Nucleic Acids Research, 2001, Vol. 29, No. 1 75–79

YPD™, PombePD™ and WormPD™: model organism volumes of the BioKnowledge™ Library, an integrated resource for protein information

Maria C. Costanzo*, Matthew E. Crawford, Jodi E. Hirschman, Janice E. Kranz, Philip Olsen, Laura S. Robertson, Marek S. Skrzypek, Burkhard R. Braun, Kelley Lennon Hopkins, Pinar Kondu, Carey Lengieza, Jodi E. Lew-Smith, Michael Tillberg and James I. Garrels

Proteome, Inc., 100 Cummings Center, Suite 435M, Beverly, MA 01915, USA

Received October 12, 2000; Accepted October 17, 2000

ABSTRACT

The BioKnowledge Library is a relational database and web site (http://www.proteome.com) composed of protein-specific information collected from the scientific literature. Each Protein Report on the web site summarizes and displays published information about a single protein, including its biochemical function, role in the cell and in the whole organism, localization, mutant phenotype and genetic interactions, regulation, domains and motifs, interactions with other proteins and other relevant data. This report describes four species-specific volumes of the BioKnowledge Library, concerned with the model organisms *Saccharomyces cerevisiae* (YPD), *Schizosaccharomyces pombe* (PombePD) and *Caenorhabditis elegans* (WormPD), and with the fungal pathogen *Candida albicans* (CalPD™). Protein Reports of each species are unified in format, easily searchable and extensively cross-referenced between species. The relevance of these comprehensively curated resources to analysis of proteins in other species is discussed, and is illustrated by a survey of model organism proteins that have similarity to human proteins involved in disease.

INTRODUCTION

Now that the goal of obtaining complete genomic sequence information has been met for several eukaryotic model organisms (1–3; see http://www.sanger.ac.uk/Projects/S_pombe/), the next great challenge lies in the organization, interpretation and utilization of these data. A wealth of information about specific proteins lies buried not only in their sequences, but also in a vast, under-utilized resource: the biological research literature. Mining this resource for experimental data and tying these data to specific proteins is a difficult task for researchers studying small sets of proteins, and it is impossible for an individual to do on a large scale. Proteome's BioKnowledge Library is intended to facilitate this task for several different types of scientist: those focusing on a particular process in a particular organism; those engaged in more global experiments generating data on hundreds or thousands of genes or proteins; and those seeking to annotate newly sequenced, uncharacterized genes by comparison with their characterized orthologs in model organisms.

This article focuses on the model organism volumes of the BioKnowledge Library: YPD, for *Saccharomyces cerevisiae*; PombePD, for *Schizosaccharomyces pombe*; and WormPD, for *Caenorhabditis elegans*. Another volume concerning the important fungal pathogen of humans, *Candida albicans* (CalPD) is also discussed; additional volumes of the BioKnowledge Library concerning mammalian proteins are currently available by subscription (HumanPSD™ Proteome Survey Databases for human, mouse and rat proteins, and GPCR-PD™ for G protein-coupled receptors and associated signaling partners).

The model organism volumes and CalPD are organized in a one Web page-per-protein format, described previously (4). Each Protein Report contains information collected from the biological literature by trained curators (Ph.Ds with research experience in the model organisms) who read the full text of articles. Information about important properties of the protein, including biochemical function, cellular role, subcellular localization and many others, is tabulated in the top section of the Protein Report, while more complex data are summarized in free-text annotations organized by topic in the lower section, and a complete reference list is displayed at the bottom of each page (see Supplementary Material for a detailed diagram of the Protein Report format). A detailed search form allows searching by many individual criteria as well as the construction of more complex queries, including limiting the search to experimentally verified results for some protein properties.

While the basic Protein Report format has remained relatively constant over the past year, the content of the BioKnowledge Library has continued to grow at a rapid pace. We have added an entirely new model organism volume, PombePD; we are collecting new types of information, with a particular emphasis on large-scale functional genomics experiments; and we update Protein Reports continuously as new literature is published. All of these factors contribute to the unparalleled utility of the BioKnowledge Library as a resource for basic research, drug discovery and comparative genomics. The model organism

*To whom correspondence should be addressed. Tel: +1 978 922 1643; Fax: +1 978 922 3971; Email: mcc@proteome.com

volumes of the BioKnowledge Library are freely accessible to academic users at http://www.proteome.com/databases; CalPD is available to academic users by registration. All volumes of the BioKnowledge Library are available by subscription to corporate users.

GROWTH OF YPD AND WormPD

Last year's report on Proteome's model organism volumes (4) described YPD and the newly introduced WormPD, based on the complete genomic sequences of *S.cerevisiae* and *C.elegans*, respectively. The protein sequences that form the foundations of YPD and WormPD continue to be revised with the latest sequence updates. A major update to the *S.cerevisiae* genomic sequence occurred with the re-sequencing of chromosome III (see http://www.mips.biochem.mpg.de/proj/yeast/). All of the sequence changes resulting from this project, affecting 68 predicted proteins, were incorporated into YPD. The predicted protein sequences in WormPD are kept consistent with WormPep releases from the Sanger Centre (see http://www.sanger.ac.uk/Projects/C_elegans/). Additionally, new GenBank entries for proteins of both species are reviewed daily by sequence editors.

The number of annotation lines in YPD had risen to nearly 145 000 by September 2000, and more than 17 000 references had been curated. WormPD contained more than 35 700 annotation lines derived from over 2400 articles, representing complete coverage of the *C.elegans* gene- and protein-specific literature. YPD coverage of the much larger body of *S.cerevisiae* literature is comprehensive for the literature of the past three years, and is more than 70% complete for the earlier part of the past decade. We are continuing efforts to identify older papers with relevance to specific *S.cerevisiae* proteins and to complete their curation.

ADDITION OF PombePD TO THE BIOKNOWLEDGE LIBRARY

In July 2000, the latest model organism volume was added to the BioKnowledge Library: PombePD, concerning the proteins of 'fission yeast'. *Schizosaccharomyces pombe* is well-studied, particularly in the areas of cell cycle and DNA repair and replication, and may be a better model for some cellular processes in higher eukaryotes (such as cell cycle and DNA repair) than *S.cerevisiae* (5). As of September 2000, the *S.pombe* genomic sequence (coordinated at the Sanger Centre, http://www.sanger.ac.uk/Projects/S_pombe/) was nearly complete, and PombePD contained approximately 4800 predicted protein sequences. The sequences in PombePD and in PomBase (http://www.sanger.ac.uk/Projects/S_pombe/pombase.shtml) are compared frequently to ensure that PombePD contains a complete, non-redundant set of protein sequence entries. New *S.pombe* sequence entries in GenBank are also incorporated into PombePD daily.

The format of PombePD is very similar to that of YPD and WormPD; similar types of information are collected and summarized in a frequently updated Title Line, in controlled-vocabulary properties and in free-text annotations. Like WormPD, PombePD has a detailed Mutant Phenotype property. Each mutation is classified by phenotype and by mutant type (i.e. null, gain-of-function or reduction-of-function). These fields are searchable, so that users can easily generate a list of genes that, when mutated, give rise to a phenotype of interest.

By September 2000, more than 2200 research articles had been curated for PombePD, generating nearly 29 900 annotation lines. The addition of PombePD brings the model organism volumes of the BioKnowledge Library, taken together, to a total of 29 506 Protein Reports, with over 210 000 annotation lines generated from coverage of over 21 800 references. They also contain over 16 000 property entries in the Biochemical Function, Cellular Role, Subcellular Localization, Molecular Environment and Mutant Phenotype properties; nearly 7000 of these protein properties are designated as derived from experimental results.

A FUNGAL PROTEIN RESOURCE

The inclusion of YPD, PombePD and CalPD in the BioKnowledge Library make it a powerful resource for fungal protein information with application to antifungal research. CalPD contains comprehensive coverage of the literature about *C.albicans* proteins, and all full sequences of *C.albicans* proteins from GenBank. In early 2001, CalPD will be expanded into a unified volume of the BioKnowledge Library containing protein information not only about *C.albicans* but also about the other major fungal pathogens of humans: *Histoplasma capsulatum*, *Blastomyces dermatitidis*, *Coccidioides immitis*, *Pneumocystis carinii*, *Cryptococcus neoformans*, *Aspergillus fumigatus*, *Aspergillus flavus*, *Aspergillus niger* and several additional *Candida* species.

FUNCTIONAL GENOMICS

As the number of large-scale functional genomics experiments increases rapidly, we are engaged in an ongoing analysis of how best to incorporate this information into the BioKnowledge Library. The nature of the automated, high-throughput techniques necessitates that the results of such experiments be differentiated from the results of experiments using 'classical' techniques to study individual genes or proteins in great detail and with greater accuracy. A new annotation topic, 'Functional Genomics', was added to the model organism volumes to hold annotations derived from several different types of high-throughput experiments: large-scale deletion mutant analyses, systematic two-hybrid interaction studies, global transcription profiling, subcellular localization and others. In September 2000, 4648 Protein Reports in YPD, 244 in PombePD and 119 in WormPD contained annotations in the Functional Genomics topic.

The property section of each Protein Report of YPD contains a link to a collection of transcript profiling results for that particular gene. The number of such datasets added to this list has continued to increase, with the results of 97 experiments contained in YPD as of September 2000 (comprising about 1 160 000 individual expression values). Next to the name of each experiment in the list appears a thumbnail summary of the transcription profile, to facilitate finding experiments of interest, and a hyperlink leads to a pop-up window containing a description of the experiment and the results for that gene. We expect to add transcriptional profiling results to WormPD, PombePD and CalPD as such results become available in the future.

Table 1. Proportions of *S.cerevisiae*, *S.pombe* and *C.elegans* proteins that are characterized to various extents

	S.cerevisiae	*S.pombe*	*C.elegans*
Total number of known and predicted proteins[a]	6145	4829	18 545
Proteins characterized by genetics or biochemistry[b]	3778 (61% of total)	889 (18% of total)	1541 (8% of total)
Proteins known by similarity to characterized proteins[c]	701 (11% of total)	2384 (49% of total)	10 227 (55% of total)
Proteins of unknown function[d]	1665 (27% of total)	1564 (32% of total)	7118 (38% of total)

[a]Total number of predicted proteins as of September 2000 (before completion of the *S.pombe* genomic sequencing effort).
[b]Proteins with experimentally determined function, role, subcellular localization, modification or protein–protein association, or with known mutant phenotypes beyond essentiality/non-essentiality.
[c]Experimentally uncharacterized proteins with similarity to other proteins of known or predicted function, or containing domains or motifs of known function.
[d]Experimentally uncharacterized proteins with no similarity to other proteins, or with similarity to uncharacterized proteins.

INTEGRATION AND USES OF THE MODEL ORGANISM VOLUMES

YPD, PombePD and WormPD encapsulate many years' research and thousands of publications on individual proteins of these model organisms. More than 61% of *S.cerevisiae* proteins have been studied and experimentally characterized, and significant proportions of the proteomes of *S.pombe* and *C.elegans* have also been subjected to experimentation (Table 1).

All Protein Reports of the BioKnowledge Library are seamlessly integrated: whenever a protein of one species covered is mentioned in a BLAST report or annotation, its name is hyperlinked to its Protein Report, so that a user can browse across species as easily as within a species. (The Protein Reports of each species are color-coded and labeled so that it is immediately obvious to which species a protein belongs.) This integration greatly facilitates the application of experimental information about characterized proteins in one species to functional predictions about uncharacterized, similar proteins in other species.

It is of interest to ask how many protein sequences are conserved in the model organisms, relative to each other and to mammals. A comparison of the predicted proteins of *S.cerevisiae*, *S.pombe* and *C.elegans*, presented in Figure 1A, shows that a significant proportion of the proteins of each organism is strongly conserved in one or both of the other species. A comparison of proteins of the model yeasts (*S.cerevisiae* and *S.pombe* combined), *C.elegans* and mammals (human, mouse and rat combined), presented in Figure 1B, reveals conservation of many proteins between the model organisms and higher eukaryotes. It is evident that, given the proportion of proteins of the model organisms that are experimentally characterized (Table 1) and the proportion that have significant similarity to proteins of non-model organisms (Fig. 1B), the model organism volumes of the BioKnowledge Library are a valuable tool for illuminating the possible functions of proteins from humans or other organisms.

To illustrate this, we used the BioKnowledge Library to look for proteins in the model organisms with moderate or greater similarity (*E* value of 1×10^{-10} or lower) to human proteins known to be involved in disease processes. Of 108 human inherited disease genes identified by positional cloning (see http://genome.nhgri.nih.gov/clone/), 77 have a match by these criteria in at least one of the organisms *S.cerevisiae*, *S.pombe* or *C.elegans.* Most of the disease proteins are conserved in *C.elegans* (74/77) but a substantial number are also conserved in *S.pombe* (36/77) and/or *S.cerevisiae* (42/77; see Supplementary Material for the complete data set). Many are quite strongly conserved, with 48 *C.elegans* proteins matching their human counterparts with *E* values of 1×10^{-50} or lower; 17 *S.pombe* proteins and 19 *S.cerevisiae* proteins also find matches of this significance. Sequence similarity of this extent between model organism proteins and human proteins does not in itself indicate orthology. However, functional studies and cross-complementation suggest that some of the proteins considered in this comparison are indeed orthologs, even in some cases where the *E* value is relatively high, indicating relatively little primary sequence similarity (see Table S1).

The integration of the BioKnowledge Library makes it straightforward to navigate from the Protein Report of a protein involved in human disease found in the subscription database HumanPSD™ or GPCR-PD™, either to the BLAST report which shows alignments of that protein with model organism proteins or directly to the Protein Report of a similar protein in the model organism volumes. It is then possible to assess rapidly whether the model organism protein has been experimentally analyzed. A significant number of the model organism proteins in this set have indeed been experimentally characterized: 21 of the *C.elegans* proteins, 12 of the *S.pombe* proteins and 33 of the *S.cerevisiae* proteins (see Supplementary Material). Some particularly striking examples of the utility of model organism research, as captured in the BioKnowledge Library, for illuminating the possible functions of proteins of other organisms, are discussed in the Supplementary Material.

CONTACTING AND CITING YPD, PombePD, WormPD AND CalPD

We welcome comments and corrections from users. Unpublished data may be submitted for inclusion in the databases, and will be cited as a personal communication. Functional genomic datasets are especially welcomed, and users with functional genomics web sites are encouraged to link to our site. Any correspondence, including requests for spreadsheets containing subsets of YPD and WormPD data, should be directed to ypd@proteome.com, pombepd@proteome.com, wormpd@proteome.com, calpd@proteome.com or by mail to the address of the authors.

Authors wishing to make use of the information provided by the BioKnowledge Library should cite this article as a general

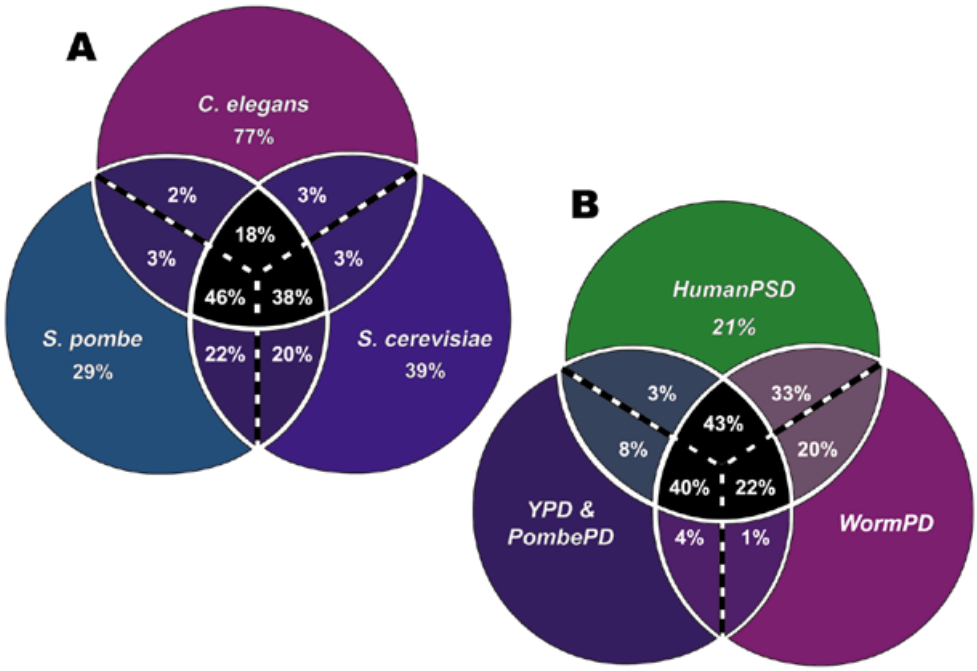

Figure 1. (**A**) Proportion of proteins conserved among the model organisms. Each circle represents the set of proteins analyzed from that species and contained in the corresponding Proteome database. Complete sets of the predicted proteins (as of September 2000) of *S.cerevisiae* (6145 proteins) and *C.elegans* (18 546 proteins) and a nearly complete set of *S.pombe* predicted proteins (4837) were compared with each other using BLAST analysis (6,7), and matches with an *E* value of 1×10^{-10} or lower were counted. In the intersection of all three circles (colored in black) are proteins that find a match in all three species, represented as the percentage of total proteins in that species. At the intersection of each pair of circles is the percentage of proteins conserved between those two species but not found in the third, by these criteria. For example, 38% of *S.cerevisiae* proteins find a match in both *S.pombe* and *C.elegans*; 20% match *S.pombe* but not *C.elegans*; and 3% match *C.elegans* but not *S.pombe*. The proportions of proteins similar to a protein of at least one of the other species are: 61% of *S.cerevisiae* proteins, 71% of *S.pombe* proteins and 23% of *C.elegans* proteins find a match in at least one of the other species. (**B**) Comparison of predicted proteins of the model yeasts (*S.cerevisiae*, 6145 proteins and *S.pombe*, 4837 proteins), *C.elegans* (18 546 proteins) and mammals (10 229 from human, 5966 from mouse and 3188 from rat), as represented in the indicated volumes of the BioKnowledge Library as of September 2000. The proteins analyzed represent the complete predicted proteomes of *S.cerevisiae* and *C.elegans*, the nearly complete proteome of *S.pombe* and a partial set of mammalian proteins. Criteria for counting similar proteins were the same as for the comparison of (A). Proteins conserved in all three species are indicated in the intersection of all three circles, as in (A). The proportions of proteins similar to a protein of at least one of the other species are: 79% of the mammalian proteins, 43% of the *C.elegans* proteins and 52% of the model yeast proteins.

reference for access to and content of YPD, PombePD, WormPD and CalPD.

SUPPLEMENTARY MATERIAL

The following Supplementary Material is available at NAR Online: (i) view a sample Protein Report; (ii) view the YPD Full Search form; (iii) Table S1: Conservation of human disease proteins in *S.cerevisiae*, *S.pombe* and *C.elegans*; (iv) Table S2: Examples illustrating the utility of model organism protein-specific information to the study of human disease.

ACKNOWLEDGEMENTS

We are grateful for the hard work and enthusiasm of our curators. We would like to thank the many members of the international *S.cerevisiae*, *S.pombe*, *C.elegans* and *C.albicans* research communities who take time to communicate updates and corrections to the databases. Several members of the *S.pombe* research community provided invaluable help in the creation of PombePD: Valerie Wood, Charles Hoffman, Susan Forsburg, Judith Potashkin, Paul Young, Mitsuhiro Yanagida, Nigel Peat and J.Richard McIntosh. We also thank the staffs of the Saccharomyces Genome Database (http://genome-www.stanford.edu/Saccharomyces), the Munich Information Centre for Protein Sequences (http://www.mips.biochem.mpg.de), the Sanger Centre *S.pombe* sequencing project (http://www.sanger.ac.uk/Projects/S_pombe/) and A *C.elegans* Database (http://www.sanger.ac.uk/Projects/C_elegans/webace_front_end.shtml and http://stein.cshl.org/elegans) for their help and cooperation. We thank Scott MacDonald for assistance with graphics. The development of PombePD was funded by a Phase II SBIR grant from the National Institute of Allergy and Infectious Diseases (5 R44 AI43728-03).

REFERENCES

1. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 563–567.
2. The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.

3. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
4. Costanzo,M.C., Hogan,J.D., Cusick,M.E., Davis,B.P., Fancher,A.M., Hodges,P.E., Kondu,P., Lengieza,C., Lew-Smith,J.E., Lingner,C. *et al.* (2000) The Yeast Proteome Database (YPD) and *Caenorhabditis elegans* Proteome Database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.*, **28**, 73–76.
5. Forsburg,S.L. (1999) The best yeast? *Trends Genet.*, **15**, 340–344.
6. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
7. Waterman,M.S. (1995) *Introduction to Computational Biology: Maps, Sequences and Genomes.* Chapman & Hall, London.