A Trans-dimensional Bayesian Model for Pattern Recognition in DNA Sequences

Sierra M. Li[1*] Jon Wakefield[2] Steve Self[3]

[1] Division of Oncology Biostatistics Sidney Kimmel Cancer Center
Johns Hopkins School of Medicine Baltimore, MD 21205-2013
email: sierrali@jhmi.edu Tel: (410) 664-4770 Fax: (410) 955-0859
[2] Department of Biostatistics University of Washington Seattle, WA 98195-7232
[3] Fred Hutchinson Cancer Research Center Seattle, WA 98109-1024
[*] To whom correspondence should be addressed.

Supplementary Material

1 UPDATE NON-DIMENSION-CHANGING PARAMETERS

The values of $\boldsymbol{q}_0$, $\boldsymbol{Q}$, $\boldsymbol{A}$, and $\boldsymbol{Z}$ do not change the dimensionality of the parameter space and with their conjugate priors, we can easily update them via the Gibbs Sampler.

Let $\boldsymbol{\Theta}_{-\boldsymbol{q}_0}$ and $\boldsymbol{\Theta}_{-\boldsymbol{q}_{mi}}$ denote the parameter set without $\boldsymbol{q}_0$ and $\boldsymbol{q}_{mi}$, respectively. Because $\pi(\boldsymbol{q}_0|\boldsymbol{\Theta}_{-\boldsymbol{q}_0}, \boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\Theta})\pi(\boldsymbol{\Theta}) \propto \boldsymbol{q}_0^{\boldsymbol{c}_0+\boldsymbol{\delta}_0}$ and $\pi(\boldsymbol{q}_{mi}|\boldsymbol{\Theta}_{-\boldsymbol{q}_{mi}}, \boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\Theta})\pi(\boldsymbol{\Theta}) \propto \boldsymbol{q}_{mi}^{\boldsymbol{c}_{mi}+\boldsymbol{\delta}}$, the conditional posteriors for $\boldsymbol{q}_0$ and $\boldsymbol{q}_{mi}$ are $\text{Dir}(\boldsymbol{c}_0 + \boldsymbol{\delta}_0)$ and $\text{Dir}(\boldsymbol{c}_{mi} + \boldsymbol{\delta})$, for $i = 1, 2, \cdots, W_m,\ m = 1, 2, \cdots, M$.

Updating $(a_{m,n}, z_{m,n})$, the location and orientation of site n in motif m, provides a means to replace a current site by a new site. The joint prior $\pi(a_{m,n}, z_{m,n}|\boldsymbol{\Theta}_{-\{a_{m,n},z_{m,n}\}})$ is $1/(2|\boldsymbol{L}(a_{m,n}|\boldsymbol{\Theta}_{-a_{m,n}})|)$, where $|\boldsymbol{L}(a_{m,n}|\boldsymbol{\Theta}_{-a_{m,n}})|$ is the number of possible positions for $a_{m,n}$. Let $\boldsymbol{c}_{mni}$ denote the 4×1 vector of nucleotide indicators for site n, at the i-th position in motif m, taking into account the orientation of site n. We use $\boldsymbol{d}_{mni}$ to denote the 4×1 vector of nucleotide indicators for corresponding "background

loss", where $\boldsymbol{d}_{mi}$ does not consider the orientation of the site. Both $\boldsymbol{c}_{mni}$ and $\boldsymbol{d}_{mni}$ are unit vectors, since they are the counts at a single nucleotide. It is easy to see that $\boldsymbol{c}_{mi} = \sum_{n=1}^{N_m} \boldsymbol{c}_{mni}$ and $\boldsymbol{d}_{mi} = \sum_{n=1}^{N_m} \boldsymbol{d}_{mni}$.

Letting $\boldsymbol{c}_Y$ denote the 4×1 vector of nucleotides counts for bases in $\boldsymbol{Y}$, including both the background and motif bases, we have $\boldsymbol{c}_Y = \boldsymbol{c}_0 + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{i=1}^{W_m} \boldsymbol{d}_{mni}$. The likelihood in can be written as $P(\boldsymbol{Y}|\boldsymbol{\Theta}) = \boldsymbol{q}_0^{\boldsymbol{c}_Y} \prod_{m=1}^{M} \prod_{n=1}^{N_m} \prod_{i=1}^{W_m} \boldsymbol{q}_{mi}^{\boldsymbol{c}_{mni}} / \boldsymbol{q}_0^{\boldsymbol{d}_{mni}}$, so the joint conditional posterior is easily derived as

$$\pi(a_{m,n}, z_{m,n}|\boldsymbol{\Theta}_{-\{a_{m,n},z_{m,n}\}}, \boldsymbol{Y}) \propto \prod_{i=1}^{W_m} \frac{\boldsymbol{q}_{mi}^{\boldsymbol{c}_{mni}}}{\boldsymbol{q}_0^{\boldsymbol{d}_{mni}}}, \tag{1.1}$$

where $\boldsymbol{c}_{mni}$ and $\boldsymbol{d}_{mni}$ depend on $(a_{m,n}, z_{m,n})$.

The computation time for updating all sites is proportional to $|L|^{\sum_{m=1}^{M} N_m}$. This step becomes expensive when the sequence data is large. Consequently, we update $(a_{m,n}, z_{m,n})$ every few hundred of iterations.

2 UPDATE DIMENSION-CHANGING PARAMETERS

The dimension of the parameter space, $\boldsymbol{\Theta} = (\boldsymbol{Q}, \boldsymbol{q}_0, \boldsymbol{A}, \boldsymbol{Z}, \boldsymbol{N}, \boldsymbol{W}, M)$, is $3\sum_{m=1}^{M} W_m 3 + 2\sum_{m=1}^{M} N_m + 2M + 4$, and is therefore a function of $\boldsymbol{N}, \boldsymbol{W}$ and M. Updating $\boldsymbol{N}, \boldsymbol{W}$ and M is associated with dimensional changes of the parameter space; therefore, simple Gibbs sampler and usual Metropolis-Hastings steps are no longer possible.

We use RJMCMC to update the parameters. Table 2.1 gives a schematic presentation of the probabilities of choosing moves, where M is the number of motifs and M_2 is the number of motifs with 2 sites and s_1 and s_2 are constants of choice. The motif birth step will be taken if there are currently no motifs. In practice, we have found that using $s_1 = 0.1$ and $s_2 = 0.5$ works well.

2.1 Move 1: change the width for a randomly selected motif.

When Move 1 is proposed, for a randomly selected motif m, its width W_m is randomly increased or decreased by 1, either at the front or the rear end of the motif.

Table 2.1: Probabilities for moves 1–4 in the RJMCMC algorithm; s_1 and s_2 are specified constants that lie in (0, 1).

M	M_2	η_W	η_N	η_b	η_d
0	0	0	0	1	0
$[1, M_{\max}-1]$	0	$(1-s_1)s_2$	$(1-s_1)(1-s_2)$	s_1	0
$[1, M_{\max}-1]$	$[1, M]$	$(1-s_1)s_2$	$(1-s_1)(1-s_2)$	$s_1/2$	$s_1/2$
$M_{\max}$	0	s_2	$1-s_2$	0	0
$M_{\max}$	$[1, M]$	$(1-s_1)s_2$	$(1-s_1)(1-s_2)$	0	s_1

Recall that $\boldsymbol{c}_{mi}$ denotes the nucleotide count at the i-th position of motif m, over all N_m sites accounting for site orientations. Let $\boldsymbol{c}_{m0}$ represent the nucleotide count for adjacent bases directly preceding the sites, and $\boldsymbol{c}_{m(W_m+1)}$ the count for bases directly following the sites.

Due to the expansion of the motif width, the loss of nucleotide count from the background is $\boldsymbol{d}_{m0}$ or $\boldsymbol{d}_{m(W_m+1)}$. A column is added to the motif composition matrix $\boldsymbol{Q}_m$ with $\boldsymbol{q}_{mx}$ proposed from $\text{Dir}(\boldsymbol{c}_{mx})$, $x = 0$ or $W_m + 1$. The acceptance probability is $\min(1, \alpha_{W+})$ and

$$\begin{aligned}\alpha_{W+} =& \frac{\text{candidate posterior}}{\text{current posterior}} \times \frac{\text{p(candidate point} \to \text{current point)}}{\text{p(current point} \to \text{candidate point)}} \times |J| \\ =& \frac{L(\boldsymbol{Y}|\boldsymbol{\Theta}^*)}{L(\boldsymbol{Y}|\boldsymbol{\Theta})} \frac{\pi(\boldsymbol{\Theta}^*)}{\pi(\boldsymbol{\Theta})} \frac{p(\boldsymbol{\Theta}^* \to \boldsymbol{\Theta})}{p(\boldsymbol{\Theta} \to \boldsymbol{\Theta}^*)} |J|.\end{aligned} \tag{2.1}$$

The Jacobian matrix $|J| = |\partial(\boldsymbol{\Theta}^*)/\partial(\boldsymbol{\Theta}, \boldsymbol{q}_{mi})| = 1$ and the proposal densities are

$$p(\boldsymbol{\Theta}^* \to \boldsymbol{\Theta}) = \frac{1}{M}\,\eta_W \qquad \text{and} \qquad p(\boldsymbol{\Theta} \to \boldsymbol{\Theta}^*) = \frac{1}{M}\,\eta_W\,\frac{\Gamma(|\boldsymbol{c}_{mi} + \boldsymbol{\delta}|)}{\Gamma(\boldsymbol{c}_{mi} + \boldsymbol{\delta})}\,(\boldsymbol{q}_{mi})^{\boldsymbol{c}_{mi}+\boldsymbol{\delta}}\,.$$

for $i = 0$ or $W_m + 1$.

It is easy to show $L(\boldsymbol{Y}|\boldsymbol{\Theta}^*)/L(\boldsymbol{Y}|\boldsymbol{\Theta}) = (\boldsymbol{q}_{mi})^{\boldsymbol{c}_{mi}}/\boldsymbol{q}_i^{\boldsymbol{d}_{mi}}$. Though it can be calculated in explicit but complicated form, we write $\pi(\boldsymbol{A}^*|\boldsymbol{N}, \boldsymbol{W}^*, M)/\pi(\boldsymbol{A}|\boldsymbol{N}, \boldsymbol{W}, M) = R_{W+}$, where R_{W+} is the ratio of the two priors, for the site locations $\boldsymbol{A}$, with and without the width expansion. When the dataset has sufficiently large $|\boldsymbol{L}|$, the prior ratio for site locations R_{W+} can be very closely approximated by 1. Therefore, the

prior ratio is $\pi(\boldsymbol{\Theta}^*)/\pi(\boldsymbol{\Theta}) = \pi(\boldsymbol{q}_{mi})R_{W+}$. Hence, the acceptance is $\min(1,\ \alpha_{W+})$, where

$$\alpha_{W+} = \boldsymbol{q}_0^{-\boldsymbol{d}_{mx}}\ R_{W+}\ \frac{\Gamma(|\boldsymbol{\delta}|)}{\Gamma(\boldsymbol{\delta})}\frac{\Gamma(\boldsymbol{c}_{mx}+\boldsymbol{\delta})}{\Gamma(|\boldsymbol{c}_{mx}+\boldsymbol{\delta}|)}, \quad x = 0 \text{ or } W_m+1. \tag{2.2}$$

Similarly, let R_{W-} be the prior ratio of $\boldsymbol{A}$ with and without the width deduction. The acceptance rate of the reversible jump is $\min(1,\ \alpha_{W-})$, where

$$\alpha_{W-} = \boldsymbol{q}_0^{\boldsymbol{d}_{mx}}\ R_{W-}\ \frac{\Gamma(\boldsymbol{\delta})}{\Gamma(|\boldsymbol{\delta}|)}\frac{\Gamma(|\boldsymbol{c}_{mx}+\boldsymbol{\delta}|)}{\Gamma(\boldsymbol{c}_{mx}+\boldsymbol{\delta})}, \quad x = 1 \text{ or } W_m. \tag{2.3}$$

2.2 Move 2: change the number of sites for a randomly selected motif.

When Move 2 is proposed, for a randomly selected motif m, the number of sites N_m is proposed to increase by 1, via the birth of a new site, or decrease by 1, via the death of a current site. When a new site of motif m, labeled N_{m+1}, is born, its location and orientation, a_{m,N_m+1} and z_{m,N_m+1}, are generated from their conditional posterior. It is easy to show $|J| = 1$, $L(\boldsymbol{Y}|\boldsymbol{\Theta}^*)/L(\boldsymbol{Y}|\boldsymbol{\Theta}) = \prod_{i=1}^{W_m} \boldsymbol{q}_{mi}^{\boldsymbol{c}_{m(N_m+1)i}}/\boldsymbol{q}_0^{\boldsymbol{d}_{m(N_m+1)i}}$ and $\pi(\boldsymbol{\Theta}^*)/\pi(\boldsymbol{\Theta}) = 2R_{N+}$, where R_{N+} is the conditional prior ratio of the site locations with or without the additional site. Let R_{N-} denote the conditional prior ratio of the site locations before and after the death of site n. The acceptance rates for the birth and death steps are $\min(1,\ \alpha_{N+})$ and $\min(1,\ \alpha_{N-})$, where

$$\alpha_{N+} = 2R_{N+}\sum_{a_{m,N_m+1}}\sum_{z_{m,N_m+1}}\left(\prod_{i=1}^{W_m}\frac{\boldsymbol{q}_0^{\boldsymbol{d}_{m(N_m+1)i}}}{\boldsymbol{q}_{mi}^{\boldsymbol{c}_{m(N_m+1)i}}}\right) \text{ and } \alpha_{N-} = \frac{1}{2R_{N-}}\sum_{a_{m,n}}\sum_{z_{m,n}}\left(\prod_{i=1}^{W_m}\frac{\boldsymbol{q}_0^{\boldsymbol{d}_{mni}}}{\boldsymbol{q}_{mi}^{\boldsymbol{c}_{mni}}}\right). \tag{2.4}$$

When the sequence data is sufficiently large, R_{N+} and R_{N-} are very close to $|\boldsymbol{L}(a_{m,N_m+1}|\boldsymbol{\Theta})|$ and $|\boldsymbol{L}(a_{m,n}|\boldsymbol{\Theta}_{-a_{m,n}})|$, respectively, where $|\boldsymbol{L}(a_{m,N_m+1}|\boldsymbol{\Theta})|$ is the number of available locations for the new site and $|\boldsymbol{L}(a_{m,n}|\boldsymbol{\Theta}_{-a_{m,n}})|$ is the number of available locations given $\boldsymbol{\Theta}_{-a_{m,n}}$. With these approximations, acceptance rates have the simple interpretation of average likelihood ratios over the available positions.

2.3 Move 3: birth of a new motif.

When we propose the birth of the motif labeled $M+1$ (with two sites), the width W_{m+1} is generated from a discrete uniform on $[W_{\min}, W_{\max}]$. The locations the two new-born sites are randomly selected from the available positions, from their conditional priors, the orientations are generated with a probability 1/2, and the motif composition $\boldsymbol{Q}_{M+1}$ is generated from the conditional posterior. Therefore, the proposal density $p(\boldsymbol{\Theta} \to \boldsymbol{\Theta}^*)$ is

$$\frac{\eta_b}{4|\boldsymbol{L}(\boldsymbol{a}_{M+1,1}, \boldsymbol{a}_{M+1,2}|\boldsymbol{\Theta})|\ (W_{\max} - W_{\min} + 1)} \prod_{i=1}^{W_{M+1}} \frac{\Gamma(|\boldsymbol{c}_{(M+1)i} + \boldsymbol{\delta}|)}{\Gamma(|\boldsymbol{c}_{(M+1)i} + \boldsymbol{\delta})} \boldsymbol{q}_{(M+1)i}^{\boldsymbol{c}_{(M+1)i}+\boldsymbol{\delta}}.$$

The proposal density $p(\boldsymbol{\Theta}^* \to \boldsymbol{\Theta}) = \eta_d$, and the likelihood ratio and the prior ratio are $L(\boldsymbol{Y}|\boldsymbol{\Theta}^*)/L(\boldsymbol{Y}|\boldsymbol{\Theta}) = \prod_{i=1}^{W_{M+1}} \boldsymbol{q}_{(M+1)i}^{\boldsymbol{c}_{(M+1)i}} / \boldsymbol{q}_0^{\boldsymbol{d}_{(M+1)i}}$, and

$$\frac{\pi(\boldsymbol{\Theta}^*)}{\pi(\boldsymbol{\Theta})} = \frac{\lambda}{4(M+1)(W_{\max} - W_{\min} + 1)} \prod_{i=1}^{W_{M+1}} R_{M+} \frac{\Gamma(|\boldsymbol{\delta}|)}{\Gamma(\boldsymbol{\delta})} \boldsymbol{q}_{(M+1)i}^{\boldsymbol{\delta}},$$

where $|\boldsymbol{L}(\boldsymbol{a}_{M+1,1}, \boldsymbol{a}_{M+1,2}|\boldsymbol{\Theta}|)$ stands for the number of available locations for $a_{M+1,1}$ and $a_{M+1,2}$ given $\boldsymbol{\Theta}$, and R_{M+} is the conditional prior ratio of site locations, with or without the 2 sites. It can be shown that $R_{M+}|\boldsymbol{L}(a_{M+1,1}, a_{M+1,2}|\boldsymbol{\Theta})|$ is very close to 1. Again, the Jacobian matrix $|J|$ is 1. Thus, the acceptance probability is min(1, α_b), where

$$\begin{aligned} \alpha_b = & \frac{\eta_d}{\eta_b} \frac{\lambda}{M+1} R_{M+} |\boldsymbol{L}(a_{M+1,1}, a_{M+1,2}|\boldsymbol{\Theta})| \\ & \times \boldsymbol{q}_0^{-\sum_{i=1}^{W_{M+1}} \boldsymbol{d}_{mi}} \left(\frac{\Gamma(|\boldsymbol{\delta}|)}{\Gamma(\boldsymbol{\delta})}\right)^{W_{M+1}} \prod_{i=1}^{W_{M+1}} \frac{\Gamma(\boldsymbol{c}_{(M+1)i} + \boldsymbol{\delta})}{\Gamma(|\boldsymbol{c}_{(M+1)i} + \boldsymbol{\delta}|)}, \end{aligned} \tag{2.5}$$

2.4 Move 4: death of a motif.

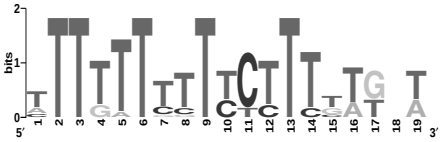
When M is proposed to decrease by 1, we randomly select a motif with 2 sites, say it is labeled m, and propose that it dies. Let R_{M-} be the conditional prior ratio of

site locations, before and after the death of motif m. The acceptance rate is $\min(1, \alpha_d)$, where

$$\begin{aligned}\alpha_d =& \frac{\eta_b}{\eta_d}\,\frac{M}{\lambda}\,\frac{1}{|\boldsymbol{L}(a_{m,1}, a_{m,2}|\boldsymbol{\Theta}_{-\{a_{m,1},a_{m,2}\}})|\, R_{M-}} \\ & \times \boldsymbol{q}_0^{\sum_{i=1}^{W_m} \boldsymbol{d}_{mi}}\,\left(\frac{\Gamma(\boldsymbol{\delta})}{\Gamma(|\boldsymbol{\delta}|)}\right)^{W_m}\,\prod_{i=1}^{W_m}\frac{\Gamma(|\boldsymbol{c}_{mi}+\boldsymbol{\delta}|)}{\Gamma(\boldsymbol{c}_{mi}+\boldsymbol{\delta})},\end{aligned} \tag{2.6}$$

and $|\boldsymbol{L}(a_{m,1}, a_{m,2}|\boldsymbol{\Theta}_{-\{a_{m,1},a_{m,2}\}})|\, R_{M-}$ is very close to 1. Details of this calculation is similar to those in Section 2.3.

3 PERSISTENT MOTIFS IN OCT4, SOX2 AND NANOG CHIP



(a) poly-T: $(\widehat{W}, \widehat{N}) = (19, 13)$

(b) poly-C: $(\widehat{W}, \widehat{N}) = (9, 27)$

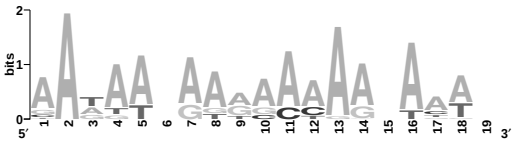
(c) poly-G: $(\widehat{W}, \widehat{N}) = (19, 11)$

(d) Sox2-Oct4: $(\widehat{W}, \widehat{N}) = (16, 10)$

(e) GC repeats: $(\widehat{W}, \widehat{N}) = (17, 10)$

Figure 3.1: Sequence Logos for 5 motifs discovered in Oct4 ChiP study

(a) G/C repeats: $(\widehat{W}, \widehat{N}) = (20, 50)$



(b) poly-A: $(\widehat{W}, \widehat{N}) = (19, 48)$

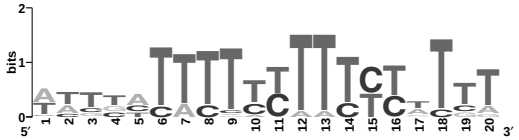
(c) poly-G: $(\widehat{W}, \widehat{N}) = (19, 16)$

(d) $(\widehat{W}, \widehat{N}) = (15, 6)$

(e) Sox2-Oct4: $(\widehat{W}, \widehat{N}) = (16, 22)$

(f) $(\widehat{W}, \widehat{N}) = (20, 12)$

Figure 3.2: Sequence Logos for 6 motifs discovered in Sox2 ChiP study



(a) Poly-T repeats: $(\widehat{W}, \widehat{N}) = (20, 36)$

(b) $(\widehat{W}, \widehat{N}) = (20, 6)$

(c) $(\widehat{W}, \widehat{N}) = (20, 4)$

(d) $(\widehat{W}, \widehat{N}) = (19, 32)$

(e) $(\widehat{W}, \widehat{N}) = (20, 34)$

(f) $(\widehat{W}, \widehat{N}) = (20, 9)$

(g) A/T:$(\widehat{W}, \widehat{N}) = (19, 11)$

(h) Oct4$(\widehat{W}, \widehat{N}) = (15, 13)$

Figure 3.3: Sequence Logos for 8 motifs discovered in NANOG ChiP study