D446–D451 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj013

The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*

Miguel C. Teixeira[1], Pedro Monteiro[2], Pooja Jain[2], Sandra Tenreiro[1], Alexandra R. Fernandes[1], Nuno P. Mira[1], Marta Alenquer[1], Ana T. Freitas[2,3], Arlindo L. Oliveira[2,3] and Isabel Sá-Correia[1,3,*]

[1]Biological Sciences Research Group, Centro de Engenharia Biológica e Química, Instituto Superior Técnico, Avenida Rovisco Pais, 1049-001 Lisbon, Portugal, [2]INESC-ID, R. Alves Redol, 9, 1000 Lisbon, Portugal and [3]Instituto Superior Técnico, Avenida Rovisco Pais, 1049-001, Lisbon, Portugal

Received July 27, 2005; Revised and Accepted September 14, 2005

ABSTRACT

We present the YEAst Search for Transcriptional Regulators And Consensus Tracking (YEASTRACT; www.yeastract.com) database, a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. This database is a repository of 12 346 regulatory associations between transcription factors and target genes, based on experimental evidence which was spread throughout 861 bibliographic references. It also includes 257 specific DNA-binding sites for more than a hundred characterized transcription factors. Further information about each yeast gene included in the database was obtained from *Saccharomyces* Genome Database (SGD), Regulatory Sequences Analysis Tools and Gene Ontology (GO) Consortium. Computational tools are also provided to facilitate the exploitation of the gathered data when solving a number of biological questions as exemplified in the Tutorial also available on the system. YEASTRACT allows the identification of documented or potential transcription regulators of a given gene and of documented or potential regulons for each transcription factor. It also renders possible the comparison between DNA motifs, such as those found to be over-represented in the promoter regions of co-regulated genes, and the transcription factor-binding sites described in the literature. The system also provides an useful mechanism for grouping a list of genes (for instance a set of genes with similar expression profiles as revealed by microarray analysis) based on their regulatory associations with known transcription factors.

BACKGROUND

The model eukaryote *Saccharomyces cerevisiae*, with the genome sequence available since 1996 (1), plays an essential role in our efforts to understand the complex biological networks that control life processes.

Since the release of the complete genome sequence of the yeast *S.cerevisiae*, a number of computational methods and tools have become available to support research related with this organism. Most significant for the Yeast community, the *Saccharomyces* Genome database (SGD) (2), and other databases specialized in Yeast, such as the Comprehensive Yeast Genome Database (3), or the Yeast Resource Center (4), make available extensive information on molecular biology and genetics of *S.cerevisiae*.

The precise coordinated control of gene expression is accomplished by the interplay of multiple regulatory mechanisms. The transcriptional machinery is recruited to the promoter leading to the transcription of the downstream gene through the binding of transcription regulatory proteins to short nucleotide sequences occurring in gene promoter regions. To support the analysis of the promoter sequences in the yeast genome, a set of software tools is provided by Regulatory Sequences Analysis Tools (RSAT) (5). RSAT makes available pattern matching methods, supporting the search for given nucleotide sequences (e.g. transcription

*To whom correspondence should be addressed. Tel: +351 218417682; Fax: +351 218489199; Email: isacorreia@ist.utl.pt

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

factor-binding sites) within the promoter region of chosen genes, thus leading to the identification of putative target genes for specific transcription factors. However, the transcription factor-binding sites, used for pattern matching in RSAT, have to be provided by the user, since RSAT does not hold a database of transcription factor-binding sites. The two databases that attempt to be such a repository of regulatory elements, The Transcription Factor Database (TRANSFAC) (6) and *Saccharomyces Cerevisiae* Promoter Database (SCPD) (7) do not fill this gap. TRANSFAC is mostly focused on transcription factor-binding sites in higher eukaryotes and provides a limited number of binding sites for *S.cerevisiae*. SCPD, although limited to Yeast, offers incomplete and non up-to-date information.

Since the occurrence of a transcription factor-binding site on the promoter region of a gene is not evidence *per se* that this specific gene is regulated by the transcription factor, further experimental evidence is required to truly establish the group of genes effectively regulated by each transcription factor. The experimental evidence underlying these regulatory associations is spread throughout hundreds of published articles and, until now, no database offered systematic and up-to-date information on documented regulatory associations between transcription factors and yeast genes.

YEAst Search for Transcriptional Regulators And Consensus Tracking (YEASTRACT; www.yeastract.com) proposes to fill these gaps by making publicly available up-to-date information on documented regulatory associations between transcription factors and yeast genes as well as between transcription factors and DNA-binding sites. In addition, it provides an adequate set of bioinformatics tools that facilitate the full exploitation of the data. The originality and usefulness of YEASTRACT rely on the integration of different sets of annotations and experimental information plus the different computational tools.

DATABASE DESCRIPTION

Given that the three principal entities involved in gene regulation are the concept of gene, protein and binding site (consensus), the internal structure of the database is organized around these three concepts.

Figure 1 describes the major entities in the database, and their relationships. The concept of open reading frame (ORF)/gene refers to a DNA region limited by start and stop codons. If the gene that corresponds to a given ORF is characterized, the gene name is stored as one of the attributes. A number of other attributes are also stored for each entity ORF/gene. Protein is the second fundamental concept and is stored as a separate entity, with specific attributes. Finally, the third main entity, consensus, models the binding sites that mediate the process of regulation of genes by transcription factors.

These three concepts are related by regulation relations. These relations document the associations between transcription factors and target genes, and can be of two types: documented and potential.

A documented association between a transcription factor and a target gene is supported by published data showing at least one of the following types of experimental evidence.

(i) Change in the expression of the target gene owing to the deletion (or mutation) of the transcription factor-encoding gene; this evidence may come from detailed gene by gene analysis or genome-wide expression analysis.
(ii) Binding of the transcription factor to the promoter region of the target gene, as supported by band-shift, footprinting or chromatin immunoprecipitation (ChIP) assays.

The user should check the corresponding references, provided in the database, to become fully aware of the nature of the evidence. Each documented association is annotated with

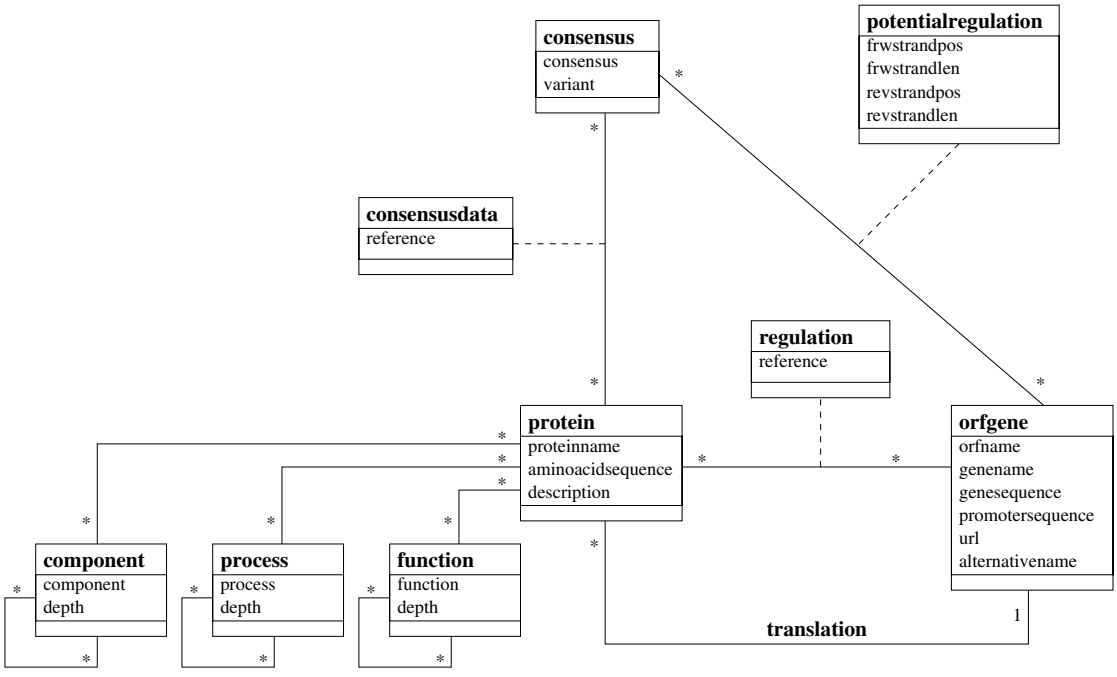

Figure 1. Main entities in the database model and their relationships.

the reference(s) that provided the information, linked to PubMed.

A potential association between a transcription factor and a target gene is based on the occurrence of the transcription factor-binding site in the promoter of the target gene. The binding sites that were considered for each transcription factor are supported by published footprinting or ChIP experiments. As above, references are provided to be checked by the user, if desired.

Figure 1 also shows, on the left, some auxiliary tables/concepts. Tables *component*, *process* and *function* store the corresponding concepts in the gene ontology (GO) (8) hierarchy, and are used by some of the queries.

The database presently contains 12 346 regulatory associations between the yeast genes and 149 described transcription factors, based on 861 bibliographic references. Each regulation has been annotated manually, after examination of the relevant references. The database also contains the description of 257 specific DNA-binding sites for a sub-group of 106 transcription factors. Since a number of transcription factors bind to the same DNA motifs, these 257 binding sites associated to different transcription factors correspond to only 181 distinct nucleotide sequences.

The information in the database is kept up-to-date by a group of curators that use a back-office facility to integrate newly published data on regulatory associations. The contribution of yeast researchers to the accuracy of the information present in the database is requested and possible by using the cooperative work tools available.

YEASTRACT has been implemented using the MySQL database server, the Apache web server and PHP to implement the user interface. The system runs on a machine with two 3.2 MHz Xeon processors and 4 GB of RAM. JavaScript has also been used to enhance usability.

SEARCHING AND MANIPULATING INFORMATION TO SOLVE BIOLOGICAL QUESTIONS

YEASTRACT was made available on the web after extensive internal testing in order to make the system intuitive and easy to use. The interface was developed by two research groups that include yeast biologists and software engineers, with the objective of improving its usefulness and usability. All pages have a context dependent help, automatic form filling using sample data and a complete tutorial on the use of the system.

Typical applications

Although the database can be used in a large number of ways to process data from different sources, and with different objectives, three major groups of processes are directly supported by the existing interfaces.

The first group is related with the identification of documented and potential regulatory associations for a given gene. The system supports a number of queries that can be used to explore documented and potential regulatory mechanisms involving the gene.

The second group of processes is related to the act of analyzing and characterizing sets of genes with identified common expression profiles (for instance, obtained from microarray data) based on their regulatory associations with known transcription factors. This tool provides an interesting mechanism for grouping global expression results and to identify the regulators putatively underlying a transcriptional response.

The third group of processes involves the search for the occurrence of candidate-binding sites, provided by the user (for instance, nucleotide sequences over-represented among co-regulated genes coming from microarray experiments), within the promoter regions of a set of genes and the comparison of those candidate-binding sites with the documented transcription factor-binding sites stored in YEASTRACT.

The tutorial available on the web presents three case studies exemplifying the use of different query options and utilities to support processes in these categories.

Available queries

YEASTRACT makes available the information stored in the database through a set of queries and a number of additional utilities.

The main queries available in the present version are (i) search transcription factors by regulated genes or by keywords, (ii) search genes regulated by specific transcription factors, (iii) group genes by transcription factor, (iv) search by DNA motifs, and (v) search regulatory associations between transcription factors and genes.

The *search transcription factors by regulated genes* query allows the user to identify documented and/or potential regulators of genes present in the given list of genes. Documented and/or potential targets of each transcription factor, present in the given gene list, are displayed in a number of ways, selected by the user. When performing a search for potential regulatory associations, the system may also display a graphic depiction (Figure 2) of the locations of the potential binding sites in the promoter region of the given list of genes. The use of this query is here exemplified for *FLR1* and *TPO1* genes (Figure 2). They encode two multidrug resistance transporters of the Major Facilitator Superfamily, belonging to cluster II of 12-spanner H^+:drug antiporter DHA12 family, and are required for multiple drug resistance (MDR) (9). The graphic display of all potential binding sites (Figure 2a) can be simplified by selecting only those known to be involved in the MDR phenomenon (Figure 2b). Under chemical stress, the role of the potential regulators Yap1p and Pdr3p in the transcriptional activation of *FLR1*, and of Pdr1p and Pdr3p in *TPO1* transcriptional up-regulation, is supported by published data [(10,11) and other references displayed in the database]. Another facility provided by this query is the search for transcription factors that match a specific set of keywords, provided by the user. This query allows the user to search for transcription factors by keywords, found to occur in their description, as extracted from SGD (2).

The *search genes regulated by specific transcription factors* query allows the user to search for genes regulated by the given transcription factors. The user may search for the genes documented as being regulated by specific transcription factors or for genes potentially regulated, based on the existence of the transcription factor-binding site in their promoter region. It is also possible to combine the results from the above two searches.

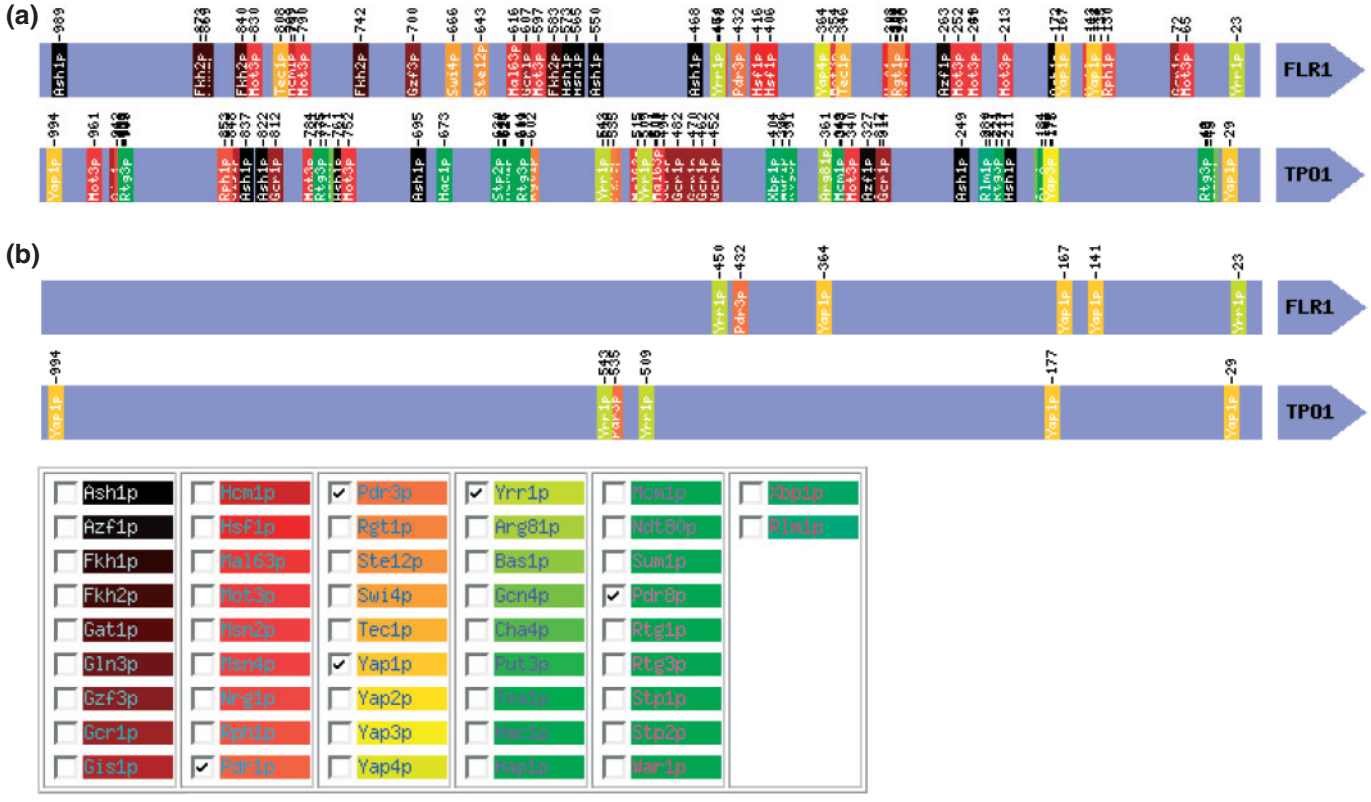

Figure 2. (**a**) Graphic display of the transcription factor-binding sites found to occur in the promoter regions of *FLR1* and *TPO1* genes. (**b**) The table below contains the list of all transcription factors identified as potential regulators of *FLR1* or *TPO1*. By selecting a subset of transcription factors known to be involved in multidrug resistance, it is possible to generate a simplified graphic display.

The *group genes by transcription factor* query allows the user to group a given gene list (for instance, a set of co-activated genes obtained from microarray experiments) according to the transcription factors which are their documented or potential regulators. The transcription factors considered during this query are either a given list of transcription factors or all the transcription factors in the database. For each established regulon, a percentage value representing the proportion of genes regulated by each transcription factor is presented. This value is calculated relative to (i) the total number of genes in the given list; this may indicate the transcription factors involved in the regulation of the referred gene list or (ii) the number of genes, in the whole yeast genome, documented as being regulated by the same transcription factor; this may indicate the transcription factor networks predominantly involved in the regulation of the referred gene list.

The *search by DNA motif* query allows the user to search one or more DNA motifs in the promoter region of one or more genes or within the described transcription factor-binding sites. Figure 3 shows an example of the output of the system to this query, illustrating the navigation menu facilities and overall look and feel of the user interface.

The first possibility, the search for DNA motifs within the promoter region of genes searches for the input DNA motifs in the promoter region of one or more genes or within the promoter region of all the genes in the database. To avoid an excessive number of matches, there is an imposed minimum length for motifs used in this query. The second possibility searches for a match between a single DNA motif with described transcription factors-binding sites. The result of this search is a list of transcription factors-binding sites where there exists a match with the given DNA motif. This search allows the user to check whether a newly identified DNA motif corresponds to a previously described transcription factor-binding site. The search also accepts DNA motifs with ambiguous bases (IUPAC code).

Finally, the *search regulatory associations between transcription factors and genes* query allows the user to identify documented as well as potential regulatory associations between input transcription factors and genes. Documented and potential targets of each transcription factor, which are present in the given gene list, are displayed. This search rejects the transcription factors for which no documented or potential regulatory associations with any of the input genes exists. By default, regulatory associations are searched for input transcription factors against input genes, but other options are available to the user.

Additional utilities

A number of additional utilities are also available in YEASTRACT. These utilities can be used by themselves, or coupled with the previously described main queries. Some of these utilities use data from the GO Consortium (8). The most relevant utilities provided are (i) find transcription factors-binding sites, (ii) group genes by GO terms, (iii) group regulations by GO terms, and (iv) transform a list of ORFs into a list of genes.

Figure 3. Output of a *search for DNA motifs* query. A set of DNA motifs, corresponding to the Yap1p-binding sites documented in the literature, was searched for in the promoter region of a Yap1p-documented target gene, *FLR1* [(9) and other references displayed in the database]. The figure shows the occurrences of the given DNA motifs in the promoter region of this gene. Information on the gene function ('Description') and a link to the SGD page dedicated to *FLR1* are also available.

The *find transcription factors binding sites* utility searches for all described yeast transcription factor-binding sites compatible with the given nucleotide sequence. The output is graphical as well as tabular, displaying all binding sites found in the input sequence on the sense and anti-sense strand.

The *group genes by gene ontology terms* groups a given set of genes according to the GO terms assigned to them by SGD. Grouping can be done by any of the three GOs: biological process, molecular function and cellular component. The specificity of grouping can be enhanced by specifying the level of the GO terms in their respective hierarchy.

The *group regulations by gene ontology terms* utility supplements the search for regulatory associations between input lists of transcription factors and regulated genes by grouping

them by GO terms. The grouping is anchored at the GO terms associated with the transcription factor. Furthermore, the user can select either the biological function or the molecular process ontology.

Finally, the *transform a list of ORFs into a list of genes* utility is a general purpose utility made available to simplify the process of querying databases and systems that accept only one of the names. When an ORF has no attributed gene name, the ORF name appears in the gene name list.

FUTURE WORK

Even though public announces of the availability of the system have not yet been made, YEASTRACT has already been extensively used by a number of research groups working with Yeast, and has demonstrated its usefulness as a tool to support research on transcription regulation processes in this organism.

Nonetheless, we plan to significantly extend the capabilities of the system by connecting it with a number of data processing tools that will increase its usefulness. We plan to interconnect YEASTRACT with tools that will make available, amongst others, the following functions. (i) Search for common motifs in the promoter region of genes, using efficient algorithms for structured motif discovery (12). (ii) Grouping of genes by analysis of gene/motif co-occurrence matrices. (iii) Analysis of microarray data, using non-supervised methods to group genes by expression profile (13).

The interconnection of these tools with the YEASTRACT database will make available to biologists an even more powerful set of tools to analyze regulation mechanisms in Yeast, supported in permanently up-to-date, manually checked, information. Such an analysis will, in the future, support mechanisms for the inference of gene regulatory networks in *S.cerevisiae*, one of the main strategic objectives of this project.

ACKNOWLEDGEMENTS

The information about Yeast genes other than documented regulations, potential regulations and the transcription factor-binding sites contained in YEASTRACT was gathered from SGD, GO Consortium and RSAT. We acknowledge the RSAT team for providing free access to different functionalities of their system to academic users. We are also grateful to colleagues and friends from the Yeast community for their encouragement and suggestions. This work was supported by FEDER, FCT and the POSI and POCTI programs (projects POSI/EIA/57398/2004, POSI/SRI/47778/2002, POCTI/BIO/56838/2004 and POCTI/BIO/38115/2001, and PhD or post-doctoral grants—FSRH/BPD/14484/2003, BPD/5649/01, SFRH/BD/17456/2004 and SFRH/BD/17260/2004—to M.C.T., S.T., N.M. and M.A., respectively). Funding to pay the Open Access publication charges for this article was provided by FEDER, FCT and the POCTI Programme.

Conflict of interest statement. None declared.

REFERENCES

1. Goffeau,A., Barrell,B., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 563–567.
2. Cherry,J.M., Adler,C., Ball,C.A., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M. *et al.* (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.*, **26**, 73–79.
3. Güldener,U., Münsterkötter,M., Kastenmüller,G., Strack,N., van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., García-Martínez,J., Pérez-Ortín,J.E. *et al.* (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.*, **33**, D364–D368.
4. Riffle,M., Malmström,L. and Davis,T.N. (2005) The yeast resource center public data repository. *Nucleic Acids Res.*, **33**, D378–D382.
5. van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.
6. Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhuser,R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.
7. Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.
8. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
9. Sá-Correia,I. and Tenreiro,S. (2002) The multidrug-resistance transporters of the major facilitator superfamily, six years after disclosure of *Saccharomyces cerevisiae* genome sequence. *J. Biotechnol.*, **98**, 215–226.
10. Tenreiro,S., Fernandes,A.R. and Sá-Correia,I. (2001) Transcriptional activation of *FLR1* gene during *Saccharomyces cerevisiae* adaptation to growth with benomyl: role of Yap1p and Pdr3p. *Biochem. Biophys. Res. Commun.*, **280**, 216–222.
11. Teixeira,M.C. and Sá-Correia,I. (2002) *Saccharomyces cerevisiae* resistance to chlorinated phenoxyacetic acid herbicides involves Pdr1p-mediated transcriptional activation of *TPO1* and *PDR5* genes. *Biochem. Biophys. Res. Commun.*, **292**, 530–537.
12. Carvalho,A.M., Freitas,A.T., Oliveira,A.L. and Sagot,M.F. (2005) In *Proceedings of the Third Asia-Pacific Bioinformatics Conference*, Imperial College Press, Singapore, pp. 273–282.
13. Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **1**, 24–45.