TESE: Generating specific protein structure test set ensembles

Sirocco F[(1)], Tosatto SCE[(1)]

[(1)] Department of Biology, University of Padova, Padova

Motivation

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.
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.

Methods

One limitation of 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., Nucleic Acids Res 2003) 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. 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 (Sirocco & Tosatto, Bioinformatics 2008), a novel server for the automatic generation of large benchmark sets both on the sequence and on the structure level.

Results

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. Keywords or a small sample of PDB files can be used to seed the TESE search for specific proteins, e.g. for alpha-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. 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. A more extensive server description and examples are available from the web site.

Contact : silvio.tosatto@unipd.it