D378–D382 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki073

The Yeast Resource Center Public Data Repository

Michael Riffle, Lars Malmström and Trisha N. Davis*

Department of Biochemistry, University of Washington, Seattle, WA 98195, USA

Received August 6, 2004; Revised and Accepted October 6, 2004

ABSTRACT

The Yeast Resource Center Public Data Repository (YRC PDR) serves as a single point of access for the experimental data produced from many collaborations typically studying *Saccharomyces cerevisiae* (baker's yeast). The experimental data include large amounts of mass spectrometry results from protein co-purification experiments, yeast two-hybrid interaction experiments, fluorescence microscopy images and protein structure predictions. All of the data are accessible via searching by gene or protein name, and are available on the Web at http://www.yeastrc.org/pdr/.

INTRODUCTION

The Yeast Resource Center (YRC) is an NCRR Biomedical Technology Resource Center that provides expertise and otherwise costly tools of research to scientists and students worldwide. This is accomplished via collaborations and technology development projects—with 231 such collaborations having been submitted since the beginning of 2002. The collaborations focus mainly on the study of *Saccharomyces cerevisiae* via four primary areas of expertise provided by the YRC: mass spectrometry, yeast two-hybrid arrays, deconvolution fluorescence microscopy and protein structure prediction. The YRC investigators, who have been responsible for fulfilling collaboration requests are Dr John Yates, Dr Ruedi Aebersold (mass spectrometry), Dr Stanley Fields (yeast two-hybrid), Dr Trisha Davis, Dr Eric Muller (fluorescence microscopy) and Dr David Baker (protein structure prediction).

Collaborative projects can involve multiple experiments carried out in one or more of these four areas. All four areas can produce large amounts of data—not all of which are necessarily used in the course of publication by the collaborator. In addition, not all collaborations necessarily lead to a publication; but data produced through the collaboration may be valuable and useful. The YRC makes available both the published and unpublished data through the YRC Public Data Repository (PDR) to the community at large.

Perhaps the most significant aspect of the YRC PDR is that it releases all of the data at a single point of access, bringing together the experimental data from many research projects into one consolidated searchable database accessible through the Web. Instead of going from website to website supporting individual papers, one can easily search the experimental data for multiple papers at once and view the results in a single interface. As more datasets from research collaborations with the YRC become public, the database will continue to grow and become an increasingly significant asset to the research community.

THE CONTENTS OF THE YRC PDR DATABASE

At the time of this writing, the YRC PDR includes data from six collaborative projects—including four publications (1–4). This includes mass spectrometry data collected through protein co-purification experiments, yeast two-hybrid protein interaction data, fluorescent microscopy images and protein structure prediction data.

Protein structures are predicted for protein domains, as parsed from the Ginzu algorithm (5). *Ab initio* structure predictions are available as Protein Data Bank (PDB) (6) formatted text, as generated using the Rosetta *de novo* structure prediction method (7–10).

In addition, the database includes images taken from silver-stained polyacrylamide gels of samples produced from protein co-purification experiments; and links to descriptions of the protocol used for the purification.

The breakdown of the amount of data presently included in the database is summarized in Table 1.

THE YRC PDR WEB INTERFACE

We have developed a simple-to-use web interface to the YRC PDR database. The primary means of interacting with the data is to perform searches based on systematic open reading frame (ORF) or gene names. Gene names are mapped onto systematic ORF names through the publicly available *Saccharomyces* Genome Database (11,12). Searching will bring the user to a page displaying an overall summary of all the experimental data we have for a given ORF. An example search result is

*To whom correspondence should be addressed. Tel: +1 206 543 5345; Fax: +1 206 685 1792; Email: tdavis@u.washington.edu

Table 1. A summary of the quantity of the different types of data currently available in the database

Mass spectrometry data	
Total runs	119
Total unique proteins identified	3138
Total peptides identified	41 397
Total gel images	45
Yeast two-hybrid data	
Total baits with significant hits	409
Total unique ORFs with significant hits	1373
Total unique significant interactions	2031
Fluorescence microscopy data	
Total unique proteins localized	122
Total full-field images	767
Total selected region images	877
Protein structure prediction	
Total ORFs with structure data	145
Total domains with structure data	255
Total *ab initio* structures	850 (for 86 domains from 63 proteins)

given in Figure 1. This 'ORF Overview Page' is separated into five sections, from which the user can view the Gene Ontology (13) description for the ORF and jump to experimental data view pages for each of the four types of data. Each of these data view pages is tailored to a specific kind of data and each has its own features that are described below. All data are clearly labeled according to publication(s) for which they were produced. In addition, data not used in any publication are clearly labeled as unpublished.

Mass spectrometry data

From the ORF Overview Page's mass spectrometry section, the user is presented with several links for viewing the mass spectrometry data.

View Protocol link: This provides the user with a text description of the protocol used for a particular protein purification, if the protocol is available.

Bait ORF link: This lists the actual purified protein and a link to that protein's ORF Overview Page. Whenever the name of an ORF is given in the website, it is linked to that ORF's overview page.

View Gel link: If the protein sample was subjected to electrophoresis on an SDS polyacrylamide gel, this link will be present and will provide an image of the silver-stained gel.

View Run link: This is a link to the results from the analysis of the protein by mass spectrometry. The data include a filtered and formatted listing produced from the DTASelect algorithm (14). The data are presented as a list of systematic ORF names for proteins that are co-purified with the bait protein, along with its sequence coverage, number of peptides, spectrum count and molecular weight. A guideline for interpretation of these columns is provided on this page. For each ORF listed, there is a link for viewing the peptides that were used to make that identification. The list of ORFs and the peptide lists may be downloaded as tab-delimited text files from the site. An example of the page displaying mass spectrometry data is provided in Figure 2.

Fluorescence microscopy (localization) data

The ORF Overview Page's localization section allows the user to view fluorescence microscopy images of each protein tagged with a fluorescent protein such as green fluorescent protein. All localization experiments involving this ORF are clearly listed here. The 'View Images' link provides the means to view all images from the localization experiment, the experimental parameters used to create these images and the localization determination expressed as a cellular component term from Gene Ontology.

Yeast two-hybrid data

The ORF Overview Page's yeast two-hybrid section provides the means to quickly jump to and view results from all yeast two-hybrid screens in which the ORF of interest was bait or prey. Screen results display the prey ORF as well as the number of hits. A number of hits greater than one are considered significant, but single hits are shown for completeness. The results from these screens are also available for download as a tab-delimited text file.

Protein structure prediction data

If structure prediction data are available for an ORF, the protein structure prediction section provides a list of computationally derived domains for the ORF. This section will give the start and stop residue for each domain, the source of the structure in the database and a link to structural information for that domain. The information in these structure links is tailored to how the structure was derived.

Domains, for which the structures were obtained through *ab initio* prediction, will contain links to the top ten predicted structures. These structures are viewable in the site itself via the WebMol Java applet (15). The structures are also downloadable as PDB text files.

AVAILABILITY

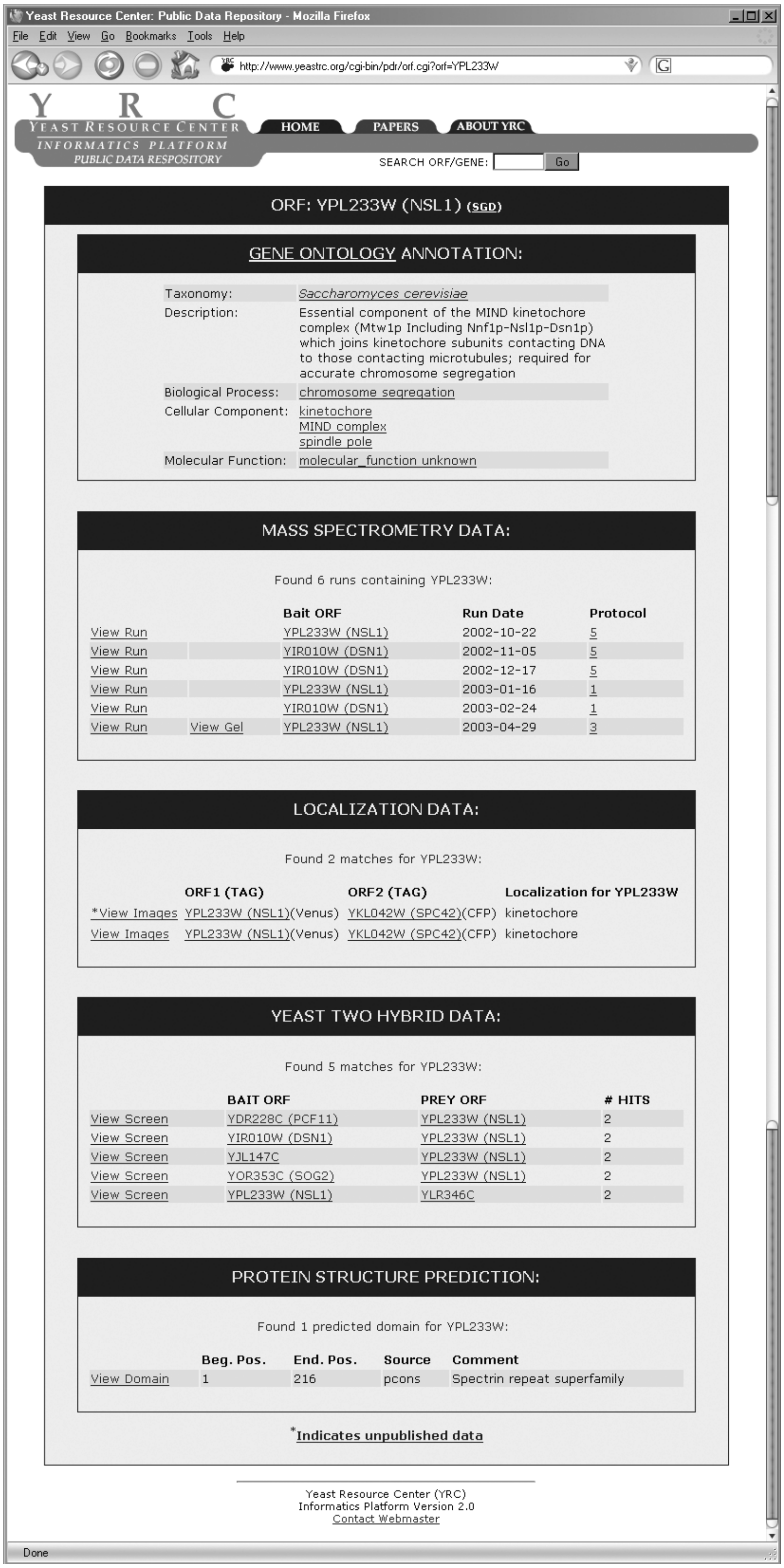
The contents of the YRC PDR are available on the Web at http://www.yeastrc.org/pdr/. From this URL the contents of the database can be viewed as HTML pages, as well as tab-delimited text files when applicable. The entire published datasets of yeast two-hybrid and mass spectrometry run results are available as tab-delimited text files, linked from the front page. The unpublished datasets are available upon request. These tab-delimited text files can easily be imported into Microsoft Excel, as well as other spreadsheet and data software.

FUTURE DIRECTIONS

The YRC will likely expand beyond providing collaborations and technology development in only these four current areas of expertise. As a result, the type of experimental data available in the YRC PDR database will also expand.

Currently, the YRC PDR only includes experimental data covering *S.cerevisiae*. The YRC has broadened its scope and has begun participating in collaborations involving other organisms. As a result, the YRC PDR will contain data from protein experiments involving multiple organisms.

Given these two main points and the fact that the YRC PDR will continue to expand by the addition of data from more and more collaborations, the functionality of the interface will be


Yeast Resource Center: Public Data Repository - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.yeastrc.org/cgi-bin/pdr/orf.cgi?orf=YPL233W
Y R C
YEAST RESOURCE CENTER
INFORMATICS PLATFORM
PUBLIC DATA RESPOSITORY
HOME
PAPERS
ABOUT YRC
SEARCH ORF/GENE:
Go
Done

ORF: YPL233W (NSL1) (SGD)

GENE ONTOLOGY ANNOTATION:

Taxonomy:	*Saccharomyces cerevisiae*
Description:	Essential component of the MIND kinetochore complex (Mtw1p Including Nnf1p-Nsl1p-Dsn1p) which joins kinetochore subunits contacting DNA to those contacting microtubules; required for accurate chromosome segregation
Biological Process:	chromosome segregation
Cellular Component:	kinetochore MIND complex spindle pole
Molecular Function:	molecular_function unknown

MASS SPECTROMETRY DATA:

Found 6 runs containing YPL233W:

		Bait ORF	Run Date	Protocol
View Run		YPL233W (NSL1)	2002-10-22	5
View Run		YIR010W (DSN1)	2002-11-05	5
View Run		YIR010W (DSN1)	2002-12-17	5
View Run		YPL233W (NSL1)	2003-01-16	1
View Run		YIR010W (DSN1)	2003-02-24	1
View Run	View Gel	YPL233W (NSL1)	2003-04-29	3

LOCALIZATION DATA:

Found 2 matches for YPL233W:

	ORF1 (TAG)	ORF2 (TAG)	Localization for YPL233W
*View Images	YPL233W (NSL1)(Venus)	YKL042W (SPC42)(CFP)	kinetochore
View Images	YPL233W (NSL1)(Venus)	YKL042W (SPC42)(CFP)	kinetochore

YEAST TWO HYBRID DATA:

Found 5 matches for YPL233W:

	BAIT ORF	PREY ORF	# HITS
View Screen	YDR228C (PCF11)	YPL233W (NSL1)	2
View Screen	YIR010W (DSN1)	YPL233W (NSL1)	2
View Screen	YJL147C	YPL233W (NSL1)	2
View Screen	YOR353C (SOG2)	YPL233W (NSL1)	2
View Screen	YPL233W (NSL1)	YLR346C	2

PROTEIN STRUCTURE PREDICTION:

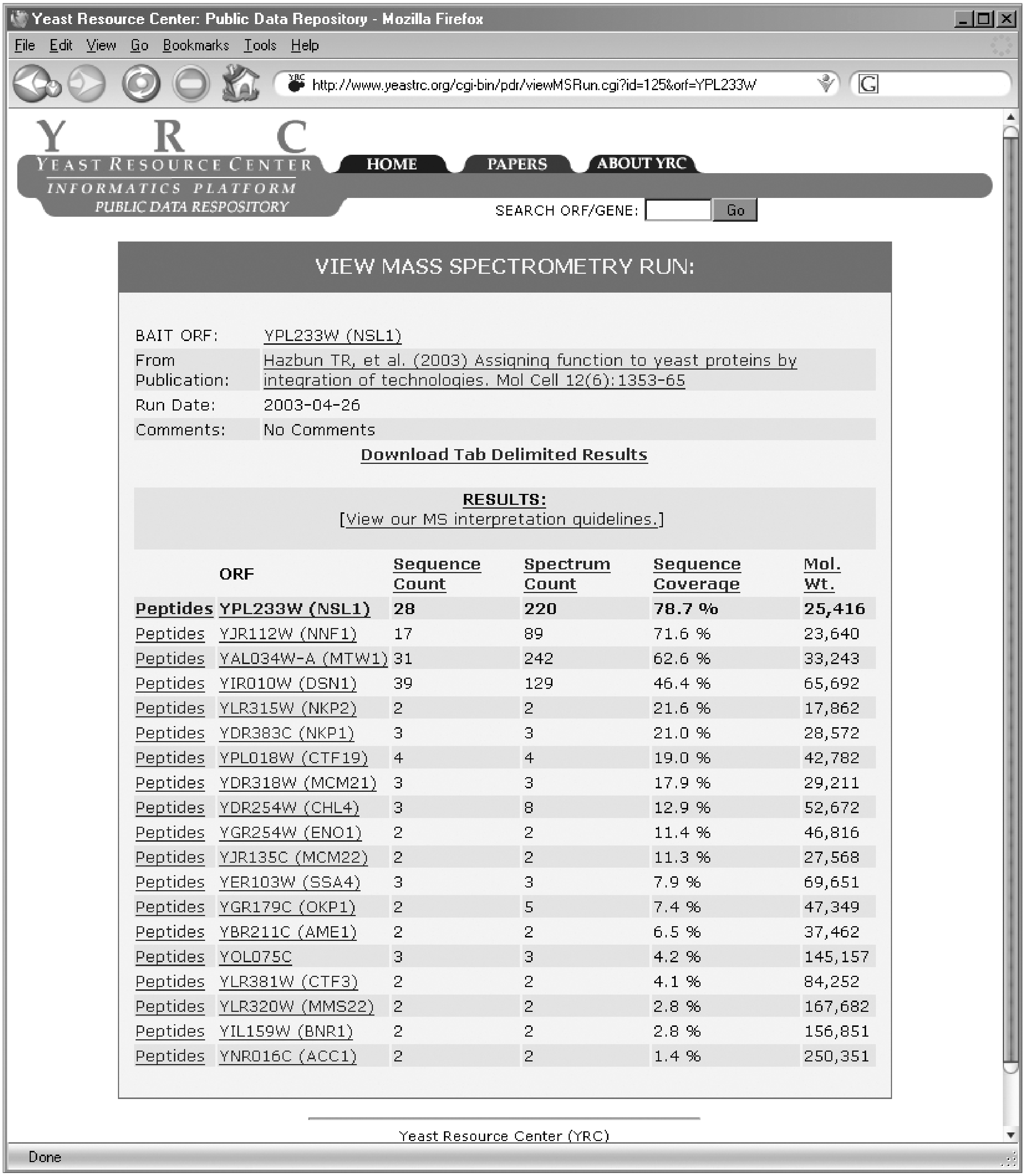
Found 1 predicted domain for YPL233W:

	Beg. Pos.	End. Pos.	Source	Comment
View Domain	1	216	pcons	Spectrin repeat superfamily

*Indicates unpublished data

Yeast Resource Center (YRC)
Informatics Platform Version 2.0
Contact Webmaster

Figure 1. A screen capture of the 'ORF Overview Page' for the *S.cerevisiae* gene *NSL1*. This screen illustrates the result of searching for NSL1 or YPL233w. The page is separated into five distinct sections—Gene Ontology annotations, mass spectrometry, localization, yeast two-hybrid and protein structure prediction. Each section contains a summary of the experimental data relevant to NSL1 and provides links to the data.


Yeast Resource Center: Public Data Repository - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.yeastrc.org/cgi-bin/pdr/viewMSRun.cgi?id=125&orf=YPL233W
Y R C
YEAST RESOURCE CENTER
INFORMATICS PLATFORM
PUBLIC DATA RESPOSITORY
HOME
PAPERS
ABOUT YRC
SEARCH ORF/GENE:
Go
Done

VIEW MASS SPECTROMETRY RUN:

BAIT ORF:	YPL233W (NSL1)
From Publication:	Hazbun TR, et al. (2003) Assigning function to yeast proteins by integration of technologies. Mol Cell 12(6):1353-65
Run Date:	2003-04-26
Comments:	No Comments

Download Tab Delimited Results

RESULTS:
[View our MS interpretation guidelines.]

	ORF	Sequence Count	Spectrum Count	Sequence Coverage	Mol. Wt.
Peptides	**YPL233W (NSL1)**	**28**	**220**	**78.7 %**	**25,416**
Peptides	YJR112W (NNF1)	17	89	71.6 %	23,640
Peptides	YAL034W-A (MTW1)	31	242	62.6 %	33,243
Peptides	YIR010W (DSN1)	39	129	46.4 %	65,692
Peptides	YLR315W (NKP2)	2	2	21.6 %	17,862
Peptides	YDR383C (NKP1)	3	3	21.0 %	28,572
Peptides	YPL018W (CTF19)	4	4	19.0 %	42,782
Peptides	YDR318W (MCM21)	3	3	17.9 %	29,211
Peptides	YDR254W (CHL4)	3	8	12.9 %	52,672
Peptides	YGR254W (ENO1)	2	2	11.4 %	46,816
Peptides	YJR135C (MCM22)	2	2	11.3 %	27,568
Peptides	YER103W (SSA4)	3	3	7.9 %	69,651
Peptides	YGR179C (OKP1)	2	5	7.4 %	47,349
Peptides	YBR211C (AME1)	2	2	6.5 %	37,462
Peptides	YOL075C	3	3	4.2 %	145,157
Peptides	YLR381W (CTF3)	2	2	4.1 %	84,252
Peptides	YLR320W (MMS22)	2	2	2.8 %	167,682
Peptides	YIL159W (BNR1)	2	2	2.8 %	156,851
Peptides	YNR016C (ACC1)	2	2	1.4 %	250,351

Yeast Resource Center (YRC)

Figure 2. A screen capture of the mass spectrometry data view page. Listed are the ORFs identified through mass spectrometry as having co-purified with the bait ORF, along with experimental data and links to peptide information.

expanded to include more sophisticated searching tools, such as searching only published data, searching by species and searching by protein or gene sequence. User-controlled filters will be added to the mass spectrometry results in order to facilitate the user in identifying more meaningful results. In addition, a probability-based algorithm for analyzing multiple mass spectrometry that runs simultaneously will be added to the site, allowing the user to discover probable protein complexes.

ACKNOWLEDGEMENTS

This work is supported by the National Center for Research Resources of the National Institutes of Health by a grant to T.N.D. entitled 'Comprehensive Biology: Exploiting the Yeast Genome', PHS No. P41 RR11823.

REFERENCES

1. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P., Qureshi-Emili,A., Li,Y., Godwin,B., Conover,D., Kalbfleisch,T., Vijayadamodar,G., Yang,M., Johnston,M., Fields,S. and Rothberg,J.M. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
2. Drees,B.L., Sundin,B., Brazeau,E., Caviston,J.P., Chen,G.C., Guo,W., Kozminski,K.G., Lau,M.W., Moskow,J.J., Tong,A., Schenkman,L.R., McKenzie,A.,III, Brennwald,P., Longtine,M., Bi,E., Chan,C., Novick,P., Boone,C., Pringle,J.R., Davis,T.N., Fields,S. and Drubin,D.G. (2001) A protein interaction map for cell polarity development. *J. Cell Biol.*, **154**, 549–571.
3. Hazbun,T.R., Malmstrom,L., Anderson,S., Graczyk,B.J., Fox,B., Riffle,M., Sundin,B.A., Aranda,J.D., McDonald,W.H., Chiu,C.H., Snydsman,B.E., Bradley,P., Muller,E.G., Fields,S., Baker,D., Yates,J.R.,III and Davis,T.N. (2003) Assigning function to yeast proteins by integration of technologies. *Mol. Cell*, **12**, 1353–1365.
4. Sundin,B.A., Chiu,C.H., Riffle,M., Davis,T.N. and Muller,E.G. (2004) Localization of proteins that are coordinately expressed with Cln2 during the cell cycle. *Yeast*, **21**, 793–800.
5. Chivian,D., Kim,D.E., Malmstrom,L., Bradley,P., Robertson,T., Murphy,P., Strauss,C.E., Bonneau,R., Rohl,C.A. and Baker,D. (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins*, **53** (Suppl. 6), 524–533.
6. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
7. Bonneau,R., Tsai,J., Ruczinski,I., Chivian,D., Rohl,C., Strauss,C.E. and Baker,D. (2001) Rosetta in CASP4: progress in *ab initio* protein structure prediction. *Proteins*, **45** (Suppl. 5), 119–126.
8. Simons,K.T., Kooperberg,C., Huang,E. and Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.
9. Simons,K.T., Bonneau,R., Ruczinski,I. and Baker,D. (1999) *Ab initio* protein structure prediction of CASP III targets using ROSETTA. *Proteins*, **37** (Suppl. 3), 171–176.
10. Simons,K.T., Ruczinski,I., Kooperberg,C., Fox,B.A., Bystroff,C. and Baker,D. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, **34**, 82–95.
11. Dwight,S.S., Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dolinski,K., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J., Hong,E.L., Issel-Tarver,L., Nash,R.S., Sethuraman,A., Starr,B., Theesfeld,C.L., Andrada,R., Binkley,G., Dong,Q., Lane,C., Schroeder,M., Weng,S., Botstein,D. and Cherry,J.M. (2004) *Saccharomyces* Genome Database: underlying principles and organisation. *Brief Bioinformatics*, **5**, 9–22.
12. 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., Hong,E.L., Issel-Tarver,L., Nash,R., Sethuraman,A., Starr,B., Theesfeld,C.L., Andrada,R., Binkley,G., Dong,Q., Lane,C., Schroeder,M., Botstein,D. and Cherry,J.M. (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**, D311–D314.
13. 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., Harris,M.A., Hill,D.P., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J.C., Richardson,J.E., Ringwald,M., Rubin,G.M. and Sherlock,G. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
14. Tabb,D.L., McDonald,W.H. and Yates,J.R.,III (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.*, **1**, 21–26.
15. Walther,D. (1997) WebMol—a Java-based PDB viewer. *Trends Biochem. Sci.*, **22**, 274–275.