A novel promoter prediction method based on multiple sigma factors model for bacterial genomes

Longshu Yang, Qi Wang, Cheng Yang, Qiuyue Wang, Li Qu, Feifei He and Huaiqiu Zhu,

October 20, 2015

In this supplementary material, we first introduce the parameter initialization procedure, then describe the iterative parameter estimation based on expectation maximization (EM) algorithm or more precisely a general EM method. Finally, the details of thresholds employed in promoter scanning is also presented. Since the probabilistic model of SigmaPromoter for characterizing different σ promoters has elaborated in Method section, and the denotations are quoted from that section as well. All these additional explanations will improve readers' understanding of model and algorithm of SigmaPromoter.

1 Parameter initialization

Though the convergence of EM method has been strictly proved, it is difficult to avoid solution from converging to the local maximum. Furthermore, the initialization of parameters may benefit for achieving the global maximum as the work of Bailey *et al.* demonstrates [1]. Therefore, we introduce a similar parameter initialization process, which is derived from 20 most frequently occurred subsequences embedded in the learning set promoter sequences. Since our model for promoter detection is divided into two PWMs that -10 region and -35 region are represented respectively, therefore the initialization will be conducted for 400 times. We then employ a quantization rule, that the appeared base in the initial subsequence is assigned with 0.52 as its probability of appearance while the probability of other absent bases are assigned with 0.16, is applied to directly convert a subsequence into a PWM (see Table S7 as an example). For all these start point, we calculate the log-likelihood without iteration of parameter re-estimation. Finally, the subsequence combination which lead to the highest log-likelihood is reserved as the true start point, and runs EM iteration still convergence. During this initialization, we assume p_j and q_d obey a uniform distribution to reduce the bias introduced by start site of a signal. Since our procedure has been proved to be effective for determine a good start point, moreover, for a complex model to recognize promoter signal in two PWMs, it is much more efficient than iterate from all 400 start points.

2 The details of SigmaPromoter algorithm

The iterative general EM algorithm applied in SigmaPromoter includes two steps: 1) E step as utilize expectation as the estimation of intermediate variables; and 2) M step: because the partial differentiation of log-likelihood can not be deduced to a analytical solution, thereby we just calculate the expectation for parameter and utilize these estimations as substitutions to calculate log-likelihood as the M step for each iteration. As long as the log-likelihood increases with step by step, this procedure coincide with the condition of general EM algorithm and is also convergent.

2.1 E step: the estimation of intermediate variables by expectation

In this subsection, we denote the parameters W, h, v, g_m as:

$$W = W_1 + W_2, h = j + W_1 + d, v = \{A, C, G, T\},$$

$$g_m = \begin{cases} j & : \quad m = 1, \\ h & : \quad m = 2. \end{cases}$$

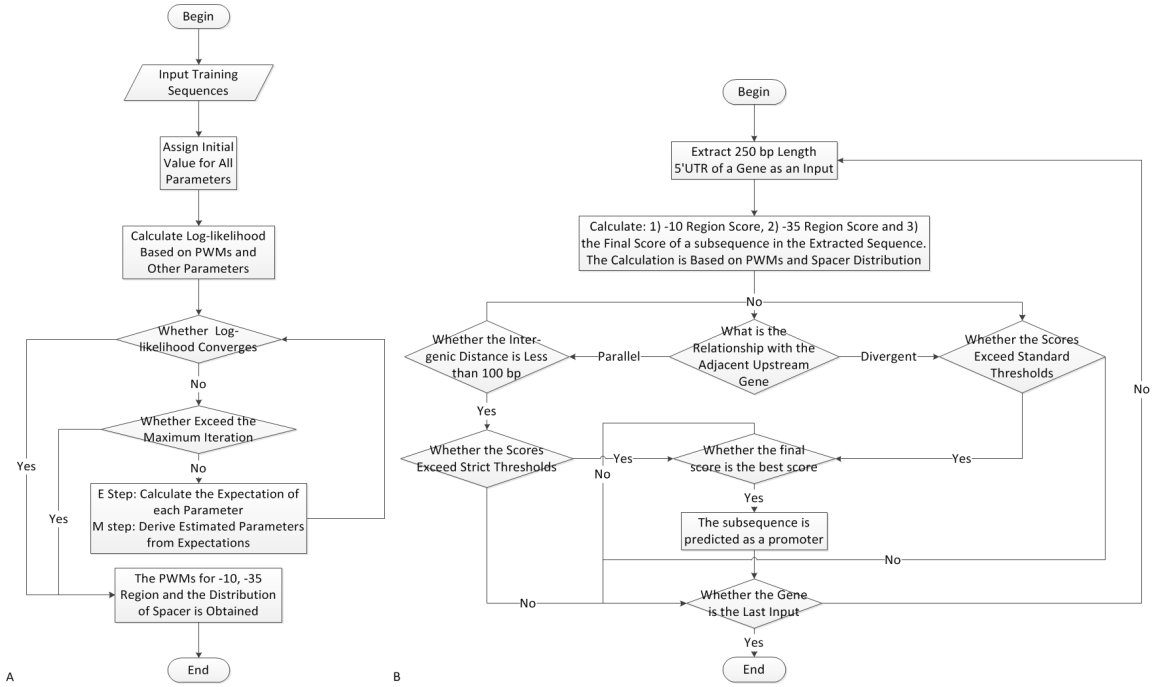

Figure S1: The flow chart of two steps. **(A)** The flow chart of parameter learning by employing EM algorithm. Through this procedure, both PWM elements and spacer distribution are iteratively calculated and these values would converge. **(B)** The flow chart of σ promoters scanning. Through this procedure, every subsequence of 5'UTR of a gene is scanned and classified as promoter sequence or not.

For each sequence S_k involved in training set, for the intermediate variable λ_{kj} as the estimation of p_j is achieved by calculating the marginal probability as follows:

$$\lambda_{kj} = \frac{p_j \sum_{d=d_{min}}^{d_{max}} q_d \prod_{\substack{l \notin [j, j+W_1-1] \\ \cup [h, h+W_2-1]}} b_{S_{kl}} \prod_{i_1=1}^{W_1} w_{1,i_1,S_{k,i_1+j-1}} \prod_{i_2=1}^{W_2} w_{2,i_2,S_{k,i_2+h-1}}}{P\left(S_k|\Omega\right)}, \tag{1}$$

$$\lambda_{k0} = \frac{p_0 \prod_{l=1}^{L} b_{S_{kl}}}{P\left(S_k|\Omega\right)}. \tag{2}$$

Similarly, the estimation of q_d can be calculated by intermediate variable η_{kd} as:

$$\eta_{kd} = \frac{q_d \sum_{j=1}^{L-W+1} p_j \prod_{\substack{l \notin [j, j+W_1-1] \\ \cup [h, h+W_2-1]}} b_{S_{kl}} \prod_{i_1=1}^{W_1} w_{1,i_1,S_{k,i_1+j-1}} \prod_{i_2=1}^{W_2} w_{2,i_2,S_{k,i_2+h-1}}}{P(S_k|\Omega) - p_0 \prod_{l=1}^{L} b_{S_{kl}}}, \tag{3}$$

These intermediate variables are employed to describe the distribution of signal start site in generalized from training set data.

Then, we calculate bases counted as background:

$$B_v = \sum_{k=1}^{N} \left(\sum_{l=1}^{L} T_{klv} - \sum_{j=1}^{L-W+1} \sum_{d=d_{min}}^{d_{max}} \lambda_{kj} \eta_{kd} \left(\sum_{i_1=1}^{W_1} T_{k,i_1+j-1,v} + \sum_{i_2=1}^{W_2} T_{k,i_2+h-1} \right) \right). \tag{4}$$

In this formula,

$$T_{klv} = \begin{cases} 1 & : \quad S_{kl} = v, \\ 0 & : \quad S_{kl} \neq v. \end{cases}$$

Here, $S_{kl} = v$ denotes that the l-th base of S_k is base v, and $S_{kl} \neq v$ denotes that they are not the same base. Furthermore, the bases counted as either -10 region or -35 region is:

$$D_{miv} = \sum_{k=1}^{N} \sum_{j=1}^{L-W+1} \sum_{d=d_{min}}^{d_{max}} \lambda_{kj} \eta_{kd} \sum_{i_m=1}^{W_m} T_{k,i_m+g_m-1,v}. \tag{5}$$

Till now, all intermediate variables for further calculation of $p_0, p_j, q_d, b_v, w_{miv}$ are given by marginal probability and

2.2 M step: variable substitution for maximizing log-likelihood function

In M step, the estimation of $p_0, p_j, q_d, b_v, w_{miv}$ is substituted as the expectation of corresponding intermediate variables as:

$$\hat{p}_j = \frac{\sum_{k=1}^{N} \lambda_{kj}}{\sum_{j=1}^{L-W+1} \sum_{k=1}^{N} \lambda_{kj} + \sum_{k=1}^{N} \lambda_{k0}}. \tag{6}$$

$$\hat{p}_0 = \frac{\sum_{k=1}^{N} \lambda_{k0}}{\sum_{j=1}^{L-W+1} \sum_{k=1}^{N} \lambda_{kj} + \sum_{k=1}^{N} \lambda_{k0}}. \tag{7}$$

$$\hat{q}_d = \frac{\sum_{k=1}^{N} \eta_{kd}}{\sum_{d=d_{min}}^{d_{max}} \sum_{k=1}^{N} \eta_{kd}}. \tag{8}$$

We hence derive the expectation estimation of b_v as:

$$\hat{b}_v = \frac{B_v}{\sum_v B_v}, \tag{9}$$

and

$$\hat{w}_{miv} = \frac{D_{miv}}{\sum_v D_{miv}}. \tag{10}$$

After all these expectation are calculated, we substitute parameters by their estimation to obtain the log-likelihood score until the iteration fulfill the condition of termination. The criterion for convergence is:

$$\delta = \left| L\left(\Omega|S\right)_{t+1} - L\left(\Omega|S\right)_t \right| < 1.0 \times 10^{-6} \cdot \left| L\left(\Omega|S\right)_{t+1} \right|, \tag{11}$$

or iteration step exceeds the maximum step limitation, which is 5000. Here, t represents t-th iterative step. In most case, the iteration converges lower than 500 steps. The flow chart of training process described here are illustrated as Figure S1 A.

Table S7. The example of conversion from sequence to PWM

Base	Sequence					
	T	A	T	A	A	T
A	0.16	0.52	0.16	0.52	0.52	0.16
C	0.16	0.16	0.16	0.16	0.16	0.16
G	0.16	0.16	0.16	0.16	0.16	0.16
T	0.52	0.16	0.52	0.16	0.16	0.52

3 The details of thresholds employed in promoter scanning

As the statistics conducted by Okuda *et al.* indicates, the spacer of two consecutive genes correlates with promoter sequence formation [2]. In Okuda's study, adjacent genes are classified into three operon pairs: operon pair (OP), sub-operon (SOP) and non-operon pair(NOP), within these categories, promoters or terminators may exist between the last two types of gene pairs [2]. Furthermore, the fact that median spacer of SOPs and NOPs are larger than median spacer of OPs demonstrates promoter sequences prefer wider intergenic spacers [2]. The study of Price *et al.* [3] also speculates that a widely spaced operon implies complex regulation and leads to differentiated expression

Table S8. Thresholds for different σ promoter

σ promoter	Relation	Inter-genic distance (bp)	Threshold		
			-35 region	-10 region	Total
σ^{70}	parallel	≥ 100	1.0	3.0	5
	divergent	none	1.0	3.0	5
	parallel	< 100	1.0	3.0	7
σ^{38a}	parallel	≥ 100	NA	NA	4.2
	divergent	none	NA	NA	4.2
	parallel	< 100	NA	NA	6.9
σ^{32}	parallel	≥ 100	2.5	3.0	8
	divergent	none	2.5	3.0	8
	parallel	< 100	2.5	3.0	11.0
σ^{24}	parallel	≥ 100	2.5	3.0	8.4
	divergent	none	2.5	3.0	8.4
	parallel	< 100	2.5	3.0	11.0

[a] σ^{38} promoter contains one binding site, so only the final score is necessary.

pattern—alternative transcription event. All these previous researches prove that the wider the intergenic is spaced, the more probable a regulatory sequence embeds within the pair of genes.

According to these analysis of previous studies, for the promoter scanning procedure, we employ a criterion that the threshold of promoter prediction is lower for gene pairs with a close spacing than for those with a wide spacing. Therein, the intergenic distance greater or equal 100bp is defined as the wide spacing, and the close spacing represents that the intergenic distance is less than 100bp. Table S8 is provided to show the detail thresholds for discrimination on a putative promoter sequence. By integrating the relationship between two consecutive genes with their spacer, these thresholds guarantee a high specificity for SigmaPromoter prediction and characterize promoters to a more rigorous extent. Our accurate prediction shows that some qualitative intrinsic rules may bury in these biological features, and a further analysis is still in urgent need to unveil the complex mechanism of the origin of multiple σ factor induced transcriptional regulatory. The flow chart of promoter scanning process is illustrated as Figure S1 B.

References

[1] Bailey,T.,L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell Syst. Mol. Biol.*, **2**, 28-36.

[2] Okuda,S., Kawashima,S., Kobayashi,K., Ogasawara,N., Kanehisa,M. and Goto,S. (2007) Characteriztion of relationships between transcriptional units and operon structures in *Bacillus subtilis* and *Escherichia coli*. *BMC Genomics*, **8**, 48.

[3] Price,M.,N., Arkin,A.,P. and Alm,E.,J. (2006) The life-cycle of operons. *Plos Genet.*, **6**, e96.