BIOINFORMATICS

Vol. 19 Suppl. 2 2003, pages ii5–ii14
DOI: 10.1093/bioinformatics/btg1052

Computational detection of cis*-regulatory modules*

Stein Aerts[1], Peter Van Loo, Gert Thijs, Yves Moreau and Bart De Moor*

[1]*Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, Leuven, 3001, Belgium*

Received on March 17, 2003; accepted on June 9, 2003

ABSTRACT

Motivation: The transcriptional regulation of a metazoan gene depends on the cooperative action of multiple transcription factors that bind to *cis*-regulatory modules (CRMs) located in the neighborhood of the gene. By integrating multiple signals, CRMs confer an organism specific spatial and temporal rate of transcription.

Results: Based on the hypothesis that genes that are needed in exactly the same conditions might share similar regulatory switches, we have developed a novel methodology to find CRMs in a set of coexpressed or coregulated genes. The ModuleSearcher algorithm finds for a given gene set the best scoring combination of transcription factor binding sites within a sequence window using an A^* procedure for tree searching. To keep the level of noise low, we use DNA sequences that are most likely to contain functional *cis*-regulatory information, namely conserved regions between human and mouse orthologous genes. The ModuleScanner performs genomic searches with a predicted CRM or with a user-defined CRM known from the literature to find possible target genes. The validity of a set of putative targets is checked using Gene Ontology annotations. We demonstrate the use and effectiveness of the ModuleSearcher and ModuleScanner algorithms and test their specificity and sensitivity on semi-artificial data. Next, we search for a module in a cluster of gene expression profiles of human cell cycle genes.

Availability: The ModuleSearcher is available as a web service within the TOUCAN workbench for regulatory sequence analysis, which can be downloaded from http://www.esat.kuleuven.ac.be/~dna/BioI.

Contact: stein.aerts@esat.kuleuven.ac.be

INTRODUCTION

Cis-regulatory modules (CRMs), either proximal to genes (called promoters) or distal (called enhancers) control complex genetic programs, such as bilaterian development (Davidson, 2001). The role of CRMs in governing the genetic developmental program can hardly be overestimated, given that, while highly diverse in developmental patterns, bilaterians share the same basic set of developmental genes—a fact that has become clear by comparative genomics and by rescue experiments on knock-outs with orthologous genes. Working with combinations of factors makes it possible to integrate multiple inputs and this further provides cross-coupling of signal transduction and gene regulatory pathways. This way, a CRM functions as an information processing device (Yuh *et al.*, 1998).

The availability of several sequenced and annotated genomes and specialized alignment algorithms designed to identify functional noncoding segments (e.g., AVID (Bray *et al.*, 2003) among others) allow for the delineation of putative regions containing CRMs in large intergenic sequences (Berman *et al.*, 2002; Aerts *et al.*, 2003). It is an effective way of reducing the search space of possible binding sites, thereby reducing the number of false positives while the associated increase of true negatives (true binding sites located outside syntenic regions cannot be detected) is limited. Yuh *et al.* (2002) obtained a success rate of 65% of syntenic regions between two sea urchins that are functional in the cis-regulation of *otx*.

As thousands of genes are activated during development it is expected that at least some genes might share one or more CRMs. Detecting DNA motifs by their statistical over-representation in a set of sequences (Thijs *et al.*, 2002; van Helden *et al.*, 1998) or detecting over-represented hits of known TFBSs (Aerts *et al.*, 2003) have been used with various degrees of success. Exploiting colocalization to find true binding sites in a particular gene yields valuable hypotheses regarding transcriptional regulation, for example in combination with sequence conservation across species (Loots *et al.*, 2002; Jegga *et al.*, 2002), and particularly for known combinations of factors (Berman *et al.*, 2002; Krivan and Wasserman, 2001; Wasserman and Fickett, 1998; Halfon *et al.*, 2002; Rebeiz *et al.*, 2002) or for multiple instances of one factor (Markstein *et al.*, 2002).

*To whom correspondence should be addressed.

Here we present a novel approach for finding combinations of TFBSs that occur several times across multiple coregulated human genes. We specifically search within syntenic regions with respective mouse orthologous genes since these have a high chance of containing real CRMs (i.e., functional evolutionary conservation). We apply a score function that combines slightly adjusted log likelihood ratios (using higher-order background models) of individual position-specific frequency matrices (PSFMs) from TRANSFAC (Wingender *et al.*, 2000). Here, attention is paid to the sensitivity and specificity of the PSFM scoring. Obviously an efficient algorithm is needed to search the enormous state space of possible combinations of binding sites (e.g., if we have 400 factors then there are $400^5/5! = 8.10^{10}$ possibilities for a CRM with 5 binding sites). The ModuleSearcher algorithm implements the score function in an A^* tree search. We show the results of the ModuleSearcher obtained on four artificial data sets and explore the sensitivity and specificity of the algorithm. Using the rare examples of CRMs in the literature, we hereby justify the methodology and the different thresholds and parameters used along the road, when applying the ModuleSearcher on real biological data. For the latter we have chosen a coherent cluster of gene expression profiles, as captured by a microarray study on the cell cycle in a human cancer cell line. The modules we find are proven to contain real regulatory information. To our knowledge, this shows for the first time that module detection in microarray clusters of *human* genes is feasible, when taking all precautions discussed here to reduce the level of noise into account.

The score function alone is used in the ModuleScanner program to detect genes that might be controlled by a certain CRM. We have tested this program using the IFN-β enhancer as a model, and using the predicted CRM of the microarray cluster. Predicted targets are validated *in silico* using Gene Ontology annotation.

DATA AND METHODS

Methodology overview Figure 1 shows a flow chart that overviews the system for detecting regulatory modules (read more below and in the figure caption).

Human-mouse syntenic regions All human-mouse orthologous pairs were selected from Ensembl release 9 (19,914 pairs). 10kb of sequence upstream of the coding sequence of the human and mouse gene were selected (18,778 pairs with successful selection). Each 10kb pair was aligned with AVID (Bray *et al.*, 2003) and the alignment output was parsed using VISTA (Mayor *et al.*, 2000) to select regions with at least 75% identity in windows of 100 bp (10,049 pairs had at least one region; 33,282 regions in total). These regions form the "Syntenic fastA" database (Fig. 1). All syntenic regions were scanned to predict transcription factor binding sites (TFBSs) using the MotifScanner algorithm (prior parameter set to 0.2, see below). Frequency matrices were taken from TRANSFAC Professional release 6.3, which contained 429 vertebrate matrices. All occurrences are stored in GFF format in the 'Syntenic GFF' database that is both used for the selection of annotated regions of coregulated genes (to find CRMs) and for 'genomic searches' to find genes containing a given CRM. In the current version we have limited the intergenic sequence space to 10kb upstream of the coding sequence, but extensions towards syntenic regions located in introns or downstream of the gene are possible.

Semi-artificial sequence sets A 3rd-order Markov model was calculated from the Syntenic fastA database (Thijs *et al.*, 2001), representing the base pair composition of conserved regions. Artificial sequences were generated by sampling symbols from this background model. Transcription factor binding sites were implanted at random locations by sampling a TFBS from position-specific frequency matrices. To reflect a more realistic biological situation, we added artificial sequences without implanted binding sites that represent false positive sequences (a real set of sequences that *all* contain the same CRM can probably never be found and sequence sets could consist of multiple classes of regulons each containing another CRM). The second column of Table 1 describes the contents of the four constructed test sets. In Art_3 multiple of these artificial sequences were implanted themselves into larger sequences. Figure 2F shows 10 such sequences with four implanted CRMs each, separated by Ns. The blanks between the modules illustrate the fact that we will consider only the syntenic regions, not other intergenic DNA.

Sets of coregulated genes Sets of coexpressed genes were selected using SOURCE (Diehn *et al.*, 2003). A typical case of coregulation is the cell cycle and we have queried the SOURCE database for cyclin B2 (CCNB2). In the 'expression view' we have chosen the data set of gene expression during the cell cycle in a human cancer cell line (HeLa) (Whitfield *et al.*, 2002). By searching for genes that have a similar profile, using the functionality provided by the application, we selected 44 genes that might share a common *cis*-regulatory element. Of these, 34 had a Ensembl identifier, and in this set we found 13 genes with at least one syntenic region with the respective mouse orthologous gene (32 regions in total).

Scoring single TFBSs The binding sites of transcription factors have been represented and predicted with matrices for the last two decades (Stormo, 2000). We start from the position-specific frequency matrix $\mathbf{\Theta}$ (PSFM) and a higher-order background model $\mathcal{B}_m$. Combining PSFM

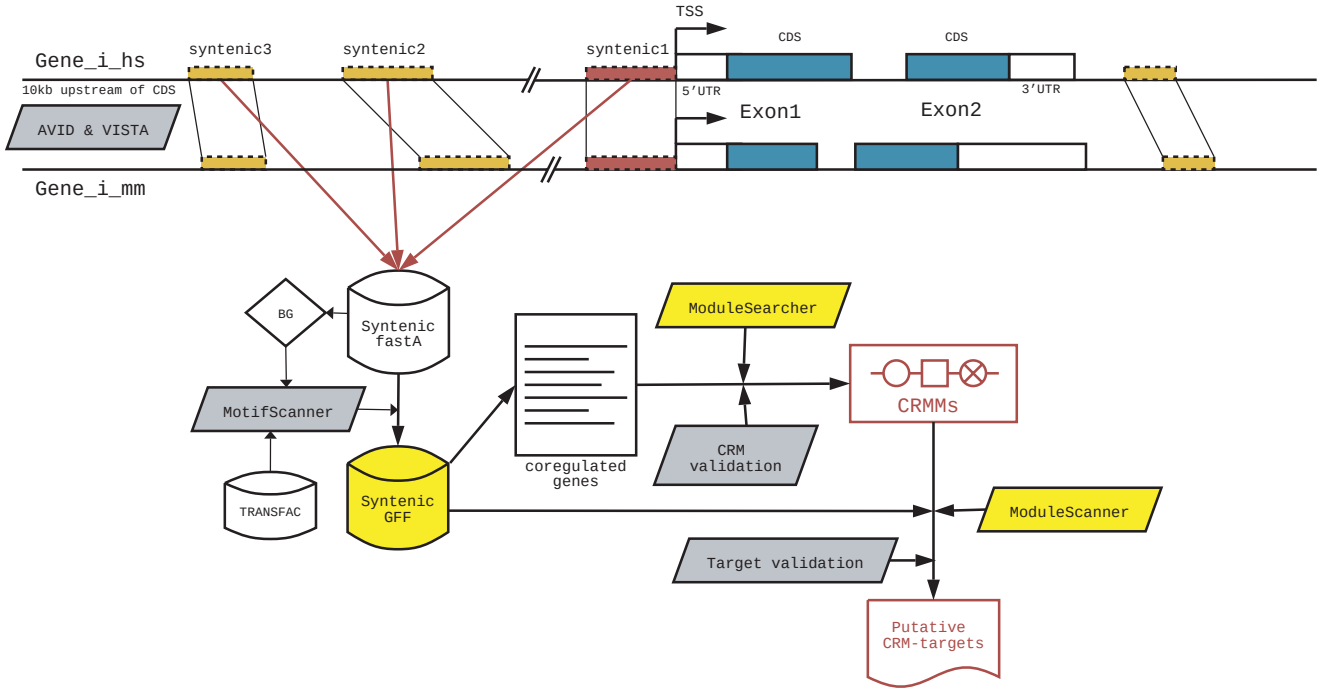

Fig. 1. Overview of the system to detect regulatory modules. All DNA regions, ranging from 100 to several hundreds of base pairs, resulting from global alignment of all human-mouse ortholog pairs are stored, as are the hits of all transcription factors of TRANSFAC, in GFF format. The GFF can be selected for a set of genes, and the ModuleSearcher finds the best module model within the set. Such a model can then be used to find putative targets, using the same GFF database.

and background model, the score $W(\mathbf{x})$ of a segment $\mathbf{x} = [b_1, b_2, \ldots, b_w]$ in a sequence s is computed as

$$W(\mathbf{x}) = \frac{\prod_{j=1}^{w} \Theta(b_j, j)}{\prod_{j=1}^{w} P(b_j|s, \mathcal{B}_m)},$$

where b_j is the nucleotide found at position j in the segment $\mathbf{x}$, $\Theta(b_j, j)$ is the probability of finding b_j at position j according to the PSFM and $P(b_j|s, \mathcal{B}_m)$ is the probability of finding b_j in the sequence according to the background model. This formula indicates how likely it is that the segment is generated by the motif model with respect to the background. The use of higher-order background models have been described extensively in (Thijs *et al.*, 2001).

These scores can be used directly, as in a PWM scoring scheme (Stormo, 2000), by computing the logarithm of $W(\mathbf{x})$ and rescaling the scores to a value between 0 and 1. By defining a threshold, we retain all segments with a score greater than this threshold. The resulting program is called MotifLocator. The second program, MotifScanner, uses a probabilistic sequence model to estimate the number of instances c of a motif model that are hidden in a noisy background sequence (Aerts *et al.*, 2003). If the estimated number of instances is c, the c sites with the highest score $W(\mathbf{x})$ in the sequence are selected.

Matrix similarity Motif models are redundant at two levels: (1) there can be multiple matrices describing the binding site of the same TF and (2) there can be distinct TFs with similar PSFMs. Consequently there is a limit on the sensitivity to distinguish some models computationally. The similarity between two motif models, Θ_1 and Θ_2, is measured with the Kullback-Leiber distance (Kullback, 1959), which is computed as

$$\max_{\mathbf{A}} \frac{1}{w} \sum_{j=1}^{w} \sum_{b=A}^{T} \Theta_1(j, b) \log \frac{\Theta_1(j, b)}{\Theta_2(j, b)}$$

where $\Theta_1(j, b)$ is the probability of finding base b at position j in Motif 1, w is the length of the motif, and $\mathbf{A}$ is the set of all possible alignments for an allowed shift (e.g., 2 base pairs). Since this equation is asymmetric, we take the average between the distance from Θ_1 to Θ_2 and from Θ_2 to Θ_1. The motif models can be grouped into classes depending on an imposed threshold on this distance. We have used threshold values of 0.2 (high stringency), 0.3 (moderate stringency), and 0.4 (low stringency) to construct classes of motif models.

Module score function Analogous with the distinction between a binding site and a motif model (a frequency

matrix is a motif model), we distinguish CRMs and CRM models. CRMs are clusters of actual binding sites on a sequence, and CRM models are sets of motif models. The score of a CRM model m on a set of sequences $\mathbf{s} = (s_1, \ldots, s_n)$ is calculated as

$$\mathcal{S}_m(\mathbf{s}) = \sum_{i=1}^{n} \mathcal{S}_m(s_i).$$

The score of a CRM model m on one sequence s is calculated as

$$\mathcal{S}_m(s) = \max_{k_{1,s},\ldots,k_{l,s}} p(\mathbf{t})b(\mathbf{t}) \times \sum_{i=1}^{l} \log \mathcal{S}(t_{i,s}^{k_{i,s}}).$$

The different elements of this formula are the following. Each *cis*-regulatory module model m is a collection of motif models $\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_l$. The set of matching binding sites is $\mathbf{t} = (t_{1,s}^{k_{1,s}}, \ldots, t_{l,s}^{k_{l,s}})$, where $t_{i,s}^{k}$ is the kth instance of $\mathbf{\Theta}_i$ on sequence s, $\mathcal{S}(t_{i,s}^{k})$ is the score of one TFBS. $b(\mathbf{t})$ is a boolean function expressing whether the given combination of TFBSs is classified as a valid CRM or not. This function is determined by two parameters: (1) overlap between different TFBSs can be allowed or not and (2) the sites should fall within the specified window length (default = 200 base pairs). The parameters $k_{i,s}, i = 1, \ldots, l$, represent a count over the occurring TFBSs of model $\mathbf{\Theta}_i$ in sequence s. If the MotifScanner algorithm returns $q_{i,s}$ sites of model $\mathbf{\Theta}_i$ on sequence s, $k_{i,s}$ can take the values $0, \ldots, q_{i,s}$. A value of $k_{i,s} = 0$ means that no instance of $\mathbf{\Theta}_i$ is found. By definition $\mathcal{S}(t_{i,s}^{0}) = 1, \forall i, \forall s$. Since $\log \mathcal{S}(t_{i,s}^{k})$ can be interpreted as the energy of binding of the kth TFBS of $\mathbf{\Theta}_i$ on sequence s, this definition makes sense ($\log 1 = 0$). The factor $p(\mathbf{t})$ functions as a penalization of CRMs that do not contain an instance of each motif model currently in the module model. It is the number of occurring sites in the module divided by the number of motif models l in the current module model. Penalization of incomplete CRMs can be enabled or disabled, as required by the user. If it is disabled, $p(\mathbf{t}) = 1, \forall s, \forall k$.

The score incorporates distance constraints in the form of a window and does not take the motif order into account. The simple score function presented here was satisfactory for our current goals. However, more complicated score functions based on hidden markov models could be tested in the future, such as COMET (Frith *et al.*, 2002).

The A search algorithm* Our search for the best CRM model on a set of coregulated genes is handled with an A^* procedure, a branch-and-bound search with an estimate of remaining distance to the solution. It is an optimal heuristic graph search algorithm (Hart *et al.*, 1968). In bioinformatics, the A^* algorithm has already been used for multiple sequence alignment (Lermen and Reinert, 2000). Each node in the implicit search tree is a CRM model. Creating child nodes involves adding TFBMs to parent CRM models. Since we do not consider the order of sites in this step, we have removed redundant nodes by allowing only alphabetically ordered CRM models. A function $\mathcal{G}_m = \mathcal{S}_m + \mathcal{H}_{m,n_\Theta}$ is used, where $\mathcal{S}_m$ is the score-function, and $\mathcal{H}_{m,n_\Theta}$ is a heuristic overestimate of the rise in score from CRM model m to the best child CRM model m_b. The algorithm, searching for the maximal score, is shown here:

1. Initialization
   (a) `Queue` contains the root node as only element (the empty CRM model).
   (b) `Solution` is null.
   (c) The parameter n_Θ is set, which is the number of sites a module should contain.
   (d) The parameters of the score function are initialized.
2. While $\mathcal{G}_m(s_1, \ldots, s_n) \geq \mathcal{S}_{\text{Solution}}(\mathbf{s})$, where m is the first CRM model in the `Queue` (or while no `Solution` is found yet), do
   (a) Remove first CRM model m from `Queue`.
   (b) Do for all *valid* models $\mathbf{\Theta}_i$ ($\mathbf{\Theta}_i$ is valid if the CRM model does not contain a $\mathbf{\Theta}$ of the same class, unless multiple copies of the same motif model are allowed, but the latter is only true for exactly the same $\mathbf{\Theta}$):
      i. Create a new CRM model $m_{\text{new},i} = m, \mathbf{\Theta}_i$ (add $\mathbf{\Theta}_i$ to m).
      ii. If the size of $m_{\text{new},i}$ is n_Θ, and if $\mathcal{S}_{m_{\text{new},i}}(\mathbf{s}) > \mathcal{S}_{\text{Solution}}(s_1, \ldots, s_n)$, then `Solution` = $m_{\text{new},i}$.
      iii. If the size of $m_{\text{new},i}$ does not equal n_Θ, add $m_{\text{new},i}$ to `Queue`.
   (c) Sort the `Queue` by descending $\mathcal{G}(\mathbf{s})$, where

$$\mathcal{G}_m(\mathbf{s}) = \mathcal{S}_m(\mathbf{s}) + \mathcal{H}_{m,n_\Theta}(\mathbf{s})$$

with $\mathcal{H}_{m,n_\Theta}(\mathbf{s})$ is a heuristic function that is an *overestimate* of the difference between the score of m and the best CRM model consisting of n_Θ matrices and containing all matrices of m:

$$\mathcal{H}'_{m,n_\Theta}(\mathbf{s}) = \max_{\mathbf{t}} b_m(\mathbf{t}) \sum_{i=l+1}^{e} \mathcal{S}_{[\mathbf{\Theta}_i]}(\mathbf{s}),$$

where l is the length of CRM model m, and $[\mathbf{\Theta}_i]$ is a CRM model containing one

matrix, Θ_i. These $\mathcal{S}_{(t_k)}(s_1, \ldots, s_n), \forall k$ can be calculated before the start of the algorithm. $b_m(t_{k_{l+1}}, \ldots, t_{k_e})$ is a boolean function expressing whether the given combination of motif models, when added to m, constitutes a valid CRM model or not (2b). In case we penalize incomplete CRMs, the heuristic becomes

$$\mathcal{H}_{m,n_\Theta}(\mathbf{s}) = \mathcal{S}'_m(\mathbf{s}) - \mathcal{S}_m(\mathbf{s}) + \mathcal{H}'_{m,n_\Theta}(\mathbf{s}),$$

where S' and $\mathcal{H}'$ are the score function without penalization and the heuristic without penalization, respectively.

3. `Solution` now contains the optimal CRM model.

Gene Ontology statistics GO4G (http://www.esat.kuleuven.ac.be/~saerts/software/go4g.html) works as follows. All annotated GO terms for a set of genes are retrieved from the GOA annotations of the EBI (http://www.ebi.ac.uk/GOA/). For each term, each path to the root of the GO tree is followed and each encountered term is added to a gene's annotation. For each term, the frequency of this term is then the number of genes that have the term in their extended annotation divided by the total number of genes in the gene set. The binomial formula is then used to calculate p values for each frequency, where the expected frequencies are calculated from a large reference set, such as the complete human genome. For the analysis described here we have used the set of human genes that have a mouse ortholog. The p values are then corrected for multiple testing. GO4G can be used for testing the functional coherence of a gene set and is therefore useful for validating predicted target genes.

Availability within Toucan The ModuleSearcher is included in Toucan (Aerts *et al.*, 2003) as a web service (Stein, 2002). Toucan is a Java tool for *cis*-regulatory sequence analysis and phylogenetic footprinting for metazoan genes. It is tightly linked with the Ensembl genome databases (Hubbard *et al.*, 2002) for the retrieval of intergenic sequences, gene annotations, and orthologous sequences. Once a sequence set is created, the MotifScanner (Aerts *et al.*, 2003) can be run to score the sequences with a database of position weight matrices or the MotifSampler (Thijs *et al.*, 2002) can be used to find over-represented motifs using Gibbs sampling. The putative binding sites that result from these actions can be sent to the ModuleSearcher to find the best combination of sites. Toucan can be stArtd from a URL using Java Web Start (http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html).

RESULTS

Semi-artificial sequences

Table 1 lists the results obtained on semi-artificial data (see Data & Methods). Analysis of Art_1 shows that the ModuleSearcher is able to detect a module of 5 elements correctly (all 5 elements are found) when it is hidden in 10 sequences of 200 bp and when another 10 random sequences of the same length are added. The results on the Art_2 set show that the ModuleSearcher can detect 2 distinct modules that are hidden in a set of 15 sequences, although some elements were misidentified: 4 out of 5 elements of Module 1 are correct, and 2 out of 3 elements of Module 2 are correct. Figure 2A–E shows Art_2 when scored with the MotifScanner. It can be seen from this figure that many implanted sites are missed in the scoring step, which causes an important limitation on the sensitivity of module detection.

We search for a combination of factors that is over-represented in a set; therefore a distinction can be made between treating all syntenic regions of one gene independently (in that case, a set contains all regions of all genes separately) and keeping all regions of a gene together (the set contains all genes, each having one or more regions). To investigate this effect, and more importantly to decide whether to keep the regions in a real biological data set together, we tested both possibilities on semi-artificial data as well. Comparing Art_3 (where all regions are added independently to a set) and Art_4 (where multiple syntenic regions of one gene are kept together, see Figure 2F) shows that the second approach is advisable, so this will be applied on the coexpressed gene set as described below.

Sensitivity to PWM scoring

Because the ModuleSearcher algorithm uses the scores of individual matrix hits, we have compared the effectiveness of the algorithm using different types of scoring (as described in Methods). The Art_1 set was scored with the MotifScanner using different values for the prior parameter. When 0.1 or 0.2 were used, the ModuleSearcher found 5 out of 5 correct CRM elements. Using 0.5 as a prior, it found 4 out of 5 elements. The same set was also scored with the MotifLocator, with varying threshold values. The MotifLocator can be compared with other programs that score frequency matrices such as Matinspector (Quandt *et al.*, 1995). Setting the threshold to 0.75 resulted in 4 out of 5 correct elements, but this threshold yields 12 times as many hits as for the MotifScanner with prior 0.2. A threshold of 0.8 resulted in 3 out of 5 correct elements; 0.85 in 1 out of 5 and 0.9 in 0 out of 5. Taken together, the MotifScanner (with its probabilistic estimation of the number of hits) confers robustness to the ModuleSearcher and will be used in the Syntenic GFF database and in the study of coexpressed genes.

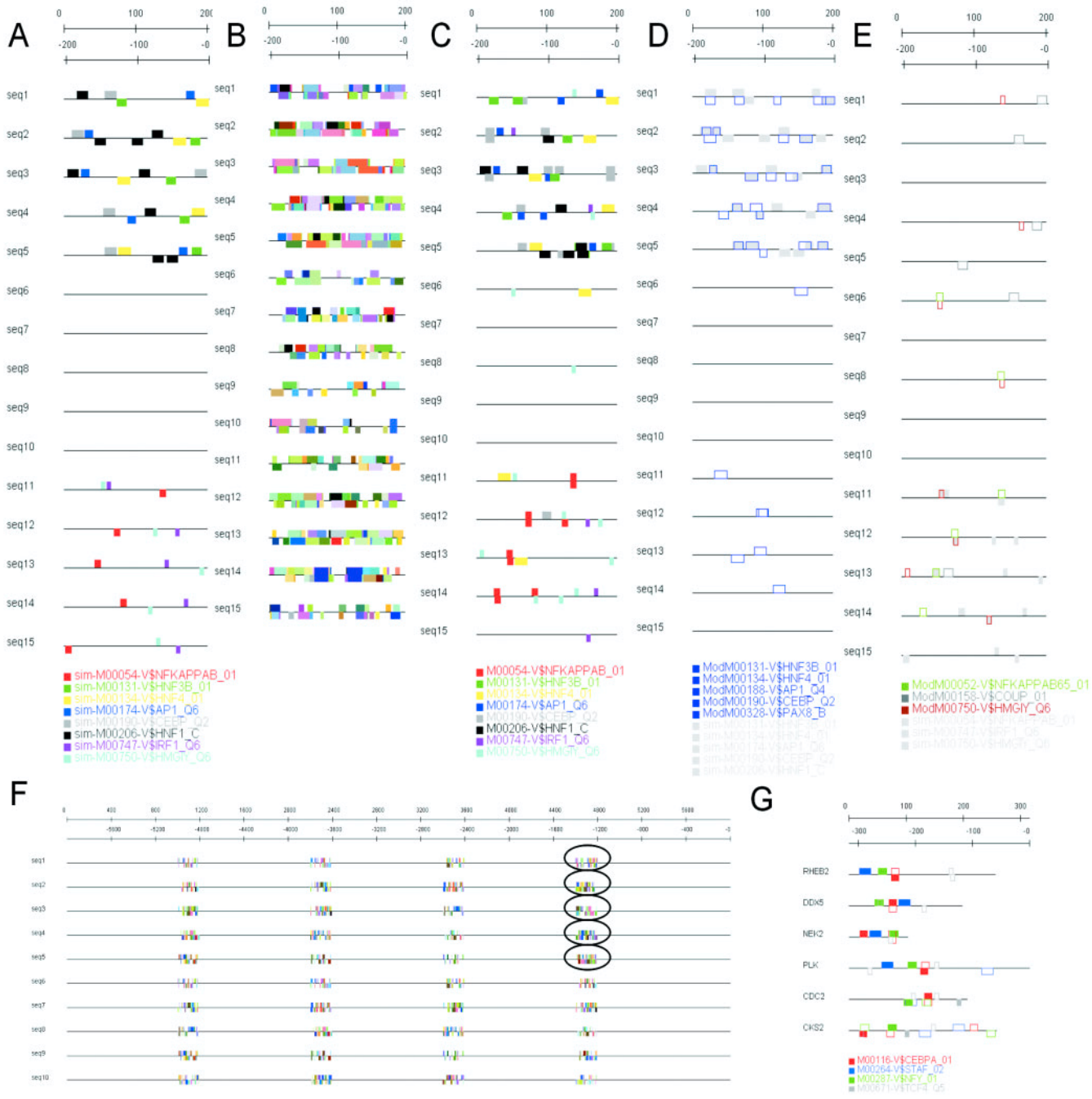

Fig. 2. Module detection in artificial data sets. A. Set Art_2 as described in Table 1, showing only the implanted binding sites, sampled from the respective matrices from TRANSFAC. B. The same set, scored with the MotifScanner using all available matrices. This is the actual data in which the ModuleSearcher will search for modules. C. The same as in B, but now only displaying the instances of the matrices that were implanted. It is clear that there are many false positives and many true negatives, a fact that obviously hinders module detection. D. In blue are the results of a first run of the ModuleSearcher and in grey the implanted sites as in A. E. In red and green are two of the three hidden matrices, as detected in a second run on the same set (masking the results of the first run) of the ModuleSearcher. F. Set Art_4 as described in Table 1, resembling the biological situation where multiple syntenic regions of one gene belong together. Only the encircled regions have implanted modules (5 out of 40 regions), and these can still be detected. G. Six of the 20 highest scoring syntenic regions with the CEBPA-STAF-NFY-TCF4 model that was found in the cyclin B2 microarray cluster. The closed boxes are the sites of the module and the open boxes are putative sites of the same factors scored with a lower threshold. Taking the open and closed boxes together, each region has at least one instance of each module factor.

Table 1. Results of the ModuleSearcher on different sequence sets

Set name	Set contents	Highest scoring module	Notes
Art_1	10 random sequences of 200bp, each with following implants: { M00134-V$HNF4_01, M00131-V$HNF3B_01, M00190-V$CEBP_Q2, M00174-V$AP1_Q6, M00206-V$HNF1_C } + 10 random sequences of 200bp without implants (i.e., noise)	{ M00134-V$HNF4_01, M00131-V$HNF3B_01, M00190-V$CEBP_Q2, [a]M00188-V$AP1_Q6, M00206-V$HNF1_C }	5 out of 5 correct. The found module contains all 5 hidden elements.
Art_2	5 random sequences of 200bp, each with following implants: { M00134-V$HNF4_01, M00131-V$HNF3B_01, M00190-V$CEBP_Q2, M00174-V$AP1_Q6, M00206-V$HNF1_C } + 5 random sequences of 200bp, each with following implants: { M00054-V$NFKAPPAB_01, M00747-V$IRF1_Q6, M00750-V$HMGIY_Q6 } + 5 random sequences of 200bp without implants (i.e., noise) See Figure 2A	First run: { M00134-V$HNF4_01, M00131-V$HNF3B_01, M00190-V$CEBP_Q2, [a]M00188-V$AP1_Q6, [b]M00328-V$PAX8_B } Second run: { [a]M00052-V$NFKAPPAB65_01, M00750-V$HMGIY_Q6, [b]M00158-V$COUP_01 }	The first module was found with 4 out of 5 elements correct. The second module was found after masking the elements of the first module; 2 out of 3 elements of the second module are correct.
Art_3	5 random sequences of 200bp, each with following implants: { M00134-V$HNF4_01, M00131-V$HNF3B_01, M00190-V$CEBP_Q2, M00174-V$AP1_Q6, M00206-V$HNF1_C } + 35 random sequences of 200bp without implants (i.e., noise)	{ [b]M00446-V$SPZ1_01, [b]M00285-V$TCF11_01, [b]M00748-V$STAT5B_Q6, [b]M00137-V$OCT1_03, [b]M00734-V$CIZ_01 }	The hidden module is not found when it is present in only 5 out of 40 sequences.
Art_4	5 genes with 1 module as in Art_1 and 3 empty modules, well separated + 5 genes with 4 empty regions. The empty stretches between the regions are not scored with TRANSFAC. See Figure 2F.	{ M00134-V$HNF4_01, M00131-V$HNF3B_01, M00190-V$CEBP_Q2, [a]M00188-V$AP1_Q6, M00206-V$HNF1_C }	When different regions of the same gene are grouped together, the level of noise is reduced and the module can be found, with 5 out of 5 elements correct.
CCNB2_clus	Set of 13 human genes coexpressed with cyclin B2 during the cell cycle in HeLa cells; selected from SOURCE. In total they have 48 conserved sequence blocks within 10kb upstream of the CDS. The blocks of a gene are grouped together as in Art_4.	{ M00116-V$CEBPA_01, M00264-V$STAF_02, M00287-V$NFY_01, M00671-V$TCF4_Q5 }	This result was validated by finding target genes of the module using the MotifScanner, see text.

[a] Motif belongs to the same class as the implanted motif;
[b] Motif that was not implanted

Genomic searches

Using the ModuleScanner we can score the complete 'Syntenic GFF' database to find syntenic regions that potentially contain a CRM. To determine the specificity of target detection, we have compared the scores of the sequences in the Art_1 set (using the best CRM found with the ModuleSearcher in this set) with the scores of the same (artificial) CRM on the database. There are 6 regions (out

of the 10 regions where we implanted it) that have a higher score than all the regions in the database.

A second test was carried out, this time using a known *cis*-regulatory module, namely the IFN-β enhancer (Munshi *et al.*, 1999). This module contains, within less than 100 base pairs, functional binding sites for NF-$\kappa\beta$, ATF2/JUN, IRF, and HMGI(Y) (four copies and one overlaps with the NF$\kappa\beta$ site). The TRANSFAC database only contains matrices for HMGI(Y), NF$\kappa\beta$, and IRF-1 so we used these three to specify a module model. The ModuleScanner scored the GFF database with this model, and the top 10 scoring genes were fed into the GO4G program. Table 2 shows the significantly over-represented GO terms within these 10 genes, and it can be seen that they are related to the response of a cell to viral infection, the process where the IFN-β enhancer is active. The IFN-β gene itself was found as fourth best scoring gene. Other high scoring genes include: EH-domain containing protein 1 (testilin, HUGO=EHD1) involved in the recycling of major histocompatibility complex class I molecules to the plasma membrane; IL-1 β precursor (catabolin, HUGO=IL1B), an important mediator of the inflammatory response; NF-$\kappa\beta$ inhibitor alpha (HUGO=NFKBIA), involved in apoptosis and possibly pointing at feedback control mechanisms; and semaphorin 3B precursor (HUGO=SEMA3B), involved in cell-cell signaling and possibly coregulated with IFN-β to mediate contacts between dendritic cells and T lymphocytes. By combining transcription factors in modules, the specificity increases to a level where genomic searches become feasible. This result opens the door to the validation of predicted modules, as illustrated in the next paragraph, because a genomic search with a false module will retrieve random top scoring genes that have an extremely low chance of statistical significant functional coherence.

Detecting modules in microarray clusters

The selected gene cluster around cyclin B2 (26 genes, see Data and Methods) is functionally tight: among the highly significantly over-represented Gene Ontology terms are cell cycle (15 genes, p value = 10^{-14}), M phase (9 genes, p value = 3.10^{-13}), and microtubule cytoskeleton (9 genes, p value = 2.10^{-7}). The best module model in the cluster, as selected by the ModuleSearcher (window=100bp and n_{Θ}=4) consisted of NFY, STAF, TCF4, and CEBPA. It has been shown that NFY (nuclear factor Y) regulates genes (e.g., cyclin B1) in a cell type specific and cell-cycle dependent fashion (Katula *et al.*, 1997). TCF4 regulates cyclin D1 expression in a complex with β-catenin (Tetsu and McCormick, 1999), so its involvement in cell-cycle specific expression of other genes is plausible. CEBPA (CCAAT/enhancer binding protein alpha) overlaps with some of the NFY sites (see Figure 2G), which could explain its presence in the module. The fourth element, STAF, is a zinc finger protein that is a promiscuous activator for enhanced transcription by RNA polymerases II and III (Schaub *et al.*, 1997).

Using the [STAF-CEBPA-NFY-TCF4] module in a genomic search with the ModuleScanner shows indeed that this combination contains cell-cycle specific regulatory information, because (1) 30.8% (4 out of 13) of the original cluster is found in the top 100 scoring genes, and (2) the GO4G statistics on the top 20 scoring genes show a significant (corrected p value smaller than 0.05) for terms like ‘mitosis’, ‘regulation of cell cycle’, and ‘cell proliferation’ (see Table 2). Figure 2G shows the actual modules in some of the top 20 scoring cell cycle genes. Polo-like kinase (PLK) is possibly active in chromosomal segregation, NEK2 is involved in chromosome segregation and centrosome separation. CDC2 (cell division cycle 2) is a catalytic subunit of the highly conserved protein kinase complex known as M-phase promoting factor (MPF), which is essential for G1/S and G2/M phase transitions of eukaryotic cell cycle. CKS1B is also known as CDC2 associated protein so its coregulation with CDC2 is plausible.

CONCLUSIONS

The ultimate goal of every technique for *cis*-regulatory sequence analysis is to detect real binding sites for transcription factors that can explain a particular expression profile of a gene. In studies in organisms with compact genomes and with a limited complexity of cooperativity, such as yeast, the detection of over-represented binding sites in the promoters of the genes in a microarray cluster (Tavazoie *et al.*, 1999) yields valuable clues in understanding several aspects of transcriptional regulation, especially in combination with genomewide location analysis (Lee *et al.*, 2002).

To test the rather ambitious thought that our methodology could help in analogous studies in human, we have first tested the proposed algorithms on artificial data and showed that we could find back the hidden modules with a high sensitivity (i.e., after adding multiple sequences without the module), even if many of the implanted sites are missed by the matrix scoring step. The influence of the latter on the robustness of module finding was also tested and it was shown that our probabilistic estimation of the number of hits is more reliable than traditional log odds scoring. Another test showed that the signal to noise ratio is much higher when the syntenic regions of a gene are kept together instead of separating them.

Our current program always finds a ‘best’ module model in a set of sequences. Therefore, it is necessary to validate the module. Some possibilities are (1) the ability to retrieve target genes in the genome, (2) functional coherence

Table 2. Validating putative target genes found by the ModuleScanner using GO4G

Genes	Significant GO terms	Corrected *p* value
Top 10 scoring genes for a simplified IFN-β enhancer: M00750-V$HMGIY_Q6 M00054-V$NFKAPPAB_0 M00747-V$IRF1_Q6	apoptosis	0.002046759
	negative regulation of cell proliferation	0.003660635
	protein amino acid dephosphorylation	0.004239594
	response to pest/pathogen/parasite	0.005201112
	protein phosphatase activity	0.006733237
	innate immune response	0.010171468
	cytokine activity	0.012253669
	response to stress	0.014523294
	phosphoric monoester hydrolase activity	0.015017083
	cell communication	0.031928904
Top 20 scoring genes for a new module found with the ModuleSearcher on a set of cyclin B2 coexpressed genes: M00116-V$CEBPA_01 M00264-V$STAF_02 M00287-V$NFY_01 M00671-V$TCF4_Q5	mitosis	0.000435808
	M phase of mitotic cell cycle	0.000452022
	cytokinesis	0.000468573
	nuclear division	0.001257531
	regulation of cell cycle	0.001395887
	protein serine/threonine kinase activity	0.010734361
	obsolete	0.024402181
	cell proliferation	0.026461498

Terms with at least 2 occurrences and corrected *p* value smaller than 0.05 are shown. When both parent and child terms were significant, only the child is shown.

of predicted target genes, (3) structure conservation of the modules in the training set and in the top scoring database modules, and (4) phylogenetic footprinting. Structure conservation can imply conserved strand preferences or distances between binding sites. Here we have only used (1) and (2). We tested these approaches using the known IFN-β enhancer model and the results show that real module models are specific enough to find back their instances in the full genome.

Lastly we predicted a module in a set of coexpressed genes and validated the prediction using the same approach. It was shown that module detection can yield valuable hypotheses and these can ultimately help in cracking the complex gene regulatory code.

How exactly the top scoring genes are related to the modules remains to be investigated. We believe however that using the described approaches, the *in silico* generated hypotheses regarding *cis*-regulation should have a higher success rate compared to approaches based on single factors or that do not take cross-species sequence conservation into account.

Regarding future developments, it will be feasible to perform module detection analysis on several kinds of gene sets (expression clusters, protein complexes, text-based clusters, and so on) to build a database of hypotheses on *cis*-regulation. Here, a biologist who is working on the regulation of his 'pet gene' could filter his list of his own wet-lab hypotheses by checking whether related *in silico* hypotheses exist in the database.

ACKNOWLEDGEMENTS

The authors thank M. Dabrowski, B. De Strooper, and B. Hassan from the Center for Human Genetics, K.U.Leuven and VIB, Belgium for helpful discussions. Research supported by Research Council KUL: GOA-Mefisto 666, IDO; Flemish Government: FWO: projects G.0115.01, G.0240.99, G.0407.02, G.0413.03, G.0388.03, G.0229.03, ICCoS, ANMMM; AWI; IWT; STWW-Genprom, GBOU-McKnow, GBOU-SQUAD, GBOU-ANA; Belgian Federal Government: DWTC (IUAP IV-02 and IUAP V-22); EU: CAGE; ERNSI; Contract Research/agreements: VIB.

REFERENCES

Aerts,S., Thijs,G., Coessens,B., Staes,M., Moreau,Y. and De Moor,B. (2003) Toucan: deciphering the *cis*-regulatory logic of coregulated genes. *Nucleic Acids Res.*, **31(6)**, 1753–1764.

Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99(2)**, 757–762.

Bray,N., Dubchak,I. and Pachter,L. (2003) AVID: A Global Alignment Program. *Genome Res.*, **13(1)**, 97–102.

Davidson,E.H. (2001) *Genomic Regulatory Systems*. Academic Press, San Diego, USA.

Diehn,M., Sherlock,G., Binkley,G., Jin,H., Matese,J.C., Hernandez-Boussard,T., Rees,C.A., Cherry,J.M., Botstein,D., Brown,P.O. and Alizadeh,A.A. (2003) SOURCE: a unified genomic resource

of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, **31(1)**, 219–223.

Frith,M.C., Spouge,J.L., Hansen,U. and Weng,Z. (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.*, **30(14)**, 3214–3224.

Halfon,M.S., Grad,Y., Church,G.M. and Michelson,A.M. (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.*, **12(7)**, 1019–1028.

Hart,P., Nilsson,N. and Raphael,B. (1968) A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Syst. Sci. Cybern.*, **SSC-4**, 100–107.

Hubbard,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30(1)**, 38–41.

Jegga,A.G., Sherwood,S.P., Carman,J.W., Pinski,A.T., Phillips,J.L., Pestian,J.P. and Aronow,B.J. (2002) Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res.*, **12(9)**, 1408–1417.

Katula,K.S., Wright,K.L., Paul,H., Surman,D.R., Nuckolls,F.J., Smith,J.W., Ting,J.P., Yates,J. and Cogswell,J.P. (1997) Cyclin-dependent kinase activation and S-phase induction of the cyclin B1 gene are linked through the CCAAT elements. *Cell Growth Differ.*, **8(7)**, 811–820.

Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, **11(9)**, 1559–1566.

Kullback,S. (1959) *Information Theory and Statistics*. John Wiley & Sons, New York, USA.

Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298(5594)**, 799–804.

Lermen,M. and Reinert,K. (2000) The practical use of the A* algorithm for exact multiple sequence alignment. *J. Comput. Biol.*, **7(5)**, 655–671.

Loots,G.G., Ovcharenko,I., Pachter,L., Dubchak,I. and Rubin,E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, **12(5)**, 832–839.

Markstein,M., Markstein,P., Markstein,V. and Levine,M.S. (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc. Natl Acad. Sci. USA*, **99(2)**, 763–768.

Mayor,C., Brudno,M., Schwartz,J.R., Poliakov,A., Rubin,E.M., Frazer,K.A., Pachter,L.S. and Dubchak,I. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **16(11)**, 1046–1047.

Munshi,N., Yie,Y., Merika,M., Senger,K., Lomvardas,S., Agalioti,T. and Thanos,D. (1999) The IFN-beta enhancer: a paradigm for understanding activation and repression of inducible gene expression. *Cold Spring Harb. Symp. Quant. Biol.*, **64**, 149–159.

Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23(23)**, 4878–4884.

Rebeiz,M., Reeves,N.L. and Posakony,J.W. (2002) SCORE: a computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc. Natl Acad. Sci. USA*, **99(15)**, 9888–9893.

Schaub,M., Myslinski,E., Schuster,C., Krol,A. and Carbon,P. (1997) Staf, a promiscuous activator for enhanced transcription by RNA polymerases II and III. *EMBO J.*, **16(1)**, 173–181.

Stein,L. (2002) Creating a bioinformatics nation. *Nature*, **417(6885)**, 119–120.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16(1)**, 16–23. Historical Article.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22(3)**, 281–285.

Tetsu,O. and McCormick,F. (1999) Beta-catenin regulates expression of cyclin D1 in colon carcinoma cells. *Nature*, **398(6726)**, 422–426.

Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor.,B., Rouze,P. and Moreau,Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17(12)**, 1113–1122.

Thijs,G., Marchal,K., Lescot,M., Rombouts,S., De Moor,B., Rouze,P. and Moreau,Y. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.*, **9(2)**, 447–464.

van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281(5)**, 827–842.

Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278(1)**, 167–181.

Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. and Botstein,D. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell.*, **13(6)**, 1977–2000.

Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28(1)**, 316–319.

1998Yuh,C.H., Bolouri,H. and Davidson,E.H. Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279(5358)**, 1896–1902.

Yuh,C.-H., Brown,C.T., Livi,C.B., Rowen,L., Clarke,P.J.C. and Davidson,E.H. (2002) Patchy interspecific sequence similarities efficiently identify positive *cis*-regulatory elements in the sea urchin. *Dev. Biol.*, **246(1)**, 148–161.