Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation

Lars Juhl Jensen and Steen Knudsen

Center for Biological Sequence Analysis, Department of Biotechnology, Building 208, The Technical University of Denmark, DK-2800 Lyngby, Denmark

Received on May 13, 1999; revised and accepted on November 3, 1999

Abstract

Motivation: *The whole genomes submitted to GenBank contain valuable information about the function of genes as well as the upstream sequences and whole cell expression provides valuable information on gene regulation. To utilize these large amounts of data for a biological understanding of the regulation of gene expression, new automatic methods for pattern finding are needed.*

Results: *Two word-analysis algorithms for automatic discovery of regulatory sequence elements have been developed. We show that sequence patterns correlated to whole cell expression data can be found using Kolmogorov–Smirnov tests on the raw data, thereby eliminating the need for clustering co-regulated genes. Regulatory elements have also been identified by systematic calculations of the significance of correlations between words found in the functional annotation of genes and DNA words occuring in their promoter regions. Application of these algorithms to the Saccharomyces cerevisiae genome and publicly available DNA array data sets revealed a highly conserved 9-mer occuring in the upstream regions of genes coding for proteasomal subunits. Several other putative and known regulatory elements were also found.*

Availability: *Upon request.*

Contact: *steen@cbs.dtu.dk*

Introduction

Although the genomic sequence of *Saccharomyces cerevisiae* has been determined (Goffeau *et al.*, 1997), and thereby the sequence of all the promoters, the regulation of gene expression is far from being fully understood. A few computational approaches have been applied to the identification of upstream regulatory sequences (Brazma *et al.*, 1998; van Helden *et al.*, 1998; Spellman *et al.*, 1998). These methods all search for motifs in the upstream regions that are shared by a given group of genes known to have a common biological function and/or regulation. However, no systematic search for correlations between upstream elements and gene function has been reported yet.

The recent availability of whole cell expression data has added a new way of finding groups of co-regulated genes: clustering the expression profiles. Although obtaining biologically meaningful clusters from the expression data is not a trivial task, once accomplished, any algorithm for finding shared sequence patterns can be used for identification of regulatory elements.

In the analysis by van Helden *et al.* (1998) of word frequencies in upstream regions of several co-regulated clusters [some of which were based on whole cell expression data for the diauxic shift (DeRisi *et al.*, 1997)], many known regulatory sequences were found. It clearly illustrated that although the word description is simplistic, it is useful for finding transcription factor binding sites and that grouping of similar words can provide a description of degenerate patterns.

Clusters of genes with similar expression profiles during the mitotic cell cycle have been investigated by Spellman *et al.* (1998). Although this approach was mainly focused on identification of co-regulated genes, several known transcription factor binding sites were found by an analysis of upstream regions.

Here, we describe a new method that does not require clustering, thereby circumventing problems like choosing the optimal number of clusters to use in the analysis. An entirely new approach that relies on functional annotation of genes is also presented. Because no experimental data is required this algorithm can be applied directly to existing databases.

Methods

Data sets

From all 6269 ORFs annotated in the *S. cerevisiae* GenBank files (accession numbers U00091, Y13134, X59720, Z71256, U00092, D50617, Y13135, U00093, Z47047, Y13136, Y13137, Y13138, Z71257, Y13139, Y13140, and U00094) two 200 bp regions were extracted: one region starting at 300 bp upstream and another starting 100 bp downstream of the ORF. The 100 bp gap between the extracted regions and the ORFs serves to avoid 5′ and 3′ UTRs.

For systematic analysis of correlations between upstream regulatory sequences and gene function, another set of 5097 upstream regions of 500 bp and the functional annotation of the corresponding genes was compiled based on the gene function annotations available from SGD (http://genome-www.stanford.edu/Saccharomyces/).

A number of different DNA chip data sets were analysed, all of which were made using the Stanford method described in DeRisi *et al.* (1997). One data set analysed is the diauxic shift data—the first DNA chip data set ever (DeRisi *et al.*, 1997). This experiment shows the gene expression response of *S. cerevisiae* to glucose starvation at seven time points. A similar time series of gene expression during sporulation by Chu *et al.* (1998) and the data for the variation in gene expression over the mitotic cell cycle measured in various synchronized cultures (Spellman *et al.*, 1998) were also analysed.

The DNA chip data sets all consist of a set of red–green ratios representing the relative expression level of a gene in two different samples—usually the expression at some time point compared to at the start of the experiment. For analysis of these data 91 data sets were made—one per time point in all experiments. These data sets consisted of the 500 bp upstream regions and the red-green ratios of all genes for which data were available at the time point in question.

Hypergeometric statistics on sequence patterns

An algorithm was developed for identification of sequence patterns—represented as non-degenerate words—that are overrepresented in a set of sequences (positive set) compared to a reference set (negative set). For each of the two sets, the number of sequences that contain each pattern was counted using a suffix tree. Counting the number of sequences rather than the number of occurrences ensures mutually independent observations, thereby allowing for the use of sample statistics as a rigorous statistical method for determining which patterns are significantly overrepresented.

Unlike most other software, the two strands are treated separately when counting patterns. By doing so we gain sensitivity on patterns that show strong preference for one orientation. The RPG box and the HOMOL1 box shown in Table 2 are examples of such patterns. The price paid is a lower sensitivity on patterns with little or no strand preference.

Four numbers are needed for deciding whether a pattern is significantly overrepresented: the number of sequences in the positive set that contain the pattern (P_t); the number in the negative set that contain the pattern (P_f); the number of sequences in each of the two sets that do not contain the pattern (N_f and N_t for the positive and negative set, respectively).

	Positive set	Negative set
With pattern	P_t	P_f
Without pattern	N_f	N_t

As the positive set is a sample of size $P_t + N_f$ from the pool of all sequences ($P_t + P_f + N_f + N_t$) which contains $P_t + P_f$ sequences with the pattern, the expected distribution for P_t is the hypergeometric distribution $H(P_t + N_f, P_t + P_f, P_t + P_f + N_t + N_f)$. Using the AS R77 algorithm (Lund, 1980; Shea, 1989) $P(x < P_t)$ was calculated for all patterns that occurred in the positive set. The exact significance potential ($p\alpha$) of any pattern of length L being overrepresented was then calculated as:

$$p\alpha = -\log(1 - P(x < P_t)^{(4^L)})$$

In this work we have analysed all correlations found to have a significance potential of at least 4. All groups of such overlapping pattern have been reported with the exception of patterns stemming from the subtelomeric regions (see discussion).

Because of numerical difficulties in calculating the above expression for very high levels of significance an alternative measure was also used for quantifying the overrepresentation of patterns namely the Mathews correlation coefficient (Mathews, 1975):

$$C = \frac{P_t N_t - P_f N_f}{\sqrt{(P_t + P_f)(P_t + N_f)(N_t + P_f)(N_t + N_f)}}.$$

The correlation coefficient attains the value 1 in case of a perfect positive correlation, −1 for a perfect negative correlation, and 0 for a completely random distribution of the k-tuple between the two sets. However, one should keep in mind that the Mathews correlation coefficient is not a measure of statistical significance—in contrast to the exact hypergeometric statistics.

Systematic analysis of functional annotation

The functional annotation associated with each gene or ORF was converted into a pseudo-sequence by removal of all non-alphanumeric characters (including spaces) and conversion of lowercase letters to uppercase. A dictionary of all words (k-tuples) up to a length of 10 letters that occurred in at least two pseudo-sequences was made.

For each of the approximately 10 000 words in the dictionary, the set of 500 bp upstream regions was divided into a corresponding positive set consisting of the sequences containing the word in their functional annotation and a negative set not containing the word. In all the positive sets and their corresponding negative set the number of DNA sequences containing each k-tuple up to length 10 bp was counted and analysed using hypergeometric statistics (the two strands were again treated separately). Because we are able to quickly perform this exact statistical test of all functional words in the dictionary compared to all DNA patterns, we do not preprocess the dictionary to reduce its size.

Kolmogorov–Smirnov statistics on DNA array data

A DNA word correlated to a specific time point can be found by Kolmogorov–Smirnov statistics. Because the Kolmogorov–Smirnov test is a rank test, the set of upstream regions is first sorted based on the red–green ratio from the time point being analysed, thereby assigning each sequence a rank in the form of its position in the dataset.

To calculate the significance of a given pattern being correlated to the observed red–green ratios, the following ratio is calculated:

$$\frac{\max_{1\leq i\leq N}\left|\frac{x_i}{n}-\frac{i}{N}\right|}{\sqrt{\frac{n+N}{n\cdot N}}}$$

N is the number of sequences in the set, n the number of sequences containing the pattern, and x_i the number of sequences of rank up to i that contain the pattern. To give a better balance between long and short patterns, the Kolmogorov–Smirnov ratio divided by the pattern length was used for evaluation of the patterns. In this work all correlations with a ratio above 0.3 were examined manually and all patterns with a ratio of 0.4 have been reported.

This test value can be calculated from the summary values N, n, and $\max|Nx_i - ni|$. By using a suffix tree with three counters per pattern, these can be obtained for all patterns in just two sweeps through the set of sorted sequences. In the first sweep, the number of sequences containing each pattern is stored in one counter. During the second sweep, the number of these sequences yet encountered and the highest value of $|Nx_i - ni|$ so far are stored in two other counters.

Implementation

The algorithms described were implemented in C++ on an SGI Power Challenge. Since all three algorithms are implemented as pattern counting in suffix trees, their time complexity is the same—$O(kL)$, where k is the maximal pattern length and L is the total length of the sequences. For the analysis of functional annotation and DNA array data this is the complexity of analysing one dictionary word/one time point respectively. Because a suffix tree is used for storing the patterns, there is no minimal pattern length; even patterns of length 1 are analysed. However, we have so far not found any patterns shorter than four nucleotides to be significant.

Table 1. Patterns found more frequently upstream than downstream of genes. The significance calculated using hypergeometric statistics is show as $p\alpha$. The column Corr contains the Mathews correlation coefficients and the two last columns show the number of upstream and downstream regions containing the pattern

	Pattern	$p\alpha$	Corr.	Up	Down
Reb1p	`CGGGTAA...`	7.6	0.063	216	94
	`CGGGTA....`	>10	0.066	304	149
	`.GGGTAA...`	5.3	0.054	421	268
	`.TTACCCG..`	>10	0.074	234	87
	`..TACCCGG.`	6.0	0.058	109	32
	`GTTACCC...`	5.9	0.058	154	60
	`..TACCCG..`	>10	0.080	330	139
	`.TTACCC...`	>10	0.069	485	279
	`...ACCCGG.`	6.4	0.058	173	73
MCB	`..ACGCG...`	5.2	0.051	466	311
	`...CGCGT..`	5.2	0.051	473	317
	`..TCGCG...`	>10	0.070	513	298
	`...CGCGA..`	5.7	0.053	469	309
	`...CGCG...`	>10	0.100	1426	937
	`....GCGA..`	>10	0.084	2057	1581
	`..TCGC....`	>10	0.081	2206	1736
Cbf1p-Met2p-Met28p	`..CACGTG..`	5.0	0.053	182	86
Curved element	`.AAAATTTTT`	4.2	0.056	240	122
	`.AAAATTTT.`	4.2	0.054	444	286

Results and discussion

We present here the results obtained by applying all three approaches to the *S. cerevisiae* data sets. The few abundant regulatory elements found by comparison of upstream and downstream regions are shown in Table 1; as is a new putative curved element. We also present both novel putative elements and known regulatory elements with significant correlation to the functional annotation of genes (Table 2) and whole cell expression data (Table 3).

Known TF binding sites

A simple comparison of patterns that are more frequently present in upstream than in downstream regions (see Table 1) reveals the recognition sites for some transcription factors known to regulate many genes. One of these is Reb1p which both plays an essential role in the termination of transcription of ribosomal RNAs (Lang and Reeder, 1993; Lang *et al.*, 1994) and functions as general transcription regulator (Remacle and Holmberg, 1992). This illustrates that although the method is very simplistic, it does actually work.

Table 2. Correlations between upstream sequence elements and functional annotation of genes found by systematic analysis. The columns $p\alpha$ and Corr contain the significance found by hypergeometric statistics and the Mathews correlation coefficient for each pattern

	Pattern	$p\alpha$	Corr.	Function
MCB	`...ACGCGT....`	>10	0.167	DNA replication
	`....CGCGTA...`	6.0	0.121	
	`...ACGCG.....`	>10	0.109	
	`....CGCGT....`	>10	0.113	
Curved element	`..AAAATTTTTT.`	>10	0.178	Protein synthesis
	`...AAATTTTTT.`	>10	0.153	
	`..AAAATTTTT..`	>10	0.144	
	`..AAAATTTT...`	6.5	0.125	
	`...AAATTTTT..`	6.3	0.126	
	`...AAATTTT...`	6.1	0.103	
Proteasomal element	`.GGTGGCAAAA..`	>10	0.420	Proteasomal subunit
	`..GTGGCAAAA..`	>10	0.358	
	`.GGTGGCAAA...`	>10	0.275	
	`..GTGGCAAA...`	>10	0.286	
	`.GGTGGCAA....`	>10	0.232	
	`..TTTGCCACC..`	>10	0.157	
	`...TTGCCACC..`	3.4	0.166	
RPG	`.ACCCATACAT..`	>10	0.144	Ribosomal proteins
	`..CCCATACAT..`	>10	0.161	
	`..CCCATACA...`	>10	0.151	
	`...CCATACAT..`	>10	0.147	
	`.ACCCATAC....`	5.8	0.142	
	`..CCCATAC....`	5.5	0.122	
HOMOL1	`CATCCGTACA...`	>10	0.177	Ribosomal proteins
	`.ATCCGTACAT..`	>10	0.181	
	`...CCGTACATTT`	>10	0.148	
	`CATCCGTAC....`	>10	0.185	
	`.ATCCGTACA...`	>10	0.187	
	`..TCCGTACAT..`	>10	0.171	
	`...CCGTACATT`	5.4	0.167	
	`CATCCGTA.....`	6.1	0.149	
	`.ATCCGTAC....`	>10	0.205	
	`..TCCGTACA...`	>10	0.210	
	`...CCGTACAT..`	>10	0.145	
Met31p, Met32p	`...AAACTGTG..`	>10	0.204	Methionine biosynthesis
	`....AACTGTG..`	4.6	0.138	
	`.....ACTGTG..`	3.6	0.100	
	`..CCACAGTT...`	2.5	0.172	
	`..CCACAGT....`	2.8	0.132	
Cbf1p-Met2p-Met28p	`....CACGTG...`	6.3	0.153	Methionine biosynthesis
	`...ACGTGA....`	4.7	0.116	
Methionine element	`...GTGACTCA..`	>10	0.157	Methionine biosynthesis
	`..CGTGACT....`	3.8	0.128	
	`....TGACTCA..`	>10	0.157	
	`....TGACTC...`	>10	0.145	
	`.....GACTCA..`	6.2	0.116	
	`.....GACTC...`	6.1	0.087	

The STRE element (Marchler *et al.*, 1993), a general stress response regulator, was not surprisingly found most strongly correlated to the diauxic shift experiment but also correlated to the elutriation cell cycle experiment (Table 3). One might speculate that the yeast cells were in fact stressed during the elutriation centrifugation.

The pattern `CACGTG`, which from functional annotation was found to be correlated to methionine biosynthesis, was surprisingly also found to be overrepresented in promoter regions in general and even weakly correlated to the cell cycle regulation (Tables 1, 2, and 3). The latter is consistent with the results of Spellman *et al.* (1998), namely that many genes involved in methionine biosynthesis are subject to cell cycle regulation. The global overrepresentation may be explained by the word `CACGTG` being the consensus sequence for both the Cbf1p–Met2p–Met28p complex and PHO4 (O'Connell and Baker, 1992). The consensus sequences of Met31p and Met32p, `AAACTGTG`, were also found to be significantly correlated to methionine biosynthesis.

The MluI cell cycle box (MCB) (McIntosh *et al.*, 1991) with the consensus sequence `WCGCGW` (`ACGCGT` being the preferred sequence) was found by the three independent methods used in this work (see Tables 1, 2, and 3). The search for patterns overrepresented in promoter regions revealed that several variations of the consensus sequence occurred about 50% more in upstream regions than in downstream regions. The systematic analysis of functional annotation correctly identified all of these MCB-like patterns as being strongly correlated to DNA replication.

Moreover, MCB was found to be correlated to regulation in all DNA chip experiments concerning the mitotic cell cycle as well as sporulation. This is reassuring since DNA replication is a part of both sporulation and the ordinary cell cycle. On the other hand the sporulation specific elements URS1 (Park *et al.*, 1992) and MSE (Ozsarac *et al.*, 1997) were both found to be strongly correlated to the regulation during sporulation, but not to regulation in any of the DNA chip data sets concerning the mitotic cell cycle.

In the cell cycle experiments the most significant correlations were found at time points during the first cycle of the experiment. This is probably because the cell cultures become less synchronized over time. This should be seen in contrast to other experiments such as diauxic shift and sporulation, where the most significant correlations are found near the end of the experiments, where the gene expression has changed the most compared to the initial time point (data not shown).

The systematic analysis of functional annotations also revealed a large number of words correlated to ribosomal proteins. By assembling the words into a longer sequence, the consensus sequence of the RPG box (`ACACCCATACAT`) (Vignais *et al.*, 1987, 1990) and HOMOL1 (`WACATCYRTRCA`) (Larkin *et al.*, 1987) were obtained (Table 2).

Table 3. Patterns found correlated to regulation by Kolmogorov–Smirnov statistics (ratio $\geq$ 0.3). All values are Kolmogorov–Smirnov ratios divided by pattern length. The Diaux and Spo columns contain the results from the diauxic shift (DeRisi *et al.*, 1997) and sporulation (Chu *et al.*, 1998) datasets. The columns Alpha, CDC15, CDC28, CLB2, CLN3, and Elu shows the results from different synchronized cultures in the the cell cycle data set by Spellman *et al.* (1998)

		Diaux	Spo	Alpha	CDC15	CDC28	CLB2	CLN3	Elu
MSE	.TTTTGTG..		0.411						
	..TTTGTG..		0.538						
	...TTGTG..		0.415						
	..CACAAAA.		0.434						
	.CCACAAA..		0.423						
	..CACAAA..		0.516						
URS1	.TAGCCGCC.		0.419						
	..AGCCGCC.		0.450						
	...GCCGCC.		0.442						
MCB	..ACGCGT..		0.690	0.845	0.665	0.853	0.507	0.694	0.558
	..ACGCGA..			0.527	0.666	0.631	0.388	0.599	0.500
	..TCGCGT..			0.509	0.477	0.579	0.421	0.619	0.449
	..ACGCG...		0.628	0.825	0.849	0.943	0.516	0.814	0.672
	...CGCGT..		0.640	0.858	0.696	0.940	0.589	0.913	0.652
	..TCGCG...		0.389	0.647	0.554	0.644	0.387	0.664	0.421
	...CGCGA..		0.362	0.535	0.778	0.769	0.347	0.637	0.525
	...CGCG...		0.493	0.669	0.737	0.937	0.481	0.820	0.534
Cbf1p-Met2p-Met28p	..CACGTG..			0.326	0.313	0.359	0.349		
STRE	..AAGGGG..	0.557			0.422				0.574
	...AGGGG..	0.630	0.329						0.689
	..CCCCT...	0.541	0.381		0.382				0.668
Curved element	AAAAATTTT.	0.549	0.553	0.513	0.407	0.472			0.574
	.AAAATTTTT	0.520	0.536	0.494	0.400			0.347	0.560
	AAAAATTT..	0.518	0.540	0.492	0.473	0.433			0.612
	.AAAATTTT.	0.666	0.685	0.720	0.539	0.546		0.422	0.706
	..AAATTTTT	0.521	0.518	0.580	0.458			0.386	0.573
	.AAAATTT..	0.519	0.575		0.474	0.474		0.303	0.615
	..AAATTTT.	0.593	0.596	0.662	0.516			0.398	0.652
Proteasomal element	GGTGGCAAA.				0.313				0.312
	GGTGGCAA..		0.325		0.340				0.328
	.GTGGCAAA.		0.307						0.316
	GGTGGCA...		0.305		0.357				0.313
	.GTGGCAA..		0.354		0.328				
	.GTGGCA...			0.312	0.323				

Putative proteasomal element

The perhaps most exciting result of the systematic analysis of functional annotations was a highly conserved pattern in upstream regions of genes encoding proteasomal subunits. Because synthesis of proteasomes require expression of the genes for all proteasomal subunits, it would make biological sense to expect a shared regulatory element for these genes.

The consensus sequence of the pattern, GGTGGCAAA, was found within 200 bp of the translation start in 25 out of 31 genes assigned as proteasomal subunits, whereas only 26 of the more than 6000 other upstream regions contain this pattern. Among these 26 apparent false positives are two proteases and three genes with relation to ubiquitin, all of which could very well be co-regulated with proteasomes.

The program DIALIGN (Morgenstern *et al.*, 1996) was used for generating an alignment of regions with local similarities from the set of 500 bp upstream regions of proteasomal genes (Table 4). Only two such regions were found—these were the consensus sequence and its reverse complement. This consensus that allows for some degeneracy was found in the upstream region of 28 of the 31 proteasomal genes.

Table 4. The two locally homologous regions found by DIALIGN (Morgenstern *et al.*, 1996) in the upstream regions of genes encoding proteasomal subunits. The two motifs are reverse complement of each other and match the upstream element that was found correlated to proteasomal subunits. The '–' characters represent gaps in the alignment

ORF	Pattern
YBL041W	GGGTGGCAAAA
YBR173C	CAGTGGCAATT
YDL147W	CGGTGGCAAAA
YDL097C	GGGTGGCAAAT
YDL007W	CGGTGGCAAAT
YDR394W	AGGTGGCAAAA
YDR427W	CGGTGGCAAAA
YER012W	CGGTGGCAAAA
YER094C	CGGTGGCAAAA
YGL048C	AGGTGGCAAAA
YGR135W	AGGTGGCAAAA
YGR253C	CGGTGGCAAAA
YHR027C	CGGTGGCAAAA
YJL001W	CGGTGGCAAAA
YOR117W	-GGTGGCAAAA
YOR157C	GAGTGGCAAAA
YOR261C	CAGTGGCAAAT
YOR362C	CGGTGGCAAAA
YPR103W	GGGTGGCAAAA
YER021W	ATTTGCCACCC
YFR050C	ATTTGCCACCT
YFR052W	ATTTGCCACCC
YGL011C	ATTTGCCACCG
YHR200W	-TTCGCCACCG
YKL145W	-TTTGCCACCC
YOL038W	-TTTGCCACCG
YOR259C	ATTTGCCACCG
YPR108W	-TTTGCCACCC

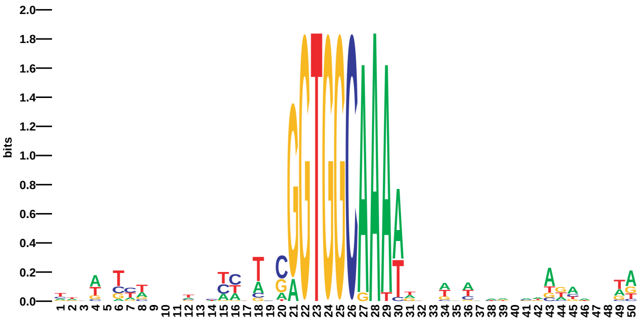

Fig. 1. Sequence logo of aligned occurrences of the proteasome pattern in 500 bp upstream regions of genes encoding proteasome subunits.

A sequence logo (Schneider and Stephens, 1990) of the 28 true positives aligned at the pattern (Figure 1) was made to investigate if this pattern could be an artifact from some larger conserved sequence. This was found unlikely as the information content dropped to near zero outside the consensus sequence. This was further verified by pairwise alignments of all 500 bp regions upstream of proteasomal coding regions (data not shown). It is also improbable that the pattern should be an RNA signal or a leader peptide in the 5′-UTR, as the pattern shows only weak strand preference.

Other putative elements

As was the case for the Cbf1p-Met2p-Met28p complex and MCB, patterns similar to AAAATTTT were picked up by all three methods used here. The pattern is two-fold overrepresented in upstream regions compared to downstream and is found most strongly correlated to genes involved in protein synthesis (Tables 1 and 2). Analysis of the expression data sets reveals strong correlation to regulation during both normal cell cycle and sporulation (Table 3). But unlike MCB and Cbf1p-Met2p-Met28p, AAAATTTT is also including diauxic shift, which implies that it is not specifically involved in regulation of the cell cycle.

One possible explanation for the apparent lack of specificity is that the sequence AAAATTTT is perhaps not a regulatory site itself, but rather an element that enhances the regulation by various transcription factors. Measurements of gel mobilities have shown that AAAATTTT has high intrinsic curvature (Hagerman, 1986), thus one possible function of this element could be to bring other regulatory sites into mutual proximity like the RPG box (Vignais *et al.*, 1987).

Like the consensus sequences for Cbf1p-Met2p-Met28p and Met31p/Met32p GTGACTCA was also found to have significant correlation to methionine biosynthesis (Table 3). The core of this pattern (TGACTC) has been shown to be involved in general control of amino acid biosynthesis with GCN4 as the most likely binding factor. However, GCN4 has a strong discrimination against GTGACTCAC for which reason it cannot be the binding factor for the methionine pattern (Tzamarias *et al.*, 1992).

In addition to these putative elements a large number of patterns were found to be correlated to the functions of genes located in the subtelomeric regions. These patterns have all been discarded, as the subtelomeric regions appear to have a different oligonucleotide composition than the rest of the genome, for which reason we do not believe that the patterns found are regulatory elements.

Comparison of the three approaches

The three methods described for finding regulatory elements each have their strengths and weaknesses, which make them applicable in different situations. The primary strength of comparing upstream and downstream regions is that only the positions of genes need to be known. It is therefore possible to identify regulatory elements in whole genomes without knowledge of the function or regulatory behavior of each gene. However, only the most abundant elements can be determined by this approach.

Analysis of functional annotation has the very important advantage over the other methods, that it finds not only the regulatory element itself but also provides information on the possible function of the element. In addition to the sequence, this method also requires the availability of functional annotations for the genes. This is not a major limitation since the function of many genes can be inferred from sequence similarity to known proteins. The most important limitation is probably, that the method relies on consistent use of words and phrases in the functional annotations.

In contrast to the first two methods, additional experimental results are required for the Kolmogorov–Smirnov approach. But when whole cell expression data is available, the method can automatically find regulatory patterns responsible for the co-regulation of sets of genes, without the need for first performing a clustering of the data.

Possible applications

Although all results presented in this paper are related to pattern finding in yeast promoters and analysis of DNA chip expression data, the algorithms presented are applicable to a much broader class of problems.

Because no experimental data are needed for the approach using hypergeometric statistics for correlating upstream patterns and functional annotation of genes, this algorithm can be applied to all of the fully sequenced and annotated genomes as well as any other database of annotated sequences. Ongoing efforts in this regard will be published elsewhere.

The Kolmogorov–Smirnov algorithm can be used for pattern finding in any set of ranked sequences. This implies that the algorithm should be equally well suited for analysing whole cell expression generated by SAGE (Velculescu *et al.*, 1995) or differential display PCR (Habu *et al.*, 1997), as for analysis of data from DNA arrays.

Acknowledgements

This work was supported by a grant from the Danish National Research Foundation. We thank Anders Gorm Pedersen for useful discussions on Kolmogorov–Smirnov statistics.

References

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.

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

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

Goffeau *et al.* (1997) The yeast genome directory. *Nature*, **387**, 5–105.

Habu,Y., Fukada-Tanaka,S., Hisatomi,Y. and Iida,S. (1997) Amplified restriction fragment length polymorphism-based mRNA fingerprinting using a single restriction enzyme that recognizes a 4-bp sequence. *Biochem. Biophys. Res. Commun.*, **234**, 516–521.

Hagerman,P. (1986) Sequence-directed curvature of dna. *Nature*, **321**, 449–450.

Lang,W. and Reeder,R. (1993) The REB1 site is an essential component of a terminator for RNA polymerase I in *S. cerevisiae*. *Mol. Cell. Biol.*, **13**, 649–658.

Lang,W., Morrow,B., Ju,Q., Warner,J. and Reeder,R. (1994) A model for transcription termination by RNA polymerase I. *Cell*, **7984**, 527–534.

Larkin,J., Thompson,J. and Woolford Jr,J. (1987) Structure and expression of the *Saccharomyces cerevisiae* CRY1 gene: a highly conserved ribosomal protein gene. *Mol. Cell. Biol.*, **7**, 1764–1775.

Lund,R. (1980) Algorithm AS 152: Cumulative hypergeometric probabilities. *Appl. Statist.*, **29**, 221–223.

Marchler,G., Schuller,C., Adam,G. and Ruis,H. (1993) A *Saccharomyces cerevisiae* UAS element controlled by protein kinase a activates transcription in response to a variety of stress conditions. *EMBO J.*, **12**, 1997–2003.

Mathews,B. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

McIntosh,E., Atkinson,T., Storms,R. and Smith,M. (1991) Characterization of a short, *cis*-acting DNA sequence which conveys cell cycle stage-dependent transcription in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **11**, 329–337.

Morgenstern,B., Dress,A. and Werner,T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl Acad. Sci. USA*, **93**, 12098–12103.

O'Connell,K. and Baker,R. (1992) Possible cross-regulation of phosphate and sulfate metabolism in *Saccharomyces cerevisiae*. *Genetics*, **132**, 63–73.

Ozsarac,N., Straffon,M., Dalton,H. and Dawes,I. (1997) Regulation of gene expression during meiosis in Saccharomyces: SPR3 is controlled by both ABFI and a new sporulation control element. *Mol. Cell. Biol.*, **17**, 1152–1159.

Park,H., Luche,R. and Cooper,T. (1992) The yeast UME6 gene product is required for transcriptional repression mediated by the CAR1 URS1 repressor binding site. *Nucl. Acids Res.*, **20**, 1909–1915.

Remacle,J. and Holmberg,S. (1992) A REB1-binding site is required for GCN4-independent ILV1 basal level transcription and can be functionally replaced by an ABF1-binding site. *Mol. Cell. Biol.*, **12**, 5516–5526.

Schneider,T. and Stephens,R. (1990) Sequence logos: A new way to display consensus sequences. *Nucl. Acids Res.*, **18**, 6097–6100.

Shea,B. (1989) Remark AS R77: A remark on algorithm AS 152: Cumulative hypergeometric probabilities. *Appl. Statist.*, **38**, 199–204.

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

Tzamarias,D., Pu,W. and Struhl,K. (1992) Mutations in the bzip domain of yeast GCN4 that alter DNA-binding specificity. *Proc. Natl Acad. Sci. USA*, **89**, 2007–2011.

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**, 827–842.

Velculescu,V., Zhang,L., Vogelstein,B. and Kinzler,K. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Vignais,M., Huet,J., Buhler,J. and Sentenac,A. (1990) Contacts between the factor TUF and RPG sequences. *J. Biol. Chem.*, **265**, 14669–146674.

Vignais,M., Woundt,L., Wassenaar,G., Mager,W., Sentenac,A. and Planta,R. (1987) Specific binding of TUF factor to upstream activation sites of yeast ribosomal protein genes. *EMBO J.*, **6**, 1451–1457.