3616–3619 *Nucleic Acids Research, 1994, Vol. 22, No. 17*

The DEF data base of sequence based protein fold class predictions

Martin Reczko and Henrik Bohr[1],*
Molecular Biophysics, DKFZ-German Cancer Research Center, Heidelberg, Germany and
[1]Center for Biological Sequence Analysis, Building 206, The Technical University of Denmark,
Lyngby, Denmark

ABSTRACT

A new method for predicting protein fold-classes and protein domains from sequence data is constructed and used for generating a data base of protein fold-class assignments. Any given sequence of amino acids is assigned a specific prediction of one out of 45 typical protein fold-classes, a prediction of one out of 4 super fold-classes for the content of secondary structures and a profile of fold-class predictions along the sequence. The prediction accuracy for the super fold-classes is around 91% correct and 82% correct for the specific fold-classes. This accuracy is maintained down to a few percent of sequence identity.

INTRODUCTION

The DEF (Database for Expected Fold-classes) contains protein fold-class predictions from sequences in the SWISS-PROT protein sequence data base and is used for making 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 out of 45 well-known folds derived from the 3-dimensional protein structures in the Brookhaven Protein Data Bank, PDB. Most of these 45 fold-classes are contained in the set given by Pascarella and Argos [1] and are roughly in accordance with similar selections of folds [2,3]. In this context folds are protein domains with a distinct back-bone topology of their 3-dimensional structure. An extra requirement for the selection of the 45 classes, used as the basis for the predictions, is that they contained at least two members in order to make an assessment of the accuracy of the prediction [4,5]. Apart from the fold-classes contained in reference 1, some extra fold-classes, for example folds found in membrane-bound structures, were added in order to cover a wider range of structures. A list of the 45 protein fold-classes is given in Table 1. In terms of secondary structure content this list of folds is rather complete and well-balanced. This can be seen from the division of these folds into 4 super classes of α-helical, β-sheet, $\alpha\beta$ and $\alpha + \beta$ structures which have α-helical and β-sheet structures equally well represented. (The third super class stands here for α-helices and β-sheets intertwined while the fourth super class has α-helices and β-sheets separated in distinct domains).

METHODOLOGY

The following is a short presentation of the methodology employed for the prediction of protein fold-classes. The main tools are artificial neural networks that can be considered as knowledge based classifiers. These networks will gradually acquire a global information processing capacity of classifying data when being exposed (trained) to many pairs of corresponding input and output data such that new output can be generated from new input. The reason for choosing these networks among many other types is due to their renowned ability to generalize molecular biology data [6–10] and their rather simple structure both with respect to processing of data and training.

In the present application a special type of feed forward neural networks called Cascade-Correlators [11] are utilized. The training algorithm optimizes the weights and the number of hidden units in a feed-forward network by adding units during the training process. The process of adding new hidden units that maximize the correlation between their activity and the error remaining at the output layer is repeated until the mapping has the desired accuracy.

IMPLEMENTATION

The actual neural networks for predicting fold classes are constructed from the SNNS (Stuttgart Neural Network Simulator) environment [12]. The networks are trained on a selection of proteins from each of 45 fold classes containing domain segments of proteins or often whole proteins. The input representation for each protein domain is a 20×20 matrix containing integer numbers corresponding to the absolute frequencies of dipeptides occuring in neighbouring positions in the primary sequence of the domain. All protein domains are transformed in this way into one input pattern of fixed size. Insertions and deletions from the protein sequence cause only small changes in the dipeptide frequencies. The same holds true for rearrangements of larger elements in the sequence. There are many cases where members of the same fold class differ mostly by permutations of sequence

*To whom correspondence should be addressed at: Molecular Bio-Physics, DKFZ, Heidelberg 69120, Germany

Table 1. Example of an entry of 9pap predicted correctly as pap

```
   DEF Expected Fold Database , Version 0.1
PROTEIN-ID 9pap
NOTATION(FOLD-ID):
1 helix-bndl., 2 cytc, 3 hmr, 4 wrp, 5 globin, 6 lzm, 7 cyp, 8 ca-bind., 9 tln, 10 cts, 11 pap, 12 crn,
13 cpp, 14 wga, 15 sns, 16 plipase, 17 gap, 18 inhibit, 19 xia, 20 kinase, 21 m-binding, 22 TIM-barrel,
23 eglin, 24 pgk, 25 dfr, 26 sbt, 27 s-protease, 28 cpa, 29 ferrodox 30 fxc, 31 pti, 32 rdx, 33 virus,
34 virus-protease, 35 gcr, 36 igb, 37 il, 38 ac-protease, 39 tox, 40 plasto, 41 ltn, 42 hoe

   SUPER CLASS I (I:α, II:αβ, III:α + β, IV:β)
TOTAL CLASS pap
DOMAINS 1-100 pap, 110-200 pap
#SEQUENCE FOLDCLASS PROFILE (entries are structural similarity score 1-100)
```

SeqNo																				*foldclass*																
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	1	1	0	0	0	0	0	0	0	17	33	0	1	2	1	1	1	1	0	3	1	2	0	0	3	1	0	0	2	1	2	0	1	0	0	1
2	1	1	0	0	0	0	0	0	0	18	34	0	1	2	2	1	1	1	0	3	1	1	0	0	3	1	0	0	1	1	2	0	1	0	0	1
3	1	1	0	0	0	0	0	0	0	18	33	0	0	1	2	1	1	0	0	3	1	1	0	0	2	1	0	0	1	1	2	0	0	0	0	1
4	1	1	0	0	0	0	0	0	0	18	28	0	0	1	1	1	0	1	0	3	1	1	0	0	2	1	0	0	1	2	2	0	0	0	0	1
5	1	1	0	0	0	0	0	0	0	16	28	1	0	1	1	1	0	1	0	3	1	1	0	0	2	1	0	0	1	2	2	0	0	0	0	1
6	1	1	0	0	0	0	0	1	0	10	27	1	0	1	1	1	0	1	0	3	1	1	0	0	1	0	0	0	1	2	2	0	1	0	0	1
7	1	1	0	0	0	0	0	0	0	14	29	1	0	1	1	0	0	1	0	3	1	1	0	0	1	0	0	0	1	1	2	0	1	0	0	1
8	1	1	0	0	0	0	0	0	0	15	28	0	0	1	1	0	0	1	0	2	2	1	0	0	1	0	0	1	1	1	3	0	1	0	0	1
9	1	0	0	0	0	0	0	0	0	14	26	0	0	1	1	0	0	1	0	2	2	1	0	0	1	0	0	1	1	2	2	0	0	0	0	1
10	1	0	0	0	0	0	0	0	0	12	24	0	0	1	1	0	0	1	0	1	2	0	0	0	1	0	0	1	0	1	2	0	0	0	0	1
11	0	1	1	0	0	0	0	0	0	14	21	0	0	2	1	0	0	1	0	1	2	0	1	0	0	0	0	2	0	1	2	0	0	0	0	1
12	0	1	1	0	0	0	0	0	0	14	22	0	0	2	1	0	0	0	0	1	2	0	1	0	0	0	0	3	0	1	3	0	0	1	0	1
13	0	1	1	0	1	0	0	0	0	14	24	0	0	2	1	0	0	0	0	1	2	0	1	0	0	0	0	3	0	1	2	0	0	1	0	1
14	0	1	1	0	1	0	0	0	0	14	26	0	0	2	1	0	0	0	0	1	2	0	1	0	0	0	0	2	0	0	3	0	0	1	0	2
15	0	1	1	0	1	0	0	0	0	12	26	0	0	2	1	0	0	0	0	1	2	0	0	0	0	0	0	2	0	0	2	0	0	1	0	2
16	1	1	0	0	0	0	0	0	0	13	25	0	0	2	1	0	0	1	0	1	1	0	0	0	0	1	0	2	0	0	2	0	0	1	0	2
17	1	1	0	0	0	0	0	0	0	12	27	0	0	2	0	0	0	1	0	1	1	0	0	1	0	1	0	2	0	0	2	0	0	0	0	2
18	1	0	0	0	1	0	1	0	0	9	29	0	0	2	0	0	0	1	0	1	1	1	0	0	0	1	0	1	0	1	2	0	0	0	0	2
19	1	0	0	0	1	0	1	0	0	4	31	1	0	2	0	0	1	2	1	1	1	1	0	0	0	1	0	1	1	1	2	0	0	0	0	2
20	2	0	0	0	1	0	1	0	0	4	31	1	0	2	0	0	1	3	1	1	1	2	0	0	0	2	0	0	1	1	3	0	0	0	0	2
21	.	.	.	.	.	.	.	.																												

elements. Such permutations of the primary sequence lead to very similar dipeptide matrices which supports similar classification results. Each fold class is represented by one output unit which should have an activation close to 1.0 if the domain coded in the input layer is a member of that fold class. In all other cases the activity should be close to 0. When an unknown sequence is classified, the fold class corresponding to the largest activation at the output unit is assigned to the sequence.

DATASET OF KNOWN FOLD-CLASSES

The selection of protein structures that were used as basis for the fold-class predictions are listed in the top of Table 1. In some cases they were whole proteins and in other cases just distinct domains of proteins. Only classes that contained more than 1 structural fold was used and each class represented a particular distinct topology of the protein backbone chain. A sufficient but not necessary requirement for protein domains to be member of a given class was that they had more than 50 % sequence similarity to the other folds in the class but members could in some cases be down to 10 % sequence identical to other members of the same class. Most of the classes were contained in the set proposed by several authors [1−3] but some newer structures were incorporated in order to represent membrane-bound proteins and nucleotide binding proteins.

FORMAT

The format of the entries in the database is best explained by an example. The example is of an entry of the protein 9pap given in Table 1. In the first line just after the database name the sequence name is given. The next five lines give the full list of all the fold-classes which are used as basis for the prediction. The names of the fold-classes are taken from reference 1 and both in reference 1 and 6 all the proteins contained in the fold-classes are given. In the next line the super class prediction is given in Roman numbers and explained in a parenthesis. The next line gives the prediction of the specific fold-class. Next comes a line with the predicted domain specification. Finally comes the table with the profile of fold-class predictions made from a window of 100 residues moved along the sequence. The different fold-classes are numbered along the horizontal line and the sequence number down along the vertical axis.

The format for a new sequence to be assigned a predicted fold-class and profile is the following: 1. line contains the identifier (name of the sender), 2. line contains the sequence name and 3. the forthcoming lines should give the sequence of amino acids in the one-letter code with 80 characters on each line.

DISTRIBUTION

The database will be available from the end of August 1994 and can be obtained from:

Department of Molecular Bio-physics,
DKFZ, Heidelberg 69120, Germany.
Telephone: +49-6221-422372,
Telefax: +49-6221-422333

and is also available through the anonymous ftp address: mbp-sgi4.inet.dkfz-heidelberg.de in the directory/pdb/databases/def. The database can be easily accessed and is in principle public domain. The net-node number is 193.174.48.50. The database on the e-mail address will also be connected to an automatic mail server that can make fold-class predictions for any incoming sequence in the GCG format mentioned above.

DATABASE PROGRAM

The program that runs the test of the network and thereby the one that generates fold-class predictions for new sequences is installed in the directory /pdb/databases/def and is written in C+.

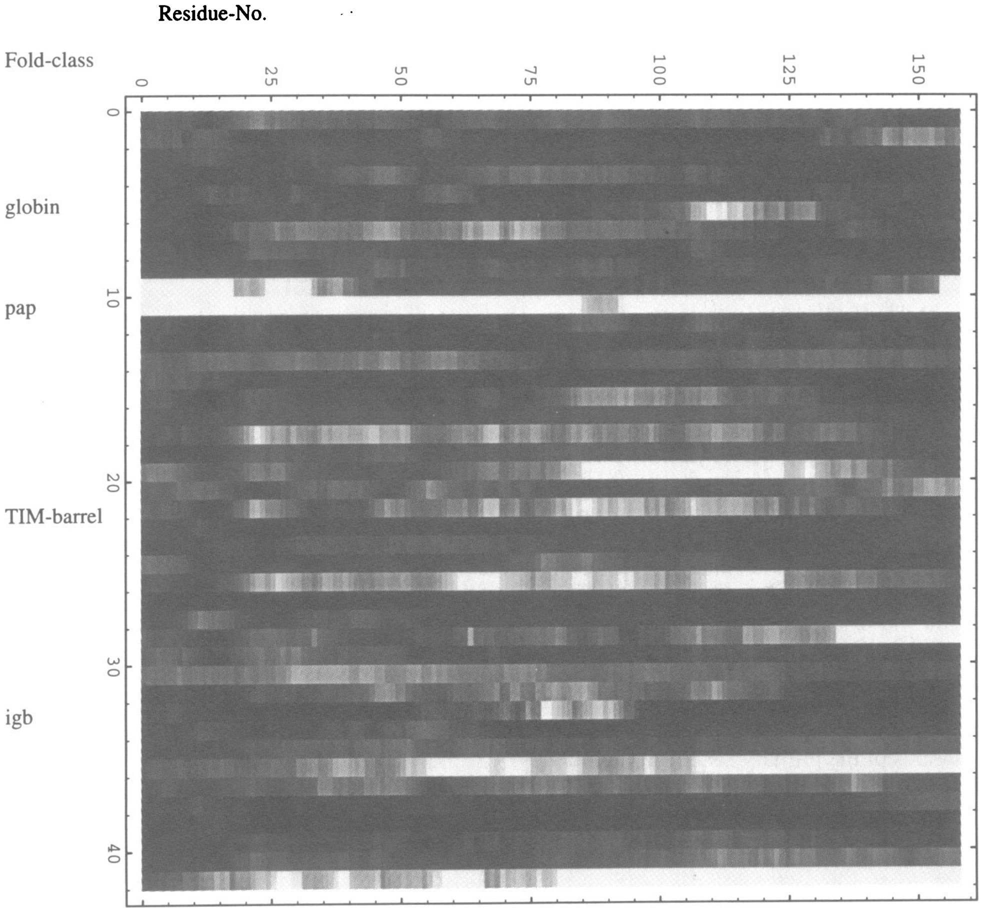

Figure 1. Graphical representation of the fold profile for 9pap. The protein is consisting of two distinct domains with separate folds which is clearly seen in the division of the white band for the pap fold (no. 11). The window of the profile prediction is 100 residues wide which means that the profile starts at the 51'th residue. A profile can also be made from the beginning of the sequence by adding a number of neutral characters corresponding to half the window size.

It is within the environment of the Stuttgartt Neural Network Simulator, SNNS, and basically assumes sequences written in the one-letter code and will give out a definite fold-class for the overall structure. However, for proteins with structures containing different domains the fold-class profile can give knowledge about the partition of the structure (see Fig. 1).

Prediction of 42 fold-classes (out of the 45 classes) is nicely illustrated in the permutation matrix shown in a figure in references 5 and 6 but similar results are seen in the Figure 1 below, where the profile or spectrum of similarities to all fold-classes and along the sequence in question. In that way each possible domain of a protein sequence around each residue can be evaluated for structural similarity. It is seen that most mistakes are made between classes close to the diagonal which means among similar classes. The classes are so arranged that the α-helical rich protein classes are at the upper end and the β-strand rich protein classes are on the bottom.

PERFORMANCE OF THE DATABASE PROGRAMS

The present networks appear to train acceptably well (about 100% correct in recall, and a similar Matthews's correlation coefficient [13,14]) on the task of predicting fold classification and distance matrix geometry [15]. Their predictive performance turned out to be rather successful with a score of around 82% for predicting fold classes (with a total of 42 classes). An improvement was obtained when the prediction of super classes was combined with the 42 fold-class predictions in a hierarchical way. In the case of predicting a correct super fold-class (out of a total of 4 classes) we obtained a correctness score of 90%. There have been other successful attempts recently in predicting fold classes on a less fine grained scale [16–18].

ACKNOWLEDGEMENTS

The authors wish to thank S.Subramaniam, P.V.Sudhakar and Laura Walsh at the Beckman Institute, University of Illinois at Urbana–Champaign and the scientists at the Department of Molecular Biophysics at DKFZ for help in constructing the database. From DKFZ especially Profs Sandor Suhai, Anil Saran and Karl Heinz Glatting and Drs Detlef Wolf, Artemis Hatzigeorgiou, Frank Herrmann and Martin Reese are acknowledged.

REFERENCES

1. Pascarella, S. and Argos, P. (1992) *Protein Engineering*, 5, 121−137.
2. Holm, L. and Sander, C. (1993) *J. Mol. Biol.*, 233, 123−138.
3. Jones, D. J., Taylor, W. R. and Thornton, J. M. (1992) *Nature*, 358, 86−89.
4. Reczko, M. and Bohr. H., Sudhakar, P. V., Hatzigeorgiou, A. and Subramaniam, S. (1993) 'Protein Structures by Distance Analysis', p. 277, Eds. Bohr, H. and Brunak, S., Published by IOS press, Maj 1994.
5. Reczko, M. and Bohr, H., Sudhakar, P. V., Hatzigeorgiou, A. and Subramaniam, S. (1994) Detailed protein fold class prediction from sequence, (paper submitted to Protein Engineering).
6. Qian, N. and Sejnowski, T. J., (1988) *J. Mol. Biol.*, 202, 865−884.
7. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Lautrup, B., Nørskov, L., Olsen, O. H. and Petersen, S. B. (1988) *FEBS Lett.*, 241, 223−228.
8. Holley, L. H. and Karplus, M. (1989) *Proc. Natl. Acad. Sci. USA*, 86, 152−156.
9. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Fredholm, H., Lautrup, B. and Petersen, S. B. (1990) *FEBS Lett.*, 261, 43−46.
10. Brunak, S., Engelbrecht, J. and Knudsen, S. (1990) *Nature*, 343, 123.
11. Fahlman, S. E. and Lebiere, C. (1990) in Advances in Neural Information Processing systems II, D.S. Touretzky, Ed. Los Altos, CA: Morgan Kaufmann, 1990, 524−532.
12. Zell, A., Mache, N., Sommer, T. and Korb, T. (1991) in 'Proc. Applications of Neural Networks Conf., SPIE, Aerospace Sensing Intl. Symposium', Orlando Florida, 1469, 708−719.
13. Mathews, B. W. (1975) *Biochem. Biophys. Acta*, 405, 442.
14. Stolorz, P., Lapedes, A. and Xia, Y. (1991) 'Predicting Protein Secondary Structure Using Neural Net and Statistical Methods', Los Alamos Preprint LA-UR-91−15.
15. Bohr, J., Bohr, H., Brunak, S., Cotterill, R. M. J., Fredholm, H., Lautrup, B. and Petersen, S. B. (1990) *J. Mol. Biol.*, 231, 861−869.
16. Goldstein, R. A., Luthey-Schulten, Z. A. and Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA*, 89, 9029−9033.
17. Dubchak, I., Holbrook, S. and Kim, S. (1993) *Prot. Struc., Func.and Genetics*, 16, 79−91.
18. Jones, D. and Thornton, J. (1993) *J. of Computer-Aided Molecular Design*, 7, 439−456.