Article No. mb981947 *J. Mol. Biol.* (1998) **281**, 827–842

JMB

Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies

J. van Helden[1]*, B. André[2] and J. Collado-Vides[1]

[1]*Centro de Investigación sobre Fijación de Nitrógeno Universidad Nacional Autónoma de México, AP565A Cuernavaca, 62100, Morelos México*

[2]*Laboratoire de Physiologie Cellulaire et de Génétique des Levures, Université Libre de Bruxelles, Campus Plaine CP 244, Bd du Triomphe B-1050, Bruxelles, Belgium*

**Corresponding author*

We present here a simple and fast method allowing the isolation of DNA binding sites for transcription factors from families of coregulated genes, with results illustrated in *Saccharomyces cerevisiae*. Although conceptually simple, the algorithm proved efficient for extracting, from most of the yeast regulatory families analyzed, the upstream regulatory sequences which had been previously found by experimental analysis. Furthermore, putative new regulatory sites are predicted within upstream regions of several regulons. The method is based on the detection of over-represented oligonucleotides. A specificity of this approach is to define the statistical significance of a site based on tables of oligonucleotide frequencies observed in all non-coding sequences from the yeast genome. In contrast with heuristic methods, this oligonucleotide analysis is rigorous and exhaustive. Its range of detection is however limited to relatively simple patterns: short motifs with a highly conserved core. These features seem to be shared by a good number of regulatory sites in yeast. This, and similar methods, should be increasingly required to identify unknown regulatory elements within the numerous new coregulated families resulting from measurements of gene expression levels at the genomic scale. All tools described here are available on the web at the site http://copan.cifn.unam.mx/Computational_Biology/yeast-tools

Keywords: computational methods; functional genomics; transcriptional regulation; yeast; DNA microarray technology

Introduction

Despite the boom in genomic sequencing, computational analysis of regulatory sequences remains a relatively marginal domain. The vast majority of the programs available have been developed to measure structural relationships between coding sequences or between proteins. Non-coding regions represent, however, a striking interest for the biologist since they govern the regulation of gene expression. Regulatory profiles of known and unknown genes are already being determined experimentally at a genomic scale, thanks to the new technologies, such as DNA microarray technology (Schena *et al*., 1995, 1996; Schena, 1996; Goffeau, 1997; Strachan *et al*., 1997; Lashkari *et al*. 1997; DeRisi *et al*., 1997). Differential expression measurements will allow us to determine which set of genes respond at the transcriptional level to virtually any change in environmental conditions or to the controlled expression of any chosen transcriptional factor. Therefore, we will soon be confronted with countless families of coregulated genes sharing unknown regulatory sites. Detecting candidate elements responsible for this common behaviour represents an important challenge for the bioinformatician.

For several reasons, the computational detection of regulatory sites is a difficult problem, specially in eukaryotes: the consensus sequences recognized by transcriptional factors are generally much shorter than in prokaryotes, they can be quite variable,

Present address: J. van Helden, European Bioinformatics Institute. EMBL Outstation, EBI, Wellcome Trust Genome. Campus Hinxton, Cambridge CB10 1SD, UK

Abbreviations used: ORF, open reading frame; sig, significance index.

E-mail address of the corresponding author: jvanheld@ebi.ac.uk

0022–2836/98/350827–16 $30.00/0

and can be dispersed over very large distances. They are generally active in both orientations.

Given the flexibility of the regulatory mechanisms, one can hardly conceive a comprehensive method that could systematically detect all kinds of regulatory signals. Dedicated methods based on well defined models should allow us to unravel unknown elements of a common type. Based on this rationale, several programs have been developed that isolate unknown patterns shared by sets of functionally related DNA sequences (Waterman *et al*., 1984; Galas *et al*., 1985; Mengeritsky & Smith, 1987; Stormo & Hartzell, 1989; Hertz *et al*., 1990; Lawrence & Reilly, 1990; Cardon & Stormo, 1992; Lawrence *et al*., 1993; Neuwald *et al*., 1995; Hertz & Stormo, 1995; Wolfertstetter *et al*., 1996). These programs have each been inspired by a particular type of signal, and are generally highly efficient for the detection of elements that follow certain rules either in their internal organization and size, or in their location.

We present here a method inspired by the regulation of nitrogen metabolism in yeast. When cultured on a medium containing an optimal nitrogen source (glutamine, glutamate or ammonia), a series of genes that permit the utilization of alternative nitrogen sources (proline, arginine, gamma-aminobutyric acid, etc.) are down-regulated. All these nitrogen-repression-sensitive genes contain multiple copies of the so-called GATA boxes in their upstream regions (Magasanik, 1992). The GATA consensus box is short (five to six conserved nucleotides), and is found almost everywhere in the genome. However, a single isolated GATA box is insufficient to properly exert the regulatory function (Bysani *et al*., 1991). The specificity of GATA-box activity comes from its repetition in several copies within the upstream region of the controlled genes.

On basis of this regulatory paradigm, we designed a simple and fast algorithm that detects over-represented oligonucleotides within a group of upstream regions of coregulated genes. The program counts all oligonucleotide occurrences within the sequence set, and estimates their statistical significance. An essential prerequisite is that the system has to be calibrated to take into account the uneven oligonucleotide representation in the yeast genome, and more specifically in non-coding sequences. The algorithm proved highly efficient for the detection of upstream regulatory sequences previously found by experimental analysis in several yeast regulons beyond that one of nitrogen. In most of these cases, the already known upstream sequences were detected with a very high statistical significance. We also detected unknown motifs with a high statistical significance which are good candidates as new putative regulatory sites. A graphical tool was set up to further assess the interest of significantly over-represented oligonucleotides on the basis of positional considerations. A web interface has been implemented (http://copan.cifn.unam.mx/Computational_Biology/yeast-tools) allowing anyone to perform a similar analysis on new sets of coregulated genes. The principles of this methodology should be applicable to families of coregulated genes in other genomes, provided the adequate calibration is done.

Results

Regulatory families

Yeast metabolism has been widely studied and provides numerous examples of known regulons. In many cases the transcriptional factor involved in the common response is known, as well as its binding site. These families of coregulated genes provide ideal datasets to calibrate the method, which, in a further step, could be extended to families whose regulatory elements are unknown. We built several families on the basis of the above-mentioned criteria. Table 1 gives the composition of each family, as well as the criterion underlying its constitution. The genes were in all cases grouped together based on a genetic criterion, without *a priori* consideration on the content of their upstream regions. Our main source of information for family constitution was the yeast red book (Jones *et al*., 1992), along with some more recent articles for some of the families.

Two families, YAP and TUP, deserve some special attention because they were derived from the results from the first DNA microarray experiment on a genome scale (DeRisi *et al*. 1997). The YAP family regroups all the genes whose expression was induced more than twofold by the controlled expression of Yap1p. The TUP family was built on the basis of two criteria: on the one hand, all genes from this family are derepressed by a factor greater than 4 in a Tup1 deletion; on the other hand, they are induced by a factor greater than 2 during the diauxic shift. The reason for using a double condition is that Tup1p protein is not a DNA-binding factor, but is involved in transcriptional repression by interacting with a series of distinct DNA-binding factors. Selecting all the genes derepressed in the Tup1 mutant would lead to a family of 230 genes, grouping together the regulons for several transcriptional factors, involved in distinct metabolic pathways. One of these factors is Mig1p, which represses a series of genes when glucose is provided in the culture medium. This repression is released during the diauxic shift. Restricting the TUP family to the subset of genes which are also activated during the diauxic shift amounts to select the targets of Mig1p.

Contrary to all other families, YAP and TUP families each contain several genes of unknown function, for which the only current knowledge is the fact that they are activated by the Yap1p transcriptional factor, or repressed by Mig1, respectively.

Table 1. Composition of the regulatory families, and criterion underlying their constitution

Family	Genes	Common regulatory property	References
NIT	DAL5, DAL80, GAP1, MEP1, MEP2, MEP3, PUT4	Repressed when good nitrogen sources (glutamine, glutamate, ammonia) are present in the medium	Magasanik (1992)
MET	MET3, MET25, MET2, MET19, MET14, MET6, SAM1 SAM2, MET1, MET30, MUP3	Repressed by methionine	Hinnebusch (1992), Blaiseau *et al*. (1997)
PHO	PHO5, PHO11, PHO8, PHO84, PHO81	Repressed by P_i	Oshima *et al*. (1996)
PDR	YOR1, PDR11, PDR10, GAS1, STE6, SNQ2, PDR5	Pleiotropic drug resistance	Balzi & Goffeau (1995)
GAL	GAL1, GAL2, GAL7, GAL80, MEL1, GCY1	Activated when galactose is present and glucose absent from the culture medium	Johnston & Carlson (1992)
GCN	ARG1, ARG3, ARG4, ARG8, ARO3, ARO4, ARO7, CPA1, CPA2, GLN1, HIS1, HIS2, HIS3, HIS4, HIS5, HOM2, HOM3, HOM6, ILV1, ILV2, ILV5, LEU1, LEU2, LEU3, LEU4, LYS1, LYS2, LYS5, LYS9, MES1, MET14, MET3, MET6, TRP2, TRP3, TRP4, TRP5, THR1	General amino acid control; genes activated by Gcn4p.	Hinnebusch (1992)
INO	ACS2, CHO1, CHO2, FAS1, FAS2, FAS3, INO1, INO2, INO4, OPI3	Repressed in presence of inositol or choline	Patlauf *et al*. (1992)
HAP	PET9, ASN1, CIT1, COX5A, CYB2, HEM3, HMG1, SOD2	Targets of the Hap2/3/4/5p complex	
YAP	YNL331C, YKL071W, YFL056C, YLL060C, YOL165C, YCR107W, ATR1, FLR1, FRM2, YJR155W, OYE3, YLR460C, ECM4, OYE2, YML131W, MDH2	All genes induced more than twofold by the expression of Yap1p, in DNA microarray experiment on a genome scale	DeRisi *et al*. (1997)
TUP	FSP2, YNR073C, YOL157C, HXT15, SUC2, YNR071C, YDR533C, YEL070W, RNR2, YER067W, CWP1, YGR243W, YDR043C, YER096W, HXT6, YLR327C, YJL171C, YGR138C, HXT4, HXT7, GSY1, YOR389W, MAL31, YML131W, RCK1	All genes which are both – derepressed by a factor >4 when TUP1 is deleted, and – induced by a factor >2 during the diauxic shift	DeRisi *et al*. (1997)

Clusters of overlapping hexanucleotides generally reveal wider regulatory sequences

For each family shown in Table 1, we extracted the set of 800 bp upstream sequences, and performed an hexanucleotide analysis as described (see Methodology). All hexanucleotides with a significance coefficient (briefly "*sig*") higher than 0 were retained. The results of this analysis are summarized in Table 2. With the chosen threshold, very few sequences are retained (about ten per family, out of the 2080 possible hexanucleotide pairs). It is clear that in most families, the hexanucleotides with the highest significance coefficient (left side of the Table) correspond to the functional regulatory sites found by experimental analysis (right side). Highly significant patterns generally appear clustered with a few additional overlapping hexanucleotides that have a weaker significance coefficient. For instance, the most salient hexanucleotide of the MET family is CACGTG ($sig = 7.0$), which can be grouped with two strongly overlapping sequences: TCACGT ($sig = 6.1$) and GTCACG ($sig = 0.7$). When combined, the two most significant hexanucleotides correspond in fact to the 7 bp consensus sequence of the binding site for the Cbf1p-Met4p-Met28p complex (TCACGTG). In most families (MET, PHO, PDR, GCN, INO), the overlapping clusters reflect the fact that the recognition domain of the transcriptional factor is wider than six nucleotides. The maximum significance indicates the most conserved core that usually corresponds to the bases directly interacting with the transcriptional factor. The decrease of significance for the lateral overlaps comes from the fact that these positions are less crucial for the binding. In some cases however, clustering simply reflects a bias due to the high frequency of a shorter recognition site. This is observed in the nitrogen family, where GATAAG appears along with AGATAA and ATAAGA, although the binding site is not considered to exceed 6 bp.

Oligonucleotide size

In most families, the simple hexanucleotide analysis of the 800 bp upstream regions allowed us to detect the regulatory sequences previously found by means of experimental analysis. Analysis performed with different oligonucleotide sizes generally reveals the same patterns with different significance indices. Table 3 shows the variation of the maximal *sig* value associated to the principal sites from each family, as a function of oligonucleotide length used in the analysis. Each regulatory sequence shows a peak at a specific length, but most of them remain detectable within a reasonable range.

The oligonucleotide sequence associated with the peak of significance (bold in Table 3) fits well with the most conserved part of the previously described consensus (right part of Table 2). Note that a higher statistical significance only indicates a stronger over-representation, which does not necessarily correspond to a functional requirement

Table 2. Detection of regulatory sites by oligonucleotide analysis

Family	Hexanucleotide analysis result: Sequence	Ms	Occ	Exp	*sig*	Sites previously characterized: Consensus	Bound factors
NIT	ATAAGA	6	20	6.0	**2.0**		Gln3p, Nillp,
	GATAAG	6	26	3.0	9.1	GATAAG	Gzf3p, Uga43p
	AGATAA	7	17	6.1	0.4		(Zn finger)
	CTGATA	6	11	3.1	0.1		
	CCGCGC	2	6	0.7	0.8	–	–
	CGGCAC	4	6	0.8	0.5	–	–
	ACATCT	4	11	2.9	0.4	–	–
MET	CACGTG	9	26	2.0	**7.0**		Cbflp-Met4p-
	TCACGT	9	19	2.9	**6.1**	TCACGTG	Met28p complex
	GTCACG	6	8	1.4	0.7		(bHLH-bLZ-bLZ)
	TGTGGC	7	10	2.4	0.5		
	CTGTGG	8	11	2.1	**1.6**	AAAACTGTGG	Met31p, Met32p
	ACTGTG	9	12	3.2	0.6		(Zn finger)
	AACTGT	10	17	5.5	0.9		
	ATATAT	19	82	42.3	0.8	–	–
	TATATA	11	80	43.9	0.2		
	GCTTCC	7	12	3.5	0.2	–	–
PHO	CGCACG	5	6	0.5	**1.5**		
	GCACGT	5	10	0.8	**4.4**		Pho4p
	CACGTG	5	12	0.9	**1.8**	GCACGTGGG	(bHLH)
	ACGTGG	5	8	0.7	**2.8**	(high affinity)	
	CGTGGG	3	5	0.5	0.5		
	CACGTT	5	7	1.2	0.3	GCACGTTTT	Pho4p
	ACGTTT	5	11	2.6	0.8	(medium affinity)	(bHLH)
	CTGCAC	4	8	1.0	**1.7**	–	–
	TGCCAA	4	12	2.0	**2.6**	–	–
PDR	TCCGTG	5	8	1.1	**1.4**		
	CCGTGG	4	12	1.1	**7.4**		
	CGTGGA	5	10	1.1	**3.3**		
	GTGGAA	6	11	2.8	0.5		
	TCCGCGG	3	10	0.8	**4.5**		Pdrlp, Pdr3p
	CCGCGG	2	12	0.6	**2.6**	TCCGCGGA	(Zn_2Cys_6
	CGCGGA	3	10	0.8	**4.5**		binuclear
	CTGCGG	2	6	0.9	0.2		cluster)
	GCGCGA	5	6	0.8	0.6	–	–
	AGGCAC	3	7	1.3	0.1	–	
	GGCACC	5	6	0.9	0.2		–
GAL	–	-	-	-	-	$CGGN_5WN_5CCG$	Gal4p (Zn_2Cys_6 binuclear cluster)
GCN	AGTGAC	17	25	10.4	0.7		
	GTGACT	16	25	9.4	**1.4**		
	ATGACT	26	42	16.1	**3.9**		
	TGACTC	26	44	7.3	**8.0**	RRTGACTCTTT	Gcn4p
	GACTCA	22	26	6.9	**4.3**		(bZip)
	GACTCT	14	20	7.5	0.6		
	CAGCGG	16	23	5.8	**4.0**	–	–
	AACCGG	10	15	5.3	0.2	–	–
	ACCGGC	9	13	3.0	**1.4**		
	CATCGA	16	26	10.9	0.8	–	–
	ATCGAA	23	33	16.8	0.2	–	–
	AGAGAG	21	32	16.1	0.2		
INO	CAACAA	9	28	8.5	**3.6**	Unknown	Unknown
	AACAAC	7	18	6.5	0.5		
	AACAAG	8	19	7.1	0.5		
	CATGTG	9	15	2.7	**3.3**	CATGTGAAWT	Ino2p/Opi1p (bHLH/leucine
	TGTGAA	9	15	4.9	0.4		zipper)

Table 2—*Continued*

Family	Hexanucleotide analysis result: Sequence	Ms	Occ	Exp	*sig*	Sites previously characterized: Consensus	Bound factors
	TCTTCA	9	21	7.6	**1.0**	–	–
	GTTCAA	8	16	5.3	0.5	–	–
	GTCGCA	7	8	1.5	0.5	–	–
HAP	–	–	–	–	–	CCAA$^{T}/_{C}$	Hap2/3/4/5p
	AGAGAG	4	14	3.4	**1.6**		
	GAGAGA	5	16	3.4	2.8	–	–
	ATGGGG	6	7	1.3	0.2	–	–
	TGGGGC	4	6	0.8	0.3		
YAP	CGTTCC	9	15	2.7	**3.3**	–	–
	TTCCGT	7	16	6.0	0.0		
	caTTAC	11	21	7.6	**1.0**	TTACTAA	Yap1p (bzip)
	CTGAAG	11	17	5.5	0.9	–	–
TUP	GGGGTC	9	9	2.0	0.2		
	GGGGTA	12	26	3.5	**8.6**	KANWWWWATSYGGGGW	Mig1p
	TGGGGT	15	19	4.2	**3.7**		(Zn finger)
	CGGGGT	10	14	2.3	**3.5**		
	AGGGGC	12	14	3.3	**1.6**		
	GTGGGG	14	21	3.2	**7.0**		
	GCGGGG	12	12	2.3	**2.0**		
	CAGGGG	9	11	2.4	**1.0**		
	CGTGGG	11	12	2.5	**1.6**		
	TTTGTG	16	33	15.0	**1.0**		
	CACGGG	8	9	1.6	**1.0**		
	GCACGG	9	11	3.0	0.2		
	GGCACG	7	11	2.4	**1.0**		
	AGGCAC	13	21	4.6	**4.3**	–	–
	AAAAAA	25	305	226.6	**3.1**	–	–
	AAAGAA	25	92	62.8	0.2		
	GGAGGA	14	16	5.8	0.1		
	AGGAGG	12	16	5.3	0.6	–	–
	AAGGAG	16	24	10.4	0.4		
	CAAACA	19	33	16.4	0.4	–	–
	ACAAAC	17	30	14.2	0.4		
	CTCCGC	8	11	3.0	0.3	–	–
	TCTGCA	13	22	9.6	0.1	–	–
	CCTGCA	12	15	5.4	0.0		
	CGTAGC	9	11	3.1	0.1	–	–

For each family, all hexanucleotides with a positive significance coefficient ($sig \geq 0$) are indicated. Significance indices higher than 1 are highlighted in bold. Hexanucleotides are clustered by sequence similarity. Substitutions within a cluster are underlined. The 2 last columns show the sites previously characterized. Note that the highest *sig* value from each family generally corresponds to the site experimentally described. Abbreviations: ms, matching sequences, i.e. the number of sequences from the family which contain at least one occurrence of the pattern, occ, number of occurrences of the pattern among all upstream regions from the family, exp, expected number of occurrences; *sig*, significance index, calculated as defined in Methodology; bHLH, basic helix-loop-helix motif; bZip, basic leucine zipper.

on size. For instance, in the NIT family, the significance of GATAA ($sig = 3.5$) is lower than that of GATAAG (9.1). However, the PUT4 upstream region does not contain any copy of the GATAAG pattern, though we know it responds to changes in nitrogen conditions (Jauniaux *et al.*, 1987). The response elements for this gene are two GATAA sites not followed by a G (one GATAAA and one GATAAC). Moreover, in the pentanucleotide analysis of the NIT family, ATAAG ($sig = 4.1$) appears with a higher *sig* value than GATAA (3.5), although the real functional core is GATAA. Reciprocally, a variant of the gata box, GATTA, does not appear over-represented in this family, even though it is known to be functional (but less efficient than the canonical GATAA: Magasanik, 1992). In summary, the highest significance index reflects a family property, whereas the regulation of individual genes can rely on variants, even if these are statistically less significant.

Even though most regulatory sites are already detected with the simple hexanucleotide analysis, a systematic scanning of various oligonucleotide lengths can be crucial for some of them, as shown for the YAP family. Hexanucleotide analysis only revealed one pattern (caTTAC) which partly overlaps the Yap1p binding site (TTACTAA). The most significant pattern from the heptanucleotide anal-

Table 3. Significance index as a function of oligonucleotide length

Family	Pattern	Oligonucleotide length 4	5	6	7	8	9
NIT	a**GATAAG**a	1.8	4.1	**9.1**	4.6	0.9	–
MET	g**TCACGTG**	4.4	4.1	7.0	**8.2**	3.2	–
	AAACTGTGg	1.5	2.3	1.6	4.8	**5.2**	4.9
PHO	**CACGT**ggg	4.7	**8.4**	4.4	4.3	4.3	–
	a**TGCCAA**	2.6	1.5	**2.6**	0.6	–	–
	CTGCAC	–	–	**1.7**	–	–	–
INO	**CAACAA**g	2.9	2.1	**3.7**	1.3	–	–
	c**CATGTGAA**	–	–	2.7	3.2	**6.4**	0.4
PDR	t**CCGTGG**a	1.5	3.3	**7.4**	6.9	4.2	1.4
	t**CCGCG**ga	6.9	**7.1**	4.5	5.6	1.8	1.0
GCN	gt**GACTC**a	5.4	**8.8**	8.2	7.7	4.7	–
	CAGCGGa	3.3	3.5	**4.0**	0.6	–	–
YAP	**CATTACTAA**	–	–	1.0	2.3	2.1	**3.2**
	c**CGTTCC**	0.1	0.5	**3.3**	0.3	–	–
YAP	ca**TTACTAA**	–	–	0.7	**4.5**	2.5	3.5
400 bp	c**CGTTCC**	0.8	0.5	**2.4**	0.7	0.2	–
TUP	gt**GGGG**ta	**10.1**	9.0	8.6	5.6	3.0	–
	cat**AGGCAC**	3.3	3.3	**4.3**	2.6	3.3	1.7

The highest *sig* value associated to each pattern for each oligonucleotide size is shown. Only positive *sig* values are shown. Each pattern shows a peak at a specific length (shown in bold). The fragment of the pattern associated to the peak is indicated in bold and uppercase letters. The row labeled YAP 400 bp shows the significance obtained with upstream regions of 400 bp. All other results were obtained with 800 bp upstream regions.

ysis was TTACTAA (*sig* = 2.3), i.e. the exact Yap1p binding site. Extending oligonucleotide size up to nine reveals the extended pattern CATTACTAA, including simultaneously the most significant hexanucleotide and heptanucleotide. Another interesting property is that the Yap1p characteristic heptanucleotide appears with a much higher significance when the analysis is performed on shorter upstream regions (400 bp). In some cases, it is thus worth trying variations on the upstream region size to detect binding sites with a preferential proximal location.

Multiple clusters reveal either multiple sites or single site variability

Several clusters generally appear from each family. In some cases, multiple clusters correspond to distinct regulatory sites. This is clearly the case in the MET family, where two independent clusters are detected, the former forming the pattern GTCACGTG, corresponding to the binding site for the Gbf1p-Met4p-Met28p complex (Hinnebusch, 1992), and the second revealing the site AAAACTGTGG, recognized by Met31p and Met32p (Blaiseau *et al.*, 1997). Alternatively, clusters can be structurally related, and represent variants of the same binding site, as in the PHO family, where one cluster forms the pattern GCACGTGGG, shown to bind Pho4p with high affinity, and another cluster defines the low affinity consensus GCACGTTTT for this same transcriptional factor (Oshima *et al.*, 1996). A similar situation is observed in the PDR family, where one cluster forms the Pdr1p binding site TCCGCGGA (Balsi & Goffeau, 1995), and another cluster forms the TCCGTGGA variant.

Unknown sites revealed by oligonucleotide analysis

Some additional hexanucleotides are identified within each family besides those belonging to known regulatory sequences. Based on the results for known sites, one can infer that the ideal unknown site should appear as a cluster of overlapping hexanucleotides with a core showing a high significance coefficient. Several unknown patterns extracted from the hexanucleotide analysis fit with these criteria, and are good candidates as new regulatory sequences.

A cluster of four hexanucleotides appears in the TUP family, forming the pattern AGGCACGG, with a maximal significance index of 4.3. Varying oligonucleotide size reveals the same pattern with a flanking CAT at left (CATAGGCAC). The GCN family also contains an unknown pattern, CAGCGG with a high significance index (*sig* = 4.0). In the PHO family, we detected the TGCCAA hexanucleotide with a good significance coefficient (2.6). This pattern is particularly represented in the PHO5 and PHO84 upstream region, where it occurs five and four times, respectively. Interestingly, the feature map (Figure 1) shows that a pair of TGCCAA sites are found at a more-or-less conserved position (−204 to −133) in three upstream regions of the family (PHO11, PHO84 and PHO8).

We have no clue about a possible regulatory function for these motifs, but some of them could reveal binding sites for an additional unknown regulatory factor, in addition to the known factor for this family. Such a coordinated regulation has already been demonstrated experimentally in case of the MET family, and it would thus not be surprising to observe it in some other families.

One pattern from the INO family fits perfectly the above criteria, but should be taken with caution. The analysis of the INO family shows an unknown hexanucleotide, CAACAA, with an even higher statistical significance (3.6) than the known CATGTGAA site (Paltauf *et al.*, 1992). CAACAA appears clustered with two weakly significant overlaps (AACAAC, AACAAG), forming the pattern CAACAA$^{C}/_{G}$. This sequence has not been detected by experimental analysis (Susan Henry, personal communication). It is, however, highly over-represented: 28 occurrences of CAACAA are observed within nine among the ten genes of the INO family. Interestingly, the pattern is especially concentrated in the upstream regions of the FAS genes (four occurrences in FAS1, six in FAS2 and five in FAS3). Most occurrences are located distally (between -365 and −800). This specific concentration in the FAS genes could suggest a possible involvement of CAACAA in the response of these genes to fatty acids, for which the regulatory elements are unknown. However, four of the six occurrences observed in FAS2 upstream sequence are located within a valine-rich coding region from the upstream gene SSO1. This, of course, does not prevent them from exerting simultaneously a regulatory effect on FAS2, but their statistical significance has to be re-evaluated taking into account their coding function. To allow this, we implemented a "non-ORF overlap" option in our program for upstream sequence extraction. On the other hand, restricting the analysis to the non-coding upstream sequence could be misleading in the opposite way, leading to the loss of some important elements. This is particularly well illustrated by the PHO family, where three genes have a very closely located upstream ORF (at 181 bp from PHO8, 238 from PHO84, and 598 from PHO11, respectively). No less than seven of the ten sites protected by Pho4p (Oshima *et al.*, 1996) are located within (five sites, see asterisks in Figure 1) or even upstream from (two sites) the neighbor predicted ORF. Consequently, when an hexanucleotide analysis is performed on the non-coding upstream sequences from this family, the CACGTG and CACGTT are not detected. The best way to proceed is probably to perform the analysis with fixed size upstream regions, but keeping in mind additional possible biases due to the coding sequences. Eventually, the regulatory activity of the unknown patterns detected will only be assessed by experimental testing.

Positional clues

As illustrated in the previous section, the feature map is helpful to highlight the interest of putative regulatory motifs on the basis of positional considerations. Two kinds of information can be considered: the conserved position of a motif across all sequences of the family; and the concentration of one signal within one or within a subset of the upstream sequences.

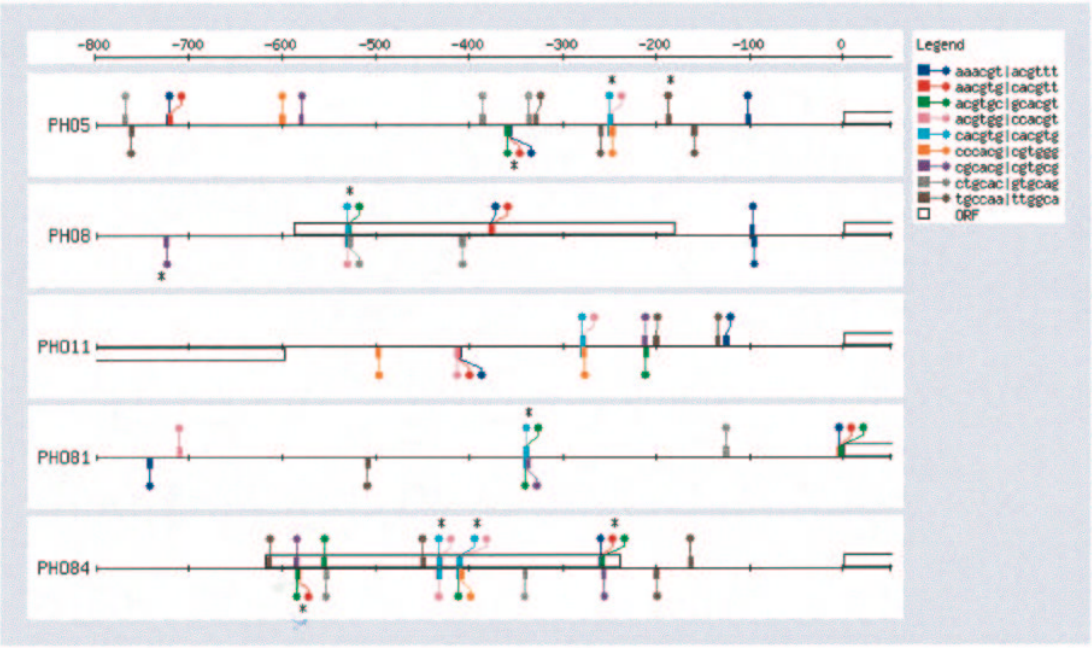

Figure 1. Feature map of over-represented hexanucleotides in the 800 bp upstream regions from the PHO genes. The scale bar provides coordinates relative to the ORF start. A specific color is associated to each hexanucleotide. The sites previously found to be bound by Pho4p are marked with an asterisk. Note the fact that five of the known Pho4p binding sites are located within predicted ORFs (empty boxes). Note also the pair of TGCCAA motifs in the −100 to −200 regions from PHO11, PHO84 and PHO5, as well as high number of occurrences of this element in the PHO5 and PHO84 upstream region.

The same kind of positional considerations can be applied to known binding sites, as illustrated by the CTGCGG motif in the PDR family. This variant of the CCGCGG core appears with a low significance coefficient (0.2). However, the feature map reveals that five of the six occurrences are concentrated in the upstream region of a single gene (PDR11), which on the other hand, does not contain any occurrence of the CCGCGG core. This variant could reflect some regulatory specificity of PDR11.

The statistical significance of these positional inhomogeneities could certainly be assessed by more elaborate statistical tools. We feel, however, that place should be left for the biologist's intuition, helped by his knowledge of regulatory families. The feature map provides a visual support for this intuition.

Known sites escaping detection

Oligonucleotide analysis enabled us to detect nine known regulatory elements among eight out of ten regulatory families. Two elements escaped detection: the binding sites for Gal4p in the GAL family, and for the Hap complex in the HAP family.

Contrary to all other families, not a single hexanucleotide had a positive significance coefficient in the GAL family. Penta- and tetranucleotide analyses revealed some patterns with a low significance, none of them reaching the value 1 for *sig* (not shown). Trinucleotide analysis reveals two clearly over-represented sequences: CCG (*sig* = 2.2) and GCC (*sig* = 2.1). Transcriptional regulation of the GAL genes is mediated by Gal4p, which binds to dyads of GC-rich trinucleotides separated by an 11 bp sequence with high internal variation. Trinucleotide analysis pointed out these GC-rich elements, but the complete Gal4p consensus could not be extracted from this analysis. This suggests that the oligonucleotide analysis is not appropriate for these kind of patterns, and points out the need for an algorithm based on the detection of spaced pairs of oligonucleotides.

The second element that escaped detection is the binding site for the Hap complex. This complex is formed by four proteins (Hap2p, Hap3p, Hap4p and Hap5p). Hap2p and Hap3p have been shown to bind DNA, while Hap4p acts as activation domain for the complex. The Hap2p-Hap3p binding site contains a conserved motif $CCAA^{T}/_{C}$ present in several copies within each upstream region of the family. We could not detect this motif under any condition of the analysis. Despite the fact that $CCAA^{T}/_{C}$ boxes are repeated in four occurrences per sequence on average, their statistical significance remains well below our threshold (−1.6 for CCAAC and −1.3 for CCAAT), and several unrelated patterns appear with a higher significance. Note that the effective binding site for the Hap complex contains a second essential region, 10 bp apart from the $CCAA^{T}/_{C}$ motif (Guarente, 1992). We are currently developing a program which would specifically detect spaced dyads as found in these two regulatory systems.

Importance of the expected frequency calibration

All our analysis was done using oligonucleotide-specific expected frequencies. As an estimate for these expected frequencies, we used the frequency observed for each oligonucleotide in the set of all non-coding sequences from the whole genome. This choice is fully justified *a posteriori* by the fact that in all families, the highest significance indices correspond to functional regulatory elements. This probabilistic model is further validated by the very low level of noise observed. Indeed, a very limited number of patterns are isolated from each family, and their majority belongs to the regulatory sequences previously characterized. The same signals would have been undetectable without frequency calibration. We illustrate this for the MET family in Table 4, where we show the rank at which the patterns selected from Table 2 would have emerged if all hexanucleotides had been considered equiprobable (column r2). For instance, the CACGTG core, which had the highest significance index for the MET family in Table 2 would have appeared at the 23rd position. Even worse, the CTGTGG core, selected at third position with the frequency tables, would have been the 99th pattern otherwise. In other words, the binding site for Met31p would have been undetectable. An acceptable alternative would have been to calculate expected frequencies based on the alphabet usage (i.e. with A and T probabilities of 0.31, C and G probabilities of 0.19). In this case, the regulatory signals would have appeared with a reasonable rank (column r3), though intermingled with many non-specific hexanucleotides (AAAAAA, AAAAAG, ...).

Table 4. Importance of the use of oligonucleotide-specific expected frequencies

Family	Pattern	r1	r2	r3
MET	**CACGTG**	**1**	**23**	**5**
	TCACGT	2	21	6
	GTCACG	6	259	24
	TGTGGC	8	131	12
	CTGTGG	**3**	**99**	**8**
	ACTGTG	7	86	17
	AACTGT	4	32	18
	ATATAT	5	3	2
	TATATA	10	1	3
	GCTTCC	9	75	7

We compared the rank of all hexanucleotides from Table 2 (r1) with the rank they would have if hexanucleotides had been considered equiprobable (column r2) or with an estimation of hexanucleotide probabilities on the basis of yeast non-coding sequence alphabet (r3). The cores of the two sites previously characterized by experimentation are highlighted in bold.

Gene *versus* family

In these oligonucleotide analyses, we used families containing from 5 to 38 genes. There is no methodological constraint on the number of genes per family, so, in principle, one could analyze smaller or larger families. The limit case is to consider each gene as a separate family, and to evaluate whether its upstream region contains over-represented elements. Since multiple copies of a regulatory sequence are commonly found in the upstream region of yeast genes, it is conceivable that some functional elements could be isolated from the single gene analysis. We performed a systematical hexanucleotide analysis of the upstream regions from the 6219 predicted genes from the yeast genome (the complete results can be consulted on the web site). In a fraction of the genes, significant patterns are detected. We did not analyze exhaustively these results, but in some cases at least the over-represented hexanucleotides correspond to known regulatory elements.

These data were used for a more detailed analysis of the genes from Table 1. We compared the hexanucleotides extracted from the gene-by-gene analysis with those resulting from the family analysis. We found that patterns that are significant for a family are not necessarily significant for the individual genes. A specially illustrative example comes from the MET family, where CACGTG has a very high significance coefficient (7.0), though it does not appear from the analysis of any single gene (individual *sig* value ranging between −0.9 and −2.2). In other cases, the family-specific pattern is significant in a fraction of the family, but its significance is lower in any gene taken alone. This is the case for the GATAAG hexanucleotide which has a significance coefficient of 9.1 for the whole nitrogen family. Gene-by-gene analysis detects it in MEP2 ($sig = 3.1$) and with a weaker score in MEP3 and GAP1 ($sig = 0.7$). In the four other genes of the family it has a negative value ($sig = -1.3$), although it is known to be functional in these genes.

Reciprocally, some motifs that were not part of the initial families appear with significant scores in individual genes. An interesting motif, ATACGA, is detected with a significance of 3.3 in the upstream sequence of FAS3, where it occurs seven times, five of which belong to a stretch of tandem repeats of the pentanucleotide ATACG. Experiments should permit us to determine whether this site in tandem is involved in the transcriptional regulation of FAS3.

The high significance observed in regulatory families for functional patterns are thus a group property, which cannot be inferred from the analysis of individual regulatory regions. A gene-by-gene analysis provides no guarantee to extract the functional regulatory elements from upstream sequences. It is, however, worth remembering that it can in some cases highlight a specific regulatory property for an individual gene.

Whole genome scanning for putative nitrogen-sensitive genes

After having determined the upstream sequences responsible for a regulatory response, one would naturally like to scan the whole yeast genome for potentially coregulated genes. As soon as the complete sequence of yeast chromosome III has been available, Fondrat & Kalogeropoulos (1994) published predictions of genes potentially regulated by various known transcriptional factors. An important insight of their work is that specific rules have to be established for each transcriptional factor, taking into account their regulatory characteristics.

A classical way to predict genes putatively regulated by a given transcriptional factor is to scan the genomic database for matches with the degenerate consensus of its binding site. Position weight matrices can also be used when the variability of the binding site is sufficiently documented. One limitation of these approaches is that short sites are likely to be found in almost every upstream region. For instance, scanning the set of 800 bp upstream regions from the 6219 predicted ORFs for the pattern GATAAG would lead to 2086 candidate genes. Even worse, the GATAA motif, shown functional for the response to nitrogen, is found in 5618 upstream regions. We can, however, build rules taking into account more information coming from experimental results, i.e. the knowledge that the highest sensitivity to nitrogen is generally associated to a high number of repetitions of the GATA boxes. Moreover, the majority of the GATA boxes for which a regulatory activity has been shown are localized within the 500 bp upstream of the ORF start.

We tested different conditions for counting GATA boxes occurrences. Either 800 bp or 500 bp was used as an upstream region size. Three motifs were tested: GATAAG (the most significant pattern of our analysis), GATAA (the common GATA box description), and GATWA (taking into account the occasional GATTA variant). For each condition, we counted the occurrences of the motifs within the upstream region from each of the 6219 predicted ORFs. Genes were then sorted according to the number of occurrences. We compared the selectivity for each set of conditions, i.e. the capacity to select in the top of the list a high number of genes known to be nitrogen-sensitive.

The best results were obtained with the GATAAG motif and a 500 bp upstream region size. Three or more occurrences of this motif (Table 5) were observed in 34 upstream regions, of which 23 belong to a gene with a known function. Twenty of these 23 genes are known to be regulated by nitrogen. Given this high proportion, it is highly probable that several of the 11 unknown genes also belong to the metabolism of nitrogen. It should be easy to test experimentally the response of these genes to nitrogen. Two GATAAG occurrences were found in 184 additional upstream regions, among which 68 belong to genes with a

Table 5. Whole genome scanning for all genes containing three or more occurrences of GATAAG in their 500 bp upstream region

	ORF	Locus	occ	Function
A. *Involved in nitrogen metabolism*				
1	YNL142W	MEP2	7	High affinity low capacity ammonia permease
2	YPR138C	MEP3	5	Low affinity high capacity ammonia permease
3	YIR032C	DAL3	5	Ureidoglycolate hydrolase
4	YKR039W	GAP1	4	General amino acid permease
5	YIR028W	DAL4	4	Allantoin permease
6	YFL021W	GAT1	4	Transcription factor for nitrogen regulation
7	YER060W	FCY21	4	Purine-cytosine permease
8	YLR160C	ASP3D	3	L-Asparaginase II
9	YLR158C	ASP3C	3	L-Asparaginase II
10	YLR157C	ASP3B	3	L-Asparaginase II
11	YLR155C	ASP3A	3	L-Asparaginase II
12	YKR034W	DAL80	3	Transcriptional repressor for several nitrogen regulated genes
13	YJR152W	DAL5	3	Allantoate permease
14	YJL110C	GZF3	3	Transcriptional repressor for several nitrogen regulated genes
15	YIR031C	DAL7	3	Malate synthase 2
16	YIR027C	DAL1	3	Allantoinase
17	YEL063C	CAN1	3	Arginine permease
18	YDL210W	UGA4	3	GABA-specific high-affinity permease
19	YBR208C	DUR1,2	3	Urea amidolyase
20	YAL062W	GDH3	3	NADP-glutamate dehydrogenase
B. *Supposedly not regulated by nitrogen*				
1	YLL039C	UBI4	4	Ubiquitin
2	YKR031C	SPO14	3	Phospholipase D
3	YGR082W	TOM20	3	Mitochondrial outer membrane import receptor subunit, 20 kDa
C. *Unknown genes*				
1	YIR033W		5	Similarity to Spt23p
2	YOL019W		4	Similarity to Rim9p and YFR012w
3	YDL237W		4	Hypothetical protein
4	YPL150W		3	Similarity to Ser/Thr protein kinases
5	YOL128C		3	Strong similarity to protein kinase Mck1p
6	YIL146C	ECM37	3	Hypothetical protein
7	YGR081C		3	Weak similarity to mammalian myosin heavy chain
8	YDL238C		3	Hypothetical protein
9	YDL068W		3	Questionable ORF
10	YBL077W		3	Questionable ORF
11	YBL049W		3	Hypothetical protein

The column labeled occ shows the number of occurrences (occ).

known function. Twelve of these 68 genes are regulated by nitrogen. Additional criteria could be considered to increase the selectivity among the genes with two occurrences of GATAAG. On the one hand, we could take into account the presence of other GATAA motifs which are not terminated by a G. Another interesting possibility would be to select those in which at least two GATA boxes are closely associated, since the characterized GATA boxes appear generally clustered.

One could use a similar approach in cases where two distinct elements are known to be involved in the same metabolic pathway (as seen in the MET family), by adding the constraint of a simultaneous presence of both elements. Positional clues could also be used, for instance in cases where a repressing sequence (URS) is known to be active only when located downstream of the activating sequences (UAS). Specific rules should thus be established for each regulatory model, taking into consideration the repetitions, combinations, and positional specificities of the known regulatory elements. The definitions of these rules and the estimation of the value of the predictions require an expert knowledge for each particular regulon. Our feeling is that in the future the emphasis should be put on incorporation of most available biological information rather than on the development of elaborated statistical methods.

Comparison with related methods

Several methods have been published for the detection of unknown elements within a set of functionally related sequences. Each of these methods has been designed on basis of a precise biological model. Several algorithms (Waterman *et al*., 1984; Galas *et al*., 1985; Mengeritsky & Smith, 1987) allow the detection of elements with high internal variation, such as the TATA box, or the −35 box. An essential prerequisite of these methods is that the regulatory elements have to share a conserved position relative to a common reference (e.g. the transcription start). They are thus well adapted to the analysis of prokaryote promoters, but would be inappropriate for our regulatory model.

Three other programs (Coresearch, Consensus and the Gibbs sampler) are publicly available and we could directly assess their efficiency on our regulatory families. Coresearch (Wolfertstetter *et al*., 1996) detects short over-represented sequences, and in a second step extends the pattern laterally. It failed to detect the GATA boxes, for two reasons: it does not allow to specify an unequal alphabet utilization, and the scores do not take into consideration the repetitions of a pattern within the same sequence.

Consensus (Stormo & Hartzell, 1989; Hertz *et al*., 1990) overcomes both of these problems: an arbitrary alphabet utilization can be specified, as well as an expected number of matches for the searched pattern. Specifying a matrix length of six and an expected number of matches of 35, consensus isolated the GATAAG motif from the NIT family. It also succeeded in the isolation of regulatory elements from our other regulatory families, including the binding site for Gal4p in the GAL family (with a matrix length of 17), which had escaped detection by oligonucleotide analysis. One drawback is that Consensus isolates a single element from each family (optionally, several matrices can be retained, but they are variants of the same regulatory element). It is thus not optimal for the detection of multiple elements shared by the same set of sequences. In the case of the MET family, this inconvenience could, however, be circumvented by playing with the matrix length: the Cbf1p-Met4p-Met28p binding site (CACGTG) appeared as the most significant matrix of length six, whilst the Met31p binding site (AAACTGTGG) appeared when matrix length was set to nine. Another drawback is that the isolation of all these regulatory elements required the *a priori* specification of two parameters: matrix length and expected number of matches. This second parameter can be crucial for the isolation of some elements. For instance in the NIT family, GATAAG is obtained when the expected number of matches is set to 35, but escapes detection when this parameter is set to 15. Therefore, the user of Consensus has to scan a two-dimensional parameter space (pattern length and expected occurrences), instead of a one-dimensional space (oligonucleotide length) with oligonucleotide analysis. Consensus is also quite slower than oligonucleotide analysis (several minutes of processing are required for each family, instead of a few seconds with our method).

The Gibbs sampler (Lawrence *et al*., 1993; Neuwald *et al*., 1995) can detect shared motifs in either proteic or nucleic acid sequences, with or without gaps. With a simple configuration (length = 6, expected number of occurrences = 15), the Gibbs sampler detected the same over-represented hexanucleotides as our method did. Moreover, increasing the pattern length to 17 allowed to isolate the Gal4p binding site, which our method failed to detect. Only the HAP site escaped detection by the Gibbs sampler, as it did with our method. Gibbs sampler is thus more sensitive than oligonucleotide analysis. However, it is essentially an heuristic method, and is not exhaustive. The program starts with a set of random positions and then converges towards an optimal matrix. Depending on the initial positions, the program can be attracted by distinct local maxima. In order to validate the significance of a pattern, it is thus necessary to run it repeatedly starting from different initial conditions. We had to run 100 repetitions for each family to extract the same range of patterns as with oligonucleotide analysis. Under these conditions, Gibbs sampler is much slower (20 minutes per family) than oligonucleotide analysis (ten seconds per family). Some important patterns (CATGTG in the INO family) were detected with a weak frequency, due to the attraction of a slightly more significant feature (CAACAA in the same family). There is thus a risk to miss some important regulatory features when they are masked by highly attractive local maxima.

In summary, Consensus and the Gibbs sampler are slower than oligonucleotide analysis and require a more precise description of the characteristics of the pattern searched (i.e. an expected number of matches), but both programs are much more efficient for the description of large motifs with higher internal variation. A good trade could be to use oligonucleotide analysis for a systematic and fast scan of simple patterns, and Consensus or the Gibbs sampler to refine the description of the patterns detected by oligonucleotide analysis or to detect other types of patterns.

Conclusions and Perspectives

Based on the model of the regulation of nitrogen metabolism, we developed an algorithm that detects over-represented oligonucleotides within the upstream sequences from a family of coregulated genes. In most of the families here analyzed, hexanucleotide analysis detected the main motifs involved in transcriptional regulation. Several unknown signals were detected, some of which present a high significance index and are good candidates as putative regulatory sites. This does not mean that the algorithm would detect all kinds of regulatory signals. For instance, the Gal4p binding site could not be detected due to its complex structure (a pair of trinucleotides separated by a non-conserved sequence). Similarly, HAP sites escaped detection due to the fact that they are not properly over-represented. In both of these cases, the analysis failed to show any signal with high significance. One can conclude that when a signal is selected as highly over-represented, it is very likely to correspond to a functional regulatory site, although the opposite is not true.

The main originality of our method, and the principal reason for its efficiency, is the use of calibrated tables with a specific estimation for each oligonucleotide expected frequency. Similar frequency tables had already been used in

another context, for the discriminant intron-exon analysis (Claverie & Bougueleret, 1986). Another distinctive feature of our approach is the use of the total count of occurrences instead of using the number of matching sequences, as most other programs do. This is justified by the fact that our biological paradigm relies on the repetition of the regulatory signal within each regulatory sequence. Our implementation also allows to assess the significance of the number of matching sequences (i.e. the number of sequences presenting at least one occurrence of the oligonucleotide), but in most families (except GCN), this option did not prove efficient in isolating regulatory elements. Indeed, due to the small number of sequences per family and to the large size of upstream regions, the fact to encounter a pattern at least once within each sequence is generally not statistically significant.

The most appealing feature of oligonucleotide analysis is its simplicity. The whole statistical analysis relies on the binomial formula. A direct consequence is the speed of execution: a family of ten genes is analyzed in ten seconds. The processing time increases linearly with the number of genes, and is weakly sensitive to oligonucleotide length. The method is straightforward (there is no iterative process), rigorous (as opposed to heuristic), and exhaustive (with a single run of the program, all over-represented patterns of the chosen size are detected). The cost of the methodological simplicity is that our method is limited to the detection of short and relatively conserved motifs: regulatory sites with a highly conserved core of five to eight nucleotides. Larger conserved patterns can, however, be reconstituted by the combination of overlapping oligonucleotides. Weak substitutive variations of the core (one variable position) can also be detected, as shown with the PHO and PDR families. Sophistication could of course be added for the detection of patterns with more internal variation. Some such options are in fact implemented: the program can include an IUPAC degenerate nucleotide code at any position, but this did not improve efficiency in the cases analyzed here. On the contrary, the significance of the functional signals is lowered by their grouping with non-over-represented variants (not shown).

The hexanucleotide analysis allowed us to detect the regulatory elements from most families. A systematic analysis of various oligonucleotide lengths (between five and nine) provides, however, a more accurate description of the pattern, indicating the most significant oligonucleotide length and sequence. An alternative way to extend the description of the putative regulatory sites would be to search within the sequence set for all matches with a short cluster core (for example the hexanucleotide showing the highest significance), and to extract them accompanied by their flanking sequences. These extended matching sequences could then be used to build a position weight matrix, as is done by the program Coresearch (Wolfertstetter *et al*., 1996).

Besides the detection of regulatory sites in families of coregulated genes, counting of oligonucleotide occurrences was used to scan the whole genome for potential nitrogen-sensitive genes. This allowed us to select a very small number of genes presenting a high probability to be regulated by nitrogen. The next step will be to validate experimentally the transcriptional response of these candidates. Another experimental avenue to follow after detection of unknown regulatory sites, is the isolation of the related transcriptional factors. This can be done experimentally, for instance by the one-hybrid method (see Blaiseau *et al*. (1997) for an application to the MET family).

How far can oligonucleotide analysis be extended to other organisms? It would probably be of weak efficiency in prokaryotes, where regulatory signals are large, and show a higher variation. On the contrary, there is good hope that it will be applicable to detect regulatory elements in higher eukaryotes, where short repeated signals are common. This feature has already been exploited for the automatic detection of periodic signals within single contiguous sequences (Bodnar & Ward, 1987). Some of these signals correspond to specific regulatory elements, while others are thought to participate in global structural properties of the DNA. Calibrated frequency tables could become even more crucial in higher organisms, where numerous cases of over-represented oligonucleotides have been shown.

The results obtained with the YAP and TUP families show that oligonucleotide analysis can be used for families derived from DNA microarray experiments. Some care should, however, be taken in the definition of the gene family. Indeed, the response to a change in medium conditions is generally complex and may involve multiple transcription factors. A particularly illustrative case is the diauxic shift, during which more than 700 genes are transcriptionally activated and more than 1000 repressed (DeRisi *et al*., 1997). It is, however, possible to refine the definition of the families on the basis of the temporal profile of response (not shown), or by concomitantly taking into account information from independent experiments, as we did for the TUP family. Thanks to the development of new techniques such as the DNA microarrays (Velculescu *et al*., 1997; DeRisi *et al*., 1997), the relative level of expression of all yeast genes will be systematically monitored in cells with different nutritional metabolisms and/or developmental states. Oligonucleotide analysis could be applied to sets of genes for which regulatory elements are totally unknown. Provided a good criterion can be found to ensure a correct definition of these new regulons, our method offers the perspective to rapidly identify candidate upstream regulatory sequences for them.

Methodology

Constitution of regulatory families

The essential criterion for the constitution of a regulatory family is that all member genes have to show a common regulatory response. Families can be defined as a regulon, i.e. a set of genes controlled by a common regulator, or as a stimulon, i.e. a set of genes whose transcription responds to a common environmental stimulus (Neidhardt & Savageau, 1996). When one builds regulatory families, one should avoid to include pairs of structurally related upstream sequences which would strongly bias the probabilistic calculation. For instance, GAL1 and GAL10 are regulated by a common intergenic region, so we had to remove one of them from the GAL family to avoid counting twice the oligonucleotides of the same sequence. Another situation to avoid would be a pair of highly similar upstream regions due to a recent duplication event.

Definition of regulatory region limits

In yeast, regulatory elements are found almost exclusively upstream from the promoter. One would be tempted to consider as putative regulatory sequence the region located between the transcription start and the immediate upstream coding sequence. However, both limits are matter of problems. In eukaryotes, and particularly in yeast, the transcription start can hardly be predicted on the sole basis of the sequence. Its position is only reliable in the few experimentally determined promoters. Consequently, we chose the beginning of the ORF as the downstream limit of the putative regulatory region, although we are aware that this would include in the analysis sequences downstream of the transcription start, leading to some loss of statistical significance.

Using the end of the immediate upstream ORF as the upstream limit of a regulatory region is not very satisfactory either. On the one hand, we have no reason *a priori* to discard the possibility that a coding sequence would simultaneously exert some regulatory action on a neighbor ORF. For instance, there are cases where a putative ORF is detected in the very close vicinity of a gene of interest (e.g. YKR032W ends up 67 bp upstream from DAL80 ORF start), though we know this gene to be regulated by a larger upstream region. For this reason, we decided to impose a fixed length for all considered upstream sequences. In order to determine the optimal size for the upstream region, we analyzed the position of the 308 yeast regulatory sites from Transfac database (Wingender *et al.*, 1996, 1997; Heinemeyer *et al.*, 1998). The vast majority (99%) of these sites are located within a 800 bp range. Consequently, we considered in our analysis the sequence comprised between coordinates −1 and −800 upstream from the ORF start.

After having retrieved the set of upstream sequences from the regulatory family, the number of occurrences of all oligonucleotides of the selected size are counted. This count takes into consideration multiple occurrences within the same upstream sequence. The number of occurrences of each oligonucleotide across the regulatory family is then compared to its expected value.

Expected oligonucleotide frequencies

The yeast genome is characterized by a sensitive bias in favor of A-T *versus* G-C bases ($f_A = f_T = 0.31$; $f_G = f_C = 0.19$). Moreover, nucleotide succession is not random, and some oligonucleotides are clearly over-represented, noticeably the poly (A), poly (T) and poly (AT) chains (Yagil, 1994). An additional bias results from the fact that oligonucleotides are differentially represented in coding *versus* non-coding sequences (Hutchinson, 1996). A specific expected frequency has thus to be used for each oligonucleotide sequence. One way to calculate the expected oligonucleotide frequencies is to use the frequency observed in the collection of all 800 bp upstream regions from the yeast genome. But there is no reason to systematically restrict the analysis to this precise length, and in fact, the user can select the length of the sequences to analyze. We took the biologically defined set of all non-coding sequences from the genome to evaluate the expected frequencies. Since the intergenic regions have an average length of 600 bp, this calibration is also adequate for regions 800 bp long (not shown). We built tables showing, for each possible oligonucleotide (b), the frequency observed throughout all non-coding segments of the whole yeast genome ($F_{nc}\{b\}$), and this for all sizes between one and nine. These frequencies were then used to estimate the oligonucleotide-specific expected frequencies ($F_e\{b\}$):

$$F_e\{b\} = F_{nc}\{b\}$$

These expected frequencies are used to calculate the number of expected occurrences for each oligonucleotide in the set of upstream sequences from the regulatory family:

$$E(\mathrm{occ}\{b\}) = F_e\{b\} \times 2 \times \sum_{i=1}^{S}(L_i - w + 1) = F_e\{b\}^{*}T \qquad (1)$$

where E(occ{b}) is the expected number of occurrences for the oligonucleotide b; w is the oligonucleotide length; S is the number of sequences in the set; L_i is the length of the ith sequence of the set. The factor 2 stands for the fact that we sum the occurrences on both DNA strands, since in our model the action of regulatory sites is orientation-insensitive. T represents the total number of possible matching positions for a pattern of length w across both strands of the sequence set. Since in our case all upstream sequences have the same length (L), T can be simplified as follows:

$$T = 2 \times \mathrm{S} \times (\mathrm{L} - \mathrm{w} + 1)$$

Statistical significance

The probability to observe exactly n occurrences of the oligonucleotide b is estimated by the binomial formula:

$$P(\mathrm{occ}\{b\} = n) = \frac{T!}{(T-n)! \times n!} \times (F_e\{b\})^n \times (1 - F_e\{b\})^{(T-n)} \qquad (2)$$

The probability to observe n or more occurrences of the oligonucleotide b is:

$$P(\mathrm{occ}\{b\} \geqslant n) = \sum_{j=n}^{T} P(\mathrm{occ}\{b\} = j) = 1 - \sum_{j=0}^{n-1} P(\mathrm{occ}\{b\} = j) \qquad (3)$$

One can impose a threshold on $P(\mathrm{occ}\{b\} \geqslant n)$ in order to single out the unexpectedly over-represented oligonu-

cleotides, i.e. those for which the occurrence probability is very low. The choice of an appropriate threshold depends on oligonucleotide length. For instance, there are 4096 possible hexanucleotides. In our case, however, we summed occurrences on both strands, so that each oligonucleotide is grouped together with its reverse complement, except for the 64 palindromic hexanucleotides. The number of palindromic oligonucleotides of size w is $n_{pal} = 4^{w/2}$, when w is even, and $N_{pal} = 0$ otherwise. When the search is performed on both strands, the number D of distinct oligonucleotides is:

$$D = 4^w - (4^w - N_{pal})/2$$

For hexanucleotides, $D = 2080$. With a probability threshold of 0.001, two of them are still expected at random within each family. A good criterion is thus to conserve only hexanucleotides whose occurrence probability is lower than $1/D$.

On basis of these considerations, we defined a significance coefficient:

$$sig = -\log_{10}[P(\text{occ}\{b\} \geqslant n) \times D] \quad (4)$$

which takes into account the number of distinct oligonucleotides. The highest values for this parameter correspond to the most over-represented patterns. When selecting only the patterns for which $sig \geqslant 0$, one expects less than one pattern to occur at random within each family. Each increment of 1 for the significance coefficient represents a drop of a factor of 10 for the occurrence probability. In other words, one expects to find at random one pattern with $sig \geqslant 1$ every ten families, one with $sig \geqslant 2$ every 100 families, and one with $sig \geqslant s$ every 10^s families. The advantage of the significance coefficient is that its threshold can be selected and its values interpreted independently of oligonucleotide size, upstream sequence size, and number of genes within the family. This parameter proved very useful to detect over-represented oligonucleotides, and we systematically used it rather than the crude probabilities in the description of the results.

Correction for auto-correlated patterns

An essential condition for the validity of the binomial formula is the independence of successive trials. This condition is not fulfilled in our case, since the oligonucleotide at each position will depend on the $w-1$ preceding ones, and affect the $w-1$ following ones. We checked the effect of this dependency with random sequences. The results of this analysis are not detailed here, but can be consulted on the web site.

For most oligonucleotides, this effect is not sensible, and the occurrences observed at random fit perfectly with the binomial curve. A deviation from the binomial frequencies is, however, observed in some well defined cases, i.e. for auto-correlated oligonucleotides (e.g. AAAAAA, ATATAT, ATGATG). Auto-correlation does not affect the expected occurrence number, but increases the variance (Kleffe & Borodovsky, 1992). In other words, the probability to observe either very high or very low occurrence values is increased for auto-correlated patterns. An auto-correlation coefficient (Pevzner *et al.*, 1989) can be calculated to estimate the importance of this effect. Corrections on occurrence probabilities (Pevzner *et al.*, 1989; Stückle *et al.*, 1990; Kleffe & Borodovsky, 1992) have been proposed in the context of Markov chain models, but these are not appropriate in our case (we do not use $(k-1)$-mer frequencies to estimate k-mer expected number of occurrences). The auto-correlation effect is neglected in most approaches based on binomial or related formulae (Waterman *et al.*, 1984; Naus & Sheng, 1997), and we did not find any related correction to the binomial formula in the literature.

In our case, the most important contribution to auto-correlation is due to the fact that we sum the occurrences of each pattern on both strands, accordingly with the biological model. This has a very drastic consequence on the probabilities of complementary palindromic patterns (e.g. CACGTG), for which each occurrence on one strand is systematically accompanied by another occurrence on the reverse complementary strand (the auto-correlation coefficient is thus multiplied by 2). A very simple correction for this is to calculate the probabilities on the basis of the single strand occurrences. The correction consists in replacing T by $^T/_2$ and n by $^n/_2$ in equations (2) and (3). With this correction, the occurrence probabilities calculated for complementary palindromic patterns fit perfectly with those observed for random sequences.

For non-palindromic patterns, we noticed that in the absence of any correction, the autocorrelation was sensitive only for patterns with a periodicity of 1 (e.g. AAAAAA and TTTTTT) or 2 (e.g. TATATA, CACACA). This effect was occasionally observed in sets of yeast upstream regions. In the few cases where a high statistical significance was assigned to such patterns, they were especially concentrated in a long continuous stretch on one or two upstream regions. Such large periodic stretches are thought to be involved in DNA structure rather than in specific regulatory functions. Since their presence is easily detected on the feature map, the risk to misinterpret them is insignificant, and we did not judge necessary to correct the statistics for this level of auto-correlation.

Feature map

After having selected the significantly over-represented oligonucleotides, their matching positions within each upstream sequence are determined, and represented graphically in a feature map.

Implementation

All programs used in this paper can be run from a public web interface (http://copan.cifn.unam.mx/Computational_Biology/yeast-tools). The only input required from the user is the name of the genes included in a regulatory family. A series of modular tools are then presented, allowing to perform successively each step of our method: extraction of upstream regions from the genomic sequences, detection of over-represented oligonucleotides, search for all matching positions within the set of upstream sequences, automatic drawing of a feature map. At each step, a series of parameters can be modified by the user. These parameters are generally intuitive (upstream region size, oligonucleotide size, search on single or both strands, etc.) and can be understood by a non-experimented user. Default values are proposed for all parameters, so that the whole analysis can be performed without any other intervention than clicking "OK" at each step. All results from Table 2 were obtained without changing any of these default parameters.

Each tool can also be used independently of the others, with custom input from the user (e.g. searching

for a known pattern in his sequences, or drawing a feature map of experimentally described regulatory sites).

All programs are written in Perl (Wall & Schwartz, 1991) and run on a Sun workstation. The web interface is written in perl-cgi (Gundavaram, 1996). Oligonucleotide analysis is fast. Processing time grows linearly with the size of sequences to analyze, with an average speed of 0.8 kb per second. Hexanucleotide analysis of a family comprising ten genes is typically performed within ten seconds.

Acknowledgments

J.v.H. was a postdoc at CIFN-UNAM when this work was done. This work was supported by grants from DGAPA-UNAM and CONACYT to J.C.-V. We acknowledge an anonymous referee for the observation of the poly-valine stretch in the ORF upstream from FAS2.

References

Balzi, E. & Goffeau, A. (1995). Yeast multidrug resistance: the PDR network. *J. Bioenerget. Biomembr.* **27**, 71–76.

Blaiseau, P.-L., Isnard, A.-D., Surdin-Kerjan, Y. & Thomas, D. (1997). Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism genes. *Mol. Cell. Biol.* **17**, 3640–3648.

Bodnar, J. W. & Ward, D. C. (1987). Highly recurring sequence elements identified in eukaryotic DNAs by computer analysis are often homologous to regulatory sequences or protein binding sites. *Nucl. Acids Res.* **14**, 1835–1851.

Bysani, N., Daugherty, J. R. & Cooper, T. G. (1991). Saturation mutagenesis of the UASNTR (GATAA) responsible for nitrogen catabolite repression-sensitive transcriptional activation of the allantoin pathway genes in *Saccharomyces cerevisiae*. *J. Bacteriol.* **173**, 4977–4982.

Cardon, L. R. & Stormo, G. D. (1992). Expectation maximisation algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J. Mol. Biol.* **223**, 159–170.

Claverie, J. M. & Bougueleret, L. (1986). Heuristic informational analysis of sequences. *Nucl. Acids Res.* **14**, 179–196.

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

Fondrat, C. & Kalogeropoulos, A. (1994). Approaching the function of new genes by detection of their potential upstream activation sequences in *Saccharomyces cerevisiae*: application to chromosome III. *Curr. Genet.* **25**, 396–406.

Galas, D. J., Eggert, M. & Waterman, M. S. (1985). Rigorous pattern-recognition methods for DNA sequences: analysis of promoter sequences from *Escherichia coli*. *J. Mol. Biol.* **186**, 117–128.

Goffeau, A. (1997). Molecular fish on chips. *Nature*, **385**, 202–203.

Guarente, L. (1992). Messenger RNA transcription and its control in *Saccharomyces cerevisiae*. In *The Molecular and Cellular Biology of the Yeast* Saccharomyces: *Gene Expression* (Jones, E. W., Pringle, J. R. & Broach, J. R., eds), pp. 193–281, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Gundavaram, S. (1996). *CGI Programming on the World Wide Web*, O'Reilly Associates, Sebastopol, CA.

Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A. E., Kel, O. V., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Kolpakov, F. A., Podkolodny, N. L. & Kolchanov, N. A. (1998). Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucl. Acids Res.* **26**, 362–367.

Hertz, G. Z. & Stormo, G. D. (1995). Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps. In *Proceeds. 3rd Internat. Conf. on Bioinformatics and Genome Research*, pp. 201–216, World Scientific Publishing Co. Ltd, Singapore.

Hertz, G. Z., Hartzell, G. W. & Stormo, G. D. (1990). Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* **6**, 81–92.

Hinnebusch, A. G. (1992). General and pathway-specific regulatory mechanisms controlling the synthesis of amino acid biosynthetic enzymes in Saccharomyces cerevisiae. In *The Molecular and Cellular Biology of the Yeast* Saccharomyces: *Gene Expression* (Jones, E. W., Pringle, J. R. & Broach, J. R., eds), pp. 319–414, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Hutchinson, G. B. (1996). The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput. Appl. Biosci.*, **12**, 391–398.

Jauniaux, J. C., Vandenbol, M., Vissers, S., Broman, K. & Grenson, M. (1987). Nitrogen catabolite regulation of proline permease in *Saccharomyces cerevisiae*. Cloning of the PUT4 gene and study of PUT4 RNA levels in wild-type and mutant strains. *Eur. J. Biochem.* **164**, 601–606.

Johnston, M. & Carlson, M. (1992). Regulation of carbon and phosphate utilisation. In *The Molecular and Cellular Biology of the Yeast* Saccharomyces: *Gene Expression* (Jones, E. W., Pringle, J. R. & Broach, J. R., eds), pp. 193–281, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Jones, E. W., Pringle, J. R. & Broach, J. R. (1992). Editors of *The Molecular and Cellular Biology of the Yeast* Saccharomyces: *Gene Expression*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Kleffe, J. & Borodovsky, M. (1992). First and second moment counts of words in random texts generated by Markov chains. *Comput. Appl. Biosci.* **8**, 433–441.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O. & Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl Acad. Sci. USA*, **94**, 13057–13062.

Lawrence, C. E. & Reilly, A. A. (1990). An expectation maximisation (EM) algorithm for the identification and characterisation of common sites in unaligned biopolymer sequences. *Proteins: Struct. Funct. Genet.* **7**, 41–51.

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

Lopes, J. M., Hirsch, J. P., Chorgo, P. A., Schulze, K. L. & Henry, S. A. (1991). Analysis of sequences in the INO1 promoter that are involved in its regulation by phospholipid precursors. *Nucl. Acids Res.* **19**, 1687–1693.

Magasanik, B. (1992). Regulation of nitrogen utilisation. In *The Molecular and Cellular Biology of the Yeast* Saccharomyces: *Gene Expression* (Jones, E. W., Pringle, J. R. & Broach, J. R., eds), pp. 283–318, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Mengeritsky, G. & Smith, T. F. (1987). Recognition of characteristic patterns in sets of functionally equivalent DNA sequences. *Comput. Appl. Biosci.* **3**, 223–227.

Naus, J. I. & Sheng, K.-N. (1997). Matching among multiple random sequences. *Bull. Mat. Biol.* **59**, 483–496.

Neidhardt, F. C. & Savageau, M. A. (1996). Regulation beyond the operon. In *Escherichia coli and Salmonella Cellular and Molecular Biology* (Curtis, R., *et al*., ed.), pp. 1325–1342, ASM Press Washington.

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

Oshima, Y., Ogawa, N. & Harashima, S. (1996). Regulation of phosphatase synthesis in *Saccharomyces cerevisiae*: a review. *Gene,* **179**, 171–177.

Paltauf, F., Kohlwein, S. D. & Henry, S. (1992). Regulation and compartmentalization of lipid synthesis in yeast. In *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression* (Jones, E. W., Pringle, J. R. & Broach, J. R., eds), pp. 415–500, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

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

Schena, M. (1996). Genome analysis with gene expression microarrays. *BioEssays,* **18**, 427–431.

Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science,* **270**, 467–470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. & Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA,* **93**, 10614–10619.

Stormo, G. D. & Hartzell, G. W. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA,* **86**, 1183–1187.

Strachan, T., Abitol, M., Davidson, D. & Beckman, J. S. (1997). A new dimension for the human genome project: towards comprehensive expression maps. *Nature Genet.* **16**, 126–132.

Stückle, E. E., Emmrich, C., Grob, U. & Nielsen, P. J. (1990). Statistical analysis of nucleotide sequences. *Nucl. Acids Res.* **18**, 6641–6647.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Basset, D. E., Jr., Hieter, P., Vogelstein, B. & Kinzler, K. W. (1997). Characterisation of the yeast transcriptome. *Cell,* **88**, 243–251.

Wall, L. & Schwartz, R. L. (1991). *Programming Perl*, O'Reilly and Associates, Inc., Sebastopol, CA.

Waterman, M. S., Arratia, R. & Galas, D. J. (1984). Pattern recognition in several sequences: consensus and alignment. *Bull. Math. Biol.* **45**, 515–527.

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

Wingender, E., Kel, A. E., Kel, O. V., Karas, H., Heinemeyer, T., Dietze, P., Knüppel, R., Romaschenko, A. G. & Kolchanov, N. A. (1997). TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation. *Nucl. Acids Res.* **25**, 265–268.

Wolfertstetter, F., Frech, K., Hermann, G. & Werner, T. (1996). Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comput. Appl. Biosci.* **12**, 71–80.

Yagil, G. (1994). The frequency of oligopurine. oligopyrimidine and other two-base tracts in yeast chromosome III. *Yeast,* **10**, 601–611.

Edited by G. von Heijne

(Received 9 February 1998; received in revised form 10 May 1998; accepted 13 May 1998)