Yeast Functional Analysis Report

A web site for the computational analysis of yeast regulatory sequences

J. VAN HELDEN[1,3]*, 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, Mexico*
[2] *Laboratoire de Physiologie Cellulaire et de Génétique des Levures, Université Libre de Bruxelles, CP300, Aéropole de Gosselies, rue des Professeurs Jeener et Brachet 12, 6041 Gosselies, Belgium*
[3] *Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP160/16, 50 av F.D. Roosevelt, B-1050 Bruxelles, Belgium*

A series of computer programs were developed for the analysis of regulatory sequences, with a special focus on yeast. These tools are publicly available on the web (http://copan.cifn.unam.mx/Computational_Biology/yeast-tools or http://www.ucmb.ulb.ac.be/bioinformatics/rsa-tools/). Basically, three classical problems can be addressed: (a) search for known regulatory patterns in the upstream regions of known genes; (b) discovery of unknown regulatory patterns within a set of upstream regions known to be co-regulated; (c) search for unknown genes potentially regulated by a known transcription factor. Each of these tasks can be performed on basis of a simple (string) or more refined (matrix) description of the regulatory patterns. A feature-map program automatically generates visual representations of the positions at which patterns were found. The site also provides a series of general utilities, such as generation of random sequence, automatic drawing of XY graphs, interconversions between sequence formats, etc. Several tools are linked together to allow their sequential utilization (piping), but each one can also be used independently by filling the web form with external data. This widens the scope of the site to the analysis of non-regulatory and/or non-yeast sequences. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS — bioinformatics; yeast; transcriptional regulation; sequence analysis

INTRODUCTION

The release of the complete genomic sequence from the yeast *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996) not only provided a huge amount of data for the molecular biologist, but also stimulated the development of new computational techniques for the analysis of these data. The first wave of analysis of these genomic sequences focused mainly on coding sequences (ORF pre-diction, classification according to sequence similarities), and lesser attention was paid to non-coding sequences. This probably reflects the history of molecular and computational biology, but is somewhat unfortunate given the biological importance of these sequences, which mediate the regulation of gene expression.

Three typical situations occur when analysing DNA regulatory regions: (a) the regulatory pattern (e.g. the consensus for a transcriptional factor), as well as the regulated genes, are known. The search is focused on finding matching positions of putative binding sites; (b) the regulatory pattern is known but the regulated genes are unknown. Complete genomes can be scanned to

*Correspondence to: J. van Helden, Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP160/16, 50 av F.D. Roosevelt, B-1050 Bruxelles, Belgium. E-mail: jvanheld@ucmb.ulb.ac.be, collado@cifn.unam.mx

identify genes possibly regulated by a given transcription factor; (c) the co-regulated or co-expressed genes are known but the regulatory pattern is unknown. The goal is to detect elements shared by a set of functionally related upstream sequences, which could reveal unknown regulatory sites. This question is raised when analysing sets of genes obtained in global gene expression experiments using micro-array technologies (deRisi *et al.*, 1997).

In a previous paper (van Helden *et al.*, 1998), we described a simple but efficient methodology for discovering unknown regulatory elements shared by a set of co-regulated genes. This methodology proved efficient for the detection of regulatory elements from most of the families studied. The program (*oligo-analysis*) has been integrated in a set of modular tools (*yeast-tools*), that can either be used individually, or combined in different ways to answer questions related to transcriptional regulation. The tools are accessible via web interfaces (http://copan.cifn.unam.mx/Computational_Biology/yeast-tools or http://www.ucmb.ulb.ac.be/bioinformatics/rsa-tools/). The aim of the present paper is to provide a brief description of these tools, and to show how they can be combined to answer different questions. Since our initial publication, the capabilities of the site have been extended by the addition of several new programs complementing the previous ones. A new program (*dyad-detector*) allows an efficient detection of spaced dyads that are characteristic binding sites for an important family of yeast transcription factors: the binuclear Zn clusters. Three other programs (patser, consensus, and the gibbs sampler) developed externally (Hertz *et al.*, 1990; Hertz and Stormo, 1999; Lawrence *et al.*, 1993; Neuwald *et al.*, 1995), have been interfaced to support analyses based on a matrix description of the regulatory elements.

BINDING SITE DESCRIPTION

Before presenting the programs available for the analysis of non-coding sequences, it is worth introducing the different representations that can be used to describe regulatory sites. Transcription factors bind to specific DNA sequences. In most yeast transcription factors, the DNA sequence in direct contact with the factor is short (5–8 nucleotides). The nucleotide–amino acid interaction is highly specific, so that these central bases are generally well conserved among all the sites bound by the same factor. Some degeneracy can, however,

occur at certain positions, as illustrated by the binding specificity of Pho4p, for which several binding sites have been isolated and characterized experimentally (Table 1A). The sites showing the highest affinity for Pho4p all follow pretty well the motif GCACGTGGG. For some sites, the G-rich 3′ end is replaced by a T-rich segment. These variants have been shown to bind Pho4p with a medium affinity, and participate in the transcriptional activation of several genes involved in phosphate utilization (Oshima *et al.*, 1996).

Various methods can be used to represent the specificity of the factor–DNA binding. The simplest representation is the exact consensus, which retains at each position the base observed with the highest frequency among the known binding sites. This description relies on only the letters A, C, G and T. String search programs can then be used to localize the matching positions, with a specified number of allowed mismatches. A slightly more refined description consists in using the IUPAC degenerate code, where additional letters represent combinations of bases. With this notation, the Pho4p consensus binding site will be noted GCACGTKKK (Table 1B), where K stands for T or G. This description reflects better the fact that more than one base can be found at a given position while another can not, but it is still not informative as to the preferred base. Such a refinement of the description can be obtained with position frequency matrices, showing the frequency observed for each base at each position of the aligned binding sites (Table 1C). Such matrices can be generated by specific programs available on the site (consensus, gibbs), or imported from databases on transcriptional regulation, such as TRANSFAC (Heinemeyer *et al.*, 1998).

THE WEB INTERFACE

The home page is divided in two frames (Figure 1). The left frame presents a menu allowing navigation among the different tools. The right frame changes depending on the item selected in the menu. We tried to make the tools accessible for users who are not familiar with bioinformatics. Several pages have been written to provide on-line help at each step. An introduction page briefly summarizes the functionality of each tool. A tutorial based on concrete and illustrative examples has been designed to guide new users in the first steps and through the modularity of the programs. A more detailed explanation can be

Table 1. Different ways to represent the binding specificity of Pho4p.

Gene	Site name	Sequence	Affinity
A. *Multiple alignment* adapted from Oshima *et al.* (1996)			
PHO5	UASp2	`---ACTCACACACGTGGGACTAGC-`	High
PHO84	Site D	`---TTTCCAGCACGTGGGGCGGA--`	High
PHO81	UAS	`----TTATGGCACGTGCGAATAA--`	High
PHO8	Proximal	`GTGATCGCTGCACGTGGCCCGA---`	High
PHO5	UASp3	`--TAATTTGGCATGTGCGATCTC--`	Low
PHO84	Site C	`-----ACGTCCACGTGGAACTAT--`	Low
PHO84	Site A	`-----TTTATCACGTGACACTTTTT`	Low
Group 1	Consensus	`---------GCACGTGGGAC-----`	High–low
PHO5	UASp1	`--TAAATTAGCACGTTTTCGC----`	Medium
PHO84	Site E	`----AATACGCACGTTTTTAATCTA`	Medium
PHO84	Site B	`-----TTACGCACGTTGGTGCTG--`	Low
PHO8	Distal	`---TTACCCGCACGCTTAATAT---`	Low
Group 2	Consensus	`-----AT-CGCACGTTTT-------`	Med–low
B. *IUPAC string*			
Groups 1 and 2	Consensus	`---------GCACGTKKK------`	High or medium

C. *Position frequency matrix (TRANSFAC matrix F$PHO4−01)*

pos	1	2	3	4	5	6	7	8	9	10	11	12
A	1	3	2	0	8	0	0	0	0	0	1	2
C	2	2	3	8	0	8	0	0	0	2	0	2
G	1	2	3	0	0	0	8	0	5	4	5	2
T	4	1	0	0	0	0	0	8	3	2	2	2

found in the manual page associated to each tool. A demo button leads to a form pre-filled with some typical data. For each form, default parameters have been selected that reflect the optimal conditions found during our testing phase.

Our site provides an integrated approach to the analysis of regulatory sequences: the programs are interconnected in such a way that the result of a request can automatically be sent as input for a subsequent request (Figure 2). Starting from a set of gene names as initial input, all the steps from sequence retrieval to drawing of a feature map can be performed in the same web site.

THE FIRST STEP: SEQUENCE RETRIEVAL

The program *upstream-sequence* returns the upstream sequences for a set of selected genes.

Queries can be entered in the form of ORF identifiers or gene names. Synonyms are recognized. The range of sequence to retrieve can be selected by the user. Given the lack of good information about transcription start, the ATG codon has been taken as coordinate 0 for all genes. The sequences of several genes can be retrieved simultaneously, allowing their collective analysis by subsequent tools of the site.

A classical dilemma when analysing upstream sequences is whether or not they should be allowed to overlap with a predicted ORF located upstream of the gene of interest. It might sound simpler to consider only non-coding sequences for analysis of transcriptional regulation, but on the one hand one cannot *a priori* exclude that a coding sequence could at the same time play a regulatory role on a neighbour gene. On the other hand,
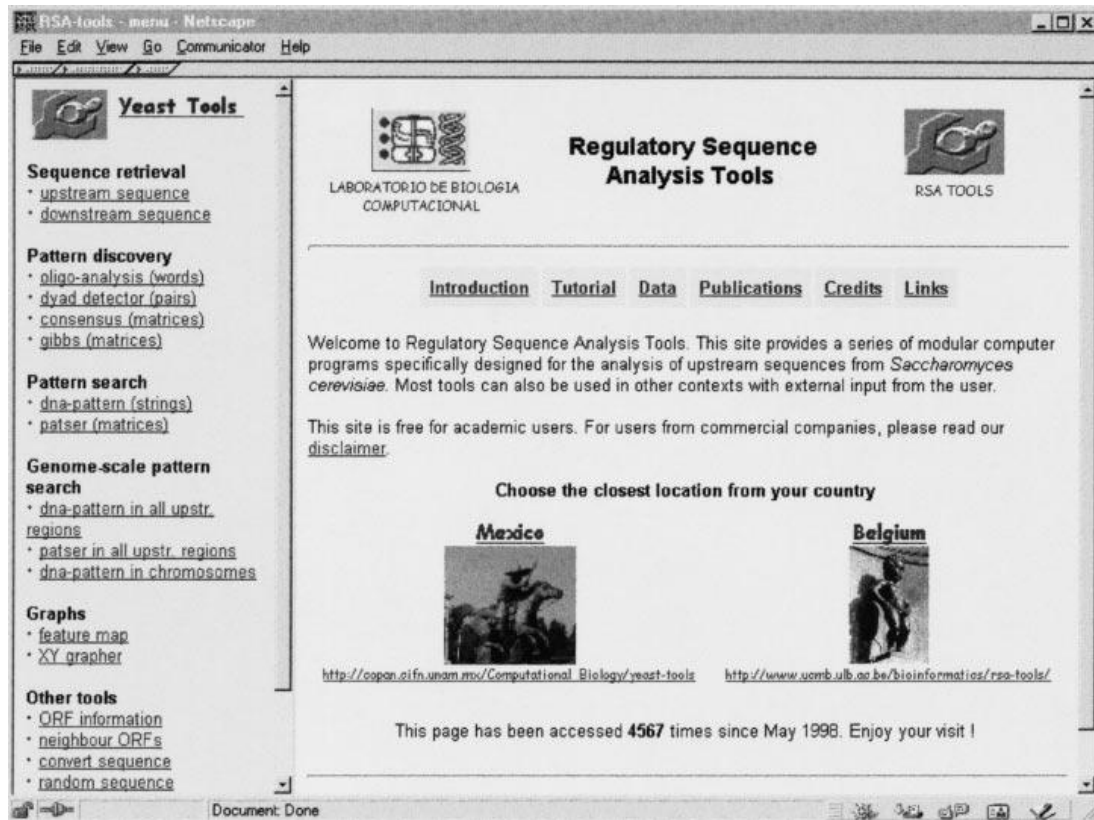
Figure 1.   Homepage of the web site. The left frame provides a menu of all available tools. Clicking on any item displays the corresponding form in the right frame.

some of the predicted ORFs are likely to be false-positives, and discarding them would prevent from detecting real regulatory sequences. By default, we allow inclusion of predicted coding sequences in the upstream sequences, but the user can inactivate this option.

PATTERN DISCOVERY: DETECTION OF UNKNOWN REGULATORY SITES FROM A SET OF CO-REGULATED GENES

Let us start from the case where we know that a set of genes is co-regulated without knowing the elements responsible for their regulation. This is a typical situation in results from large-scale gene expression measurements. These experiments isolate a family of genes that coordinately respond to a change in the culture medium (e.g. the diauxic shift), or to a change in the expression of a precise

transcription factor (controlled expression or deletion).

Building gene families

A crucial issue for the success of any pattern discovery program is to select families that are likely to be regulated by a common transcription factor. Certainly, an important lesson from the first published gene expression measurement (deRisi *et al.*, 1997) was that a tremendous number of genes show a significant response: during the diauxic shift, 1000 genes were upregulated, and 700 downregulated, by at least a factor of 2. This complex response involves several transcription factors; therefore, it does not make sense to analyse all these genes as a single family for the purpose of discovering a common regulatory motif. Additional information can, however,
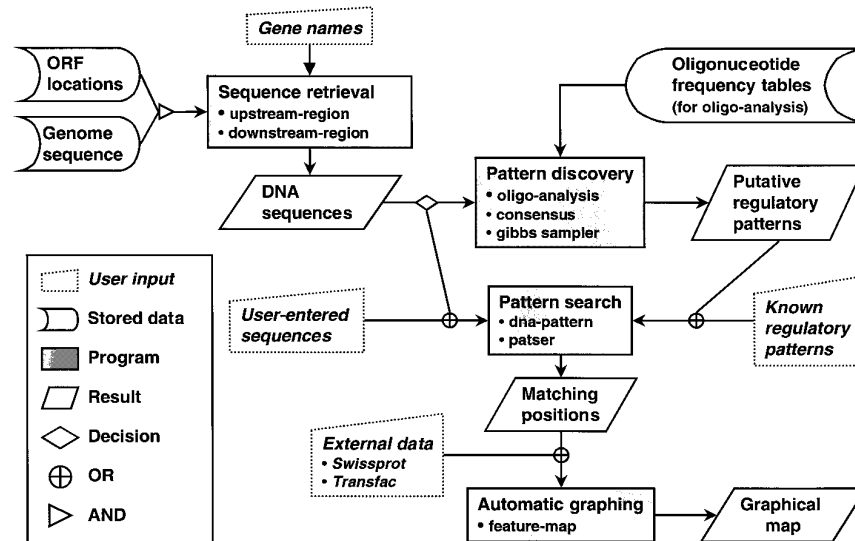
Figure 2. Flow chart of the tools.

be exploited to select smaller subfamilies, which are likely to associate genes in closer functional relationships. Indeed, deRisi *et al.* (1997) took seven measurements at 2 h intervals, so that each gene is associated with a temporal profile of expression. Among the upregulated genes, distinct types of profiles are observed: some genes are activated early and remain so until the end of the experiment, while others are activated only during the last steps. More complex profiles are also observed such as a transient activation followed by a decrease in expression down to the initial level. Restricted families selected on the basis of expression profile similarity are more likely to respond to a common transcription factor, and indeed they lend themselves much better to pattern discovery (unpublished results).

Families should also include a reasonable minimum of members. Pattern discovery programs detect the fact that a pattern is unexpectedly repeated among the set of upstream sequences. Some patterns can, however, be found in several sequences by chance, without being involved in a common transcriptional response. Analysing larger families increases the number of occurrences of the real regulatory site, while reducing the importance of random patterns, thus improving the signal-to-noise ratio. Empirically, we found that significant patterns were extracted from families of five genes or more.

However, smaller families can sometimes provide good results, due to the repetition of the regulatory pattern within the same upstream sequence.

String discovery programs: oligo-analysis and dyad-detector

We developed two programs that detect over-represented elements in a set of DNA sequences. The first program (*oligo-analysis*) detects regulatory sequences built around a central well-conserved core. The second program (*dyad-detector*) detects pairs of words spaced by a non-conserved region.

The word detector (*oligo-analysis*) has been described before (van Helden *et al.*, 1998). The method is based on a systematic counting of occurrences for all the possible words of a given length. The program has been optimized for yeast upstream sequences by the use of specific oligo-nucleotide frequency tables for the estimation of the significance of each pattern. Despite its simplicity, the method has proved efficient for the detection of the regulatory patterns for most functional families from the testing set. These families were built on the basis of experimental information, and the program was able to isolate the known regulatory motifs, plus a few other motifs which could represent binding sites for unknown transcription factors. An advantage of

the method is that it is able to detect all the over-represented patterns of a given length in a single run. Variants of a pattern are sometimes revealed by a combination of related words. For instance, the analysis of the Pho4p targets isolates two significant patterns: CACGTGGG and CACGTTT. Thanks to its simplicity, the analysis is very fast and can deal with large sequence sets. For instance, the program has been used to analyse the complete set of 6217 downstream regions encompassing 200 bp each, for the detection of signal termination signals (van Helden, del Olmo and Perez-Ortin, in preparation), and is able to perform the task in less than 1 min.

The single-word detector was sufficient to extract the vast majority of yeast regulatory elements, but failed in a specific case, the GAL family (van Helden *et al.*, 1998), and this for an understandable reason: the conserved bases of the Gal4p binding site are two trinucleotides separated by a 11 base-wide non-conserved region. Such spaced pairs are very often found in regulatory sites and generally reflect the fact that a transcription factor binds DNA in a homodimeric form: each monomer contacts half of the conserved site. Some constraints may exist on the composition of the intervening region, such as an AT richness, which influences DNA conformation, but these bases are much less conserved than those entering into direct contact with transcription factor residues. We developed a program, *dyad-detector*, that specifically scans a set of co-regulated sequences for over-represented word pairs, with a variable spacing region. This program detects a Gal4p binding motif with a very high significance level (Table 2). Gal4p is a member of the large family of C_6 binuclear zinc-

cluster transcription factors (Schjerling and Holmberg, 1996). Binding to spaced dyads is a recurrent feature of these transcription factors, opening a large field of application for the dyad-detector in the analysis of new yeast gene families (van Helden *et al.*, in preparation). The same methodology is also efficient to extract binding sites of numerous *Escherichia coli* helix–turn–helix transcription factors (Ríos *et al.*, in preparation).

Matrix discovery: consensus and gibbs sampling

Several programs are available that extract a position frequency matrix representing patterns shared by a set of unaligned sequences. We tested them with the same dataset as in van Helden *et al.* (1998) and created a web interface for two of them, which gave satisfactory results in the analysis of yeast upstream sequences (not shown): *consensus* (Hertz *et al.*, 1990; Hertz and Stormo, 1999), and the *gibbs sampler* (Lawrence *et al.*, 1993; Neuwald *et al.*, 1995). Another powerful approach for pattern discovery is the use of Hidden Markov Models (HMM; Durbin *et al.*, 1998). We did not install the HMM algorithm on our site, but one implementation is already available on the web (YEBIS; Yada *et al.*, 1998), which provides good results when tested with the same data set (not shown).

Matrix-based pattern discovery offers, compared with string-based algorithms, the advantage of providing a more sensitive description of putative binding sites, accounting for the variability observed at each position. However, the use of these programs requires some experience and the interpretation of the results is less straightforward than in the case of the oligonucleotide analysis. These algorithms are also much more

Table 2.   Example of pattern discovery with the dyad-detector.

Alignment	Pattern		Exp freq	Obs occ	Exp occ	Sig
`cgg............ccg.`	`cggn{11}ccg`	`cggn{11}ccg`	0.00007	20	0.62	7.0
`cgg............cga`	`cggn{12}cga`	`tcgn{12}ccg`	0.00006	10	0.55	4.7
`cgg..........tcc..`	`cggn{10}tcc`	`ggan{10}ccg`	0.00011	10	1.02	2.2
`gcc..........tccga`	**assembly**					

A family of six genes regulated by Gal4p was analysed (GAL1, GAL2, GAL7, GAL80, MEL1 and GCY1). The 800 bp upstream sequence was retrieved for each gene and the number of occurrences was counted for all the trinucleotide pairs spaced by a region of 0–20 nucleotides. Occurrences were counted on both strands, stating that the regulatory elements were acting in an orientation-insensitive way, as is generally the case in yeast. Among the ∼40 000 possible patterns, not more than three are significantly over-represented. These three dyads are mutually overlapping and can be assembled into a contiguous pattern (assembly), which fits the known Gal4p binding site (Schmerling and Holmberg, 1996). Exp freq, expected frequency; Obs occ, observed number of occurrences; Exp occ, expected number of occurrences; Sig, significance index.

computer-demanding than word counting, so that the amount of sequences they can analyse is more limited. A good trade-off is to combine the use of string and matrix programs: oligo-nucleotide and dyad analyses provide a fast and systematic scanning of all over-represented patterns, but the pattern description generated is simpler. It allows, however, detection of the patterns that should be expected with the matrix search programs, and it provides an estimate of the optimal size and number of occurrences of the significant patterns, both of which are important input parameters for consensus and gibbs.

PATTERN SEARCH: LOOKING FOR KNOWN PATTERNS IN UPSTREAM SEQUENCES

If the regulatory pattern is already known from experimental sources, or once a putative pattern has been discovered with the above methods, one can then search all matching positions within a given set of upstream sequences. The search can be performed on a selection of yeast upstream sequences, or on the complete set of 6217 upstream sequences from the yeast genome. The web site allows performing searches with either a string or a matrix description of the regulatory pattern. A valuable source of known pattern descriptions is TRANSFAC, the database of transcription factors (Heinemeyer *et al.*, 1998). The current version holds a string description for around 300 yeast binding sites, and 21 matrix descriptions for yeast transcription factors (*S. cerevisiae* represents ~7% of TRANSFAC contents).

String searches: dna-pattern

DNA and IUPAC strings can be searched with the program *dna-pattern*. The program accepts a set of several sequences as input, allowing to search the pattern within several co-regulated genes at once. Several patterns can also be entered simultaneously in the web query form. Searches can be performed on either or both strands. Separate forms allow using the same program for searching on whole chromosomes (*genome-search*) or on the set of upstream sequences from the 6217 predicted ORFs (*all-upstream-search*).

Matrix searches: patser

For matrix searches, we implemented a web interface for the program *patser* (Hertz and Stormo, 1999). As for string searches, the program can accept custom input sequences, or can be used to scan the set of all upstream sequences (*all-upstream-patser*).

Searches with a library

Instead of providing a single pattern description, one might want to scan an upstream sequence for matches with any known yeast transcription factor. This facility is already provided by several web sites: TRANSFAC (Heinemeyer *et al.*, 1998), ooTFD (Ghosh, 1999). These sites combine pattern search programs with a library of pattern descriptions (strings or matrices). All these programs allow the selection of a subset of the library limited to yeast transcription factors. Currently, our site does not provide these options, so the interested user is referred to the existing sites.

GRAPHICAL REPRESENTATION: THE FEATURE MAP PROGRAM

After having detected all matching positions for the regulatory patterns, the *feature-map* program creates a graphical representation of their relative location (Figure 3). The program can represent several sequences on the same map, allowing a synthetic view of the family of upstream sequences. The graph displayed is dynamic; i.e. positioning the cursor upside a given feature displays the corresponding information (position, feature type, sequence) on the bottom of the browser window. *feature-map* accepts input not only from our pattern search programs, but also from a variety of external sources. For instance, the feature field of the Swissprot entries can be automatically extracted to draw a comparative map of protein features. Feature maps can also be generated from the output of the programs that scan regulatory sequences with a library of known patterns (MatInspector, Signal Scan).

CONCLUSION

Whilst most genome informatic centres provide mainly information on the coding sequences, our tools are focused on the analysis of transcriptional regulation. We offer a set of computer programs, each of which can be accessed independently. The tools can also be combined in a pre-defined concatenation, following an order that depends
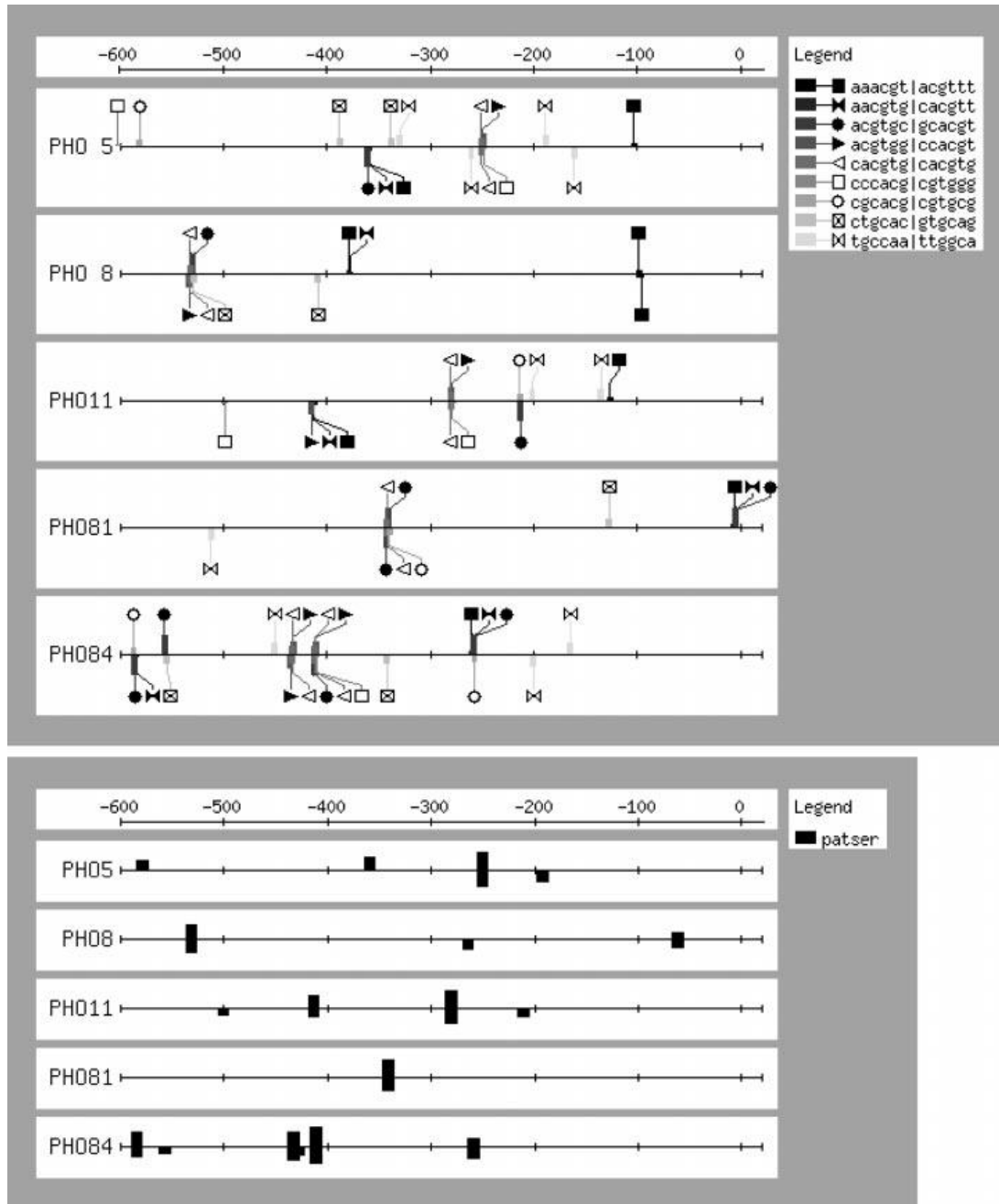
Figure 3.   Feature-map of predicted regulatory sites in the PHO genes. **A**: result of a string discovery (oligo-analysis) followed by a string search (dna-pattern). **B**: matrix search (patser) with the Pho4p position frequency matrix shown in Table 1C. Several features and several genes can be represented at once, allowing visual comparison. Each *symbol* corresponds to a different word, the *thickness and gray scale* reflect the prediction score. *On the web site, colors can also be used to represent the distinct patterns.* Coordinates are counted in negative units, taking the start codon as reference, according to the usual representation for upstream sites.

Table 3.   Summary of the tools currently available on the web site.

Sequence retrieval	
upstream-region	Retrieves the sequence located upstream from a gene or predicted ORF. The user defines the limits relative to the ORF start. Users may exclude segments that overlap an upstream ORF
downstream-region	*Idem* for downstream sequences
Pattern discovery	
oligo-analysis	Detects over-represented oligonucleotides within a set of upstream sequences. The method allows discovery of unknown regulatory sites from a set of co-regulated upstream regions (van Helden *et al.*, 1998)
dyad-detector	Detects over-represented spaced pairs of conserved words. The method allows discovery of a specific class of regulatory sites that escape detection by oligo-analysis, typically the binding sites for transcription factors having a C6 binuclear Zn cluster DNA binding domain
consensus	Detects shared motifs in unaligned sequences, based on a greedy algorithm. Developed by Jerry Hertz (Hertz *et al.*, 1990; Hertz and Stormo, 1999)
gibbs	Detects shared motifs in unaligned sequences on the basis of a Gibbs sampling strategy. Developed by Andrew Neuwald (Lawrence *et al.*, 1993; Neuwald *et al.*, 1995)
Pattern search	
dna-pattern	String search program dedicated to DNA sequences. Supports IUPAC code for partially specified nucleotides. Several patterns can be searched simultaneously in several sequences, allowing a fast detection of multiple features. Searches can be performed on a single strand or on both
patser	Matrix search program. Developed by Jerry Hertz (Hertz and Stormo, 1999)
genome-search	*dna-pattern* applied to the whole genomic sequence, or to a specific chromosome
all-upstream-search	*dna-pattern* applied to all upstream sequences of each of the 6200 predicted ORFs
all-upstream-patser	*patser* (Hertz and Stormo, 1999) applied to all upstream sequences of each of the 6200 predicted ORFs
Miscellaneous	
feature-map	Generates a physical map with matching positions from a pattern search. Several sequences can be represented in parallel, allowing visual comparison of matching positions
neighbour-orfs	Given a chromosomal position, returns the immediate neighbour ORFs
convert-seq	Interconverts different sequence formats
XYgraph	Generates a 2-D graph from a set of numerical data
random-seq	Generates random sequences. Different probabilistic models can be used (equiprobable nucleotides, specific alphabet utilization, Markov chains)

on which of the three main situations one is facing. (a) In the first situation, the genes are known but the patterns are unknown. This problem is typically encountered with the recent DNA-chip technology: one can now isolate a set of genes on the basis of their coordinated response to some transcription factor. Four pattern discovery programs are available, each using a different strategy to detect over-represented motifs within sets of co-regulated upstream sequences. (b) In the second situation, the search aims to find the locations of matching positions within a restricted set of upstream regions. Depending on the number of known binding sequence alternatives for one given transcription factor, one can use either a string or a matricial description of the binding specificity. (c) In the third situation, a pattern is known, but the regulated genes are unknown. A search for genes potentially regulated by a transcription factor has to be performed in the whole genome. The same search programs are used as in the first situation, except that they operate on the whole set of upstream sequences of the yeast genome.

A challenging issue for the future will be the fourth situation, when both genes and patterns are ignored. The question might seem somehow pointless, but this is, however, the typical approach for analysing coding sequences in every genomic project. As soon as new genomic sequences are released, the first task is to predict the location of potentially coding sequences, and to align them all with each other, as well as with known sequences from other organisms. This already provides a preliminary insight into the potential function of the predicted genes. However, primary structure is insufficient for a real understanding of protein function. Therefore, extracting additional information about the regulatory properties of the predicted genes would be precious. Unfortunately, non-coding sequences do not lend themselves to the same kind of analysis. Regulatory elements are very short, their sequence can vary, and their position with respect to the transcription start is not conserved. The distance range in which these elements can be found is quite large. During evolution, regulatory sequences vary from organism to organism much more rapidly than amino acid sequences. It is thus not possible to perform multiple alignments of complete upstream regions. Moreover, contrary to protein domains, there is no relationship between structure and function of a regulatory element. The insufficiency of the sequence alone to assign gene function was soon realized and, in parallel to the genomic sequencing projects, large-scale projects were undertaken for a systematic functional analysis of the newly sequenced genes. Thanks to the recently developed technologies, allowing measurement of the expression of thousands of genes in a single experiment, one can switch from the fourth situation to the first (known families of co-regulated genes, unknown patterns) by determining experimentally families of genes that respond to a controlled environmental stimulus. The battery of pattern discovery and pattern search programs (Table 3) installed on our site hopefully provide an appropriate resource for interpreting the results of these experiments.

WEB SITES CITED

- *yeast-tools Mexican site*
  http://copan.cifn.unam.mx/Computational_
  Biology/yeast-tools

- *yeast-tools Belgian site*
  http://www.ucmb.ulb.ac.be/bioinformatics/
  rsa-tools/
- *MIPS*
  http://www.mips.biochem.mpg.de/
- *Stanford SGD*
  http://genome-www.stanford.edu/
  Saccharomyces/
- *TRANSFAC*
  http://transfac.gbf-braunschweig.de/
  TRANSFAC/index.html
- *ooTFD*
  http://www.isbi.net/
- *IMD*
  http://bimas.dcrt.nih.gov/molbio/matrixs/
- *GD graphical library*
  http://www.boutell.com/gd/
- *consensus, patser*
  ftp://beagle.colorado.edu/pub/Consensus
- *gibbs sampler*
  ftp://ncbi.nlm.nih.gov/pub/neuwald/
- *YEBIS*
  http://www-scc.jst.go.jp/YEBIS/

CREDITS AND ACKNOWLEDGMENTS

REFERENCES

deRisi JL, Iyer VR, Brown PO. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**: 680–686.

Durbin R, Eddy S, Krogh A, Mitchion G. 1998. *Biological Sequence Analysis—Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press: Cambridge, MA.

Ghosh D. 1999. Object-oriented transcription factors database (ooTFD). *Nucleic Acids Res* **27**: 315–317.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B,

Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. 1996. Life with 6000 genes. *Science* **274** (5287): 546, 563–567.

Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, Podkolodny NL, Kolchanov NA. 1998. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res* **26**: 362–367.

Hertz GZ, Hartzell GW, Stormo GD. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput Appl Biosci* **6**: 81–92.

Hertz GZ, Stormo GD. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15** (in press).

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. 1993. Detecting subtle signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**: 208–214.

Neuwald AF, Liu JS, Lawrence CE. 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.

Schjerling P, Holmberg S. 1996. Comparative amino acid sequence analysis of the C6 zinc cluster family of transcriptional regulators. *Nucleic Acids Res* **24**: 4599–4607.

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

Yada T, Totoki Y, Ishikawa M, Asai K, Nakai K. 1998. Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences. *Bioinformatics* **14**: 317–325.