Occurrence probability of structured motifs in random sequences

Robin S.[1,2,*], Daudin J.-J.[1], Richard H.[3], Sagot M.-F.[4], Schbath S.[2]

([1]) INA-PG / INRA, UMR Biométrie et Intelligence Artificielle, 16, rue Claude Bernard, F-75005 Paris, France

([2]) INRA, Unité Mathématique, Informatique & Génome, F-78026 Versailles Cedex, France

([3]) CNRS-UPRESA 8071, Laboratoire Statistique et Génomes, 523, place des Terrasses de l'Agora, Tour Evry 2, F-9100 Évry, France

([4]) INRIA Rhônes-Alpes - Laboratoire de Biométrie Évolutive, Univ. Claude Bernard, 43 bd du 11 novembre 16918, F-69622, Villeurbanes cedex, France

Running head: Probability occurrence of structured motifs
Key-words: Markov models, motif occurrences, promoters, structured motifs.

Abstract

The problem of extracting from a set of nucleic acid sequences motifs which may have biological function is more and more important. In this paper, we are interested in particular motifs that may be implicated in the transcription process. These motifs, called structured motifs, are composed of two ordered parts separated by a variable distance and allowing for substitutions. In order to assess their statistical significance, we propose approximations of the probability of occurrences of such a structured motif in a given sequence. An application of our method to evaluate candidate promoters in *E. coli* and *B. subtilis* is presented. Simulations show the goodness of the approximations.

1 Introduction

For many years, biologists, statisticians and computer scientists have been concerned about the problem of extracting from a set of nucleic acid sequences motifs which may have a biological function (typically, such motifs represent DNA binding sites). The motifs that may be considered in terms of algorithms are becoming increasingly sophisticated. In particular, various motifs may be extracted simultaneously. This enables us to start addressing a possible cooperative effect between binding sites implicated in related biological processes such as transcription for instance. However, it is in the statistical evaluation of such motifs that we have being lagging behind. Yet it is important to be able to evaluate with enough accuracy how unexpected are such motifs given a model for the sequences. Indeed, exhaustively extracting motifs that satisfy certain constraints (e.g., maximum error rate allowed, minimum frequency of sequence occurrence) is not enough in most cases. Especially when trying to identify weakly conserved motifs, flexibility is required and this may lead to many motifs which satisfy the constraints. Most of these motifs can be "explained" by the sequence composition and only a few stand a chance of being related to a "true" biological object, such as a binding site.

The complex motifs that we are becoming increasingly more efficient in extracting from a set of sequences correspond to what has been called "structured motifs" (see Marsan and Sagot (2000a), (2000b)). They are motifs that may be composed of two or more ordered parts, called "boxes". Each box is separated from the next one by a distance which may take any value inside an interval. The intervals may be different for two pairs of consecutive boxes in a structured motif. Each box may also exhibit a different degree of conservation, i.e., the maximum number of errors allowed (in general, only substitutions) against the corresponding part in the motif may vary from one box to the other. Finally, the structured motif must appear in a minimum number of the input sequences.

Various ways of calculating the probability for a motif composed of just one box to occur in a sequence, possibly allowing substitutions, have been known for some time (Blom and Thorburn (1982), Fu (1996), Robin and Daudin (2001)). Recently, the problem of calculating the probability of exact occurrence of a motif composed of two boxes separated by a fixed distance has also been addressed by van Helden *et al.* (2000). However, the problem of calculating the probability of occurrence of a motif composed of two or more boxes separated by variable distances and allowing for substitutions has not yet been faced. This paper proposes a first method for addressing this problem in the case of structured motifs composed of two boxes separated by a variable distance. It is well known that the probability of occurrence strongly depends on the overlapping structure of the motif, i.e. on the possibility for a motif to overlap itself. Since this overlapping structure cannot be completely described in the case of structured motifs, the method proposed in this paper is not exact but uses generalized geometric approximations in order to calculate the required probability with a good precision.

The paper is organized as follows. Section 2 defines a structured motif and recalls a combinatorial algorithm for extracting structured motifs given certain constraints they must satisfy. Section 3 presents a way of approximating the probability of occurrence of structured motifs. Finally, Section 4 presents an application of the method to obtain the statistical significance of the occurrences of some particular structured motifs in sequences coming just upstream from genes of *B. subtillis* (first dataset) and *E. coli* (second dataset). The particular structured motifs are candidate promoters. The results are compared, both in terms of performance and accuracy, to statistics obtained by using simulation.

2 Structured motifs

2.1 Basic definitions

Let Σ be the alphabet of nucleotides, Σ^+ the set of all non-empty words on the Σ alphabet. We denote by

$$\mathbf{m} = [\mathcal{V}(d_1 : d_2)\mathcal{W}]$$

a structured motif composed of two boxes (sets of words) $\mathcal{V} \in \Sigma^+$ and $\mathcal{W} \in \Sigma^+$ separated from one another by a distance d between d_1 and d_2 ($d_1 \leq d \leq d_2$), for d_1 and d_2 two non negative integers that are fixed. In the particular case where $\mathcal{V}$ and $\mathcal{W}$ are reduced to unique words $\mathbf{v}$ and $\mathbf{w}$, the structured motif is simply denoted by

$$\mathbf{m} = [\mathbf{v}(d_1 : d_2)\mathbf{w}].$$

We now define what it means for a structured motif $\mathbf{m}$ to *occur* in a sequence $\mathbf{S}$ of length ℓ. We must first define what it means for a box $\mathcal{V}$ of a motif $\mathbf{m}$ to occur in $\mathbf{S}$. $\mathcal{V}$ is said to occur in $\mathbf{S}$ if $\mathbf{S}$ contains at least one occurrence of a word $\mathbf{x}$ of $\mathcal{V}$, that is, if there exists $\mathbf{S}', \mathbf{S}'' \in \Sigma^+ \cup \emptyset$ such that $\mathbf{S} = \mathbf{S}'\mathbf{x}\mathbf{S}''$. In our case, we are interested in $\mathcal{V} = \mathcal{N}_e(\mathbf{v})$ and $\mathcal{W} = \mathcal{N}_e(\mathbf{w})$ where $\mathbf{v}$ and $\mathbf{w}$ are fixed words of Σ^+ and $\mathcal{N}_e$ denotes the e-neighborhood. The e-neighborhood of $\mathbf{v}$ is the set of all words that may be obtained from $\mathbf{v}$ by making at most e substitutions; $\mathbf{v}$ and all its e-neighbors have the same length denoted by $|\mathbf{v}|$. Our approach could be generalized to insertions or deletions (indels) since allowing indels simply enlarges the neighborhood. However, neither the algorithm nor the probabilistic approximations presented here can be directly applied to this case, mainly because the words included in the neighborhood would have not all the same length. The set $\mathcal{N}_e(\mathbf{v})$ contains therefore all the words of $\Sigma^{|\mathbf{v}|}$ that are at a Hamming distance at most e from $\mathbf{v}$. A structured motif $\mathbf{m}$ is now said to *occur* in a sequence $\mathbf{S}$ if there is at least a pair of occurrences $(\mathbf{x}, \mathbf{x}')$ of $\mathcal{V} \times \mathcal{W}$, such that the distance between the end position of $\mathbf{x}$ and the start position of $\mathbf{x}'$ is between d_1 and d_2.

Finally, a structured motif $\mathbf{m}$ is said to be *valid* in relation to a set of sequences $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \cdots\}$ if, given a percentage q, $\mathbf{m}$ occurs in at least $q|\mathcal{S}|$ sequences of $\mathcal{S}$ (it may occur more than once in a sequence). In the remainder, q is called the *quorum*.

2.2 Extracting structured motifs

The basic idea of the algorithm consists in efficiently spelling all valid structured motifs, that is, motifs satisfying the constraints (maximum error rate, distance between the two boxes and quorum). This is obtained by dynamic programming between a lexicographic tree $\mathcal{M}$ of the potential motifs with an index of the sequences. The index is given under the form of a generalized suffix tree $\mathcal{T}$ (Bieganski *et al.* (1994), Crochemore and Rytter (1994), Gusfield (1997)) and is built in a preprocessing step. The tree $\mathcal{M}$ of potential motifs is virtual: it is never built. Its traversal is done on the fly at the same time as a traversal of $\mathcal{T}$. The dynamic programming recurrence formulas are easy to obtain when substitutions only are allowed. They indicate how to find the occurrences of a motif $\mathbf{m}\alpha$, $\alpha \in \Sigma$, from the occurrences of $\mathbf{m}$. For ease of exposition, we assume that the structured motifs $\mathbf{m}$ sought are composed of two boxes, each of same length k, separated by a fixed distance d. Occurrences of $\mathbf{m}$ are nodes, possibly virtual (Marsan and Sagot (2000a), (2000b)) in $\mathcal{T}$. They are identified by a pair (v, e_v) where v is the node and e_v is the number of substitutions between the motif $\mathbf{m}$ and the string labeling the path from the root of $\mathcal{T}$ to v (such path is noted s_v). Then:

Algorithm 2.1 *Let e be the maximum number of substitutions allowed, and let us note $ancestor_\lambda(v)$ the ancestor at distance λ of node v in $\mathcal{T}$ (when λ is 1, such ancestor is the father of v – it is denoted by $father(v)$). The pair (v, e_v) is an occurrence of motif $\mathbf{m}\alpha$ if:*

case $|s_v| \leq k$ or $|s_v| \geq (k+d)$: *the pair $(ancestor_\lambda(v), e_{ancestor_\lambda(v)})$ is an occurrence of $\mathbf{m}$ where λ is d if $|s_v| = (k+d)$, 1 otherwise and:*

- *there exists an arc leaving $father(v)$ with a label starting with the letter α (in this case e_v will be equal to $e_{father(v)}$);*
- *such an arc does not exist, but $e_{father(v)} < e$ (in this case, e_v will be equal to $e_{father(v)} + 1$);*

case $|s_v| = (k+1)$: *the pair $(father(v), e_{father(v)})$ is an occurrence of $\mathbf{m}$.*

In the last case, a jump is made in the tree directly to all the k^{th} descendants of $father(v)$. Such nodes represent a potential starting point for a second box of the motif being spelled. The jump may be performed in two different ways. It is not essential for the purposes of this paper to describe them. Details may be found in Marsan and Sagot (2000a) or (2000b).

It is important to observe that the algorithm is exact. It will exhaustively enumerate all the structured motifs that are valid given the parameters. The algorithm is also efficient in practice, as expected considering its theoretical complexity given in next section.

2.3 Complexity of the algorithm

Depending on how the jump is done in the tree $\mathcal{T}$ once a first box for a structured motif $\mathbf{m}$ has been identified, the time complexity of the algorithm may be $O(|\mathbf{S}|\nu_{2k+d_2}n^2(e,k))$ (tree $\mathcal{T}$ remains unchanged) or $O(|\mathbf{S}|\nu_k n^2(e,k) + |\mathbf{S}|\nu_{2k+d_2}n(e,k))$ (tree $\mathcal{T}$ is temporarily and locally modified) where $|\mathbf{S}|$ is the total number of sequences, ν_λ is the number of (possibly virtual) nodes at level λ in $\mathcal{T}$, and $n(e, k)$ is the number of e-neighbors of a k-letter word. Note that $\nu_\lambda \leq \nu_{\lambda+1} < \ell + |\mathbf{S}|$ where ℓ is the average length of the $|\mathbf{S}|$ sequences. The asymptotic space complexity is the same in both cases, $O(|\mathbf{S}|^2\ell)$, although the second approach has a higher constant.

3 Occurrence probability of a structured motif

In this section, we deal with the problem of calculating the probability for a structured motif $\mathbf{m} = [\mathcal{V}(d_1 : d_2)\mathcal{W}]$ to occur in a sequence $\mathbf{S}$ of length ℓ. This problem is equivalent to the waiting time problem raised in the original paper of Blom and Thorburn (1982). The exact distribution of the waiting time before a given word occurs in a random sequence is given in the Markovian case in Robin and Daudin (1999) and (2001), together with the distribution of the distance between successive occurrences. All these distributions are characterized by their generating functions.

From now on we shall assume that the sequence $\mathbf{S} = (S_1, S_2, \ldots, S_\ell)$ is a stationary Markov chain with transition probabilities $\pi(a, b) = \Pr\{S_{x+1} = b \mid S_x = a\}$, where $a, b \in \Sigma$. Each letter S_x of $\mathbf{S}$ is distributed according to the stationary distribution μ of the Markov chain; μ is defined by $\mu(b) = \sum_{a\in\Sigma} \mu(a)\pi(a, b)$ for all $b \in \Sigma$.

Overlapping structure. One of the most important results given in the papers cited above is that the distribution of the waiting time of a word $\mathbf{v}$ depends on the overlapping structure of the word, i.e. on the possibility for $\mathbf{v}$ to overlap itself. In the case of complex motifs like those studied here, this overlapping structure cannot be completely described so the exact distribution of the waiting time can not be provided. However, we are able to take into account the overlapping structures of the words $\mathbf{v}$ and $\mathbf{w}$.

3.1 Generalized geometric approximations

We want to calculate the probability $\gamma(\ell) = \Pr\{\mathbf{m} \in \mathbf{S}\}$. We use the convention that $\{\mathbf{m} \text{ at } x\}$ means that a word of $\mathcal{V}$ starts at x and a word of $\mathcal{W}$ starts y positions later with $|\mathbf{v}| + d_1 \leq y \leq |\mathbf{v}| + d_2$, and $\{\mathbf{m} \text{ not at } x\}$ means either that no word of $\mathcal{V}$ starts at x or that a word of $\mathcal{V}$ starts at x but no word of $\mathcal{W}$ starts y positions later with $|\mathbf{v}| + d_1 \leq y \leq |\mathbf{v}| + d_2$. This probability $\gamma(\ell)$ can be decomposed as follows:

$$\begin{aligned} 1 - \gamma(\ell) &= \Pr\{\mathbf{m} \text{ not at } 1, \mathbf{m} \text{ not at } 2, \ldots, \mathbf{m} \text{ not at } \ell - |\mathbf{m}| + 1\} \\ &= \Pr\{\mathbf{m} \text{ not at } 1\} \times \Pr\{\mathbf{m} \text{ not at } 2 \mid \mathbf{m} \text{ not at } 1\} \times \ldots \\ &\quad \times \Pr\{\mathbf{m} \text{ not at } \ell - |\mathbf{m}| + 1 \mid \mathbf{m} \text{ not at } 1, \ldots, \mathbf{m} \text{ not at } \ell - |\mathbf{m}|\} \end{aligned} \quad (1)$$

where $|\mathbf{m}|$ is the maximal length of the motif (i.e. $|\mathbf{v}| + |\mathbf{w}| + d_2$). We propose to approximate $\Pr\{\mathbf{m} \in \mathbf{S}\}$ by considering the past only up to a fixed order.

Order 0. Approximating $\Pr\{\mathbf{m}$ not at $x \mid \mathbf{m}$ not at $1, \ldots, \mathbf{m}$ not at $x-1\}$ by $\Pr\{\mathbf{m}$ not at $x\} = \Pr\{\mathbf{m}$ not at $1\}$, we get the usual geometric approximation

$$\gamma(\ell) \simeq \gamma_0(\ell) = 1 - [1 - \mu(\mathbf{m})]^{\ell - |\mathbf{m}| + 1} \simeq 1 - \exp[-(\ell - |\mathbf{m}| + 1)\mu(\mathbf{m})] \tag{2}$$

where $\mu(\mathbf{m}) = \Pr\{\mathbf{m}$ at $x\}$. The last approximation is valid for a small $\mu(\mathbf{m})$.

Order 1. Approximating $\Pr\{\mathbf{m}$ not at $x \mid \mathbf{m}$ not at $1, \ldots, \mathbf{m}$ not at $x-1\}$ by $\Pr\{\mathbf{m}$ not at $x \mid \mathbf{m}$ not at $x-1\}$ and denoting $\mu_1(\mathbf{m}) = \Pr\{\mathbf{m}$ at $x \mid \mathbf{m}$ not at $x-1\}$, we get

$$\begin{aligned}\gamma(\ell) \simeq \gamma_1(\ell) &= 1 - [1 - \mu(\mathbf{m})][1 - \mu_1(\mathbf{m})]^{\ell - |\mathbf{m}|} \\ &\simeq 1 - [1 - \mu(\mathbf{m})] \exp[-(\ell - |\mathbf{m}|)\mu_1(\mathbf{m})].\end{aligned} \tag{3}$$

The last approximation is valid for a small $\mu_1(\mathbf{m})$.

Order k. Denoting $\mu_k(\mathbf{m}) = \Pr\{\mathbf{m}$ at $x \mid \mathbf{m}$ not at $x-1, \ldots, \mathbf{m}$ not at $x-k\}$ and $\mu_0(\mathbf{m}) = \mu(\mathbf{m})$, we define the approximation of order k:

$$\gamma(\ell) \simeq \gamma_k(\ell) = 1 - [1 - \mu_k(\mathbf{m})]^{(\ell - |\mathbf{m}| - k + 1)} \prod_{i=0}^{k-1} [1 - \mu_i(\mathbf{m})]. \tag{4}$$

We use the results on the exact distribution of the distance between words to calculate the probabilities $\mu_k(\mathbf{m})$. For the ease of the reader, we first show how to calculate the probability $\mu_k(\mathbf{m})$ when the boxes $\mathcal{V}$ and $\mathcal{W}$ are reduced to single words $\mathbf{v}$ and $\mathbf{w}$, i.e. when the number e of allowed errors is null (see Section 2.1). We then consider the case of multiple words, i.e. when $e > 0$.

In the application part (Section 4), we limit ourselves to the case $k \leq 1$ since this order seems good enough. in our applications.

3.2 Positions and distances

We first have to define some random variables (r.v.'s). Let us consider the sets of words $\mathcal{V} = \{\mathbf{v}_i\}_{1 \leq i \leq I}$ and $\mathcal{W} = \{\mathbf{w}_j\}_{1 \leq j \leq J}$. According to Section 2.1, $\mathcal{V}$ (respectively $\mathcal{W}$) is the e-neighborhood of a word $\mathbf{v}$ (respectively $\mathbf{w}$), so all its elements have the same length $|\mathbf{v}|$ (respectively $|\mathbf{w}|$). Let us define the following r.v.'s:

- $X_{\mathcal{V}}$: the position of the first occurrence of any word of $\mathcal{V}$. This position is given by the position of the first letter of the word: $X_{\mathcal{V}} = \inf\{x : (S_x, \ldots S_{x+|\mathbf{v}|-1}) \in \mathcal{V}\}$, with the convention that $X_{\mathcal{V}}$ is infinite if no element of $\mathcal{V}$ occurs in $\mathbf{S}$.

- $X_{\mathbf{v}_i}(\mathcal{V})$: the position of the first occurrence of $\mathbf{v}_i$ if no other element of $\mathcal{V}$ has occurred before. Formally

$$X_{\mathbf{v}_i}(\mathcal{V}) = \begin{cases} X_{\mathcal{V}} & \text{if } \mathbf{v}_i \text{ is the first word of } \mathcal{V} \text{ to occur,} \\ \infty & \text{otherwise.} \end{cases}$$

  The distribution of $X_{\mathbf{v}_i}(\mathcal{V})$ is denoted $p_{\mathbf{v}_i}(x) = \Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = x\}$.

- $Y_{\mathbf{v}_i,\mathcal{W}}$: the distance between the start of an occurrence of $\mathbf{v}_i$ and the first following occurrence of an element of $\mathcal{W}$. Let x be the position of (the start of) an occurrence of $\mathbf{v}_i$; we have $Y_{\mathbf{v}_i,\mathcal{W}} = \inf\{y \geq 0 : (S_{x+y}, \dots S_{x+y+|\mathbf{w}|-1}) \in \mathcal{W}\}$.

- $Y_{\mathbf{v}_i,\mathbf{w}_j}(\mathcal{W})$: the distance between an occurrence of $\mathbf{v}_i$ and the first following occurrence of $\mathbf{w}_j$ if no other element of $\mathcal{W}$ has occurred before. Formally

$$Y_{\mathbf{v}_i,\mathbf{w}_j}(\mathcal{W}) = \begin{cases} Y_{\mathbf{v}_i,\mathcal{W}} & \text{if } \mathbf{w}_j \text{ is the first word of } \mathcal{W} \text{ to occur after } \mathbf{v}_i, \\ \infty & \text{otherwise.} \end{cases}$$

  This distance can be null, for example if $\mathcal{V}$ is reduced to one word of one letter (say a) and if $\mathcal{W}$ is reduced to one word w beginning with a. The distribution of $Y_{\mathbf{v}_i,\mathbf{w}_j}(\mathcal{W})$ is denoted $q_{\mathbf{v}_i,\mathbf{w}_j}(y) = \Pr\{Y_{\mathbf{v}_i,\mathbf{w}_j}(\mathcal{W}) = y\}$.

- $Y_{A,\mathbf{w}_j}(\mathcal{W})$: the distance between the random letter A and the first following occurrence of $\mathbf{w}_j$ if no other element of $\mathcal{W}$ has occurred before (the distribution of $Y_{A,\mathbf{w}_j}(\mathcal{W})$ depends on the distribution of A). The cumulative probability function of $Y_{A,\mathbf{w}_j}(\mathcal{W})$ depends on the distribution of A:

$$\Pr\{Y_{A,\mathbf{w}_j} \leq y\} = \sum_{a \in \Sigma} \Pr\{Y_{a,\mathbf{w}_j} \leq y\} \Pr\{A = a\}.$$

The distributions of all these r.v.'s are given in Robin and Daudin (2001). They can be calculated thanks to the recursive formulas given in appendix A.1.

The r.v.'s $Y_{\mathbf{v}_i,\mathbf{w}_j}$ (resp. $Y_{a,\mathbf{w}_j}$) are defined conditionally to the occurrence of the word $\mathbf{v}_i$ (resp. of the letter a).

3.3 Calculation of μ and μ_1 for single words

We consider here the case where the sets $\mathcal{V}$ and $\mathcal{W}$ are respectively reduced to $\mathbf{v}$ and $\mathbf{w}$.

Calculation of $\mu(\mathbf{m})$. To calculate the probability $\mu(\mathbf{m})$, we have to consider only the occurrences of $\mathbf{w}$ starting between positions $|\mathbf{v}| + d_1$ and $|\mathbf{v}| + d_2$ after the start of $\mathbf{v}$. We decompose this probability according to all the possible values of the random letter $A(\mathbf{v})$ that occurs in position $|\mathbf{v}| + d_1 + 1$ after the start of $\mathbf{v}$:

$$\begin{aligned} \mu(\mathbf{m}) &= \mu(\mathbf{v}) \Pr\{Y_{A(\mathbf{v}),\mathbf{w}} \leq d_2 - d_1\} \\ &= \mu(\mathbf{v}) \sum_{a \in \Sigma} \pi^{d_1+1}(v_{|\mathbf{v}|}, a) \sum_{y \leq d_2 - d_1} q_{a,\mathbf{w}}(y). \end{aligned} \tag{5}$$

where $\mu(\mathbf{v})$ is the probability that the word $\mathbf{v}$ occurs at a given position, namely, $\mu(\mathbf{v}) = \mu(v_1) \prod_{u=2}^{|\mathbf{v}|} \pi(v_{u-1}, v_u)$ where v_u is u-th letter of $\mathbf{v}$. The distribution of the letter $A(\mathbf{v})$ is given by $\Pr\{A(\mathbf{v}) = a\} = \pi^{d_1+1}(v_{|\mathbf{v}|}, a)$ where $\pi^d(a, b)$ is the transition probability from the letter a to the letter b in d steps, i.e. the element (a, b) of the power d of the transition matrix π.

In the case of a fixed distance d_1 $(= d_2)$ between $\mathbf{v}$ and $\mathbf{w}$, the equation (5) simply is:

$$\mu(\mathbf{m}) = \mu([\mathbf{v}(d_1)\mathbf{w}]) = \mu(\mathbf{v})\pi^{d_1+1}(v_{|\mathbf{v}|}, w_1)\frac{\mu(\mathbf{w})}{\mu(w_1)}.$$

In the general case, the probability in the right hand side of (5) can be calculated using the formulas given in Appendix A.1.

Calculation of $\mu_1(\mathbf{m})$. To calculate $\mu_1(\mathbf{m}) = \Pr\{\mathbf{m} \text{ at } x \mid \mathbf{m} \text{ not at } x-1\} = \Pr\{\mathbf{m} \text{ at } 2 |$ $\mathbf{m}$ not at $1\}$, we decompose the event $\{(\mathbf{m} \text{ at } 2) \cap (\mathbf{m} \text{ not at } 1)\}$ as

$$\big\{(X_{\mathbf{v}} = 2) \cup [(X_{\mathbf{v}} = 1) \cap (Y_{A(\mathbf{v}),\mathbf{w}} > d_2 - d_1)]\big\} \cap \{\mathbf{m} \text{ at } 2\}.$$

We then get

$$\begin{aligned}
&\Pr\{(\mathbf{m} \text{ at } 2) \cap (\mathbf{m} \text{ not at } 1)\} \\
&= \Pr\{X_{\mathbf{v}} = 2\} \Pr\{Y_{A(\mathbf{v}),\mathbf{w}} \leq d_2 - d_1\} + \Pr\{X_{\mathbf{v}} = 1\} \Pr\{Y_{\mathbf{v},\mathbf{v}} = 1\} \Pr\{Y_{A(\mathbf{v}),\mathbf{w}} = d_2 - d_1\} \\
&= p_{\mathbf{v}}(2) \sum_{a\in\Sigma} \pi^{d_1+1}(v_{|\mathbf{v}|}, a) \sum_{y \leq d_2 - d_1} q_{a,\mathbf{w}}(y) + p_{\mathbf{v}}(1) q_{\mathbf{v},\mathbf{v}}(1) \sum_{a\in\Sigma} \pi^{d_1+1}(v_{|\mathbf{v}|}, a) q_{a,\mathbf{w}}(d_2 - d_1)
\end{aligned}$$

and therefore, finally

$$\mu_1(\mathbf{m}) = \Pr\{(\mathbf{m} \text{ at } 2) \cap (\mathbf{m} \text{ not at } 1)\}/[1 - \mu(\mathbf{m})].$$

3.4 Calculation of μ and μ_1 for multiple words

We now consider the general case of two boxes $\mathcal{V} = \{\mathbf{v}_i\}_{1\leq i\leq I}$ and $\mathcal{W} = \{\mathbf{w}_j\}_{1\leq j\leq J}$. We recall that all the elements of $\mathcal{V}$ (respectively $\mathcal{W}$) have the same length $|\mathbf{v}|$ (respectively $|\mathbf{w}|$). We hence consider motifs of the form

$$\mathbf{m} = [\mathcal{V}(d_1 : d_2)\mathcal{W}].$$

Typically, the sets $\mathcal{V}$ and $\mathcal{W}$ will be respectively of the form $\mathcal{N}_e(\mathbf{v})$ and $\mathcal{N}_e(\mathbf{w})$.

Calculation of $\mu(\mathbf{m})$. $\mu(\mathbf{m}) = \Pr\{\mathbf{m} \text{ at } 1\}$ can be calculated decomposing the event $\{\mathbf{m} \text{ at } 1\}$ as $\bigcup_i \bigcup_j \big\{(X_{\mathbf{v}_i} = 1) \cap (Y_{A(\mathbf{v}_i),\mathbf{w}_j}(\mathcal{W}) \leq d_2 - d_1)\big\}$, therefore we get

$$\begin{aligned}
\mu(\mathbf{m}) &= \sum_i \Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = 1\} \sum_j \Pr\{Y_{A(\mathbf{v}_i),\mathbf{w}_j}(\mathcal{W}) \leq d_2 - d_1\} \\
&= \sum_i p_{\mathbf{v}_i}(1) \sum_{a\in\Sigma} \pi^{d_1+1}(v_{i,|\mathbf{v}_i|}, a) \sum_j \sum_{y\leq d_2-d_1} q_{a,\mathbf{w}_j}(y)
\end{aligned} \tag{6}$$

where $v_{i,u}$ is the u-th letter of $\mathbf{v}_i$.

Calculation of $\mu_1(\mathbf{m})$. As in the single word case, we decompose the event $\{(\mathbf{m}$ not at $1) \cap (\mathbf{m}$ at $2)\}$ and we get

$$
\begin{aligned}
&\Pr\{(\mathbf{m} \text{ not at } 1) \cap (\mathbf{m} \text{ at } 2)\} \\
&= \sum_i \Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = 2\} \sum_j \Pr\{Y_{A(\mathbf{v}_i),\mathbf{w}_j}(\mathcal{W}) \leq d_2 - d_1\} \\
&\quad + \sum_i \Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = 1\} \sum_k \Pr\{Y_{\mathbf{v}_i,\mathbf{v}_k} = 1\} \sum_j \Pr\{Y_{A(\mathbf{v}_k),\mathbf{w}_j}(\mathcal{W}) = d_2 - d_1\} \\
&= \sum_i p_{\mathbf{v}_i}(2) \sum_{a\in\Sigma} \pi^{d_1+1}(v_{i,|\mathbf{v}_i|}, a) \sum_j \sum_{y \leq d_2 - d_1} q_{a,\mathbf{w}_j}(y) \\
&\quad + \sum_i p_{\mathbf{v}_i}(1) \sum_k q_{\mathbf{v}_i,\mathbf{v}_k}(1) \sum_{a\in\Sigma} \pi^{d_1+1}(v_{k,|\mathbf{v}_k|}, a) \sum_j q_{a,\mathbf{w}_j}(d_2 - d_1).
\end{aligned}
\tag{7}
$$

Special motifs. In the case where the number of admitted errors e is a global one, meaning that it is simultaneously defined on the words $\mathbf{v}$ and $\mathbf{w}$, the elements of $\mathcal{W}$ that are valid to complete the motif $\mathbf{m}$ depend on the element of $\mathcal{V}$ that occurred in the first box. Hence, the calculation of $\mu(\mathbf{m})$ and $\mu_1(\mathbf{m})$ are modified. For example, if we globally accept at most one error ($e = 1$), two cases have to be considered:

1. if $\mathbf{v}$ occurs exactly in the first box then any element of $\mathcal{W} = \mathcal{N}_1(\mathbf{w})$ is valid to complete $\mathbf{m}$;

2. if $\mathbf{v}$ occurs with one error in the first box then only $\mathbf{w}$ is valid to complete $\mathbf{m}$.

For this case, the formulas of $\mu(\mathbf{m})$ and $\mu_1(\mathbf{m})$ are given in appendix A.2.

For an arbitrary global number of admitted errors e, $\mathcal{W} = \mathcal{N}_1(\mathbf{w})$ has to be replaced by $\mathcal{W} = \mathcal{N}_e(\mathbf{w})$ in case 1; Case 2 becomes: if $\mathbf{v}$ occurs with exactly $e' \leq e$ errors then only the elements of $\mathcal{N}_{e-e'}(\mathbf{w})$ are valid to complete $\mathbf{m}$.

3.5 Number of occurrences in a set of sequences.

In the application, we consider a set of sequences $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \cdots\}$ and we are interested in the number N of sequences of $\mathbf{S}$ that contain the motif $\mathbf{m}$. Since $\gamma(\ell_s)$ is the probability for the motif $\mathbf{m}$ to occur in a sequence of length ℓ_s, N is distributed like a sum of $|\mathcal{S}|$ independent Bernoulli variables of respective parameters $\gamma(\ell_s)$:

$$
N \sim \sum_{s=1}^{|\mathcal{S}|} \mathcal{B}[1, \gamma(\ell_s)]. \tag{8}
$$

The approximation of order k gives $N \approx \sum_{s=1}^{|\mathcal{S}|} \mathcal{B}[1, \gamma_k(\ell_s)]$ where $\gamma_k(\ell_s)$ is given by equation (4).

If all the sequences have the same length ℓ, then N has a binomial distribution: $\mathcal{B}[|\mathcal{S}|, \gamma(\ell)]$.

4 Application

4.1 Data sets

The purpose of the application presented now is not to exhibit new results but to evaluate the statistical approach introduced in this paper. This is done on data that is clean in the sense that we know the motifs it should contain.

The data consist in two sets of non coding sequences coming just upstream from genes in two well-studied bacterial organisms, *Escherichia coli* and *Bacillus subtilis*. The sets were obtained from Ozoline *et al.* (1998) and Helmann (1995). The transcriptional starting point of the genes have been experimentally determined, or, more rarely, the promoter itself. The sequences in both sets are therefore aligned (on the start of transcription) which enables an easy judgment "by eye" of the biological pertinence of the motifs found. The *E. coli* dataset contains 441 sequences between 60 and 80 bps; the *B. subtilis* dataset contains 131 of length 100bps (except 1 of 99 bps).

The algorithm described in section 2 provides a list of candidate promoters for each datasets. For a quorum $q = 4\%$ and a total maximal number of error $e = 1$ on the two boxes, 40 motifs are selected in the *E. coli* dataset and 564 in the *B. subtilis* dataset. We then have to determine their statistical significance, that is the probability that these structured motifs occur in at least 4% of the sequences.

4.2 Calculations and simulations

For each candidate motif, we calculate according to (8) the approximate expected number of sequences in the set containing the motif

$$\mathbb{E}_k(N) \simeq \sum_{s=1}^{|\mathcal{S}|} \gamma_k(\ell_s)$$

for $k = 0, 1$. The corresponding p-value is calculated in the same way to select relevant motifs. For example, considering the motif $\mathbf{m} = \texttt{TTGACAC}(16:18)\texttt{TATAATA}$, we get $\gamma_1(100) = 1.4885 \times 10^{-4}$. The expected number of sequences containing $\mathbf{m}$ is $131 \times \gamma_1(100) = 1.95 \times 10^{-2}$. This motif appears actually in 6 sequences, so the corresponding p-value is

$$\Pr\{\mathcal{B}[131, \gamma_1(100)] \geq 6\} = 7.67 \times 10^{-14}.$$

On the other hand, the expected numbers can be estimated using simulations: 2000 simulations give a satisfying precision (see the confidence intervals in Figure 1). However, the estimation of the p-values requires a huge number of simulations for an acceptable precision.

Fig. 1

The simulations are performed using a first order Markov model. The parameters $\pi(a, b)$ are estimated on the whole set of sequences, and the $|\mathcal{S}|$ sequences are then simulated with respective lengths $(\ell_1, \ldots \ell_{|\mathcal{S}|})$.

4.3 Comparison

Figure 1 shows that the approximation of order 0 is biased for some motifs. For example, in the *E. coli* dataset, the motif `TTTTTT`(5 : 7)`TTAATT` is expected to occur 2.75 times among the 441 sequences according to the order 0 approximation, while simulations give a value of 1.94. The same observation can be made in relation to the *B. subtilis* dataset. A careful analysis shows that all these motifs have a first box $\mathcal{V}$ with high probability of self-overlaps.

The approximation of order 1 corrects this bias and fits very well to the simulated values (see Figure 1). 7.8% of the motifs are outside the 95% confidence interval based on the simulated values.

Computation times. The computations were carried out on a Celeron 1Ghz, with an average time of 5×10^{-3} sec. per motif. Hence, the computation time is about 30 sec. for 5 000 motifs using the approximation of order 1 while simulations require 90 mn for the same problem.

4.4 Results

Table 1 gives the motifs with a p-value smaller than 10^{-16} using the approximation of order 1. This table is a practical result of our method. Applying the same threshold on the p-value with intervals (16 : 18), (17 : 19), (18 : 20) and (19 : 21), we get a list of motifs given in Table 2 that seems to agree with the known consensus for the site at -35 and the `TATA`-box (see Record *et al.* (1996)). It has to be pointed out that motifs obtained with different intervals may be redundant. P-values presented in Table 2 have been calculated separately for each structured motifs. No multiple testing correction has been applied.

Tab. 1, 2

Software availability. These formulas were implemented in a C program (for unix systems), which is available upon request at the following address :

`richard@genopole.cnrs.fr`.

Given parameters (*i.e.* transition matrix and lengths of sequences) and a list of motifs (with their respective number of occurrences N) this program computes the expectation and the p-value of order 1 for each of them.

A Formulas for probability calculation

A.1 Recursive formulas

The distribution of the positions $X_{\mathbf{v}_i}(\mathcal{V})$ and the distances $Y_{\mathbf{v}_i,\mathbf{w}_j}(\mathcal{W})$ can be deduced from the results of Robin and Daudin (2001). These distributions are characterized by the matrix containing the generating functions of all the r.v.'s $X_{\mathbf{v}_i}$ and $Y_{\mathbf{v}_i,\mathbf{w}_j}$ for all $\mathbf{v}_i$ and $\mathbf{w}_j$. Unfortunately, the calculation of this matrix requires intensive computation in the case of numerous or long words. We propose here an alternative method based on recurrence formulas. We recall two notations and define a third one:

$$\begin{array}{rcl} p_{\mathbf{v}_i}(x) & = & \Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = x\}, \\ q_{\mathbf{v}_i,\mathbf{w}_j}(y) & = & \Pr\{Y_{\mathbf{v}_i,\mathbf{w}_j}(\mathcal{W}) = y\}, \end{array} \qquad r_{\mathbf{v}_i,\mathbf{w}_j}(y) = \Pr\left\{ \begin{array}{c} \mathbf{w}_j \text{ occurs } y \\ \text{positions after } \mathbf{v}_i \end{array} \right\}.$$

The difference between $q_{\mathbf{v}_i,\mathbf{w}_j}(y)$ and $r_{\mathbf{v}_i,\mathbf{w}_j}(y)$ is that, in $r_{\mathbf{v}_i,\mathbf{w}_j}(y)$, the occurrence of $\mathbf{w}_j$ at y is not necessarily the first one after $\mathbf{v}_i$. Remember that the position of a word is given by the position of its first letter.

The probabilities $r_{\mathbf{v},\mathbf{w}}(y)$'s are given for any words $(\mathbf{v}, \mathbf{w})$ by Lemma 2 of Robin and Daudin (2001):

$$\begin{aligned} r_{\mathbf{v},\mathbf{w}}(y) = \varepsilon_{\mathbf{v},\mathbf{w}}(|\mathbf{v}| - y) & \prod_{u=|\mathbf{v}|-y+1}^{|\mathbf{w}|} \pi(w_{u-1}, w_u)\mathbb{I}\{y < |\mathbf{v}|\} \\ & + \pi^{(y-|\mathbf{v}|+1)}(v_{|\mathbf{v}|}, w_1)\frac{\mu(\mathbf{w})}{\mu(w_1)}\mathbb{I}\{y \geq |\mathbf{v}|\} \end{aligned} \tag{9}$$

where the indicator function $\mathbb{I}\{C\}$ equals 1 if C is true and 0 otherwise, and $\varepsilon_{\mathbf{v},\mathbf{w}}(u)$ is the indicator that the first u letters of $\mathbf{w}$ are identical to the last u letters of $\mathbf{v}$: $\varepsilon_{\mathbf{v},\mathbf{w}}(u) = \mathbb{I}\{(w_1 \dots w_u) = (v_{|\mathbf{v}|-u+1} \dots v_{|\mathbf{v}|})\}$. We use the convention $\prod_a^b = 1$ if $a > b$.

Recursive formulas for the positions. The probabilities $p_{\mathbf{v}_i}(x)$'s can be calculated according to Theorem 1 of Robin and Daudin (1999) in the single word case. The generating functions corresponding to the multiple word case are given in Robin and Daudin (2001), but not the recursive formula. We give it here:

$$p_{\mathbf{v}_i}(x) = \mu(\mathbf{v}_i) - \sum_{k=1}^{I}\sum_{z=1}^{x-1} p_{\mathbf{v}_k}(z) r_{\mathbf{v}_k,\mathbf{v}_i}(x - z). \tag{10}$$

The probabilities $p_{\mathbf{v}_i}(x)$'s can therefore be calculated thanks to the probabilities $r_{\mathbf{v}_i,\mathbf{w}_j}(y)$'s using (9). Note that, in any case, $p_{\mathbf{v}_i}(1) = \mu(\mathbf{v}_i)$.

Recursive formulas for the distances. The probabilities $q_{\mathbf{v}_i,\mathbf{w}_j}(y)$'s are given by Lemma 1 of Robin and Daudin (2001) only in the case where $\mathcal{V} = \mathcal{W}$. We generalize here this result to the case of different lists:

$$q_{\mathbf{v}_i,\mathbf{w}_j}(y) = r_{\mathbf{v}_i,\mathbf{w}_j}(y) - \sum_{k=1}^{J}\sum_{z=1}^{y-1} q_{\mathbf{v}_i,\mathbf{w}_k}(z) r_{\mathbf{w}_k,\mathbf{w}_j}(y - z). \tag{11}$$

The probabilities $q_{\mathbf{v}_i,\mathbf{w}_j}(y)$'s can be calculated according to the following procedure:

1. calculate the $r_{\mathbf{v}_i,\mathbf{w}_j}(y)$'s and $r_{\mathbf{w}_j,\mathbf{w}_k}(y)$'s using (9),

2. calculate recursively the $q_{\mathbf{w}_j,\mathbf{w}_k}(y)$'s according to (11),

3. calculate recursively the $q_{\mathbf{v}_i,\mathbf{w}_j}(y)$'s according to (11).

The order of the recurrence formulas involved in this procedure is proportional to the distance y. Thus the computation time of $q_{\mathbf{v}_i,\mathbf{w}_j}(y)$ is in $O(y^2)$. In the problem we study here, the distances y are always very short ($y \leq 30$) so the computation time remains very reasonable.

A.2 Formulas for the probabilities μ and μ_1 for $e = 1$

We consider the case described at the end of section 3.4 where at most one error on both $\mathbf{v}$ and $\mathbf{w}$ is accepted. In this case $\mathcal{V} = \mathcal{N}_1(\mathbf{v})$ and $\mathcal{W} = \mathcal{N}_1(\mathbf{w})$. Both sets $\mathcal{V}$ and $\mathcal{W}$ are reordered in such a way that $\mathbf{v}_1 = \mathbf{v}$ and $\mathbf{w}_1 = \mathbf{w}$, so $\mathcal{V} \setminus \mathbf{v} = \{\mathbf{v}_2, \dots \mathbf{v}_I\}$ and $\mathcal{W} \setminus \mathbf{w} = \{\mathbf{w}_2, \dots \mathbf{v}_J\}$.

In the following, the term $(\mathcal{V})$ will be dropped for the event $X_{\mathbf{v}_i}(\mathcal{V}) = 1$ because $\Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = 1\}$ simply equals $\mu(\mathbf{v}_i)$ that does not depend on $\mathcal{V}$.

Calculation of $\mu(\mathbf{m})$. The element of $\mathcal{W}$ needed to complete $\mathbf{m}$ depends on the element of $\mathcal{V}$ the motif $\mathbf{m}$ starts with. Equation (6) has to be decomposed according to the element of $\mathcal{V}$ that occurs in the first box and we get

$$\begin{aligned}
\mu(\mathbf{m}) &= \Pr\left\{(X_{\mathbf{v}_1} = 1) \cap \left[\bigcup_{j\geq 1} \{Y_{A(\mathbf{v}_1),\mathbf{w}_j}(\mathcal{W}) \leq d_2 - d_1\}\right]\right\} \\
&\quad + \Pr\left\{\bigcup_{i>1} (X_{\mathbf{v}_i} = 1) \cap (Y_{A(\mathbf{v}_i),\mathbf{w}_1} \leq d_2 - d_1)\right\} \\
&= \Pr\{X_{\mathbf{v}_1} = 1\} \sum_{j\geq 1} \Pr\{Y_{A(\mathbf{v}_1),\mathbf{w}_j}(\mathcal{W}) \leq d_2 - d_1\} \\
&\quad + \sum_{i>1} \Pr\{X_{\mathbf{v}_i} = 1\} \Pr\{(Y_{A(\mathbf{v}_i),\mathbf{w}_1} \leq d_2 - d_1)\}.
\end{aligned}$$

Calculation of $\mu_1(\mathbf{m})$. As for the calculation of $\mu(\mathbf{m})$, each term of (7) has to be decomposed according to the element of $\mathcal{V}$:

$$
\begin{aligned}
&\Pr\{(\mathbf{m}\text{ not at }1)\cap(\mathbf{m}\text{ at }2)\}\\
&=\Pr\{X_{\mathbf{v}_1}=1\}\Pr\{Y_{\mathbf{v}_1,\mathbf{v}_1}=1\}\sum_j \Pr\{Y_{A(\mathbf{v}_1),\mathbf{w}_j}(\mathcal{W})=d_2-d_1\}\\
&\quad+\Pr\{X_{\mathbf{v}_1}=1\}\sum_{i>1}\Pr\{Y_{\mathbf{v}_1,\mathbf{v}_i}=1\}\Pr\{Y_{A(\mathbf{v}_i),\mathbf{w}_1}(\mathcal{W})=d_2-d_1\}\\
&\quad+\sum_{i>1}\Pr\{X_{\mathbf{v}_i}=1\}\sum_{j>1}\Pr\{Y_{\mathbf{v}_i,\mathbf{v}_j}=1\}\Pr\{Y_{A(\mathbf{v}_j),\mathbf{w}_1}=d_2-d_1+1\}\\
&\quad+\sum_{i>1}\Pr\{X_{\mathbf{v}_i}=1\}\Pr\{Y_{\mathbf{v}_i,\mathbf{v}_1}=1\}\\
&\qquad\times\left[\Pr\{Y_{A^-(\mathbf{v}_1),\mathbf{w}_1}(\mathcal{W})=d_2-d_1+1\}+\sum_{j>1}\Pr\{Y_{A(\mathbf{v}_1),\mathbf{w}_j}(\mathcal{W})\le d_2-d_1\}\right]\\
&\quad+\Pr\{X_{\mathbf{v}_1}(\mathcal{V})=2\}\sum_j\Pr\{Y_{A(\mathbf{v}_1),\mathbf{w}_j}(\mathcal{W})\le d_2-d_1\}\\
&\quad+\sum_{i>1}\Pr\{X_{\mathbf{v}_i}(\mathcal{V})=2\}\Pr\{Y_{A(\mathbf{v}_i),\mathbf{w}_1}\le d_2-d_1\},
\end{aligned}
$$

where $A^-(\mathbf{v})$ is the (random) letter coming d_1 positions after the end of $\mathbf{v}$, i.e. the letter that just precedes $A(\mathbf{v})$.

Acknowledgement. The authors thank L. Marsan for providing his programs.

References

Bieganski, P., Riedl, J., Carlis, J. V. and Retzel, E. (1994). Generalized suffix trees for biological sequence data: applications and implementations. In *Proc. of the 27th Hawai Int. Conf. on Systems Sci.*, 35–44. IEEE Computer Society Press.

Blom, G. and Thorburn, D. (1982). How many random digits are required until given sequences are obtained ? *J. Appl. Prob.* **19** 518–531.

Crochemore, M. and Rytter, W. (1994). *Text Algorithms.* Oxford University Press.

Fu, C. J. (1996). Distribution of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica.* **6** 957–974.

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge University Press.

van Helden, J., Rios, A. F. and Collado-Vides, J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucl. Acids Res.* **28** 1808–1818.

Helmann, J. D. (1995). Compilation and analysis of *Bacillus subtilis* α-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucl. Acids Res.* **23** 2351–2360.

Marsan, L. and Sagot, M.-F. (2000a). Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Comp. Biol.* **7** 345–362.

Marsan, L. and Sagot, M.-F. (2000b). Extracting structured motifs using a suffix tree – algorithms and application to promoter consensus identification. In *RECOMB'00. Proceedings of Fourth Annual International Conference on Computational Molecular Biology*, (R. Shamir, S. Miyano, S. Istrail, P. Pevzner, and M. Waterman, ed.), 210–219. ACM Press.

Ozoline, O. N., Deev, A. A. and Arkhipova, M. V. (1998). Non-canonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by *Escherichia coli* RNA polymerase. *Nucl. Acids Res.* **25** 4703–4709.

Record, M. T., Reznikoff, W. S., Craig, M. L., McQuade, K. L. and Schlax, P. J. (1996). *Escherichia coli RNA polymerase σ^{70} promoters, and the kinetics of the steps of transcription initiation.* (F. C. Neidhardt, ed.), volume 1. ASM Press.

Robin, S. and Daudin, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36** 179–193.

Robin, S. and Daudin, J.-J. (2001). Exact distribution of the distances between any occurrences of a set of words. *Ann. Inst. Statist. Math.* **36 (4)** 895–905.

Tables and figures

$\mathbf{v}$	$(d_1 : d_2)$	$\mathbf{w}$	observed N	$\mathbb{E}_1(N)$
GTTGACA	$(16 : 18)$	ATATAAT	7	$2.43\ 10^{-2}$
GTTGACA	$(16 : 18)$	TATAATA	8	$2.23\ 10^{-2}$
TGTTGAC	$(16 : 18)$	TATAATA	10	$2.12\ 10^{-2}$
TTGACA	$(16 : 18)$	TTATAAT	12	$1.53\ 10^{-1}$
TTGACAA	$(16 : 18)$	TACAAT	9	$9.82\ 10^{-2}$
TTGACAA	$(16 : 18)$	TATAATA	10	$5.07\ 10^{-2}$
TTGACAG	$(16 : 18)$	TATAAT	9	$7.12\ 10^{-2}$
TTGACG	$(16 : 18)$	TATAAT	11	$2.01\ 10^{-2}$

Table 1: Highly significant motifs (p-value $< 10^{-16}$) for the *B. subtilis* dataset when the length of the intervals is $(16 : 18)$.

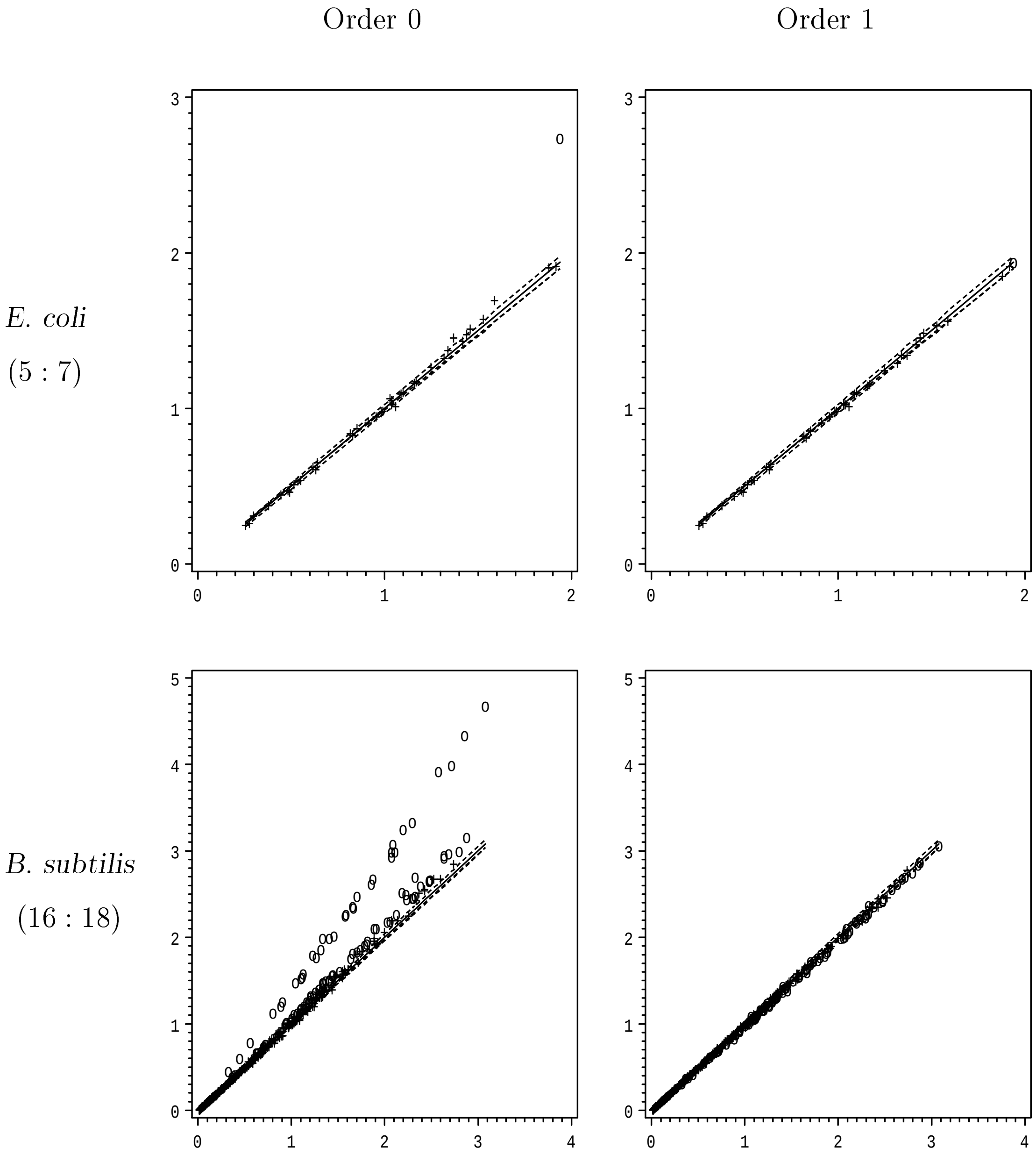

Figure 1: Comparison of approximated (y-axis) and simulated (x-axis) expected numbers of occurrences $\mathbb{E}(N)$ for two sets of structured motifs and order 0 and 1. Motifs for which the approximation is far from the simulation for order 0 are marked with a circle.

$\mathbf{v}$	$(d_1 : d_2)$	$\mathbf{w}$	observed N	$\mathbb{E}_1(N)$
GTTGACA	$(16 : 18)$	ATATAAT	7	2.43×10^{-2}
GTTGACA	$(16 : 18)$	TATAATA	8	2.23×10^{-2}
TGTTGAC	$(16 : 18)$	TATAATA	10	2.12×10^{-2}
TTGACAA	$(16 : 18)$	TACAAT	9	9.82×10^{-2}
TTGACAA	$(16 : 18)$	TATAATA	10	5.07×10^{-2}
TTGACAG	$(16 : 18)$	TATAAT	9	7.12×10^{-2}
TTGACAA	$(17 : 19)$	ATAATAA	9	6.97×10^{-2}
TTGTTGA	$(17 : 19)$	TATAATA	8	5.17×10^{-2}
GTTGACA	$(17 : 19)$	ATAATAA	8	3.09×10^{-2}
GTTGACA	$(17 : 19)$	TATAATA	8	2.19×10^{-2}
CTTGACA	$(17 : 19)$	TATAAT	8	6.04×10^{-2}
TGTTGAC	$(17 : 19)$	TATAATA	12	2.09×10^{-2}
TGTTGAC	$(17 : 19)$	ATATAAT	7	2.29×10^{-2}
TTGTTGA	$(18 : 20)$	TATAATA	8	5.09×10^{-2}
GTTGACA	$(18 : 20)$	ATAATGA	7	1.79×10^{-2}
GTTGTTG	$(18 : 20)$	TATAATA	7	2.53×10^{-2}
TGTTGAC	$(18 : 20)$	ATAATAA	10	2.90×10^{-2}
TGTTGAC	$(18 : 20)$	ATACTA	7	2.77×10^{-2}
TGTTGAC	$(19 : 21)$	ATAATAA	10	2.86×10^{-2}
TGTTGAC	$(19 : 21)$	ATACTA	7	2.73×10^{-2}
TGTTGAC	$(19 : 21)$	TATAAT	10	6.53×10^{-2}
GTTGACT	$(19 : 21)$	ATAATA	8	6.25×10^{-2}

Table 2: Putative alignment of highly significant motifs (p-value $< 10^{-16}$) for the *B. subtilis* dataset when the length of the intervals is successively $(16 : 18)$, $(17 : 19)$, $(18 : 20)$ and $(19 : 21)$