74–78 Nucleic Acids Research, 1999, Vol. 27, No. 1

Using the *Saccharomyces* Genome Database (SGD) for analysis of protein similarities and structure

Stephen A. Chervitz, Erich T. Hester, Catherine A. Ball, Kara Dolinski, Selina S. Dwight, Midori A. Harris, Gail Juvik, Alice Malekian, Shannon Roberts, TaiYun Roe, Charles Scafe, Mark Schroeder, Gavin Sherlock, Shuai Weng, Yan Zhu, J. Michael Cherry* and David Botstein

Department of Genetics, Stanford University, Stanford, CA 94305-5120, USA

Received October 1, 1998; Revised and Accepted October 8, 1998

ABSTRACT

The *Saccharomyces* Genome Database (SGD) collects and organizes information about the molecular biology and genetics of the yeast *Saccharomyces cerevisiae*. The latest protein structure and comparison tools available at SGD are presented here. With the completion of the yeast sequence and the *Caenorhabditis elegans* sequence soon to follow, comparison of proteins from complete eukaryotic proteomes will be an extremely powerful way to learn more about a particular protein's structure, its function, and its relationships with other proteins. SGD can be accessed through the World Wide Web at http://genome-www.stanford.edu/Saccharomyces/

INTRODUCTION

The *Saccharomyces* Genome Database (SGD) exists to provide the scientific community with access to the *Saccharomyces cerevisiae* sequence and the wealth of associated information. This database includes a variety of biological information, including the complete, annotated DNA and protein sequence along with several tools for sequence analysis. Many of these features have been recently described (1,2). Here we focus on features of SGD that provide users with tools for comparing yeast protein sequences and examining protein structure. Sequence comparisons play a critical role in the initial process of determining the function of specific proteins and also in interpreting new protein sequence data from large-scale genome sequencing projects. There are several sequence comparison tools at SGD. Here, we discuss the Genome-wide Protein Similarity View program, which is a powerful tool for examining protein similarities. Like the expanding base of sequence information, there is also a growing amount of structural information. Sacch3D is a feature of SGD that organizes and presents structural information about yeast proteins and their putative homologs. Familiarity with tools such as those described here will enable molecular biologists and geneticists to gain insight into the function and possible evolution of their protein of interest.

EXAMINING PROTEIN SIMILARITIES AT SGD

The Genome-wide Protein Similarity View (GPSV), at URL http://genome-www.stanford.edu/cgi-bin/SGD/SWA/swaEntryForm.pl , displays, either graphically or in a table, all the ORFs in the *S.cerevisiae* genome that are similar to a given query ORF based on a Smith–Waterman protein sequence comparison (3). Smith–Waterman comparisons were conducted on a TimeLogic DeCypher II machine using the affine Smith–Waterman application (4). This system uses the pscorer program to calculate a *P*-value (5). More details of the Smith–Waterman alignment are available at the GPSV help page (URL in Table 1).

The GPSV graphic view, a Java applet, represents all 16 yeast nuclear chromosomes as horizontal black bars, with centromeres and positional coordinates indicated (Fig. 1A). Superimposed on the black chromosome bars are small vertical colored bars (similarity bars) that represent ORFs predicted by the Smith–Waterman analysis to have significant protein sequence similarities. A small black rectangle surrounds the bar for the query ORF itself. Its ORF name and associated standard gene name are displayed in the upper right hand corner. The color of the bars indicates the relative similarity shared with the query ORF. The warm colors (red) indicate high similarity while the cool colors (blue) indicate lower similarity. The user can switch between different query ORFs, add ORFs to the query list, and change several parameters of the similarity display.

Immediately below the graphic display are seven fields that contain additional information about the query and target ORFs (Fig. 1B). The first field displays a constantly updated readout of the current location of the mouse cursor in terms of base pairs along the chromosomes and the names of genes or ORFs selected by the mouse. The remaining fields contain information when the cursor is positioned over a similarity bar. The classes of information are: (i) *P*-Value (the *P*-value for the similarity between the query ORF and the target ORF); (ii) % Aligned (the percent of the query sequence that is aligned with the target sequence); and (iii) Gaps (the number of gaps inserted in the query sequence to achieve the alignment).

Each similarity bar can be clicked to reach more information about the target ORF. Options include links to the SGD Locus and

*To whom correspondence should be addressed. Tel: +1 650 723 7541; Fax: +1 650 723 7016; Email: cherry@genome.stanford.edu

Table 1. URLs mentioned in this review

Site	URL
Saccharomyces Genome Database (SGD)	http://genome-www.stanford.edu/Saccharomyces/
Genome-wide Protein Similarity View (GPSV)	http://genome-www.stanford.edu/cgi-bin/SGD/SWA/swaEntryForm.pl
GPSV help page	http://genome-www.stanford.edu/Saccharomyces/help/sw_details.html
Sacch3D	http://genome-www.stanford.edu/Sacch3D/
Sacch3D help page	http://genome-www.stanford.edu/Sacch3D/help/index.html
Chime (MDL Information Systems, Inc.)	http://www.mdli.com/download/chimedown.html
Cn3D	http://www.ncbi.nlm.nih.gov/Structure/cn3d.html
MolMovDB	http://bioinfo.mbb.yale.edu/MolMovDB/
Entrez	http://www3.ncbi.nlm.nih.gov/Entrez/
Clusters of Orthologous Groups (COGs) from NCBI	http://www.ncbi.nlm.nih.gov/COG/
Protein Data Bank (PDB)	http://www.pdb.bnl.gov/
ModBase Comparative Protein Structure Models for Yeast Proteins	http://guitar.rockefeller.edu/gmodels/
RasMol	http://www.umass.edu/microbio/rasmol/
Java PDB Viewer	http://genome-www.stanford.edu/Sacch3D/help/3dviewers.html
Structural Classification of Proteins (SCOP)	http://www.pdb.bnl.gov/scop/
SWISS-Model	http://www.expasy.ch/swissmod/SWISS-MODEL.html
EMOTIF	http://motif.Stanford.EDU/emotif/
Pfam	http://pfam.wustl.edu/
SWISS-Prot	http://expasy.hcuge.ch/sprot/sprot-top.html
TimeLogic	http://www.timelogic.com

Gene/Sequence Resources pages for the target ORF, an alignment of the query and target amino acid sequences, the DNA Similarity View, which displays the alignment of the target and query ORF DNA sequences, and the Protein Similarity View, where the selected target ORF is used as the query ORF.

The protein similarity data can also be displayed as a table, which can be accessed from the graphic display page or from the ORF input form. The table lists target ORFs in order of decreasing similarity to the query ORF, as determined by *P*-value; the target ORFs can also be sorted by percent identity. For each target ORF, the table lists the same information and links as the graphic display.

PROTEIN STRUCTURE AT SGD

The Sacch3D feature at SGD (at URL http://genome-www.stanford.edu/Sacch3D/) provides structural information for *S.cerevisiae* proteins by integrating data from SGD and structural databases and presenting it via up-to-date, concise summaries and links to structural resources. Sacch3D supplies researchers both within and outside the yeast community with insight into the structure and putative function of yeast proteins. Structural information for Sacch3D is obtained primarily by BLASTP analysis (6,7) of the Brookhaven Protein Database (PDB) (8,9) to identify all PDB structures with significant sequence (and therefore likely structural) similarity to yeast proteins. Results are updated monthly to keep pace with the growth of the PDB. To reduce the redundancy in the PDB and thus simplify the BLAST analysis, all PDB protein sequences are first clustered into groups of closely related sequences (see the on-line help at the Sacch3D website for details; Table 1) before the BLAST is run. As of September 1998, 18% of yeast proteins have either a known structure or putative homolog in a clustered version of the PDB (Table 2). The Sacch3D search utility provides a structural information page for all ORFs in the yeast genome (example shown in Fig. 2). This page contains information provided by both internal and external resources. A summary table is presented showing PDB structures for the yeast protein (if a structure can be identified) and proteins with which it shares significant sequence similarity. For each structure, there are links to a variety of freely available 3D viewers (Fig. 3) and external structural databases. 3D viewers include RasMol (10), Webmol Java viewer (11), Chime (MDL Information Systems, www.mdli.com) and Cn3D (12). External structural databases include PDB (8,9), SCOP (13,14), CATH (15), PDBsum (16), ModBase (17), Macromolecular Movements Database (18) and MMDB (19). The PDB similarities are listed from best to worst (based on BLASTP *P*-value) and are clustered to facilitate browsing. That is, one representative structure is listed in cases where there are multiple variants of the same structure (mutants or complex forms). Access to the neighboring structures is also provided.

For yeast proteins without a known structure but with significant sequence similarity to proteins with structures contained within PDB, links are available on the structural information page for homology-based models of the yeast protein structure. These models are accessed by links to the external

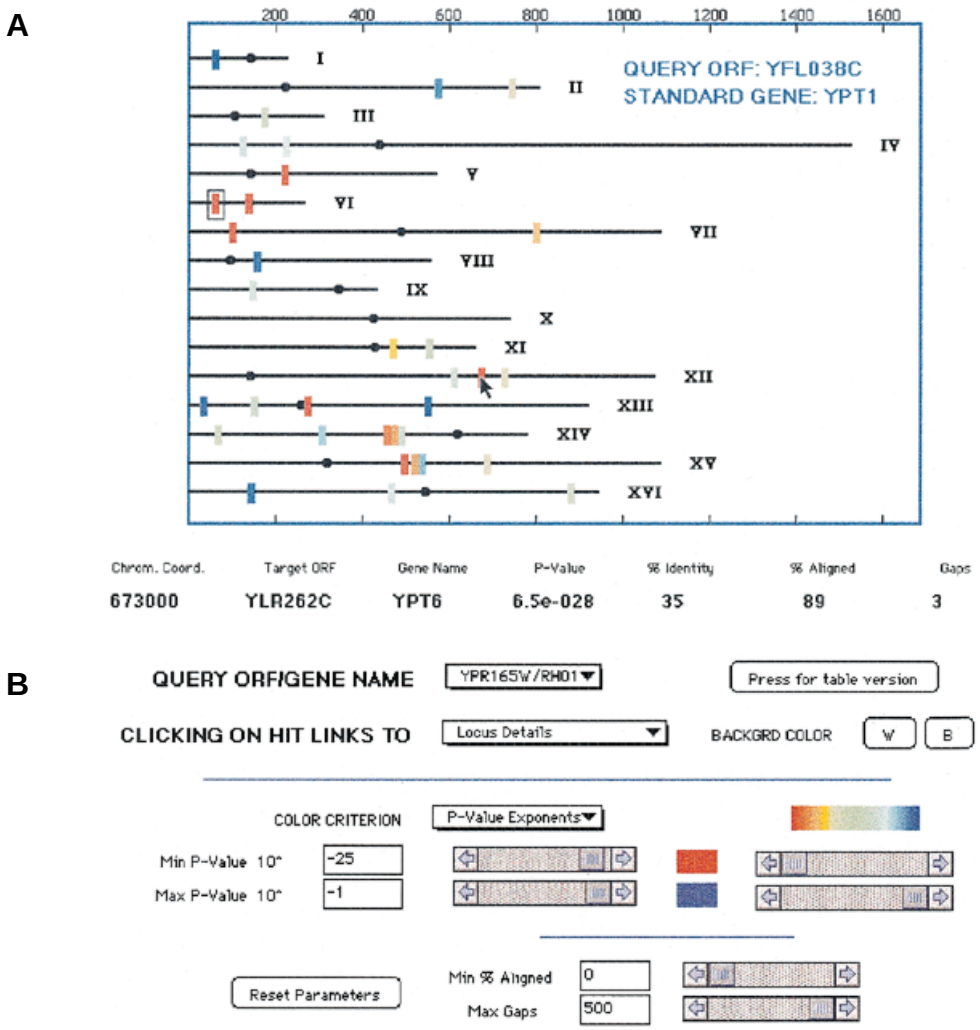

Figure 1. The Genome-Wide Protein Similarity View page. Features are discussed in detail in the text. (**A**) The Genome-Wide Protein Similarity View graphic view; (**B**) Genome-Wide Protein Similarity View parameters.

resources ModBase (17) and Swiss-Model (20). Even for yeast proteins that lack significant similarities in the PDB, a variety of useful links are presented. These include links to secondary structure predictions, several pre-computed BLAST reports, and the Emotif (21) and Pfam (22) sequence search programs. Links to Swiss-Prot, Entrez and the NCBI COGs site for the yeast protein are also included.

Table 2. Sacch3D summary statistics for protein-encoding ORFs

	N	Percent
Total yeast proteins	6215	100
Identified genetically	3086	50
With known 3D structure(s)[a]	52	0.8
With PDB homolog(s)	1110	18
With PDB homolog and identified genetically	848	13

Data in this table are current as of the 2 September 1998 release of the PDB.
[a]Yeast proteins with known structures are current as of the 27 May 1998 release of the PDB.

Other Sacch3D features include: (i) flexible search options using yeast gene or ORF name, PDB identifier, Swiss-Prot identifier/accession, or text; (ii) a special page devoted to *S.cerevisiae* structures in PDB showing the number of different structures for each yeast protein with links to SGD and Sacch3D; (iii) a structural URLs page. Sacch3D maintains this list of URLs for web sites relevant to the analysis of protein structure and/or function, including links to structural biology resources, 3D viewers, genome-analysis web sites, journals and research groups; (iv) a domains page providing access to yeast proteins based on their SCOP-classified domains. Users can search for domains using a yeast gene/ORF name, a SCOP class number, or a SCOP fold number. Links are also provided to the WebMol Java viewer (11) to illustrate the location of the domain within the context of the 3D structure; (v) an analysis page that performs an electronic version of a Southern blot using the yeast genomic sequence; (vi) a What's New page that lists new features in Sacch3D as well as new yeast protein structures and new protein structures with homology to yeast proteins.

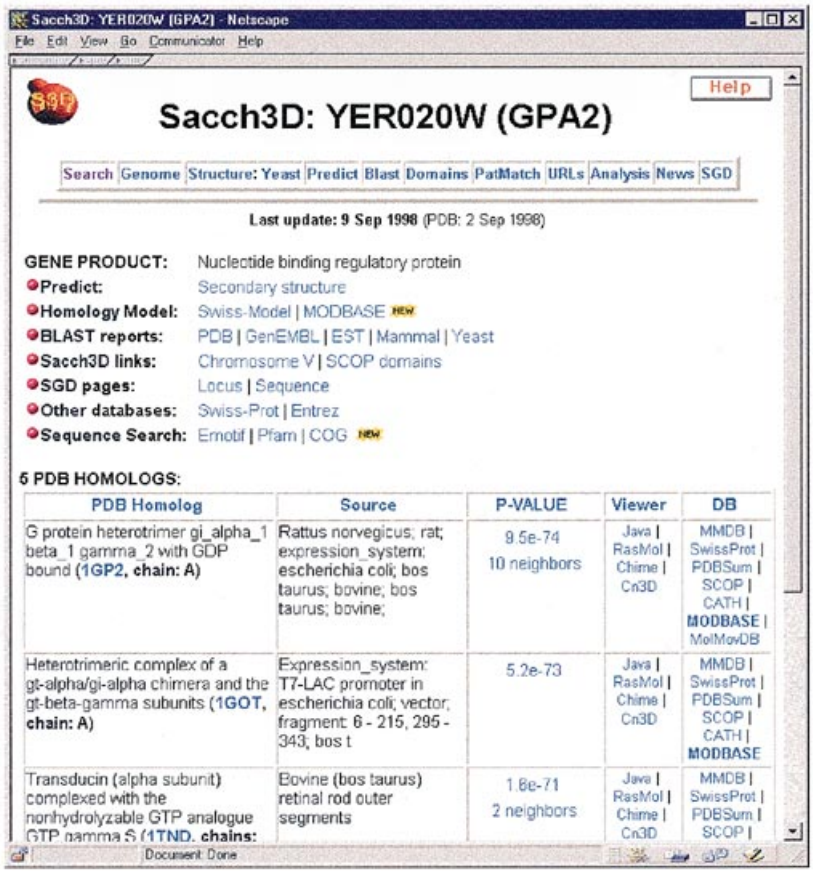

Figure 2. Structural information page at Sacch3D. Structural information page for yeast GTP-binding protein GPA2/YER020W.

FUTURE DIRECTIONS

Protein sequence analysis and structure prediction will continue to be updated. However, in the future the major new analysis features of SGD will be associated with the many types of functional genomic results that are beginning to be released.

CITING SGD

When referring to Sacch3D and Genome-wide Similarity View at SGD, please cite this publication.

REFERENCES

1 Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M., Weng,S. and Botstein,D. (1998) *Nucleic Acids Res.*, **26**, 73–79.
2 Dolinski,K., Ball,C., Chervitz,S.A., Dwight,S.S., Harris,M., Roberts,S., Roe,T., Cherry,J.M. and Botstein,D. (1998) *Yeast*, **14**, in press.
3 Smith,T.F. and Waterman,M.S. (1981) *J. Mol. Biol.*, **147**, 195–197.
4 TimeLogic: http://www.timelogic.com
5 Gish,W. and Altschul,S.F. (1996) *Methods Enzymol.*, **266**, 460–490.
6 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
7 Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
8 Abola,E.E., Sussman,J.L., Prilusky,J. and Manning,N.O. (1997) *Methods Enzymol.*, **277**, 556–571.
9 Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
10 Sayle,R.A. and Milner-White,E.J. (1995) *Trends Biochem. Sci.*, **9**, 374.
11 Walther,D. (1997) *Trends Biochem. Sci.*, **22**, 274–275.
12 Hogue,C.W. (1997) *Trends Biochem. Sci.*, **22**, 314–316.
13 Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
14 Hubbard,T.J.P., Murzin,A.G., Brenner,S.E. and Chothia,C. (1997) *Nucleic Acids Res.*, **25**, 236–239.
15 Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.
16 Laskowski,R.A., Hutchinson,E.G., Michie,A.D., Wallace,A.C., Jones,M.L. and Thornton,J.M. (1997) *Trends Biochem. Sci.*, **22**, 488–490.
17 Sánchez,R. and Sali,A. (1997) *Proteins*, Suppl. 1, 50–58.
18 Gerstein,M. and Krebs,W. (1998) *Nucleic Acids Res.*, **26**, 4280–4290.
19 Ohkawa,H., Ostell,J. and Bryant,S. (1995) *ISMB*, **3**, 259–267.
20 Peitsch,M.C. (1996) *Biochem. Soc. Trans.*, **24**, 274–279.
21 Nevill-Manning,C.G., Wu,T.D. and Brutlag,D.L. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 5865–5871.
22 Sonnhammer,E.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) *Nucleic Acids Res.*, **26**, 320–322.

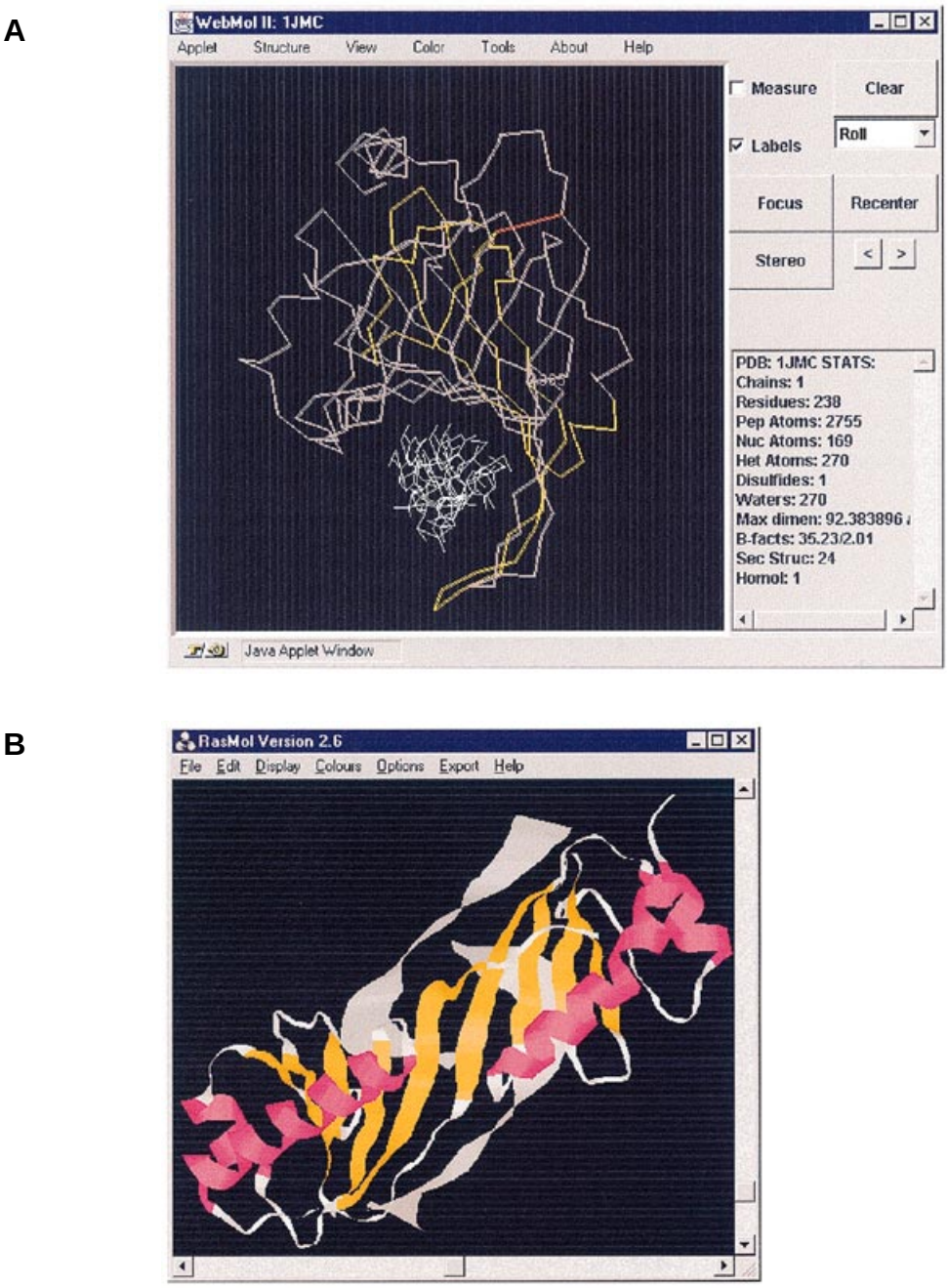

Figure 3. (**A**) A screen shot of the Java Viewer (11) showing the structure of a human single-stranded DNA-binding domain of the RPA70 subunit bound to ssDNA (PDB structure 1JMC; yeast homolog RFA1/YAR009C). Yellow, region of similarity to yeast protein; white, ssDNA; gray, region without similarity to yeast protein; red, disulfide bond. (**B**) A screen shot of the RasMol viewer (10) showing the human TATA-binding protein complex with TATA element DNA (PDB structure 1TGH; yeast homolog SPT15/YER148W). Gold, β sheets; red, α helices; thin white ribbon, coil; wide gray ribbon, DNA.