Appendix 2: Mixture Evolution Model and the Conditional Probabilities to Update Motif Instances

Before we go into details, we restate all the notations in Table 2 and give an example of some terms on the phylogenetic tree for three species in Fig. 5 below.

m	Number of current species; there are $m-1$ ancestral species.
n	Number of genes in one species.
I	Species set. $I = \{0, 1, \cdots, 2m-2\}$.
Θ	Motif model for species 0, the root species.
Θ_0	Background model for species 0.
M_{0i}	Background substitution matrix at the *i*th branch of the tree, $i = 0, 1, ..., 2m-3$.
M_{1i}	Motif substitution matrix at the *i*th branch of the tree, $i = 0, 1, ..., 2m-3$.
p_i	Probability of a gene containing motif instances in the *i*th species, $i \in I$.
w	Motif width.
$A^{(i)}$	Motif instance set in the *i*th species, $i \in I$.
$A_j^{(i)}$	The motif instance in the *j*th gene of *i*th species, $i \in I$.
$A_{jk}^{(i)}$	The *k*th nucleotide in the motif instance in the *j*th gene of *i*th species, $i \in I$.
S	The set of regulatory sequences from all the current species.
$S_j^{(i)}$	The regulatory sequence of the *j*th gene in the *i*th species, $i \in I$.

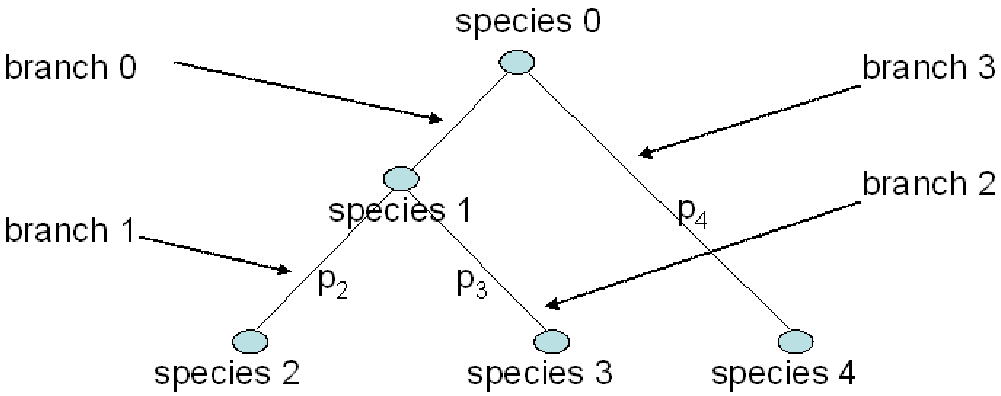

Fig. 5. Current species 1, 2, and 3 are at the bottom of the tree, from the left to the right. Note that the current species 1, 2, and 3 are the same as species 2, 3, and 4, respectively. p_2 is the probability that the genes in current species 1 will inherit the motif instances. p_3 and p_4 have similar meanings.

We assume that the sequences of coregulated genes in the ancestral species at the root of the phylogenetic tree were generated from a mixture model, in which background sequences were generated from the multinomial distribution with parameter Θ_0 while motif instances were generated from the product multinomial distribution with parameter $\Theta = (\Theta_1, \cdots, \Theta_w)$, where each Θ_i is a multinomial distribution and w is the motif width. The background sequences and motif instances evolved according to two different continuous Markov chain models, i.e., at the *i*th branch of the tree, the background sequences in the parent species evolved according to a background substitution matrix M_{0i} while the nucleotides in the motif instances evolved according to a motif substitution matrix M_{1i}.

To understand this mixture model, we now describe the likelihood of observing the sequences in the current species, for the case $m = 2$. Similar procedures can be performed for other cases.

For two species, the likelihood, given all motif instances including the ancestral ones given, for all parameters, is as follows

$$
\begin{aligned}
&\Pr(S,\ A_j^{(i)}, i = 0, 1, 2;\ j = 1, \cdots, n \mid p_i, \Theta, \Theta_0, w, M_{0i}, M_{1i}, i = 1,2) \\
&= \prod_{j=1}^{n} \{\Pr(A_j^{(0)} \mid \Theta) \prod_{i=1}^{2} \Pr(A_j^{(i)} \mid A_j^{(0)}, M_{1i}, p_i) \sum_{(S_j^{(0)})^c} [\Pr((S_j^{(0)})^c \mid \Theta_0) \prod_{i=1}^{2} \Pr((S_j^{(i)})^c \mid (S_j^{(0)})^c, M_{0i})]\} \quad (0)
\end{aligned}
$$

where $(S_j^{(i)})^c$ means all other nucleotides except those in the motif instances in the jth gene of the ith species. Note that the randomness of $(S_j^{(0)})^c$ will not affect that of $A_j^{(i)}$. This is so because, according to the correspondences among sequences of the same gene across different species, motif instances and background sequences in the current species evolved independently from the motif instances and background sequence in the ancestral species, respectively.

Assume there are n' motif instances in $A^{(0)}$ and all the n' motif instances are aligned, there are n_{kl} nucleotides l (l = A, C, G, or T) at the kth position in the alignment ($k = 1, \cdots, w$). Then we can integrate out Θ, in the above formula (**0**) by multiplying a Dirichlet prior (with parameter 0.5) for Θ, with

$$
C_j = \prod_{i=1}^{2} \Pr(A_j^{(i)} \mid A_j^{(0)}, M_{1i}, p_i) \sum_{(S_j^{(0)})^c} [\Pr((S_j^{(0)})^c \mid \Theta_0) \prod_{i=1}^{2} \Pr((S_j^{(i)})^c \mid (S_j^{(0)})^c, M_{0i})],
$$

to obtain the formula for a marginal likelihood:

$$\begin{aligned}
&\Pr(S,\ A_j^{(i)}, i = 0, 1, 2;\ j = 1, \cdots, n \mid p_i, \Theta_0, w, M_{0i}, M_{1i}, i = 1,2) \\
&= \int \prod_{j=1}^{n} C_j \Pr(A_j^{(0)} \mid \Theta) \Pr(\Theta) d\Theta \\
&= \prod_{j=1}^{n} C_j \int \Pr(A_j^{(0)} \mid \Theta) \Pr(\Theta) d\Theta \\
&= \prod_{j=1}^{n} C_j \prod_{k=1}^{w} \int_0^1 \theta_{k1}^{n_{k1}+0.5} d\theta_{k1} \int_0^{1-\theta_{k1}} \theta_{k2}^{n_{k2}+0.5} d\theta_{k2} \int_0^{1-\theta_{k1}-\theta_{k2}} \theta_{k3}^{n_{k3}+0.5} d\theta_{k3} \int_0^{1-\theta_{k1}-\theta_{k2}-\theta_{k3}} \theta_{k4}^{n_{k4}+0.5} d\theta_{k4} \\
&= \prod_{j=1}^{n} C_j \prod_{k=1}^{w} \frac{\Gamma(2) \prod_{l=1}^{4} \Gamma(n_{kl} + 0.5)}{[\Gamma(0.5)]^4 \Gamma(n'+2)} \qquad (1)
\end{aligned}$$

Formula **1** gives the marginal likelihood, given all motif instances, for the remaining parameters with Θ integrated out. From formula **1**, it is easy to derive the conditional distribution of $A_1^{(0)}$ given all other motif instances and all parameters (without Θ) as follows:

$$
\begin{aligned}
&\frac{\Pr(S,\ A_j^{(i)}, i=0,1,2;\ j=1,\cdots,n \mid p_i,\Theta_0,w,M_{0i},M_{1i},i=1{,}2)}{\sum_{A_1^{(0)}}\Pr(S,\ A_j^{(i)}, i=0,1,2;\ j=1,\cdots,n \mid p_i,\Theta_0,w,M_{0i},M_{1i},i=1{,}2)}\\
&=\frac{[\prod_{j=1}^{n}\prod_{i=1}^{2}\Pr(A_j^{(i)}\mid A_j^{(0)},M_{1i},p_i)]\prod_{k=1}^{w}\frac{\Gamma(2)\prod_{l=1}^{4}\Gamma(n_{kl}+0.5)}{[\Gamma(0.5)]^4\,\Gamma(n'+2)}}{\sum_{A_1^{(0)}}[\prod_{j=1}^{n}\prod_{i=1}^{2}\Pr(A_j^{(i)}\mid A_j^{(0)},M_{1i},p_i)]\prod_{k=1}^{w}\frac{\Gamma(2)\prod_{l=1}^{4}\Gamma(n_{kl}(A_1^{(0)})+0.5)}{[\Gamma(0.5)]^4\,\Gamma(n'(A_1^{(0)})+2)}}\\
&\quad\times\frac{\prod_{j=1}^{n}\sum_{(S_j^{(0)})^c}[\Pr((S_j^{(0)})^c\mid\Theta_0)\prod_{i=1}^{2}\Pr((S_j^{(i)})^c\mid(S_j^{(0)})^c,M_{0i})]}{\prod_{j=1}^{n}\sum_{(S_j^{(0)})^c}[\Pr((S_j^{(0)})^c\mid\Theta_0)\prod_{i=1}^{2}\Pr((S_j^{(i)})^c\mid(S_j^{(0)})^c,M_{0i})]}\\
&=\frac{[\prod_{j=1}^{n}\prod_{i=1}^{2}\Pr(A_j^{(i)}\mid A_j^{(0)},M_{1i},p_i)]\prod_{k=1}^{w}\frac{\Gamma(2)\prod_{l=1}^{4}\Gamma(n_{kl}+0.5)}{[\Gamma(0.5)]^4\,\Gamma(n'+2)}}{\sum_{A_1^{(0)}}[\prod_{j=1}^{n}\prod_{i=1}^{2}\Pr(A_j^{(i)}\mid A_j^{(0)},M_{1i},p_i)]\prod_{k=1}^{w}\frac{\Gamma(2)\prod_{l=1}^{4}\Gamma(n_{kl}(A_1^{(0)})+0.5)}{[\Gamma(0.5)]^4\,\Gamma(n'(A_1^{(0)})+2)}}\\
&=\frac{\prod_{i=1}^{2}\Pr(A_1^{(i)}\mid A_1^{(0)},M_{1i},p_i)\prod_{k=1}^{w}\frac{\Gamma(2)\prod_{l=1}^{4}\Gamma(n_{kl}+0.5)}{[\Gamma(0.5)]^4\,\Gamma(n'+2)}}{\sum_{A_1^{(0)}}\prod_{i=1}^{2}\Pr(A_1^{(i)}\mid A_1^{(0)},M_{1i},p_i)\prod_{k=1}^{w}\frac{\Gamma(2)\prod_{l=1}^{4}\Gamma(n_{kl}(A_1^{(0)})+0.5)}{[\Gamma(0.5)]^4\,\Gamma(n'(A_1^{(0)})+2)}}
\end{aligned}
\tag{2}
$$

Note that we can cancel the probabilities for the background evolution in the denominator and nominator at the second step above. Thus, even though the definition of the likelihood in the formula **0** depends on the ancestral background sequences, these unobserved sequences are actually not needed in the Gibbs updating step.

According to the formula **2**, we can sample $A_1^{(0)}$ in two steps. We calculate the weight matrix by using all $A_j^{(0)}$ except $A_1^{(0)}$. Then we calculate the probability of $A_1^{(0)}$ based on the weight matrix and $A_1^{(i)}$ for $i=1,2$. More precisely, the second step can be implemented position by position for all the positions in $A_1^{(0)}$. An example is given for the *met10* orthologous gene in two yeast species, *Saccharomyces cerevisiae* and *Saccharomyces mikatae* in the paper.

To calculate the conditional probability of $A_1^{(1)}$ given all other motif instances and all parameters is straightforward from formula **1**, which is $\Pr(A_1^{(i)} \mid A_1^{(0)}, M_{1i})$. See the example in the section of the motif instance sampling step in the main paper.