

Gene Function, Metabolic Pathways and Comparative Genomics in Yeast

Qing Dong, Rama Balakrishnan, Gail Binkley, Karen R. Christie, Maria Costanzo, Kara Dolinski, Selina S. Dwight, Stacia Engel, Dianna G. Fisk, Jodi Hirschman, Eurie L. Hong, Rob Nash, Laurie Issel-Tarver, Anand Sethuraman, Chandra L. Theesfeld, Shuai Weng, David Botstein, and J. Michael Cherry

Saccharomyces Genome Database, Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305 USA

yeast-curator@genome.stanford.edu

Abstract

The budding yeast, Saccharomyces cerevisiae, has been experimentally manipulated for several decades. Much of the information generated is available in the Saccharomyces Genome Database (SGD, <http://www.yeastgenome.org/>). SGD contains large datasets of both genomic and proteomic information, as well as tools for data analysis. This paper will highlight three datasets that are maintained by SGD. First, a large dataset of hand-curated information is provided in machine readable format for each gene of the Saccharomyces genome. These hand-curated annotations use the Gene Ontology (GO) controlled vocabularies for Biological Process, Molecular Function and Cellular Component and each contains categorical evidence codes and literature references. A second area of focus is on metabolic pathways. A new dataset of hand-curated information on metabolic pathways within budding yeast was released in May 2003. This resource can be searched to view biochemical reactions and pathways and their component gene products. This resource also maps data from genome-wide expression analyses onto the pathway overview providing a visualization of the changes in gene expression in the context of cellular metabolism. These pathways are created and edited using the Pathway Tools software but the content is reviewed and updated by SGD. A third dataset has recently become available as the result of two comparative genomic analyses. Two groups sequenced the genomes of several yeasts closely related to S. cerevisiae, and then completed a gene-by-gene comparison of these genomes. These genome comparisons were combined with available experimental evidence by SGD. Using these data the annotations for the S.cerevisiae reference genome were improved. All these datasets are freely available from the SGD ftp site (see Online Resources section).

1. Introduction

The *Saccharomyces* Genome Database (SGD) project collects information and maintains a database of the molecular biology of the budding yeast *Saccharomyces*

cerevisiae. This database includes a variety of genomic and biological information and is maintained and updated by SGD staff.

The sequence of the yeast genome was completed in 1996. The yeast community has been enjoying the complete genome sequence of strain S288C since 1996. Since that time the emphasis has shifted from identification of genes to determination of the role of the respective gene products in the cell. To meet the needs of the community, SGD provide its users with annotations that allow relationships to be made between gene products, both within *S. cerevisiae* and other fungal species. To this end, SGD has been annotating genes to the Gene Ontology (GO, <http://www.geneontology.org/>), a structured representation of biological knowledge that is shared across species [1]. The GO consists of three separate ontologies describing the Molecular Function of the gene product, the Biological Process that it is involved in and a Cellular Component or place in the cell where it is located [2]. Each hand-curated annotation includes an evidence code and a literature reference. SGD has also developed tools that facilitate both the display of GO annotations and the analysis of large sets of genes through shared function. Two newly developed tools are the GO Term Finder and the GO Term Mapper. The GO Term Finder searches for terms that are shared among a group of genes entered by the user. The search results are displayed in graphic or tabular form and may reveal that a subset of the user identified genes encode proteins that share a similar function, location or that may function in a common pathway. The graphic view illustrates the parent-child relationships of the GO terms that have been annotated to the gene products. The GO Term Mapper identifies major branches of the ontologies common to a set of user-defined genes or ORFs based on their GO annotations. The major branches of the ontology are represented by high-level terms, also known as GO-slim terms, which represent the parentage of the granular terms associated with individual genes.

Another recent focus of SGD is the metabolic pathways. The Yeast Biochemical Pathways (<http://pathway.yeastgenome.org/>) were created using the Pathway Tools software developed and maintained by Peter Karp and his colleagues at the Bioinformatics

Research Group at SRI International (<http://ecocyc.org/>). The Pathways Tools facilitates the retrieval and visualization of biochemical pathways and genomic annotations based on information in a Pathway/Genome Database. SGD is using the software to create yeast biochemical pathways. The yeast biochemical pathways were automatically generated using PathoLogic, a pathway prediction program built into the Pathway Tools. PathoLogic used biochemical information available at SGD (GO function annotations, sequence similarities, and Enzyme Commission (EC) classification) to initialize MetaCyc [3], a database that contains pathways from 150 different organisms and several hundred reactions, to create an *S. cerevisiae* specific pathway dataset. The automatically generated pathways are being manually reviewed by SGD curators to ensure that the pathways are correct and when incomplete, that new pathways are added. In addition, the Overview Expression Viewer feature of the Pathway Tools allows the user to superimpose expression data onto the metabolic Overview Diagram. Imported data files containing the expression levels for metabolic enzymes can then be layered or superimposed onto the Metabolic Overview Diagram, allowing the user to analyze the expression of metabolic pathway enzymes under various conditions or to create an animation of how expression levels change over time.

The third dataset, recently provided by SGD, is the sequence data described in two recently published papers, one from Mark Johnston's group at Washington University at St. Louis [4] and the other from Eric Lander's group at Whitehead Institute/MIT Center for Genome Research [5]. Both groups performed comparative genomic analysis of the *S. cerevisiae* genome relative to that of closely related. Combining this information with available experimental evidence, we were able to substantially improve the quality of annotations of *S. cerevisiae* reference genome. Currently, the assembled sequences for each species are available, and additional data will be available soon. In addition, SGD is working to provide this data in a common format to facilitate bioinformatic use of these files. These results are viewable on the web via the Fungal Alignments and Synteny Viewer option under the Comparison Resources menu from the SGD Locus pages.

All these datasets mentioned above, plus much more can be freely downloaded from SGD ftp site (see Online Resources section). SGD is funded as a Biotechnology Research Resource by the US National Human Genome Research Institute, US National Institutes of Health.

2. Online Resources

SGD home: <http://www.yeastgenome.org/>

SGD ftp site:

ftp://ftp.yeastgenome.org/yeast/data_download/

Go Annotation and Metabolic Pathway data:

ftp://ftp.yeastgenome.org/yeast/data_download/literature_curation/

Yeast comparative genome sequence data:

ftp://ftp.yeastgenome.org/yeast/data_download/sequence/fungal_genomes/

Please read the README file under each directory for more info.

3. References

- [1] S.S. Dwight, M.A. Harris, K. Dolinski, C.A. Ball, G. Binkley, K.R. Christie, D.G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethuraman, S. Weng, D. Botstein, and J.M. Cherry, "*Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO)", *Nucleic Acids Res.*, 2002 Jan 1, 30(1):69-72.
- [2] Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology", *Nat Genet.*, 2000 May, 25(1):25-9.
- [3] P.D. Karp, M. Riley, S.M. Paley, and A. Pellegrini-Toolehe, "The MetaCyc Database", *Nucleic Acids Res.*, 2002 Jan 1, 30(1):59-61.
- [4] P. Cliften, P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B.A. Cohen, and M. Johnston, "Finding Functional Features in *Saccharomyces* Genomes by Phylogenetic Footprinting", *Science*, 2003 May 29 (published on line).
- [5] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E.S. Lander, "Sequencing and comparison of yeast species to identify genes and regulatory elements", *Nature*. 2003 May15, 423(6937):241-54.