BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 22 2008, pages 2632–2633
doi:10.1093/bioinformatics/btn488

Structural bioinformatics

TESE: generating specific protein structure test set ensembles

Francesco Sirocco and Silvio C. E. Tosatto*
Department of Biology, University of Padova, Viale G. Colombo 3, 35131 Padova, Italy

Received on June 5, 2008; revised on August 26, 2008; accepted on September 10, 2008
Advance Access publication September 16, 2008
Associate Editor: Burkhard Rost

ABSTRACT

Summary: TESE is a web server for the generation of test sets of protein sequences and structures fulfilling a number of different criteria. At least three different use cases can be envisaged: (i) benchmarking of novel methods; (ii) test sets tailored for special needs and (iii) extending available datasets. The CATH structure classification is used to control structural/sequence redundancy and a variety of structural quality parameters can be used to interactively select protein subsets with specific characteristics, e.g. all X-ray structures of α-helical repeat proteins with more than 120 residues and resolution <2.0 Å. The output includes FASTA-formatted sequences, PDB files and a clickable HTML index file containing images of the selected proteins. Multiple subsets for cross-validation are also supported.

Availability: The TESE server is available for non-commercial use at URL: http://protein.bio.unipd.it/tese/.

Contact: silvio.tosatto@unipd.it

1 INTRODUCTION

Creating representative ensembles of sufficiently diverse proteins is a recurring problem in bioinformatics. Any novel method has to be trained and benchmarked on a test set of protein sequences and/or structures ensuring wide coverage of the protein universe and solid statistical evaluation. At least three different use cases can be envisaged: (i) The benchmarking of novel sequence alignment protocols and statistical potentials. (ii) The generation of test sets for specialized protein classes, e.g. transmembrane proteins. (iii) Extending datasets from previous publications with new structures to enhance statistical significance, e.g. for novel repeat proteins. The benchmarking problem has been recently addressed in the area of protein–ligand docking for instance (Jain and Nicholls, 2008). Given the exponential growth in available information, it is increasingly necessary to generate representative test sets large enough to allow solid statistical evaluation of the results. One of the earliest methods for the systematic selection of reduced protein lists from the Protein Data Bank (PDB; Berman *et al*., 2002) is PDBSELECT (Hobohm and Sander, 1994). It produces a list of protein sequences selected for a maximum percentage of sequence identity and reasonable structural quality. PDB-REPRDB (Noguchi and Akiyama, 2003) and UniqueProt (Mika and Rost, 2003) were developed to automate and facilitate the sequence selection process with more stringent similarity filters. More recently, the PISCES server (Wang and Dunbrack, 2003, 2005) has seen extensive usage for the generation of benchmark sets. PISCES combines both sequence similarity and structure quality filters to produce annotated lists of protein sequences. Structural alignments are used to improve the discrimination of proteins with weak sequence similarities. One limitation of the currently available services is the lack of an underlying structural classification throughout the selection process. This becomes increasingly important in the low sequence similarity range, where it is desirable to eliminate homology, and limits the usefulness of current methods in fold recognition for instance. On the other hand, the structural classification schemes, e.g. CATH (Pearl *et al*., 2003) and SCOP (Andreeva *et al*., 2004), are readily used for the selection of similar structures in absence of sequence similarity. However, only the full classifications are distributed and it is the developer's responsibility to extract meaningful subsets in a similar way to the previously mentioned services (e.g. PISCES). This process can become rather cumbersome in practice, e.g. when selecting structures with short tandem repeats or representatives of the Rossman fold. A lack of standardization and the relevance of many technical details in the selection process, frequently also complicates the unbiased assessment of novel methods to avoid 'cherry-picking' of the data. For these reasons, we have developed TESE, a novel server for the automatic generation of large benchmark sets both on the sequence and on the structure level.

2 FEATURES

TESE is a method to derive meaningful *ad hoc* test sets from proteins of known structure. The CATH structural classification is used to control sequence/structural redundancy at various levels, e.g. <35% pairwise sequence identity corresponds to the 'S' level. Queries may be started in three different ways, as in the schematic overview of Figure 1. Keywords or a small sample of PDB files can be used to seed the TESE search for specific proteins, e.g. for α-helical repeats or oxidoreductases, or to extend previously published datasets. Alternatively, the user may specify search parameters related to the desired CATH similarity level, e.g. topology, the experimental method and quality, e.g. maximum X-ray resolution or protein size, e.g. minimum length, to initiate the search. It is possible to select all structures or a randomly chosen subset of any size. For sets of less than 600 proteins, a clickable list of protein structures and their CATH classification is produced. New proteins may be selected by directly choosing a different protein subset or by adding additional search parameters. When satisfied, the user may save the protein list as a compressed archive containing the relevant FASTA-formatted sequences, PDB files and a HTML index of the selected proteins.

*To whom correspondence should be addressed.

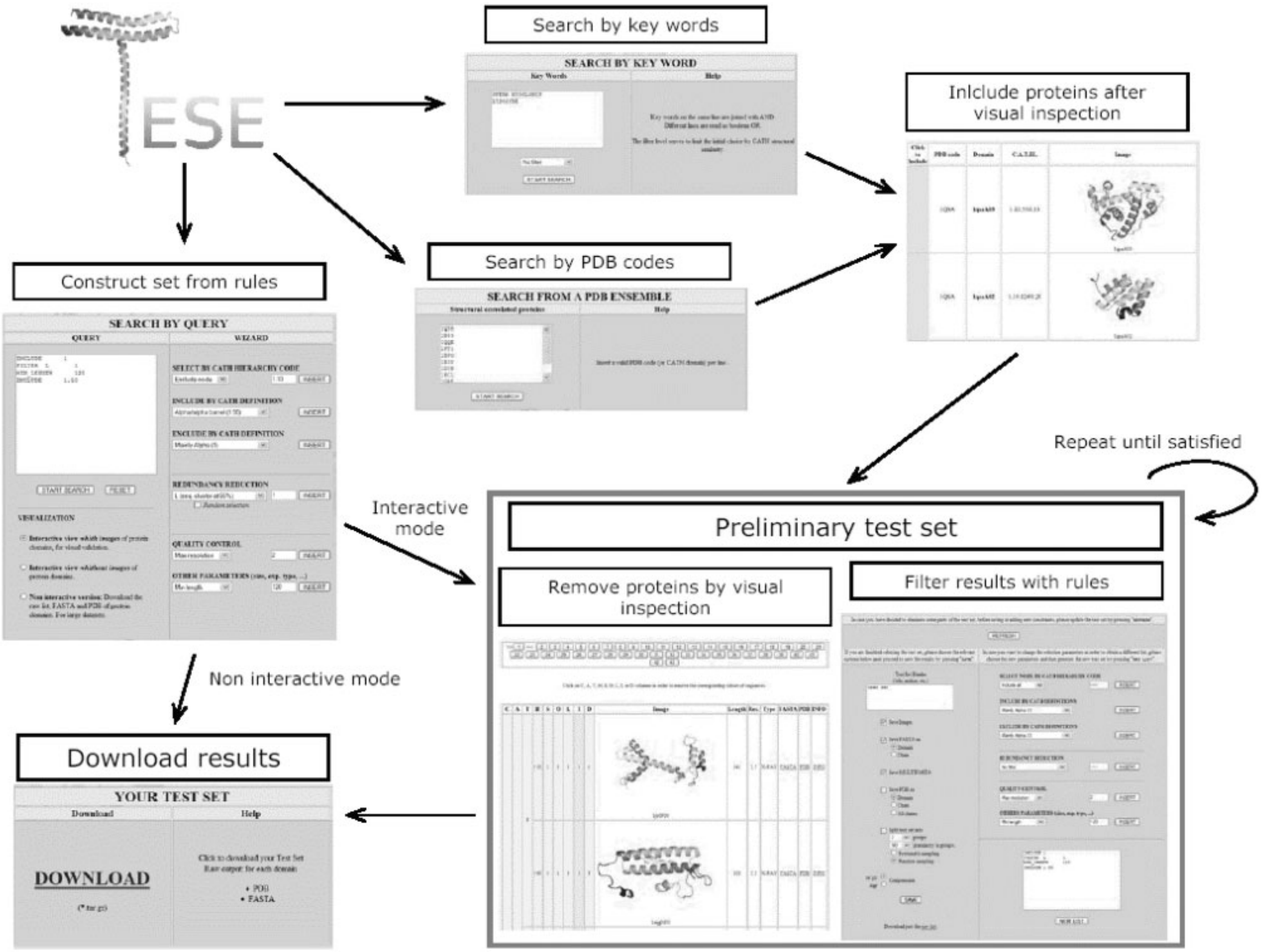

Fig. 1. Overview of TESE. The server has three main modes of operation: structural filters, list of PDB entries or keywords. These serve to generate a dynamically generated clickable list of structures from which to choose adequate structures. The process can be repeated iteratively, refining the search with additional structural filters, before saving the results as a compressed archive containing a HTML index with pictures, sequence and structure information.

The test set may be automatically split to create subsets for cross-validation. Large datasets of more than 600 proteins are treated in a non-interactive way to limit bandwidth usage. Some widely used test sets are available as precompiled archives. An online help is provided to guide the user through the process.

TESE uses a MySQL database containing information from the latest CATH release and PDBFINDERII (Hooft *et al*., 1996) to derive the relevant structural parameters with Perl scripts used for data conversion. The underlying databases are updated weekly and the TAP score (Tosatto and Battistutta, 2007) is calculated locally. Pictures of PDB structures are drawn using PyMol (DeLano Scientific LLC, URL: http://www.pymol.org/). A more extensive server description and examples are available from the web site.

ACKNOWLEDGEMENTS

The authors are grateful to Dr. Ingolf Sommer, Dr. Stefano Toppo and members of the BioComputing UP lab for insightful discussions.

Conflict of Interest: none declared.

REFERENCES

Andreeva,A. *et al*. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32** (Database issue), D226–D229.

Berman,H.M. *et al*. (2002) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.

Hooft,R.W. *et al*. (1996) The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. *Comput. Appl. Biosci.*, **12**, 525–529.

Jain,A.N. and Nicholls,A. (2008) Recommendations for evaluation of computational methods. *J. Comput. Aided Mol. Des.*, **22**, 133–139.

Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.

Noguchi,T. and Akiyama,Y. (2003) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucleic Acids Res.*, **31**, 492–493.

Pearl,F.M. *et al*. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.

Tosatto,S.C. and Battistutta,R. (2007) TAP score: torsion angle propensity normalization applied to local protein structure evaluation. *BMC Bioinformatics*, **8**, 155.

Wang,G. and Dunbrack,R.L., Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Wang,G. and Dunbrack,R.L., Jr (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.*, **33**, W94–W98.