Biometrics 000, 000–000 DOI: 000
000 0000

Supplementary materials for "A Bayesian Hidden Markov Model for Motif Discovery through Joint Modeling of Genomic sequence and ChIP-chip Data"

Jonathan A.L. Gelfond*

Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, USA

**email:* gelfondjal@uthscsa.edu

and

Mayetri Gupta*

Department of Biostatistics, University of North Carolina at Chapel Hill, NC, USA

**email:* gupta@bios.unc.edu

and

Joseph G. Ibrahim*

Department of Biostatistics, University of North Carolina at Chapel Hill, NC, USA

**email:* ibrahim@bios.unc.edu

Supplementary materials

0.1 *Computing the HMM likelihood*

Hidden Markov model likelihoods generally cannot be written in a closed form so that a recursive procedure based upon the law of total probability is used in the likelihood computation (Juang and Rabiner, 1991). We use a *forward summation* recursive formula for computing the HMM likelihood, described below. We define $g_p(s)$ as the probability of the sequence of probes up to probe p with the p^{th} state as s:

$$g_p(s) = P((\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_p, \boldsymbol{y}_p), s_p = s) = P(\boldsymbol{x}_p, \boldsymbol{y}_p|s) \sum_{s_{p-1} \in \{0,1\}} g_{p-1}(s_{p-1}) \tau_{s_{p-1},s} \qquad (1)$$

where $P(\boldsymbol{x}_p, \boldsymbol{y}_p|s) = p_s(\boldsymbol{x}_p) f_s(\boldsymbol{y}_p)$ for $s \in \{0, 1\}$, corresponding to the *not enriched* and *enriched* states. The τ_{ij} parameters represent the transition probability between states i and j $(i, j \in \{0, 1\})$. The total likelihood is calculated through recursively computing equation (1) for $p = 1, \ldots, P$ and given by $g_P(0) + g_P(1) = P((\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_P, \boldsymbol{y}_P))$.

The sequence likelihoods $p_0(\boldsymbol{x}_p)$ and $p_1(\boldsymbol{x}_p)$ do not have closed forms and can also be calculated using a recursive formula for each sequence taken to correspond to a probe. We achieve synchrony between the probe intensity and probe sequence by expanding the probe sequence to include the sequences of the binding DNA fragments. The length of the probe sequence is taken to be 1500 bp about the center of the probe regardless of the probe length on the microarray. This could potentially lead to overlap between shorter probes, thus invalidating the independent sequence assumption between probes- however with a small number of short probes we ignore this overlap in the enrichment estimation step in the interests of model simplicity. The few longer probes are truncated to 1500 bp to avoid biases due to probe length. In motif estimation step, the overlapping probes are truncated to avoid double counting of motif sites. The likelihood of subsequence $\boldsymbol{x}_p[i : j]$ given that it

was emitted from motif Θ_v is denoted as $p(\boldsymbol{x}_p[i:j]|\Theta_v)$ where $j-i+1=w_v$ and is given by

$$p(\boldsymbol{x}_p[i-w_v+1:i]|\Theta_v) = I[i-w_v+1>0]I[i\leqslant K]\prod_{j=i-w_v+1}^{i}\prod_{l\in\{A,C,G,T\}}\Theta_{v,lj}^{I[\boldsymbol{x}_p[j]=l]}.$$

The term $I[i-w_v+1>0]I[i\leqslant K]$ makes the probability 0 when the motif would not fit within the sequence of length K. Let $\phi_p(k)$ denote the probability of $\boldsymbol{x}_p$ up to position $k\in[0\ldots K]$. Then $p_1(\boldsymbol{x}_p)$ are calculated by recursive summation allowing for for all possible motif site locations, as below:

$$\begin{aligned}\phi_p(0) &= 1\\ \phi_p(k) &= p_1(\boldsymbol{x}_p[1:k]) = \sum_{v=1}^{V}\pi_v p(\boldsymbol{x}_p[k-w_v+1:k]|\Theta_v)\phi_p(k-w_v).\end{aligned} \tag{2}$$

$p_0(\boldsymbol{x}_p)$ is found similarly by allowing $v=1,..,V-1$.

0.2 *MCMC fitting procedure*

After initialization of parameters, we fit the model with a Data Augmentation (DA) method, involving the following iterative steps:

(1) Sample intensity parameters (μ_1, ν_1^2, ν_0^2, σ_a^2 | $\mathbf{s},\mathbf{x},\mathbf{y}$) using a Metropolis-Hastings (MH) random walk procedure.

(2) The enrichment states s_p are then sampled jointly from the posterior distribution of ($\mathbf{s}$ | $\mathbf{x},\mathbf{y}$, *parameters*) through the recursive *backward sampling* procedure, below:

$$\begin{aligned}s_P|(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots(\boldsymbol{x}_P,\boldsymbol{y}_P) &\sim \text{Bern}\left(\frac{g_P(1)}{g_P(1)+g_P(0)}\right),\\ s_p|(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_p,\boldsymbol{y}_p),s_{p+1} &\sim \text{Bern}\left(\frac{g_p(1)\tau_{1,s_{p+1}}}{g_p(1)\tau_{1,s_{p+1}}+g_p(0)\tau_{0,s_{p+1}}}\right) \quad \text{for } p=\{P-1,\ldots,1\}.\end{aligned}$$

(3) The transition parameters τ_{ij} | $\mathbf{x},\mathbf{y},\mathbf{s}$, where t_{ij} are the $i\to j$ transitions given by s_p ($s_p=0,1$), are drawn from the complete conditional (Beta) distributions.

(4) The enriched segments are formed by the overlapping regions of the genome that correspond to probes with $s_p=1$, and these segments are denoted by $\boldsymbol{x}_e$ where the index e stands for "enriched". A recursive DA step is applied to the $\boldsymbol{x}_e$ in order to

iteratively sample the motif site locations: $(\mathbf{A} \mid \Theta_V, \mathbf{x}, \mathbf{y}, \textit{parameters})$, the motif matrix: $(\Theta_V \mid \mathbf{A}, \mathbf{x}, \mathbf{y}, \textit{parameters})$ and motif prevalence: $(\boldsymbol{\pi} \mid \mathbf{A}, \mathbf{x}, \mathbf{y}, \textit{parameters})$.

0.3 *Data Preprocessing*

The data consist of four arrays and $11,575$ non-telomeric probes of various lengths spanning the yeast genome of 17 chromosomes with a total of 12 million base pairs. The corresponding yeast genome was filtered for large base pair repeats using the RepeatMasker software of Smit et al. (2004). We used median centering and variance standardization to normalize the data. Shifting the median of each array to 0 is important because the proposed model assumes that the majority of the observations will arise from a distribution that is symmetric about 0. Variance standardization allows us to assume that the within array variance is equal.

Appendix: Computational Details

The sampling scheme for the joint intensity sequence model is given below.

(1) Initialize all parameters μ_1, ν_1^2, ν_0^2, σ_a^2, Θ_v, $\boldsymbol{\pi}$, τ_{ij}, and $s_1, \ldots s_P$.
(2) Sample μ_1, ν_1^2, ν_0^2, and σ_a^2 with MH random walk.
(3) Compute $P(\boldsymbol{y}_p, \boldsymbol{x}_p|s_p) = f_{s_p}(\boldsymbol{y}_p)p_{s_p}(\boldsymbol{x}_p)$ for $s_p \in \{0, 1\}$.
(4) Compute $g_p(0)$ and $g_p(1)$ for $p \in \{1, \ldots, n\}$ with Forward Algorithm.
(5) Sample Backwards s_P, $s_{P-1} \ldots s_1$.
(6) Count the number transitions t_{ij} where $i \to j$ in $s_{1\ldots P}$.
(7) Sample $\tau_{ij} \sim \text{Beta}(t_{ij} + \delta_{ij}, \sum_{k \neq j} t_{ik} + \delta_{ij})$.
(8) We define the $K_e \times V$ matrices A_e corresponding to the segments $\boldsymbol{x}_e$ of length K_e. The elements $A_{e,jv}$ indicate the sampling of the motif or PSWM v at position j such that $A_{e,jv} = 1$ if the v^{th} PSWM was sampled with $A_{e,jv} = 0$ otherwise. We may sample $A_{e,jv}$ using the backward algorithm described below:
    (a) Initialize $A_{e,jv} = 0$, $n_v = 0$.
    (b) Let $j = K_e$ the last position in sequence $\boldsymbol{x}_e$.
    (c) Sample $A_{e,j\cdot} \sim \text{Multinomial}(\frac{\phi_e(j-w_1)\pi_1 p(\boldsymbol{x}_e[j-w_1+1:j]|\Theta_v)}{\phi_e(j)}, .., \frac{\phi_e(j-w_V)\pi_V p(\boldsymbol{x}_e[j-w_V+1:j]|\Theta_v)}{\phi_e(j)})$ so that $A_{e,jv} = 1$ iff the v^{th} PSWM was sampled $A_{e,jv} = 0$ otherwise.
    (d) Decrement j by $(w_v - 1)$ iff $A_{e,jv} = 1$.
    (e) Increment n_v by 1 iff $A_{e,jv} = 1$.
    (f) Return to 3 until $j = 0$.
(9) The motif matrix Θ_V depends on the letter counts from the sampled TFBS where $A_{e,jV} = 1$, and we will call this $4 \times w_V$ count matrix $\boldsymbol{C}$ where the element c_{ij} is the number of the symbol i at motif position j. Θ_V has conditional distribution $PD(\boldsymbol{B} + \boldsymbol{C})$ where PD is the product Dirichlet distribution. Sample $\Theta_V \sim PD(\boldsymbol{B} + \boldsymbol{C})$.
(10) Next, the $\boldsymbol{\pi}$ parameter depends on the number of sampled realizations of each PSWM $n_1, .., n_V$ given by A_e so that $\pi_V \sim \text{Beta}(\delta_0(1-\gamma) + \sum_{v=1}^{V-1} n_v, \delta_0\gamma + n_V)$, and the complete conditional for $[\pi_1, \ldots, \pi_{V-1}]$ becomes $\text{Dirichlet}(n_1 + \delta_1, .., n_{V-1} + \delta_{V-1})$. Sample $\pi_V \sim \text{Beta}(\delta_0(1-\gamma) + \sum_{v=1}^{V-1} n_v, \delta_0\gamma + n_V)$.
(11) Sample $[\pi_1, \ldots, \pi_{V-1}] \sim \text{Dirichlet}(n_1 + \delta_1, .., n_{V-1} + \delta_{V-1})$.
(12) Return to 2.

The intensity only sampling scheme would skip steps 8-11, and step 3 would only compute $P(\boldsymbol{y}_p|s) = f_{s_p}(\boldsymbol{y}_p)$. The computations of step 3 may be prohibitive because of the terms $p_{s_p}(\boldsymbol{x}_p)$ if the number of background motifs $V - 1$ is large. The ratio $p_0(\boldsymbol{x}_p)/p_1(\boldsymbol{x}_p)$ is what is necessary for the computation of $g_p(s_p)$, and this ratio may be approximated by reducing the number of background motifs in this step. The method is implemented with the C programming language, and the executable program will be available upon request made to gelfondjal@uthscsa.edu.

0.4 *Model Initialization*

The initialization of the sequence model requires a reasonable estimate of the TFBS motif to facilitate convergence. The sequences selected by the above procedure are likely to have the highest concentration of the motif binding sites, but it is evident that there are many non-random patterns in the DNA that correspond to different modes in the likelihood and can lead to the failure of the stochastic dictionary model to find the motif which gives the highest likelihood for these sequences. To get the initial motif estimate, an *accumulating stochastic dictionary model* was fit to the sequences in which successive motifs are estimated and added to the dictionary. First, the dictionary was initialized with PSWMs of length one representing A's, C's, G's, and T's as well as repeat words of A's and T's of both of length 4 and length 8, which appeared sufficient to capture the dependence in the background, i.e. did not lead to further "repeat" motifs being predicted. These 8 motifs were considered part of the fixed background model with motif matrices $\Theta_1, \ldots \Theta_8$. The search for the "interesting" motif (Θ_V) was restricted to the assumed motif width of 13 (Lieb et al., 2001), and a motif of length 13 with uniform probability across all letters at all positions was added to the dictionary and updated using the data augmentation method described in Section 2 ($V = 9$). This motif is considered the foreground motif Θ^* and is the only motif updated in each cycle of the DA sampler. After approximate convergence, the updated motif is added to the fixed background dictionary, and another motif of length 13 with uniform probability across all letters at all positions is added to the dictionary so that $V = 10$, and this new word becomes the new foreground motif. The procedure of iteratively adding words to the background allows the model to consider different modes in the space of potential motifs.

Two likelihoods of the sequences are plotted across the iterations in order to find a reasonable motif for initialization. The first is the likelihood of the sequences given the full dictionary up to that point which may be denoted as $\prod_{X_i \in \text{Top Sequence}} p(X_i|\Theta_1, \ldots, \Theta_{8+m}, \Theta^*)$ where $m \geqslant 0$ is the number of accumulated words and Θ^* is the updated motif. The likelihood increases as motifs are added to the dictionary, and after a few iterations a plateau is reached signifying entrapment in a likelihood mode. The second likelihood computed is based on the original eight-PSWM background with only the current foreground motif and may be denoted as $\prod_{X_i \in \text{Top Sequence}} p(X_i|\Theta_1, \ldots \Theta_8, \Theta^*)$. This likelihood is an indication of the improvement in model fit given the addition of only the current foreground motif (Figure 1 in the Supplementary material). The motif that gives the largest increase in sequence likelihood is taken as a reasonable choice for the initial estimate of Θ_V in the joint sequence and intensity model, while the PSWMs $\Theta_1, \ldots, \Theta_8$ are used in the background model.

[Figure 1 about here.]

[Table 1 about here.]

References

Juang, B.-H. and Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics* **33,** 251–272.

Lieb, J. D., Liu, X. L., Botstein, D., and Brown, P. O. (2001). Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genetics* **28,** 327–334.

Smit, A. F. A., Hubley, R., and Green, P. (2004). Repeatmasker open-3.0. *http://www.repeatmasker.org* .

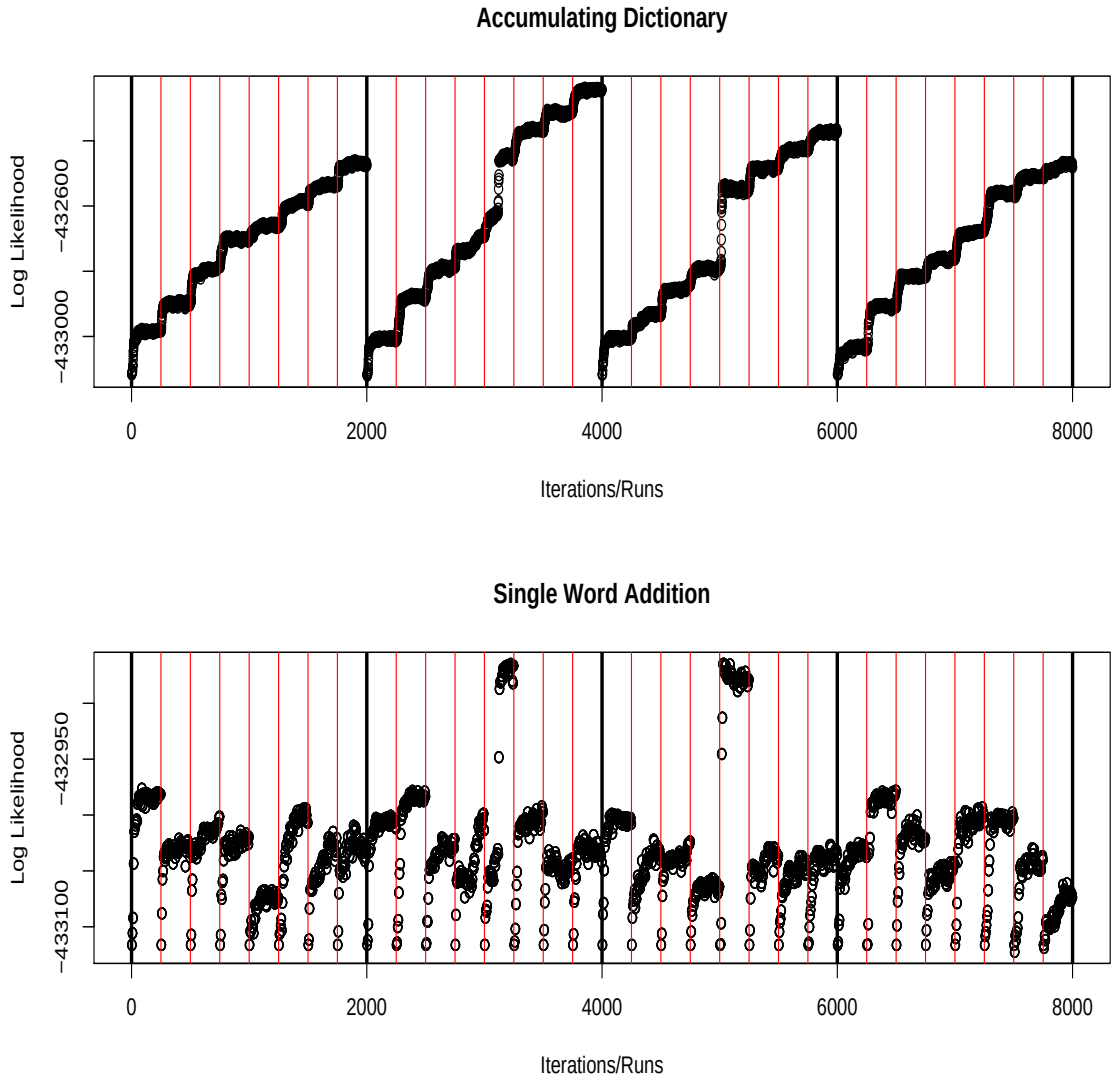

Figure 1. Likelihood trace plots for Accumulating Dictionary (Upper), and Single Word Addition (Lower). Independent runs are distinguished by bold vertical bars, and subsequent motifs are separated by light vertical bars. Plots are shown for 4 runs of length 2000 in which a total of 8 words are added to the dictionary every 250 iterations. One can see that runs 2 and 3 have the highest likelihood and that the fifth motifs added in both of the runs give a hugely significant improvement in model fit. The fifth motifs added in both of these runs are similar to the Rap1 motif.

Table 1
Comparison of binding sites $\gamma = 0.00006$

	IS	IO	Chipotle
IS	284		
IO	256	262	
Chipotle	256	247	271

Table 2
Comparison of binding sites $\gamma = 0.00007$

	IS	IO	Chipotle
IS	295		
IO	263	269	
Chipotle	259	251	274

Table 3
Comparison of binding sites $\gamma = 0.00008$

	IS	IO	Chipotle
IS	299		
IO	270	276	
Chipotle	266	255	277

Table 4
Comparison of binding sites $\gamma = 0.00009$

	IS	IO	Chipotle
IS	305		
IO	275	283	
Chipotle	271	263	289

Table 5
Comparison of binding sites $\gamma = 0.00010$

	IS	IO	Chipotle
IS	309		
IO	279	284	
Chipotle	277	267	293