BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 14 2004, pages 2312–2314
doi:10.1093/bioinformatics/bth223

SPrCY: comparison of structural predictions in the Saccharomyces cerevisiae genome

T. J. Dolinsky[1], P. M. J. Burgers[1], K. Karplus[2] and N. A. Baker[1,*]

[1]Department of Biochemistry and Molecular Biophysics, Center for Computational Biology, Washington University in St Louis, St Louis, MO 63110, USA and [2]Biomolecular Engineering Department, University of California at Santa Cruz, Santa Cruz, CA 95064, USA

Received on January 27, 2004; revised and accepted on February 5, 2004
Advance Access publication April 1, 2004

ABSTRACT

Summary: SPrCY is a web-accessible database which provides comparison of structure prediction results for the *Saccharomyces cerevisiae* genome. This web service offers the ability to search, analyze and compare the yeast structural predictions from sequence-only (Superfamily, PDBAA BLAST and Pfam) and sequence-structure-based (SAM-T02, 3D-PSSM, mGenTHREADER) methods.

Availability: The service is freely available via web at http://agave.wustl.edu/yeast/

Contact: baker@biochem.wustl.edu

INTRODUCTION

This note presents the SPrCY (Structure Prediction Comparison for Yeast) database which compares the predictions of several fold-recognition techniques to the *Saccharomyces cerevisiae* genome. Protein structure prediction is a diverse and rapidly changing field with the state of the art assessed every two years at the CASP competition (Moult *et al*., 2003). There have been a number of previous structure and function prediction efforts (Hegyi and Gerstein, 1999; Sanchez and Sali, 1998) encompassing a wide range of goals and subjects; however, the focus of the SPrCY service is to provide an insight into three specific questions with respect to the *S.cerevisiae* genome: (1) for what fraction of the yeast genome can significant structural assignments be made using several different state-of-the-art structure-prediction methods? (2) To what extent do the various prediction methods provide consistent *and* accurate structural annotation of the genome? (3) To what extent can the predicted structures be used to suggest functional roles for yeast genes? We anticipate the comparison of structure prediction methods provided by SPrCY to be of interest with the computational biology community, and the new structural and functional annotations for the yeast genome to help guide new experimental research on this important model organism.

*To whom correspondence should be addressed.

POPULATION OF THE DATABASE

All predictions were obtained from protein translations of open reading frames (ORFs) of the *S.cerevisiae* genome (Cherry *et al*., 1997) obtained from the SGD website (http://www.yeastgenome.org) (Dwight *et al*., 2002; Issel-Tarver *et al*., 2002). Each ORF was then processed with a number of methods as outlined in the following sections. In addition to the initial population of the database described here, a subset of the results (Superfamily, SAM-T02, PDBAA and Pfam) are updated on a monthly basis or as new database versions become available.

PDBAA. Each ORF was searched against databases of sequences of proteins with known structures in the Protein Data Bank (PDB) (Berman *et al*., 2000a,b) using NCBI BLAST 2.2.4 (Altschul *et al*., 1997) using default options.

Superfamily. The SPrCY database also includes *S.cerevisiae* results from the Superfamily database (Gough *et al*., 2001), which maintains a set of genome matches to a large number of SCOP-based Hidden Markov Models (HMMs).

Pfam. The local HMMs from the Pfam database (Bateman *et al*., 2004) were used with the HMMER software (http://hmmer.wustl.edu/) to search the yeast genome using an *E*-value cut-off of 10.0. Pfam provides HMMs for both structural and non-structural domains and therefore complements data returned from the Superfamily searches.

SAM-T02. The SAM-T02 prediction method (Karplus *et al*., 2003) was run on all ORFs (including long ones and some not accepted as genes by the SGD database). The predictions were carried out at UC Santa Cruz, using a slightly modified version of the SAM-T02 web server (http://www.cse.ucsc.edu/research/compbio/HMM-apps/T02-query.html) and are updated approximately monthly, based on the changes in the PDB database and the template library.

mGenTHREADER. Each of the 5336 yeast ORFs with less than 800 amino acids was analyzed via the mGenTHREADER structure-prediction method (Jones, 1999;

McGuffin and Jones, 2003). These runs were performed on the PSIPRED structure-prediction server (http://bioinf.cs.ucl.ac.uk/psipred/) using the mGenTHREADER fold-recognition option. The few predictions which matched sequences to structural templates listed by the PDB as 'theoretical' were ignored.

3D-PSSM. Each of the 5336 yeast ORFs with less than 800 amino acids was analyzed with the 3D-PSSM method (Fischer *et al*., 1999; Kelley *et al*., 2000). Runs were performed by submission to the 3D-PSSM structure-prediction server (http://www.sbg.bio.ac.uk/~3dpssm/) using default options (global-local search, low-complexity filtering, and five iterations of PSI-BLAST). The few predictions which matched sequences to structural templates listed by the PDB as 'theoretical' were ignored.

DATABASE AND WEB SERVER FEATURES

The SPrCY website (http://agave.wustl.edu/yeast/) allows users to search, browse and analyze the generated predictions. The data obtained from each prediction method was parsed, cross referenced with different ORF naming schemes (allowing the user greater search options) and entered into a MySQL database. This database serves as a backend for the web server, which uses a Python frontend (via the MySQLdb package) to query the database as prompted by the CGI scripts available on the main website. The available scripts provide users with several ways to view the results of these calculations, including searches of ORFs and predictions by *E*-value, ORF name and PDB template. All available predictions can be viewed for each ORF, thereby allowing users to compare results between prediction methods and check for consistency. All website features are described to facilitate their use.

Additionally, users can browse and analyze the results in the context of the SCOP structural hierarchy (Lo Conte *et al*., 2002; Murzin *et al*., 1995). The SCOP tree can be traversed to identify ORFs placed in specific structural family and the consistency of predictions from the various methods can be assessed at each level of the structural hierarchy.

Finally, putative functional annotation was added to allow searching/browsing by ORF functional class. All functional assignments were based on the Gene Ontology (GO; http://www.geneontology.org/) (Ashburner *et al*., 2000) classification scheme due to its ease of access and widespread use in genome annotation. SPrCY also provides utilities to compare ORF GO IDs with GO IDs associated with predicted structures, thereby offering an additional tool for user assessment of structural predictions.

CONCLUSIONS

The SPrCY database and website presents the results of several structure-prediction methods applied to the *S.cerevisiae* genome. Users are able to search the database, browse by structural and functional classification and compare structure-prediction results between methods, the level of specific ORFs as well as structural and functional classes. Given the importance of yeast as a model organism and the large number of yeast ORFs with uncharacterized structures, it is anticipated that SPrCY will be a useful service for the yeast community.

ACKNOWLEDGEMENTS

N.A.B. would like to thank Gary Stormo, Garland Marshall and Sean Eddy for several useful conversations. N.A.B. and T.J.D. were supported by a grant from the National Partnership for Advanced Computational Infrastructure. P.M.J.B. was supported by NIH GM32431. K.K. was supported in part by Department of Energy grant DE-FG0395-99ER62849 and NIH grant 1R01GM068570-01.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. *Nat. Gene.*, **25**, 25–29.

Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

Berman,H.M., Bhat,T.N., Bourne,P.E., Feng,Z., Gilliland,G., Weissig,H. and Westbrook,J. (2000a) The Protein Data Bank and the challenge of structural genomics. *Nat. Struct. Biol.*, **7**(Suppl.), 957–959.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000b) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Cherry,J.M., Ball,C., Weng,S., Juvik,G., Schmidt,R., Adler,C., Dunn,B., Dwight,S., Riles,L., Mortimer,R.K. *et al*. (1997) Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature*, **387**(Suppl. 6632), 67–73.

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.

Fischer,D., Barret,C., Bryson,K., Elofsson,A., Godzik,A., Jones,D., Karplus,K.J., Kelley,L.A., MacCallum,R.M., Pawowski, K. *et al*. (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins* (Suppl. 3), 209–217.

Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.

Hegyi,H. and Gerstein,M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.*, **288**, 147–164.

Issel-Tarver,L., Christie,K.R., Dolinski,K., Andrada,R., Balakrishnan,R., Ball,C.A., Binkley,G., Dong,S., Dwight,S.S., Fisk,D.G. *et al*. (2002) Saccharomyces Genome Database. *Methods Enzymol.*, **350**, 329–346.

Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.

Karplus,K., Karchin,R., Draper,J., Casper,J., Mandel-Gutfreund,Y., Diekhans,M. and Hughey,R. (2003) Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction. *Proteins*, Suppl. 6, 491–496.

Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.

Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002 refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.

McGuffin,L.J. and Jones,D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, **19**, 874–81.

Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, **53**, 334–339.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Sanchez,R. and Sali,A. (1998) Large-scale protein structure modeling of the *Sacchromyces cerevisiae* genome. *Proc. Natl Acad. Sci., USA*, **95**, 13597–13602.