Finding gapped motifs within a Gibbs sampling framework

Gert Thijs[1] Kathleen Marchal[1] Bart De Moor[1] Yves Moreau[1]
[1] ESAT-SISTA/COSIC, KULeuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
Gert.Thijs@esat.kuleuven.ac.be

ABSTRACT
We present an extension of the Gibbs sampling algorithm for motif finding to search for motifs with a gap of variable length. Gapped motifs consist of two smaller conserved parts separated by a variable gap or spacer. These motifs are hard to find with classical motif finding methods that search for motifs of fixed length. To overcome this problem we adapt the Gibbs sampling algorithm. We incorporate the gap together with the alignment position as a missing value and derive the predictive update formula for this model. This formula is used in an iterative sampling procedure to define the actual algorithm. The algorithm is first tested on simulated data sets to give a proof of concept. Finally we show results on real biological data sets where a gapped motif has been described in the literature.
Keywords: Gibbs sampling, motif finding, gapped motifs, cis-acting regulatory elements

1. INTRODUCTION
When searching for over-represented motifs in set of upstream sequences of coregulated genes different methods are available [1, 2, 3, 4, 9, 13, 11, 14, 16, 17]. Most methods are restricted to retrieving ungapped motifs of a fixed length. An important class of regulatory elements are the gapped motifs. Such motifs consists of two smaller, well-conserved parts, separated by a gap or spacer of variable length. Few implementations of algorithms can deal with motifs with a variable gap. There are some methods based on oligo-nucleotide frequency analysis [15, 16]. There also exists some methods based on Gibbs sampling. Rocke et al. [10] have combined Gibbs sampling with a dynamic programming strategy for sequence alignment. Liu et al. [7] have included the possibility of finding gapped motifs in their motif finding algorithm *BioProspector*.

In this paper we describe an extension of our previously described implementation of the Gibbs sampling algorithm, which finds gapped motifs [12, 11]. The gap is incorporated as an additional missing value in the probabilistic sequence model. The start position of the first motif part and length of the gap define unambiguously the start position of the second motif part. Given this observation we adapt the probabilistic sequence model to find the gapped motifs. We develop a Gibbs sampling strategy based on the algorithm described by Liu et al. (1995) to sample simultaneously the starting position and corresponding gap in a sequence conditioned on the current set of alignment positions and corresponding gaps. To speed up the process we simplify the expressions to only use the gap that maximizes the motif probability at a given position. In the remainder of the manuscript we give an elaborate description of the parameters of the model and the derivation of the predictive update formula and we discuss some aspects of the practical implementation of the algorithm.

Further, results on simulated data sets are shown to give a proof of concept of our implementation. To further test the algorithm, we also perform tests on biological data sets which are known to share a common gapped motif. The experiments show that our implementation of a gapped motif sampling algorithm is capable of retrieving this type of motifs. Different examples also show the influence of the parameter settings on the performance of the motif finding algorithm and on the type of retrieved motifs.

2. PROBLEM STATEMENT
Given a set of sequences, S, consisting of N sequences $S_i = \{b(i,j)|j = 1,\ldots,L_i\}$ with $b(i,j)$ the residue at the jth position in sequence S_i, we would like to find the most probable gapped motif in this data set. In this manuscript we define the nucleotides as

$$\begin{aligned}
\mathrm{A} &= [1,0,0,0]^T \\
\mathrm{C} &= [0,1,0,0]^T \\
\mathrm{G} &= [0,0,1,0]^T \\
\mathrm{T} &= [0,0,0,1]^T
\end{aligned}$$

Given a the vectors $q = [q_1, q_2, q_3, q_4]^T$ and $v = [v_1, v_2, v_3, v_4]^T$ we define $q^v = q_1^{v_1} q_2^{v_1} q_3^{v_1} q_4^{v_1}$.

2.1 Motif representation
As in the original Gibbs sampling algorithm for motif finding [4, 6], the motif is described by a product multinomial model and can be represented with a position probability matrix

θ:

$$\text{Motif } \theta = \begin{pmatrix} q_{1,\mathrm{A}} & q_{2,\mathrm{A}} & \cdots & q_{W,\mathrm{A}} \\ q_{1,\mathrm{C}} & q_{2,\mathrm{C}} & \cdots & q_{W,\mathrm{C}} \\ q_{1,\mathrm{G}} & q_{2,\mathrm{G}} & \cdots & q_{W,\mathrm{G}} \\ q_{1,\mathrm{T}} & q_{2,\mathrm{T}} & \cdots & q_{W,\mathrm{T}} \end{pmatrix},$$

where W is the fixed length of the motif. Each entry in this matrix, $q_{i,b}$, gives the probability of finding the nucleotide b at position i in the motif. Each column, $Q_j = [q_{j,\mathrm{A}}, q_{j,\mathrm{C}}, q_{j,\mathrm{G}}, q_{j,\mathrm{T}}]^T$, in this matrix corresponds to the parameters of a product multinomial distribution. As a prior for the parameters of θ we define a product Dirichlet distribution, $\mathrm{PD}(\alpha)$, with $\alpha = (\alpha_1, \ldots, \alpha_W)$ and $\alpha_j = [\alpha_{j,A}; \alpha_{j,C}; \alpha_{j,G}; \alpha_{j,T}]^T$. This leads to

$$\begin{aligned} f(\theta) &= \mathrm{PDir}(\alpha) \\ &= \prod_{j=1}^{W} \mathrm{Dir}(\alpha_j) \\ &= C \prod_{j=1}^{W} (Q_j)^{\alpha_j} (1 - \sum_{b=A}^{T} q_{i,b})^{\alpha_{W+1}}. \end{aligned} \tag{1}$$

A gapped motif is defined as consisting of two shorter motif parts, θ_1 and θ_2, separated by a variable gap. To find the parameters of the motif models θ_1 and θ_2 we use a Gibbs sampling strategy as proposed by Liu et al. [5, 6]. For the simplicity of notation and without loss of generality, we assume that both motif parts are of the same length W.

2.2 Background representation

In the remainder of the manuscript we assume that the background model is fixed. Assuming a single nucleotide background model, $B_0 = [q_{0,\mathrm{A}}, q_{0,\mathrm{C}}, q_{0,\mathrm{G}}, q_{0,\mathrm{T}}]^T$, the probability that the sequence S_i is generated by this background model can be computed as

$$P_{\mathrm{bg}}(S_i) = P(S_i|B_0) = \prod_{j=1}^{L} \theta_0^{b(i,j)}, \tag{2}$$

where $b(i,j)$ is the nucleotide at position j in sequence S_i. For a higher-order background model of order m this background probability becomes

$$\begin{aligned} P_{\mathrm{bg}}(S_i) &= P(S_i|B_m) \\ &= P(b(i,1)\ldots b(i,m)|B_m) \prod_{j=m+1}^{L} P(b(i,j)|S_i, B_m) \end{aligned}$$

where $P(b(i,j)|S_i, B_m$ is the probability of finding the base $b(i,j)$ at position j in the sequence S_i according to the sequence context and the background model [11].

2.3 Introducing the gap as missing value

To extend the sequence model, we introduce the gaps as a set of extra missing values together with the positions of each motif occurrence. If the length of the two conserved parts is kept fixed, knowing the starting position of these two parts defines the gap between the two parts within a particular sequence. We can reformulate this by saying that knowing the position of the first part and the length of the gap defines the starting position of the second part. Translating this formulation to the probabilistic sequence model defines a new set of missing values $[A, G]$, with $A = \{(i, a_i)|i = 1, \ldots, N\}$ the start positions of the motif and $G = \{(i, g_i)|k = 1, \ldots, N\}$ the length of the gap in each sequence.

Treating both the alignment vector and the gap length as missing values and assuming that the distribution of the nucleotides within the gap resembles the single nucleotide frequency, the likelihood for one sequence S_i with one copy of the motif is written as

$$\begin{aligned} \pi(S_i, a_i, g_i|\theta_1, \theta_2, B_m) \propto\ & \frac{P_{\mathrm{bg}}(S_i)}{P_{\mathrm{bg}}(S_i(a_i \ldots a_i + 2W + g_i - 1))} \\ & \times \prod_{j=1}^{W} \theta_1^{b(i,a_i+j-1)} \prod_{j=1}^{g_i} \theta_0^{b(i,a_i+W+j-1)} \\ & \times \prod_{j=1}^{W} \theta_2^{b(i,a_i+W+g_i+j-1)} \end{aligned} \tag{3}$$

where $S_i(a_i \ldots a_i + 2W + g_i - 1)$ is the segment starting at position a_i in sequence S_i, which covers the complete motif (two motif parts and the gap). The complete-data likelihood of the full sequence set is given by

$$\pi(S, A, G|\theta_1, \theta_2, B_m) = \prod_{i=1}^{N} \pi(S_i, a_i, g_i|\theta_1, \theta_2, B_m).$$

To further extend this expression we define $A(j) = \{(i, a_i + j - 1)|i = 1, \ldots, N\}$ to denote all the jth positions of the segments that start at the positions a_i. Using Equation 3 and the defined indices we derive the following formula:

$$\begin{aligned} & \pi(S, A, G|\theta_1, \theta_2, B_m) \propto \\ & \underbrace{\prod_{i=1}^{N} \Big[\frac{P_{bg}(S_i)}{P_{bg}(S_i(a_i, a_i + 2W + g_i - 1))} \Big]}_{\mathcal{P}_{\mathrm{bg}}} \times \\ & \prod_{i=1}^{N} \Big[\prod_{j=1}^{W} \theta_1^{b(i,a_i+j-1)} \prod_{j=1}^{g_i} \theta_0^{b(i,a_i+W+j-1)} \prod_{j=1}^{W} \theta_2^{b(i,a_i+W+g_i+j-1)} \Big]. \end{aligned}$$

Let us define two counting functions, $\mathcal{H}_{\mathrm{c}}$ and $\mathcal{H}_{\mathrm{gap}}$,

$$\begin{aligned} \mathcal{H}_{\mathrm{c}}(A(j)) &= \sum_{i=1}^{N} b(i, a_i + j - 1) \\ \mathcal{H}_{\mathrm{gap}} &= \sum_{i=1}^{N} \sum_{j=1}^{g_i} b(i, a_i + W + j - 1) \end{aligned}$$

to denote the total number of nucleotides at a given position (resp. in the motif parts and the gap). Given these functions the previous equation simplifies to

$$\begin{aligned} & \pi(S, A, G|\theta_1, \theta_2, B_m) \propto \\ & \qquad \mathcal{P}_{\mathrm{bg}} \prod_{j=1}^{W} \theta_1^{\mathcal{H}_c(A(j))}\, \theta_0^{\mathcal{H}_{\mathrm{gap}}} \prod_{j=1}^{W} \theta_2^{\mathcal{H}_c(A(W+g_i+j))}. \end{aligned} \tag{4}$$

3. ALGORITHM AND IMPLEMENTATION

3.1 Predictive update formula

In the previous section we defined the model parameters and the priors of the model parameters as product Dirichlet distributions. We also introduced an expression for the complete-data likelihood of the model parameters, the missing values and the sequence data. Following the rationale of Liu et al. [6], this results in a predictive update formula. By using the predefined model descriptions, Bayes' theorem and integrating out the θ_1 and θ_2 we obtain a predictive distribution of the missing values A and G:

$$\begin{aligned}\pi([A,G]|S) &\propto \pi(S,[A,G]) \\ &= \int\!\!\int \pi(S,[A,G]|\theta_1,\theta 2)\, f_1(\theta_1)\, f_2(\theta_2)\, d\theta_1 d\theta_2 \\ &\propto \mathcal{P}_{\mathrm{bg}} \prod_{j=1}^{W} \Gamma(\mathcal{C}_1(j))(\theta_0)^{\mathcal{H}_{\mathrm{gap}}} \prod_{j=1}^{W} \Gamma(\mathcal{C}_2(j)), \quad (5)\end{aligned}$$

where $\mathcal{C}_1(j)$ and $\mathcal{C}_2(j)$ are defined as

$$\begin{aligned}\mathcal{C}_1(j) &= \mathcal{H}_c(A(j)) + \alpha_j \\ \mathcal{C}_2(j) &= \mathcal{H}_c(A(W+g_i+j)) + \alpha_j,\end{aligned}$$

with a_j the coefficients of the product Dirichlet distribution defined in Equation 1. If $[A,G]_{\bar{k}}$ denotes the set of alignment positions and gaps in all sequences excluding those of sequence S_k, then we can write the predictive formula of the Gibbs sampler as

$$\begin{aligned}\pi([a_k,g_k]|[A,G]_{[\bar{k}]},S) &= \frac{\pi([a_k,g_k],[A,G]_{\bar{k}}|S)}{\pi([A,G]_{\bar{k}}|S)} \\ &= \frac{\pi([A,G]|S)}{\pi([A,G]_{\bar{k}}|S_{\bar{k}})}. \quad (6)\end{aligned}$$

By using Equation 5 and by applying a similar approximation as proposed by Liu et al. [6], this Equation 6 can be rewritten as

$$\pi([a_k,g_k]|[A,G]_{\bar{k}},S) \propto \frac{\prod_{j=1}^{W} \hat{\theta}_1^{b(k,a_k+j-1)} \theta_0^{\mathcal{H}_{\mathrm{gap}}} \prod_{j=1}^{W} \hat{\theta}_2^{b(k,a_k+W+g_k+j-1)}}{P_{bg}(S_k(a_k, a_k+g_k+2W+1))}, \quad (7)$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the motif models calculated from the current set of alignment positions for the first and second part of the motif. This is the probability of a segment being generated by the gapped motif model divided by the same segment being generated by the background model.

Figure 1 gives an example of how the predictive update distribution $\pi([a_k,g_k]|[A,G]_{\bar{k}},S)$, with $a_k = 1,\ldots,L_k$ and $g_k = \text{min\_gap},\ldots,\text{max\_gap}$, changes from initialization to convergence for the first sequence in the simulated data set (see Section 4.1). The gap ranges from min_gap $= 5$ till max_gap $= 15$. The distribution at the initialization is generated with a random motif model for both parts that resembles the single nucleotide frequency. After convergence, the distribution has evolved to one large pronounced peak where the motif is found.

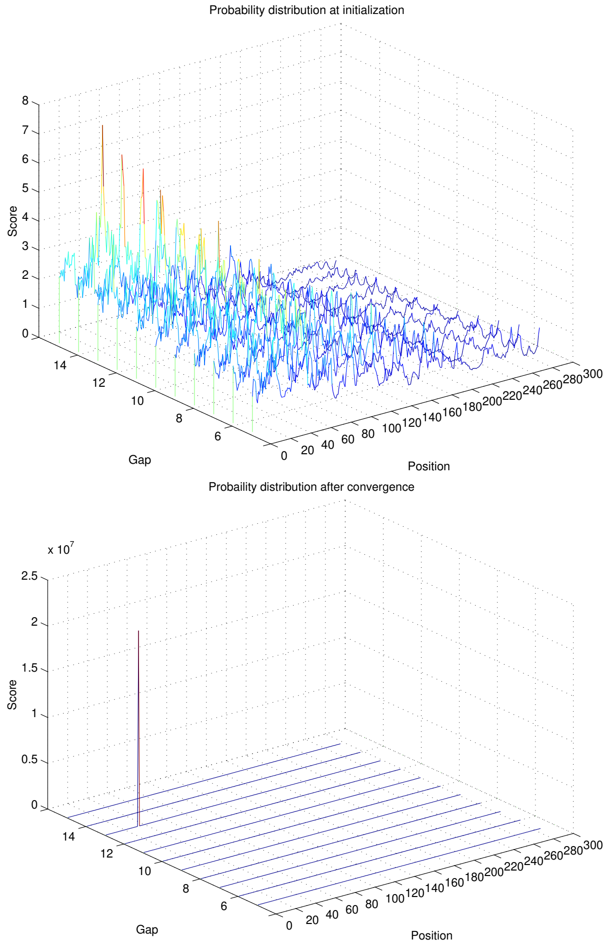

Figure 1: Comparison of the probability distribution of the motif starting at a given position (x-axis) and the corresponding gap (y-axis) at initialization (top) and after convergence (bottom).

3.2 Algorithm

Given the previously derived Equation 7 we could simultaneously sample the starting position of the motif and the gap from a 2-dimensional discrete distribution as shown in Figure 1. Instead of sampling directly from this 2-dimensional distribution, this distribution is mapped to a 1-dimensional distribution from which subsequently samples are drawn. The mapping consists of selecting at each position a_i the gap g_i that maximizes the probability calculated with Equation

7. Using the proposed approximation results in a distribution which is easier to sample from. Consequently, in each step of the iteration the gap that maximizes the motif probability is selected.

Finally, integrating all of the previously described steps results in the following algorithm:

1. Select or compute the background model B_m.
2. Compute the background probability for all segments x that fall within the gap range.
3. Initialize the alignment vector $A = \{a_k | k = 1 \dots N\}$ and the gap $G = \{g_k | k = 1 \dots N\}$, both from a flat prior.
4. For each sequence S_z, $z = 1, \dots, N$
   (a) Create subsets $S_{\bar{z}} = \{S_i | i \neq z\}$ and $[A_{\bar{z}}, G_{\bar{z}}] = \{[a_i, g_i] | i = 1, \dots, N; i \neq z\}$.
   (b) Calculate θ_1 and θ_2 based on $S_{\bar{z}}$ and $[A_{\bar{z}}, G_{\bar{z}}]$.
   (c) Assign to each segment defined by $[a_i, g_i]$, with $a_i = 1, \dots, L_z$, and $g_i = \text{min\_gap}, \dots, \text{max\_gap}\}$, the weight $w(a_i, g_i) = \pi([a_i, g_i] | [A, G]_{\bar{i}}, S)$.
   (d) Create distribution W_z with each $W_z(a_i)$ the maximum of $\{w(a_i, g_i) | g_i = \text{min\_gap}, \dots, \text{max\_gap}\}$.
   (e) Sample new position a_z from normalized probability distribution W_z. The corresponding gap g_z is known from the mapping in the previous step.
5. Repeat from 4 until convergence is reached.

In this algorithm there are a few parameters that are user defined:

1. The length W of the two motif parts, which might be different for both parts.
2. The range of the gap which goes from min_gap to max_gap.
3. The order and type of the background model B_m.

4. RESULTS

4.1 Simulated data set

As a first test case, a gapped motif is inserted in a set of simulated sequences. The simulated sequences are created according to the 4th-order background model of *Arabidopsis thaliana* [11]. While the first part of the inserted motif has GCTCG as consensus, the second part has AGCGA as consensus. Adding a limited amount of noise to the models allows to mimic slightly more degenerate motif parts. The range of the gap is set between 9 and 13. The number of occurrence c of the motif in a sequence is limited to 0, 1 or 2 and was sampled according to the distribution $p(c = 0) = 0.1, p(c = 1) = 0.7$ and $p(c = 2) = 0.2$. An alignment plot of the simulated sequences and the inserted boxes is shown in Figure 2. Sequence 4 and 9 contain two copies of the motif, while there is no copy inserted in sequence 3.This set of simulated sequence is intended to give a proof of concept and to test the performance of our implementation.

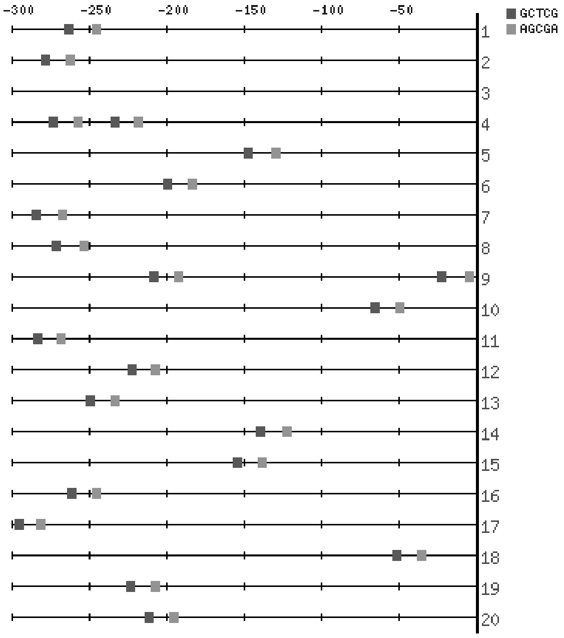

Figure 2: Alignment plot of the inserted gapped motif GCTCG$\cdots$AGCGA in the simulated sequences.

In a first set of experiments we search for a gapped motif where both parts have a fixed length of 5bp and the gap is allowed to range from 5 to 20bp. We repeat this experiment several times. In each run 3 different motifs are searched for. The inserted gapped motif is a rather strong motif. This is reflected by the fact that it pops up as the first motif found in most repeated runs. The other motifs found are less stable and are not considered for further analysis. Table 1 gives an overview of all the detected instances of the retrieved motif that resembles the true inserted motif. This table shows that the algorithm has retrieved all the inserted instance of the gapped motif. However, there are a few sequences where additional instances are found that are not intentionally inserted there (those are indicated with a star in the table). A special case occurs in sequence 20, where a shifted version of the second motif is selected. Apparently, the segment AGAGA matches the second motif part better than the true inserted site, ACCGA.

To further explore the influence of the parameters we run this data set on our original ungapped motif sampler [12, 11]. When looking for a motif with a length of 5 or 6 bp, the two inserted parts are easily found. The fact that both parts are easily picked separately with the ungapped motif sampler, reflects the strong degree of conservation of the both parts. This explains why the gapped motif sampler picked up the inserted motif even with a large range of the gap.

Let us now create another data set where a less conserved motif is inserted. In this case we use the same method to generate the sequences but we insert two shorter motifs, ACTT and GAGT. The range of the gap is set between 6 and 9bp. Figure 3 gives the alignment plot of the inserted boxes.

Nbr.	Pos.	Box 1	gap	Box 2
1	34	GCTCG	tcaaattaataca	AGCGA
2	19	GCTCG	atttaggttca	AGCGA
*3	137	ACTCG	atatacgtttaaa	AGCGA
4	24	GCTCG	atttattatta	AGCGA
4	64	GCTCG	ccacaaaatg	AGCGA
5	150	GCTCG	gcttaatataata	AGCGA
6	98	GCTCG	gataatccaag	AGCGA
7	13	GCTCG	atgtcgtctata	AGCGA
8	26	GCTCG	ctctcccacgcga	AGCGA
9	89	GCTCG	tttaaactggc	AGCGA
9	275	GCTAG	gttccagagataa	AGCGA
10	232	GCTCG	tatggctgtct	AGCGA
11	14	GCTCG	ggttttgagt	AGCGA
12	75	GCTCG	aagtaaaaac	AGCGA
13	48	GCTCA	aagctcttcta	AGCGA
14	158	GCTCG	agaggaatactg	AGAGA
*14	23	GCTCA	agaacgtt	AGCGT
15	143	GCTCG	agtgttgtaga	AGCGA
16	36	GCTCG	ctatccacgag	AGCGA
17	2	GCACG	ttctgtccc	AGCGA
18	246	GCTCG	aaagaagtagt	AGCGA
19	74	GCTCG	actaatttctt	AGCGA
*20	86	GCTCG	agatcccacgt*ACCG*	AGAGA
Consensus:		GCTCG		AGCGA

Table 1: Instances of the gapped motif GCTCG···AGCGA found in the simulated data set.

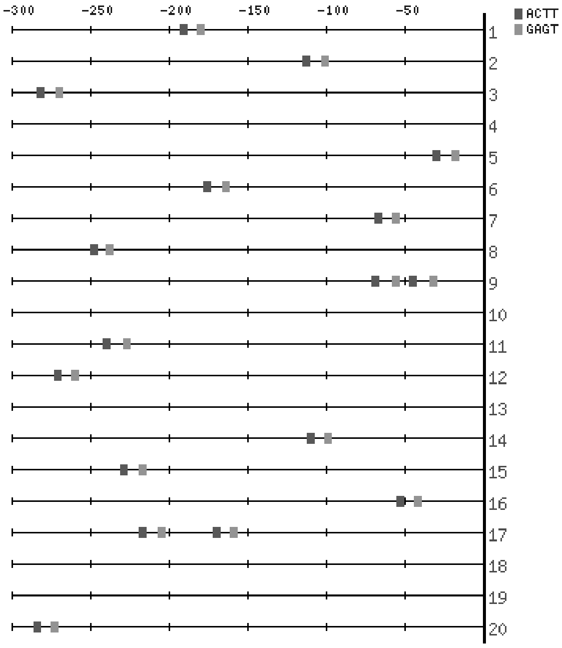

Figure 3: Alignment plot of the inserted gapped motif ACTT···GAGT in the simulated sequences.

If we perform a search on this new set of simulated sequences with the ungapped motif sampler, it is much harder to find both inserted motif parts. Even if we search for motifs with a length of 4bp, the true inserted motif parts are only found occasionally.

Next, we test if it is possible to retrieve the gapped motif. We start this experiment with the same parameter settings as in the previous experiment. In this case the algorithm is unable to retrieve the inserted motif. In the next set of experiments we start to decrease the gap range. When the gap is set to range between 5 and 10bp, which is one shorter or longer than the range of the inserted gap, and the length of both parts is set to 4bp, the algorithm is able to retrieve the inserted motif. The instances found are described in Table 2, which shows that most instances perfectly resemble the inserted motifs. However, if we compare Table 2 with Figure 3, it shows that there are several instances found that were not inserted (indicated with a star). This can be explained by the fact that the prior chance of finding such a small site is much greater than when searching for a larger site. There are also a few instances that are missed. These instance have typically one or two mismatches when compared to the consensus.

4.2 GAL regulated gene family in yeast

A first set of biological data comes from a family of coregulated genes in *Saccharomyces cerevisiae* as described by van Helden et al. [14]. The GAL-family contains genes that are active when galactose is present in and glucose absent from the culture medium. The upstream region of the genes contains a gapped motif *Gal4p* which has the consensus CGGN$_5$WN$_5$CCG. The data set contains six upstream sequences of 800bp. In their article, van Helden et al. were unable to retrieve the motif described in the literature since it is constituted from two small conserved parts and a large gap relative to the total motif length. Although the gap is described as being of a fixed length, we are also unable to retrieve this motif with the ungapped motif sampler. Therefore, we opt to use the gapped motif sampler to search for this gapped motif.

Since this motif is difficult to find, exhaustively testing different parameter settings allows to identify those parameters for which we are able to retrieve the motif. The most important parameters are the length of the conserved motif parts and the range of the gap. In general, when searching for this type of motifs we encounter two kinds of problems. First, the conserved parts of the motif are rather short and are well hidden within the sequences. Counting all the trimers in the data set shows that CGG occurs 41 times and CCG occurs 38 times on a total of 4788 possible trimers. These numbers indicate that the two motif parts are actually underrepresented. Second, the length of the conserved parts is very short as compared to the gap length. However, choosing the range of the gap too large introduces too much noise in the models, which complicates finding the correct motifs.

First we experiment with the length W of both parts set to 3bp (the length of the conserved parts in the described motif). The range of the gap is varied from a rather large gap to a more narrow gap range. The results show that when the range of the gap is choosen too large (e.g., 5–15bp) relative to the motif length, the algorithm is still able to retrieve a motif that has two perfectly conserved motif parts. However, the two parts of the motif tend to be selected from the set of most frequent trimers. For instance, Table 3 shows 18 occurrences of the site TCA-ATT. This site contains two frequent trimers: TCA has 104 instances in the data set and ATT has 147 instances. When the gap is set too large,

Nbr.	Pos.	Box 1	gap	Box 2
1	107	ACTT	taagaca	GAGT
2	185	ACTT	tctacaac	GAGT
*4	68	ACTT	tctccgtac	GTGT
5	268	ACTT	tatatcga	GAGT
6	122	ACTT	tgcatgaa	GAGT
7	231	ACTT	tttttag	GAGT
8	50	ACTT	aaagtc	GAGT
9	229	ACTT	ggccattta	GAGT
9	253	ACTT	ttaaaacac	GAGT
11	58	ACTT	cacataaat	GAGT
*12	55	ACTC	caattttttt	GAGT
14	188	ACTT	ggcactt	GAGT
15	69	ACTT	cgacacat	GAGT
*15	124	CCTT	tggatg	GAGT
17	128	ACTT	aaatgtg	GAGT
17	81	ACTT	cccatctg	GAGT
*18	232	ACTT	ctgacatg	GTGT
*19	50	CCTC	tgattt	GAGT
20	14	ACTT	tggtgct	GAGT
*20	222	ACTT	tgctcatg	GAGT
Consensus:		ACTT		GAGT

Table 2: Instances of the gapped motif `ACTT···GAGT` found in the simulated data set.

several random conformations of two frequent trimers can be perfectly fitted in the proposed motif model.

Name	ID	pos.	box 1	Gap	box 2
GAL1	ybr020w	543	TCA	aattaacgaatcaa	ATT
GAL1	ybr020w	23	TCA	tatccattctca	ATT
GAL1	ybr020w	620	TCA	gcgaagcgatg	ATT
GAL1	ybr020w	194	TCA	tattacatggc	ATT
GAL2	ylr081w	477	TCA	ctccgaaccgag	ATT
GAL2	ylr081w	672	TCA	gtaattgg	ATT
GAL2	ylr081w	340	TCA	cgtgatctat	ATT
GAL2	ylr081w	215	TCA	aactcca	ATT
GAL7	ybr018c	219	TCA	catgtagggaccga	ATT
GAL7	ybr018c	474	TCA	tacattcttaa	ATT
GAL7	ybr018c	358	TCA	ctcacaact	ATT
GAL80	yml051w	145	TCA	tcgtagtgccca	ATT
MEL1	ybr184w	441	TCA	cttgaact	ATT
MEL1	ybr184w	525	TCA	taatgtataggcat	ATT
MEL1	ybr184w	352	TCA	acgtcagctgca	ATT
GCY1	yor120w	527	TCA	atctagcacgcag	ATT
GCY1	yor120w	777	TCA	ttactaa	ATT
GCY1	yor120w	487	TCA	tagcaaggtc	ATT
Consensus:			TCA		ATT

Table 3: Example of a gapped motif `TCA···ATT` found in the GAL data set with the gap range set between 5 and 15bp.

If we however decrease the range of the gap we will see that the number of possible conformations that fit within the proposed structure decreases and that the true motif will pop up. Table 4 shows the results obtained with a more stringent setting of the gap range. The gap range is in this example restricted to range from 9 to 13 bp. The algorithm is now able to retrieve the known motif. But, no motif instance is selected in the sequence *ybr184w*.

A second possibility is to change the length of the motif. Since the actual core of the two motif parts is only 3bp,

Name	ID	pos.	box 1	Gap	box 2
GAL1	ybr020w	348	CGG	attagaagccg	CCG
GAL1	ybr020w	385	CGG	aagactctcct	CCG
GAL1	ybr020w	367	CGG	gcgacagccct	CCG
GAL2	ylr081w	473	CGG	atcactccgaa	CCG
GAL2	ylr081w	468	CGG	ggcggatcact	CCG
GAL2	ylr081w	385	CGG	cggtctttcgt	CCG
GAL2	ylr081w	404	CGG	agatatctgcg	CCG
GAL2	ylr081w	271	CGG	aaagcttcctt	CCG
GAL7	ybr018c	609	CGG	acaactgttga	CCG
GAL7	ybr018c	522	CGG	agcactgttga	GCG
GAL80	yml051w	629	CGG	cgcactctcgc	CCG
GAL80	yml051w	498	CGG	acgatgtct	CTG
GCY1	yor120w	432	CGG	ggcagactatt	CCG
MEL1	ybr184w	-	-	-	-
Consensus:			CGG		CCG

Table 4: The *Gal4p* motif found in the GAL data set.

widening the motif window will introduce more noise into the motif model and less informative position will be added to the model. This means that the quality of the two retrieved motif parts will decrease. If we take a closer look at Table 4 we can see for instance that adding the first nucleotide of the gap to end of the first motif model will certainly deteriorate the motif model. Again, the experiments show that also in this case the gap range is better chosen close to the true gap range.

4.3 O_2-response in bacterial sequences

As a second example, a data set with intergenic sequences from bacteria is created. The selected data set contains a subset of bacterial genes all shown to be regulated by the bacterial O_2-responsive protein FNR [8]. The genes are selected from several bacterial species: *Azospirillum brasilense*, *Paracoccus denitrificans*, *Rhodobacter sphaeroides*, *Rhodobacter capsulatus*, *Sinorhizobium meliloti* and *Escherichia coli*. The data set contains 10 intergenic sequences with varying length.

We search for a gapped motif with the length of the two motif parts set to 5bp and the gap could vary between 3 and 10. Table 5 gives an overview of the motif retrieved with these parameter settings. The motif corresponds to the motif described in literature [8]. When we searched in the same data set for a motif with a fixed length [12] we also were only able to find the occurrences of the motif that have a fixed gap of 4bp. In this case, it was possible to find the motif with the ungapped motif sampler since the size of the gap is small relative to the complete motif length.

5. DISCUSSION

We have implemented an algorithm to detect gapped motifs in the upstream region of coregulated genes based on the Gibbs sampling algorithm for motif finding. Several experiments on simulated data as well as on biological data show that our implementation is able to retrieve the gapped motifs. However, the examples show that defining the correct set of parameters is a rather difficult task and has a profound impact on the performance of the motif finding algorithm. Setting the range of the gap too large as compared to the length of the conserved motif parts allows for too much vari-

Accn	Gene	pos.	box 1	Gap	box 2
af016223	ccoN	60	TTGAC	gcgg	ATCAA
af016223	ccoN	122	TTGTC	aagagtttg	ATCAA
af016223	ccoN	133	TTGAT	caag	GTCAA
af054871	cytN	255	TTGAC	gtag	ATCAA
pdu34353	ccoN	131	TTGAC	gcag	ATCAA
pdu34353	ccOt	84	GTGTT	ttt	ATCAA
pdu34353	ccOt	210	TTGAC	gcag	ATCAA
af195122	bchE	89	TTGAC	atgc	ATCAA
af016236	dorS	8	TTGAC	gtca	ATCAA
af016236	dorS	25	TTGTC	ccgc	ATTAA
ae000220	narK	229	ATGAT	aaat	ATCAA
ae000220	narK	267	TTGAT	ttac	ATCAA
rlfixnc	fixNc	104	TTGAT	gtag	ATCAA
rlfixnd	fixNd	240	TTGAC	gcag	ATCAA
pdu34353	fnr	36	TTGAC	ccaa	ATCAA
Consensus:			TTGAY		ATCAA

Table 5: Gapped motif found in the FNR data set.

ability. Motifs are then picked up that are merely artifacts of the data set and not the true motifs.

The experiments show that finding a good balance between motif length and gap length is very important. If the gapped motif consists of two strong parts it is much easier to find the gapped motif even when allowing a large range of the gap. However, if the two motif parts get weaker, selecting the correct motif length and defining the corresponding gap range will become crucial.

In the described version of the algorithm we assume a flat non-informative prior distribution for the gap. By looking at the biological data sets it seems that typically the gap has a rather small variability and is only shifted one or two positions to the left or right. Therefore, it might be useful to incorporate this knowledge in the algorithm. This fact was shown in the examples where choosing a smaller gap range improved the performance of the algorithm. We could for instance multiply the predictive update formula (Eq. 7) with a weight that favors the gap that is closer to the mean of the gaps. This weight might be adapted during the iterations when the motif is being formed.

Another difficulty that we encounter is the fact that the sampler might get stuck in a local optimum. This local optimum occurs when the first part of the true motif is found as the second part of a motif or vice versa. This kind of problem can be solved by initializing two new runs of the algorithm. The first run will start with a new gapped motif where the first part of this motif corresponds to the second part of the originally found motif. The second run will use the first part of the old motif as the second part of the new motif. After convergence of the two extra runs, we compare the motif scores and select the motif with the highest score.

Acknowledgments

Gert Thijs is a research assistant of the IWT; Kathleen Marchal is a post-doctoral researcher of the FWO; Yves Moreau is a post-doctoral researcher of the FWO; Prof. Bart De Moor is full time professor at the KULeuven. This work is partially supported by: 1. IWT projects: STWW-980396; 2. Research Council KULeuven: GOA Mefisto-666 (2000-2004); IDO 3e010271(2001-2004); 3. FWO project: G.0115.01; 4. IUAP P4-02 (1997-2001); The scientific responsibility is assumed by its authors.

6. REFERENCES

[1] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51–80, 1995.

[2] J.D. Hughes, P.W. Estep, S. Tavazoie, and G.M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, 296:1205–1214, 2000.

[3] L.J. Jensen and S. Knudsen. Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, 16(4):326–333, 2000.

[4] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subbtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[5] J.S. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Stat. Assoc.*, 89(427):958–966, 1994.

[6] J.S. Liu, A.F. Neuwald, and C.E. Lawrence. Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *J. Am. Stat. Assoc.*, 90(432):1156–1170, 1995.

[7] X. Liu, D.L. Brutlag, and J.S. Liu. Bioprospector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Proc. Pacific Symposium on Biocomputing*, volume 6, pages 127–138, 2001.

[8] Kathleen Marchal. *The O_2 paradox of Azospirillum brasilense under diazotrophic conditions.* PhD thesis, FLTBW, KULeuven, 1999.

[9] L.A. McCue, W. Thompson, C.S. Carmack, M.P. Ryan, J.S. Liu, V. Derbyshire, and C.E. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, 29:774–782, 2001.

[10] E. Rocke and M. Tompa. An algorithm for finding novel gapped motifs in DNA sequences. In *Proc. 2nd Annual International Conference on Computational Biology*, volume 2, pages 228–233, 1998.

[11] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouzé, and Y. Moreau. A higher order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 2001. in press.

[12] G. Thijs, K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouzé, and Y. Moreau. A Gibbs sampling meyhod to detect over-represented motifs in the upstream regions of co-expressed genes. In T. Lengauer, D. Sankoff, S. Istrail, P. Pevzner, and

M.S. Waterman, editors, *Fifth Annual International Conference on Computational Biology*, volume 5, pages 305–312. ACM press, 2001.

[13] M. Tompa. An exact method for finding short motifs in sequences, with apllication to the ribosome binding site problem. In *7th Intl. Conf. Intelligent Systems for Molecular Biology*, volume 7, pages 262–271, Heidelberg, Germany, August 1999.

[14] J. van Helden, B. André, and L. Collado-Vides. Extracting regulatory sites from upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281:827–842, 1998.

[15] J. van Helden, A.F. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, 28(8):1808–1818, 2000.

[16] A. Vanet, L. Marsan, A. Labigne, and M.F. Sagot. Inferring regulatory elements from a whole genome. an analysis of helicobacter pylori $sigma^{80}$ family of promoter signals. *J. Mol. Biol.*, 297(2):335–353, 2000.

[17] C.T. Workman and G.D. Stormo. Ann-spec: a method for discovering transcription binding sites with improved specificity. In *Proc. Pacific Symposium on Biocomputing*, volume 5, pages 464–475, Honolulu, Hawai, 2000.