46–49 Nucleic Acids Research, 1996, Vol. 24, No. 1

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.

YPD SPREADSHEET FIELDS

a. Gene Names
- YPD gene name
- SWISS-PROT/LISTA gene name
- *Saccharomyces* Genome Database name
- Synonym list

b. Technical flags (used primarily in database development)
- Identical sequence flags (entries with same number are identical)
- Closely-related sequence flags (entries with same number are related)
- Database source (0 = PIR, not in GenBank; 1 = GenBank major release; 2 = GenBank cumulative update; 3 = GenBank daily update)
- Systematic sequence flag (1 = derived from systematic sequencing)

c. Calculated data
- Isoelectric point
- Isoelectric point after adding 1 positive charge
- Isoelectric point after adding 1 negative charge
- Molecular weight
- Codon bias
- Codon adaptation index

d. Genetic data
- Chromosome number
- Presence or absence of intron in gene
- Knockout mutation (L = lethal, V = viable)

e. Accession numbers
- GenBank
- PIR-International
- SWISS-PROT
- YEPD (2D gel database numbers)

f. Subcellular localization and functional classification
- Major localization category (nuclear, mitochondrial etc)
- Minor localization category (mitochondrial inner membrane, etc)
- Molecular environment (integral membrane, DNA-associated, etc)
- Functional classification (protein kinase, transcription factor, etc)

g. Post-translational modifications and length
- N-terminal modification (acetylation, myristoylation)
- C-terminal modification (farnesylation, geranylgeranylation, etc.)
- Phosphorylation
- N- or O-linked glycosylation
- N-terminal precursor length
- Mature protein length (in amino acids after removal of N- and C-terminal precursor peptides)

h. Amino acid composition
- 20 amino acid fields (number of residues in mature protein)
- Met-adjust field (1 indicates N-met is predicted but not known to be removed)

i. Motifs
- Potential sites for phosphorylation by Cdc28 protein kinase
- Potential sites for phosphorylation by CKII protein kinase
- Potential sites for phosphorylation by PKA protein kinase
- Potential sites for N-linked glycosylation
- Potential transmembrane domains

j. N- and C-terminal sequence fragments
- N-terminal sequence of precursor protein
- N-terminal sequence of mature protein
- C-terminal sequence of mature protein

k. Protein name/description and references
- Protein name and descriptive phrases
- List of references (Refers to numbered references in YPD REFS file)

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 are obtained after making selections on the search form by gene name, by keywords, or by protein property categories.

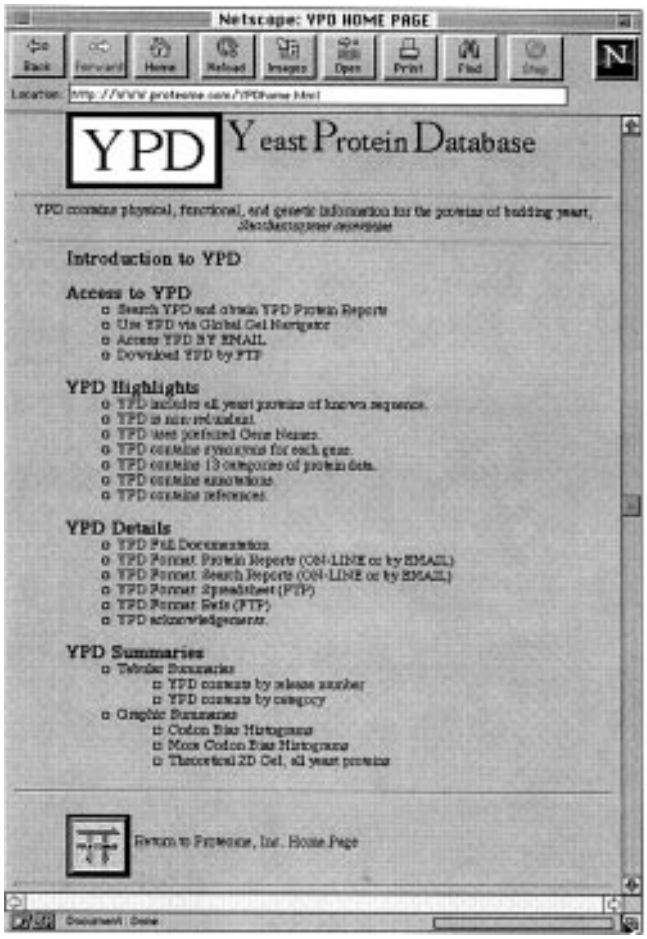


Figure 2. The YPD Home Page. World-Wide Web users can access the page at http://www.proteome.com/YPDhome.html. The YPD Home Page can be used to access YPD in any of its formats or to display further information about YPD.

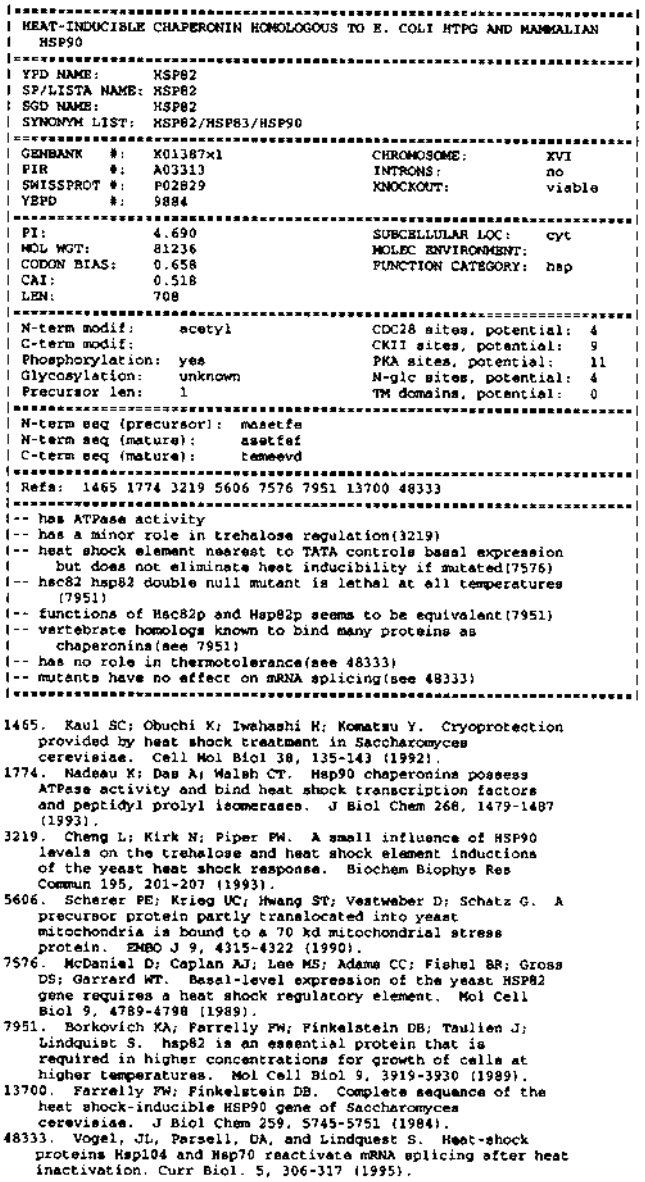
THE YPD PROTEIN REPORTS

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 is does not contain the functional annotations. The Email and WWW versions have search forms that allow



```
HEAT-INDUCIBLE CHAPERONIN HOMOLOGOUS TO E. COLI HTPG AND MAMMALIAN
  HSP90

YPD NAME:          HSP82
SP/LISTA NAME:     HSP82
SGD NAME:          HSP82
SYNONYM LIST:      HSP82/HSP83/HSP90

GENBANK    #:      X01387x1                CHROMOSOME:          XVI
PIR        #:      A03313                  INTRONS:             no
SWISSPROT  #:      P02829                  KNOCKOUT:            viable
YEPD       #:      9884

PI:                4.690                   SUBCELLULAR LOC:     cyt
MOL WGT:           81236                   MOLEC ENVIRONMENT:
CODON BIAS:        0.658                   FUNCTION CATEGORY:   hsp
CAI:               0.518
LEN:               708

N-term modif:      acetyl                  CDC28 sites, potential:  4
C-term modif:                              CKII sites, potential:   9
Phosphorylation:   yes                     PKA sites, potential:    11
Glycosylation:     unknown                 N-glc sites, potential:  4
Precursor len:     1                       TM domains, potential:   0

N-term seq (precursor):  masetfe
N-term seq (mature):     asetfef
C-term seq (mature):     teneevd

Refs:  1465 1774 3219 5606 7576 7951 13700 48333

-- has ATPase activity
-- has a minor role in trehalose regulation(3219)
-- heat shock element nearest to TATA controls basal expression
     but does not eliminate heat inducibility if mutated(7576)
-- hsc82 hsp82 double null mutant is lethal at all temperatures
     (7951)
-- functions of Hsc82p and Hsp82p seems to be equivalent(7951)
-- vertebrate homologs known to bind many proteins as
     chaperonins(see 7951)
-- has no role in thermotolerance(see 48333)
-- mutants have no effect on mRNA splicing(see 48333)

1465.  Kaul SC; Obuchi K; Iwahashi H; Komatsu Y.  Cryoprotection
    provided by heat shock treatment in Saccharomyces
    cerevisiae.  Cell Mol Biol 38, 135-143 (1992).
1774.  Nadeau K; Das A; Walsh CT.  Hsp90 chaperonins possess
    ATPase activity and bind heat shock transcription factors
    and peptidyl prolyl isomerases.  J Biol Chem 268, 1479-1487
    (1993).
3219.  Cheng L; Kirk N; Piper PW.  A small influence of HSP90
    levels on the trehalose and heat shock element inductions
    of the yeast heat shock response.  Biochem Biophys Res
    Commun 195, 201-207 (1993).
5606.  Scherer PE; Krieg UC; Hwang ST; Vestweber D; Schatz G.  A
    precursor protein partly translocated into yeast
    mitochondria is bound to a 70 kd mitochondrial stress
    protein.  EMBO J 9, 4315-4322 (1990).
7576.  McDaniel D; Caplan AJ; Lee MS; Adams CC; Fishel BR; Gross
    DS; Garrard WT.  Basal-level expression of the yeast HSP82
    gene requires a heat shock regulatory element.  Mol Cell
    Biol 9, 4789-4798 (1989).
7951.  Borkovich KA; Farrelly FW; Finkelstein DB; Taulien J;
    Lindquist S.  hsp82 is an essential protein that is
    required in higher concentrations for growth of cells at
    higher temperatures.  Mol Cell Biol 9, 3919-3930 (1989).
13700.  Farrelly FW; Finkelstein DB.  Complete sequence of the
    heat shock-inducible HSP90 gene of Saccharomyces
    cerevisiae.  J Biol Chem 259, 5745-5751 (1984).
48333.  Vogel, JL, Parsell, DA, and Lindquest S.  Heat-shock
    proteins Hsp104 and Hsp70 reactivate mRNA splicing after heat
    inactivation. Curr Biol. 5, 306-317 (1995).
```

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.

The spreadsheet data file is just over 3 megabytes in size and the collection of all YPD Protein Reports is about 13 megabytes in size, although each are available in compressed format. A summary of the growth of YPD since its release in 1994 is shown in Table 1.

Table 1. YPD contents versus release number

Release	Date	Total	Known[a]	Homol[b]	Unknown[c]
1.2	Nov. 23, 1994	3020	1729	387	904
2.0	Dec. 8, 1994	3142	1750	450	942
3.0	Feb. 1, 1995	3512	1871	524	1117
4.0	Jun. 6, 1995	4046	1951	667	1428
4.1	Jul. 7, 1995	4305	2012	729	1564

[a]Proteins characterized through genetic or biochemical experiments.
[b]Proteins that have not been characterized but have sequence similarity to characterized proteins.
[c]Proteins 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

(mewes@mips.embnet.org). 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

1 Garrels, J.I., Futcher, B, Kobayashi, R., Latter, G.I., Schwender, B., Volpe, T., Warner, J.R., and McLaughlin, C.S. *Electrophoresis* **15**, 1466–1486 (1994).
2 Latter, G.I., Boutell, T., Monardo, P.J., Kobayashi, R., Futcher, B., McLaughlin, C.S., and Garrels, J.I. *Electrophoresis* **16**, 1170–1174 (1995).
3 Cherry, J.M. In Stewart, A. (ed.) *Trends in Genetics, Genetic Nomenclature Guide including Information on Genetic Databases*, p. 12 (1995).
4 Linder P. *et al. Nucleic Acids Res* **21**, 3001–3002 (1993).
5 Schuler, G.D., Epstein, J.A., Ohkawa, H., and Kans, J.A. In *Methods in Enzymology* (R. Doolittle, ed.), Academic Press, Inc. In press.