YPD—A database for the proteins of Saccharomyces cerevisiae

James I. Garrels

Proteome, Inc., 181 Elliott Street, Suite 909, Beverly, MA 01915, USA

Received September 8, 1995; Revised and Accepted November 3, 1995

ABSTRACT

YPD is a database for the proteins of the budding yeast, Saccharomyces cerevisiae. YPD has two formats: (i) a spreadsheet which tabulates many of the physical and functional properties of yeast proteins, and (ii) the YPD Protein Reports which are formatted pages containing the protein properties, annotations gathered from the literature, and references with titles. YPD is available through the World-Wide Web, through an Email server, and by anonymous FTP. New releases of the YPD spreadsheet are produced every two to four months, and the on-line information is updated daily.

INTRODUCTION

YPD is a protein database specialized in the collection of the physical and functional information for the proteins of budding yeast Saccharomyces cerevisiae. YPD evolved from a project at the QUEST Protein Database Center of the Cold Spring Harbor Laboratory to identify yeast proteins on two-dimensional gels (1,2). For this project, knowledge of the predicted isoelectric point, molecular weight, codon bias, post-translational modifications, subcellular localizations, precursor peptide lengths, and amino acid composition of mature proteins was needed to assist in protein identifications. A tabulation of data from the protein sequence databases and from the literature was begun. The tabulation was made non-redundant by screening multiple versions of each sequence from the database. YPD has now been expanded into a general resource that is available (i) as a spreadsheet for downloading into personal computers, (ii) through an Email server, and (iii) through the World-Wide Web (WWW).

YPD is complementary to other biological databases in content and in format. The Saccharomyces Genome Database (SGD), a primary resource for yeast genetic information maintained by J. Michael Cherry at the Saccharomyces Genome Information Resource (3) at Stanford University, has links to YPD information through its ‘protein-info’ class. YPD does not contain sequence information, but instead gives accession numbers to GenBank, SWISS-PROT, and PIR-International. The spreadsheet format of YPD allows users to load and manipulate the entire database of protein properties and references on a personal computer. A different format, the YPD Protein Reports, is used for presentation on the WWW and Email servers. Each YPD Protein Report displays protein properties, functional annotations, and references for one protein. The YPD Protein Reports are searchable by gene name, by keywords, and by the protein property categories.

YPD was first released as a spreadsheet for downloading by anonymous FTP in November, 1994. New releases of the spreadsheet have been provided every 2 to 4 months. YPD has been available since 1994 on the QUEST World-Wide Web server at the Cold Spring Harbor Laboratory, where it is interfaced to two-dimensional gel maps through the Global Gel Navigator Software (2), and on the Saccharomyces Genome Database (SGD) WWW server. The YPD Protein Reports were made available in 1995 through a new World-Wide Web server and Email server at Proteome, Inc. The YPD database is maintained at Proteome, Inc. as a free public resource.

THE YPD SPREADSHEET

The YPD spreadsheet contains one record (row) for each yeast protein of known sequence, including the open reading frames discovered by systematic genome sequencing. The identifier for each record is the gene name. The genetic names for budding yeast are coordinated at the Saccharomyces Genome Database at Stanford, using the published literature, the LISTA database of Patrick Linder (4), and direct submissions from investigators. For each record, the YPD spreadsheet lists the SGD gene name, the SWISS-PROT gene name, and a list of synonyms that includes all known names used for the gene, including temporary and permanent names assigned by systematic sequencing.

Accession numbers to GenBank, PIR-International and SWISS-PROT are given as fields in the spreadsheet. When multiple sequences for a gene are available in the databases, the sequence from systematic sequencing is selected or, if that is not available, the most recent sequence is usually selected. Calculations of isoelectric point, molecular weight, codon bias, codon adaptation index, and amino acid composition, etc. are based on the designated GenBank sequence. Fragments of N- and C-terminal sequence are given in the spreadsheet to help users verify the identity of the protein in cases where the nomenclature is confusing.

All calculations of predicted properties, motifs, and amino acid composition are based on the mature protein sequence after removal of N- and C-terminal precursor peptides. YPD relies on the experimental or strongly predicted cleavage sites reported in the literature. YPD itself does not make cleavage site predictions. The complete list of fields (columns) used in the spreadsheet format of YPD are listed in Figure 1.
Figure 1. Protein information fields maintained in the YPD spreadsheet. Each item represents one column of the spreadsheet. On the spreadsheet preformatted for Microsoft Excel, the columns are presented in the order and the groupings shown here.

THE YPD HOME PAGE

The YPD Home Page on the World-Wide Web, shown in Figure 2, provides access to all formats of YPD, including the online search form for access to YPD Protein Reports, the Global Gel Navigator on the QUEST WWW server (2), the Email server, and the FTP server. It also contains links to pages with introductory material, documentation, and summaries of the YPD contents.

THE YPD PROTEIN REPORTS

An example YPD Protein Report is shown in Figure 3. This format presents 34 fields of data from the spreadsheet, annotations gathered from the literature, and the reference list with titles. This format is used to present information for each protein as an Email report or as a single WWW page. These are recompiled daily so that the latest updates are always available on-line. In addition, a file containing all the YPD Protein Reports on the date of the latest spreadsheet release can be downloaded by anonymous FTP (see below).

For each YPD Protein Report obtained through WWW, hypertext links are available for immediate access to SGD, GenBank, PIR-International, SWISS-PROT, and Entrez (5). The Entrez server from NCBI provides abstracts for most references used in YPD.

SEARCHING YPD

The spreadsheet version of YPD is inherently searchable by its categorized protein properties, however, it must be downloaded to be used and it does not contain the functional annotations. The Email and WWW versions have search forms that allow...
Figure 3. A sample YPD Protein Report. The top field is the protein name/description field from the spreadsheet. The following fields present many of the protein properties from the YPD spreadsheet. After the property fields, annotations from the literature and the list of references are given.

convenient searching of the YPD Protein Reports by gene name or synonym, by keywords, and by the protein properties. The form allows ‘AND’ and ‘OR’ modes for construction of queries based on multiple criteria. These Email and WWW servers always use data from the latest daily updates. The result of each search is a page containing a synopsis of the search strategy, and a list of the protein ‘hits’ by gene name, synonyms, and protein name/description. On the WWW page, clicking any protein in the ‘hit’ list brings up the corresponding YPD Protein Report.

CONTENTS OF THE CURRENT RELEASE

Release 4.1 contains 4305 entries including sequences from GenBank through July 30, 1995, SWISS-PROT through May 23, 1995, and PIR-International Release 43. Of these sequences, 3789 derive from systematic genomic sequencing projects. YPD currently lists 2012 proteins that have been characterized through genetics or biochemistry, 729 proteins that are known only by homology to characterized proteins, and 1564 that have unknown function. In the current release, YPD tabulates 477 nuclear proteins, 249 mitochondrial proteins, 136 transcription factors, 94 protein kinases, 16 cyclins, and many other categories relating the function, localization, and modification of the proteins. A complete summary of the YPD contents by category is found in the spreadsheet documentation and under YPD Summaries on the YPD Home Page.

Table 1. YPD contents versus release number

<table>
<thead>
<tr>
<th>Release</th>
<th>Date</th>
<th>Total</th>
<th>Knowna</th>
<th>Homolb</th>
<th>Unknownc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>Nov. 23, 1994</td>
<td>3020</td>
<td>1729</td>
<td>387</td>
<td>904</td>
</tr>
<tr>
<td>2.0</td>
<td>Dec. 8, 1994</td>
<td>3142</td>
<td>1750</td>
<td>450</td>
<td>942</td>
</tr>
<tr>
<td>3.0</td>
<td>Feb. 1, 1995</td>
<td>3512</td>
<td>1871</td>
<td>524</td>
<td>1117</td>
</tr>
<tr>
<td>4.0</td>
<td>Jun. 6, 1995</td>
<td>4046</td>
<td>1951</td>
<td>667</td>
<td>1428</td>
</tr>
<tr>
<td>4.1</td>
<td>Jul. 7, 1995</td>
<td>4305</td>
<td>2012</td>
<td>729</td>
<td>1564</td>
</tr>
</tbody>
</table>

*Proteins characterized through genetic or biochemical experiments.

aProteins that have not been characterized but have sequence similarity to characterized proteins.

bProteins of completely unknown function.

HOW TO ACCESS YPD

The FTP server contains the spreadsheet data files (YPD_Excel) formatted for Microsoft Excel on Macintosh and formatted as tab-delimited ASCII text (YPD_ASCII) for loading into any spreadsheet. The associated files include documentation (YPD.doc), a references file (YPD_REFS), and a file (YPD_FORMATTED) containing all the YPD Protein Reports. The address of the FTP server is isis.cshl.org, and the directory is/pub/yeast/YPD. The FTP server can also be accessed from the YPD Home Page.

The Email server is accessed by sending Email to yeast@proteome.com. YPD Protein Reports are requested by placing one or more gene names in the subject line. The search form is requested by placing ‘HELP’ in the subject line. Documentation is available by placing ‘DOC’ in the subject line. The YPD Protein Reports and other documents are automatically returned, with each report in a separate Email message.

YPD can be reached on the World-Wide Web through the YPD Home Page (http://www.proteome.com/YPDhome.html). Data from YPD has also been incorporated into the WWW server of the QUEST Protein Database Center (http://siva.cshl.org/#ypd), the Saccharomyces Genome Database (SGD) (http://genome-www.stanford.edu), and the MIPS Protein Database/ Yeast Genome Database (http://www.mips.biochem.mpg.de). The QUEST site is managed by Gerald I. Latter (latter@cshl.org), SGD is managed by J. Michael Cherry (cherry@genome.stanford.edu), and MIPS is managed by Werner Mewes
The implementation of YPD on the MIPS WWW server is managed by Karl Kleine (kleine@mips.embnet.org) and Cynthia Harris (harris@mips.embnet.org). The YPD Home Page and the contents of YPD are maintained by the author (jg@proteome.com).

Authors wishing to cite YPD should use this article as a general reference for the latest release available electronically.

ACKNOWLEDGEMENTS

I am grateful to Gerald Latter and Bruce Futcher for discussions that led to the YPD Project, and to Gerald Latter and Tom Boutell for their pioneering implementation of YPD on the QUEST WWW server. I thank Michael Cusick, Les Grivell, Bruno André, Jonathan Warner, and Rick Moerschell for expert review and assistance with portions of the database. Finally I thank Irene Ong for WWW assistance, and Rachelle Hecht and Shelley Lengieza for assistance in building YPD, and Cheryl Lengieza for assistance in preparing the manuscript.

REFERENCES