D374–D377 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki023

Fungal BLAST and Model Organism BLASTP Best Hits: new comparison resources at the *Saccharomyces* Genome Database (SGD)

Rama Balakrishnan, Karen R. Christie, Maria C. Costanzo, Kara Dolinski[1], Selina S. Dwight, Stacia R. Engel, Dianna G. Fisk, Jodi E. Hirschman, Eurie L. Hong, Robert Nash, Rose Oughtred[1], Marek Skrzypek, Chandra L. Theesfeld, Gail Binkley, Qing Dong, Christopher Lane, Anand Sethuraman, Shuai Weng, David Botstein[1] and J. Michael Cherry*

Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305-5120, USA and [1]Lewis-Sigler Institute for Integrative Genomics, Carl Icahn Laboratory, Princeton University, Washington Road, Princeton, NJ 08544, USA

Received September 15, 2004; Revised and Accepted September 21, 2004

ABSTRACT

The *Saccharomyces* Genome Database (SGD; http://www.yeastgenome.org/) is a scientific database of gene, protein and genomic information for the yeast *Saccharomyces cerevisiae*. SGD has recently developed two new resources that facilitate nucleotide and protein sequence comparisons between *S.cerevisiae* and other organisms. The Fungal BLAST tool provides directed searches against all fungal nucleotide and protein sequences available from GenBank, divided into categories according to organism, status of completeness and annotation, and source. The Model Organism BLASTP Best Hits resource displays, for each *S.cerevisiae* protein, the single most similar protein from several model organisms and presents links to the database pages of those proteins, facilitating access to curated information about potential orthologs of yeast proteins.

INTRODUCTION

The *Saccharomyces* Genome Database (SGD) collects and organizes biological information about genes and proteins of *Saccharomyces cerevisiae*, and presents this information on individual Locus Pages for each yeast gene (1,2). In addition to assembling a detailed library of information about *S.cerevisiae*, we continually strive to develop tools and resources that allow users to identify connections between *S.cerevisiae* genes and proteins and those from different species. These connections may help researchers studying other organisms to glean knowledge from more extensively studied *S.cerevisiae* genes, or may enhance the study of *S.cerevisiae* genes with data from other organisms. These resources include the Gene Ontology (3) as well as comparison resources such as PSI-BLAST analyses and the Synteny and Fungal Alignment Viewers (1).

The Fungal BLAST and Model Organism BLASTP Best Hits tools, described here, are two new SGD resources that extend the users' reach beyond *S.cerevisiae* by allowing a variety of sequence comparisons. The Fungal BLAST tool may be used for comparison of any sequence of choice with a wide range of fungal nucleotide or protein sequences, while the Model Organism BLASTP Best Hits resource specifically makes connections between each *S.cerevisiae* protein and its best hit in protein sets from several other model organisms.

FUNGAL BLAST

The numerous publicly available fungal sequences in GenBank provide a rich source of information for the identification of conserved, functionally important coding and non-coding sequences. For example, recent large-scale comparisons among related fungal species have sparked new insights into the evolution of chromosome structure and regulatory sequences (4,5). This information also initiated revisions of the *S.cerevisiae* genome sequence (4–8). While bioinformatics approaches to sequence comparisons have been invaluable for gaining a broad understanding of genomes, single gene comparisons across species are often useful to researchers focused on particular areas of biology. The Fungal BLAST tool is designed to put these sorts of comparisons into

*To whom correspondence should be addressed. Tel: +1 650 723 7541; Fax: +1 650 725 1534; Email: cherry@genome.stanford.edu

Model Organism BLASTP Best Hits Summary

Species	Database	Hit	Description	E–value	% Aligned	Source Range	Target Range
A. gossypii	AGD	AGR101C	Syntenic homologue of S. cerevisiae YER179W (DMC1) ; 1-intron	1.0e-163	98.5	6..334	4..333
H. sapiens	ENSEMBL (HUMAN)	ENSP00000216024	Database:core Gene:ENSG00000100206 Clone:AL022320 Contig:AL022320.23.1.85500 Chr:22 Basepair:37157993 Status:known	1.0e-95	94.6	19..334	24..340
A. thaliana	TAIR	At3g22880.1	meiotic recombination protein, putative similar to Swiss-Prot:Q14565 meiotic recombination protein DMC1/LIM15 homolog [Homo sapiens]; contains non-consensus AT/AC non-consensus splice sites at intron 14	1.0e-90	94.6	19..334	31..344
D. melanogaster	FlyBase	CG7948-PA	type=protein; loc=3R:complement(25869957..25870886, 25870950..25871027); ID=Rad51-PA; name=Rad51-PA; db_xref='CG7948,FlyBase:FBgn0011700'; len=336	5.0e-77	95.2	17..334	20..335
C. elegans	WormBase	Y43C5A.6a	locus:rad-51 Helix-hairpin-helix motif. status:Confirmed TR:Q95Q25 protein_id:CAB61038.2	1.0e-76	96.1	14..334	74..394
S. cerevisiae	SGD	YER095W	RAD51 SGDID:S0000897, Chr V from 349976-351178, Verified ORF	3.0e-73	96.7	9..331	73..394

Figure 1. Summary table of the Model Organism BLASTP Best Hits page. A summary table similar to this representative table is generated for each locus having a 'hit' in one or more model organism databases. In this figure, *Saccharomyces* protein Yer179wp results are shown as an example. Columns of the table are as follows: species of the hit protein; name of the database for the hit protein, hyperlinked to the home page of that database; name of the hit protein from its database, hyperlinked to the database page of that protein or its gene; description of the hit protein, as found in its database; *E*-value (expectation value), reflecting the number of hits expected to be found by chance; percent aligned, showing the percentage of the length of the query protein over which it aligns with the hit protein; source range, showing the amino acid coordinates of the region of the *S.cerevisiae* query protein that was aligned; and target range, showing the amino acid coordinates of the region of the 'hit' protein that was aligned with the *S.cerevisiae* query protein.

the hands of researchers who concentrate on single loci or gene families.

The Fungal BLAST tool uses the WU-BLAST software (9) to compare any query nucleotide or protein sequence to fungal sequence datasets at GenBank. These include genome sequences from multiple *Saccharomyces* species (including *S.cerevisiae*, *S.bayanus*, *S.castellii*, *S.kluyveri*, *S.kudriavzevii*, *S.mikatae*, *S.paradoxus*) as well as sequences from genome projects, ESTs and other available sequences from all phyla in the kingdom Fungi. The sequences are updated periodically, and new fungal sequence datasets are added. The current list of species whose genomic sequences are included in Fungal BLAST analysis at SGD is provided on the Fungal BLAST help page (http://www.yeastgenome.org/help/fungal-blast.html). Although these sequences are available for searching at NCBI, the Fungal BLAST tool facilitates faster, directed searching by dividing the sequences into searchable subcategories according to organism, the status of genome sequencing and annotation (Complete Genomes, Annotated Genomes, Assembled Genomes, etc.), and type of sequence (e.g. Mitochondrial, EST, etc.).

In order to accommodate different types of query and target sequence, the Fungal BLAST tool offers four BLAST programs:

(i) BLASTN compares a nucleotide query sequence against a nucleotide sequence dataset;
(ii) TBLASTN compares a protein query sequence against a nucleotide sequence dataset dynamically translated in all six reading frames (both strands);
(iii) BLASTP compares an amino acid query sequence against a protein sequence dataset;
(iv) BLASTX compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence dataset.

The Fungal BLAST search is accessible from the 'Comparison Resources' pull-down menu on each Locus Page of all *S.cerevisiae* genes and structural features such as ARS and CEN elements as 'BLASTN vs fungi' or 'TBLASTN vs fungi'. When the form is accessed from a Locus Page, the query sequence box contains the nucleotide or protein sequence of that locus. In addition, links to the interface from

both the Analysis & Tools and the Homology & Comparisons contents pages allow input of any sequence of interest, either by simply pasting a text sequence into the dialog box, or by uploading a sequence file. The interface accepts sequences in FASTA, GCG or raw text formats.

The ability to analyze this broad spectrum of fungal sequence data provides a powerful means to identify sequences conserved through evolution and presumably important for the biology of the organisms. Limiting the genomes to those from fungi allows researchers to identify signatures in fungal-specific genes and gene products that might not be discernible by comparing sequences across kingdoms.

MODEL ORGANISM BLASTP BEST HITS

Comparison of sequences across diverse taxa is a powerful technique for finding universally conserved domains. If a curated database page exists for a protein of another organism to which a *S.cerevisiae* protein has similarity, the significance of the sequence conservation may become clear. To help our users find curated information concerning protein sequences conserved between *S.cerevisiae* and other organisms, SGD has developed the Model Organism BLASTP Best Hits page.

The Model Organism BLASTP Best Hits page (Figure 1) displays the results of NCBI BLASTP analyses, with the default parameters, using each *S.cerevisiae* protein sequence as the query against the complete set of predicted protein sequences from several model organisms. The single best BLASTP hit with an *E*-value of $\leqslant$0.01 is shown for each organism (more than one hit may be shown if the top hits have identical *E*-values). Protein datasets used for comparison are limited to completely sequenced and annotated genomes where curated database web pages are available for the individual proteins. As of September 2004, BLASTP analyses had been run against predicted protein sequences from six model organisms (Table 1). *S.cerevisiae* is one of the model organisms used for comparison, but in this case the target sequence identical to the query sequence is excluded from the Best Hits display, and the next best hit is shown.

Out of the 6591 protein coding *S.cerevisiae* genes, 5368 have a hit in at least one other model organism database (MOD) while 2387 ORFs have a hit in all 5 MODs (excluding SGD). Of the 78 dubious ORFs (considered unlikely to encode a protein), five had hits in the *Ashbya gossypii* dataset and the rest had hits only within the *S.cerevisiae* protein dataset (Table 2).

The Model Organism BLASTP Best Hits results page (Figure 1) shows the best hits in the other organisms for the *S.cerevisiae* query protein along with details about the alignments and links to the relevant database pages for the proteins from other organisms. The BLASTP analyses are run periodically and the model organisms included in these analyses will be updated as new datasets are available from other model organism databases. The Model Organism BLASTP Best Hits page can be accessed from the 'Comparison Resources' pull-down menu on the right-hand side of the Locus Page and from a link on the Homology & Comparisons contents page. A file containing all of the Best Hits data is available for download from our FTP site (ftp://ftp.yeastgenome.org/yeast/).

Table 1. Summary of best BLASTP hits for *S.cerevisiae* proteins in selected model organism databases

Organism (database name)	Total predicted proteins	Predicted proteins similar to *S.cerevisiae* query proteins	*S.cerevisiae* proteins with a hit in the target organism
Drosophila melanogaster (FlyBase)	18 746	2949 (15.73%)	2929 (44.44%)
Ashbya gossypii (AGD)	4726	4231 (89.53%)	4980 (75.56%)
Caenorhabditis elegans (WormBase)	22 254	2176 (9.77%)	2834 (43.00%)
Arabidopsis thaliana (TAIR)	29 161	2718 (9.32%)	3109 (47.17%)
Homo sapiens (ENSEMBL)	29 802	2730 (9.16%)	3137 (47.60%)
Saccharomyces cerevisiae (SGD)	6703	2246 (33.51%)	2984 (45.27%)

For each model organism, the table displays: total number of predicted proteins, as of September 2004; predicted proteins that are 'hit' (*E*-value $\leqslant$ 0.01) by an *S.cerevisiae* query protein, expressed as the number of proteins and as the percentage of total proteins for that organism; and *S.cerevisiae* proteins that find a hit in the predicted proteins of that model organism, expressed as the number of proteins and as the percentage of total *S.cerevisiae* open reading frames. The *S.cerevisiae* set of predicted protein sequences comprised 6591 open reading frames predicted as of September 2004. For comparisons to *S.cerevisiae*, the best hit is defined as the most similar protein not identical to the query sequence.

Table 2. BLASTP best hits for *S.cerevisiae* proteins, sorted by ORF classification

ORF classification	Total number in the *S.cerevisiae* genome	*S.cerevisiae* ORFs with a hit in one or more of the MODs (including SGD)	*S.cerevisiae* ORFs with a hit in all 5 MODS (excluding SGD)
Verified	4231	4069 (96.17%)	2105 (49.75%)
Uncharacterized	1546	1221 (78.98%)	282 (18.24%)
Dubious	814	78 (9.58%)	0
Total	6591	5368 (81.44%)	2387 (36.22%)

ORFs are classified at SGD as verified, uncharacterized or dubious, depending on the likelihood that they are expressed as protein products. For each ORF class and for the whole ORF set, the table displays: the total number of ORFs in that set; the ORFs that find a hit in one or more of the model organism protein sequence datasets including *S.cerevisiae* proteins, expressed as the number of ORFs and as the percentage of total ORFs in that set; and the ORFs that find a hit in all of the model organism protein datasets excluding *S.cerevisiae* proteins, expressed as the number of ORFs and as the percentage of total ORFs in that set.

SUMMARY

SGD is continually expanding its resources to increase the ease of access to information about genes and proteins from fungi and other organisms. The Fungal BLAST and the Model Organism BLASTP Best Hits resources allow easy identification and examination of the conserved sequence regions in fungal genomes and facilitate the use of *S.cerevisiae* as a model organism and reference for comparison with other species. This will further aid in understanding the function and evolution of these sequences.

REFERENCES

1. Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman, *et al*. (2004) *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids. Res*., **32**, D311–D314.
2. Dwight,S.S., Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dolinski,K., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E., Hong,E.L. *et al*. (2004) *Saccharomyces* genome database: underlying principles and organisation. *Brief. Bioinformatics*, **5**, 9–22.
3. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al*. (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res*., **30**, 69–72.
4. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
5. Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
6. Blandin,G., Durrens,P., Tekaia,F., Aigle,M., Bolotin-Fukuhara,M., Bon,E., Casaregola,S., de Montigny,J., Gaillardin,C., Lepingle,A. *et al*. (2000) Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. *FEBS Lett*., **487**, 31–36.
7. Brachat,S., Dietrich,F.S., Voegeli,S., Zhang,Z., Stuart,L., Lerch,A., Gates,K., Gaffney,T. and Philippsen,P. (2003) Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. *Genome Biol*., **4**, R45.
8. Dietrich,F.S., Voegeli,S., Brachat,S., Lerch,A., Gates,K., Steiner,S., Mohr,C., Pohlmann,R., Luedi,P., Choi,S. *et al*. (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science*, **304**, 304–307.
9. Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol*., **266**, 460–480.