 Nucleic Acids Research, 1998, Vol. 26, No. 1 73–79

SGD: *Saccharomyces* Genome Database

J. Michael Cherry*, Caroline Adler, Catherine Ball, Stephen A. Chervitz, Selina S. Dwight, Erich T. Hester, Yankai Jia, Gail Juvik, TaiYun Roe, Mark Schroeder, Shuai Weng and David Botstein

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

Received September 15, 1997; Revised and Accepted October 3, 1997

ABSTRACT

The *Saccharomyces* Genome Database (SGD) provides Internet access to the complete *Saccharomyces cerevisiae* genomic sequence, its genes and their products, the phenotypes of its mutants, and the literature supporting these data. The amount of information and the number of features provided by SGD have increased greatly following the release of the *S.cerevisiae* genomic sequence, which is currently the only complete sequence of a eukaryotic genome. SGD aids researchers by providing not only basic information, but also tools such as sequence similarity searching that lead to detailed information about features of the genome and relationships between genes. SGD presents information using a variety of user-friendly, dynamically created graphical displays illustrating physical, genetic and sequence feature maps. SGD can be accessed via the World Wide Web at http://genome-www.stanford.edu/Saccharomyces/

INTRODUCTION

The *Saccharomyces* Genome Database (SGD) was established to provide a convenient means for accessing the rapidly expanding knowledge available for the budding yeast *Saccharomyces cerevisiae*. The completion of the *S.cerevisiae* genomic DNA sequence in 1996 (1) provided the sequence of each of its genes and currently represents the only complete sequence of a eukaryotic genome. Systematic efforts to identify *S.cerevisiae* genes, describe their role within the cell's life cycle, and reveal their interactions with other gene products are now underway. Such experimental approaches are changing how basic biological research is conducted and are resulting in an explosion of information.

The data in SGD are organized around the genome's sequence and its genes. SGD has as its primary goals the provision of information about the DNA sequence and its individual components, RNAs, encoded proteins and the structures and biological functions of any known gene products. An equally important goal of SGD has been to create tools which allow the user to easily retrieve and display these types of information. This has resulted in graphical interfaces which are geared towards biologists using the database, irrespective of their familiarity with yeast. By knowing a bit of sequence, a gene name, a function (e.g. enzymatic activity), or a map position, one can efficiently query for information about a gene. In addition, SGD serves as the *S.cerevisiae* community's repository for gene nomenclature. SGD in its present form is consulted ~45 000 times weekly via the World Wide Web (http://genome-www.stanford.edu/Saccharomyces/).

SEQUENCE INFORMATION

SGD is not a primary sequence database (2), but instead collects DNA and protein sequence information from primary providers (GenBank, EMBL, DDBJ, SwissProt and PIR). It then assembles it into datasets (described below) that make the sequence information more useful to molecular biologists. These datasets are available from SGD through the World Wide Web and Anonymous FTP.

The genomic sequence. In April 1996, the complete genomic sequence of the *S.cerevisiae* strain S288C, determined through a world-wide collaboration, was released to the public (1). In the time since this first release there have been many updates to this 'systematic sequence.' It is anticipated that these updates will continue for some time, albeit at a decreased rate. In addition, sequence annotations must be continually updated to reflect the current knowledge of the genome. The rate of accumulation of such annotations, particularly with regard to gene function, is increasing. The European Union funded database at MIPS (3,4) can also be consulted for information about the yeast systematic genomic sequence.

Much of the information about each ORF and/or gene in SGD is presented through the Web on a 'Locus page,' as the gene and ORF names are the organizing centers of the biological information in SGD. Information about a gene and/or ORF is presented either directly on the Locus page or by means of hyperlinks to other text pages, literature references or graphical displays. Conversely, the graphical displays and tables of results from sequence searches are hyperlinked via the gene and/or ORF names to the biological information contained on the locus page. The 'Locus,' 'Clone' and 'Sequence' pages provide a mini-map graphic allowing the user to get back to the maps at any time.

The latest version of the systematic sequence is available in chromosomal pieces via anonymous FTP from SGD (Table 1) and forms the underlying basis for SGD's organization and many of it's displays as found on the Web. Arbitrary segments of this sequence defined by the user can be retrieved from the Web using SGD DNA/Protein Sequence retrieval forms. In general, SGD creates its current version of systematic sequence by assembling the many entries submitted by the systematic sequencers. When a GenBank (5), EMBL (6) or DDBJ (7) sequence database entry is updated, SGD adopts the changes contained with the update

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

into its current representation of the systematic sequence. There are, however, occasions when the sequence database entries are not updated as quickly as the community requires. In such instances, the genomic sequence provided by SGD is modified to match the best information available from the community in advance of the sequence database update. SGD then works with the sequence databases and authors of the original sequence to facilitate an update so that the yeast community will have access to the best information, regardless of which database they choose.

Yeast GenBank. This is an archival dataset containing all sequences from GenBank/EMBL/DDBJ with the Organism labels '*Saccharomyces cerevisiae*', 'Mitochondrion *Saccharomyces cerevisiae*' or 'Killer virus of *S. cerevisiae*'. This dataset is updated three times a week. This dataset is both quite large and diverse. It includes the sequences submitted by individual laboratories over the past two decades of DNA sequencing. It therefore contains many mutant, authentically variant and erroneous sequences as well as the larger sequences submitted by systematic sequencing groups.

Yeast ORF. SGD provides two datasets that represent the predicted hypothetical Open Reading Frames (ORFs) contained within the genome. The 'systematic ORFs' are those initially defined by the sequencing groups and refined with time by the community. At present there are many ORFs that are questionable on the basis of one or more computational tests (e.g. the SAGE results; 8). However, each of these ORFs will remain in SGD until experimental evidence supports its removal. The first yeast ORF dataset SGD provides is a collection of DNA sequences for the ORFs ('orf_coding.fasta') including the coding sequences from the systematic sequence with their introns removed. The second, an amino acid dataset ('orf_trans.fasta'), contains the hypothetical translations of the ORF coding regions.

ORF Locations Table. Yeast ORF information is also summarized in a handy table available via SGD's FTP site (see Table 1). It lists the ORFs identified by the systematic sequencing effort. For each ORF, the table includes its systematic name, its SGDID number (an accession number for that ORF within SGD), its standard gene name, the chromosome upon which it resides, its base pair coordinates, the number of introns, the relative coordinates of its exons and a brief description of this product.

In many cases, researchers are less interested in ORFs than in the other genetic features of the genome. The non-ORF Features Table (see Table 1) lists genetic features such as centromeres, tRNAs, other RNA genes, Ty elements and LTRs. This table provides similar information as the ORF tables. Both tables are tab-delimited and can be combined or viewed with most spreadsheet software.

In addition to providing sequence information to users, SGD is responsible for updating all the non-EU systematic sequencing results. Cooperation with the yeast community and the sequence databases places SGD in an excellent position to update the sequence, as well as the biological annotations of these sequences. The sequences SGD updates include *Saccharomyces* genomic sequencing results of the Stanford DNA Sequencing & Technology Center, the Washington University Genomic Sequencing Center, the University of Montreal, RIKEN Institute, and the Sanger Centre. The updates therefore include all or part of chromosomes I (9), IV (10), V (11), VI (12), VIII (13), IX (14), XII (15), XIII (16) and XVI (17).

To enhance the utility of *S.cerevisiae* as a model organism for other systems, SGD is expanding into the area of comparative genomics. A first step in this direction is the addition of an area dedicated to the exploration of homology between yeast and mammalian protein sequences. This area, termed Sacch3D, offers BLAST reports between yeast and all mammalian sequences in GenBank as well as tables summarizing the results (18). This page also hyperlinks to XREFdb (19,20) and contains additional information about human–yeast functional homologs and yeast genes with similarity to human disease genes.

Table 1. Useful URLs for the *Saccharomyces* Genome Database

Description	World Wide Web URLs
SGD FTP	ftp://genome-ftp.stanford.edu
SGD Home Page	http://genome-www.stanford.edu/Saccharomyces/
SGD BLAST	http://genome-www2.stanford.edu/cgi-bin/SGD/nph-blast2sgd
SGD FASTA	http://genome-www2.stanford.edu/cgi-bin/SGD/nph-fastasgd
Pattern Matching	http://genome-www.stanford.edu/cgi-bin/SGD/PATMATCH/patmatch
Sequence Similarity View	http://genome-www.stanford.edu/Saccharomyces/SSV/viewer_start.html
Sequence Retrieval and Displays	http://genome-www.stanford.edu/cgi-bin/SGD/seqDisplay
Sacch3D	http://genome-www.stanford.edu/Sacch3D/
Yeast-Mammalian Comparison	http://genome-www.stanford.edu/Saccharomyces/mammal/
Genomic Sequence chromosome files	ftp://genome-ftp.stanford.edu/pub/yeast/genome_seq/
ORF Locations Table	ftp://genome-ftp.stanford.edu/pub/yeast/tables/ORF_Locations/
non-ORF Features Table	ftp://genome-ftp.stanford.edu/pub/yeast/tables/Other_Features_Locations/
Gene Naming Guidelines	http://genome-www.stanford.edu/Saccharomyces/gene_guidelines.html
SGD Gene Registry Request Form	http://genome-www.stanford.edu/Saccharomyces/gene_list.html
Global Gene Hunter	http://genome-www.stanford.edu/cgi-bin/SGD/geneform
Gene Registry Form	http://genome-www.stanford.edu/Saccharomyces/forms/gene-registry.html
Example of the Gene_Info for *AAP1*	http://genome-www.stanford.edu/cgi-bin/dbrun/SacchDB?find+Gene_Info+AAP1

Sequence similarity searches

SGD provides a variety of sequence similarity search services, including BLAST (21,22), FASTA (23), Pattern Matching, Sequence Similarity View and Stripe View. Pattern Matching, Sequence Similarity View and Stripe View are all programs created by SGD staff. Pattern Matching allows users to perform a variety of motif searches, using degenerate search sequences. Sequence Similarity View and Stripe View provide a visual display of sequence similarities within the yeast genome.

A variety of datasets are available for searching with BLAST, FASTA and Pattern Matching: the genomic sequence, all GenBank *S.cerevisiae* sequences, a non-redundant set of *S.cerevisiae* protein sequences combined from NCBI's GenPept (24), PIR (25) and SwissProt (26), the DNA coding sequence of all hypothetical ORFs identified by the systematic sequencing project, the hypothetical translation of all ORF sequences, and the non-ORF DNA sequences. In addition to the protein sequence datasets, SGD also includes the topic category Protein_Info which contains information curated by the YPD (27,28) resource.

Pattern Matching allows the user to query the nucleic acid or amino acid databases maintained at SGD for a regular expression that defines a motif or sequence pattern of interest. The results are displayed in a tabular format with hyperlinks that provide direct access to detailed biological information via the locus page.

Sequence Similarity View (SSV) (see Table 1) is a derivative of the dot-matrix analysis of DNA sequences (29). In developing this feature, the entire genomic sequence was compared to itself to create a similarity matrix. The user begins with a view comparing the sequence of any two chromosomes. Direct and indirect repeat regions between the two chromosomes are presented. The user can 'zoom' into a particular region by clicking on a section of the displayed image. Close-up views reveal the names and locations of the ORFs, which are hyperlinked to the locus pages containing detailed information about them. SSV displays sequence similarity over a large region between two chromosomes and graphically represents the organization of repetitive regions contained on both chromosomes.

The Stripe View displays the similarity matrices from Sequence Similarity View in a different way. When starting with a small region of the genome, the user can request the Stripe View presentation. This presentation illustrates the similarity of the selected small region along the various chromosomes, represented by 16 bars. The Stripe View allows the distribution of repetitive sequence elements to be displayed very effectively. It is also useful for visually portraying the distribution of duplicated regions within the genome. Hyperlinks to the locus page guarantee ready access of biological information about any of the ORFs or genes shown.

Viewing and retrieving sequence data

Sequences can be retrieved, analyzed and displayed in many ways using the feature 'Sequence Retrieval and Displays'. All the graphic presentations are designed to be clickable and thus direct the user to more detailed information. This program, also written by SGD staff, allows users to select a standard gene name, a systematically named ORF, a GenBank sequence or any region of a chromosome. After specifying the region of interest, all available options for that item are displayed. Options can include text information for genes or sequences. Sequences can be viewed as DNA sequences, restriction maps or amino acid translations. In addition, there are several tables and maps that display a sequence in the context of the whole genomic sequence. The restriction maps and sequence retrieval are provided in GCG format (30).

Chromosomal Features Table. The Chromosomal Features Table lists details of the genetic features of the genomic sequence. These features include ORFs, tRNAs, centromeres, RNA genes and Ty transposable elements and LTRs. For each feature, the table lists the systematic name of the element, the chromosome base pair coordinates, its coding strand (either Watson or Crick; see below), several sequence and information retrieval options, the standard gene name (if assigned), and a brief description (when available).

The data used for this table is compiled from several sources. The positions of ORFs and centromeres come from the systematic sequence annotations, while gene names and descriptions come from SGD's set of loci and sequences. The tRNAs were identified using tRNAscan-SE (version 1.0.2) software from Todd Lowe & Sean Eddy (31). Ty elements and LTRs are provided through an active collaboration with Dan Voytas and his lab at Iowa State University (32).

Genomic View. The Genomic View is a Java-based atlas providing graphical hyperlinks to various types of maps of the *S.cerevisiae* genome. Each of the 16 yeast chromosomes is represented graphically by horizontal bars containing several mapped genes as landmarks. Genomic View allows the user to view a selected chromosomal region in one of three different map displays. These displays include (i) Chromosomal Features Map, (ii) Physical Map, and (iii) Combined Genetic and Physical Map. The user then selects the region to be viewed by clicking on one of the chromosomal bars in the desired location. Since each map is generated on demand from current information contained within the database, standard gene names will replace uncharacterized ORF names as they are described. The chromosomal position for each map view is indicated above its respective graphic, and hyperlinks to the locus page are provided for each gene and/or ORF name.

Chromosomal Features Map. The Chromosomal Features map (Fig. 1) has a clickable interface which displays the position of a genetic element relative to its chromosomal neighbors. This representation of a user-selected region of a chromosome includes the same data used in the Chromosomal Features Table: characterized genes' open reading frames (ORFs), centromeres, Ty and LTR repetitive elements, transfer RNAs (tRNAs), ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs). Clicking on the text or graphic representation of a named ORF brings up its Locus page, while clicking on another type of sequence feature retrieves it's sequence information. The features to the left or right along the chromosome can be browsed either by using the 'Retrieve Features' buttons located beneath the map or by selecting a new region on the chromosomal representation at the top of the graphic. The resulting map orients viewers by overlapping with the previously-viewed map. Each feature represented in the map is hyperlinked to descriptive information contained in SGD that can be accessed by using the mouse to select an object or text on the map.

The Features Map for a 5-kilobase pair (kb) fragment containing a particular gene or ORF is also accessible from abbreviated 'minimaps' found on appropriate Locus, Sequence and Clone pages. The Features minimap shows both DNA strands of a 5 kb region surrounding a specific locus, ORF, sequence or clone. Genetic elements within a minimap can be selected,

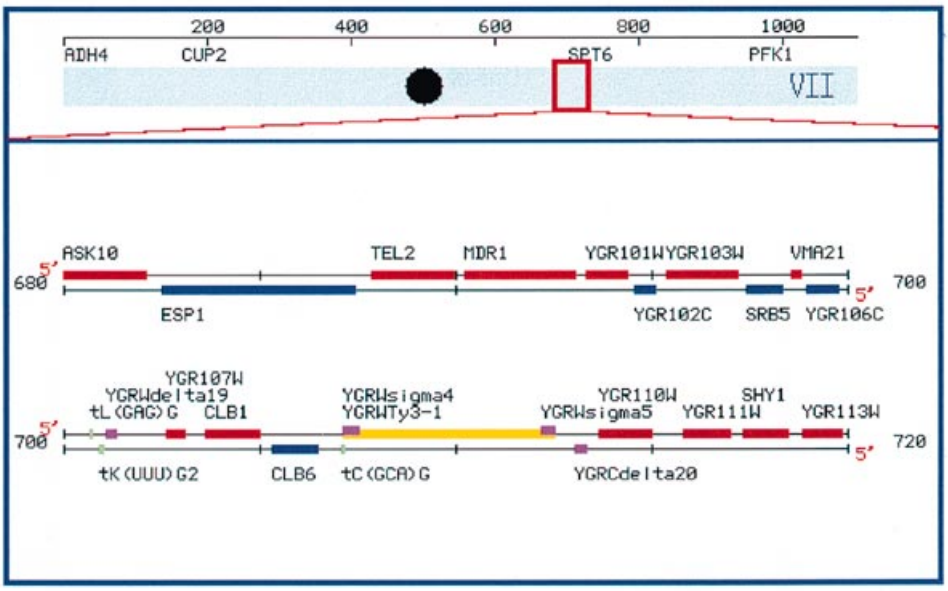

Figure 1. Clickable Features Map. This map shows the region of Chromosome VII, from 680 000 to 720 000 bp. This map, generated on demand, shows the location of physical map elements within this region of Chromosome VII. The map elements are color coded by type: dark green, ATCC clones; red, ORF encoded on the Watson strand; blue, ORFs encoded on the Crick strand; lime, tRNAs; grey, other RNA genes; yellow, transposons; burgundy, snRNAs; pink, LTRs.

enabling users to see an expanded view of the Chromosomal Features Map, from which hyperlinks to the biological information are provided. These minimaps can be incorporated into Web pages at other sites by using a simple URL to the minimap GIF producing cgi software.

Chromosomal Physical Map. The Chromosomal Physical Map illustrates the position of clones available from the American Type Culture Collection (ATCC, 33) as well as ORF locations relative to the chromosomal features for every region of the entire *S.cerevisiae* genome. From the Genomic View, the map displays these features for user-specified coordinates in 100 000 base pair (bp) increments. The entire chromosome, represented as an orange bar above the Physical Map, shows the region currently displayed in the graphic. The Physical Map display can be changed to view different regions of the chromosome or genome either by retrieving the features to the left or right of the displayed area, by using the mouse to select another area within the chromosome, or by specifying a locus or ORF name elsewhere in the genome. In the Physical map, Zoom features enable users to magnify the center of the currently displayed physical map. The chromosomal features, depicted as objects or text at the bottom of the Physical map are also hyperlinks to information contained within SGD. This information is retrieved simply by clicking on an item displayed in the graphic.

Combined Physical and Genetic Map. The Combined Physical and Genetic Map (Fig. 2) presents both genetic mapping data and data from the systematic genomic sequencing project as a parallel comparison of the genetic and physical maps for each chromosome (34–37). The entire chromosome (in blue) is represented on the left next to the scales for the genetic and physical map distances. The Physical map (in kb) is shown to the right of the chromosomal representation. ORFs are depicted as colored boxes. ORFs shown in red are on the Watson strand. Those in blue are on the Crick strand. Where biological function for an ORF has been defined, a gene name is given. An ORF displayed on the right side of the physical map indicates a locus that has been mapped both physically and genetically, an ORF on the left has been only physically mapped. The location of genes that have been both physically and genetically mapped are linked by a line connecting the ORF location on the physical map to a tick mark denoting its position on the genetic map. The map location can be adjusted by choosing an area along the chromosomal representation with the mouse, or by using the Retrieve Map options below the graphic. As in the Physical and Features maps, hyperlinks have been provided from the text and objects on the maps to additional information in the database.

GUIDE TO STRUCTURAL INFORMATION

Sacch3D is a facility offered by SGD for collecting and describing currently available structural information for *S.cerevisiae* proteins. Knowledge of a protein's structure can provide insights into its function and can be a useful aid in designing experiments. Sacch3D presents a 'yeast-centric' view of other structural databases as well as a structure-centric view of the yeast genome in a non-intimidating manner. It is intended to add value to SGD and to provide researchers both within and outside of the yeast research community insight into yeast proteins.

Structural information for Sacch3D is obtained by analysis of the Brookhaven PDB database (38,39) of crystallographic coordinates to identify all yeast structures currently deposited as well as other PDB structures with significant sequence (and likely structural) similarity to yeast proteins. The analysis covers all experimentally identified *S.cerevisiae* proteins as well as those predicted based on the systematic sequencing of the yeast genome.

Due to the relatively small number of solved 3D protein structures and the inherent difficulties in predicting structure from sequence using current techniques, any given set of criteria will have tradeoffs. For example, the BLASTP analysis used for

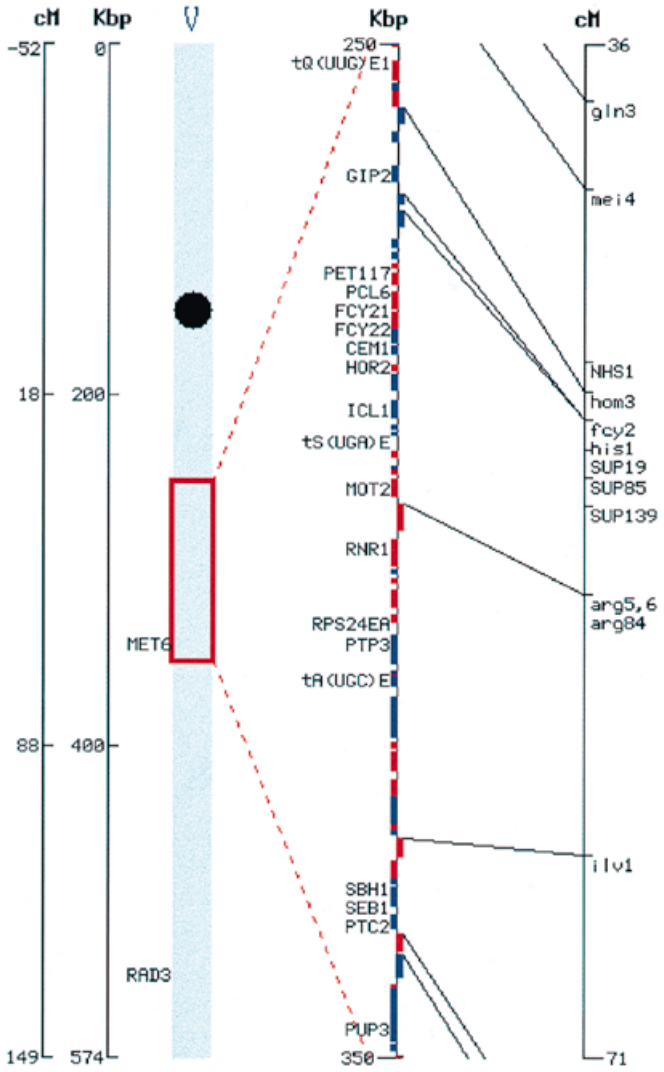

Figure 2. Clickable Combined Physical and Genetic Map. This map shows the region of chromosome V, from 250 000 to 350 000 bp. The left hand side of the map represents the entire chromosome, with vertical bars indicating genetic distance from the centromere in centimorgans (cM) and the physical length of the chromosome (in kb). The blue bar in the middle of the figure represents chromosome V, with the centromere position indicated by a black circle. The red square indicates the region of the chromosome for which the combined Physical and Genetic Map is displayed to the right. In the physical map, distance is measured in kb and ORFs are indicated by red (encoded on the Watson strand) or blue (encoded on the Crick strand) boxes. When available, the standard gene name is indicated. The genetic map is measured in centimorgans. The positions of genetically mapped genes are indicated by tick-mark next to the standard gene name. Demonstrated correlations between the genetic and physical maps are indicated by the black lines that extend from a location on the physical map to location on the genetic map.

Sacch3D is quick and detects regions of significant sequence similarity but will lead to false negatives in cases where structural similarity exists in the absence of significant sequence similarity. Additional criteria are being explored for potential future software to detect such cases.

The Sacch3D search utility provides a 'structural information page' for any given yeast gene or ORF name. This page summarizes the PDB homologs for the yeast protein, listing the name and source organism for the homolog and the BLAST p-value when compared against the yeast protein. For each homolog, hyperlinks to interactive 3D viewers [RasMol (40), Java viewer and Cn3D (41,42)] as well as external structural databases [PDB, MMDB (43), SCOP (44,45), CATH (46), Swiss-Model (47,48), PDBsum (49)] are offered for learning more about any particular protein structure listed. The Java viewer (WebMol II) co-developed by SGD staff (50), for example, provides an interactive 3D structure that is colorized based on sequence similarity to the yeast protein. The PDB homologs themselves are listed from best to worst and are clustered in order to reduce the redundancy present in the PDB. Thus, one representative structure is listed in cases where there are multiple variants of the same structure (mutants or complex forms). Only 14% of yeast proteins currently have homologs in the PDB using BLASTP with a p-value cutoff of 1×10^{-4}.

From the structural information page, one may also obtain a BLASTP report between the yeast protein and the PDB as well as a variety of other datasets (GenBank, ESTs, mammalian). A few hyperlinks are also provided for predicting secondary and tertiary structure for the yeast protein (although structure prediction is not the focus of Sacch3D). Hyperlinks to SGD, SwissProt and NCBI's Entrez (51) are also provided for the yeast protein.

To begin the process of collecting and analyzing domains in yeast proteins, Sacch3D has collected data for structurally defined domains in yeast using data contained within the SCOP database. The Sacch3D 'Domains' page provides access to all yeast proteins that can be classified according to the SCOP domain classification system. This data may be searched using a yeast gene/ORF name, a SCOP class number or a SCOP fold number. A fold number search, for example, will retrieve a table of all yeast proteins that either contain or show significant sequence similarity to proteins containing that fold. Hyperlinks are also provided to the PDB as well as to the WebMol II Java viewer to illustrate the location of the fold within the context of the 3D structure.

GUIDE TO YEAST LITERATURE

SGD houses over 16 000 journal citations pertinent to *S.cerevisiae* loci, including their abstracts (when available). The section of the database referred to as Gene_Info was created to guide users through this large amount of literature so that they may efficiently locate publications relevant to a gene of interest. In addition, Gene_Info can potentially serve as a tool for introducing the researcher to new information about a gene and its relationship to other *S.cerevisiae* genes. It may also be useful as a resource for directing researchers to information about various constructs, techniques, and mutant strains specific to a locus of interest. Gene_Info is prominently featured on the Locus page for all standard gene names in SGD.

Gene_Info categorizes published journal citations into various scientific topics based on the information within abstracts. Gene_Info is curated by first searching the NCBI literature resource PubMed (52) for all papers which contain the gene name, its systematic ORF name, any of its secondary names, the gene product name associated with it and, in addition to any one of these, the term *S.cerevisiae*. The abstracts of the resulting papers are then read by scientific curators at SGD, and each abstract is categorized via its citation into one or more appropriate

topics. There are currently 28 topics which have been carefully selected to represent major areas of biological information in a comprehensive manner. Any papers which are found to be irrelevant to the *S.cerevisiae* gene are not categorized, and thus do not appear under any of the topic headings. In this manner, Gene_Info presents a thorough set of references, categorized under appropriate biological topic(s), that are pertinent to the *S.cerevisiae* gene of interest. To view the unedited PubMed search results, there is a hyperlink entitled 'Search PubMed for this Gene Name' provided at the bottom of the Gene_Info pages which performs the PubMed search result described above in 'real time.'

The primary feature of Gene_Info is a list of standard topic names which represent the categorization of citations according to biological topics, as described above. As an example, the Gene_Info page for the standard locus *AAP1* (Table 1) contains the following topic headings: Alias, Protein-protein Interactions, Mutants/Phenotypes, Cellular Location, Non-Yeast Related Proteins, Constructs, Techniques and Reagents, DNA/RNA Sequence Features, Protein Sequence Features, Function, RNA Levels and Processing, Protein Physical Properties, Other features, Regulated by, and Transcription. Note that only those topics for which there is relevant published information are listed for an individual gene. Each of the above topic headings is hyperlinked to a list of citations whose abstracts contain some piece of information relevant to the topic. The abstract for a citation may be obtained by clicking on the citation. A hyperlink entitled 'Curated PubMed References' found at the bottom of the Gene_Info page is the sum of all the citations found in the various topics listed for a given gene.

One goal of Gene_Info is to guide the user to desired information. For instance, to find information about the localization of Aap1p, a user can go to the Locus page for *AAP1*, follow the Gene_Info hyperlink for *AAP1*, and select the topic 'AAP1 Cellular Location.' A list of citations whose abstracts mention Aap1p's location within the cell will be presented, and each abstract may be obtained by clicking on its corresponding citation. In this manner, the user is efficiently guided towards a piece of information without having to read through many different abstracts. The Gene_Info feature is particularly useful to researchers outside the yeast community that study genes with homology to characterized yeast genes. The function of the *S.cerevisiae* homolog may provide clues to the function of orthologous or paralogous genes.

Gene_Info not only directs users to general information about a gene, but also alerts users that specific information about a gene or its product may exist in the published literature. For instance, seeing the Gene_Info topic heading 'Protein Processing/Regulation' indicates that a gene is post-translationally regulated. Thus, the mere presence of a topic provides information.

Another goal of Gene_Info is to uncover biological relationships between different gene products. The above examples describe how to obtain basic information directly related to a locus. Taking this one step further, the Gene_Info topics 'Genetic Interactions,' 'Protein-protein Interactions,' 'Yeast Related Proteins,' 'Non Yeast Related Proteins' and 'Disease Gene Related' help the user explore relationships between genes and gene products. This information is particularly useful to users unfamiliar with a given yeast gene. Gene_Info provides further information about relationships between genes via a hyperlink entitled 'Genes Mentioned in Curated Papers with this Gene' found at the bottom of Gene_Info pages with curated references. This leads to a list of loci mentioned in at least one abstract determined to be pertinent to the original locus of interest.

Lastly, Gene_Info provides a resource for locating techniques pertinent to the study of a given gene (i.e. protein purification, protein localization) and for finding out where to locate constructs, mutant strains and reagents. Citations for abstracts containing such information can be found under the topics 'Techniques and Reagents,' 'Mutants/Phenotypes' and 'Strains/Constructs.'

GENE NOMENCLATURE

SGD was chosen by the yeast community to maintain a consistent nomenclature for *S.cerevisiae*, one of it primary goals. SGD maintains a complete list of all *S.cerevisiae* gene names, called the Gene Name Registry. This list includes all the 'standard' locus names in the database, the 'non-standard' locus names (i.e., aliases), and the 'reserved' locus names (see 'Reserving a gene name with SGD,' below). The SGD curators search NCBI's PubMed database as well as the sequence databases GenBank/EMBL/DDBJ to verify that gene names are unique and to identify new gene names. If a gene name is not unique, the SGD curators work to resolve the nomenclature conflict.

SGD and the community of yeast researchers have together established Gene Naming Guidelines that describe the policies that SGD follows in accepting gene name reservations and converting a reserved name into a standard or non-standard name upon publication of that name. The full text of the Gene Naming Guidelines is available at our web site (see Table 1).

Yeast gene names

There are two types of names that exist for a yeast locus, a Systematic ORF Name and a Standard Gene Name.

Systematic ORF Name. In accordance with policies adopted by the yeast systematic sequencing groups, all *S.cerevisiae* ORFs are designated by a symbol consisting of three uppercase letters followed by a number and then another letter, as follows: Y (for 'Yeast'); A – P for the chromosome upon which the ORF resides (where 'A' is chromosome I, up to 'P' for chromosome XVI); L or R (for Left or Right arm of the chromosome); a 3-digit number corresponding to the sequential order of the open reading frame on the chromosome arm (starting from the centromere and counting out to the telomere); and W or C for whether the open reading frame is encoded on the 'Watson' or 'Crick' strand (where 'Watson' runs 5′ to 3′ from left telomere to right telomere). The Watson strand is the strand deposited within the sequence databases by the systematic sequencing groups. For example, 'YFL039C', is the 39th ORF to the left of the centromere on chromosome VI and is encoded on the Crick strand. Most ORF designations by the systematic sequencing groups use a predicted 100 amino acid polypeptide as the minimum size limit, except when a smaller gene has already been characterized and localized to the chromosomal sequence, or when very strong similarity is observed to an other gene from any organism. When there is evidence suggesting that a new ORF should be added to the systematic sequence, it is named by taking the name of the centromere-proximal adjacent ORF and adding a hyphenated 'A' or 'B' to the end of it (this avoids re-numbering all the distal ORFs).

Standard Gene Name. A standard gene name is a unique gene name that is published in a scientific journal and is the primary accepted name for that gene. Other published names for that gene

are 'Not Standard' names, or aliases. The Gene Naming Guidelines discuss at greater length the policy that SGD follows in designating a gene name the Standard Gene Name. A Standard Gene Name should consist of three letters (the gene symbol) followed by a number (e.g. *ADE12*). Dominant alleles of the gene (most often wild-type) are denoted by all uppercase letters, while recessive alleles are denoted by all lowercase letters. The 3-letter gene symbol should stand for a description of a phenotype, gene product or gene function. In addition, a given gene symbol should have only one associated description, i.e., all genes that use a given 3-letter symbol should have a related phenotype, gene product or gene function.

SGD has created tools to help researchers choose an appropriate name that does not conflict with existing names. The researchers should consult both the SGD gene registry and Global Gene Hunter (see Table 1), which can be used to search both SGD and several other public databases (see Table 1). In cooperation with the yeast research community, SGD has developed a system for reserving gene names. Reserved names are made publicly available to discourage researchers from using a single name to describe multiple genes or creating several names to describe the same gene. To allow users to reserve gene names, register gene names that are in use and submit relevant genetic or biological information, SGD has created a Gene Registry Form.

ACKNOWLEDGEMENTS

We would like to thank Julie Sneddon, Joy Chen and Maud Morshedi for assistance in preparing information, Web pages, and software for the database. SGD is supported by a P41, national resources, grant from the National Human Genome Research Institute at the US National Institutes of Health.

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) *Science*, **274**, 546.
2 International Nucleotide Sequence Database Collaboration: http://www.ncbi.nlm.nih.gov/collab/
3 MIPS: http://www.mips.biochem.mpg.de/mips/yeast/
4 Mewes,H.W., Albermann,K., Heumann,K., Liebl,S. and Pfeiffer,F. (1997) *Nucleic Acids Res.* **25**, 28–30 [see also this issue (1998) *Nucleic Acids Res.* **26**, 33–37].
5 GenBank: http://www.ncbi.nlm.nih.gov/Web/Genbank/
6 EMBL: http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html
7 DDBJ: http://www.ddbj.ac.jp/
8 Velculescu,V.E., Zhang,L., Zhou,W., Vogelstein,J., Basrai,M.A., Bassett,D.E., Jr., Hieter,P., Vogelstein,B. and Kinzler,K.W. (1997) *Cell*, **88**, 243–251.
9 Bussey,H., Kaback,D.B., Zhong,W., Vo,D.T., Clark,M.W., Fortin,N., Hall,J., Ouellette,B.F., Keng,T., Barton,A.B., *et al.* (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 3809–3813.
10 Jacq,C., Alt-Morbe,J., Andre,B., Arnold,W., Bahr,A., Ballesta,J.P., Bargues,M., Baron,L., Becker,A., Biteau,N., *et al.* (1997) *Nature*, **387**, 75–78.
11 Dietrich,F.S., Mulligan,J., Hennessy,K., Yelton,M.A., Allen,E., Araujo,R., Aviles,E., Berno,A., Brennan,T., Carpenter,J., *et al.* (1997) *Nature*, **387**, 78–81.
12 Murakami,Y., Naitou,M., Hagiwara,H., Shibata,T., Ozawa,M., Sasanuma,S., Sasanuma,M., Tsuchiya,Y., Soeda,E., Yokoyama,K., *et al.* (1995) *Nature Genet.*, **10**, 261–268.
13 Johnston,M., Andrews,S., Brinkman,R., Cooper,J., Ding,H., Dover,J., Du,Z., Favello,A., Fulton,L., Gattung,S., *et al.* (1994) *Science*, **265**, 2077–2082.
14 Churcher,C., Bowman,S., Badcock,K., Bankier,A., Brown,D., Chillingworth,T., Connor,R., Devlin,K., Gentles,S., Hamlin,N., *et al.* (1997) *Nature*, **387**, 84–87.
15 Johnston,M., Hillier,L., Riles,L., Albermann,K., Andre,B., Ansorge,W., Benes,V., Bruckner,M., Delius,H., Dubois,E., *et al.* (1997) *Nature*, **387**, 87–90.
16 Bowman,S., Churcher,C., Badcock,K., Brown,D., Chillingworth,T., Connor,R., Dedman,K., Devlin,K., Gentles,S., Hamlin,N., *et al.* (1997) *Nature*, **387**, 90–93.
17 Bussey,H., Storms,R.K., Ahmed,A., Albermann,K., Allen,E., Ansorge,W., Araujo,R., Aparicio,A., Barrell,B., Badcock,K., *et al.* (1997) *Nature*, **387**, 103–105.
18 Botstein,D., Chervitz,S.A. and Cherry,J.M. (1997) *Science*, **277**, 1259.
19 Bassett,D.E., Jr., Boguski,M.S. and Hieter,P. (1996) *Nature*, **379**, 589–590.
20 XREFdb: http://www.ncbi.nlm.nih.gov/XREFdb/
21 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
22 BLAST2: ftp://blast.wustl.edu/blast2/
23 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
24 GenPept: ftp://ncbi.nlm.nih.gov/genbank/genpept.fsa.Z
25 George,D.G., Dodson,R.J., Garavelli,J.S., Haft,D.H., Hunt,L.T., Marzec,C.R., Orcutt,B.C., Sidman,K.E., Srinivasarao,G.Y., Yeh,L-S.L., *et al.* (1997) *Nucleic Acids Res.* **25**, 24–27 [see also this issue (1998) *Nucleic Acids Res.* **26**, 27–32].
26 Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res.* **25**, 31–36 [see also this issue (1998) *Nucleic Acids Res.* **26**, 38–42].
27 YPD: http://www.proteome.com/YPDhome.html
28 Payne,W.E. and Garrels,J.I. (1997) *Nucleic Acids Res.* **25**, 57–62 [see also this issue (1998) *Nucleic Acids Res.* **26**, 68–72].
29 Maizel,J.V. and Lenk,R.P. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7665–7669.
30 Wisconsin Package Version 9.0, Genetics Computer Group (GCG), Madison, Wisconsin, USA
31 tRNAscan-SE: http://genome.wustl.edu/eddy/tRNAscan-SE/
32 Voytas Lab Ty information: http://www.public.iastate.edu/~voytas/ltrstuff/ltrtables/yeast.html
33 American Type Culture Collection, ATCC: http://www.atcc.org/
34 Mortimer,R.K. and Schild,D. (1980) *Microbiol. Rev.*, **44**, 519–571.
35 Mortimer, R.K., Schild,D., Contopoulou,C.R. and Kans,J. (1989) *Yeast*, **5**, 321–404.
36 Mortimer, R.K., Contopoulou,C.R. and King,J.S. (1992) *Yeast*, **8**, 817–902.
37 Cherry,J.M., Ball,C., Weng,S., Juvik,G., Schmidt,R., Adler,C., Dunn,B., Dwight,S., Riles,L., Mortimer,R.K. and Botstein,D. (1997) *Nature*, **387**, 67–73.
38 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.
39 Protein Database (PDB), http://www.pdb.bnl.gov/
40 RasMol: http://www.umass.edu/microbio/rasmol/
41 Hogue,C.W.V. (1997) *Trends Biochem. Sci.*, **22**, 314–316.
42 Cn3D: http://www.ncbi.nlm.nih.gov/Structure/cn3d.html
43 MMDB: http://www.ncbi.nlm.nih.gov/Structure/
44 Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
45 Structural Classification Of Proteins (SCOP): http://www.pdb.bnl.gov/scop/
46 CATH: http://pdb.pdb.bnl.gov/bsm/cath/
47 Peitsch,M.C. (1996) *Biochem. Soc. Trans.*, **24**, 274–279.
48 Swiss-Model. http://www.expasy.ch/swissmod/SWISS-MODEL.html
49 PDBsum: http://www.biochem.ucl.ac.uk/bsm/pdbsum/
50 WebMol II: http://genome-www.stanford.edu/~sac/java/pdb/
51 NCBI Entrez: http://www3.ncbi.nlm.nih.gov/Entrez/
52 PubMed: http://www.ncbi.nlm.nih.gov/PubMed/