BIOINFORMATICS

Vol. 17 no. 12 2001
Pages 1113–1122

A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling

Gert Thijs [1,*], Magali Lescot [1], Kathleen Marchal [1], Stephane Rombauts [2], Bart De Moor [1], Pierre Rouzé [3] and Yves Moreau [1]

[1]ESAT-SISTA/COSIC, KULeuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium, [2]Department of Plant Genetics, VIB, UGent, Ledeganckstraat 35, 9000 Gent, Belgium and [3]INRA Associated Laboratory, VIB, UGent, Ledeganckstraat 35, 9000 Gent, Belgium

Received on February 6, 2001; revised on June 4, 2001; accepted on June 6, 2001

ABSTRACT

Motivation: Transcriptome analysis allows detection and clustering of genes that are coexpressed under various biological circumstances. Under the assumption that coregulated genes share *cis*-acting regulatory elements, it is important to investigate the upstream sequences controlling the transcription of these genes. To improve the robustness of the Gibbs sampling algorithm to noisy data sets we propose an extension of this algorithm for motif finding with a higher-order background model.

Results: Simulated data and real biological data sets with well-described regulatory elements are used to test the influence of the different background models on the performance of the motif detection algorithm. We show that the use of a higher-order model considerably enhances the performance of our motif finding algorithm in the presence of noisy data. For *Arabidopsis thaliana*, a reliable background model based on a set of carefully selected intergenic sequences was constructed.

Availability: Our implementation of the Gibbs sampler called the Motif Sampler can be used through a web interface: http://www.esat.kuleuven.ac.be/~thijs/Work/MotifSampler.html.

Contact: gert.thijs@esat.kuleuven.ac.be; yves.moreau@esat.kuleuven.ac.be

INTRODUCTION

Recent high-throughput techniques to monitor gene expression levels constitute an important advance in the identification of coexpressed genes (for a review, see Lockhart and Winzeler, 2000). The commonly accepted assumption that coregulated genes share similarities in their regulatory mechanism has led to a major challenge for the computational biologist: detecting novel regulatory elements (motifs) in such sets of coexpressed genes (Brazma and Vilo, 2000; Bucher, 1999; Chu *et al.*, 1998; DeRisi *et al.*, 1997; Spellman *et al.*, 1998; Wolfsberg *et al.*, 1999; Zhang, 1999). These similarities at transcriptional level imply that the promoter region might contain consensus motifs recognized by the same regulatory proteins. In the upstream regions of such sets of coregulated genes, the common consensus motifs are statistically over-represented as compared to their frequency in a background set (of non-coregulated genes).

Several methods to search for over-represented motifs in the upstream region of a set of coregulated genes have been developed and tested (Ohler and Niemann, 2001). These methods can be divided in two major classes: methods based on word counting (Jensen and Knudsen, 2000; Vanet *et al.*, 2000; van Helden *et al.*, 1998, 2000) and methods based on probabilistic sequence models (Bailey and Elkan, 1995; Hughes *et al.*, 2000; Lawrence *et al.*, 1993; Liu *et al.*, 1995; Neuwald *et al.*, 1995; Roth *et al.*, 1998; Workman and Stormo, 2000). Word counting methods are based on the frequency analysis of oligonucleotides in the upstream sequences. Over-representation is measured by comparing the counted number of occurrences of a word to the expected number of occurrences. A common motif is then compiled by grouping similar words. In the probabilistic methods, the motif model is represented as a position probability matrix and the motif is assumed to be hidden in a noisy background sequence. To find the parameters of such a model, maximum likelihood estimation is used. The most frequent methods to do so are Expectation Maximization (EM) and Gibbs sampling. EM is a maximum likelihood algorithm for estimating the parameters of a probabilistic model. Gibbs sampling is a stochastic equivalent of EM.

The drawback of these algorithms is that they tend to be

*To whom correspondence should be addressed.

sensitive to noise. Noise is due to the presence of upstream sequences in the data set that do not contain the motif. These sources of noise have either an experimental origin or are artefacts of the clustering process and are difficult to avoid.

Another source of noise comes from the large size of the upstream sequences of the selected genes as compared to the small size of the motifs. Parts of the sequence not containing a motif can indeed be considered as noise. This second source of noise obviously depends on the compactness of the genome. For higher eukaryotes, the size of intergenic regions varies considerably between organisms, being much larger on average in humans than in *Arabidopsis thaliana*. Even within the same species the size of the intergenic region can vary (e.g., by at least 2 orders of magnitude for *A.thaliana*, from $<10^2$ to $>10^4$; Pavy *et al.*, 1999). Therefore the influence of the noise can be expected to be reasonably low for bacteria and still limited for lower eukaryotes such as yeast, but more pronounced for higher eukaryotes.

Conceivably, it is important to have a motif detection algorithm that can cope with this noise and discriminate between motifs that are over-represented by chance and motifs that are biologically functional. An improved background model (model of non-coregulated genes) can considerably improve this discrimination.

The most popular probabilistic models published so far (Bailey and Elkan, 1995; Hughes *et al.*, 2000; Lawrence *et al.*, 1993; Liu *et al.*, 1995; Neuwald *et al.*, 1995; Roth *et al.*, 1998) use a simple background model based on the frequency of the nucleotides A, C, G, and T in the data set to represent an intergenic sequence. However, a background model solely based on single nucleotide frequencies poorly reflects the complex structure of genome sequences. A description of DNA sequences as higher-order Markov chains on the other hand has been used in most of the state-of-the-art gene recognition software to represent coding and non-coding regions, (e.g., in Glimmer and the *A.thaliana* specific GlimmerA (Delcher *et al.*, 1999), HMMgene (Krogh, 1997) and GeneMark.hmm (Lukashin and Borodowsky, 1998)). In this paper we describe the extension of the Gibbs sampling algorithm with a complex context-dependent background model.

Recently other researchers have also proposed the use of advanced background models in the Gibbs sampling algorithm (Liu *et al.*, 2001; McCue *et al.*, 2001). Here we like to address the specific issues associated with the background models used in these methods. McCue *et al.* (2001) extended the Gibbs sampler with a position specific background model estimated with a Bayesian segmentation algorithm that was presented by Liu and Lawrence (1999). This model tries to capture the varying GC and AT-content of the different regions in DNA sequences. The model parameters and corresponding change points are found using a Bayesian sequence segmentation algorithm that maximizes the joint likelihood of the data, parameters and missing values (change point positions). These model parameters can be used to calculate the probability that a certain site in the sequence is generated with this background model. Furthermore, Liu and Lawrence have situated the problem of sequence alignment within a Bayesian framework that we also favour. Liu *et al.* (2001) have developed an extended version of the Gibbs sampler called BioProspector. They have proposed the use of a context-depend Markov model to represent the probability that a site is generated by the background model. We will comment on some of the specific technicalities of this method in the next section.

Together with the development of the algorithm a selected data set of intergenic sequences from *A.thaliana* was used to construct a reliable higher-order background model of gene upstream regions of this model plant. The influence of different background models on the robustness of the motif sampler in the presence of noisy data was exhaustively tested. We will describe the construction of the background models and the use of this background model with the Gibbs sampling algorithm. To test the influence of these models, we used simulated data and several well-described data sets of genes from *A.thaliana* for which different motifs are documented.

HIGHER-ORDER BACKGROUND MODEL

As stated in the introduction, most of the state-of-the-art gene detection software uses a context-dependent model based on a higher-order Markov process to represent DNA sequences. Based on the rationale of these algorithms, the use of such a model to detect motifs in the upstream region of coregulated genes seems a logical decision. Using a context-dependent model of order m means that the probability of finding a nucleotide b at position l in a sequence depends on the m previous nucleotides in the sequence. The probability of the sequence being generated by this background model B_m is given by

$$P(S \mid B_m) = P(b_1, \ldots, b_m) \prod_{l=m+1}^{L} P(b_l \mid b_{l-1}, \ldots, b_{l-m}).$$

The probabilities $P(b_l \mid b_{l-1}, \ldots, b_{l-m})$ are stored in a transition matrix and the prior frequency of the oligonucleotides of length m is given by $P(b_1, b_2, \cdots, b_m)$. The construction of the transition matrix of an mth-order background model is based on the counting of all oligonucleotides of length $(m+1)$ in the data set. To compensate for zero occurrences of certain oligonucleotides a pseudocount is added. Based on Bayesian statistics, we assume that the more data are available the more we can rely on these data to approximate the true biological

model. Based on the comparison of experimental results, the pseudocounts are chosen proportional to the single nucleotide frequency and inverse proportional to the square root of the size of the data set.

The background model can be either constructed based on the input sequences or based on an independent data set of intergenic sequences. The latter approach seems the more sensible one to produce a reliable background model. The quality of the background model depends on the quality of the data set. In this paper a set of carefully selected intergenic sequences from *A.thaliana* is used to construct a reliable background model.

Extension of the motif sampler

The implementation of our motif finding algorithm is based on the original Gibbs sampling algorithm previously described by Lawrence *et al.* (1993). An elaborate description of our algorithm is given elsewhere (Thijs *et al.*, 2001). In this paper we like to emphasize the specific aspects of the higher-order model and we will only summarize the aspects of the algorithm to facilitate further understanding.

The calculation of the background model is done as an initialization step of the algorithm. The background model B_m is computed either from the input sequences, making it useful for any organism, or from an independent data set. This model B_m is not updated during the algorithm since there is no need to re-estimate the background model at each iteration step of the algorithm. The background model will be used to calculate for each site x of length W in a sequence the probability that this site was generated by the background model. This probability is referred to as P_x:

$$P_x = P(\text{Site} \mid B_m) = \prod_{l=1}^{W} P(b_l \mid b_{l-1}, \ldots, b_{l-m}).$$

The motif of length W is represented with a position probability matrix θ_W where the entry $q_{i,b}$ contains the probability of finding nucleotide b at position i in the motif:

$$\theta_W = \begin{bmatrix} q_{1,A} & q_{2,A} & \cdots & q_{W,A} \\ q_{1,C} & q_{2,C} & \cdots & q_{W,C} \\ q_{1,G} & q_{2,G} & \cdots & q_{W,G} \\ q_{1,T} & q_{2,T} & \cdots & q_{W,T} \end{bmatrix}.$$

For each site x of length W in a sequence the probability Q_x of site x being generated by the motif model θ_W is calculated:

$$Q_x = P(\text{Site} \mid \theta_W) = \prod_{l=1}^{W} q_l, b_l.$$

Based on these probabilities a weight A_x is then assigned to each segment x in the sequence

$$A_x = \frac{Q_x}{P_x}.$$

Subsequently the alignment vector of the motif in this sequence is sampled according to the distribution of normalized weights A_x. By updating this distribution we can find the alignment that maximizes the ratio of the corresponding site probability to the background probability.

As stated in the introduction, Liu *et al.* (2001) have also proposed the use of a context-dependent Markov model in a Gibbs sampling algorithm. In their algorithm the probability of a site being generated by this background model is computed as

$$\begin{aligned} P(b_l, b_{l+1}, b_{l+2}, \cdots, b_{l+W-1}) &= P(b_l)P(b_{l+1} \mid b_l) \\ &\times P(b_{l+2} \mid b_l, b_{l+1})P(b_{l+3} \mid b_l, b_{l+1}, b_{l+2}) \\ &\cdots P(b_{l+W-1} \mid b_{l+W-2}, b_{l+W-3}, b_{l+W-4}). \end{aligned}$$

This approach resembles the one proposed in this paper but differs in the calculation of the probability of the site being generated by the background model. In our algorithm we take the m preceding bases of the site into account to compute the background probability of the site since the complete sequence information is available at the time of computation. This also means that we only need to compute the parameters of the mth-order background model and not the parameters for all the models from 1 to m.

DATA SETS

Intergenic data set

To construct the best possible representation of promoter sequences or intergenic sequences a data set consisting of carefully selected intergenic sequences was constructed, following a previously described rationale to build Araset (Pavy *et al.*, 1999). To define clean intergenic sequences, all complete cDNAs were downloaded through SRS and aligned on BAC sequences. The aligned genes were manually checked by an expert (S. Aubourg, personal communication). Each time the cDNAs matched two consecutive genes on the BAC, the intergenic sequence was extracted. The sequences with a length below 10 kb were then extensively checked for any unannotated potential coding sequences, using BLAST (Altschul *et al.*, 1990) for homology searches and prediction software such as EuGène (Schiex *et al.*, 2000). 78 intergenic sequences were retained, representing a total of 156 087 bp. These sequences were added to the 94 intergenic sequences retrieved from Araset, resulting in a data set with 341 248 bp. Figure 1

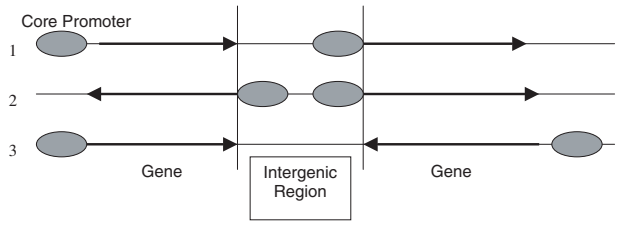

Fig. 1. Representation of the three different configuration of the intergenic region in DNA. The genes are represented with an arrow and the core promoter for each gene is indicated with an oval box. (1) The two genes are pointing in the same direction and there is one core promoter in the intergenic region. (2) Two genes are pointing in opposite directions and the intergenic region contains the two core promoters. In the last case (3) the genes are pointing towards each other and in the intergenic region there is no core promoter present.

shows the three different configurations in which neighboring genes can occur. 105 sequences have the first configuration (1) consisting of two genes coding in tandem on the same strand. In this case the intergenic region is expected to contain only one promoter. 38 sequences have the second configuration (2) where both genes are pointing away from each other. In this intergenic region divergent promoters are expected to control transcription. In the last case (containing 29 sequences) the transcription of the two genes is convergent. No promoter regulatory element is expected to occur in the intergenic region. The transition matrix was only built from the intergenic sequences of the classes (1) and (2), which likely contain either one or two promoters.

Data sets for testing

To test the performance of our implementation we constructed several data sets. The data sets are accessible on the web: http://www.plantgenetics.rug.ac.be/∼males/Datasets/Data.html.

- Simulated data: these sequences were used to test the influence of the background model on the detection of different types of motifs. By sampling according to the 4th-order background model 20 sequences of 500 bp were generated. In several tests, instances of predefined motifs were inserted in these sequences at random positions.
- G-box sequences: this set of sequences was extracted from PlantCARE (Rombauts *et al.*, 1999) and contains the upstream region of genes that are known to be regulated by G-box binding proteins in dicots. The consensus of the G-box is CACGTG. The position of the G-box is well defined in this data set. The set contains 33 sequences of 500 bp. The G-box (CACGTG) is a well-conserved ubiquitous *cis*-acting regulatory element found in plant genomes and is bound by the GBF (G-box binding factors) family of bZIP proteins (Donald and Cashmore, 1990).
- Light induced: this set contains the upstream region of 28 coexpressed *A.thaliana* genes. Coexpression was based on the cluster analysis of a microarray experiment (Desprez *et al.*, 1998).
- Random: set of randomly selected *A.thaliana* upstream sequences of at least 150 bp, not described to be involved in light regulation and not containing a known G-box.

RESULTS

Construction of an independent background model

The construction of a Markov process can rely either on the upstream sequences from the input data or from an independent data set. This independent data set consists of a well-defined set of intergenic regions of *A.thaliana* genes (see Section **Data sets**). It should be noted that the number of nucleotides used to construct the Markov model limits the order of the background model that can be used. Indeed, when a transition matrix of order m is constructed, all oligonucleotides of length $(m + 1)$ are counted. The number of possible different oligonucleotides equals 4^{m+1} and increases exponentially with m. The data set used for the construction of the background model should, under the assumption of an equal nucleotide distribution, at least contain 4^{m+1} different base pairs to have a single count for each nucleotide. In reality the assumption of equal nucleotide distribution does not hold and a much larger data set will be needed. When an oligonucleotide does not occur in the data set, it will be replaced by a pseudocount. When the order of the background is too high relative to the size of the data set on which this background model was based, less frequent motifs will be encountered which deteriorates the motif model. Following this reasoning the improvement of using a Markov chain background model will be more explicit when its construction is based on a large data set (such as the one used in this study). A thorough study of the intergenic data set shows that when all hexamers are counted, there are 430 388 examples in this data set. The hexamer with highest number of occurrences, 2018, was AAAAAA, while there was no instance at all found of the hexamer GCGGGC. The consequence of this observation is that the 5th-order background model is less reliable, as will be shown in the tests.

Simulated sequences

As a first set of tests, simulated sequences were used. The sequences were generated according to the 4th-order background model. Although randomly generated

Table 1. Results of the motif finding in the simulated sequences with all four motifs inserted. Each number corresponds to number of runs, out of 20, in which the corresponding consensus was found and this for the different orders of the background models

	Input sequences				Intergenic sequences					
	3	**2**	**1**	**snf**	**snf**	**1**	**2**	**3**	**4**	**5**
GCTGCAGC	0	18	14	19	19	19	20	20	18	15
TAGAATA	4	12	14	8	9	17	16	20	20	15
GsCCGnnnCGGsC	2	4	7	7	13	11	14	14	18	10
ATAnwCCwnTAA	9	9	5	4	3	9	13	13	14	12

sequences do not fully resemble biological sequences, the use of the 4th-order model ensures the conservation of the pentamer composition of the sequences as compared to the composition of the intergenic data set. This also means that these simulated sequences are AT-rich. Subsequently the different motifs were inserted at random positions throughout the sequences. Motifs were represented with a position probability matrix and each base of a created instance is sampled according to the distribution in the corresponding column of the matrix. Distinct types of motifs can be observed: AT-rich versus GC-rich and well conserved versus more degenerate. Using these simulated data sets we are able to test the influence of the different background models on the rate of detecting the inserted motifs.

In a first set of tests we included all four types of motifs in 20 sequences of 500 bp. Figure 2 gives an overview of the logos of the four created motifs based on the inserted instances. We tried several parameter settings and the best results were obtained when searching for six different motifs with a length of 14 bp that can have 0, 1 or 2 copies. Since our method is probabilistic, each test was repeated 20 times. Table 1 indicates for the different types of the background model how many times the corresponding consensus term was found in the 20 repeated runs. The background model was computed from either the input sequences or the precompiled intergenic models.

The first motif, GCTGCAGC, a well-conserved, GC-rich motif was easily found with each of the background model used, except with the 3rd-order model based on the input sequences. To explain this phenomenon we plotted in Figure 3 the transition matrices of the two 3rd-order background models. Each point in the plot corresponds to the same entry in both transition matrices. The x axis shows the value for the matrix based on the intergenic sequences while the value for the matrix based on the input sequences is depicted on the y axis. For a perfect 3rd-order background model a one-to-one correspondence is expected, however some outliers are clearly visible. For these points the corresponding representation of the entry in the transition matrix is indicated. The points that are largely overestimated by the background model based on the input sequences correspond to the words part of the inserted GC-rich motifs. Therefore, the computed background model based on the input sequences is biased towards this motif and thus it will overestimate P_x for these sites.

The second inserted motif is a frequent, short, well-conserved and AT-rich motif: TAGAATA. When using the single nucleotide background model the motif was only found in 50% of the runs, while it was found in all 20 runs when using the intergenic either the 3rd or 4th-order models. Finding a short AT-rich motif seems to be much harder when using the single nucleotide model. Even when searching for shorter motifs (8 bp) it was picked up less frequently with the single nucleotide model.

The third motif consists of two GC-rich blocks and it has a fixed gap of 3 bp. This motif was retrieved with the same consistency for all background models except when the background models were compiled from the input sequences. In this case there is no difference between the higher order models and the single nucleotide models. We repeated this test also with a data set where only this degenerated GC-rich motif was inserted and there the same consistent performance was observed.

Finally a rather degenerate, AT-rich motif, resembling the background, was inserted. This motif is assumed to be the hardest to find. Using the single nucleotide background model resulted in a fairly poor performance. The motif was detected in only 3 or 4 out of 20 runs. Using higher-order background models, the motif was more frequently found. This effect was even more pronounced when we inserted only this degenerated motif in the random sequences. The number of detections with the single nucleotide model increased to 6 out of 20 runs, but for the 3rd-order model this was 18 out of 20 and even 20 out of 20 runs for the 4th-order model. To further test this difficult case we inserted slightly modified versions of this motif, TTAAwCwATA and wTAATsTATA, in the generated sequences. In these two test cases we only used the intergenic background models. Again a similar performance was measured, as is shown in Table 2. Using the single nucleotide model it is difficult to retrieve the inserted motifs, while using the higher-order models significantly improves the performance.

In this section we described the analysis of several tests with simulated data. Although the simulated data do not fully resemble biological sequences, we can still draw some conclusion from these tests. It was shown in these tests that the use of higher order background model, compiled from an intergenic data set, improves the rate of detection of difficult motifs as compared to the use of a single nucleotide model. The results show that the 3rd and 4th-order models based on the intergenic data

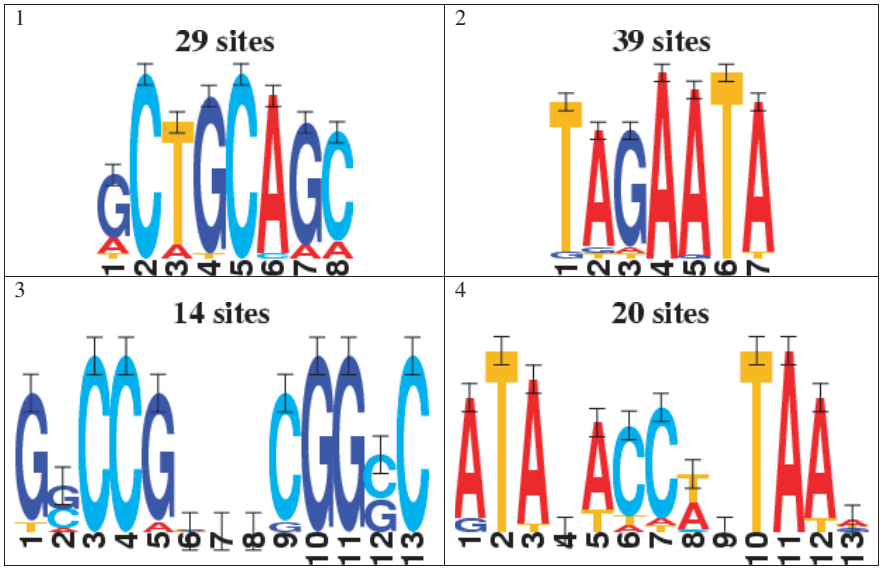

Fig. 2. Logo representation of the four inserted sites in the simulated sequences. These logos are created from all the inserted instances of the motifs in the sequences.

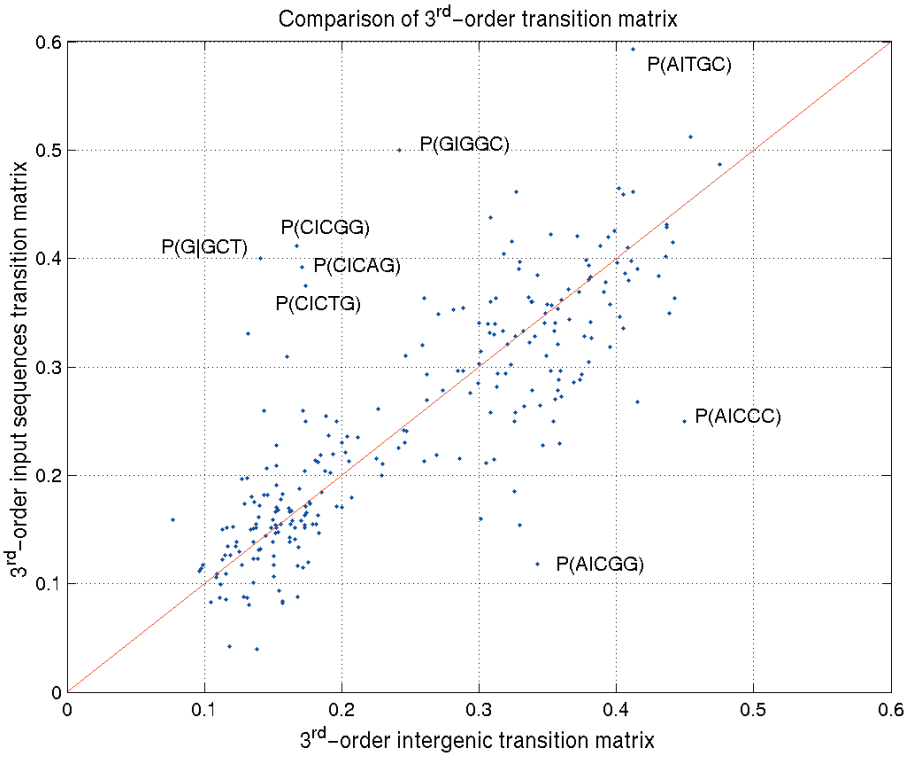

Fig. 3. Comparison of the 3rd-order transition matrices computed from the intergenic sequences and the input sequences. Each point represents a corresponding entry in both transition matrices. The x axis indicates the value in the intergenic matrix while the y axis depicts the value in the matrix based on the input sequences. For some of the outliers the corresponding entry in the matrix is given.

Table 2. Results of the motif finding in the simulated sequences with only the AT-rich motif inserted. Each number corresponds to number of runs, out of 20, in which the correct consensus was found and this for all the background models based on the intergenic data set

		snf	**1**	**2**	**3**	**4**	**5**
TTAAwCwATA	Correct	5	9	14	10	13	5
	Shifted	1	2	3	4	3	2
wTAATsTATA	Correct	0	12	7	12	8	8
	Shifted	1	5	6	4	5	3

set have the most positive influence on the detection of distinct motifs. The effect of the 1st and 2nd-order models is less pronounced. The 5th-order model on the other hand has in some cases a rather poor performance. This is due to the fact that the transition matrix is built by counting hexamers. As was shown in the section about the construction of the background model this count is not reliable enough. The influence of using the intergenic data set instead of using the input sequences is obvious especially for the 3rd-order model. However we should take the necessary care when interpreting the results in Tables 1 and 2. These results show that tests with the 4th-order model have the best results, but this is possibly related to the fact that this background model was used to generate the sequences.

Influence of noisy sequences

To test the influence of the complex background model on the robustness of the motif sampler in the presence of noise, another set of tests was performed. In this set of tests we used a set of biologically relevant sequences upstream of 33 genes (G-box data set, see Section **Data sets**). In subsequent tests, the number of noisy sequences added to the G-box data set was progressively increased (10 at a time). The set of noisy sequences, from which each time 10 sequences were sampled, consisted of a random mixture of the light-induced (Desprez *et al.*, 1998) and random data sets (see Section **Data sets**). Preliminary tests showed that the 3rd-order background model had a better performance than the other higher-order background models on this data set. Therefore in the next set of exhaustive tests only the single nucleotide and the 3rd-order background model were compared. All parameters of the motif sampler algorithm were kept fixed except for the order of the background model (we tried either single nucleotide frequency, 3rd-order Markov model computed from the input data or 3rd-order Markov model computed from the intergenic data set). In each test we searched for 10 different motifs with a length of 8 bp that can have 0 or 1 copy in the sequence and each test was repeated 10 times.

To evaluate the results we checked in which runs the G-box consensus **CACGTG** was detected. Based on this definition of the G-box sequence, we calculated the number of times the G-box was found in each test (group of 10 runs). Figure 4a describes the behaviour of the algorithm in the presence of an increasing number of noisy sequences for different background models. The 3rd-order background model clearly outperforms the single nucleotide background model. With the single nucleotide model, the algorithm only detects the G-box consensus in a small number of runs even in the presence of only a limited number of noisy sequences. Both higher-order background models can find the G-box consensus in the presence of a large number of noisy sequences. To further validate the outcome, the positions of the G-box motifs predicted by the algorithm were compared with the positions of G-box of the documented 33 G-box sequences. Three different possibilities can be distinguished:

(1) the predicted motif is located at the same position as the known G-box motif (true positive);

(2) the algorithm could not detect a motif although the presence of a motif was described (false negative);

(3) a potential G-box motif detected by the algorithm is located at a different position than the described G-box (ambiguous case).

In the last case, the predicted position might represent a yet undetected G-box and is therefore inconclusive. Figure 4a shows the average number of correctly predicted motif positions. The calculation was based for each experiment only on the runs in which a G-box consensus was detected. Figure 4b demonstrates that the number of correctly predicted motifs (true positives) decreases slightly with increasing noise. However, the order of the background model does not interfere drastically with the number of correctly predicted motifs. On average, approximately 70% of the G-boxes were correctly predicted. This indicates that, if a motif is detected, it is in 70% of the cases the right motif, irrespective of the background model. The background model does not improve the performance of the algorithm in making correct predictions. However, the influence of the higher-order background model on the robustness and performance of the algorithm in the presence of noise becomes obvious when taking into account the number of missed true positives. Figure 4c depicts the average number of sequences in which the algorithm could not predict the right G-box motif (false negatives). This number of false negatives consists of all the sequences in a run in which the algorithm could not detect a G-box consensus although a G-box was described in the sequence. Figure 4c shows that the more noise is added to the data set, the higher the percentage

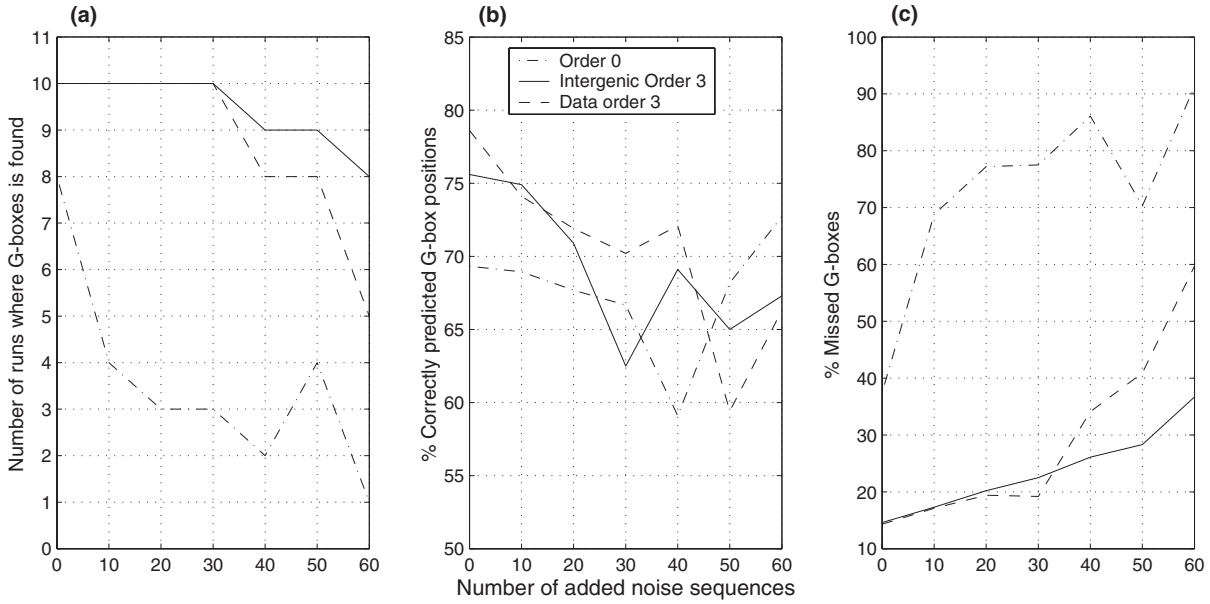

Fig. 4. (a) Total number of times the G-box consensus is found in 10 runs. The horizontal axis shows the number of noisy sequences added to the G-box data set. (b) Average number of correctly predicted G-box positions. This number is based on comparison of the described G-box positions and the predicted positions of the G-box motif in all the runs where a G-box consensus was found. (c) Average percentage of wrongly classified motifs. This number is based the number of sequences that are indicated as not having a G-box although a G-box was documented (including the runs where no G-box consensus is found).

of missed G-boxes. Moreover, this effect is considerably more pronounced for the single nucleotide background model than for the 3rd-order background model. The 3rd-order model based on the set of intergenic sequences performs better than the 3rd-order model based on the input data.

We usually observed during the tests that when using a 3rd-order background model, the algorithm retrieved the G-box consensus as one of the first motifs, while this was not the case for the single nucleotide model. The rapid convergence of the algorithm to the G-box indicates that it is a very stable motif in the presence of a 3rd-order model. This was further corroborated by the fact the G-box motif was in these cases also the motif with the highest log-likelihood score.

DISCUSSION

We aimed at improving the performance of a probabilistic implementation of a motif finding algorithm in the presence of noisy data. To this end, the existing algorithm was extended with a more complex background model. We anticipated that the description of the background sequences as single nucleotide frequencies was not sufficient to capture the complex information in the inherently non-random sequence code. Therefore we used higher-order Markov models to represent the intergenic sequences in DNA. We adapted the original Gibbs sampling algorithm in such a way that we can incorporate the higher-order background model to update the probabilities of finding a motif at a certain position in the sequence. A set of carefully selected intergenic regions was used to construct a higher-order background model for *A.thaliana*. As was shown in the Section **Results** the quality and the size of this intergenic data set determine the reliability of the order of the model. We tested our implementation on different simulated data sets, where the sequences were generated from the 4th-order background model and different types of motifs were inserted. These tests have shown that the use of a higher order background model, especially 3rd or 4th-order, can improve the performance of the algorithm. We also showed with the simulated data that when using an unreliable higher-order model (e.g., the 3rd-order model based on the input sequences) the performance decreases significantly.

The behaviour of the algorithm in the presence of an increasing amount of noisy data has extensively been tested. The use of a 3rd-order model was shown to be considerably more robust than a single nucleotide background model. The overall recovery of the motifs was higher in the presence of a higher-order model, though the number of correctly predicted motifs was only marginally affected by the complexity of the background model.

Future work will concentrate on the improvement of the *A.thaliana* background model through extending the intergenic data set and also by using interpolated Markov chains to augment the significance of the transition matrix. Focus will be on the automatic selection of the best background model. We will also compile background models of other organisms.

ACKNOWLEDGEMENTS

Gert Thijs is research assistant with the IWT; Yves Moreau is a post-doctoral researcher of the FWO; Professor Bart De Moor is a full time professor at the KULeuven; Pierre Rouzé is Research Director of INRA (Institut National de la Recherche Agronomique, France). This work is partially supported by: 1. IWT project: STWW-980396; 2. Research Council KULeuven: GOA Mefisto-666; 3. FWO projects: G.0240.99 and G.0256.97; 4. IUAP P4-02 (1997–2001); 5. Industrial Contract Research: Data4s. The scientific responsibility is assumed by its authors.

REFERENCES

Altschul,S., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using Expectation Maximization. *Mach. Learn.*, **21**, 51–80.

Brazma,A. and Vilo,J. (2000) Gene expression data analysis. *FEBS Lett.*, **480**, 17–24.

Bucher,P. (1999) Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.*, **9**, 400–407.

Chu,S., DeRisi,J., Eisen,M.B., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

Delcher,A.L., Harman,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with Glimmer. *Nucleic Acids Res.*, **27**, 4636–4641.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Desprez,T., Amselem,J., Caboche,M. and Hofte,H. (1998) Differential gene expression in *Arabidopsis* monitored using cDNA arrays. *Plant J.*, **14**, 643–652.

Donald,R.G.K. and Cashmore,A.R. (1990) Mutation of either G-box or I-box sequences profoundly affects expression from the *Arabidopsis rbcS*-1A promoter. *EMBO J.*, **9**, 1717–1726.

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

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

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

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

Krogh,A. (1997) Two methods for improving performance of an HMM and their application for gene finding. *ISMB*, **5**, 179–186.

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

Liu,J.S. and Lawrence,C.E. (1999) Bayesian inference on biopolymer models. *Bioinformatics*, **15**, 38–52.

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

Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **6**, 127–138.

Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

Lukashin,A.V. and Borodowsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.

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

Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, **4**, 1618–1632.

Ohler,U. and Niemann,H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.*, **17**, 56–60.

Pavy,N., Rombauts,S., Déhais,P., Mathé,C., Ramana,D.V.V., Leroy,P. and Rouzé,P. (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics*, **15**, 887–899.

Rombauts,S., Déhais,P., Van Montagu,M. and Rouzé,P. (1999) PlantCARE, a plant *cis*-acting regulatory element database. *Nucleic Acids Res.*, **27**, 295–296.

Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole genome mRNA quantitation. *Nat. Biotech.*, **16**, 939–945.

Schiex,T., Moisan,A., Duret,L. and Rouzé,P. (2000) EuGene: a eucaryotic gene finder that combines several sources of evidence. In *Proc. JOBIM'2000*. http://www.inra.fr/bia/T/schiex/Export/EuGene2.pdf

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridisation. *Mol. Biol. Cell.*, **9**, 3273–3297.

Thijs,G., Marchal,K., Lescot,M., Rombauts,S., De Moor,B., Rouzé,P. and Moreau,Y. (2001) A Gibbs sampling method to detect over-represented motifs in upstream regions of co-expressed genes. *RECOMB*, **5**, 305–312.

Vanet,A., Marsan,L., Labigne,A. and Sagot,M.F. (2000) Inferring

regulatory elements from a whole genome. An analysis of *Helicobacter pylori* sigma(80) family of promoter signals. *J. Mol. Biol.*, **297**, 335–353.

Wolfsberg,T.G., Gabrielian,A.E., Campbell,M.J., Cho,R.J., Spouge,J.L. and Landsman,D. (1999) Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*. *Genome Res.*, **9**, 775–792.

Workman,C.T. and Stormo,G.D. (2000) ANN-SPEC: a method for discovering transcription binding sites with improved specificity. *Pac. Symp. Biocomput.*, **5**, 467–478.

Zhang,M.Q. (1999) Large-scale gene expression data analysis: a new challenge to computational biologist. *Genome Res.*, **9**, 681–688.