Volume 261, number 1, 43-46 FEBS 08101 February 1990

A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks

H. Bohr, J. Bohr, S. Brunak*, R.M.J. Cotterill*, H. Fredholm°+, B. Lautrup† and S.B. Petersen°

*Risø National Laboratory, DK-4000 Roskilde, *The Technical University of Denmark, B. 307, DK-2800 Lyngby, °NOVO-Nordisk Research Institute, Novo Alle, DK-2880 Bagsværd, +UNI-C, DK-2800, Lyngby* and *†Niels Bohr Institute, Blegdamsvej 17, DK-2100 København Ø, Denmark*

Received 9 December 1989

Three-dimensional structures of protein backbones have been predicted using neural networks. A feed forward neural network was trained on a class of functionally, but not structurally, homologous proteins, using backpropagation learning. The network generated tertiary structure information in the form of binary distance constraints for the C_α atoms in the protein backbone. The binary distance between two C_α atoms was 0 if the distance between them was less than a certain threshold distance, and 1 otherwise. The distance constraints predicted by the trained neural network were utilized to generate a folded conformation of the protein backbone, using a steepest descent minimization approach.

Protein secondary structure prediction; Protein tertiary structure prediction; Neural network; Binary distance matrix; Steepest descent minimization

1. INTRODUCTION

One current aim of molecular biology is determination of the three-dimensional (3D) tertiary structures of proteins in their folded native state from their sequences of amino acid residues [1,2].

All secondary structure prediction methods [3–6] have reached a performance ceiling around 50–70%. This indicates the importance of long-range interactions between different local folding domains in the chain of amino acid residues [7]. The neural network method has been reported [8–10] to perform better than the Chou and Fasman method [3]. Recently, neural networks have also been applied to predict specifically the beta turns in proteins [11].

The 3D structure of proteins can be determined using either X-ray diffraction patterns of the crystalline phase, or NMR for proteins in solution [12]. The latter is currently limited to the study of proteins smaller than approximately 150 amino acid residues. The rate at which (3D) structures are being solved is at least one order of magnitude lower than the rate at which new protein sequences are being determined. Since the functional characteristics of a protein are intimately linked to its 3D structure [2,13–15], it is important to develop tools that can predict the structures corresponding to new sequences on the basis of knowledge acquired from known tertiary structures.

If significant sequence and functional homology exists between a protein of interest and proteins for which the 3D structures are already known, it is possible (but cumbersome) to build a reasonable 3D model of the protein's structure [16–18].

Correspondence address: S.B. Petersen, NOVO-Nordisk Research Institute, Novo Alle, DK-2880 Bagsvaerd, Denmark

2. MATERIALS AND METHODS

We here describe a method for predicting the 3D structure of a protein backbone from its amino acid sequence. A neural network (fig.1) was trained on matching sets of amino acid sequences and structural information of two different types, one being the corresponding secondary structure and the other the binary distance constraints, in the form of diagonal bands of binary versions of the C_α distance matrices [19]. In the case of a training set consisting of 13 proteases [20], the network was capable of learning to a level of 99.9% on the distance matrix output and 100% on the associated secondary structure assignment. The distance matrix for a protein novel to the network was generated by the trained network. The width of the diagonal band of the distance matrix was chosen to be 61. Because of symmetry, only the lower half was used, giving 30 distance constraints for each amino acid residue. Subsequently, steepest descent minimization was used to fold the protein's backbone until a maximally attainable number of the distance constraints were satisfied. These constraints are derived from the distance matrix. Each trace in the distance matrix comprises a prescription for local folding features, within the limitation imposed by the 30-residue horizon. In order not to exclude any potentially useful information, we adopted a procedure which took into account the entire number of constraints arising from the distance matrix prediction (which in principle could be as large as thirty times the number of amino acid residues). The combination of the two techniques thus constitutes a full attempt at protein folding from primary structure to tertiary structure of the backbone. In the training set consisting of the 13 proteases, all 61-residue windows were unique. Most secondary structural elements are defined within such a 61-amino-acid window. Exceptions are the linings of beta barrel structures [21] and very long-range parallel beta sheets.

3. RESULTS

The following figures present the main results of the current investigation. Fig.2a shows the binary distance matrix of the trypsin 1TRM (rat trypsin), which is 223

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/90/$3.50

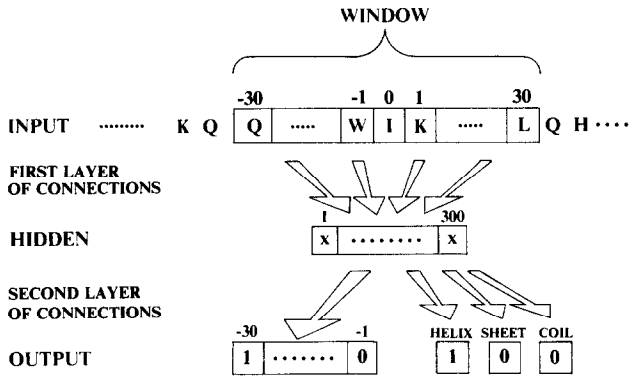

Fig.1. The architecture of the 2-layer (of connections) feed forward neural network used for predicing distance constraints from the sequence of amino acid. The input to the network is the amino acid sequence (here indicated in the one letter code). This sequence is moved stepwise through a window of 61 letters. The input information passes through a hidden level of neurons (needed for processing data internally), down to the output level. At the latter, a set of binary distances, between the centrally positioned amino acid and those lying to the left of it in the input window, is produced. Secondary structure assignments, in the 3 categories of helix, sheet and coil, are also given at the output level. Regarding the binary distance matrix, the network is trained to report which of the 30 preceding C_α atoms are positioned within a threshold distance of 8 Å to the centrally placed amino acid. The number of hidden units was 300 for an 8 Å threshold and 400 when the threshold was set to 12 Å. The input level had 1220 (20 × 61) units, the hidden level had 300 units, and the output had 33 units.

residues long. The network's ability to correctly assign structural information is amply illustrated in fig.2b, where the network is predicting the structure for a trypsin-like sequence, although the training set consisted of both trypsin and subtilysin. Although a significant degree of homology exists between the trypsin in question and the trypsins included in the training set, not a single input window presented to the network was identical to any window in the training set.

In the figure, there is a clear distinction between alpha-helices, anti-parallel and parallel beta sheets, as well as other tertiary motifs; the helices being bands parallel to the diagonal and anti-parallel sheets being stripes orthogonal to the diagonal.

Fig.2c shows the binary distance matrix of the folded backbone structure of this trypsin, using 4PTP as the starting configuration and the predicted binary distance matrix for the distance constraints. Although 1TRM is 74% homologous [22] to 4PTP, none of the 223 window configurations with 61 consecutive residues that could be generated from 1TRM, were represented in the training set. The result of that minimization is shown in fig.3a,b, and it agrees with the correct 3D structure of 1TRM, to within 3.0 Å rms. In this particular case, the length of the sequence used for the starting configuration was identical to that of the protein to be fitted. When the sequences are of unequal length, on the other hand, it is clear that additional considerations would have to be taken into account during the fitting process.

In order to test the performance of the steepest descent method, we also used a 223-residue coiled alpha helix as the initial configuration for the folding of 1TRM. The steepest descent minimization generates two subdomains separated by a 20-residue random coil domain. This result is remarkable because the folding motif of 1TRM (as well as that of other trypsins) clearly consists of two domains separated by a similar length of random coil segment at approximately the same position.

For large proteins, where the band of distance constraints does not cover all spatial contacts, local folding domains may acquire different chiralities, leading to improper packing of the domains in the protein. However, for 6PTI (bovine pancreatic trypsin inhibitor), which is only 56 residues long, we have been able to generate a correctly folded backbone structure

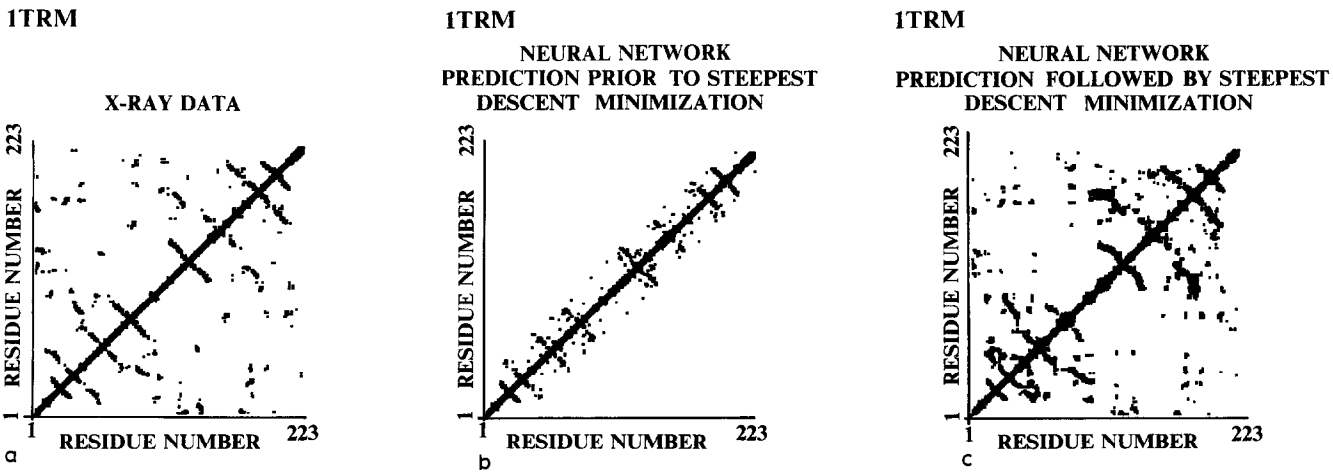

Fig.2. Binary distance matrices for 1TRM. The matrices (223 × 223) show which C_α atoms are within an 8 Å distance to each other C_α atom in the folded protein. a) The matrix corresponding to the structure determined from the X-ray data (8 Å threshold). b) Neural network prediction of an 8 Å distance matrix. A 61-residue band centered along the diagonal is generated. The network predicts this band with an accuracy of 96.6% c) The matrix corresponding to the structure produced by steepest descent minimization, using the neural network prediction as a starting point.

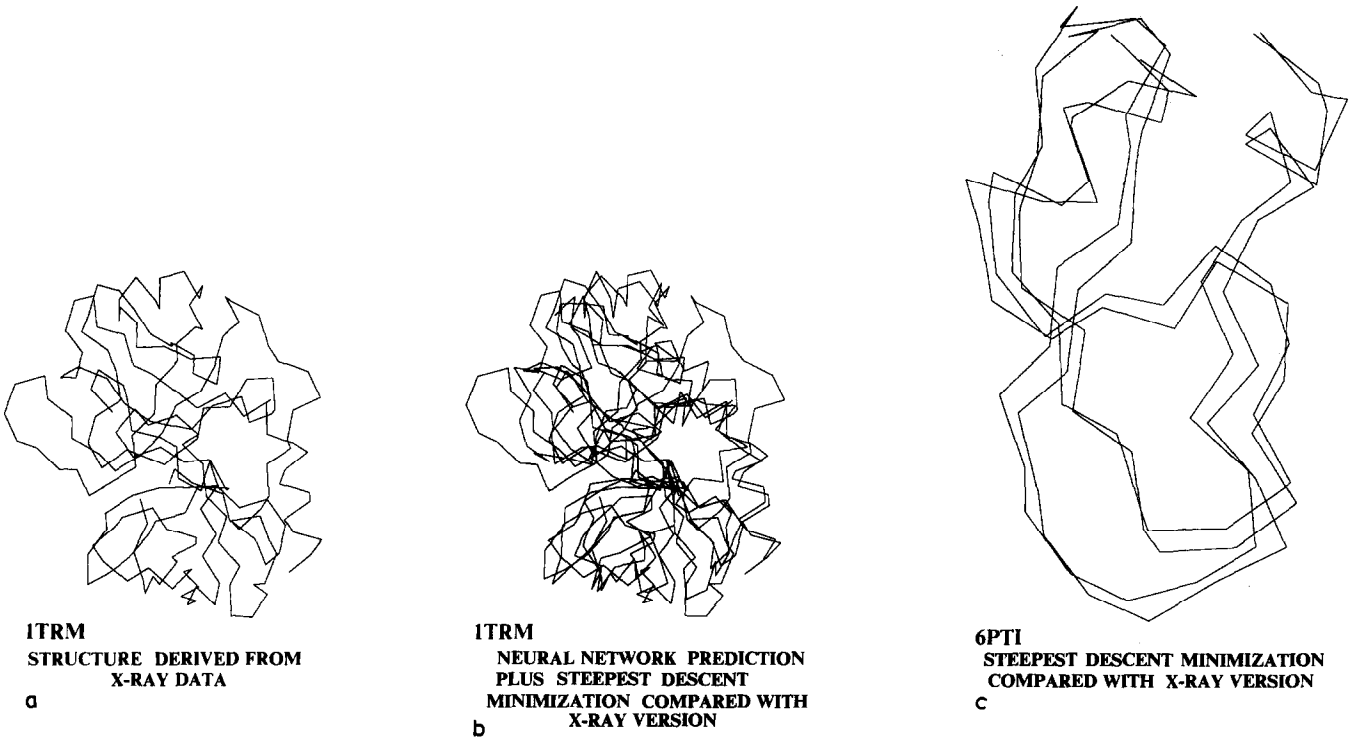

Fig.3. Backbone conformations for the trypsin 1TRM (223 residues long) and the trypsin inhibitor 6PTI (56 residues long). a) The crystal structure for 1TRM, as determined by the X-ray data. This corresponds to fig.2a. b) The predicted structure of 1TRM superimposed on the crystal structure. The rms deviation was 3.0 Å for all the C_α atoms. The largest deviations were present in surface loops, which in the crystal structure are fixed by several disulphide bridges. c) The crystal structure for 6PTI superimposed on the structure generated from the full distance matrix (56 × 56) at 8 Å, using the steepest descent method with a totally randomized initial configuration of the backbone. The distance matrix was obtained from the crystal data for 6PTI. The rms of the fit was 1.2 Å. This demonstrates that the steepest descent approach is capable of producing an essentially perfect fit to the X-ray data.

using the steepest descent method. The full binary distance matrix used for the minimization was generated from crystallographic data for 6PTI. Following convergence, the errors between the fitted and the correct structure lay within 1.2 Å rms (fig.3c).

4. DISCUSSION

The main achievement of this study has been the generation of a 3D structure of a protein from its amino acid sequence; the novel approach involving first the prediction of distance matrices using a neural network and thereafter a minimization fitting procedure. The results reported here are predictions of folded conformations, illustrated with the trypsin 1TRM. Our neural network is clearly capable of generalizing the folding information stemming from known proteins with homologous function. We are presently investigating in detail, how sequence homology between a protein and the proteins in the training set influences the quality of this approach's predictions. Distance constraints can also be derived from experimental procedures such as NMR, in which they take the form of nuclear Overhauser enhancement (NOE) factors. Structural information can be successfully derived from such data using restraint dynamics [23–26] which in its essential form bears some resemblance to the approach employed here, the most salient difference being that the potential energy function in our work is much simpler.

Finally, we note that further studies are necessary to clarify whether proteins with low levels of homology to other proteins of known structures can be predicted, if the neural network is trained on a much larger set of protein structures than is reported here.

Acknowledgements: We thank the Danish Ministry of Agriculture, Thomas B. Thrige's Foundation and The Danish Natural Science Research Council for grants.

REFERENCES

[1] Wetlaufer, D.B. (1984) The Protein Folding Problem, AAAS Selected Symposium 89.
[2] Jaenicke, R. (1987) Prog. Biophys. Mol. Biol. 49, 117–237.
[3] Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 211–245.
[4] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97.
[5] Lim, V.I. (1974) J. Mol. Biol. 8, 873.
[6] Robson, B. and Paine, R.H. (1974) Biochem. J. 141, 883.
[7] Taylor, W.R. and Thornton, J.M. (1984) J. Mol. Biol. 173, 487–512.
[8] Quian, N. and Sejnowski, T.J. (1988) J. Mol. Biol. 202, 865–884.

[9] Bohr, H., Bohr, J., Brunak, S., Cotterill, R.J.M., Lautrup, B., Norskov, L., Olsen, O.H. and Petersen, S.B. (1988) FEBS Lett. 241, 223-228.
[10] Holley, L.H. and Karplus, M. (1989) Proc. Natl. Acad. Sci. USA 86, 152-156.
[11] M.J. McGregor, Flores, T.P. and Sternberg, M.J.E. (1989) Protein Engineering 2, 521-526.
[12] Kline, A.D., Braun, W. and Wüthrich, K. (1988) J. Mol. Biol. 204, 675-724.
[13] Kikuchi, T., Nemethy, G. and Scheraga, H.A. (1988) J. Prot. Chem. 7, 427-471.
[14] Kikuchi, T., Nemethy, G. and Scheraga, H.A. (1988) J. Prot. Chem. 7, 473-490.
[15] Kikuchi, T., Nemethy, G. and Scheraga, H.A. (1988) J. Prot. Chem. 7, 491-507.
[16] Weber, I.T., Miller, M., Jaskolski, M., Leis, J., Skalka, A.M., and Wlodawer, A. (1989) Science 243, 928-931.
[17] Navia, M.A., Fitzgerald, P.M.D., McKeever, B.M., Leu, C.-T., Heimbach, J.C., Herber, W.K., Sigal, I.S., Darke, P.L. and Springer, J.P. (1989) Nature 337, 615-620.
[18] This is well illustrated by a recent successful attempt in our laboratory to build a subtilisin-like protease, savinase, based on the 62% homologous structure of subtilisin carlsberg. The later-determined X-ray structure was within 1 Å rms from the predicted structure on the whole protein backbone.
[19] Rossmann, M.G. and Liljas, A. (1974) J. Mol. Biol. 85, 177-181.
[20] Brookhaven Protein Data Bank entry codes: 1SGT (streptomyces trypsin), 2EST (porcine pancreatic elastase), 4PTP (bovine pancreatic beta trypsin), 2KAI (porcine pancreatic kallikrein A), 1CHG (bovine chymotrypsin A), 2PRK (fungal proteinase K), 1SEC (subtilisin carlsberg), 1SGC (streptomyces proteinase A), 2ALP (lysobacter alfalytic protease), 3APR (rhizopus acid proteinase), 3RP2 (rat mast cell proteinase), 2SBT (subtilisin NOVO). In addition we included the crystal structure data for the subtilisin savinase, which is not yet published.
[21] Lasters, I., Wodak, S.H., Allard, P. and Cutsem, E.V. (1988) Proc. Natl. Acad. Sci. USA 85, 3338-3342.
[22] The homology was measured using the Protein Identification Resource (PIR) 'ialign' program with the unitary protein matrix (bias = 6 and penalty = 1000).
[23] Wüthrich, K. (1989) NMR on Proteins and Nucleic Acids, Wiley.
[24] Clore, G.M., Gronenborn, A.M., Brünger, A.T. and Karplus, M. (1985) J. Mol. Biol. 186, 435.
[25] Van Gunsteren, W.F. and Berendsen, H.J.C. (1977) Mol. Phys. 34, 1311.
[26] Levitt, M. (1983) J. Mol. Biol. 170, 723.