An update of the DEF database of protein fold class predictions

Martin Reczko, Dimitris Karras[1] and Henrik Bohr[2,*]

Synaptic Ltd, Aristotelous 313, 13671 Acharnai, Greece, [1]Computer Science Department, University of Ioannina, 45110 Ioannina, Greece and [2]Center for Biological Sequence Analysis, B. 206, DTU, DK-Lyngby, Denmark

Received October 8, 1996; Accepted October 9, 1996

ABSTRACT

An update is given on the Database of Expected Fold classes (DEF) that contains a collection of fold-class predictions made from protein sequences and a mail server that provides new predictions for new sequences. To any given sequence one of 49 fold-classes is chosen to classify the structure related to the sequence with high accuracy. The updated prediction system is developed using data from the new version of the 3D-ALI database of aligned protein structures and thus is giving more reliable and more detailed predictions than the previous DEF system.

INTRODUCTION

The DEF system (1,5) consists of protein fold-class predictions for sequences in, for example, the SWISSPROT protein sequence data base and of means to make predictions of fold-classes for any new sequence. In the DEF database a sequence of amino acids is assigned a specific overall fold-class, a super fold-class with respect to secondary structure content and a profile of possible fold-classes along the sequence. The assignment of a fold-class is one of 49 well-known folds derived from the three-dimensional protein structures in the Brookhaven Protein Data Bank, PDB. The 49 fold-classes were contained in the set given by Pascarella and Argos (2) and roughly in accordance with similar selections of folds (3,4) and consisted of protein domains with a distinct back-bone topology in their three-dimensional structure.

FOLD-CLASS SELECTION

For a reliable verification of the prediction system, it is required that each fold-class is represented by at least three structures with low pairwise sequence identity. In the new release of 3D-ALI avaliable on the WWW under http://www.embl-heidelberg.de/argos/ali/ali.html there are 49 classes with this requirement.

In terms of secondary structure content this list of folds is rather complete and well-balanced. That can be seen from the division of these folds into super classes of alpha-helical, beta-sheet, alpha-beta and alpha + beta structures (where alpha-helices and beta-sheets are intertwined in the third category but positioned in distinct domains in the fourth super class) which have alpha-helical and beta-sheet structures equally well represented.

METHODS

The prediction of fold-classes is produced from constrained neural networks. These networks were trained on half the members of the known structures in each class and tested on the remaining members in order to make an assessment of the quality of the prediction. The accuracy of the predictions is very high and gives valuable information for further structure determination of proteins from their sequences.

ADDRESS AND FORMAT OF THE DATABASE

The address for the DEF database system is:
http://zeus.cs.uoi.gr/neural/biocomputing/def.html

The format of the entries in the database is explained by an example given in the WWW address above.

REFERENCES

1 Reczko,M. and Bohr,H. (1994) *Nucleic Acids Res.* **23**, 3616–3618.
2 Pascarella,S. and Argos,P. (1992) *Protein Engng.* **5**, 121–137.
3 Holm,L. and Sander,C. (1993) *J. Mol. Biol.* **233**, 123–138.
4 Jones,D.J., Taylor,W.R. and Thornton,J.M. (1992) *Nature* **358**, 86–89.
5 Reczko,M., Bohr,H. *et al.* (1994) *Protein Structures by Distance Analysis*, p. 277. IOS Press.

* To whom correspondence should be addressed. Tel: +45 4525 2468; Fax: +45 4593 4808; Email: hbohr@cbs.dtu.dk