we get back the original multinomial model in CENTIPEDE.

We will derive and explore the following three variations of the above model.

(1) B_{jk} is a fixed parameter to be estimated by maximizing the likelihood of the model.
(2) B_{jk} is a latent random variable, $B_{jk} \sim \text{beta}(\mu_j, \mu_j)$.
(3) To account for overdispersion across sites, we allow for site-specific values for R, drawn from a beta distribution with scale-dependent mean and scale- and location-dependent variances.

Specifically, the three variations are:

(1)
$$\begin{aligned} Y_{njk} | T_{njk}, R_{jk} &\sim \text{binom}(R_{jk}; T_{njk}) \\ R_{jk} &= \gamma_{jk}\delta\left(\frac{1}{2}\right) + (1 - \gamma_{jk})\delta\left(B_{jk}\right) \\ \gamma_{jk} &\sim \text{Ber}(\pi_j) \end{aligned}$$

Thus, $Y_{njk} | T_{njk}, \pi_j, B_{jk} \sim \pi\text{binom}(\frac{1}{2}; T_{njk}) + (1 - \pi)\text{binom}(B_{jk}; T_{njk})$

(2)
$$\begin{aligned} Y_{njk} | T_{njk}, R_{jk} &\sim \text{binom}(R_{jk}; T_{njk}) \\ R_{jk} &= \gamma_{jk}\delta\left(\frac{1}{2}\right) + (1 - \gamma_{jk})\delta\left(B_{jk}\right) \\ \gamma_{jk} &\sim \text{Ber}(\pi_j) \\ B_{jk} &\sim \text{beta}(\mu_j, \mu_j) \end{aligned}$$

Thus, $Y_{njk} | T_{njk}, \pi_j, B_{jk} \sim \pi\text{binom}(\frac{1}{2}; T_{njk}) + (1 - \pi)\text{BetaBinom}(\mu_j, \mu_j; T_{njk})$

$$
\begin{aligned}
(3) \qquad Y_{njk}|T_{njk}, R_{njk} &\sim \text{binom}(R_{njk}; T_{njk}) \\
R_{njk} &\sim \text{beta}(R_{jk}\tau_{jk}, (1-R_{jk})\tau_{jk}) \\
R_{jk} &= \gamma_{jk}\delta\left(\frac{1}{2}\right) + (1-\gamma_{jk})B_{jk} \\
\gamma_{jk} &\sim \text{Ber}(\pi_j) \\
\text{Thus,} \quad Y_{njk}|T_{njk}, \pi_j, B_{jk} &\sim \pi\text{BetaBinom}(\frac{1}{2}\tau_{jk}, \frac{1}{2}\tau_{jk}; T_{njk}) \\
&\quad + (1-\pi)\text{BetaBinom}(B_{jk}\tau_{jk}, (1-B_{jk})\tau_{jk}; T_{njk})
\end{aligned}
$$

In the framework of CENTIPEDE, we now have two sets of latent variables: one specifying whether a factor is bound at a site or not, and the other specifying the degree of smoothness in DNase I cleavage rates at different scales and locations around a specific motif. Let us assume that the latent variable Z_n is observed to be 1 for a set of sites. Maximum likelihood estimates for the parameters in each model can be calculated by maximizing the lower bound on the likelihood, obtained by proposing a family of posterior distributions $q(\gamma_{jk})$, using the EM algorithm.

$$q(\gamma_{jk}) = \text{Ber}(\tilde{\gamma}_{jk}) \tag{2.7}$$

2.1. **Model 1.** For this model, the lower bound to the log-likelihood can be derived as follows:

$$\mathcal{L} = \sum_{n,j,k} \log p(Y_{njk}|\pi_j, B_{jk}; T_{\cdot jk}) \tag{2.8}$$

$$= \sum_{n,j,k} \log \sum_{\gamma_{jk}} p(Y_{njk}|\gamma_{jk}, B_{jk}; T_{njk}) p(\gamma_{jk}|\pi_j) \tag{2.9}$$

$$\geq \sum_{n,j,k} \sum_{\gamma_{jk}} q(\gamma_{jk}) \left(\log p(Y_{njk}|\gamma_{jk}, B_{jk}; T_{njk}) + \log \frac{p(\gamma_{jk}|\pi_j)}{q(\gamma_{jk})} \right) \tag{2.10}$$

$$= \sum_{n,j,k} \mathbf{E}_{q(\gamma)} \left[\log p(Y_{njk}|\gamma_{jk}; T_{njk}) \right] + \mathbf{E}_{q(\gamma)} \left[\log \frac{p(\gamma_{jk}|\pi_j)}{q(\gamma_{jk})} \right] \tag{2.11}$$

$$= \sum_{n,j,k} \mathbf{E}_{q(\gamma)} \left[\gamma_{jk} \log p(Y_{njk}|\frac{1}{2}; T_{njk}) + (1-\gamma_{jk}) \log p(Y_{njk}|B_{jk}; T_{njk}) \right] + \mathbf{E}_{q(\gamma)} \left[\log \frac{p(\gamma_{jk}|\pi_j)}{q(\gamma_{jk})} \right] \tag{2.12}$$

$$= \sum_{n,j,k} \tilde{\gamma}_{jk} \log p(Y_{njk}|\frac{1}{2}; T_{njk}) + (1-\tilde{\gamma}_{jk}) \log p(Y_{njk}|B_{jk}; T_{njk}) + \mathbf{E}_{q(\gamma)} \left[\log \frac{p(\gamma_{jk}|\pi_j)}{q(\gamma_{jk})} \right]. \tag{2.13}$$

The distributions in the first and second terms are binomial distributions and the relevant likelihood functions can be written as follows.

$$\mathcal{L}^b_{njk} = \log p(Y_{njk}|\frac{1}{2}; T_{njk}) = \mathcal{C}_{njk} + T_{njk} \log\left(\frac{1}{2}\right) \tag{2.14}$$

$$\mathcal{L}^{bb}_{njk}(B_{jk}) = \log p(Y_{njk}|B_{jk}; T_{njk}) = \mathcal{C}_{njk} + Y_{njk} \log(B_{jk}) + (T_{njk} - Y_{njk}) \log(1-B_{jk}) \tag{2.15}$$

Thus,

$$\mathcal{L} = \sum_{n,j,k} \tilde{\gamma}_{jk} \mathcal{L}^b_{njk} + (1-\tilde{\gamma}_{jk}) \mathcal{L}^{bb}_{njk}(B_{jk}) + \tilde{\gamma}_{jk} \log \frac{\pi_j}{\tilde{\gamma}_{jk}} + (1-\tilde{\gamma}_{jk}) \log \frac{(1-\pi_j)}{(1-\tilde{\gamma}_{jk})} \tag{2.16}$$

Maximizing $\mathcal{L}$ with respect to $\tilde{\gamma}_{jk}$ while keeping other parameters fixed gives

$$\frac{\partial \mathcal{L}}{\partial \tilde{\gamma}_{jk}} = \frac{1}{N}\sum_n \mathcal{L}^b_{njk} - \mathcal{L}^{bb}_{njk}(B_{jk}) + \log \pi_j - \log(1-\pi_j) - \log \tilde{\gamma}_{jk} + \log(1-\tilde{\gamma}_{jk}) = 0 \tag{2.17}$$

$$\log \frac{\tilde{\gamma}_{jk}}{1-\tilde{\gamma}_{jk}} = \log \frac{\pi_j}{1-\pi_j} + \frac{1}{N}\sum_n \mathcal{L}^b_{njk} - \mathcal{L}^{bb}_{njk}(B_{jk}) \tag{2.18}$$

Maximizing $\mathcal{L}$ with respect to π_j while keeping other parameters fixed gives

$$\frac{\partial \mathcal{L}}{\partial \pi_j} = \frac{1}{\pi_j}\sum_k \tilde{\gamma}_{jk} - \frac{1}{1-\pi_j}\sum_k (1-\tilde{\gamma}_{jk}) = 0 \tag{2.19}$$

$$\pi_j = \frac{1}{K}\sum_k \tilde{\gamma}_{jk} \tag{2.20}$$

Maximizing $\mathcal{L}$ with respect to B_{jk} while keeping other parameters fixed gives

$$\frac{\partial \mathcal{L}}{\partial B_{jk}} = \frac{1-\gamma_{jk}}{B_{jk}}\sum_n Y_{njk} - \frac{(1-\gamma_{jk})}{1-B_{jk}}\sum_n (T_{njk} - Y_{njk}) = 0 \tag{2.21}$$

$$B_{jk} = \frac{\sum_n Y_{njk}}{\sum_n T_{njk}} \tag{2.22}$$

2.2. **Model 2.** Following the derivation in the previous model, the lower bound to the log-likelihood can be written as follows:

$$\mathcal{L} \geq \sum_{n,j,k} \tilde{\gamma}_{jk} \log p(Y_{njk}|\frac{1}{2}; T_{njk}) + (1-\tilde{\gamma}_{jk}) \log p(Y_{njk}|\mu_j; T_{njk}) + \mathbf{E}_{q(\gamma)}\left[\log \frac{p(\gamma_{jk}|\pi_j)}{q(\gamma_{jk})}\right]. \tag{2.23}$$

The distribution in the first term is a binomial distribution and the distribution in the second is a "symmetric" beta-binomial distribution. Since the parameters of these distributions are fixed, we will replace the relevant likelihood functions as follows.

$$\mathcal{L}^b_{njk} = \log p(Y_{njk}|\frac{1}{2}; T_{njk}) \tag{2.24}$$

$$\mathcal{L}^{bb}_{njk} = \log p(Y_{njk}|\mu_j; T_{njk}) \tag{2.25}$$

Thus,

$$\mathcal{L} = \sum_{n,j,k} \tilde{\gamma}_{jk}\mathcal{L}^b_{njk} + (1-\tilde{\gamma}_{jk})\mathcal{L}^{bb}_{njk} + \tilde{\gamma}_{jk} \log \frac{\pi_j}{\tilde{\gamma}_{jk}} + (1-\tilde{\gamma}_{jk}) \log \frac{(1-\pi_j)}{(1-\tilde{\gamma}_{jk})} \tag{2.26}$$

Maximum likelihood estimates for γ_{jk} and π_j, as derived earlier, can be written as

$$\log \frac{\tilde{\gamma}_{jk}}{1-\tilde{\gamma}_{jk}} = \log \frac{\pi_j}{1-\pi_j} + \frac{1}{N}\sum_n \mathcal{L}^b_{njk} - \mathcal{L}^{bb}_{njk} \tag{2.27}$$

$$\pi_j = \frac{1}{K}\sum_k \tilde{\gamma}_{jk} \tag{2.28}$$

2.3. **Model 3.** Again, as before, the lower bound to the log-likelihood can be written as follows:

$$\mathcal{L} \geq \sum_{n,j,k} \tilde{\gamma}_{jk} \log p(Y_{njk}|\frac{1}{2}, \tau_{jk}; T_{njk}) + (1-\tilde{\gamma}_{jk}) \log p(Y_{njk}|B_{jk}, \tau_{jk}; T_{njk}) + \mathbf{E}_{q(\gamma)}\left[\log \frac{p(\gamma_{jk}|\pi_j)}{q(\gamma_{jk})}\right]. \tag{2.29}$$

The distributions in the first and second terms are beta-binomial distributions and the relevant likelihood functions can be written as follows.

$$\begin{aligned}\mathcal{L}^b_{njk}(\tau_{jk}) &= \log p(Y_{njk}|\frac{1}{2}, \tau_{jk}; T_{njk}) = \mathcal{C}_{njk} + \log\Gamma(Y_{njk}+0.5\tau_{jk}) + \log\Gamma(T_{njk}-Y_{njk}+0.5\tau_{jk}) \\ &\quad - \log\Gamma(T_{njk}+\tau_{jk}) + \log\Gamma(\tau_{jk}) - 2*\log\Gamma(0.5\tau_{jk}) \qquad (2.30)\\ \mathcal{L}^{bb}_{njk}(B_{jk}, \tau_{jk}) &= \log p(Y_{njk}|B_{jk}; T_{njk}) = \mathcal{C}_{njk} + \log\Gamma(Y_{njk}+B_{jk}\tau_{jk}) + \log\Gamma(T_{njk}-Y_{njk}+(1-B_{jk})\tau_{jk}) \\ &\quad - \log\Gamma(T_{njk}+\tau_{jk}) + \log\Gamma(\tau_{jk}) - \log\Gamma(B_{jk}\tau_{jk}) - \log\Gamma((1-B_{jk})\tau_{jk}) \qquad (2.31)\end{aligned}$$

Thus,

$$\mathcal{L} = \sum_{n,j,k} \tilde{\gamma}_{jk}\mathcal{L}^b_{njk}(\tau_{jk}) + (1-\tilde{\gamma}_{jk})\mathcal{L}^{bb}_{njk}(B_{jk},\tau_{jk}) + \tilde{\gamma}_{jk}\log\frac{\pi_j}{\tilde{\gamma}_{jk}} + (1-\tilde{\gamma}_{jk})\log\frac{(1-\pi_j)}{(1-\tilde{\gamma}_{jk})} \tag{2.32}$$

Maximum likelihood estimates for γ_{jk} and π_j are the same as derived in the earlier two models. Since the remaining parameters B_{jk} and τ_{jk} occur within $\log\Gamma(\cdot)$ functions, closed form update equations for these parameters cannot be derived. Instead, we'll maximize the likelihood with respect to these parameters using generalized convex optimization algorithms. The gradient of the likelihood with respect to each of these parameters can be derived as follows.

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial B_{jk}} &= (1-\tilde{\gamma}_{jk})\sum_n [\tau_{jk}\Psi(Y_{njk}+B_{jk}\tau_{jk}) - \tau_{jk}\Psi(T_{njk}-Y_{njk}+(1-B_{jk})\tau_{jk}) \\ &\quad - \tau_{jk}\Psi(B_{jk}\tau_{jk}) + \tau_{jk}\Psi((1-B_{jk})\tau_{jk})] \qquad (2.33)\\ \frac{\partial\mathcal{L}}{\partial\tau_{jk}} &= \sum_n \tilde{\gamma}_{jk}\,[0.5\Psi(Y_{njk}+0.5\tau_{jk}) + 0.5\Psi(T_{njk}-Y_{njk}+0.5\tau_{jk}) \\ &\quad - \Psi(T_{njk}+\tau_{jk}) + \Psi(\tau_{jk}) - \Psi(0.5\tau_{jk})] \\ &\quad + (1-\tilde{\gamma}_{jk})\,[B_{jk}\Psi(Y_{njk}+B_{jk}\tau_{jk}) + (1-B_{jk})\Psi(T_{njk}-Y_{njk}+(1-B_{jk})\tau_{jk}) \\ &\quad - \Psi(T_{njk}+\tau_{jk}) + \Psi(\tau_{jk}) - B_{jk}\Psi(B_{jk}\tau_{jk}) - (1-B_{jk})\Psi((1-B_{jk})\tau_{jk})] \qquad (2.34)\end{aligned}$$

2.4. **Posterior distribution of R_{jk}.** The posterior distribution of R_{jk} can be computed as follows.

$$\begin{aligned}p(R_{jk}|Y_{\cdot jk}, T_{\cdot jk}) &= p(R_{jk}, \gamma_{jk}=1|Y_{\cdot jk}, T_{\cdot jk}) + p(R_{jk}, \gamma_{jk}=0|Y_{\cdot jk}, T_{\cdot jk}) \qquad (2.35)\\ &= p(R_{jk}|\gamma_{jk}=1, Y_{\cdot jk}, T_{\cdot jk})p(\gamma_{jk}=1|Y_{\cdot jk}, T_{\cdot jk}) \\ &\quad + p(R_{jk}|\gamma_{jk}=0, Y_{\cdot jk}, T_{\cdot jk})p(\gamma_{jk}=0|Y_{\cdot jk}, T_{\cdot jk}) \qquad (2.36)\end{aligned}$$

where

$$\begin{aligned}p(\gamma_{jk}=0|Y_{\cdot jk}, T_{\cdot jk}) &= \frac{p(Y_{\cdot jk}|\gamma_{jk}=0, T_{\cdot jk})(1-\pi_j)}{p(Y_{\cdot jk}|\gamma_{jk}=1, T_{\cdot jk})\pi_j + p(Y_{\cdot jk}|\gamma_{jk}=0, T_{\cdot jk})(1-\pi_j)} \qquad (2.37)\\ &= 1 - p(\gamma_{jk}=1|Y_{\cdot jk}, T_{\cdot jk}). \qquad (2.38)\end{aligned}$$

We will derive $p(Y_{\cdot jk}|\gamma_{jk}=0, T_{\cdot jk})$, $p(Y_{\cdot jk}|\gamma_{jk}=1, T_{\cdot jk})$, $p(R_{jk}|\gamma_{jk}=0, Y_{\cdot jk}, T_{\cdot jk})$, and $p(R_{jk}|\gamma_{jk}=1, Y_{\cdot jk}, T_{\cdot jk})$ for each model.

2.4.1. *Model 1.*

$$p(Y_{\cdot jk}|\gamma_{jk}=0,T_{\cdot jk}) = [\prod_n \binom{T_{njk}}{Y_{njk}}](B_{jk})^{\sum_n Y_{njk}}(1-B_{jk})^{\sum_n (T_{njk}-Y_{njk})} \tag{2.39}$$

$$p(Y_{\cdot jk}|\gamma_{jk}=1,T_{\cdot jk}) = [\prod_n \binom{T_{njk}}{Y_{njk}}](\frac{1}{2})^{\sum_n T_{njk}} \tag{2.40}$$

$$p(R_{jk}|\gamma_{jk}=0,Y_{\cdot jk},T_{\cdot jk}) = \delta\left(B_{jk}\right) \tag{2.41}$$

$$p(R_{jk}|\gamma_{jk}=1,Y_{\cdot jk},T_{\cdot jk}) = \delta\left(\frac{1}{2}\right) \tag{2.42}$$

Then, the posterior mean can be written as

$$\mathbf{E}\left[R_{jk}|Y_{\cdot jk},T_{\cdot jk}\right] = \frac{(\frac{1}{2})^{(1+\sum_n T_{njk})}\pi_j + (B_{jk})^{(1+\sum_n Y_{njk})}(1-B_{jk})^{\sum_n (T_{njk}-Y_{njk})}(1-\pi_j)}{(\frac{1}{2})^{\sum_n T_{njk}}\pi_j + (B_{jk})^{\sum_n Y_{njk}}(1-B_{jk})^{\sum_n (T_{njk}-Y_{njk})}(1-\pi_j)}. \tag{2.43}$$

2.4.2. *Model 2.*

$$p(Y_{\cdot jk}|\gamma_{jk}=0,T_{\cdot jk}) = [\prod_n \binom{T_{njk}}{Y_{njk}}]\frac{B(\mu_j+\sum_n Y_{njk},\mu_j+\sum_n (T_{njk}-Y_{njk}))}{B(\mu_j,\mu_j)} \tag{2.44}$$

$$p(Y_{\cdot jk}|\gamma_{jk}=1,T_{\cdot jk}) = [\prod_n \binom{T_{njk}}{Y_{njk}}](\frac{1}{2})^{\sum_n T_{njk}} \tag{2.45}$$

$$p(R_{jk}|\gamma_{jk}=0,Y_{\cdot jk},T_{\cdot jk}) = \frac{\mu_j+\sum_n Y_{njk}}{2\mu_j+\sum_n T_{njk}} \tag{2.46}$$

$$p(R_{jk}|\gamma_{jk}=1,Y_{\cdot jk},T_{\cdot jk}) = \delta\left(\frac{1}{2}\right) \tag{2.47}$$

Then, the posterior mean can be written as

$$\mathbf{E}\left[R_{jk}|Y_{\cdot jk},T_{\cdot jk}\right] = \frac{(\frac{1}{2})^{(1+\sum_n T_{njk})}\pi_j + \frac{B(1+\mu_j+\sum_n Y_{njk},\mu_j+\sum_n (T_{njk}-Y_{njk}))}{B(\mu_j,\mu_j)}(1-\pi_j)}{(\frac{1}{2})^{\sum_n T_{njk}}\pi_j + \frac{B(\mu_j+\sum_n Y_{njk},\mu_j+\sum_n (T_{njk}-Y_{njk}))}{B(\mu_j,\mu_j)}(1-\pi_j)}, \tag{2.48}$$

where $B(\cdot,\cdot)$ is the beta function.

2.4.3. *Model 3.*

$$p(Y_{\cdot jk}|\gamma_{jk}=0,T_{\cdot jk}) = \prod_n \int \binom{T_{njk}}{Y_{njk}} R_{njk}^{Y_{njk}}(1-R_{njk})^{T_{njk}-Y_{njk}}\frac{R_{njk}^{B_{jk}\tau_{jk}-1}(1-R_{njk})^{(1-B_{jk})\tau_{jk}-1}}{B(B_{jk}\tau_{jk},(1-B_{jk})\tau_{jk})}\mathrm{d}R_{njk} \tag{2.49}$$

$$= \prod_n \binom{T_{njk}}{Y_{njk}} \frac{B(B_{jk}\tau_{jk}+Y_{njk},(1-B_{jk})\tau_{jk}+T_{njk}-Y_{njk})}{B(B_{jk}\tau_{jk},(1-B_{jk})\tau_{jk})} \tag{2.50}$$

$$p(Y_{\cdot jk}|\gamma_{jk}=1,T_{\cdot jk}) = \prod_n \binom{T_{njk}}{Y_{njk}} \frac{B(\frac{1}{2}\tau_{jk}+Y_{njk},\frac{1}{2}\tau_{jk}+T_{njk}-Y_{njk})}{B(\frac{1}{2}\tau_{jk},\frac{1}{2}\tau_{jk})} \tag{2.51}$$

$$p(R_{jk}|\gamma_{jk}=0,Y_{\cdot jk},T_{\cdot jk}) = \delta\left(B_{jk}\right) \tag{2.52}$$

$$p(R_{jk}|\gamma_{jk}=1,Y_{\cdot jk},T_{\cdot jk}) = \delta\left(\frac{1}{2}\right) \tag{2.53}$$

Then, the posterior mean can be written as

$$\mathbf{E}\left[R_{jk}|Y_{\cdot jk},T_{\cdot jk}\right]=\frac{\frac{1}{2}\pi_j\prod_n\frac{B(\frac{1}{2}\tau_{jk}+Y_{njk},\frac{1}{2}\tau_{jk}+T_{njk}-Y_{njk})}{B(\frac{1}{2}\tau_{jk},\frac{1}{2}\tau_{jk})}+B_{jk}(1-\pi_j)\prod_n\frac{B(B_{jk}\tau_{jk}+Y_{njk},(1-B_{jk})\tau_{jk}+T_{njk}-Y_{njk})}{B(B_{jk}\tau_{jk},(1-B_{jk})\tau_{jk})}}{\pi_j\prod_n\frac{B(\frac{1}{2}\tau_{jk}+Y_{njk},\frac{1}{2}\tau_{jk}+T_{njk}-Y_{njk})}{B(\frac{1}{2}\tau_{jk},\frac{1}{2}\tau_{jk})}+(1-\pi_j)\prod_n\frac{B(B_{jk}\tau_{jk}+Y_{njk},(1-B_{jk})\tau_{jk}+T_{njk}-Y_{njk})}{B(B_{jk}\tau_{jk},(1-B_{jk})\tau_{jk})}} \tag{2.54}$$

3. Centipede-PBM

Each of the above three models can be straightforwardly incorporated into CENTIPEDE's learning framework by first deriving how the Poisson-Binomial model modifies the likelihood function. The key change is restricted to the multinomial part of CENTIPEDE's likelihood function. Here, we derive the form of the change for Model 1, and then apply the change directly to Models 2 and 3

$$\mathcal{L}=\sum_n\log\sum_{Z_n}p(X_n|Z_n,T_n^X)p(T_n^X|Z_n)p(Z_n|S_n,\beta) \tag{3.1}$$

$$\geq\sum_n\sum_{Z_n}q(Z_n)\log p(X_n|Z_n,T_n^X)+\sum_{Z_n}q(Z_n)\log p(T_n^X|Z_n)+\mathbf{E}_{q(Z_n)}\left[\log\frac{p(Z_n)}{q(Z_n)}\right] \tag{3.2}$$

$$=\sum_n\tilde{z}_n\log p(X_n|Z_n=1,T_n^X)+(1-\tilde{z}_n)\log p(X_n|Z_n=0,T_n^X)$$

$$+\sum_{Z_n}q(Z_n)\log p(T_n^X|Z_n)+\mathbf{E}_{q(Z_n)}\left[\log\frac{p(Z_n)}{q(Z_n)}\right] \tag{3.3}$$

$$=\sum_n\tilde{z}_n\log\left[\prod_{j,k}\sum_{\gamma_{jk}}p(Y_{njk}|\gamma_{jk},B_{jk};T_{njk})p(\gamma_{jk}|\pi_j)\right]+(1-\tilde{z}_n)\log\prod_{j,k}p(Y_{njk}|\frac{1}{2};T_{njk})$$

$$+\sum_{Z_n}q(Z_n)\log p(T_n^X|Z_n)+\mathbf{E}_{q(Z_n)}\left[\log\frac{p(Z_n)}{q(Z_n)}\right] \tag{3.4}$$

$$\geq\sum_{n,j,k}\tilde{z}_n\sum_{\gamma_{jk}}q(\gamma_{jk})\left(\log p(Y_{njk}|\gamma_{jk},B_{jk};T_{njk})+\log\frac{p(\gamma_{jk}|\pi_j)}{q(\gamma_{jk})}\right)+(1-\tilde{z}_n)\log p(Y_{njk}|\frac{1}{2};T_{njk})$$

$$+\sum_{Z_n}q(Z_n)\log p(T_n^X|Z_n)+\mathbf{E}_{q(Z_n)}\left[\log\frac{p(Z_n)}{q(Z_n)}\right] \tag{3.5}$$

$$=\sum_{n,j,k}\tilde{z}_n\left(\tilde{\gamma}_{jk}\log p(Y_{njk}|\frac{1}{2};T_{njk})+(1-\tilde{\gamma}_{jk})\log p(Y_{njk}|B_{jk};T_{njk})+\mathbf{E}_{q(\gamma)}\left[\log\frac{p(\gamma_{jk}|\pi_j)}{q(\gamma_{jk})}\right]\right)$$

$$+(1-\tilde{z}_n)\log p(Y_{njk}|\frac{1}{2};T_{njk})+\sum_{Z_n}q(Z_n)\log p(T_n^X|Z_n)+\mathbf{E}_{q(Z_n)}\left[\log\frac{p(Z_n)}{q(Z_n)}\right] \tag{3.6}$$

$$=\sum_{n,j,k}\tilde{z}_n\left(\tilde{\gamma}_{jk}\mathcal{L}_{njk}^b+(1-\tilde{\gamma}_{jk})\mathcal{L}_{njk}^{bb}(B_{jk})+\tilde{\gamma}_{jk}\log\frac{\pi_j}{\tilde{\gamma}_{jk}}+(1-\tilde{\gamma}_{jk})\log\frac{(1-\pi_j)}{(1-\tilde{\gamma}_{jk})}\right)$$

$$+(1-\tilde{z}_n)\mathcal{L}_{njk}^b+\sum_{Z_n}q(Z_n)\log p(T_n^X|Z_n)+\mathbf{E}_{q(Z_n)}\left[\log\frac{p(Z_n)}{q(Z_n)}\right] \tag{3.7}$$

$$\tag{3.8}$$

The last two terms include the negative binomial contribution and the KL-divergence between the prior and posteriors, exactly as in CENTIPEDE. Thus, the modified likelihood terms for the three models, and the relevant modified update equations, can be written as follows.

3.1. **Model 1.** The modified likelihood terms include

$$\mathcal{L} = \sum_{n,j,k} \tilde{z}_n \left(\tilde{\gamma}_{jk} \mathcal{L}^b_{njk} + (1-\tilde{\gamma}_{jk}) \mathcal{L}^{bb}_{njk}(B_{jk}) + \tilde{\gamma}_{jk} \log \frac{\pi_j}{\tilde{\gamma}_{jk}} + (1-\tilde{\gamma}_{jk}) \log \frac{(1-\pi_j)}{(1-\tilde{\gamma}_{jk})} \right) + (1-\tilde{z}_n) \mathcal{L}^b_{njk}, \tag{3.9}$$

where

$$\mathcal{L}^b_{njk} = \log p(Y_{njk}|\frac{1}{2}; T_{njk}) = \mathcal{C}_{njk} + T_{njk} \log \left(\frac{1}{2}\right) \tag{3.10}$$

$$\mathcal{L}^{bb}_{njk}(B_{jk}) = \log p(Y_{njk}|B_{jk}; T_{njk}) = \mathcal{C}_{njk} + Y_{njk} \log(B_{jk}) + (T_{njk} - Y_{njk}) \log(1-B_{jk}) \tag{3.11}$$

In the update equation for $\tilde{z}$, the likelihood functions of the multinomial parameters can be replaced by the following terms.

$$\log \frac{\tilde{z}_n}{1-\tilde{z}_n} = \sum_{j,k} (1-\tilde{\gamma}_{jk})(\mathcal{L}^{bb}_{njk}(B_{jk}) - \mathcal{L}^b_{njk}) + \tilde{\gamma}_{jk} \log \frac{\pi_j}{\tilde{\gamma}_{jk}} + (1-\tilde{\gamma}_{jk}) \log \frac{(1-\pi_j)}{(1-\tilde{\gamma}_{jk})} + \text{remaining terms} \tag{3.12}$$

A similar modification applied for the other two models. The update equations for $\tilde{\gamma}$, B and π can be modified as follows:

$$\log \frac{\tilde{\gamma}_{jk}}{1-\tilde{\gamma}_{jk}} = \log \frac{\pi_j}{1-\pi_j} + \frac{\sum_n \tilde{z}_n \left(\mathcal{L}^b_{njk} - \mathcal{L}^{bb}_{njk}(B_{jk}) \right)}{\sum_n \tilde{z}_n} \tag{3.13}$$

$$\pi_j = \frac{1}{K} \sum_k \tilde{\gamma}_{jk} \tag{3.14}$$

$$B_{jk} = \frac{\sum_n \tilde{z}_n Y_{njk}}{\sum_n \tilde{z}_n T_{njk}} \tag{3.15}$$

3.2. **Model 2.**

$$\mathcal{L} = \sum_{n,j,k} \tilde{z}_n \left(\tilde{\gamma}_{jk} \mathcal{L}^b_{njk} + (1-\tilde{\gamma}_{jk}) \mathcal{L}^{bb}_{njk} + \tilde{\gamma}_{jk} \log \frac{\pi_j}{\tilde{\gamma}_{jk}} + (1-\tilde{\gamma}_{jk}) \log \frac{(1-\pi_j)}{(1-\tilde{\gamma}_{jk})} \right) + (1-\tilde{z}_n) \mathcal{L}^b_{njk}, \tag{3.16}$$

where

$$\mathcal{L}^b_{njk} = \log p(Y_{njk}|\frac{1}{2}; T_{njk}) \tag{3.17}$$

$$\mathcal{L}^{bb}_{njk} = \log p(Y_{njk}|\mu_j; T_{njk}). \tag{3.18}$$

Update equations for $\tilde{\gamma}$ and π can be given as

$$\log \frac{\tilde{\gamma}_{jk}}{1-\tilde{\gamma}_{jk}} = \log \frac{\pi_j}{1-\pi_j} + \frac{\sum_n \tilde{z}_n \left(\mathcal{L}^b_{njk} - \mathcal{L}^{bb}_{njk} \right)}{\sum_n \tilde{z}_n} \tag{3.19}$$

$$\pi_j = \frac{1}{K} \sum_k \tilde{\gamma}_{jk} \tag{3.20}$$

3.3. **Model 3.**

$$\mathcal{L} = \sum_{n,j,k} \tilde{z}_n \left(\tilde{\gamma}_{jk} \mathcal{L}^b_{njk}(\tau_{jk}) + (1-\tilde{\gamma}_{jk}) \mathcal{L}^{bb}_{njk}(B_{jk}, \tau_{jk}) + \tilde{\gamma}_{jk} \log \frac{\pi_j}{\tilde{\gamma}_{jk}} + (1-\tilde{\gamma}_{jk}) \log \frac{(1-\pi_j)}{(1-\tilde{\gamma}_{jk})} \right) + (1-\tilde{z}_n) \mathcal{L}^o_{njk}, \tag{3.21}$$

where

$$\begin{aligned}
\mathcal{L}^b_{njk}(\tau_{jk}) &= \mathcal{C}_{njk} + \log\Gamma(Y_{njk} + 0.5\tau_{jk}) + \log\Gamma(T_{njk} - Y_{njk} + 0.5\tau_{jk}) \\
&\quad - \log\Gamma(T_{njk} + \tau_{jk}) + \log\Gamma(\tau_{jk}) - 2 * \log\Gamma(0.5\tau_{jk}) && (3.22) \\
\mathcal{L}^{bb}_{njk}(B_{jk}, \tau_{jk}) &= \mathcal{C}_{njk} + \log\Gamma(Y_{njk} + B_{jk}\tau_{jk}) + \log\Gamma(T_{njk} - Y_{njk} + (1-B_{jk})\tau_{jk}) \\
&\quad - \log\Gamma(T_{njk} + \tau_{jk}) + \log\Gamma(\tau_{jk}) - \log\Gamma(B_{jk}\tau_{jk}) - \log\Gamma((1-B_{jk})\tau_{jk}) && (3.23) \\
\mathcal{L}^o_{njk}(\tau^o_{jk}) &= \mathcal{C}_{njk} + \log\Gamma(Y_{njk} + 0.5\tau^o_{jk}) + \log\Gamma(T_{njk} - Y_{njk} + 0.5\tau^o_{jk}) \\
&\quad - \log\Gamma(T_{njk} + \tau^o_{jk}) + \log\Gamma(\tau^o_{jk}) - 2 * \log\Gamma(0.5\tau^o_{jk}) && (3.24)
\end{aligned}$$

Update equations for $\tilde{\gamma}$ and π can be given as

$$\log \frac{\tilde{\gamma}_{jk}}{1-\tilde{\gamma}_{jk}} = \log \frac{\pi_j}{1-\pi_j} + \frac{\sum_n \tilde{z}_n \left(\mathcal{L}^b_{njk}(\tau_{jk}) - \mathcal{L}^{bb}_{njk}(B_{jk}, \tau_{jk}) \right)}{\sum_n \tilde{z}_n} \tag{3.25}$$

$$\pi_j = \frac{1}{K} \sum_k \tilde{\gamma}_{jk} \tag{3.26}$$

The gradient of the likelihood with respect to B and τ can be derived as follows.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial B_{jk}} &= \tau_{jk}(1-\tilde{\gamma}_{jk}) \sum_n \tilde{z}_n \left[\Psi(Y_{njk} + B_{jk}\tau_{jk}) - \Psi(T_{njk} - Y_{njk} + (1-B_{jk})\tau_{jk}) \right. \\
&\quad \left. - \Psi(B_{jk}\tau_{jk}) + \Psi((1-B_{jk})\tau_{jk}) \right] && (3.27) \\
\frac{\partial \mathcal{L}}{\partial \tau_{jk}} &= \sum_n \tilde{z}_n \tilde{\gamma}_{jk} \left[0.5\Psi(Y_{njk} + 0.5\tau_{jk}) + 0.5\Psi(T_{njk} - Y_{njk} + 0.5\tau_{jk}) \right. \\
&\quad \left. - \Psi(T_{njk} + \tau_{jk}) + \Psi(\tau_{jk}) - \Psi(0.5\tau_{jk}) \right] \\
&\quad + \tilde{z}_n(1-\tilde{\gamma}_{jk}) \left[B_{jk}\Psi(Y_{njk} + B_{jk}\tau_{jk}) + (1-B_{jk})\Psi(T_{njk} - Y_{njk} + (1-B_{jk})\tau_{jk}) \right. \\
&\quad \left. - \Psi(T_{njk} + \tau_{jk}) + \Psi(\tau_{jk}) - B_{jk}\Psi(B_{jk}\tau_{jk}) - (1-B_{jk})\Psi((1-B_{jk})\tau_{jk}) \right] && (3.28)
\end{aligned}$$

4. Inference

Inference on binding sites can be performed by using

$$\frac{\mathsf{P}(\text{binding site}|\text{data})}{\mathsf{P}(\text{not binding site}|\text{data})}, \tag{4.1}$$

where

$$\begin{aligned}
\mathsf{P}(\text{not binding site}|\text{data}) &= \mathsf{P}(Z_n = 0 | X_n, T^X_n) + \mathsf{P}(Z_n = 1, \gamma_{jk} = 1 \ \forall j,k | X_n, T^X_n) && (4.2) \\
\mathsf{P}(\text{binding site}|\text{data}) &= 1 - \mathsf{P}(\text{not binding site}|\text{data}). && (4.3)
\end{aligned}$$

Here,

$$\begin{aligned}
\mathsf{P}(Z_n = 1, \gamma_{jk} = 1 \ \forall j,k | X_n, T^X_n) &= \mathsf{P}(\gamma_{jk} = 1 \ \forall j,k | Z_n = 1, X_n, T^X_n) \mathsf{P}(Z_n = 1 | X_n, T^X_n) && (4.4) \\
&= \prod_{jk} [\mathsf{P}(\gamma_{jk} = 1 | Z_n = 1, X_n, T^X_n)] \mathsf{P}(Z_n = 1 | X_n, T^X_n). && (4.5)
\end{aligned}$$