BIOINFORMATICS

Vol. 17 no. 11 2001
Pages 1019–1026

Combining frequency and positional information to predict transcription factor binding sites

Szymon M. Kiełbasa, Jan O. Korbel*, Dieter Beule, Johannes Schuchhardt and Hanspeter Herzel*

Innovationskolleg Theoretische Biologie, Humboldt-Universität zu Berlin, Invalidenstraße 43, D-10115 Berlin, Germany

Received on December 10, 2000; revised on April 27 and July 8, 2001; accepted on July 25, 2001

ABSTRACT

Motivation: Even though a number of genome projects have been finished on the sequence level, still only a small proportion of DNA regulatory elements have been identified. Growing amounts of gene expression data provide the possibility of finding coregulated genes by clustering methods. By analysis of the promoter regions of those genes, rather weak signals of transcription factor binding sites may be detected.

Results: We introduce the new algorithm ITB, an Integrated Tool for Box finding, which combines frequency and positional information to predict transcription factor binding sites in upstream regions of coregulated genes. Motifs are extracted by exhaustive analysis of regular expression-like patterns and by estimating probabilities of positional clusters of motifs. ITB detects consensus sequences of experimentally verified transcription factor binding sites of the yeast *Saccharomyces cerevisiae*. Moreover, a number of new binding site candidates with significant scores are predicted. Besides applying ITB on yeast upstream regions, the program is run on human promoter sequences.

Availability: ITB is available upon request.

Contact: s.kielbasa@itb.biologie.hu-berlin.de

INTRODUCTION

A growing number of complete genome sequences provides the possibility of studying function and regulation of genes. High throughput differential expression measurements allow scientists to analyze the transcription of various human and model organism genes in response to environmental stimuli or selected transcription factors—numerous families of coregulated genes will be discovered in the near future. Promoter regions of genes regulated by a common factor can be studied in order to extract transcription factor binding sites (Brazma *et al.*, 1998; Roth *et al.*, 1998; van Helden *et al.*, 1998; Zhang, 1999). Even though hundreds of binding sites are described in public databases, the detailed regulation of most genes remains unknown. Algorithms capable of detecting regulatory elements may provide insights into important molecular processes and help to model the complexity of genetic networks.

Several different approaches are used to detect regulatory elements in the promoter regions of coregulated genes. An exhaustive algorithm RSA-tools-oligo-analysis (van Helden *et al.*, 1998) compares the frequency of conserved words in a given set of promoter sequences to the frequencies in a reference set (in the following termed training set). This method is sensitive in detecting over-represented words in the upstream regions of coregulated yeast genes. Unfortunately, regulatory elements lacking a conserved core sequence might be not detected by this method. Weight matrix based methods like AlignACE (Roth *et al.*, 1998), an algorithm based on the Gibbs sampling algorithm (first described by Lawrence *et al.*, 1993), or MEME (Bailey and Elkan, 1994) use a multiple alignment strategy to detect the signals of DNA regulatory elements. In this way, elements lacking a conserved core may be found. However, signals due to regulatory elements involved in transcription are rather weak. If only a small number of motifs occur in the coregulated sequence set, weight matrices are of limited use. In such a case, mono- or dimeric repeats or non-specific signals like the motif `AAATAA` are more likely to be aligned than functional regulatory elements. An intermediate solution, usable for small sets of coregulated genes, is to search exhaustively for regulatory elements expressed with letters matching multiple bases. Such a method was proposed by (Sinha and Tompa, 2000) and here a similar approach for scoring frequencies of motifs is used.

We introduce the algorithm ITB, an Integrated Tool for Box finding, which integrates frequency and positional information to predict transcription factor binding sites in upstream regions of coregulated genes. Regular expression-like patterns are analyzed exhaustively, allowing small gaps and the matching of more than one base at any position of the motifs. The patterns having the highest

*Both authors contributed equally.

© Oxford University Press 2001

frequency and positional scores are selected as candidates for regulatory elements.

METHODS

Scoring motif frequencies

ITB written in C++ compares observed frequencies of conserved elements in the given promoter set to their expected frequencies, estimated based on a training (reference) set (here, the complete set of yeast upstream regions of a length of 800 bp is used). The program performs an exhaustive search—scores in this paper are calculated for all possible 6-mers built from the alphabets ACGT or ACGTWRKSYMN[†]. Finally, a list of the motifs with the best scores is created, containing the most over-represented patterns in comparison to the training set.

As the scoring formula for a motif W we use the Z-score function $Z(W)$ (van Helden *et al.*, 2000a), which characterizes the difference between the observed number of occurrences $n_{\mathrm{obs}}(W)$ of the motif W (on any strand) and the expected mean number $\mu(W)$, scaled by the expected standard deviation of the motif $\sigma(W)$:

$$Z(W) = \frac{n_{\mathrm{obs}}(W) - \mu(W)}{\sigma(W)}. \qquad (1)$$

To estimate the standard deviation of the motif we use expressions (introduced by Pevzner *et al.*, 1989) taking into account the possibility of motif self-overlapping (e.g. AAAAAA or TATATA). We have adapted these formulae to double-strand analysis and the extended alphabet case:

$$\begin{aligned}\mu(W) &= \sum_{w\in\mathcal{W}} N p(w),\\ \sigma^2(W) &= \sum_{w\in\mathcal{W}} N p(w)(1-p(w)) \\ &\quad +2\sum_{s=1}^{L_W-1}(N-sM)\sum_{w,v\in\mathcal{W}}(\pi_s^{w,v}-p(w)p(v))\end{aligned} \qquad (2)$$

where N is the total number of possible motif positions in the coregulated set of M promoters. By definition, $\mathcal{W}$ contains all oligonucleotides w (expressed in the four-letter alphabet) that match the pattern W or the pattern complementary to it[‡]. The number of letters in the pattern W is represented by L_W.

The remaining two symbols $p(w)$ and $\pi_s^{w,v}$ link the Z-score formula to the background model: $p(w)$ is the probability of an ACGT-type sequence w and $\pi_s^{w,v}$ expresses the probability of a sequence built by overlapping the sequences v at position $s+1$ and w at the first position[§]. These probabilities are estimated with Markov models built on the training set.

ITB ranks over-represented patterns according to their Z-score. Self-overlapping patterns with a periodicity of one or two (e.g. AAAAAA or AWAWAW) are removed from the output (if this option is selected).

Scoring positional information

As Figure 2 illustrates, transcription factor binding sites are often clustered or appear at preferred positions. To consider such information, we implemented a positional scoring procedure. For a motif observed $n_{\mathrm{obs}}(W)$ times in a coregulated sequence set, the score reveals positional clusters for subsets of matches.

First, all motif positions extracted from upstream regions of a regulatory family are accumulated to generate an 'artificial' sequence (see Figure 2 for illustration of the artificial sequence generation). Subsets of motif matches may form clusters of n_{cl} motifs—so-called n_{cl}-clusters. ITB extracts the lengths l_{cl} of the shortest n_{cl}-clusters ($n_{\mathrm{cl}} = 2, 3, \ldots, n_{\mathrm{obs}}(W)$) in the artificial sequence of length L_{seq}. This is illustrated below ($n_{\mathrm{obs}}(W) = 4$; motifs are symbolized by 'x'; the shortest $n_{\mathrm{cl}} = 2$, 3, and 4-clusters in the sequence are highlighted by brackets):

```
.......x..x....x.............x....
...{([x..x]...x).............x}..
```

Using stochastic simulations, we estimate the probability $p_{\mathrm{pos}}(n_{\mathrm{cl}}, l_{\mathrm{cl}})$ of observing n_{cl} motifs within a stretch of length l_{cl} bases assuming equidistributed positions over the complete stretch of L_{seq}. Simulations are repeated Q times (we applied $Q = 10^4$). The number of observations of a n_{cl}-cluster of at most l_{cl} bases divided by Q gives an estimation of $p_{\mathrm{pos}}(n_{\mathrm{cl}}, l_{\mathrm{cl}})$. The cluster corresponding to the smallest probability is selected and $-\log_{10}(p_{\mathrm{pos}}(n_{\mathrm{cl}}, l_{\mathrm{cl}}))$ applied as the positional score of the motif candidate. If a n_{cl}-cluster of length l_{cl} or shorter is not observed at all, the highest score of $\log_{10}(Q) = 4$ is assigned.

Data sets

ITB was run on 11 yeast regulatory families. These gene sets (except the MAT family) were identical to clusters used by van Helden *et al.* (1998). The MAT family was constructed from DNA microarray experiments that revealed genes with a mating type specific transcriptional regulation (Roth *et al.*, 1998). We selected the 10 transcripts whose levels increased the most in mating type 'a' relative to 'α' (genes: MFA1, MFA2, AGA2, STE2, BAR1, PMP1, SRA3, VPS13, YKR071C and

[†] These symbols are taken from the standard IUB nucleotide code: W=A or T, R=A or G, K=G or T, S=C or G, Y=C or T, M=A or C, and N=any of ACGT. Symbols matching three letters were not used—for small numbers of matches the difference between them and N can be neglected.

[‡] For example, for W = ASTG the described expansion gives $\mathcal{W}$ = (ACTG, AGTG, CAGT, CACT).

[§] For example $\pi_2^{\mathrm{ATAC,ACGG}} = p(\mathrm{ATACGG}), \pi_2^{\mathrm{ATAC,TTGG}} = 0.$

YBR147W). Sequences were retrieved from the MIPS database (Mewes *et al.*, 1997) by extracting regions upstream of the corresponding ORFs.

Two human sequence sets were studied. Zuber *et al.* (2000) reported the downregulation of several genes via the human H-Ras protein involving the RAF/MEK/ERK cascade of cytoplasmic kinases. Five promoter regions of coregulated genes (LOX, LOXL1, LOXL2, RIL/LIM and TSP1) were kindly provided by the group. Moreover, four promoter sequences of genes (EP11114, EP15041, EP36018 and EP37001) upregulated via the c-Myc protein (Coller *et al.*, 2000) were extracted from the EPD database (Perier *et al.*, 1998). ITB was trained using all 271 human promoter sequences available from EPD. These do not include the studied genes regulated via H-Ras. For the analysis of c-Myc-controlled genes, the four promoters extracted from EPD were removed from the training set.

RESULTS

Evaluation of the frequency score

Table 1 shows the results of an analysis of 11 coregulated yeast gene sets. ITB was run with the alphabets ACGTWRKSYMN (the 'extended' mode) and ACGT (the 'ACGT' mode) using the default settings (word length 6, analyzed upstream sequence length 800, Markov chain model order 3). An option to remove self-overlapping patterns with a periodicity of 1 or 2 was applied. The 10 highest Z-scores were considered. Run several times on randomly generated sets of upstream regions the Z-scores were distributed nearly Gaussian, with not more than 0.5% of motifs with the Z-score in the range (3, 5) and with less than 0.0001% of motifs with the Z-score above 5.

ITB predicted most previously characterized elements correctly. When the families NIT, MET, INO, PHO, PDR, GCN, YAP, and TUP were analyzed in the 'extended' mode, the highest scoring motifs matched previously characterized transcription factor binding sites. Analysis of the MAT family revealed a motif corresponding to a previously described consensus sequence at rank 3. In the MET family, a second site previously characterized that is not directed by the Cbf1/Met4/Met28-complex but by Met31/32 (Kuras *et al.*, 1996) was also found by ITB. In most families analyzed, several of the top predictions partly matched previously identified consensus sequences. Applying ITB in the 'ACGT' mode instead of the 'extended' mode principally led to similar results.

Besides matches to previously identified consensus sequences, a number of sequences not matching known elements were predicted. In particular, the motifs CAACAA predicted in the INO family (Z-score 6.9, information content 5.2 bits; according to Schneider *et al.*, 1986[†]) and CGTTCC (6.3, 8.0 bits) that was found in the YAP family stand out with significant scores. Some additional high scoring motifs, which self-overlap several times, were not considered as promising candidates.

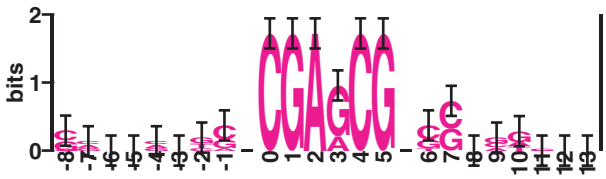

Fig. 1. Logo of the significant pattern CGARCG found in promoters of genes regulated via H-Ras. Bars show sample size error corrections.

ITB failed to detect two known elements—the consensus sequences of the transcription factors regulating the GAL and HAP family, respectively.

The publicly available algorithms AlignACE version 3.0 (Roth *et al.*, 1998), RSA-tools-oligo-analysis (van Helden *et al.*, 1998), and MEME version 2.2 (Bailey and Elkan, 1994) were also applied to detect motifs in the 11 yeast regulatory families. No algorithm was able to detect all of the described motifs.

In order to evaluate the applicability of ITB for the prediction of functional sites in the human genome, the algorithm was applied to human sequences. No significant motif similar to the expected element CACGTG was found in the promoters of genes regulated via c-Myc. In contrast, the analysis of five genes regulated via H-Ras (Table 1) revealed the significant pattern CGARCG (Z-score = 10.1, information content = 10.9 bits, see Figure 1). The motif does not match any previously characterized site listed in the TRANSFAC database (Wingender *et al.*, 1996) and does not self-overlap in the promoter sequences matched.

Variation of the parameters

In order to verify the motif predictions and to test the robustness of the algorithm, analyses were repeated several times, changing some parameters of ITB. In each run, a single parameter of ITB run in the 'ACGT' mode was changed while all other parameters were kept at their default values (word length 6, upstream region of 800 bp, Markov model of order 3).

Applying different word lengths revealed highly ranked motifs that match previously identified consensus sequences for most analyzed patterns (see Table 2 for results regarding the YAP and INO families). In this

[†] The information content is determined for the pattern region and 8 letters on each side of the motif. The formula used by ITB is the one described for skewed genomes, applying the sampling error correction introduced by Miller (1955).

Table 1. Highly ranked motifs as computed by the program ITB are similar to the consensus sequences of previously characterized transcription factor binding sites. Consensus sequences are taken from TRANSFAC (Wingender *et al.*, 1996)[1] or from van Helden *et al.* (1998)[2] (see also references therein). Besides detecting matches to known consensus sequences, ITB predicted new candidates for transcription factor binding sites. For an analysis using the 'extended' mode, predicted 6-mers are presented along with the pattern rank, the total number of motifs, the number of sequences matched, the Z-score, and the information content. In the last column, predictions of ITB run in the 'ACGT' mode are presented. Self-overlapping patterns with a periodicity of one or two were automatically removed from the output

Family name (no. of genes)	Experimentally verified motifs	Bound factors	ITB prediction in 'extended' mode (rank)	No. of motifs/sequences matched	Z-score (Information content [bits])	ITB prediction in 'ACGT' mode (rank)
NIT(7)	GATAAG[2]	Gln3	GATAAG(1)	26/6	13.9 (7.2)	GATAAG(1)
MET_1(11)	TCACGTG[2]	Cbf1/Met4/Met28	CACGTG(1)	13/11	13.6 (12.8)	CACGTG(1)
MET_2(11)	AAAACTGTGG[2]	Met31/Met32	ACYSKG(4)	39/8	4.8 (9.2)	CTGTGG(2)
PHO(5)	CACGTKNG[1]	Pho4	ACGTGS(1)	18/5	12.1 (7.2)	ACGTGC(1)
PDR(7)	TCCGCGGA[2]	Pdr1/Pdr3	CCGYGG(1)	18/4	15.3 (12.9)	CCGTGG(1)
INO_1(10)	CATGTGAAWT[2]	Ino2/Opi1	CATGTG(1)	15/9	7.9 (10.5)	CATGTG(1)
INO_2(10)	–	unknown	CAACAA(2)	28/10	6.9 (5.2)	CAACAA(2)
TUP(25)	$KANW_4ATSYG_4W$[2]	Mig1	GYGGGG(1)	33/18	11.7 (8.3)	GGGGTA(1)
YAP_1(16)	TTACTAA[2]	Yap1	MTTASK(1)	99/16	5.1 (6.5)	CATTAC(2)
YAP_2(16)	–	unknown	CGTTCC(2)	15/16	6.3 (8.0)	CGTTCC(1)
GAL(6)	$CGGN_5WN_5CCG$[2]	Gal4	–	–	–	–
HAP(8)	YCNNCCAATNANM[1]	Hap2/Hap3/Hap4	–	–	–	–
GCN(38)	RTGACTCATNS[1]	Gcn4	TGACTC(1)	44/26	12.5 (7.1)	TGACTC(1)
MAT(10)	CRTGTNNW[1]	Matα2	CATGYA(3)	21/7	5.1 (6.3)	CATGTA(2)
c-Myc(4)	CACGTG[1]	c-Myc/Max	–	–	–	–
Ras(5)	–	unknown	CGARCG(1)	9/4	10.1 (10.9)	CGAGCG(1)

Table 2. Z-scores of motifs predicted by ITB ('ACGT' mode) as a function of the analyzed motif length. Only complete or shifted by 1 bp matches to previously identified consensus sequences were considered. Z-scores of the most over-represented oligonucleotides matching known consensus sequences are presented besides the ranks of the motifs (in brackets). Patterns marked with an asterisk represent new motif candidates

Family	Motif	4	5	6	7	8
INO_1	CATGTGAAWT	1.7 (12)	3.0 (10)	**7.9** (1)	**11.0** (1)	**19.3** (1)
INO_2	aCAACAAs*	**4.1** (1)	**5.2** (1)	6.9 (2)	6.8 (7)	6.9 (40)
YAP_1	TTACTAA	1.8 (8)	4.3 (2)	4.5 (3)	**7.7** (1)	9.9 (2)
YAP_2	cCGTTCCs*	1.6 (16)	3.3 (4)	**6.3** (1)	6.5 (5)	9.1 (5)

Table 3. Z-scores of motifs predicted by ITB ('ACGT' mode) as a function of the background model. Ranks of predicted patterns are given in brackets. Peaks of the score are indicated by bold face symbols. Symbols mean: E = equiprobable base distribution; B = single nucleotide probability based on the GC-content of the training set (Markov model order 0); M1, ..., M4 = Markov chain models of the orders 1 to 4; F = probability of motifs based on the 6-mer frequency in the training set (Markov model order 5)

Family	Motif	E	B	M1	M2	M3	M4	F
INO_1	CATGTG	5.4 (17)	6.3 (4)	6.6 (2)	6.5 (2)	**7.9** (1)	7.5 (1)	7.7 (1)
INO_2	CAACAA*	**12.0** (2)	9.7 (1)	7.8 (1)	8.4 (1)	6.9 (2)	5.9 (2)	6.0 (2)
YAP_1	CATTAC	5.9 (25)	4.3 (16)	5.4 (2)	**6.5** (1)	5.1 (2)	5.2 (2)	5.0 (2)
YAP_2	CGTTCC*	3.5 (72)	6.6 (5)	**6.8** (1)	6.3 (2)	6.3 (1)	6.5 (1)	6.4 (1)

analysis, only complete matches to previously identified consensus sequences were considered (shifts of motifs by a maximum of 1 bp were allowed). When word lengths of 5–7 were selected, the top ITB predictions for 6 regulatory families revealed complete matches to previously identified transcription factor binding sites. Choosing larger or smaller motif lengths led to a decrease in ITB predictions corresponding to known sites, but often provided additional information for motifs with wider conserved cores (e.g. the pattern CATTACTAA of the YAP family).

Moreover, ITB predictions were analyzed for varying background models (results regarding the YAP and INO families are presented in Table 3). When an equiprobable base distribution was assumed, lower motif ranks and often significantly lower Z-scores were observed. When instead single nucleotide probabilities were considered, or when a Markov chain model of the order 1 was applied, the resulting predictions were better, but still biased towards non-specific, partly self-overlapping patterns. Applying Markov chain models of the orders 2–5 revealed the highest numbers of motifs that match previously identified consensus sequences.

Finally, the robustness of the predictions was tested for variations of the analyzed sequence length upstream of the

translation start. While Z-score and ranks of some motifs depend rather weakly on the analyzed sequence length, ITB detected elements with strong maximums of the Z-score for particular lengths. These peaks indicate strong positional preferences of particular motifs. For instance, the pattern CATGTA matching a known regulatory element of the MAT family revealed the highly significant Z-score of 7.7 and the top rank for an upstream region length of 400 bp, while a score of only 5.1 and the rank 2 resulted applying the default length of 800 bp.

Evaluation of positional information

The dependence of motif predictions on the analyzed upstream sequence length points to positional peculiarities of transcription factor binding sites. Experimentally verified sites of 52 yeast transcription factors were extracted from the SCPD database (Zhu and Zhang, 1999). While some motif types appear randomly distributed over the respective promoters, a high proportion of the sites listed in SCPD reveal a strong position bias. Clusters of binding sites of four transcription factors are shown in Figure 2. For Gal4, Gcn4, and Matα2 a clustering was observed, despite the alignment of sequences according to the translation start instead of the transcription initiation site.

Positional scoring was implemented in ITB as a result of these observations. Figure 3 presents outcomes obtained by applying ITB in the 'ACGT' mode—positional score and Z-score are combined in two-dimensional scatter plots. Signals of motifs corresponding to known consensus sequences are often located at the top right corner of the plots. Moreover, sequences very similar to expected patterns were frequently detected with high scores. When the MET, GCN, INO, TUP and NIT families were analyzed, signals corresponding to previously identified motifs represented the most appealing predictions clearly separated from the other spots. In the case of the PDR and PHO families, several variations of the expected sequences TCCGCGGA and CACGTKNG were predicted—at least one of each having both significant Z-scores and positional scores.

Moreover, ITB extracted new candidates for transcription factor binding sites. For instance, an analysis of the PHO family revealed the significant candidate CGTATA (Z-score 4.4, positional score 3.2). However, significant positional scores are often caused by strongly self-overlapping signals. In the MAT family, the three highest-scoring motif candidates represent shifted variations of the strongly self-overlapping GACGAC. The previously identified motif CATGTA represents the most appealing prediction of this family, if the former three signals are not considered.

The new candidate CAACAA of the INO family revealed lower scores than the expected CATGTG. However, a calculated Z-score of 6.9 along with a significant positional score of 3.1 stress the potential importance of the candidate. An analysis of the YAP family revealed the following result: Not the most over-represented oligonucleotides occupied the top right corner, but other motifs that represent good matches to the expected motif TTACTAA or the potential candidate CGTTCC.

Analyzing the human genes coregulated via H-Ras revealed a number of potential motif candidates. However, most of those signals represent strongly self-overlapping sequences or rather non-specific sequences like GGGCGG or GGGGCG. The sequence CCGAGC, a good match to the previously identified candidate CGARCG (see above), revealed a significant Z-score of 6.1 along with a positional score of 1.2.

DISCUSSION

ITB is a sensitive and powerful algorithm that integrates frequency and positional information to detect transcription factor binding sites in coregulated yeast genes. ITB performs an exhaustive search for regular expression-like patterns, allowing gaps and the matching of more than one base at any position of the motifs. Such an approach may be considered a good compromise between searches for frequent oligonucleotides and weight-matrix based methods. It allows the detection of motifs that are not completely conserved—and it guarantees to find the most significant elements due to the exhaustive search strategy. In order to correct an enlarged variance of the number of motifs due to self-overlapping, we apply an appropriate correction formula as in Sinha and Tompa (2000) (an analysis of the algorithm complexity as the function of the alphabet and the length of motifs is also presented there).

ITB detected highly significant motifs that correspond to functional regulatory elements found by experimental analysis. Only a limited set of additional patterns was predicted. Even when positional information was not considered, known regulatory elements of 8 out of 11 yeast regulatory families were predicted applying the 'extended' mode. Similar results were obtained using the 'ACGT' mode.

Combining both Z-score and positional score increases the specificity of ITB. Motifs with significant positional scores that correspond to previously identified consensus sequences were extracted from 9 out of 11 families. Only few new candidate motifs having both high positional and Z-scores were predicted.

A couple of reasons might lead to the observed positional preferences. Among these are interactions of transcription factors with the pre-initiation complex, the removal of single nucleosomes within a promoter, protein–protein interactions within factors that stabilize weak protein–DNA interactions, or recent duplications of regulatory regions. Moreover, an accumulation of functional sites might serve to increase the local concentration

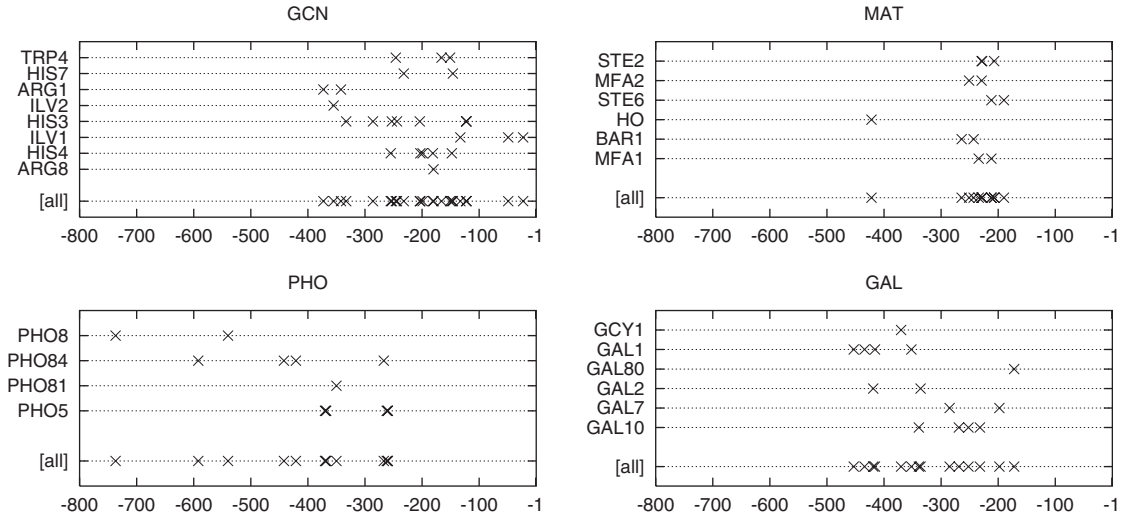

Fig. 2. Map showing the locations of experimentally verified binding sites of Matα2, Gcn4, Pho4, and Gal4 in upstream regions. All positions were extracted from SCPD (Zhu and Zhang, 1999) and are relative to the translation start site. '[all]' stands for 'artificial' upstream sequences accumulating all binding site positions of the respective factors. Gene sets shown here are not identical to the regulatory families analyzed in this study (see van Helden *et al.*, 1998) for detailed gene lists and descriptions on the generation of the regulatory families).

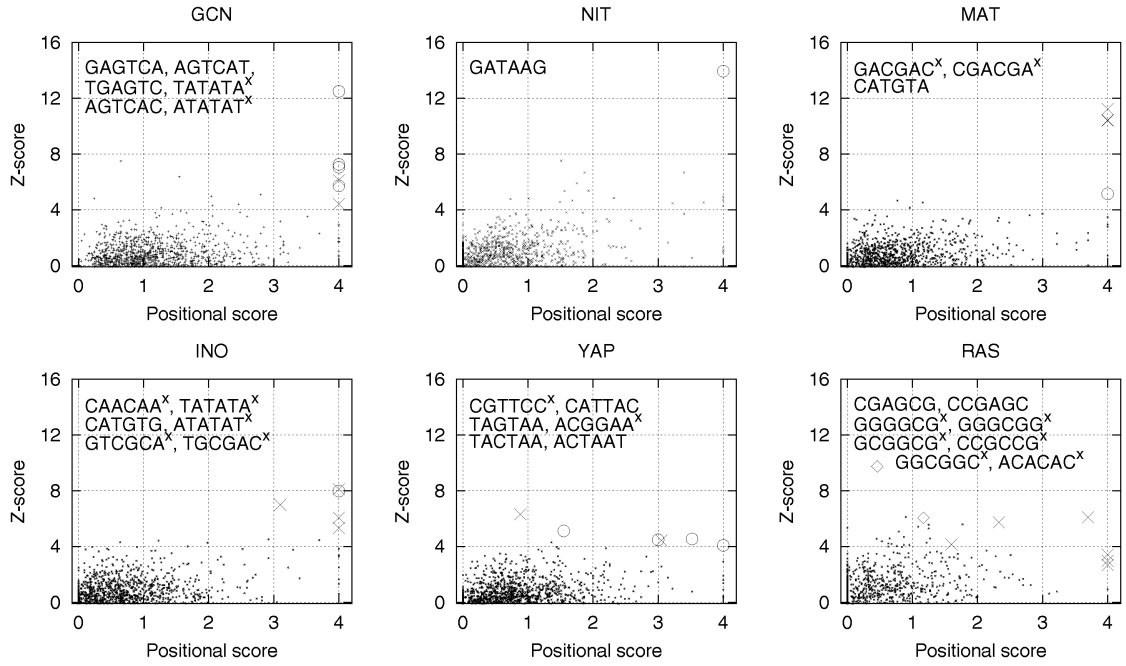

Fig. 3. Scatter plots combining Z-score and positional score generated by ITB ('ACGT' mode). Yeast coregulated gene families and human genes coregulated via H-Ras were analyzed. Circles correspond to motifs matching previously identified consensus sequences, while crosses indicate new candidates for transcription factor binding sites (marked with 'x'). Diamonds indicate matches to the motif CGARCG previously identified by ITB (for genes coregulated via H-Ras). Oligonucleotides are listed from the left to the right (or—if signals are one upon the other—from the top to the bottom). Self-overlapping motifs were not removed from the output.

of transcription factors. In our opinion, highly significant positional clusters are not necessarily a property of functional promoter elements, but rather serve to support the hypothesis of the functionality of medium candidates.

While the ITB algorithm was being implemented, a new paper by Hughes *et al.* (2000) appeared describing a new version of AlignACE. Their algorithm calculates a 'positional bias score', which is determined using a strategy different from the approach described in this manuscript. In the Hughes approach, for each candidate weight matrix extracted, the best candidates matching within 600 bp upstream of all existing ORFs are extracted from the analyzed genome. Then, a clustering of these motifs is evaluated. By this method, the group observed a significant positional bias for over one third of all transcription factor binding sites analyzed.

ITB failed to detect described consensus sequences of 2 families. Some tools like AlignACE, Dyad-detector (van Helden *et al.*, 2000b) or the method in Sinha and Tompa (2000) are capable of predicting the consensus described for the GAL family. Currently, ITB does not detect sequences like the Gal4 binding site, but is limited to motifs without longer gaps. According to van Helden *et al.* (1998) these features are shared by a number of yeast regulatory elements.

Some motifs predicted by ITB are potential candidates for novel transcription factor binding sites. The elements `CGTTCC` and `CAACAA` represent strong candidates for regulatory elements involved in the regulation of the YAP and INO family. Moreover, analyzing the PHO family revealed the significant candidate `CGTATA`. The predictions are largely robust against variations of most algorithm parameters. Furthermore, the calculation of high positional scores for the candidates or at least for other motifs matching the candidates supports a selection of these patterns. Searching TRANSFAC for the elements `CGTTCC`, `CAACAA` and `CGTATA` did not reveal matches to known yeast consensus sequences.

Based on the parameter variations performed we conclude that motif lengths of 6 or 7 are a reasonable choice for using ITB. However, analyzing wider sequences may still provide further information. Applying Markov models of the orders 2–5 led to comparable predictions. Variation of upstream sequence lengths led to significant changes of the Z-score—indicating positional peculiarities of motifs.

Since ITB succeeded in detecting regulatory elements in yeast upstream regions, it is reasonable to assume that motifs may also be detected in human promoters. Since even long human upstream sequences may not contain the transcription initiation site, our analysis focused on experimentally mapped promoters. While the expected box was not found for genes controlled by c-Myc, we introduce the motif `CGARCG`—a candidate transcription factor binding site of genes coregulated via H-Ras.

We stress that a regulation via the same protein cascade within a similar time scale does not automatically imply the regulation by a single transcription factor. Moreover, functional elements in mammalian genomes may act far away from the promoter.

More data will be analyzed to make a clearer statement about the applicability of ITB for human promoters. Prerequisites for an analysis of promoters in the genomes of higher eukaryotes are the usage of proper training sets and the reliable detection of coregulated genes (Herzel *et al.*, 2001).

Combining frequency and positional scores provides a good starting point for extensions of the algorithm. We will include the matching of words with longer gaps and the consideration of palindromes. Distinct factors bound in close proximity may interact—therefore, clusters of different binding site types should also be taken into account, particularly when higher eukaryotes are analyzed.

In the future, ITB might be applied for an exhaustive search for regulatory elements in whole genomes. If at least a few regulatory families of an organism are known, ITB could be used to extract consensus sequences and positional information of motifs.

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximisation to discover motifs in biopolymers. In *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, AAAI Press, Menlo Park, CA, pp. 28–36.

Brazma,A., Jonassen,I., Vilo,J. and Ukkonen,E. (1998) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, **8**, 1202–1215.

Coller,H.A., Grandori,C., Tamayo,P., Colbert,T., Lander,E.S., Eisenman,R.N. and Golub,T.R. (2000) Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc. Natl Acad. Sci. USA*, **97**, 3260–3265.

Herzel,H., Beule,D., Kielbasa,S., Korbel,J., Sers,C., Malik,A., Eickhoff,H., Lehrach,H. and Schuchhardt,J. (2001) Extracting information from cDNA arrays. *Chaos*, **11**, 98–107.

Hughes,T., Marton,M., Jones,A., Roberts,C., Stoughton,R., Armour,C., Bennett,H., Coffey,E., Dai,H., He,Y., Kidd,M., King,A., Meyer,M., Slade,D., Lum,P., Stepaniants,S., Shoemaker,D., Gachotte,D., Chakraburtty,K., Simon,J., Bard,M. and Friend,S. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Kuras,L., Chrest,H., Surdin-Kerjan,Y. and Thomas,D. (1996) A heteromeric complex containing the centromere binding factor 1 and two basic leucine zipper factors, met4 and met28, mediates the transcription activation of yeast sulfur metabolism. *EMBO J.*, **15**, 2519–2529.

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wooton,J.C. (1993) Detecting subtle

sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Mewes,H.W.K.A., Heumann,K., Liebl,S. and Pfeiffer,F. (1997) MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.*, **25**, 28–30.

Miller,G. (1955) *Information Theory in Psychology*. Free Press, Glencoe, IL.

Perier,R.C., Junier,T. and Bucher,P. (1998) The eukaryotic promoter database EPD. *Nucleic Acids Res.*, **26**, 353–357.

Pevzner,P.A., Borodovsky,M.Y. and Mironov,A.A. (1989) Linguistics of nucleotide sequences I: the significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J. Biomol. Struct.*, **6**, 1013–1026.

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

Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.

Sinha,S. and Tompa,M. (2000) A statistical method for finding transcription factor binding sites. In *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, Vol. 8, AAAI Press, Menlo Park, CA, pp. 344–354.

van Helden,J., André,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**, 827–842.

van Helden,J., del Olmo,M. and Perez-Ortin,J.E. (2000a) Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.*, **28**, 1000–1010.

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

Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.

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

Zhu,J. and Zhang,M. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.

Zuber,J., Tchernitsa,O.I., Hinzmann,B., Schmitz,A.C., Grips,M., Hellriegel,M., Sers,C., Rosenthal,A. and Schäfer,R. (2000) A genome-wide survey of RAS transformation targets. *Nature Genet.*, **24**, 144–152.