CYGD: the Comprehensive Yeast Genome Database.

U. Güldener, M. Münsterkötter, G. Kastenmüller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. García-Martínez, J. E. Pérez-Ortín, et al.

D364–D368 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki053

CYGD: the Comprehensive Yeast Genome Database

U. Güldener[1], M. Münsterkötter[1], G. Kastenmüller[1], N. Strack[11], J. van Helden[2], C. Lemer[2], J. Richelles[2], S. J. Wodak[2], J. García-Martínez[3], J. E. Pérez-Ortín[3], H. Michael[4], A. Kaps[5], E. Talla[6], B. Dujon[7], B. André[8], J. L. Souciet[9], J. De Montigny[9], E. Bon[10], C. Gaillardin[10] and H. W. Mewes[1,11,*]

[1]Institute for Bioinformatics, GSF National Research Center for Environment and Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany, [2]Service de Conformation des Macromolécules Biologiques et Bioinformatique, Université Libre de Bruxelles, CP 263, Blvd du Triomphe, B-1050 Bruxelles, Belgium, [3]Departamento de Bioquímica y Biología Molecular, Universitat de València, C/Dr Moliner 50, E-46100 Burjassot, Spain, [4]Department of Bioinformatics, UKG, University of Göttingen, Goldschmidtstrasse 1, D-37077 Göttingen, Germany, [5]Biomax Informatics AG, Lochhamerstrasse 11, 82152 Martinsried, Germany, [6]CNRS-LCB, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France, [7]Unité de Génétique moléculaire des levures (URA 2171 CNRS and UFR 927 Université Pierre and Marie Curie), Institut Pasteur, Paris, France, [8]Laboratoire de Physiologie cellulaire CP300, Universite Libre de Bruxelles, IBMM, rue des Pr. Jeener et Brachet, 12, 6041 Gosselies, Belgium, [9]Laboratoire de Dynamique, Evolution et Expression des Génomes de Microorganismes, Université Louis Pasteur/CNRS, FRE 2326, 28 rue Goethe, 67000 Strasbourg, France, [10]Laboratoire de Génétique Moléculaire et Cellulaire (URA 1925 CNRS and UMR 216 INRA), INA-PG PO Box 01, F-78850 Thiverval Grignon, France and [11]Technische Universität München, Chair of Genome Oriented Bioinformatics, Center of Life and Food Science, D-85350 Freising-Weihenstephan, Germany

Received August 10, 2004; Revised and Accepted October 1, 2004

ABSTRACT

The Comprehensive Yeast Genome Database (CYGD) compiles a comprehensive data resource for information on the cellular functions of the yeast *Saccharomyces cerevisiae* and related species, chosen as the best understood model organism for eukaryotes. The database serves as a common resource generated by a European consortium, going beyond the provision of sequence information and functional annotations on individual genes and proteins. In addition, it provides information on the physical and functional interactions among proteins as well as other genetic elements. These cellular networks include metabolic and regulatory pathways, signal transduction and transport processes as well as co-regulated gene clusters. As more yeast genomes are published, their annotation becomes greatly facilitated using *S.cerevisiae* as a reference. CYGD provides a way of exploring related genomes with the aid of the *S.cerevisiae* genome as a backbone and SIMAP, the Similarity Matrix of Proteins. The comprehensive resource is available under http://mips.gsf.de/genre/proj/yeast/.

INTRODUCTION

The MIPS yeast genome database was the home of the initial annotation of the first sequenced eukaryotic genome (1). It serves as a primary resource on the yeast genome and its related or derived information and builds the repository for the European functional analysis projects (2). The vast amount of publications on yeast includes a burst of data resulting from high-throughput experiments that are not easily accessible in the literature and demands for thorough annotation. With the sequencing of further yeast genomes the challenge for comparative analysis grows (3–5). To cope with these challenges, the Comprehensive Yeast Genome Database (CYGD) was developed and maintained by a group of European databases and yeast laboratories forming a decentralized network of expertise in order to provide detailed information on protein-coding sequences as well as other genetic elements.

*To whom correspondence should be addressed. Tel: +49 89 3187 3580; Fax: +49 89 3187 3585; Email: w.mewes@gsf.de
Present address:
E. Bon, CBiB–Centre de Bioinformatique de Bordeaux/LBMA–Laboratoire de Biotechnologie et Microbiologie Appliquée, Faculté d'Oenologie, Université Victor Segalen (Bordeaux 2), 146 rue Léo Saignat, F-33076 Bordeaux Cedex, France

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

Table 1. Usage and population of CYGD catalogs

Catalog	Used categories	Annotated entries	Annotated data points
FunCat	495	4660	14 136
EC	613	1330	1401
Protein classes	149	1017	1057
Protein complexes	1051	2728	8495
Localization	52	5164	13 319
Phenotypes	142	1464	3037
Transporter/membrane	234	841	841

All valid CYGD entries are considered.

ANNOTATION IN STRUCTURED CATALOGS

The compilation of sequence related data, in particular of data including different types of relationships is hard to achieve in a system based on the annotation of individual genes. Therefore, a set of catalogs was built to enable systematic classifications of genetic elements. The Functional Catalog (FunCat), a hierarchically structured, organism-independent, flexible and scalable controlled classification system, enabling the functional description of proteins has been developed and first used for the annotation of the yeast genome (1). Owing to its hierarchical architecture, the FunCat has also proved to be useful for many subsequent downstream bioinformatics applications where it served as a reference system for functional prediction. This was also illustrated by the analysis of large-scale experiments from various investigations in transcriptomics and proteomics, where the FunCat was used to project experimental data onto functional units (6,7). Beside the functional classification, catalogs concerning localization, protein classes, phenotypes and complexes were developed (Table 1). The EC nomenclature as well as the TC/MC classification systems also are implemented as catalogs. All classifications can be inspected for their topology and assigned entries as well as from any individual entry. Recently, the functional classification was updated by mapping the latest GO annotation onto FunCat categories (8).

ANNOTATION INFRASTRUCTURE AND ADDITIONAL VALUE

To be able to represent complex data of fungal genomes, we use the Genome Research Environment (GenRE) as our annotation data structure. GenRE allows for the combination of information on different classes of genetic elements and their relationships, such as protein–protein interactions or common regulatory features; it provides annotation features as well as flexible data retrieval interfaces. As nearly all annotation is performed using those catalogs, free text information is reduced to a minimum, although some remarks and phenotypic information are provided in detail.

For the CYGD project, the commercial BioRS^{TM} Integration and Retrieval System (Biomax Informatics AG) has been applied as an integration platform. The BioRS system is a data retrieval system that allows the integration of relational and flat-file oriented databases, both public and proprietary, which are based on different formats, into a common environment. It allows rapid retrieval of data (e.g. sequence, structure and literature) from multiple databanks. By using convenient forms, searches can be as simple or complicated as necessary, providing a sub-query option for search results' refinement. Cross-references between related information in different databanks ensure convenient accessibility to all available information.

Recently added information: an up-to-date review of the *Saccharomyces cerevisiae* introns and the analysis of introns in seven related species can be found in the review section (9). Manually curated Blast alignments and comparison to *S.cerevisiae* genes allowed the identification of 153 introns in seven ascomycetous yeasts partially sequenced during the Genolevures project, as well as of 16 additional introns in *S.cerevisiae* genes previously supposed to be intron-free. Flat files containing the corresponding intron sequences are available for downloading, as well as sequences of other splicing components (e.g. SR protein homologs). These data will be updated using information from additional fully sequenced yeast genomes. An overview on intron structure and splicing mechanism is also available with hypertext links to the corresponding data.

The sequence structure of yeast 3′ flanking regions was also analyzed. This study was based on a previous work (10) in which a consensus model for poly(A) signals was determined. This model was then experimentally confirmed (11,12). It includes three kinds of signals: alternating TA (S1), U-rich (S2) and A-rich (S3). A review includes a list of experimentally determined poly(A) signals for 17 genes and a browser for searching the three kinds of 3′ signals for all the yeast genes. This analysis is currently being improved using information from the genome annotations from other species of the genus *Saccharomyces* sequenced recently (J.van Helden, J.García-Martínez and J.E.Pérez-Ortín, manuscript in preparation). In contrast, the data of the experimentally determined 1540 poly(A) sites for 927 genes has been incorporated into individual CYGD entry pages.

The organization and sequence of the centromere responsible for the proper chromosome segregation were analyzed among the hemiascomycetous yeasts (3). The study is based on the *S.cerevisiae* model organization in which a 126 bp consensus sequence was identified with three blocks separated by two sequences: a 76–86 bp AT-rich DNA stretch and a 26 bp DNA stretch, respectively (13). Searches for orthologous *trans*-acting factors binding to the different DNA centromere blocks were also achieved. This model appears to be conserved only among the *Saccharomyces sensu lato* group and the *Kluyveromyces* group. As far as the evolutionary distances increased after the separation from these two groups, different types of centromeres and of *cis*-acting-related proteins evolved. This analysis is currently being improved using data from other hemiascomycetous yeasts.

TRANSPORTERS AND MEMBRANE PROTEINS

For information on membrane transport proteins, the Yeast Transport Protein DB is integrated in CYGD (14). For 282 transporters recognized on the basis of experimental and sequence criteria, the literature has been scanned to retrieve two kinds of information: (i) the chemical compound(s) recognized by the protein and (ii) the subcellular location of the protein. For both types of information, controlled vocabularies were used to define lists of terms organized as trees and linked

to tables of synonyms. Additionally, transporters were classified according to the TC/MC (see http://tcdb.ucsd.edu/tcdb/) and YTPdb (see http://alize.ulb.ac.be/YTPdb) phylogenetic classification of transporters and other membrane proteins are integrated in CYGD as a catalog (15). For each of the 282 proteins, a specific Boolean formula was designed for a PubMed search for literature.

TRANSCRIPTION FACTORS AND THEIR BINDING SITES

The collection of yeast transcription factors, their respective target genes and binding sites in CYGD is structurally based on the TRANSFAC® database (16). Thus it comprises not only relevant information about transcription factors, their target genes and regulating binding sites, but also has in addition a table with position weight matrices derived from collections of binding sites for given factors. The data used to provide this resource were extracted manually from the literature and evaluated, resulting presently in 370 factor- and 563 gene-entries. The binding site table contains 825 entries, 592 of which are experimentally proven sites, 209 binding sites are artificial, e.g. random oligonucleotides and 24 are consensus sequences. A total of 42 nucleotide distribution matrices have been constructed. The data compiled have been put to use in a variety of studies, e.g. about the prediction of co-regulated genes (17). In parallel to the version integrated into the CYGD framework, the TRANSFAC® yeast data are also freely accessible as the TRANSFAC® *Saccharomyces* Module (TSM). TSM is located at http://www.bioinf.med.uni-goettingen.de/ as part of services provided by the Department of Bioinformatics.

METABOLIC PATHWAYS AND CELLULAR PROCESSES

Information on cellular pathways and processes in *S.cerevisiae* is provided through a link to the Web interface of the aMAZE database (18). The aMAZE database contains information on the chemical reactions, genes and enzymes involved in metabolic pathways, as well as on the transcriptional regulation of the corresponding genes. It also stores information on protein–protein interactions and protein modification involved in signal transduction pathways and implements a generic ontology suitable for storing useful classifications such as the NCBI taxonomy (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy) and Gene Ontology (19). All the information on pathways in aMAZE has been expert curated from the scientific literature. Currently aMAZE contains a comprehensive set of pathways for three organisms (*Escherichia coli*, *S.cerevisiae* and human).

In the context of the CYGD project, access is provided to the data on *S.cerevisiae* only. These data comprise 31 metabolic pathways listed in Table 2. For these pathways, the stored information comprises the aMAZE identifier as well as the custom name for the pathway and *BiologicalReaction*; the *BiochemicalEntities* acting as *Substrates* or *Product* of the *BiologicalReaction*; and the EC number of the *BiologicalReaction* and the PUBMED_ID's of the publications related to the step.

Table 2. Metabolic pathways for *S.cerevisiae*

S. no.	Pathway name	No. of steps
1	Tricarboxylic acid cycle	26
2	Serine biosynthesis	3
3	Synthesis of PRPP	5
4	Glutamine biosynthesis	1
5	Tyrosine biosynthesis	3
6	Aspartate biosynthesis	2
7	Glycine biosynthesis	2
8	Riboflavin biosynthesis	9
9	Isoleucine and valine biosynthesis	11
10	Biotin biosynthesis	3
11	Leucine biosynthesis	9
12	Threonine biosynthesis	6
13	Phenylalanine biosynthesis	2
14	Lysine biosynthesis	9
15	Sulfur incorporation and transsulfuration	5
16	Glutamate biosynthesis	1
17	Methionine biosynthesis II	5
18	Histidine biosynthesis	9
19	Tryptophan biosynthesis	6
20	Alanine biosynthesis	2
21	Heme biosynthesis	8
22	Asparagine biosynthesis	2
23	Sulfate assimilation—yeast	6
24	Aromatic amino acid path	9
25	Methionine and adoMet biosynthesis	3
26	SuccinylCoA ligase	2
27	Proline biosynthesis	4
28	Lanosterol biosynthesis	14
29	Ubiquinone biosynthesis	8
30	Arginine metabolism	9
31	Methionine biosynthesis I	8

A pathway is composed of reaction steps that are connected to one another through *ProcessIntermediates*. A *ProcessIntermediate* is a *BiochemicalEntity*(*molecule*) acting as the *Product* or *Subsrate*. The *BiochemicalEntity* corresponds to a KEGG COMPOUND (whenever defined in KEGG) (20). The *BiologicalReaction* corresponds to a KEGG REACTION (whenever defined in KEGG). The order of the reaction steps in the pathway is determined by the annotator and checked against other sources including the KEGG pathways. The gene name and EC number associated with each reaction was obtained from the Incyte BioKnowledge Library. The *Biochemical Pathways* book by Gerhard Michal was used as a reference for all the annotation work (21).

In addition to the metabolic pathways, information on 18 signal transduction pathways and composing sub-pathways is also provided (listed in Table 3). This information is organized in a similar way as for the metabolic pathways except that all the interactions are modeled as specialized transformations, such as *Expression* (of genes), *Assembly* (of biochemical entities), *Translocation* (of biochemical entities between cellular localization) and *Reaction* (mainly modifying biochemical entities).

PROTEIN–PROTEIN INTERACTIONS

The Catalog of Protein–Protein Interactions, the Protein Complex Catalog and the Protein Localization Catalog allow information related to the proximity of proteins in yeast to be obtained. More than 15 600 protein–protein interaction records (~9200 physical, ~6400 genetic) were compiled manually

Table 3. Signal transduction pathways in *S.cerevisiae*

Pathway name	No. of steps
Calcineurin pathway	8
Cell wall integrity	15
Checkpoint pathway	14
G_1 phase	32
G_2 phase	10
Glucose response	26
HOG pathway	33
M phase	29
Pheromone response	31
Phosphate response	15
Pseudohyphal growth	23
S phase	33
Sporulation—early	31
Sporulation—late	4
Sporulation—mid	7
TOR pathway	42

from the literature (~3680 from single experiments) and published large-scale experiments. Furthermore, 268 manually extracted protein complexes as well as 783 complexes derived from large-scale experiments can be split up into 87 000 putative binary interactions. The vast majority of the records are documented by PubMed reference IDs and by information on the nature of the experimental evidence, which correlates with the confidence of the assignment used in probabilistic computations. The PPI data are accessible from single protein reports or through the MPact interface, which supports retrieval of the data in the standardized PSI-MI format (22).

ANALYSIS OF PARALOGOUS PROTEINS BY SESAM

Paralogous proteins from other species can be retrieved not only using the pre-computed SIMAP (SImilarity MAtrix of Proteins) database (see below) but also using the integrated SESAM tool (Seed Extraction Sequence Analysis Method) (23). The SESAM was developed to achieve better selectivity and sensitivity for the characterization of proteins at large scale without being dependent on secondary data collections, such as InterPro. The selectivity and sensitivity particularly addresses the challenging 'twilight zone' of <30% overall pairwise sequence identity. The manual adjustment of parameters is not required in SESAM and it copes well with different cases of highly conserved as well as distantly related homologs. A subsequent clustering step starts from SESAM seed-based alignments and leads to 'SESAM feature clusters'.

RELATED SPECIES AND FILAMENTOUS FUNGI

As the number of sequenced yeast as well as filamentous fungal genomes is rising steadily, as many possible genomes were analyzed using the PEDANT system and interlinked to the *S.cerevisiae* core database. The analyzed complete genomes include *Schizosaccharomyces pombe* (24), *Candida albicans* (Pasteur Institute), *Saccharomyces bayanus*, *Saccharomyces castellii*, *Saccharomyces kluyveri*, *Saccharomyces kudriavzevii*, *Saccharomyces mikatae*, *Saccharomyces paradoxus* (Whitehead Genome Center; http://www-genome.wi.mit.edu/ and George Washington University, St Louis, MO; http://www.genetics.wustl.edu/), *Candida glabrata*, *Debaryomyces hansenii*, *Kluyveromyces lactis*, *Yarrowia lipolytica* [Génolevures II; http://cbi.labri.fr/Genolevures/about.php (25)], as well as the genomes of filamentous fungi annotated at MIPS: *Neurospora crassa* (MNCDB), *Fusarium graminearum* (FGDB), *Ustilago maydis* (MUMBD) and their relatives: *Magnaporthe grisea*, and *Aspergillus nidulans* (Broad Institute; http://www.broad.mit.edu/annotation/fungi/fgi/). Further genomes will be added to enable a comprehensive comparative fungal data resource.

Additionally, the partial sequenced genomes of the Génolevure I project are also integrated and analyzed in PEDANT databases (3,26). An extensive comparative dataset on these yeast species as well as PEDANT analysis were used to refine the original annotation of the *S.cerevisiae* genome. In particular, comparative genomics between the translation product of overlapping/opposite CDS regions and the Génolevures RST datasets revealed in 449 cases that one CDS (considered as the coding genes) showed similarity to sequences of several other yeast species whereas its partner (considered as the spurious coding genes) remained entirely devoid of homolog. This study leads to 5803 coding sequences including new genes identified in *S.cerevisiae* (27). All these data as well as results from comparative analysis of completely sequenced genomes are used to refine the gene calls on *S.cerevisiae* in the CYGD database (4,5,28,29). Retrieval of the RST information starts at the single *S.cerevisiae* entry using BioRS or from a graphical chromosome display of the fungal orthologs.

SEARCHING THE FUNGAL PROTEIN SEQUENCE SPACE USING SIMAP

As the number of completely sequenced fungal genomes is already remarkable and will substantially increase through ~100 in the not so far future the demand for a centralized tool for similarity based analysis is covered by SIMAP. The SImilarity MAtrix of Protein Sequences provides a pre-calculated all-against-all comparison of the protein sets of all genomes analyzed by PEDANT as well as from other sources like Swiss-Prot. The similarity searches were carried out using the FASTA package (30). Beside the general list of all similar proteins over all taxa, the matrix is used to provide views on similar proteins of related species in specified taxonomic areas, e.g. 'Hemiascomycetes', 'Ascomycetes', etc. The result lists can be clustered to build protein families using MCL on the fly.

DOWNLOAD/LINKS

Complete sets of *S.cerevisiae* sequences and annotation can be downloaded from ftp://ftpmips.gsf.de/yeast/. This includes lists of genetic elements and the contig sequences. The functional classification as well as all other catalogs can be found on ftp://ftpmips.gsf.de/yeast/catalogues/. The protein–protein interaction data can be downloaded from ftp://ftpmips.gsf.de/yeast/PPI/. If you wish to link to the gene reports from your own site, please only use the URL: http://mips.gsf.de/genre/

proj/yeast/searchEntryAction.do?text=YAL036c with a systematic locus code.

SUMMARY

The CYGD database is a frequently used public resource for yeast related information. Yeast as the best understood and annotated eukaryotic organism serves as a reference for the exploration of fungi and higher eukaryotes. An exhaustive, comprehensive classification scheme (FunCat) has been implemented and manually verified. The entire structure of the databases has been revised using GenRE to allow for the annotation of complex relationships such as protein–protein interactions. We use a collaborative approach to incorporate external sources and newly sequenced organisms (25). Additional species will be included soon after publication and an elaborative system for the systematic cross-genome analysis will be introduced.

ACKNOWLEDGEMENTS

This work was supported by the Federal Ministry of Education, Science, Research and Technology (HNB: 01 SF 9985/6), the European Commission (QLRI-CT 1999-01333), the Deutsche Forschungsgemeinschaft (MNCDB) and the Government of the Brussels Region, Belgium, for the aMAZE project.

REFERENCES

1. Mewes,H.W., Albermann,K., Bahr,M., Frishman,D., Gleissner,A., Hani,J., Heumann,K., Kleine,K., Maierl,A., Oliver,S.G. *et al*. (1997) Overview of the yeast genome. *Nature*, **387**, 7–8.
2. Dujon,B. (1998) European Functional Analysis Network (EUROFAN) and the functional analysis of the *Saccharomyces cerevisiae* genome. *Electrophoresis*, **19**, 617–624.
3. Souciet,J.L., Aigle,M., Artiguenave,F., Blandin,G., Bolotin-Fukuhara,M., Bon,E., Brottier,P., Casaregola,S., de Montigny,J., Dujon,B. *et al*. (2000) Genomic exploration of the Hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. *FEBS Lett*., **487**, 3–12.
4. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
5. Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
6. Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B., Bangham,R., Benito,R., Boeke,J.D., Bussey,H. *et al*. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.
7. Giaever,G., Chu,A.M., Ni,L., Connelly,C., Riles,L., Veronneau,S., Dow,S., Lucau-Danila,A., Anderson,K., Andre,B. *et al*. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.
8. Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al*. (2004) *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res*., **32** (Database issue), D311–D314.
9. Bon,E., Casaregola,S., Blandin,G., Llorente,B., Neuveglise,C., Münsterkotter,M., Güldener,U., Mewes,H.W., van Helden,J., Dujon,B. *et al*. (2003) Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res*., **31**, 1121–1135.
10. van Helden,J., del Olmo,M. and Perez-Ortin,J.E. (2000) Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res*., **28**, 1000–1010.
11. Gross,S. and Moore,C.L. (2001) Rna15 interaction with the A-rich yeast polyadenylation signal is an essential step in mRNA 3′-end formation. *Mol. Cell. Biol*., **21**, 8045–8055.
12. Dichtl,B. and Keller,W. (2001) Recognition of polyadenylation sites in yeast pre-mRNAs by cleavage and polyadenylation factor. *EMBO J*., **20**, 3197–3209.
13. Clarke,L. (1998) Centromeres: proteins, protein complexes, and repeated domains at centromeres of simple eukaryotes. *Curr. Opin. Genet. Dev*., **8**, 212–218.
14. Van Belle,D. and Andre,B. (2001) A genomic view of yeast membrane transporters. *Curr. Opin. Cell Biol*., **13**, 389–398.
15. De Hertogh,B., Carvajal,E., Talla,E., Dujon,B., Baret,P. and Goffeau,A. (2002) Phylogenetic classification of transporters and other membrane proteins from *Saccharomyces cerevisiae*. *Funct. Integr. Genomics*, **2**, 154–170.
16. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al*. (2003) TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.
17. Simonis,N., Wodak,S.J., Cohen,G.N. and van Helden,J. (2004) Combining pattern discovery and discriminant analysis to predict gene co-regulation. *Bioinformatics*, **20**, 2370–2379.
18. Lemer,C., Antezana,E., Couche,F., Fays,F., Santolaria,X., Janky,R., Deville,Y., Richelle,J. and Wodak,S.J. (2004) The aMAZE LightBench: a web interface to a relational database of cellular processes. *Nucleic Acids Res*., **32** (Database issue), D443–D448.
19. 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. *Nature Genet*., **25**, 25–29.
20. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32** (Database issue), D277–D280.
21. Michal,G. (1998) *Biochemical Pathways: An Altlas of Biochemistry and Molecular Biology*. Wiley and Sons, Inc.
22. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,R., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al*. (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol*., **22**, 177–183.
23. Strack,N. and Mewes,H.W. (1999) SESAM: Seed Extraction Sequence Analysis Method. *Proc. GCB*, **1**, 59–65.
24. Wood,V., Gwilliam,R., Rajandream,M.A., Lyne,M., Lyne,R., Stewart,A., Sgouros,J., Peat,N., Hayles,J., Baker,S. *et al*. (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.
25. Dujon,B., Sherman,D., Fischer,G., Durrens,P., Casaregola,S., Lafontaine,I., de Montigny,J., Marck,C., Neuveglise,C., Talla,E. *et al*. (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.
26. Frishman,D., Mokrejs,M., Kosykh,D., Kastenmuller,G., Kolesov,G., Zubrzycki,I., Gruber,C., Geier,B., Kaps,A., Albermann,K. *et al*. (2003) The PEDANT genome database. *Nucleic Acids Res*., **31**, 207–211.
27. Talla,E., Tekaia,F., Brino,L. and Dujon,B. (2003) A novel design of whole-genome microarray probes for *Saccharomyces cerevisiae* which minimizes cross-hybridization. *BMC Genomics*, **4**, 38.
28. Kellis,M., Birren,B.W. and Lander,E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, **428**, 617–624.
29. Dietrich,F.S., Voegeli,S., Brachat,S., Lerch,A., Gates,K., Steiner,S., Mohr,C., Pohlmann,R., Luedi,P., Choi,S.D. *et al*. (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science*, **304**, 304–307.
30. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol*., **132**, 185–219.