1808–1818 *Nucleic Acids Research, 2000, Vol. 28, No. 8*

Discovering regulatory elements in non-coding sequences by analysis of spaced dyads

Jacques van Helden*, Alma. F. Rios[1] and Julio Collado-Vides[1]

Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP 160/16, 50 av. F. D. Roosevelt, B-1050 Bruxelles, Belgium and [1]Centro de Investigación sobre Fijación de Nitrógeno, Universidad Nacional Autónoma de México, AP565A, Cuernavaca 62100, Morelos, Mexico

Received November 24, 1999; Revised and Accepted March 1, 2000

ABSTRACT

The application of microarray and related technologies is currently generating a systematic catalog of the transcriptional response of any single gene to a multiplicity of experimental conditions. Clustering genes according to the similarity of their transcriptional response provides a direct hint to the regulons of the different transcription factors, many of which have still not been characterized. We have developed a new method for deciphering the mechanism underlying the common transcriptional response of a set of genes, i.e. discovering *cis*-acting regulatory elements from a set of unaligned upstream sequences. This method, called dyad analysis, is based on the observation that many regulatory sites consist of a pair of highly conserved trinucleotides, spaced by a non-conserved region of fixed width. The approach is to count the number of occurrences of each possible spaced pair of trinucleotides, and to assess its statistical significance. The method is highly efficient in the detection of sites bound by C_6 Zn_2 binuclear cluster proteins, as well as other transcription factors. In addition, we show that the dyad and single-word analyses are efficient for the detection of regulatory patterns in gene clusters from DNA chip experiments. In combination, these programs should provide a fast and efficient way to discover new regulatory sites for as yet unknown transcription factors.

INTRODUCTION

With the multiplication in genome projects, the rapid development of sequencing methods has generated an explosive growth of available data, which, however, has not resulted in a comparable increase in biological knowledge, due to the bottleneck of interpretation limits. This discrepancy has stimulated the rapid development of new methods to address, in a systematic way, the function of thousands of newly sequenced genes (1). Among these, microrarray technologies allow direct measurement of the level of expression of each single gene in a cell (2,3). One can characterize sets of genes involved in a defined cellular process by such methods. Once a cluster of co-regulated genes has been isolated by gene expression measurements, the next step is to understand the mechanism that is responsible for the coordinated response. The coherence of the cellular response to a perturbation is ensured by the action of transcription factors, which bind to conserved sequences upstream of the transcription start and interact with the RNA polymerase to activate or repress the expression of a selected set of target genes. The selectivity of this effect is ensured by the fact that each transcription factor recognizes a specific DNA sequence, the regulatory site, and accordingly, regulates only the genes with such upstream binding capability.

Starting from a family of genes characterized by their common response to a stimulus, the question we address here is how to extract motifs that are shared by their upstream sequences, and that could therefore be responsible for the co-regulation. This problem of pattern discovery has been addressed by a variety of approaches (4–15). In a previous paper (16), we developed a method called oligo-analysis, which allows extraction of regulatory sites based on an analysis of oligonucleotide frequencies. Despite its simplicity, this method turned out to be very efficient at detecting regulatory sites from many families of co-regulated genes in yeast. The method can also been applied to large sequence sets, and has been used to detect 3′ end signals in the set of 6217 yeast downstream sequences (17). The success of this word-counting algorithm is due to the structure of protein–DNA interfaces: the most common case in yeast is that the DNA-binding domain of the transcription factor makes direct contact with a limited number of highly specific adjacent nucleotides. Consequently, many regulatory sites in yeast are short strands of adjacent nucleotides (6–8 bp) with a very low internal variation.

There is, however, a notable exception to this rule: some transcription factors like Gal4p bind to a pair of conserved trinucleotides, separated by a spacer of fixed length but variable content. The short conserved trinucleotides correspond to residues that enter into direct contact with the DNA-binding domain of the transcription factor. Their pairing is due to the fact that the transcription factor forms a dimer, with each unit binding to a similar small element, accounting for the symmetry of the site. The fixed spacing in the DNA site is due to the existence of a linker domain in the transcription factor, separating the DNA-binding and dimerization domains. Since the factor generally does not enter in direct contact with the spacer, this region has a much less conserved base composition than at the two contact sites. Such spaced dyad elements are

*To whom correspondence should be addressed. Tel: +32 2 650 20 13; Fax: +32 2 648 89 54; Email: jvanheld@ucmb.ulb.ac.be

common to a large class of transcription factors, found so far almost exclusively in fungi: the C_6 Zn_2 binuclear cluster proteins (18,19). This class encompasses 56 of the 170 transcription factors predicted from the genome of *Saccharomyces cerevisiae*. The best characterized factor of this family is Gal4p, the activator of a set of genes expressed in presence of galactose.

The single word counting approach was found to be inefficient in detecting such spaced pairs (16), and failed to detect any motif in the GAL family. We present here a new method, called dyad analysis, specifically designed for the detection of spaced pairs shared by a set of upstream regions. The method is based on a systematic counting of pairs of short words separated by a fixed distance (spaced dyads), followed by a calculation of their statistical significance. We evaluated the performance of dyad analysis on 11 families of genes for which there is experimental evidence of binding sites for Zn cluster proteins. Dyad analysis is able to extract the correct regulatory sites for either 9 or for all the 11 different families, depending on the parameters used. We also evaluated the performances of the program on several families coming from DNA chip experiments.

Dyad analysis is an ideal complement to the previously published oligo-analysis program (16), which extracts most yeast regulatory sites but specifically fails in detecting those from C_6 Zn_2 binuclear cluster proteins. Given their common statistical foundation and complementary scope, the combination of these two methods provides a powerful tool for the extraction of putative regulatory sites from sets of co-regulated genes coming from genome-scale gene expression measurements. Although the analysis presented is focused on yeast regulatory regions and sites, the method can in principle be applied to identify motifs in other organisms. Dyad analysis has been added to the Regulatory Sequence Analysis Web site (http://www.ucmb.ulb.ac.be/bioinformatics/rsa-tools ; http://copan.cifn.unam.mx/Computational_Biology/yeast-tools/) (20).

MATERIALS AND METHODS

Dyad analysis

A spaced dyad D is formed by a pair of short conserved words separated by a region of fixed size and variable content, as follows:

$$D = w_1.n_s.w_2$$

where D is the sequence of a dyad (the dot represents string concatenation), w_1 and w_2 the first and second words of the dyad respectively, s the width of the spacing, and n_s any sequence of s unspecified nucleotides.

The occurrences in the input sequence set of each word and spaced dyad were counted. Since yeast upstream elements act in a strand-insensitive way, each dyad is grouped with its reverse complement and their occurrences added. In agreement with the biological model, the default word length is fixed to three. Since the spacer width can take different values depending on the transcription factor, we systematically scan all possible spacing values between 0 and16.

We evaluated two alternative estimators of the expected frequencies. The first approach was to calculate expected dyad frequencies based on the monad (word) frequencies observed in the input sequence set:

$$F_{obs}(w) = Occ(w) / \Sigma_i [Occ(w_i)]$$

$$F_{exp}(D) = F_{obs}(w_1) \times F_{obs}(w_2)$$

where Occ(w) is the number of occurrences of the word w in the input sequence, $F_{obs}(w)$ the relative frequency observed for the word w in the input sequence and $F_{exp}(D)$ the expected frequency for the dyad D.

In the second approach, we measured the frequency of each dyad in the complete set of non-coding yeast sequences and used this frequency as a direct estimate of the dyad frequencies expected in any family of upstream sequences.

$$F_{exp}(D) = F_{ncf}(D)$$

where $F_{ncf}(D)$ is the frequency of dyad D in the complete set of non-coding yeast sequences.

The statistical significance of the observed number of occurrences is obtained from the binomial:

$$P(D, \geq n) = \Sigma^T_{i=n} C_i^T [F_{exp}(D)]^i \times [1 - F_{exp}(D)]^{(T-i)}$$

$P(D, \geq n)$ being the probability of observing at least n occurrences of the dyad D in the input set:

$$C_i^T = T!/[i!\ (T - i)!]$$

is the binomial coefficient for T trials and i successes. T is the total number of positions where a dyad can be found in the input sequence set, i.e.

$$T = \Sigma_j [L_j - 2k - s + 1]$$

where L_j is the length of the j^{th} sequence of the input set, k the word length, and s the width of the spacing.

As discussed by van Helden *et al.* (16), the binomial statistic is appropriate except for self-overlapping patterns, which have an aggregative behaviour. To circumvent this problem, we performed the count without allowing overlapping matches. Each occurrence of a pattern prevented counting the same afternoon within the next K – 1 positions, where K = 2k + s is the dyad length. Accordingly, the number of possible positions, T′, has to be corrected as follows:

$$T' = \Sigma_j [L_j - 2k - s + 1] - Occ(w) \times (2k + s - 1)$$

The detection of biologically significant patterns in a set of sequences strongly depends on the choice of an appropriate threshold. This threshold must be set to a very restrictive value, given the fact that a high number of possible patterns are taken into consideration for each individual analysis. There are $4^3 = 64$ possible trinucleotides, which can be combined in $64^2 = 4096$ distinct dyads. These dyads are grouped by pairs of reverse complements, reducing the number of independent patterns to 2080 (see 16 for the details of calculation). Since we systematically scanned 17 possible spacing values (from 0 to16), $N_p = 2080 \times 17 = 35\,360$ spaced dyads are considered in each analysis. Consequently, even with a probability threshold as low as 0.001, one would still expect 35 patterns to appear at random in each family. This illustrates clearly that the threshold has to be adapted to the number of possible patterns considered, which depends on the word size and range of spacing values scanned. A good basis is to take a threshold of probability $\leq 1/N_p$.

We defined a significance index:

$$sig = -\log_{10} [P(D, \geq n) \times N_p]$$

This significance index provides an intuitive estimator of over-representation. The more over-represented the pattern, the higher is the sig value. By fixing the threshold of significance to 0, one would expect no more than one pattern at random within each family, and this is independent of sequence length, word size and range of spacing values considered. The significance index can thus be used to compare probabilities among patterns of different sizes.

Restriction to symmetric patterns

Many DNA-binding proteins (transcription factors, restriction enzymes) bind to symmetric DNA sites. Two types of symmetries are observed: tandem repeats and reverse complementary palindromes. An option of dyad analysis is to restrict the analysis to such symmetric patterns. When this option is activated, the probabilities are calculated as above, except for the number of possible patterns. Indeed, under the symmetry model, the first word determines the second one. Consequently, there are no more than $4^3 = 64$ possible tandem repeats of trinucleotides. Assuming that signals act in a strand-insensitive way, each tandem repeat is grouped with its reverse complement, reducing the number of distinct trinucleotide pairs to 32. In the same way, there are $4^3 = 64$ possible reverse complementary palindromes of two trinucleotides. This number remains unchanged, however, even when the analysis is performed on two strands, because by definition each palindrome can only be grouped with itself. When combining the two types of symmetries, one has thus to consider $N_p = 96 \times$ (max. spacing – min. spacing + 1) possible spaced dyads. The significance index is calculated by using this N_p value in the above formula.

Pattern assembly

Several patterns related by sequence similarity are generally selected from a single run of the program, and reveal overlapping segments of the same regulatory site. We wrote a program (pattern-assembly) to assemble sequence-related patterns in a contig. This program performs the same type of operations as contig assembly programs used in genome sequencing, but it is optimized to assemble short patterns with a very high level of similarity.

Implementation and availability

A prototype version of the dyad analysis was written in Mathematica. The program was then ported to Perl on Unix and its functionality was extended. The web interface is in cgi-perl. All programs are available on the web (http://www.ucmb.ulb.ac.be/bioinformatics/rsa-tools/ or http://copan.cifn.unam.mx/Computational_Biology/yeast-tools/) (20).

RESULTS

Families of co-regulated genes

We collected from the literature a list of target genes (Table 1A), as well as a description of the known binding sites for each characterized zinc cluster protein (18,19,21). This list includes a second family of genes responding to galactose: the 'GAL-chip' family results from a DNA-chip experiment (14) that evaluates the ratio of expression of all yeast genes between galactose-rich and glucose-rich culture conditions. The upstream sequences were retrieved from positions –1 to –800 relative to the ORF start positions.

Analysis of known regulons

Table 2 shows the result of dyad analysis in the GAL family. Among the 35 360 possible dyads, no more than three are significantly over-represented (two of the five rows show the reverse complementary sequence of another pattern). Moreover, these dyads are related to each other, and can be assembled into a common pattern (last row of Table 2). The resulting contig (tCGGan$_9$tCCGa) fits with the known consensus CGGrnnrcynyncnCCG (18). The spaced dyad detected with the highest significance, CGGn$_{11}$CCG, corresponds to the nucleotides that enter into direct contact with Gal4p. This dyad is found in 20 occurrences, when 0.59 would be expected at random. The probability of observing 20 occurrences or more when expecting 0.59 is $1.9e^{-12}$. This probability has to be evaluated taking into consideration the fact that each family potentially comprises 35 360 distinct dyads. Thus, when analyzing all dyads with spacing between 0 and16, the fate is challenged 35 360 times, and an individual pattern with a probability of 1/35 360 would still be expected on average once in every family. For this reason, we introduced the significance index (Materials and Methods), which better reflects the over-representation. For CGGn$_{11}$CCG, the significance index sig = $-\log_{10}(1.9e^{-12} \times 35\ 360) = 7.2$, indicating that, in random sequences, one would observe a pattern with such a high significance once every $10^{7.2}$ families.

Table 3 summarizes the results of dyad analysis for the 11 families known to be regulated by zinc cluster factors. In this table, the assemblies are directly shown rather than the individual patterns.

In the third column ('monad calibration'), expected dyad frequencies were calculated on the basis of the monad frequencies observed in the input sequence set (Materials and Methods). A very small number of patterns are selected within each family. Patterns matching the known consensus are highlighted in bold and uppercase. Within each family, the most significant patterns generally correspond to the known consensus. The program succeeds in detecting 9 known motifs among the 11 different ones. The exceptions are the binding sites for Uga3p and Hap1p. The UGA3 family contains a very small number of genes, illustrating one of the limits in sensitivity of our method: the program needs a sufficient number of sequences for the shared pattern to reach the significance threshold. Note that the correct pattern is however extracted from family PPR1, which also contains three genes, and even (albeit with a very weak significance) from the PUT3 family, which contains only two genes. The HAP1 family is an interesting case because the Hap1p binding site is the only one where the trinucleotides are not well conserved. This loss of specificity in the Zn cluster binding sequences is compensated by the fact that the linker region of Hap1 enters into specific contact with the intermediate DNA region (22). The variability of this trinucleotide sequence might be responsible for the failure of the dyad analysis in this family.

In the fourth column of Table 3 ('dyad calibration'), dyad frequencies measured in the whole non-coding genome were used as expected dyad frequencies. In this case, the program is able to isolate 11 of the 11 known motifs. This increase in sensitivity is, however, at the cost of selectivity, and a few

Table 1. Composition of the gene families analyzed in this article

A. Genes regulated by Zn cluster factors

family	genes
GAL4	GAL1 GAL2 GAL7 GAL80 MEL1 GCY1
GAL4 chips	GAL1 GAL7 GCY1 GAL2 YPL066W YMR318C PBI2 ARG1 GAL3
CAT8	ACR1 ICL1 MLS1 PCK1 FBP1
HAP1	CYB2 CYC1 CYC7 CTT1 CYT1 ERG11 HEM13 HMG1 ROX1
LEU3	GDH1 ILV1 LEU1 LEU2 LEU4
LYS	LYS1 LYS2 LYS4 LYS9 LYS20 LYS21
PDR	YOR1 PDR11 PDR10 GAS1 STE6 SNQ2 PDR5
PPR1	URA1 URA3 URA4
PUT3	PUT1 PUT2
UGA3	UGA1 UGA4 YBR006W
UME6	BAR1 CAR1 CAR2 DMC1 GAL1 HOP1 HSF1 ILV2 IME1 IME2 INO1 MEI4 MER1 REC102 REC114 RED1 RME1 SPO11 SPO13 SPO16 TOP1 ZIP1

B. Cell-cycle-regulated genes

Family	Genes
CLN2	EST1 YBR070C SMC1 CAC2 YJL181W POL12 YBR089W YCL022C YCL024W SPT21 CLB6 TOF1 RHC18 SPH1 CDC21 SMC3 CDC45 YDL011C YDL010W ASF2 YFR027W UNG1 RAD27 RFA2 RNR1 YDL161W YML133C RNH35 SWE1 YNL300W KIM2 RAD53 SPO16 YGR221C YOL017W RFA1 MSH2 RSR1 YGR151C HIF1 MSH6 PRI2 SRO4 CSI2 CLN2 YOX1 POL30 MCD1 CLB5 YPL208W YPR174C DPB2 STB1
Y' (purged)	YBL112C YEL073C YEL075C YEL077C YFL046C YFL066C YFL067W YGR296W YLL067C YLR446W YPR202W
Histone (purged)	HHF1 HHF2 HHO1 HTB1 HTB2
cell cycle MET	YPL250C YER042W YLR302C YPL274W MET28 YLL061W MET1 YIL074C YLL062C MET14 MET16 MET3 MET10 ECM17 YNL276C MUP1 MET17 MET6
CLB2	NUM1 YCL063W YLR057W APC1 BUD8 YCL012W BUD3 YPL141C KIP2 IQG1 YPR156C BUD4 SHE2 TEM1 YNL058C CHS2 MYO1 YJL051W YIL158W YML033W YML034W MOB1 HST3 ACE2 CDC20 CYK2 YML119W YLR190W SWI5 ALK1 CLB2 CDC5 CLB1 YLR084C
MCM	YIL167W MRPS28 YML050W RGT2 YHL026C UTH1 SED1 YCR041W YGP1 PRY1 YLR297W MCM3 YOR066W CDC46 GPA1 HST4 MCM6 TSM1 WTM2 BEM1 YMR031C SKN1 YLR254C FAR1 YGR230W DBF2 SPO12 KIN3 YOL070C CDC47 YRO2 YDR033W PHO12 PHO5 CDC54 MCM2 YDR190C
SIC1	YDR055W YNL078W YNR067C EGT2 PCL9 YOR263C YNL046W YOR264W FAA3 TEC1 HSP150 PIR1 PIR3 ASH1 YPL158C YKL116C YDL117W SIC1 YGR086C YER124C YHR143W CTS1 YGL028C PRY3 RME1 YBR158W
MAT	YDR493W LIF1 MF(ALPHA)2 SAG1 YKL177W STE3 YLR040C KAR4 AGA1 SST2 FUS1 MF(ALPHA)1 SRD1

C. Genes regulated by non-Zn cluster factors

family	genes
NIT	DAL5 GAP1 MEP1 MEP2 PUT4 MEP3 DAL80
MET	MET1 MET2 MET3 MET6 MET14 ZWF1 MET17 MET30 MUP3 SAM1 SAM2
PHO	PHO5 PHO11 PHO8 PHO84 PHO81
GCN4	ARG1 ARG3 ARG4 ARG8 ARO3 ARO4 ARO7 CPA1 CPA2 GLN1 HIS1 HIS2 HIS3 HIS4 HIS5 HOM2 HOM3 ILV1 ILV2 ILV5 LEU1 LEU2 LEU3 LEU4 LYS1 LYS2 LYS5 LYS9 TRP2 TRP3 TRP4 TRP5 MET14 MET3 MET6 MES1 THR1 HOM6
INO	CHO1 CHO2 FAS1 FAS2 ACC1 INO1 OPI3
HAP2-5	CYC1 HEM1 LPD1 KGD1 KGD2 SOD2 ASN1 ASN2 GDH1 CIT1 COX4 COX5A COX5B
TUP	FSP2 YNR073C YOL157C HXT15 SUC2 YNR071C YDR533C YEL070W RNR2 YER067W CWP1 YGR243W SSN21 SHC1 HXT6 YLR327C YJL171C YGR138C HXT4 HXT7 GSY1 YOR389W MAL31 YML131W RCK1
YAP	AAD14 YKL071W AAD6 GTT2 AAD15 AAD3 ATR1 FLR1 FRM2 AAD10 OYE3 YLR460C ECM4 OYE2 YML131W MDH2

(**A**) Regulons of C_6 Zn_2 binuclear cluster factor. (**B**) Gene clusters extracted from a DNA chip experiment (24). Each family corresponds to one figure of the original paper. The Y′ and histone families were purged to avoid redundant upstream regions (see text for details). (**C**) Regulons of transcription factors belonging to other classes than Zn cluster proteins, same families as in (16). Note that YAP and TUP families are also based on DNA chip experiments (16).

more patterns appear that do not seem to be related to known regulatory sites. The known signals remain nevertheless associated to the highest significance within each family. This less stringent approach may thus be an interesting alternative when no patterns are selected with the default (monad) settings.

With the monad calibration, some patterns are detected that are not likely to be involved in the family-specific regulation. These are generally AT-rich sequences. Their significance is lower than that of the known regulatory sites, except in the LYS family, where ATATATA is the top-ranking pattern.

Table 2. Dyad analysis result for the GAL family

pattern	total occurrences	overlaps	non-overlapping occurrences	expected occurrences	proba	sig
.CGG.....CCG	22	2	20	0.59	$1.9e^{-12}$	7.2
.CGG.....CGa	12	2	10	0.50	$2.1e^{-10}$	5.1
tCG.....CCG	12	2	10	0.50	$2.1e^{-10}$	5.1
.CGG.....tCC	12	3	9	0.91	$6.7e^{-07}$	1.6
..GGa.....CCG	12	3	9	0.91	$6.7e^{-07}$	1.6
tCGGa.....tCCGa	**Assembly**					

Note that in this precise case, the ATATATA motif is not only statistically over-represented, but also concentrated within a narrow range of positions (–90 to –130 from the start codon) in each member of the family. These motifs might thus represent a highly conserved TATA-box in this family. The other unknown patterns (TTTn{7}CTT, AAGAAA, . . .) are detected with low significance and without any specific distribution, and can probably be considered as noise or play a non-specific role in these promoter regions. These AT-rich motifs are filtered out with the dyad calibration, showing that their high frequency reflects a general tendency to aggregate poly-A and poly-T strands in yeast upstream sequences, rather than some family-specific feature. For such patterns, the direct measurement of dyad frequencies in all non-coding sequences thus provides a more selective calibration.

Figure 1 compares the localization of the sites predicted by dyad analysis (Fig. 1A and B) with the efficiency of the UAS (Fig. 1C), as measured experimentally (23). All sites with a significance index >0 are shown. A good putative site is generally characterized by the presence of multiple boxes of distinct colors, indicating an overlap between several significant patterns. In Figure 1A, several sites combine a red (AAAnnCCG) and a magenta box (CCGnnGGA), together with a tiny orange or violet box. All these sites but one (at position –530 in the LYS20 upstream sequence) are indeed known to be efficient UAS (23). Only one of the known UAS (at position –160 in the LYS9 upstream sequence) escapes detection. The sites with a single red box are probably not functional. Note also the conserved position for the poly-TA sites, found between –90 and –130 in all genes.

Using different ranges of spacing values (0–12, 0–14, 0–20) does not much affect the set of detected patterns, and has only a slight effect on the significance (not shown).

A more critical parameter is the choice of the upstream sequence length. Too small sequences bring the risk of losing some regulatory sites, and reduce the signal. Too large sequences enhance the noise and reduce the significance. The optimum for yeast in our experience is to analyze upstream sequences 600–800 bp long (not shown).

Many transcription factors bind DNA in the form of homodimers, resulting in the recognition of symmetric sites. It is tempting to use this information to restrict the analysis to symmetric patterns only. We implemented this as an option of the program and tested it on all the families (not shown). Our observation was that this approach is much less sensitive than when allowing any type of dyad. Indeed, some patterns are not symmetric at all, and escape detection. Moreover, even for symmetric patterns, the combination of symmetric and asymmetric fragments leads to a more complete description. In Table 2, for example, the asymmetric dyads $TCGn_{12}CCG$ and $GGAn_{10}CCG$ contribute to the formation of the contig $tCGGan_9tCCGa$. We recommend thus to systematically extend the analysis to all possible dyads.

The last column of Table 3 shows that oligo-analysis would detect only 6 out of the 11 binding sites of C_6 Zn_2 binuclear cluster factors. The dyad analysis thus provides a much higher sensitivity for this class of transcription factors.

Analysis of DNA chip data

After having calibrated the method on a set of known regulons, it is important to evaluate how the program performs with families extracted from DNA chip experiments. Indeed, it is expected for these families to be noisier, and some families could contain a mixture of genes belonging to different regulons. We collected a set of families (Table 1B) from the cell cycle analysis performed by Spellman *et al.* (24). These authors measured the level of expression of all genes during the cell cycle, selected those showing periodic fluctuations, and clustered them according to their peak of expression. Results are summarized in Table 4. Each family corresponds to one cluster from the original paper (24).

An important issue is the choice of the significance threshold. In the previous analyses, we considered sets of independent families, which had been collected from different sources, and selected a significance threshold of 0, so that we expected no more than one pattern to appear randomly per analysis. In contrast, when analyzing together a set of genes collected from the same experiment, one should increase the threshold according to the number of families considered. For example, if one analyzes 100 families, one expects to find at random one pattern with significance index ≥2, 10 with significance index ≥1 and 100 with significance index ≥0. More generally, when analyzing together F families, one should set the threshold of significance to $\log_{10}(F)$. For the eight families extracted from the cell cycle experiment, we set the threshold to 0.9. We used as calibration the dyad frequencies measured in whole-genome intergenic sequences.

For most families, a small number of patterns were detected as significant. The Y′ family, however behaves differently: when the 32 genes from the original Y′ cluster are analyzed, no less than 225 hexanucleotides (oligo-analysis) and 1659 dyads (dyad analysis) appear as significant. This answer is obviously biased by the fact that this family includes many genes resulting from recent duplications, and having almost identical upstream regions. The inclusion of multiple copies of a sequence in the dataset provokes a multiplication of the noise

Table 3. Summary of the dyad analysis in families of genes regulated by a C_6 Zn_2 binuclear cluster factor

family	known sites	dyad-analysis (monad calibration)		dyad-analysis (non-coding dyad frequency calibration)		oligo-analysis	
		dyad reverse comp.	sig	dyad reverse comp.	sig	word reverse comp.	sig
GAL4	CGGRnnRCYnYnCnCCG (ref 18)	TCGGAn9TCCGA TCGGAn9TCCGA	7.0	TCGGAn9TCCGA TCGGAn9TCCGA	7.8		
				TCGGAn8CGCCGA TCGGCGn8TCCGA	7.8	*no significant pattern*	
				CCGGAn9TCCGA TCGGAn9TCCGG	7.8		
				CCGCGCGG CCGCGCGG	0.6		
				AGAn5CCG CGGn5TCT	0.2		
				CGGnGCA TGCnCCG	0.1		
GAL4 chips	CGGRnnRCYnYnCnCCG (ref 18)	TCGGAn9TCCGA TCGGAn9TCCGA	7.8	TCGGAn9TCCGA TCGGAn9TCCGA	8.4	TCCCCGAT ATCGGGGA	1.1
				CACn8CTG CAGn8GTG	0.6		
				CAGn4CCGAC GTCGGn4CTG	0.5		
				AGGn2CAT ATGn2CCT	0.4		
				GGGn16CGA TCGn16CCC	0.2		
				GCCn13CCC GGGn13GGC	0.2		
				AGCn7CGG CCGn7GCT	0.0		
CAT8	CGGnnnnnnnGGA	CGGn4ATGGA TCCATn4CCG	3.2	CGGn4ATGGAA TTCCATn4CCG	6.0	CCCGGAG CTCCGGG	0.5
		AAGn7AAA TTTn7CTT	0.1	GGGCn4GCCC GGGCn4GCCC	0.9	CATCCG CGGATG	0.4
				GAAn4GCC GGCn4TTC	0.2		
				AAGn14AGC GCTn14CTT	0.1		
				GCGn12CCC GGGn12CGC	0.0		
HAP1	CGGnnnTAnCGG (ref 18) CGGnnnTAnCGGnnnTA (ref 21)	GAAnAAA TTTnTTC	0.7	GCCGn5TCC GGAn5CGGC	1.8	TGTGAA TTCACA	0.4
				GCCn12CCCCC GGGGGn12GGC	1.4	GCAGGA TCCTGC	0.3
				GCCTn9AGGC GCCTn9AGGC	1.1	ACGTCG CGACGT	0.0
				CGGGn8CCCG CGGGn8CCCG	0.1		
				CTTn3CCC GGGn3AAG	0.0		
LEU3	RCCGGnnCCGGY (ref 18)	CACACA TGTGTG	0.4	ACCGGCGCCGGT ACCGGCGCCGGT	1.0	ACCGGC GCCGGT	2.1
		AAGAAA TTTCTT	0.4				
		CCGGn2CCGG CCGGn2CCGG	0.3				
LYS	WWWTCCRnYGGAWWW (ref 23)	ATATATA TATATAT	3.4	AAATTCCG CGGAATTT	1.9	AATTCCG CGGAATT	2.9
		AAATTCCG CGGAATTT	1.1	TCCAGCGGA TCCGCTGGA	1.0	GCCAGCG CGCTGGC	2.2
		TCCGnCGGA TCCGnCGGA	0.4	CAAn3CAC GTGn3TTG	0.2	GTGTGA TCACAC	1.0
						AGGCAG CTGCCT	0.3
						AGCCAA TTGGCT	0.0
PDR	TYTCCGCGGARY (ref 18) TCCGCGGA (ref 26) TCCGTGGA (Balzi, pers. Comm.)	TTCCGCGGAA TTCCGCGGAA	3.7	TTCCGCGGAA TTCCGCGGAA	6.7	TTCCACGGA TCCGTGGAA	7.4
		TTCCACGGA TCCGTGGAA	3.7	TTCCACGGAG CTCCGTGGAA	6.7	TCCGCGGA TCCGCGGA	2.6
		ATTTTG CAAAAT	0.3	TTCCGCGGAG CTCCGCGGAA	6.7	GCGCGA TCGCGC	0.6
				GGAn12GTA TACn12TCC	0.7	AGGCACC GGTGCCT	0.2
				CCGn12CAC GTGn12CGG	0.5	CCGCAG CTGCGG	0.2
				CCCn7CCG CGGn7GGG	0.3		
				AGCCn10GGCT AGCCn10GGCT	0.2		
				CCGn6TCC GGAn6CGG	0.1		
				AAAn8GCC GGCn8TTT	0.0		
PPR1	WYCGGnnWWYKCCGAW (ref 18)	CGGn6CCG CGGn6CCG	0.6	CGGn6CCG CGGn6CCG	0.7	TTGAAA TTTCAA	0.9
PUT3	YCGGnAnGCGnAnnnCCGA (ref 18) CGGnAnGCnAnnnCCGA (ref 19)	CGCn10CCG CGGn10CCG	0.2	CGGn10CCG CGGn10CCG	1.2	TCAAGA TCTTGA	0.2
				CCGCn8GCGG CCGCn8GCGG	0.1		
				ATCTAGAT ATCTAGAT	0.0		
UGA3	AAARCCGCSGGCGGSAWT (ref 19)	CCTn14CCG CGGn14AGG	0.2	CCTn14CCG CGGn14AGG	1.7	TCCCGCGGGA TCCCGCGGGA	1.3
				GCCn11TCC GGAn11GGC	1.0		
				GCCGCCGnCGGC GCCGnCGGCGGC	0.9		
				ACCGn2GGC GCCn2CGGT	0.9		
				AGCn2CGG CCGn2GCT	0.1		
				GCGGGA TCCCGC	0.0		
UME6	TAGCCGCCGA (ref 18)	TAGCCGCCGA TCGGCGGCTA	6.7	TAGCCGCCGA TCGGCGGCTA	6.1	TAGCCGCCGA TCGGCGGCTA	6.1
		AGAGAAAA TTTTCTCT	2.7	ACAACA TGTTGT	0.5	ACAACA TGTTGT	0.5
		ACAACA TGTTGT	0.7	CTGTTA TAACAG	0.2	CTGTTA TAACAG	0.2
				AAGCGC GCGCTT	0.1	AAGCGC GCGCTT	0.1

Patterns matching the known binding sites (18,19,21,23,26) are highlighted in blue. In the column with known consensus, the segments that are extracted from the analysis are highlighted in violet.

and leads to an over-estimation of the significance of all words included in the duplicated regions. This can be circumvented by purging the original data set, i.e. suppressing all upstream sequences that are too close to another one. After purging, the Y′ family was reduced to 11 genes, and the number of significant patterns to two hexanucleotides and three dyads. Another family required purging for a different reason: the original histone family contained five pairs of divergently transcribed genes sharing a common upstream region. For each of these pairs, we discarded one gene from the original cluster.

We combined dyad and oligonucleotide analysis on the eight families of cell cycle regulated genes (Table 4). The most significant patterns extracted from the CLN2 family are all made of a core CGCG, preceded by a poly-T and followed by a poly-A strand. In some of the patterns, the core extends to the known MBF consensus (ACGCGT), but some variants are

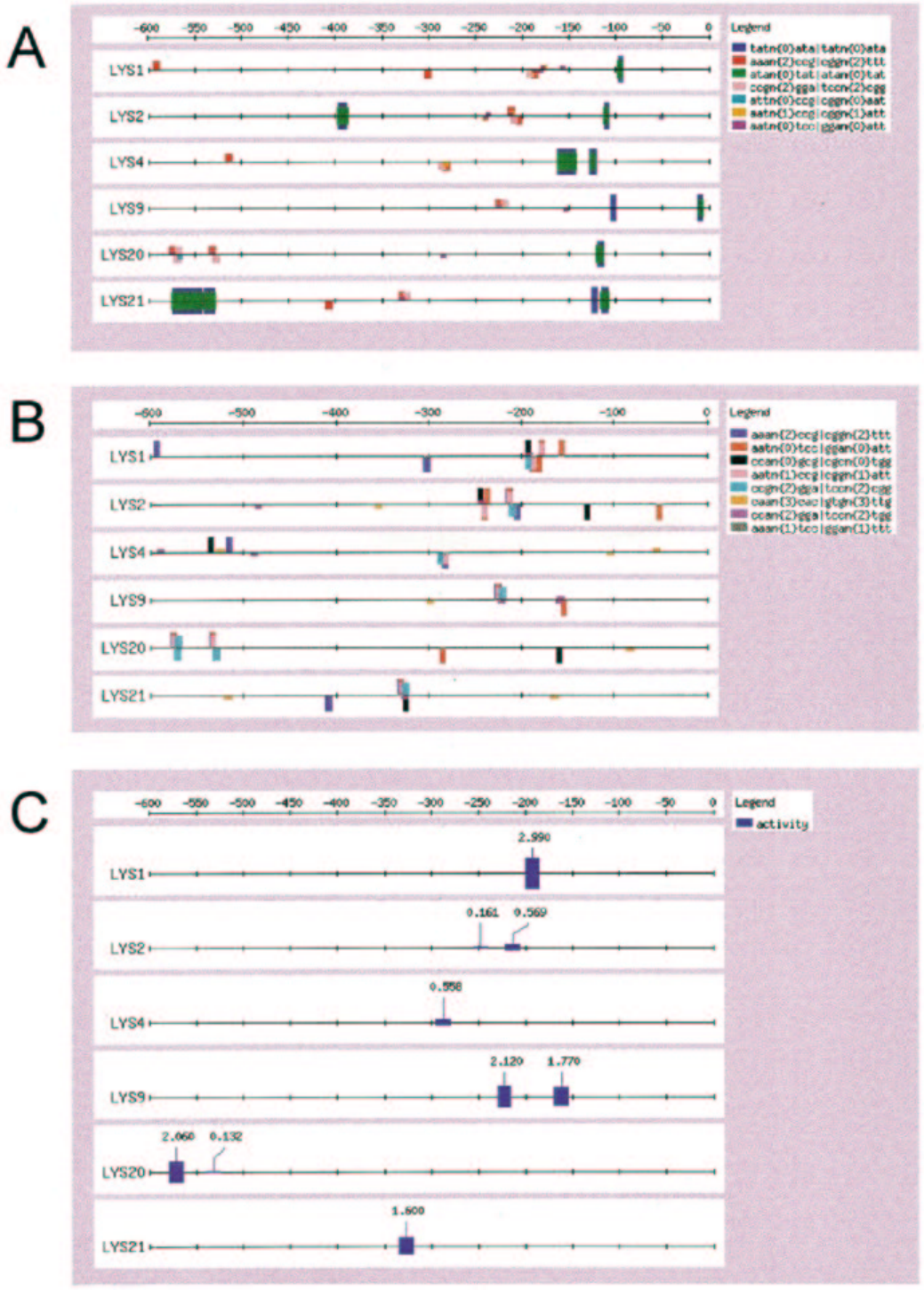

Figure 1. Comparison of predicted and known sites in the Lys family. (**A**) Patterns detected by the dyad analysis, with monad calibration. The box width reflects the statistical significance, calculated as described in Materials and Methods. (**B**) Patterns extracted with dyad analysis, using non-coding dyad frequencies as calibration. Note that the poly-TA boxes are filtered out. (**C**) Location of sites whose activity has been measured experimentally. The box thickness (or height) reflects the experimental activity as measured by Becker *et al.* (23). The label above each site shows the experimental value of activity. Only sites showing an activity >0.1 are displayed.

extracted with a similar significance: TCGCGA, ACGCGA, TCGCGT. Aligning all the CGCG-centered patterns suggests a consensus TTTKWCGCGWWWW, which would probably better reflect the properties of the regulatory sites common for this

Table 4. Pattern discovery in the upstream sequences of families extracted from a DNA chip experiment

family	oligo-analysis				dyad-analysis (non-coding dyad frequency calibration)		
	word	reverse	clpt	sig remark	dyad	reverse	clpt	sig remark
CLN2	TACGCGAA.	.TTCGCGTA		20.0 MBF; SBF variant	TTTACGCGAAAA	TTTTCGCGTAAA		20.0 MBF; SBF variant
	TACGCGTA.	.TACGCGTA		20.0 MBF; SBF	GAAAACGCGTAAA	TTTACGCGTTTTC		20.0 MBF; SBF
	TTCGCGTCG	CGACGCGAA		20.0 MBF; SBF variant	TTTTCGCGTCA.	.TGACGCGAAAA		20.0 MBF; SBF variant
	AAACGCGAA.	.TTCGCGTTT		20.0 MBF; SBF variant	TTTACGCGTCA.	.TGACGCGTAAA		20.0 MBF; SBF
	TTCGCGTCA.	TGACGCGAA		20.0 MBF; SBF variant	CGACGCGAAAA	TTTTCGCGTCG		20.0 MBF; SBF variant
	TGCCAA	TTGGCA		1.8	GAAAACGCGTCA.	.TGACGCGTTTTC		8.1 MBF; SBF
	ATCAAG	CTTGAT		1.3	AAAn8CGC	GCGn8TTT		1.9
					CAAn5CGC	GCGn5TTG		1.1
Y' (purged)	CTCGTC	GACGAG		1.8	AGTnGAG	CTCnACT		3.0
	AGTATC	GATACT		1.2	CAGn{10}ATC	GATn{10}CTG		2.0
					ATCn{12}GAG	CTCn{12}GAT		1.2
histone (purged)	CGCCCG	CGGGCG		2.6	GCGn8AGAAC	GTTCTn8CGC		3.0
	CCAGAA	TTCTGG		1.7 Mcm1	CGCCCG	CGGGCG		1.3
					ATTn2GCG	CGCn2AAT		1.3
Cell cycle MET	TGCCACAGTT	AACTGTGGCA		20.0 Met31; Met32	GCCACAGTT	AACTGTGGC		20.0 Met31; Met32
	TCACGTGA	TCACGTGA		20.0 Met4/Met28/Cbf1	GTCACGTGAC	GTCACGTGAC		20.0 Met4/Met28/Cbf1
	ACAGAG	CTCTGT		1.9	GACCACAGTTn9GTGGTC	GACCACn9AACTGTGGTC		20.0 Met31; Met32
	GACTCA	TGAGTC		0.9	CCAGn2CTGG	CCAGn2CTGG		2.1
CLB2	CCAAAG	CTTTGG		1.3	CCCn6GAA	TTCn6GGG		2.5 ECB
	CCTTCA	TGAAGG		0.9 NEG	CAAn13GCC	GGCn13TTG		0.9
					ACCn14AAT	ATTn14GGT		0.9
MCM	AGAGCA	TGCTCT		1.4	TCCCn4GGGA	TCCCn4GGGA		3.9 ECB variant
	TCCTAA	TTAGGA		1.0 Mcm1	AAAnAGG	CCTnTTT		2.8 ECB ?
					AGGn10ACT	AGTn10CCT		1.2
SIC1	AACCAGCAA	TTGCTGGTT		20.0 Swi5;Ace2	AACCAGCA....	....TGCTGGTT		20.0 Swi5; Ace2
	AGCCAGCAA	TTGCTGGCT		20.0 Swi5;Ace2	AACCAGCCAGCA	TGCTGGCTGGTT		20.0 Swi5; Ace2
	AACCAGCC	GGCTGGTT		8.0 Swi5;Ace2				
MAT	GTTTCA	TGAAAC		4.0	GTTTCA	TGAAAC		2.7
	ATCCAA	TTGGAT		2.1				

Matches between extracted patterns and known sites are indicated.

gene cluster. A related pattern is found, with a weak significance in the histone family.

The MET family contains a set of genes which, in addition to their periodic fluctuation of expression, are related by their common involvement in methionine biosynthesis and sulfur assimilation. Consistently, the most significant patterns extracted from this family correspond to the two elements regulating these metabolic pathways: TCACGTG (bound by the Met4/Met28/Cbf1 complex), and AACTGTGGCT (bound by two homolog factors: Met31p and Met32p). The dyad analysis extracts an additional variant, GACTGTGGCCG, with the highest significance.

Most regulatory elements known to play a role in the cell cycle (Table 5) were extracted from some of the families. The only exception is the SCB consensus, which could not be detected from any of the clusters. This fits with the observation by Zhang (25), who could extract this motif with the Gibbs sampler only after masking of the MCB and poly-A/poly-T tetranucleotides. The statistical assessment performed here shows that in fact the SCB consensus cannot be considered significant. This of course does not prevent the site being active in the upstream regions where it is found, but what the program tells is that this site is not a significant feature of any of the families.

Table 5. Known consensus for transcription factors involved in cell cycle and methionine biosynthesis (25)

Site	pattern	reverse compl. pattern
Mcm1	TCCTAAT	ATTAGGA
Mcm1	DCCYWWnnRG	CYNNWWRGGH
ECB	TTWCCCnnnnAGGAAA	TTTCCTNNNNGGGWAA
SFF	GTMAACAW	WTGTTKAC
Swi5;Ace2	RCCAGC	GCTGGY
SCB	CACGAAA	TTTCGTG
NEG	TnnACGCTnAAnGnC	GNCNTTNAGCGTNNA
MCB	ACGCGT	ACGCGT
Met4/Met28/Cbf1	TCACGTGA	TCACGTGA
Met31,Met32	AAAACTGTGG	CCACAGTTTT

DISCUSSION

Different approaches have been used to extract shared motifs from functionally related sequences. Each approach is based on an underlying model for the motif to be searched. In a previous work (16), we designed a detector optimized for the extraction of over-represented oligonucleotides. This was based on the biology of transcriptional regulation in yeast, where many regulatory sites consist of a highly conserved word occurring in multiple copies within an upstream region. The highest efficiency was obtained by hexanucleotide analysis, due to the structural modalities of protein–DNA interactions. Indeed, many protein–DNA-binding domains establish contact

with a limited number of adjacent nucleotides. For example, the zinc finger domain, often found in yeast transcription factors, is able to impose specificity over 3 nt. Zinc finger-containing transcription factors generally combine two tandem zinc fingers, extending the conserved DNA region to more or less 6 nt. Another important class of transcription factors contain a basic helix–loop–helix (bHLH) motif and bind DNA as dimers. The dimer acts on the DNA like tweezers, where each monomer interacts physically with a limited number of nucleotides. In the case of bHLH proteins, the two monomers contact adjacent trinucleotides, so that the DNA pattern is also a conserved hexanucleotide. Based on these general features of DNA–protein interfaces, we had previously designed the oligo-analysis program (16) and indeed, in our testing set, it performed with high efficiency in the discovery of regulatory sites with equivalent results to those of sophisticated methods. This oligonucleotide-centered approach is, however, not universal in its application. A good number of transcription factors bind to a pair of very short conserved sequences (typically three highly conserved nucleotides) separated by a non-conserved region of fixed width. This is the case for the C_6 Zn_2 binuclear cluster proteins in yeast, but also for the bacterial helix–turn–helix (HTH) factors. Motivated by these biological DNA-binding motifs, and in order to overcome the limitations of the oligo-analysis method, we developed a new algorithm addressing specifically these cases. This new program, dyad analysis, performs a systematic counting of all pairs of very short oligonucleotides (typically 3 nt), scanning a range of spacers that correspond to the known cases (e.g. 0–16 nt). When the dyad frequencies are calculated based on monad frequencies as observed in the input sequence set, the program succeeds in 9 out of the 11 families in our testing set. When using as expected dyad frequencies those observed in the complete set of genomic non-coding sequences, all regulatory sites could be extracted, but this may generate more false positives.

Despite the fact that only non-degenerated dyads are counted, the dyad analysis is to some extent able to describe internal variability in the regulatory sites. In the PDR family, the program detects two variants: TTCCgCGGA and TTCCaCGGA. Pdr1p has been shown to bind to both sequences (26). The same is true for the UME6 family, where TAGCCGCCga and TAGCCGCCc are detected, and in the CLN2 family where the dyad combination suggests a degenerate consensus TTTKWCGCGWWWW. A more exhaustive approach would be to directly assess the statistical significance of all possible degenerated patterns. The statistics presented here and in our previous single-word analysis (16) could be extended to degenerate motifs. The time of calculation would however be dramatically increased, since the number of possible trinucleotide pairs would be of the order of 15^6 instead of 4^6. Given the rate of success of the non-degenerate approach in yeast families, expanding the method to deal with degeneracy does not seem necessary, but such an extension might be valuable in the detection of bacterial regulatory sites, which present a wider degree of degeneracy. This is currently under evaluation.

Since all spacing values between 0 and 16 are scanned, one would expect the dyad analysis to be able to extract sites that are non-spaced dyads in addition to spaced ones, and thus to detect the binding sites for other factors than Zn cluster proteins as well. We tested this hypothesis on the same set of families that had been analyzed with the oligo-analysis program (16). The dyad analysis proved efficient in most of these families (Table 6). For these kinds of motifs, however, oligo-analysis remains the most efficient approach, in that the known sites are detected with a higher significance. The difference in significance for the same pattern might appear surprising at first sight. Indeed, the number of occurrences, expected occurrences and the pattern probabilities are identical in both cases. The difference is that the same probability has to be evaluated among 2080 possible hexanucleotides (grouped by pairs of reverse complements) when using oligo-analysis, whereas the dyad analysis considers 35 360 possible dyads. Thus, the same probability can result in distinct values of significance, depending on the number of patterns considered together. In summary, we recommend to use systematically both programs for the analysis of families whose sites are unknown.

The dyad analysis (as well as the previous oligo-analysis) is a rigorous algorithm (as opposed to heuristic). All possible dyads are examined and their statistical significance evaluated in a single run. The program is able to return multiple patterns in a family that would be co-regulated by multiple transcription factors. We did not find any example of a family regulated by multiple Zn cluster proteins, but in the MET family, the dyad analysis detected both Met28p/Met4p/Cbf1p and Met31p binding sites (Tables 4 and 6).

The dyad analysis program calculates the probability and the statistical significance for each pattern, permitting the extraction of candidate sites on a probabilistic basis. An important practical consequence of this probabilistic foundation is the ability of the dyad analysis and oligo-analysis programs to return a negative answer. To know that, within a family, there is not a single pattern that can be considered over-represented may be quite relevant in future combined approaches to the analysis of global expression levels. This capability to return a negative answer strongly reduces the ratio of false positives, in comparison with other programs like MEME (27), coresearch (12) or the Gibbs sampler (28), which systematically return the top-ranking patterns, irrespective of their statistical significance.

Another advantage of using a uniform statistical foundation for both oligo-analysis and dyad analysis is that it enables in principle one to directly compare the significance values obtained by the respective programs, in cases where the pattern searched is absolutely unknown and there is no hint as to whether the relevant pattern follows any of the two biological models supporting these alternative methods.

So far, C_6 Zn_2 binuclear cluster proteins have only been found in fungal organisms. It is, however, expected that the dyad analysis can have a wide range of applicability given the role of symmetry in regulatory patterns in general. One large potential field of application is the discovery of regulatory sites in prokaryotic genomes where the HTH is by far the dominant motif of regulatory proteins and where homomultimers are dominant. In *Escherichia coli* alone, HTH proteins represent 290 out of the 314 predicted transcription factors (28). Similarly to the case of C_6 Zn_2 cluster proteins, HTH proteins form dimers or multimers that bind to direct or inverted repeats separated by a short variable sequence. We are currently evaluating the efficiency of the dyad analysis in *E.coli* (manuscript in preparation).

Global transcriptome and micro-array methodologies are currently generating an explosion of experimental data and emphasizing the need for analytic methods of the kind reported

Table 6. Dyad analysis is also able to extract binding sites recognized by other transcription factors than C_6 Zn_2 binuclear cluster proteins; gene families from (16). It is, however, less efficient than oligo-analysis for these classes of transcription factors

family	known	oligo-analysis			dyad-analysis (non-coding dyad freq. calibration)		
		word	rev.comp.	sig	dyad	reverse complement	sig
NIT	GATAAG	TCTTATC	GATAAGA	20.0	TCTTATC	GATAAGA	20.0
		ACATCT	AGATGT	0.4	CAGn2CGG	CCGn2CTG	0.6
		CTGATA	TATCAG	0.1			
MET	TCACGTG	ATCACGTGAT	ATCACGTGAT	9.0	GTCACGTGAC	GTCACGTGAC	7.7
		ATCACGTGAC	GTCACGTGAT	9.0	GCCACAGTT	AACTGTGGC	2.4
	AAAACTGTGG	CGCCACAGTT	AACTGTGGCG	3.7	GCAn6GGC	GCCn6TGC	0.4
		GCTTCC	GGAAGC	0.4	GGGn4CAC	GTGn4CCC	0.2
		CCGGAG	CTCCGG	0.3	CAGn7ATC	GATn7CTG	0.0
		CCCATC	GATGGG	0.2			
		CACGCC	GGCGTG	0.1			
		CTTGAC	GTCAAG	0.0			
PHO	GCACGTGGG	CGCACGTGCG	CGCACGTGCG	3.3	CTGCACGTGCAG	CTGCACGTGCAG	2.1
		CGCACGTGGG	CCCACGTGCG	3.3	CTGCACGTGCG	CGCACGTGCAG	2.1
	GCACGTTTT	AAACGTGCG	CGCACGTTT	3.3	CTGCACGTGG	CCACGTGCAG	2.1
		TGCCAA	TTGGCA	2.6	TGCCAA	TTGGCA	1.3
		CTGCAC	GTGCAG	1.8	ACAn14TGCA	TGCAn14TGT	1.0
		ATTAGC	GCTAAT	0.0	CACAn14GGCA	TGCCn14TGTG	1.0
					ACAn6GCA	TGCn6TGT	0.4
					GCTn8CCC	GGGn8AGC	0.3
					ACGn6AGC	GCTn6CGT	0.3
GCN4	RRTGACTCTTT	TGAGTCAT	ATGACTCA	8.5	TGAGTCAT	ATGACTCA	7.3
		TGAGTCACT	AGTGACTCA	8.5	TGAGTCAC	GTGACTCA	7.3
		AGAGTCAT	ATGACTCT	8.5	AGAGTCAT	ATGACTCT	7.3
		ATGACTCC	GGAGTCAT	8.5	CAGCGG	CCGCTG	2.5
		CAGCGG	CCGCTG	3.8	GCCn10CAC	GTGn10GGC	0.9
		TAGTCAT	ATGACTA	3.7	CCGn14GCA	TGCn14CGG	0.5
		ACCGGC	GCCGGT	1.3	GCCGGTACCGGC	GCCGGTACCGGC	0.4
		CCACAG	CTGTGG	0.8	CACn16GTC	GACn16GTG	0.2
		CACGTG	CACGTG	0.8	AGCn5CGG	CCGn5GCT	0.0
		GCGTGA	TCACGC	0.1			
		CATCGAA	TTCGATG	0.1			
INO	CATGTGAAWT	CAACAAG	CTTGTTG	3.2	CAACAA	TTGTTG	1.9
		CACATG	CATGTG	2.0	CACATGTG	CACATGTG	0.7
		TGTGAA	TTCACA	0.9	GACn4ACC	GGTn4GTC	0.6
		ACAACG	CGTTGT	0.7	CCAn10GTC	GACn10TGG	0.5
		TCTTCA	TGAAGA	0.6			
		TGACCA	TGGTCA	0.1			
HAP2-5	CCAAY	GCAGGAA	TTCCTGC	1.6	CCAn4ATC	GATn4TGG	1.3
		CCCGCG	CGCGGG	0.9	CCAn13GCC	GGCn13TGG	1.3
		CAGCAG	CTGCTG	0.9	CCCn11CCA	TGGn11GGG	0.7
		CGAACG	CGTTCG	0.2	GCCGn7CGGC	GCCGn7CGGC	0.6
		TGGACA	TGTCCA	0.0	GGCn2CGA	TCGn2GCC	0.5
					ACCnTGC	GCAnGGT	0.4
					GCAGGA	TCCTGC	0.4
					TTGGCCAA	TTGGCCAA	0.1
TUP	KAnWWWWATSYGGGGW	CGTGGGGTA	TACCCCACG	20.0	ATCGTGGGGTAT	ATACCCCACGAT	20.0
		TACCCCGC	GCGGGGTA	20.0	ATACCCCGCA	TGCGGGGTAT	20.0
		GACCCCACG	CGTGGGGTC	7.1	AGGCAC	GTGCCT	3.1
		AGGCACGGG	CCCGTGCCT	4.4	TCCn13AAAA	TTTTn13GGA	2.8
		CAGGGGC	GCCCCTG	1.6	GGAnAAAT	ATTTnTCC	1.7
		CACAAACA	TGTTTGTG	1.1	GCACn9GCAC	GTGCn9GTGC	1.6
		AAGGAGGA	TCCTCCTT	0.6	CCCACGATn10TGGG	CCCAn10ATCGTGGG	1.3
		CTCCGC	GCGGAG	0.3	ATCAGGGGCA	TGCCCCTGAT	1.1
		TCTGCA	TGCAGA	0.2	CGCGn14CCC	GGGn14CGCG	1.0
		CGTAGC	GCTACG	0.1	CCCn7AAGA	TCTTn7GGG	0.9
		ACGGAG	CTCCGT	0.0	GCAn14GGC	GCCn14TGC	0.7
					GCAnACA	TGTnTGC	0.4
					GGAn10GAA	TTCn10TCC	0.3
					AATn7GGG	CCCn7ATT	0.1
					CCCn7GGG	CCCn7GGG	0.0
YAP	TTACTAA	CGTTCC	GGAACG	3.4	CCGTTCC	GGAACGG	2.1
		CATTAC	GTAATG	1.0	GCTn3TAA	TTAn3AGC	1.7
		CTGAAG	CTTCAG	1.0			

Patterns matching the known binding sites are highlighted in bold uppercase.

here (29). We applied our pattern discovery programs to 11 gene clusters that had been collected from DNA chip experiments (the eight families of Table 4 plus the GAL4-chip, YAP and TUP families). The algorithms performed in these families with similar efficiency as in known regulons. A critical step to reach this efficiency is to purge the upstream sequence sets, in order to discard whole duplications of upstream regions (frequently found in telomeric regions), and to avoid including twice an intergenic region shared by two divergently transcribed genes. Predictions will have to be validated experimentally, and computational methods could in turn take benefit from such evaluations to improve their performance. By combining the analysis of upstream regions as performed here with global approaches to functional assignment such as microarray technologies (30), operon organization (31), gene fusion analysis (32–34) and phylogenetic profiles (35), one can hope to generate a stronger biological foundation to the immense task ahead in the integrated understanding of the biology of a single cell.

ACKNOWLEDGEMENTS

We acknowledge referees for useful comments and suggestions. J.v.H. was partly funded by the European Commission (EU grant QLRI-CT-1999-01333) and the Actions de Recherche Concertées de la Communauté Française de Belgique. A.F.R. was supported by a Ph.D. fellowship from DGAPA-UNAM. J.C.-V. was supported by grants from DGAPA-UNAM and Conacyt.

REFERENCES

1. Hieter,P. and Boguski,M. (1997) *Science*, **278**, 601–602.
2. DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) *Science*, **278**, 680–686.
3. Chu,S., DeRisi,J., Elsen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) *Science*, **282**, 699–705.
4. Waterman,M.S., Arratia,R. and Galas,D.J. (1984) *Bull. Math. Biol.*, **46**, 515–527.
5. Mengeritsky,G. and Smith,T.F. (1987) *Comput. Appl. Biosci.*, **3**, 223–227.
6. Stormo,G.D. and Hartzell,G.W.d. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.
7. Hertz,G.Z., Hartzell,G.W.d. and Stormo,G.D. (1990) *Comput. Appl. Biosci.*, **6**, 81–92.
8. Lawrence,C.E. and Reilly,A.A. (1990) *Proteins*, **7**, 41–51.
9. Cardon,L.R. and Stormo,G.D. (1992) *J. Mol. Biol.*, **223**, 159–170.
10. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) *Science*, **262**, 208–214.
11. Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) *Protein Sci.*, **4**, 1618–1632.
12. Wolfertstetter,F., Frech,K., Herrmann,G. and Werner,T. (1996) *Comput. Appl. Biosci.*, **12**, 71–80.
13. Quandt,K., Grote,K. and Werner,T. (1996) *Comput. Appl. Biosci.*, **12**, 405–413.
14. Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) *Nature Biotechnol.*, **16**, 939–945.
15. Hertz,G.Z. and Stormo,G.D. (1999) *Bioinformatics*, **15**, 563–577.
16. van Helden,J., André,B. and Collado-Vides,J. (1998) *J. Mol. Biol.*, **281**, 827–842.
17. van Helden,J., Olmo,M. and Perez-Ortin,J.E. (2000) *Nucleic Acids Res.*, **28**, 1000–1010.
18. Schjerling,P. and Holmberg,S. (1996) *Nucleic Acids Res.*, **24**, 4599–4607.
19. Todd,R.B. and Andrianopoulos,A. (1997) *Fungal Genet. Biol.*, **21**, 388–405.
20. van Helden,J., André,B. and Collado-Vides,J. (2000) *Yeast*, **16**, 177–187.
21. Svetlov,V.V. and Cooper,T.G. (1995) *Yeast*, **11**, 1439–1484.
22. Vuidepot,A.L., Bontems,F., Gervals,M., Guiard,B., Shechter,E. and Lallemand,J.Y. (1997) *Nucleic Acids Res.*, **25**, 3042–3050.
23. Becker,B., Feller,A., el Alami,M., Dubois,E. and Pierard,A. (1998) *Mol. Microbiol.*, **29**, 151–163.
24. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) *Mol. Biol. Cell*, **9**, 3273–3297.
25. Zhang,M.Q. (1999) *Comput. Chem.*, **23**, 233–250.
26. Balzi,E and Goffeau,A. (1995) *J. Bioenerg. Biomembr.*, **27**, 71–76.
27. Grundy,W.N., Bailey,T.L. Elkan,C.P. and Baker,M.E. (1997) *Comput. Appl. Biosci.*, **13**, 397–406.
28. Pérez-Rueda,E. and Collado-Vides,J. (2000) *Nucleic Acids Res.*, **28**, in press.
29. Goffeau,A. (1998) *Nature Biotechnol.*, **16**, 907–908.
30. Brown,P.O. and Botstein,D. (1999) *Nature Genet.* (Suppl. 1), **21**, 33–37.
31. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
32. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) *Science*, **285**, 751–753.
33. Ouzounis,C. (1999) *Trends Genet.*, **15**, 445.
34. Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) *Nature*, **402**, 83–86.
35. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.