Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2009, Article ID 219743, 11 pages
doi:10.1155/2009/219743

Research Article

Bayesian Unsupervised Learning of DNA Regulatory Binding Regions

Jukka Corander,[1] Magnus Ekdahl,[2] and Timo Koski[3]

[1] *Department of Mathematics, Åbo Akademi University, 20500 Turku, Finland*
[2] *Department of Mathematics, University of Linköping, 58183 Linköping, Sweden*
[3] *Department of Mathematics, The Royal Institute of Technology, 100 44 Stockholm, Sweden*

Correspondence should be addressed to Jukka Corander, jukka.corander@abo.fi

Received 13 February 2009; Revised 6 June 2009; Accepted 2 July 2009

Recommended by Djamel Bouchaffra

Identification of regulatory binding motifs, that is, short specific words, within DNA sequences is a commonly occurring problem in computational bioinformatics. A wide variety of probabilistic approaches have been proposed in the literature to either scan for previously known motif types or to attempt de novo identification of a fixed number (typically one) of putative motifs. Most approaches assume the existence of reliable biodatabase information to build probabilistic a priori description of the motif classes. Examples of attempts to do probabilistic unsupervised learning about the number of putative de novo motif types and their positions within a set of DNA sequences are very rare in the literature. Here we show how such a learning problem can be formulated using a Bayesian model that targets to simultaneously maximize the marginal likelihood of sequence data arising under multiple motif types as well as under the background DNA model, which equals a variable length Markov chain. It is demonstrated how the adopted Bayesian modelling strategy combined with recently introduced nonstandard stochastic computation tools yields a more tractable learning procedure than is possible with the standard Monte Carlo approaches. Improvements and extensions of the proposed approach are also discussed.

1. Introduction

A major body of genomic information coded in the DNA is represented by the regulatory regions, which control the chronology, extent, and way in which genes are expressed. A promoter is the most important regulatory region as it controls the first step of gene expression, that is, the transcription of messenger RNA. A brief summary of and introduction to the biology of promoters are given in [1]. The basic characteristic of a promoter is that it contains binding sites for proteins called transcription factors (TFs). The binding sites in the DNA reside near the gene controlled by the promoters. The interaction of several TFs with their corresponding binding sites regulates the activation or repression of the gene. Hence, the promoter architecture is fundamental for understanding gene expression patterns, regulation networks, and cell specificity. The binding sites of one single TF are not identical, but contain (i.e., tolerate) some variation. The shared content of the binding sites representing the same TF is typically summarized by specifying statistically the degree of conservation in a short (5–20 bases) DNA pattern. The DNA pattern can be understood as a word or a string on a 4-letter alphabet which may contain some variation over its positions when multiple realizations from the same generating source are considered. The substrings of DNA that correspond to the multiple realizations are here called *motif instances* from a gene regulatory binding *motif*. A set of long DNA strings may simultaneously harbor multiple different such motifs, that is, TF binding sites for different genes. In [2, 3] motifs are defined as sets of words of certain length, such that the number of mismatches to a consensus word is smaller than a prescribed threshold. Here, we use instead a probabilistic classification framework to specify the motifs by assigning them to classes in an unsupervised manner.

An important task in computational biology is to detect novel motifs using algorithms that are capable of reasoning from noisy measurements in terms of artificial intelligence. The core of the motif detection problem is to discover novel words whose length is within a biologically meaningful range, such that multiple copies of the words are hidden inside the promoter regions of a set of genes, when the genes have shared properties of, for example, the expression profile. Putative motifs found in a computational manner could thus correspond to binding sites of some common transcription factors regulating a particular set of genes. The algorithms for motif discovery must also be able to avoid suggesting as novel motifs such DNA patterns that represent the background variation in the sequences. In a nutshell, the discovery problem consists of detecting multiple copies of only partially conserved short words in a long string of the characters $\{A, C, G, T\}$, such that the words are highly unlikely to occur under a stochastic null model describing general variation over consecutive letters within the string.

The problem may appear as rather straightforward at first sight. However, several of the involved computational challenges make motif discovery and identification a sincerely complex task [4]. The main obstacles can be listed as follows. A motif exhibits a certain degree of random variation (lack of conservation) in its contents, word lengths are unknown, the background (i.e., areas not containing the target words) structure of the DNA sequences is not random, and the space of putative candidates has an astronomic size. In addition, the motifs may appear in composite patterns, for example, as pairs of words with a fixed distance between them.

A considerable amount of effort has been devoted to solve such problems by statistical model families and by scanning algorithms detecting DNA pattern overrepresentation under a null background model; see, for example, [5–17]. Several additional related methodological papers can be found from the references of the quoted papers. In broad terms the methodological efforts for the above-mentioned problem can be divided into two categories, one where the target is to scan for the existence of *a priori* determined motifs within a set of sequences, and the other where one attempts to identify novel motifs in a given set of sequences. Our approach is solely concerned with the latter category, that is, we consider an unsupervised pattern recognition problem, although some brief remarks on the possibility of extending our method to a partially supervised situation are given in the final section.

Most statistical models for motif discovery use a Markovian probabilistic machinery, for example, ordinary Markov chains or hidden Markov models, to describe the observed patterns of the background DNA and to improve the motif specificity. However, some of the recent works, such as [11, 14], have demonstrated the particular potential of so-called variable-length (or variable-order) Markov (VLMC) models. Such models exemplify a rich class of probabilistic structures for which statistical learning theory has been developed much earlier in the general context [18–20], and which are able to compress information efficiently through a relatively sparse parameterization, while not bargaining the expressiveness of the probabilistic characterization.

In the current paper, the VLMC model is used in two ways. Firstly, candidates of motif instances are obtained by fitting the VLMC model to the sequence data using the algorithm from [20] with the implementation due to [19], and then calculating probabilities for word counts using a compound Poisson approximation. The words which are most improbable under the null background model represent natural candidates of motif instances and can thus be used for an efficient initialization of an unsupervised learning process. Special attention has been devoted in the motif identification literature to the calculation of word count probabilities associated with any particular substrings within the investigated sequences under a null background DNA model see; [17, 21–23]. However, the earlier works have not considered the calculation of such probabilities under a VLMC model, but only under ordinary Markov models of fixed length. Secondly, sufficient statistics (multinomial counts) are obtained from the sequence data under the VLMC context tree identified by the algorithm of [19], and these are used in the Bayesian unsupervised model learning to identify different types of motifs. However, it should be noticed that the sufficient statistics under the VLMC model are recalculated in the unsupervised stochastic learning steps when putative motif instances are shifted along the sequence and when they are reinserted to the background model.

Many Bayesian pattern recognition methods are principally based on the concept of latent class models, which were already early used in motif detection [5, 8]. Standard Bayesian Markov chain Monte Carlo (MCMC) computation tools, such as the Gibbs sampler [24], are available for models with *a priori* fixed numbers of classes. However, numerical convergence and mixing problems for such methods are notorious for challenging applications and motif discovery is not an exception in this respect, for example, see the discussion in [8]. The computational problems may arise already in a situation where only a single motif type is considered in the *de novo* detection context. Such problems are further accentuated in the currently considered framework, where the aim is to make formal statistical inference about the simultaneous presence of multiple putative *a priori* unknown motif types. Bayesian model-based inference in such a setting has been previously considered only by [15], where a reversible jump MCMC algorithm [25] was used for model learning. Motivated by the findings of [26] on general convergence issues with reversible Markov chains, we apply here instead the nonreversible Metropolis-Hastings algorithm they introduced for complex model learning applications.

The structure of this paper is as follows. First, we introduce the background DNA model based on a variable-length Markov process and consider a framework for obtaining putative candidates of motif instances by calculating DNA pattern probabilities under the background model. Thereafter, the unsupervised classification model is derived and in Section 4 we develop a parallel MCMC algorithm for making posterior inferences using a general class of nonreversible Metropolis-Hastings algorithms. Section 5 provides some illustrations of the developed method, and some remarks are given in the final section.

2. A Variable Length Markov Process as the Background DNA Model

Markov chains of higher-order memory were proposed a while ago for computing the expected frequencies of the nucleotides observed outside the motifs, that is, the background information in DNA; see [27] and the references therein. In particular, a third order stationary Markov chain is often chosen; see, for example, [8]. Here we consider a more general model that can handle longer Markov memories in a parsimonious manner. The background model will be formulated by means of the notion of a context of a symbol. This has been introduced to define a variable-length Markov chain (VLMC) as well as a universal finite memory source [18–20, 28]. Only brief details of VLMC models are given here, for a comprehensive treatment we refer the reader to the original references.

We consider a DNA sequence $\mathbf{x}$, which can be a long concatenation of strings of possibly varying length. Let x_r denote a finite substring in $\mathbf{x}$. The values of this string are denoted as $x_{rj} \in \{1,2,3,4\} = \mathcal{X}$, where j refers to an arbitrary position within the string x_r, and $\mathcal{X}$ refers to the set of bases in the DNA. Let $z_n^m = z_m, \ldots, z_n$ be a string in $\mathcal{X}^{m-n+1}$, which starts at position $t = n$ and ends at position $t = m$. Notice that, for $m > n$ the string is written in the reverse order. In the sequel both t and r will be used as generic indices of sequence positions, when defining random processes. A repeated use of the chain rule for a random process yields (written using the short notation $P(Z = z) = P(z)$)

$$P(z_1^n \mid z_{-\infty}^0) = \prod_{r=1}^{n} P(z_r \mid z_{-\infty}^{r-1}). \tag{1}$$

We introduce next the idea of a memory of variable order, in the sense that the conditional probabilities in the above product can depend only on a context relevant for them. Formally, we define the context as a (variable projection) function which maps the whole sequence past into a possibly shorter string.

Definition 1. A context is a map $h : z_{-\infty}^0 \to z_{-q+1}^0$, where q (the context length) is defined by

$$q(z_{-\infty}^0) = \min\left\{ l \in \mathbb{Z}^+ \mid P(z_1 \mid z_{-\infty}^0) = P\left(z_1 \mid z_{-l+1}^0\right) \ \forall z_1 \in \mathcal{X} \right\}. \tag{2}$$

Note that the superscript in $z_{-\infty}^0$ and z_{-q+1}^0 refers to the first element of any string that is given as an input argument to the function h. Such an input string itself may in the sequel be indexed with arbitrary sub- and superscripts, such as $z_{-\infty}^{r-1}$, when it refers to a particular part of a string z. It should also be noted that the context is independent of t, and it can thus be called a stationary context.

Definition 2. Let

$$o = \sup_{z_{-\infty}^0 \in \mathcal{X}^\infty} q(z_{-\infty}^0). \tag{3}$$

If $o < \infty$, then $\{Z_t\}_{t\in\mathbb{Z}}$ is a stationary variable-length Markov chain (VLMC) of order o.

Since the context is stationary, (1) can be written as

$$P(z_1^n \mid z_{-\infty}^0) = \prod_{r=1}^{n} P(z_r \mid h(z_{-\infty}^{r-1})). \tag{4}$$

We recall the kth-order Markov property defined by the equality of the conditional probabilities

$$P(z_1 \mid z_{-\infty}^0) = P\left(z_1 \mid z_{-k+1}^0\right) \tag{5}$$

for all $z_{-\infty}^0$. If the context function turns out to be $h(z_{-\infty}^0) = z_{-k+1}^0$ for all $z_{-\infty}^0$, then the process $\{Z_t\}_{t\in\mathbb{Z}}$ is a full Markov chain of order k. In other words, a VLMC of finite order is then a full Markov chain of order o with a variable memory, and the minimal state space represented by the VLMC contexts has no impact. However, in practice, the minimal state space represented by the VLMC contexts has a considerable impact, since in the genomic applications an estimated VLMC typically has an order of 7 while the size of the set of contexts $|\{h(z_{-\infty}^0)\}|$ is small. In contrast, the memory requirements for the implementation of a full 7th-order ordinary Markov chain are enormous, which hampers the estimation and use of such models.

Using the contexts $\{h\}$ from a VLMC as such may lead to certain problems for a practical implementation with regard to predicting the next context based on the last one. Such problems are illustrated by the following small example.

Example 1. Let $z_1^5 = 43412$. If $q(z_{-\infty}^4) = 1$ and $q(z_{-\infty}^5) = 4$, then $h(z_{-\infty}^4) = 3$ $(3412 \to 3)$. However, we cannot determine the probability of transition from $h(z_{-\infty}^4) = 3$ to $h(z_{-\infty}^5) = 4341$. Note that the superscript in the argument of $q(\cdot)$ refers to a position in the original string, whereas the superscript in the Definition 1 refers to the first element in an arbitrary input argument.

A definition that holds for finite data is also needed to extend the VLMC model to a finite string, such that probability transitions between the contexts can be calculated. The finiteness of real data and transition problems in Example 1 is handled by replacing the context length function q with $\tilde{q}$ according to the following definition.

Definition 3. Let

$$\tilde{q}\left(z_1^\gamma\right) = \min\left[\max\left(q\left(z_{-\infty}^\gamma\right), \tilde{q}\left(z_{-\infty}^{\gamma+1}\right) - 1\right), \gamma\right], \tag{6}$$

such that $\tilde{h}$ corresponding to $\tilde{q}(z_1^\gamma)$ replaces h.

Note that $\widetilde{q}(z_1^\gamma) = q(z_1^\gamma)$ for the largest contexts, implying that $\widetilde{q}(z_1^\gamma)$ can be constructed recursively starting from the largest contexts. The presence of γ in (6) ensures that the context length is well defined for finite sequences and the part with the maximum ensures that the next context $\widetilde{h}(z_1^{\gamma+1})$ can be predicted from the current context $\widetilde{h}(z_1^\gamma)$.

The motif instance candidates are identified by assuming the corresponding words to be overrepresented in the string $\mathbf{x}$ with respect to the background VLMC model. Let R_n be a stochastic number of nonoverlapping instances of a word calculated from the left, that is, more specifically $R_n(\widetilde{z}_1^l)$ equals the number of $\widetilde{z}_1^l$, such that no prefix of a $\widetilde{z}_1^l$ is a suffix to another $\widetilde{z}_1^l$ in Z_1^n. We then seek to compute

$$P\left(R_n\left(\widetilde{z}_1^l\right) \geqslant r_n\right) \tag{7}$$

for the observed occurrences $r_n = r_n(z_j^{j+l-1})$, within the range $l \in [l_{\min}, l_{\max}]$ and $1 \leqslant j \leqslant n+1-l$.

While it is theoretically possible to calculate $P(R_n(\widetilde{z}_1^l) \geqslant r_n)$ exactly for small n as shown in [29], this does not work for the values of n encountered in practice. Firstly, this is due to computational difficulties, and more importantly, the numerical difficulties of the techniques currently available. Instead, we approximate (7) using a compound Poisson distribution

$$R_n \approx \hat{R}_n \sim \mathrm{CP}(\lambda_1, \lambda_2, \ldots) \tag{8}$$

by invoking the results in [30]. The compound Poisson distribution $\mathrm{CP}(\lambda_1, \lambda_2, \ldots)$ refers to the probability distribution of the sum of random number T of random variables T_i

$$\sum_{i=1}^{T} T_i, \tag{9}$$

where $\{T_i\}_{i \geqslant 1}$ are independent, and T is independent of T_i, $P(T_i = k) = \lambda_k/\lambda$ for any positive integer k and every i, and T is Poisson distributed, that is, $T \sim \mathrm{Po}(\lambda)$.

The compound Poisson approximation is here implemented for sequences that are similar to those considered in [31]. We recall that the background model can be seen as a oth-order Markov chain. Any oth-order Markov chain with state space Ω can be transformed into a first-order Markov chain on the state space Ω^o. In the actual implementation the state space $\Omega = \{\widetilde{h}(z_{-\infty}^0)\}$ is designed under the target to keep the number of states low. However, for the discussion below, we assume simply that the model is represented with a state space Ω such that the corresponding r.v. $Y \in \{1, \ldots, |\Omega|\}$ can be constructed to imply a first-order Markov chain $\{Y_1, \ldots, Y_n\}$ with respect to Ω (see Example 2 below).

It is assumed that any putative motif instance $\widetilde{z}_1^l$ can be unambiguously defined as a sequence of states w_1^a (i.e., strings), where any individual element $w_i \in \Omega$. Additional states explicitly representing the transitions in the path $w_1 \to w_2, w_2 \to w_3, \ldots, w_{a-1} \to w_a$ are added to the state space Ω, which leads to the state space Ω'. The probability of generating the sequence $w_1 \to w_2, w_2 \to w_3, \ldots, w_{a-1} \to w_a$ from other states than these special states is set to 0. Then, every time the first-order Markov chain visits the end state of the path, a motif instance candidate has been visited in the DNA sequence. This generic principle is illustrated in Example 2 where the candidate string equals $\widetilde{z}_1^2 = 11$.

Example 2. Suppose that $\Omega = \{0, 1\}$ and $\widetilde{z}_1^2 = 11$. Then $\Omega' = \{0, 1, 11\}$, and the original transition probability matrix

$$P = \begin{pmatrix} P(0 \mid 0) & P(1 \mid 0) \\ P(0 \mid 1) & P(1 \mid 1) \end{pmatrix} \tag{10}$$

is transformed into

$$P = \begin{pmatrix} P(0 \mid 0) & P(1 \mid 0) & 0 \\ P(0 \mid 1) & 0 & P(11 \mid 1) \\ P(0 \mid 11) & P(1 \mid 11) & 0 \end{pmatrix}, \tag{11}$$

where $P(0 \mid 11) = P(0 \mid 1)$ and $P(1 \mid 11) = P(11 \mid 1) = P(1 \mid 1)$.

As seen from the above, the compound Poisson random variable $R_n = \sum_{i=1}^{T} T_i$ approximates the number of occurrences of $\widetilde{z}_1^l$ in a sequence. In the implemented model the First success distribution (Fs) is used for T_i, that is, $T_i \sim \mathrm{Fs}(\theta)$ and $M \sim \mathrm{Po}(\lambda^*)$. Let $\tau_{y_i'} = \inf\{j > 1 \mid Y_j' = y_i'\}$ and set the compound Poisson approximation parameters as in [30], that is, $\theta := P(\tau_{y_l'} < \tau_{y_k'} \mid Y_1' = y_k')$ and $\lambda^* = nP(Y_1' = y_k')P(\tau_{y_l'} < \tau_{y_k'} \mid Y_1' = y_k')$. In general y_l' can be chosen rather freely, for example, to minimize the bound in [30]. In the practical implementation y_l' is instead set to the random variable representing the starting context of the motif, as the execution time for the currently available algorithm for evaluating the approximation bound in [30] is $\sim$1 day on a standard computer, and in this context the typical number of putative motif instance candidates to be evaluated is larger than 10000.

3. An Unsupervised Stochastic Learning Framework for Motif Discovery

The present learning model for motif detection operates by sequentially inserting randomly positioned contiguous windows onto the considered total DNA string and classifying their contents into groups representing putative motif types, such that the number of groups is unknown *a priori*. From an evolutionary perspective, the substrings inside the windows can also be viewed as fragments inserted into specific positions of a previously existing DNA sequence.

We now summarize the concepts for the motif detection task as follows.

(i) A putative gene regulatory *motif* can be interpreted as a statistical model of a binding site which defines for any particular DNA string of an appropriate length how well it fits to that site.

(ii) A *motif instance* can be interpreted as an actual realization of the DNA under a particular regulatory motif model, which specifies a probability distribution over the bases at the positions belonging to

the motif. The unknown (multinomial) parameters that specify this joint distribution are nuisance parameters and they are thus integrated out when making Bayesian inference.

(iii) A *motif type* refers here to any particular motif. The type can be represented by a class of motif instances that are judged by the Bayesian model to correspond to the same motif. We assume that the number of motif models discoverable from the considered DNA sequence is *a priori* unknown and a primary target of the statistical learning. A putative gene regulatory motif type will be here indexed by c, $c = 1, \ldots, K$, such that $K \in [0, k_{\max}]$ is a stochastic number of motif types, each of which corresponds to an *a priori* unknown number of motif instances that are localized in $\mathbf{x}$.

A representation of the components included in our unsupervised model is shown in Figure 1, with the used notation explained in detail below. Now, given $K = k$, let $U_c, c = 1, \ldots, k$, be a stochastic number, or the abundance, of motif instances representing the motif type c, governed by the probability $P(U_c = u_c)$. Further, we let the common length l_c of the U_c motif instances be within the range $[l_{\min}, \ldots, l_{\max}]$, such that $l_{\min}$ and $l_{\max}$ are positive integers with $l_{\min} < l_{\max}$. Here $l_{\min}$, $l_{\max}$ are the lower and upper bounds, respectively, on the motif length that can be modified using biological domain knowledge; see, for example, [11].

Given the k outcomes of (u_c, l_c), there are in total $\sum_{c=1}^{k} u_c$ motif instances to be allocated onto $\mathbf{x}$, such that an arbitrary motif instance of type c contains l_c contiguous positions. Let $i_c \in \{1, \ldots, u_c\}$ index an arbitrary motif instance of type c. Within $\mathbf{x}$, a motif instance is represented by the pair (r_{i_c}, l_c) identifying a substring x_r which contains the motif instance i_c of the length l_c.

The substring representing the motif instance is explicitly denoted as $x_{i_c} = x_{r_{i_c}} x_{r_{i_c}+1} \cdots x_{r_{i_c}+l_c-1}$. The random variables corresponding to the bases observed within a motif instance are governed by the probabilities $\mathbf{p}_{cj} = (p_{cja})_{a \in \mathcal{X}}$, such that p_{cja} represents the probability of observing base a at position j among the instances of type c. Generally, this parametric construction is referred to as a weight matrix in the bioinformatics literature. Given the motif types with the unknown underlying distributions, the bases observed at each position are here modeled as conditionally independent multinomial trials with respect to any type.

An allocation of the motif instances onto $\mathbf{x}$ implies that the sequence $\mathbf{x}$ is split into two distinct parts, one containing the $\sum_{c=1}^{k} u_c$ strings x_{i_c}, and the other containing the remainder of $\mathbf{x}$ representing the background. Hence, we consider the sequence structure in terms of a model of randomly inserted motif instances as in [32].

Let now m denote a particular configuration of a motif allocation model as described above. Thus, m designates u_c motif instances of type c, $c = 1, \ldots, k$, allocated in random positions in a generic sequence. In probabilistic terms, m partitions the data $\mathbf{x}$ into conditionally independent groups. Let $\mathcal{M}$ be the space of all eligible such partitions. The partitions are defined using a set of random variables Y_i according to

$$P\left(Y_i = z_i \mid z_{-\infty}^{i-1}, m\right) = \begin{cases} P_{cj}(z_i = a) = p_{cja}, \\ \quad \text{if } Y_i \text{ is represented by motif type } c \text{ at position } j, \\ P(Z_i = z_i \mid z_{-\infty}^{i-1}), \\ \quad \text{if } Y_i \text{ is in the background.} \end{cases} \tag{12}$$

Both the motif instance classification structure and the sequence realizations are thus embedded into the set of Y_i. An arbitrary realization of these is denoted in the sequel by $\mathbf{y}$. Note here that the motif data depend only on the motif type and position in the motif instance, not the surrounding data. The background model is restarted, with the same probability law, after every motif instance, but we do not burden the notation further by this.

The unsupervised motif discovery model is now formally defined by the joint probability

$$P(\mathbf{y}, m) = P(\mathbf{y} \mid m) P(m), \tag{13}$$

where $P(\mathbf{y} \mid m)$ is the marginal distribution (likelihood) of the data given model m and $P(m)$ is a probability distribution representing our prior uncertainty about different models.

Here, motif discovery task is formulated as joint maximization of the posterior distribution of the background model and the partition of the motif instance candidates into classes representing distinct motif types, which leads to

$$\widehat{m} \in \arg\max_{m \in \mathcal{M}} P(m \mid \mathbf{y}), \tag{14}$$

where

$$P(m \mid \mathbf{y}) = \frac{P(\mathbf{y} \mid m) P(m)}{\sum_{m \in \mathcal{M}} P(\mathbf{y} \mid m) P(m)}. \tag{15}$$

The joint maximization reflects the fact that the marginal likelihood of the sequence data under the background model increases when substrings that have a low probability of occurring are instead considered as motif instances. On the other hand, increasing the number of motif types and their lengths increases the number of multinomial parameters that are needed to represent the probabilities of observing the different bases over the motifs. Thus, the model must target a balance between these two aspects when aiming at a simultaneous identification of the posterior optimal number of motif types k, as well as allocation and alignment of motifs into the classes representing the different types.

To enable efficient identification of the posterior optimal model structure we will use a constant prior $P(m) = 1/|\mathcal{M}|$ for the structural layer of our model. Notice that the space of eligible model structures depends on the *a priori* limits specified for motif lengths and number of motif types. A constant prior can be described to conceptually arise through the following generating model for a set of fragments

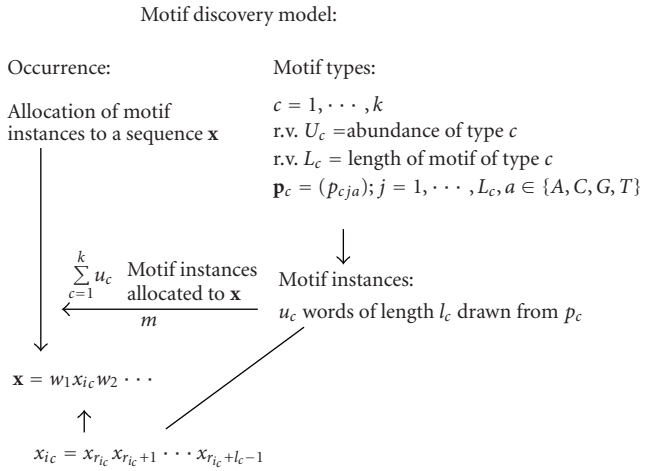

Figure 1: A schematic representation of the unsupervised motif discovery model.

randomly inserted within a DNA sequence. Consider an initial sequence of length w. Given a stochastic total number U of fragments (i.e., all motif instances), one can choose the first positions for the fragments in $\binom{w}{U}$ ways, such that for each of them, a corresponding motif instance of a certain length l_c will be inserted into the sequence z between that position and the subsequent one. The probability for any particular arrangement of the fragments is thus $\binom{w}{U}^{-1}$. Furthermore, every fragment can now be chosen with the probability k^{-1} from the k alternative sources, which specifies the generating model for the prior probabilities of the structural layer. By specifying suitable distributions for U and k, a uniform distribution for the structure m is implied, similarly to the random urn model considered in [33].

Next we derive an expression for the marginal likelihood under the classification model using an approach similar to that adopted in [33]. Let $I_{x_{r_{ic}},j}(a)$ be an indicator function that has value one if the value $a \in \mathcal{X}$ is found at position j in the motif $x_{r_{ic}}$ and has value zero otherwise. Then

$$n_{cja} = \sum_{i_c=1}^{u_c} I_{x_{r_{ic}},j}(a) \tag{16}$$

is the number of times the symbol a appears at position j in the motifs in class c and $n_{cj} = n_c$ is the total number of motif instances of type c. Similarly, let n_{ga} and n_g be the number of times a appears after the context denoted by g and the number of times the context g appears, respectively, where the context can have parts belonging to both the background and one or several motif instances. Recall that the context g refers to a particular string with elements from the alphabet $\mathcal{X}$. Now we can continue with the actual calculation of $P(\mathbf{y} \mid m)$.

We use different priors for the probabilities in the background and the motifs. Both are Dirichlet, however, with hyperparameters α (for the background) and λ (for the motifs). Dirichlet distribution is the standard choice of a prior in Bayesian modeling of DNA sequences, and in addition to the computational convenience related to this prior, there are theoretical arguments supporting it in a variety of different contexts, see, for example, the discussion in [33, 34]. Let

$$\beta_{ga} = P\left(Z_r = a \mid Z_{r-1-q(z_{-\infty}^{r-1})}^{r-1} = g\right) \tag{17}$$

and let us next introduce

$$\mathcal{V}_g = \left\{ \boldsymbol{\beta}_g = \left\{\beta_{ga}\right\}_{a\in\mathcal{X}} \mid \beta_{ga} \geqslant 0, \sum_{a\in\mathcal{X}} \beta_{ga} = 1 \right\}, \tag{18}$$

and the Dirichlet density

$$\phi\left(\boldsymbol{\beta}_g\right) = \Gamma\left(\sum_{a=1}^{4} \alpha_{ga}\right) \prod_{a=1}^{4} \frac{\beta_{ga}^{\alpha_{ga}-1}}{\Gamma\left(\alpha_{ga}\right)} \tag{19}$$

on $\mathcal{V}_g$. Another prior is considered for the motifs

$$\phi\left(\mathbf{p}_{cj}\right) = \Gamma\left(\sum_{a=1}^{4} \lambda_{cja}\right) \prod_{a=1}^{4} \frac{p_{cja}^{\lambda_{cja}-1}}{\Gamma\left(\lambda_{cja}\right)} \tag{20}$$

on the simplex

$$\mathcal{V}_{cj} = \left\{ \mathbf{p}_{cj} = \left\{p_{cja}\right\}_{a\in\mathcal{X}} \mid p_{cja} \geqslant 0, \sum_{a\in\mathcal{X}} p_{cja} = 1 \right\}. \tag{21}$$

Let us set

$$\underline{\boldsymbol{\beta}} = \prod_{g=1}^{|\{g\}|} \boldsymbol{\beta}_g, \qquad \underline{\mathbf{p}} = \prod_{c=1}^{k} \prod_{j=1}^{l_c} \mathbf{p}_{cj} \tag{22}$$

as well as

$$\phi\left(\underline{\boldsymbol{\beta}}, \underline{\mathbf{p}}\right) = \prod_{g=1}^{|\{g\}|} \prod_{c=1}^{k} \prod_{j=1}^{l_c} \phi\left(\boldsymbol{\beta}_g\right) \phi\left(\mathbf{p}_{cj}\right). \tag{23}$$

Consider now the conditional probability $P(Y_{o+1}^n = z_{o+1}^n \mid z_1^o, \underline{\boldsymbol{\beta}}, \underline{\mathbf{p}})$, which can be factorized using the chain rule of probabilities as

$$P\left(Y_{o+1}^n = z_{o+1}^n \mid z_1^o, \underline{\boldsymbol{\beta}}, \underline{\mathbf{p}}\right) = \prod_{i=o+1}^{n} P\left(Y_i = z_i \mid z_1^{i-1}, \underline{\boldsymbol{\beta}}, \underline{\mathbf{p}}\right). \tag{24}$$

Here (12) and the VLMC property in the sense of (4) are used, resulting in

$$P\left(Y_{o+1}^n = z_{o+1}^n \mid z_1^o, \underline{\boldsymbol{\beta}}, \underline{\mathbf{p}}\right) = \prod_{g=1}^{|\{g\}|} \prod_{c=1}^{k} \prod_{j=1}^{l_c} \prod_{a=1}^{4} \beta_{ga}^{n_{ga}} p_{cja}^{n_{cja}}. \tag{25}$$

By combining the above distributions, it is possible to derive the marginal data distribution analytically for our classification model. The formula for the marginal distribution is given in Theorem 1, for which a proof is provided in appendix.

Theorem 1. *If* $P(Y_{o+1}^n = z_{o+1}^n \mid z_1^o, \underline{\boldsymbol{\beta}}, \underline{\mathbf{p}})$ *is given in* (25) *and* $\phi(\underline{\boldsymbol{\beta}}, \underline{\mathbf{p}})$ *is as in* (23)*, then the marginal likelihood under the classification model equals*

$$\begin{aligned} P(Y_{o+1}^n = z_{o+1}^n \mid z_1^o, m) &= \prod_{g=1}^{|\{g\}|} \frac{\Gamma\left(\sum_{a=1}^4 \alpha_{ga}\right)}{\Gamma\left(n_g + \sum_{a=1}^4 \alpha_{ga}\right)} \\ &\times \prod_{c=1}^{k} \prod_{j=1}^{l_c} \frac{\Gamma\left(\sum_{a=1}^4 \lambda_{cja}\right)}{\Gamma\left(\sum_{a=1}^4 \lambda_{cja} + n_c\right)} \\ &\times \prod_{a=1}^{4} \frac{\Gamma\left(n_{ga} + \alpha_{ga}\right)}{\Gamma\left(\alpha_{ga}\right)} \frac{\Gamma\left(\lambda_{cja} + n_{cja}\right)}{\Gamma\left(\lambda_{cja}\right)}. \end{aligned} \tag{26}$$

4. Learning Sequence Partitions by Stochastic Search Operators

Standard MCMC algorithms such as the Gibbs sampler or Metropolis-Hastings algorithm [24] have regularly been used for the earlier inferential problems associated with motif discovery [5]. However, numerical convergence and mixing problems for such methods are burdening the motif discovery task. Furthermore, as the objectives here are even more challenging due to the *a priori* unknown number of motif types, standard stochastic computation is not expected to provide a feasible strategy. The unsupervised learning model for motif discovery developed in the previous section has the characteristics that enable the use of the nonstandard Bayesian computational strategy as developed in [26]. We will now describe the search operators embedded in the nonreversible algorithm.

Let $p_{\text{prop}}(\cdot \mid m)$ denote a generic fixed distribution that assigns probabilities on the space $\mathcal{M}$, conditional on the model structure m. A nonreversible Metropolis-Hastings algorithm may then be constructed by defining its transition kernel according to the acceptance probability

$$\nu(m^* \mid m) = \min\left(1, \frac{P(\mathbf{y} \mid m^*)}{P(\mathbf{y} \mid m)}\right), \tag{27}$$

where m^* is a candidate state (model structure), generated from a current state m under $p_{\text{prop}}(\cdot \mid m)$. The important difference between this algorithm and the ordinary reversible Metropolis-Hastings is the lack of the ratio of the proposal probabilities $p_{\text{prop}}(m \mid m^*)$, $p_{\text{prop}}(m^* \mid m)$ which ensures that the stationary distribution of the generated Markov chain equals the sought posterior distribution. However, as shown by [26], consistent estimates of the model posterior probabilities can still be constructed from a realization of the nonreversible chain. The major strength of this algorithm is that it can use more complex proposal mechanisms than the ordinary reversible MCMC algorithms, when it is not feasible in practice to calculate analytically the proposal probabilities. Also, the direct utilization of the analytically calculated marginal likelihood $P(\mathbf{y} \mid m)$ avoids the effects of Monte Carlo error in the estimation of the underlying parameters, for example, as compared to ordinary Gibbs sampler estimation where values for all model parameters are sequentially sampled. The latter advantage has been discussed also in the earlier motif discovery literature, where collapsed Gibbs samplers have been used for the model learning [5, 8].

In addition to the nonreversible transition kernel of the algorithm, we utilize n parallel interacting search processes analogously to [26]. The search process as a whole can be defined as follows.

Let $\{m_{tj}, t = 0, 1, \ldots; j = 1, \ldots, n\}$ and $\{I_t, t = 0, 1, \ldots\}$ be $n + 1$ stochastic processes defined as follows.

(1) Define a sequence of strictly decreasing probabilities $\{\alpha_t, t = 1, 2, \ldots\}$, such that $\alpha_t > \alpha_{t+1}$, and $\alpha_t \to 0$ as $t \to \infty$.

(2) Define the stochastic process $\{I_t, t = 0, 1, \ldots\}$ as $I_0 = 0$, and $P(I_t = 1) = \alpha_t$, $P(I_t = 0) = 1 - \alpha_t$, independently for $t = 1, 2, \ldots$.

(3) Let m_{0j}, $j = 1, \ldots, n$, be arbitrary initial states of $\{m_{tj}, t = 0, 1, \ldots; j = 1, \ldots, n\}$. Given a realization $\{I_t, t = 0, 1, \ldots\}$, the transition mechanism of the processes $\{m_{tj}, t > 0; j = 1, \ldots, n\}$ depends on values of I_t according to steps (4) and (5).

(4) For each t, such that $I_t = 0$, transition from m_{tj} to the next state $m_{(t+1)j}$ is determined according to the probability (27) under the proposal distribution $p_{\text{prop}}(\cdot \mid m)$, for $j = 1, \ldots, n$.

(5) For each t, such that $I_t = 1$, transition from m_{tj} to the next state $m_{(t+1)j}$ is determined according to the following distribution over the set $\{m_{tj}, j = 1, \ldots, n\}$ of candidate states:

$$P_t\left(m_{(t+1)j} = m_{tj}\right) = \frac{P\left(\mathbf{y} \mid m_{tj}\right)}{\sum_{j=1}^n P\left(\mathbf{y} \mid m_{tj}\right)}, \tag{28}$$

independently for $j = 1, \ldots, n$.

The n processes $\{m_{tj}, t = 0, 1, \ldots; j = 1, \ldots, n\}$ defined above are not time homogeneous Markov chains. However, as $t \to \infty$, their transition probabilities converge to those of the time homogeneous Markov chain defined in (27). The parallel interacting processes are defined to yield an efficient, yet consistent scheme for exchange of information between them. The processes have a tendency to coalesce towards the states which are associated with higher marginal likelihoods. Also, when multiple model structures with roughly equal marginal likelihoods are present, the probabilities (28) lead to a more dispersed proposal distribution.

One of the major challenges in the MCMC computation is the need to design proposal operators for new models. Therefore, our emphasis is on developing intelligent proposals that are able to discriminate the information content of different segments of the data. We now specify

the search operators that generate candidate states according to the distribution $p_{\text{prop}}(\cdot \mid m)$, which in fact consists of several components as typically in MCMC. Some of these are intelligent operators similar to those used in protein sequence classification by [35], and some are locally random operators similar to those used in [26].

Assume first that a nonempty set $\{\mathbf{x}_{r_i}, i = 1, 2, \ldots\}$ of initial candidates of motif instances has been made available out of the total sequence $\mathbf{x}$. Such candidates could also be continuously chosen and/or discarded as a part of the nonreversible algorithm according to a stochastic search operator under our learning model. However, to limit the computational burden, we chose in our experiments to utilize a number of results from renewal theory and compound Poisson approximations to provide a set of putative motif instances (as explained in Section 2).

To identify an optimal classification of the candidate motif instances we use for each process repeatedly the following four search operators in the transition kernel: Merge, Split, Slide, and Move. These are defined as follows.

(1) *Merge*. Let $c_{\max}$ be the motif type with the largest index. For each pair of motif types find the optimal alignment of the motif instances with respect to the marginal likelihood, when the instances are merged into a single group. When the lengths of the motif instances in the two classes are distinct, the bases in the shorter strings that are lacking in a column of any particular alignment are treated as missing data. The marginal likelihood $P(\mathbf{y} \mid m_{tj})$ can then still be calculated as in (26), because the expression is based on the sufficient statistics (counts) arising for each column in the alignment. If any of the $\binom{c_{\max}}{2}$ mergings results in a higher marginal likelihood than $P(\mathbf{y} \mid m_{tj})$, use that model structure as the proposal value m_j^*, else use m_{tj}.

(2) *Split*. Choose a motif type c randomly. Calculate the pairwise Hamming distances of the corresponding motif instances and cluster them using the standard hierarchical single linkage algorithm. Split the group of motif instances optimally into two subgroups according to the hierarchical clustering.

(3) *Slide*. Choose a motif type c randomly. Slide the corresponding motif instances randomly backwards or forwards and change their length randomly. Both the sliding and length change are preformed with respect to a simple uniform proposal distribution. The sliding step relocates the motif instance maximally 5 bases backwards or forwards with equal probabilities for all possible configurations, apart from those that would place the motif instance outside of the sequence. The length is changed by 1 or 2 bases, by randomly either shortening or extending the motif instances with equal probabilities. If the resulting sequence does not satisfy the prior limits $l_{\min}$, $l_{\max}$, the proposal will be discarded and a new value is generated from the original motif instance.

(4) *Move*. Choose a motif instance randomly among all instances and choose a motif type c randomly among the remaining $c_{\max} - 1$ types. Slide the motif instance randomly backwards or forwards and group it with the motif type c.

In addition to the above operators, both randomly chosen single motif instances and motif types are proposed to be reinserted into the background model. Since there is a strictly positive probability of associating each motif instance with any type or the background by a successive use of the proposal steps, the MCMC framework is irreducible. Consequently, the following result holds for the stochastic learning, as the algorithm introduced here satisfies the general conditions stated in [26].

Theorem 2. *Let $\mathcal{M}'$ be a model space for any given set of motif instance candidates, and let $\mathcal{M}_t \subseteq \mathcal{M}'$ be the part of the space explored at time t by the search process defined by the parallel nonreversible Metropolis-Hastings algorithm defined above. Let*

$$\hat{P}_t(m \mid \mathbf{y}) = \begin{cases} \dfrac{P(\mathbf{y} \mid m)}{\sum_{m \in \mathcal{M}_t} P(\mathbf{y} \mid m)}, & \text{if } m \in \mathcal{M}_t, \\ 0, & \text{elsewhere.} \end{cases} \tag{29}$$

Then

$$\hat{P}_t(m \mid \mathbf{y}) \xrightarrow{a.s.} P(m \mid \mathbf{y}), \quad m \in \mathcal{M}' \tag{30}$$

as $t \to \infty$.

5. Empirical Illustration

To briefly illustrate the unsupervised learning framework we have generated simulated data for the motifs reb1, matalpha2, pdr1/3 from the SCPD database [36]. The simulation was performed by first estimating a VLMC based on the SCPD background sequence for these motifs. Then, a sequence was generated using the estimated VLMC model, into which motif data was randomly inserted. Each inserted motif instance was generated using the positional weight matrices of the corresponding motifs [36]. This resulted in a sequence of approximately 10000 nucleotides.

To obtain a set of motif instance candidates, we applied the following procedure. First, we neglected the words z_j^{j+l-1} that occur in $\mathbf{x}$ less than 4 times, because they are deemed to have very small chances of being conclusively judged not to belong to the background DNA sequence. Then, probabilities for the remaining substrings z_j^{j+l-1} were ranked. Those among the 100 most improbable according to the background model, with respect to $P(\hat{R}_n(z_j^{j+l-1}) \geqslant r_n)$, were finally considered as candidates for being motifs. The prior limits of motif length were set according to the interval $[7, 15]$ and the upper bound $K = 10$ was used for the number of motif types.

The behavior of the MCMC optimization for parallel 50 processes is shown visually in Figure 2. Out of the 80 motif instances (30 reb1, 30 matalpha2, 20 pdr1/3) embedded

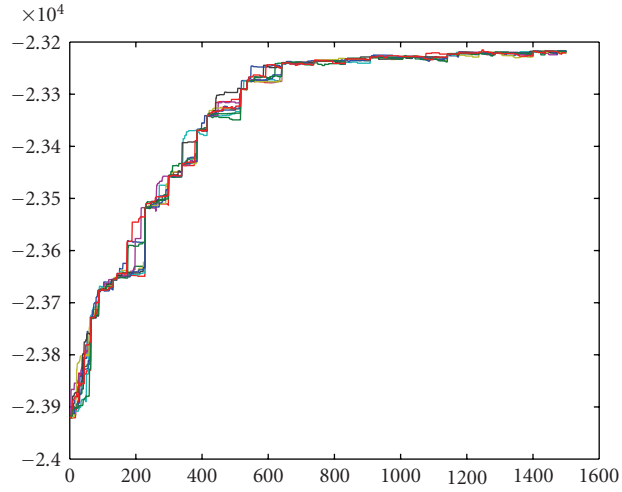

FIGURE 2: An illustration of the behavior of the parallel stochastic search process. The vertical axis corresponds to the $\log P(\mathbf{y} \mid m_{tj})$, and the horizontal axis is the time index.

in the SCPD-like sequence of length approximately 10000 nucleotides, our algorithms were able to identify 64. The adjusted Rand index (see, [37, 38]) between the optimal unsupervised classification of the candidates and the underlying true classification is 0.9623, which indicates a high fidelity of the results for those motif instances that were found.

We also tested the algorithm on real data under comparable prior settings for the motifs MCB and PDR1/3 from SCPD. This was done by providing the upstream regulating regions containing MCB and PDR1/3 (with a total length of 6149 nucleotides) as an input data set for the algorithm. This resulted in 7/32 motifs identified (2 MCB and 5 PDR1/3), which were all correctly grouped with respect to the classification in SCPD.

6. Discussion

The unsupervised classification model was developed under the assumption that the bases at the different positions of a motif instance are conditionally independent given the motif type. This assumption is analogous to the strategy used in a majority of motif identification methods, however, a more elaborate model structure could also be developed by a factorization of the marginal likelihood under a low-order dependence structure, such as a sparse Bayesian network or a context tree similar to that utilized in the construction of the VLMC. This type of a factorization principle has earlier been successfully used in the ordinary likelihood-based motif scanning; see, for example, [11, 13]. The advantage of such likelihood factorizations would in the Bayesian setting be that the marginal likelihoods could still be analytically calculated under suitable Dirichlet priors, thus enabling the utilization of the nonreversible MCMC computation for model learning. However, such a strategy would nevertheless significantly increase the computational complexity associated with the learning procedure, as the size of the model space would increase considerably.

The statistical learning method described here could be modified to incorporate more specific biological knowledge in terms of informative prior distributions for occurrence of certain nucleotides in any given part of a motif instance. Such priors may be designed from well curated biological databases using the basic properties of the Dirichlet distribution. To reduce the computational intensity of our experiments, we used the stochastic search initialized only by the candidate motif instances. Also, the VLMC model for the background was first fitted to the data, whereafter the context tree was not re-estimated during the motif discovery phase. In a fully coherent Bayesian learning procedure it would necessary to consider both aspects entirely as parts of the structural layer of the model and to learn them in parallel with the motif discovery. Similarly, it would be preferable to also monitor the learning process by creating an extensive set of posterior summaries of the motif configurations and visited segments of the sequence as in [15], instead of solely targeting the posterior mode of the model structure.

The fact that sequence scanning yields a large number of motifs that cannot be explained with the knowledge built from the current databases is an open problem in the scanning field, as confirmed by similar studies using YMF [39] and Weeder [40], as well as metastudies such as [41] and even supervised clustering algorithms such as LOGOS [10]. Despite such high "noise" level, we have illustrated that a model-based approach has potential to simultaneously discover multiple motif types hidden in the sequences. From an intuitive perspective such an approach should make more sense than screening for a single motif type at a time. The latter can be suboptimal, for example, in cases where separate motif types are present in the data, such that they are closely related in an evolutionary sense. Here our focus has been more on the mathematical formulation of a statistical framework that has potential for a simultaneous discovery of multiple motif types. We wish to emphasize the advantages of formulating the statistical problem using a Bayesian learning framework which enables the use of nonstandard MCMC computation. In future our aim is to develop more concrete practical tools extending the basic model and the implementation of the learning algorithm, by exploiting a massively parallel computation architecture to pursue large-scale validatory and exploratory experiments.

Appendix

Mathematical details of the derivation of the result in Theorem 1 are given below. For convenience of reference we recall the definition of the Dirichlet integral

$$\int_{\{\mathbf{x} \mid x_i \geqslant 0,\, i=1,\ldots,n,\, \sum x_i = 1\}} \cdots \int x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \cdots x_n^{\alpha_n - 1} d\mathbf{x} = \frac{\prod_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma(\sum \alpha_i)}, \tag{A.1}$$

where $\alpha_i \in \mathbb{R}$ and for all $i, \alpha_i > 0$.

A marginal distribution for z_{o+1}^n given z_1^o is

$$P\left(Y_{o+1}^n = z_{o+1}^n \mid z_1^o, m\right) = \int_{\mathcal{V}_1\times\cdots\times\mathcal{V}_{|\{g\}|}\times\mathcal{V}_{11}\times\cdots\times\mathcal{V}_{kl_k}} \times P\left(Y_{o+1}^n = z_{o+1}^n \mid z_1^o, \underline{\boldsymbol{\beta}}, \underline{\mathbf{p}}\right)\phi\left(\underline{\boldsymbol{\beta}}, \underline{\mathbf{p}}\right)d\left(\underline{\boldsymbol{\beta}}, \underline{\mathbf{p}}\right). \tag{A.2}$$

Now (23) and (25) yield that (A.2) is equal to

$$= \int_{\mathcal{V}_1\times\cdots\times\mathcal{V}_{|\{g\}|}\times\mathcal{V}_{11}\times\cdots\times\mathcal{V}_{kl_k}} \times \prod_{g=1}^{|\{g\}|}\prod_{c=1}^{k}\prod_{j=1}^{l_c}\prod_{a=1}^{4}\beta_{ga}^{n_{ga}}\phi\left(\boldsymbol{\beta}_g\right)p_{cja}^{n_{cja}}\phi\left(\boldsymbol{\beta}_{cj}\right)d\left(\underline{\boldsymbol{\beta}}, \underline{\mathbf{p}}\right). \tag{A.3}$$

The Dirichlet integral can now be used to establish (for each context g)

$$\int_{\mathcal{V}_g}\prod_{a=1}^{4}\beta_{ga}^{n_{ga}}\phi\left(\boldsymbol{\beta}_g\right)d\boldsymbol{\beta}_g = \frac{\Gamma\left(\sum_{a=1}^{4}\alpha_{ga}\right)}{\Gamma\left(n_g + \sum_{a=1}^{4}\alpha_{ga}\right)}\prod_{a=1}^{4}\frac{\Gamma\left(n_{ga}+\alpha_{ga}\right)}{\Gamma\left(\alpha_{ga}\right)} \tag{A.4}$$

and similarly for each motif type c this yields

$$\int_{\mathcal{V}_{cj}}\prod_{a=1}^{4}p_{cja}^{n_{cja}}\phi\left(\mathbf{p}_{cj}\right)d\mathbf{p}_{cj} = \frac{\Gamma\left(\sum_{a=1}^{4}\lambda_{cja}\right)}{\Gamma\left(n_{cj} + \sum_{a=1}^{4}\lambda_{cja}\right)}\prod_{a=1}^{4}\frac{\Gamma\left(n_{cja}+\lambda_{cja}\right)}{\Gamma\left(\lambda_{cja}\right)}. \tag{A.5}$$

Further, using (A.4) and (A.5) the stated result is established.

Acknowledgments

The authors thank Dr. Torkel Erhardsson, Linköpings Universitet, for valuable discussions about compound Poisson approximation. The authors would also like to remember their late friend and colleague Björn Larsson, graduate student in Linköping, who participated in the initial stage of this work before losing his life in a traffic accident. The work was supported by grant no. 4042140 from the Swedish Research Council (VR/NT) and by grant no. 121301 from Academy of Finland.

References

[1] T. Werner, "Models for prediction and recognition of eukaryotic promoters," *Mammalian Genome*, vol. 10, no. 2, pp. 168–175, 1999.

[2] E. Eskin and P. A. Pevzner, "Finding composite regulatory patterns in DNA sequences," *Bioinformatics*, vol. 18, supplement 1, pp. S354–363, 2002.

[3] L. Marsan and M.-F. Sagot, "Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 345–362, 2000.

[4] U. Ohler and H. Niemann, "Identification and analysis of eukaryotic promoters: recent computational approaches," *Trends in Genetics*, vol. 17, no. 2, pp. 56–60, 2001.

[5] J. S. Liu, A. F. Neuwald, and C. E. Lawrence, "Bayesian models for multiple local sequence alignment and Gibbs sampling strategies," *Journal of the American Statistical Association*, vol. 90, pp. 1156–1170, 1995.

[6] W. Thompson, E. C. Rouchka, and C. E. Lawrence, "Gibbs recursive sampler: finding transcription factor binding sites," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3580–3585, 2003.

[7] M. Gupta and J. S. Liu, "Discovery of conserved sequence patterns using a stochastic dictionary model," *Journal of the American Statistical Association*, vol. 98, no. 461, pp. 55–66, 2003.

[8] S. T. Jensen, X. S. Liu, Q. Zhou, and J. S. Liu, "Computational discovery of gene regulatory binding motifs: a Bayesian perspective," *Statistical Science*, vol. 19, no. 1, pp. 188–204, 2004.

[9] S. T. Jensen and J. S. Liu, "BioOptimizer: a Bayesian scoring function approach to motif discovery," *Bioinformatics*, vol. 20, no. 10, pp. 1557–1564, 2004.

[10] E. P. Xing, W. Wu, M. I. Jordan, and R. M. Karp, "Logos: a modular Bayesian model for de novo motif detection," *Journal of Bioinformatics and Computational Biology*, vol. 2, no. 1, pp. 127–154, 2004.

[11] I. Ben-Gal, A. Shani, A. Gohr, et al., "Identification of transcription factor binding sites with variable-order Bayesian networks," *Bioinformatics*, vol. 21, no. 11, pp. 2657–2666, 2005.

[12] L. Hertzberg, O. Zuk, G. Getz, and E. Domany, "Finding motifs in promoter regions," *Journal of Computational Biology*, vol. 12, no. 3, pp. 314–330, 2005.

[13] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan, "Modeling dependencies in protein-DNA binding sites," in *Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB '03)*, pp. 28–37, ACM Press, Berlin, Germany, April 2003.

[14] X. Zhang, H. Huang, M. Li, and T. Speed, "Finding short DNA motifs using permuted Markov models," *Bioinformatics*, vol. 21, pp. 894–906, 2005.

[15] S. M. Li, J. Wakefield, and S. Self, "A transdimensional Bayesian model for pattern recognition in DNA sequences," *Biostatistics*, vol. 9, no. 4, pp. 668–685, 2008.

[16] J. Hawkins, C. Grant, W. S. Noble, and T. L. Bailey, "Assessing phylogenetic motif models for predicting transcription factor binding sites," *Bioinformatics*, vol. 25, no. 12, pp. i339–i347, 2009.

[17] T. Marschall and S. Rahmann, "Efficient exact motif discovery," *Bioinformatics*, vol. 25, no. 12, pp. i356–i364, 2009.

[18] P. Bühlmann and A. J. Wyner, "Variable length Markov chains," *Annals of Statistics*, vol. 27, no. 2, pp. 480–513, 1999.

[19] M. Mächler and P. Bühlmann, "Variable length Markov chains: methodology, computing, and software," *Journal of Computational and Graphical Statistics*, vol. 13, no. 2, pp. 435–455, 2004.

[20] J. Rissanen, "A universal data compression system," *IEEE Transactions on Information Theory*, vol. 29, no. 5, pp. 656–664, 1983.

[21] I. Abnizova and W. R. Gilks, "Studying statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the eukaryotic genomes," *Briefings in Bioinformatics*, vol. 7, no. 1, pp. 48–54, 2006.

[22] M. C. Frith, Y. Fu, L. Yu, J.-F. Chen, U. Hansen, and Z. Weng, "Detection of functional DNA motifs via statistical

over-representation," *Nucleic Acids Research*, vol. 32, no. 4, pp. 1372–1381, 2004.

[23] J. Zhang, B. Jiang, M. Li, J. Tromp, X. Zhang, and M. Q. Zhang, "Computing exact P-values for DNA motifs," *Bioinformatics*, vol. 23, no. 5, pp. 531–537, 2007.

[24] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, New York, NY, USA, 1999.

[25] P. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.

[26] J. Corander, M. Gyllenberg, and T. Koski, "Bayesian model learning based on a parallel MCMC strategy," *Statistics and Computing*, vol. 16, no. 4, pp. 355–362, 2006.

[27] E. E. Stückle, C. Emmrich, U. Grob, and P. J. Nielsen, "Statistical analysis of nucleotide sequences," *Nucleic Acids Research*, vol. 18, no. 22, pp. 6641–6647, 1990.

[28] B. Ron, Y. Singer, and N. Tishby, "The power of amnesia: learning of probabilistic automata with variable memory lengths," *Machine Learning*, vol. 25, pp. 117–149, 1996.

[29] M. Régnier, "A unified approach to word occurrence probabilities," *Discrete Applied Mathematics*, vol. 104, pp. 259–280, 2000.

[30] T. Erhardsson, "Compound Poisson approximation for Markov chains using Stein's method," *Annals of Probability*, vol. 27, no. 1, pp. 565–596, 1999.

[31] T. Erhardsson, "Compound Poisson approximation for counts of rare patterns in Markov chains and extreme sojourns in birth-death chains," *Annals of Applied Probability*, vol. 10, no. 2, pp. 573–591, 2000.

[32] J. L. Thorne, H. Kishino, and J. Felsenstein, "Inching towards reality: an improved likelihood model for sequence evolution," *Journal of Molecular Evolution*, vol. 34, pp. 3–16, 1992.

[33] J. Corander, M. Gyllenberg, and T. Koski, "Random partition models and exchangeability for bayesian identification of population structure," *Bulletin of Mathematical Biology*, vol. 69, no. 3, pp. 797–815, 2007.

[34] D. Geiger and D. Heckerman, "A characterization of the Dirichlet distribution through global and local parameter independence," *Annals of Statistics*, vol. 25, no. 3, pp. 1344–1369, 1997.

[35] P. Marttinen, J. Corander, P. Törönen, and L. Holm, "Bayesian search of functionally divergent protein subgroups and their function specific residues," *Bioinformatics*, vol. 22, no. 20, pp. 2466–2474, 2006.

[36] J. Zhu and M. Q. Zhang, "SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*," *Bioinformatics*, vol. 15, no. 7-8, pp. 607–611, 1999.

[37] B. G. Mirkin and L. B. Chernyi, "Measurement of the distance between distinct partitions of a finite set of objects," *Automation and Remote Control*, vol. 31, pp. 786–792, 1970.

[38] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[39] S. Sinha and M. Tompa, "Discovery of novel transcription factor binding sites by statistical overrepresentation," *Nucleic Acids Research*, vol. 30, no. 24, pp. 5549–5560, 2002.

[40] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole, "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes," *Nucleic Acids Research*, vol. 32, web server issue, pp. W199–W203, 2004.

[41] M. Tompa, N. Li, T. L. Bailey, et al., "Assessing computational tools for the discovery of transcription factor binding sites," *Nature Biotechnology*, vol. 23, no. 1, pp. 137–144, 2005.