Protein Structure Prediction System based on Artificial Neural Networks[1]

J. Vanhala and K. Kaski

Tampere University of Technology/Microelectronics Laboratory
P.O.Box 692
33101 Tampere, Finland
email jv@ee.tut.fi, tel. +358-31-161407, fax +358-31-162620

Abstract

Methods based on the neural network techniques are among the most accurate in the secondary structure prediction of globular proteins. Here the same principles have been used for the tertiary structure prediction problem. The map of dihedral ϕ and ψ angles is divided into 10 by 10 squares each spanning 36 by 36 degrees. By predicting the classification of each residue in the protein chain in this map a rough tertiary structure can be deduced. A complete prediction system running on a cluster of workstations and a graphical user interface was developed. Keywords: artificial neural networks, protein structure prediction, distributed computing.

Introduction

Neural network techniques have been successfully used in the prediction of the secondary structure of the globular proteins. Although secondary structures are normally defined by the presence or absence of certain hydrogen bonds between amino acid residues, they could also be described by the dihedral angles of the backbone. When the protein chain adopts a certain secondary structure the angles are forced to corresponding canonical values. Thus predicting the secondary structure is equivalent to predicting the classification of the backbone dihedral angles into classes corresponding to helix, sheet, and coil -structures. Each class include residues whose dihedral angles may differ over 90 degrees. Thus the secondary structure classification is not sufficiently accurate for deducing the tertiary structure. However by refining the classification the network can be taught to predict more accurate dihedral angles which in turn can be used as starting point for a search of the three dimensional structure.

The geometry of a protein backbone is relatively rigid in the sense that the bond lengths are practically constant and the variations of the bond angles are relatively small, Figure 1 (see *e.g.* Cantor & Schimmel 1980 or Momany *et al.* 1975). Since the backbone is a periodic linear structure of the three atoms, (-N-C-C-), there are three dihedral angles per residue. One of these, ω, is normally in trans-conformation, which leaves only two degrees of freedom, the dihedral angles ϕ and ψ. Note that the conformational flexibility of the side chain is much higher. To define the three dimensional structure of a protein backbone it suffices to list the ϕ and ψ angles. This makes the prediction of the conformation relatively well defined task: find a mapping from the space of amino acid sequences to a space of dihedral angle sequences, which will correctly map protein sequences onto their three dimensional structures. An empirical way of defining the mapping is to look at the proteins with a known sequence and a known structure and try to extract the information required for the mapping from them. This task can be approached from different viewpoints, using *e.g.* statistical methods, pattern matching, or as in this work using artificial neural network techniques.

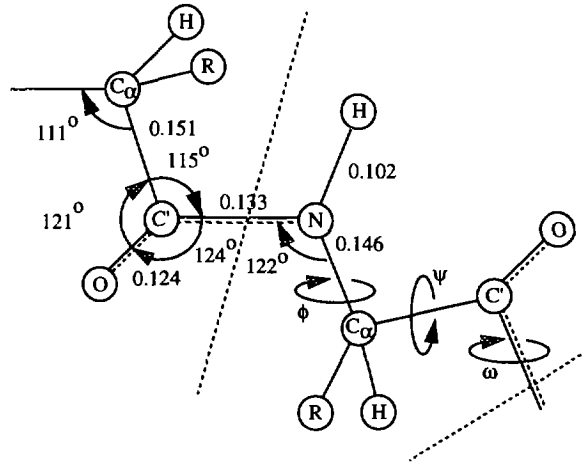


Figure 1. Protein backbone structure. Bond length are in nanometers. Bond angles are in degrees. ϕ and ψ torsion angles define the back bone conformation. Torsion angle ω of the peptide bond is normally in trans-conformation thus the consecutive Cα-C'-N-Cα atoms form a rigid planar peptide group. R denotes the side group. Dashed lines separate the amino acid residues.

[1]This paper describes the research based on the work done for the MOTECC computational chemistry software package described in Clementi 1991 and continued at the Tampere University of Technology.

An artificial neural network is used to predict the main chain conformation of globular proteins from their amino acid sequence. The dihedral ϕ and ψ angles of the protein main chain are calculated from the X-ray coordinates of a set of proteins from the Brookhaven Protein Data Bank (Bernstein *et al.* 1977) and has been given to the network during the learning phase. In prediction phase an amino acid sequence not included in the training set is input to the network and the network gives its prediction for the (or a set of) dihedral ϕ and ψ angles for each residue. The results can be used as a starting point for computationally intensive energy minimization calculations. Therefore neural network approaches may serve as a significant time saving by suggesting conformations relatively close to an energy minimum.

Neural networks in globular protein secondary structure prediction

Several research groups have used neural network techniques to predict the secondary structure of globular proteins from their amino acid sequence *e.g.* Qian and Sejnowski (1988), Holley and Karplus (1989), Bohr *et al.* (1988), and Mejia and Fogelman-Soulie (1990). Although the groups worked independently they used basically the same network architecture. The network was a feedforward layered network with full connection topology between layers and with sigmoid type nonlinear activation function and error back-propagation learning rule *i.e.* a standard back-prop network. The input to the network is a fragment of the amino acid sequence of a protein. The output gives the secondary structure classification for the residue in the middle of the fragment, see Figure 2. Secondary structure for the whole protein can be produced by sliding a window over the whole length of the amino acid sequence. The input to the network is represented using singular coding where each residue is coded with 20 input nodes. Each of these nodes correspond to one of the 20 naturally occurring amino acids. Thus the active unit in a input group flags the presence of a given residue type in that sequence position. For a window of n residues the input layer has the total of 20*n units.

In the output layer there is a separate unit for each of the secondary structure class used *e.g.* α helix, β sheet, random coil or turn. The level of activity of a given output unit can be interpreted as a probability of the corresponding class. Classification is then done by choosing the class with the highest probability *i.e.* activity. In addition to the input and output layers there is also a hidden layer. This is needed because a network with only input and output layers is capable of extracting only the first order features from the training set. These are the features that are caused by each input unit individually. Since the secondary structure of the proteins clearly depends on the local context of the amino acid

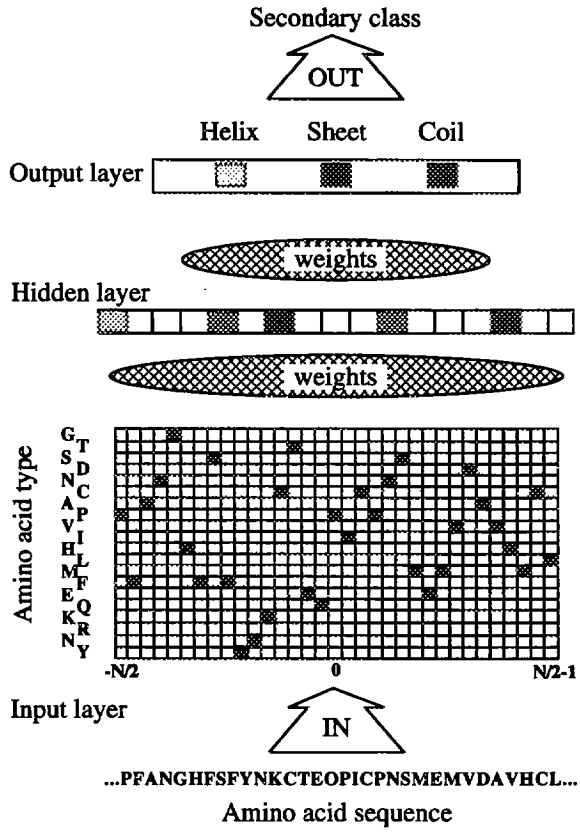


Figure 2. Network architecture for the secondary structure prediction. The input layer has one active node (shown in black) coding the residue type at each sequence position. The window width is N. Every node in the input layer is connected to every node in the hidden layer trough the lower set of weights. Also every node in the hidden layer is connected to every node in the output layer trough the upper set of weights. Amino acid types are given in one-letter code. The classification for the proline in the middle of the window is sheet, which has the highest activity level in the out-

sequence a hidden layer is needed. This gives the network the ability to represent also second order features *i.e.* features that depend on more than one input unit at a time.

Both Qian and Sejnowski and Holley and Karplus used also alternative ways of representing the input to the network. They used physicochemical properties of the amino acid residues such as hydrophobicity, charge, side chain bulk, and backbone flexibility. Qian and Sejnowski also provided the network with global information such as average hydrophobicity of the protein and the position of the residue in the sequence. All these attempts apparently failed to improve the prediction reliability over the basic scheme.

Bohr *et al.* used a separate network for each secondary structure class. Also they used two nodes to represent the output class. One node gives the probability p for belonging to the class and the other gives the probability (1-p) *i.e.*

the probability for **not** belonging to the class. In this way it is possible to recognize those cases where the network is not able to give a valid prediction.

Neural networks in globular protein tertiary structure prediction

Bohr *et al.* (1990) have developed a method where they predict the distance matrix (or rather a contact matrix) of a globular protein from its amino acid sequence using a feed forward neural network. The network tells if the distance between two α– carbons is less than a preassigned threshold value, *e.g.* 8 angstroms. All pairs of α– carbons up to 30 residues apart are considered, thus the network gives distance constraints for a diagonal band which is 30 residues wide. To generate the full distance matrix they use a deepest decent optimization to satisfy as many of the distance constraints as possible. Since the distance matrix can be generated from the back bone internal coordinates and *vice verse*, this method bears resemblance to the work described below, but uses a different way to represent the three dimensional structure of the protein backbone.

Their network is large compared to the size of their training set. Indeed the network is capable of reproducing the training set 100 percent correct. This will reduce the ability to generalize to unseen examples. However they use the network to predict the structure of a protein that has highly homologous counterparts in the training set obtaining good results. They do not report results for proteins that do not have homologous proteins in the training set. Wilcox and Poliac (1989) report even bigger network and use fewer proteins in the training set. Their network is capable of correctly recalling the distance matrices of the proteins in the training set when given the hydrophobicities of the amino acid sequence; however they do not test the network with unknown proteins. Fiedrics and Wolynes (1989) use a method closely related to Hopfield type neural networks to predict the tertiary structure of globular proteins. They call their method associative memory hamiltonian.

Methods

Our approach draws ideas from two lines of development in protein structure prediction. The first is the work done with artificial neural networks aimed at predicting the secondary structure. This has given us the tool. The second source of inspiration has been the methods developed by Lambert and Scheraga (1989a, b, c) to predict tertiary structures. This has given us some practical guide-lines and a reference for comparing our initial results.

Our training set is the same as the one used by Lambert and Scheraga in their pattern recognition-based importance-sampling minimization (PRISM) method. This choice offers the advantage for fair comparison of results.

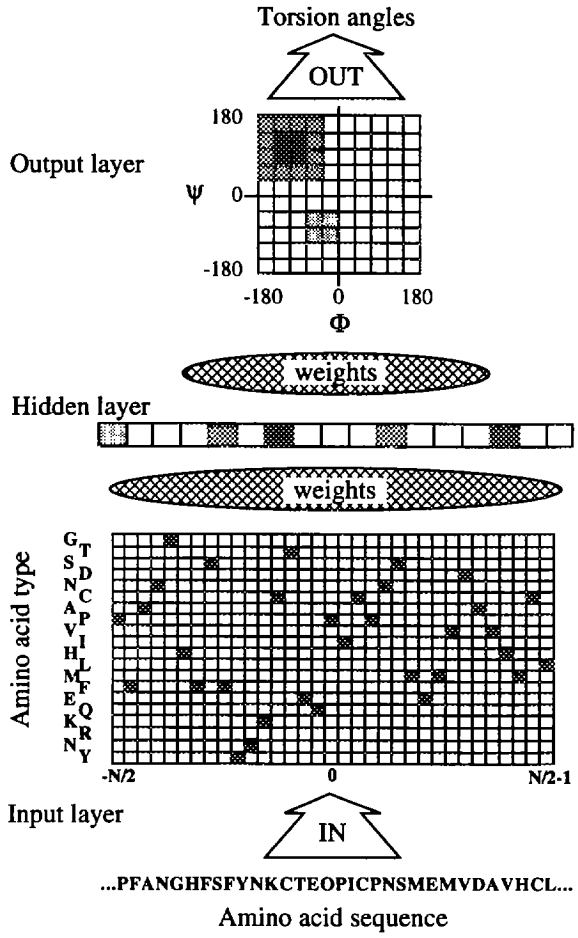


Figure 3. The network architecture for the tertiary structure prediction. The output layer in Figure 2 has been replaced with a map of 10 by 10 output nodes. Each node correspond to 36 by 36 degrees square of the ϕ–ψ diagram. Every node in the hidden layer is connected to every node in the output map trough the upper set of weights. In this case the output for the proline in the middle of the window can be interpreted approximately as ϕ=-100°, ψ=100°

Also we could skip the time consuming phase of studying the protein structure data banks and making fair and unbiased selections from them. We could as well have took either of the training sets collected by Kabsch and Sander (1983) or by Qian and Sejnowski (1988). The training set is collected from the Brookhaven Protein Data Bank by Bernstein *et al.* (1977) and it contains 46 strands from 43 proteins and has 6964 residues.

Because of the differences in the methods we had to make some slight modifications to the training set. Lambert and Scheraga calculated conformational probabilities for tripeptides *i.e.* the window width is three. Because of this choice they had to drop the first and the last residue from each amino acid sequence. Instead we have used a much larger window of residues. The width of the window varies

from case to case from 7 to 51 residues. Since neglecting one half of the window size of the terminating residues at both ends would decrease the training set too much, we define a new null residue that can be appended to both ends of the amino acid chain (as is also done by *e.g.* Qian and Sejnowski). In addition we could not simply remove from the training set those residues that are in cis-conformation or those that are of nonstandard type. Instead we define another new null residue which is skipped during the training phase. Note that in the prediction phase all amide bonds are also considered to be in the planar trans-conformation. All protein sequences have been concatenated into one long sequence with the null spacer residues between the proteins. The file that contains the training set has the residue identifiers together with the corresponding ϕ and ψ angles. These angles were calculated from the protein X-ray chrystallografic data in the protein structure data bank. Note that although the training set file contains the actual angles, they are classified into discrete output classes before being given to the network in the training phase.

We have used a feed forward layered neural network and the error back propagation algorithm. Lambert and Scheraga (1989a, b, c) did not report on prediction for the ϕ and ψ angles. The four classes is the upper limit for the Lambert's and Scheraga's approach since they collect the statistics for each of the 64 ($=4^3$) possible tripeptide conformational combinations. For even four classes the frequency of more rare combinations is not sufficient to obtain good prior probabilities. In our case in order to make more refined predictions for the angles we divide the ϕ-ψ map into a 10 by 10 grid. Each square spans an area of 36 degrees by 36 degrees, see Figure 3. As the output layer in the network may have several active units, we need a postprocessing phase that identifies the most probable answers and also gives the actual angles instead of a classification into the hundred output classes. First the output is low pass filtered to smooth out excessive roughness. Then all peaks are identified and their relative strength is calculated to give them a rank. At this point the output

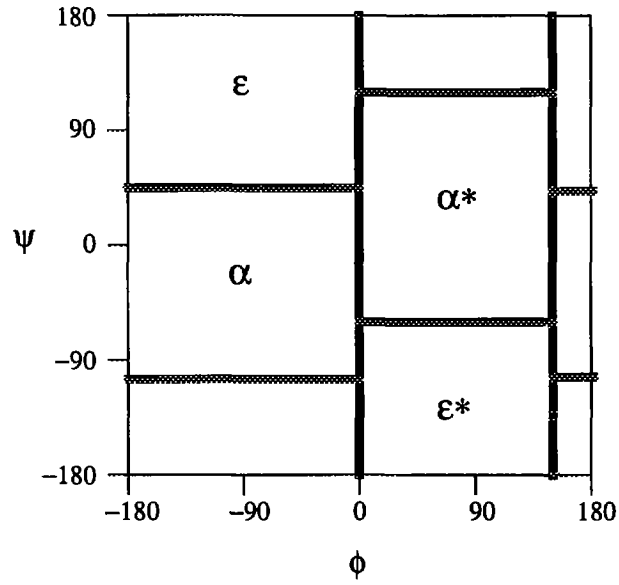


Figure 3. Conformational quadratures. Region a contains α helices, π helices, and isolated turns. Region ε contains β bridges, β sheets, and proline rich strands. Region α* contains left handed α helices. Region ε* contains glycine strands.

map is still in discrete form. The actual ϕ and ψ angles are produced by computing the "center of mass" for a neighborhood of each peak.

Results

Predicting conformational quadratures of PPT

To be able to make comparison to work done previously we choosed to try to predict the conformational classes of the avian pancreatic polypeptide as have been done by Lambert and Scheraga (1989a). They divide the two dimensional ϕ-ψ map into four regions as shown in Figure 4. Their prediction method PRISM gives the classification

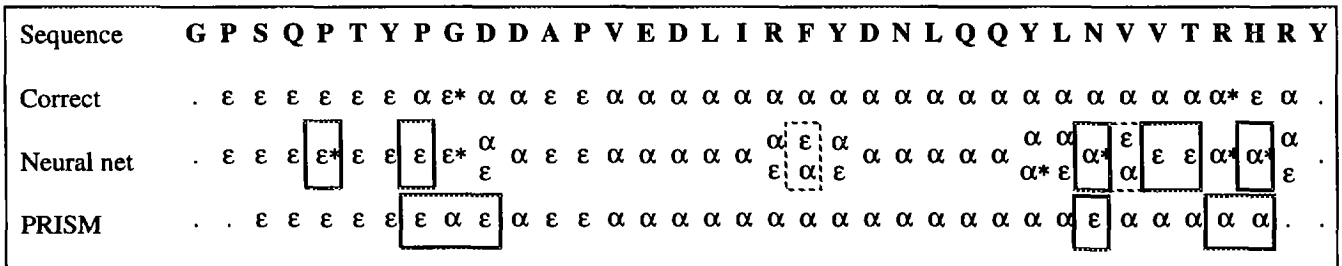


Figure 4. Secondary structure prediction for avian pancreatic polypeptide 1PPT. The top line gives the amino acid sequence in one-letter code and below that is the correct classification into the four conformational quadratures. The predictions with the neural network and with the PRISM method are on the bottom two lines. Hard errors are shown in solid boxes and soft errors in dashed boxes. The dots at the both ends of the classification denote the unused residues.

into one of the conformational classes for each residue in the sequence. This approximate description can be used as an input to a subsequent energy minimization step. In the similar manner we created a neural network with four output nodes each corresponding to one of the conformational quadratures. Using the same protein structure data base we taught the network to map the amino acid sequence to output classes. The results for both methods are shown in Figure 5. The four conformational classes α, ε, α^* and ε^* are shown. The first line gives the amino acid sequence in one letter code. The second line gives the correct classification derived from the X-ray coordinates. Below those there are the prediction results from the neural network algorithm and from the PRISM method. The neural network does not simply give one class as an answer but rather a probability of the correct classification of being in some quadrature. In this way one can have a multiple choices in situations were the network is not able to make its mind. Errors are classified in two classes: a box with thin edges is an error where the alternate answer is correct and a box with thick edges is an hard error where for all alternatives the prediction is wrong. Both the neural network and the reference has 6 hard errors, but the network gives also two soft errors. The performance of the network is a bit worse than that reported for PRISM. One explanation for this is that they used elaborate statistical means in dealing with rare cases in the amino acid sequence data, something that the neural network in its extreme simplicity is not able to do. Since the classification is not fully correct, it is not possible to obtain a tertiary structure which is relatively close to the native one. In PRISM, the problem is solved by generating a multitude of dihedral angle classification sequences and choosing the most probable ones for further processing. This tends to be a very time consuming procedure with an exponential run time complexity with respect to the chain length. Indeed it becomes unpractical to process proteins with more than 100 residues. The neural network approach does not have this limitation, since it gives multiple answers simultaneously and it is easy to sort these according to their probabilities.

Predicting ϕ and ψ angles of BPTI

Bovine Pancreatic Trypsin Inhibitor (BPTI, 5PTI) is a much studied protein that has 58 residues. The prediction by the network has been analysed by finding the fragments of the protein chain that has (near) correct ϕ and ψ angles. In the Figure 6 the predicted backbone structure of BPTI is shown together with fragments of the native BPTI structure from the X-ray data. About a half of the residues fall into correctly predicted regions. Both α helices (residues 2-7 and 47-56) and the β sheet fragment (residues 29-35) are correct. Also four ends (residues 5, 14, 38 and 51) of the three sulphur bridges (S_1, S_2, and S_3) are inside these regions. The predicted structure for BPTI is not globular but elongated. However since the cysteine residue pairs that form the sulphur bridges are normally known it "should be possible" to fold the structure into more globular shape by forcing the correct cysteine pairs together and at the same time trying to maintain the predicted dihedral angles. The native and the predicted structure of BPTI is shown in Figure 7.

The distance matrices shown here display all distances of the α -carbons of the protein chain. The matrix is symmetric, thus only one half of it is shown. This makes it easy to compare two structures by placing them on each side of the diagonal. The distances are gray scale coded. The change of colour corresponds to a distance difference of 2 Ångströms. The 16 shades cover a region from 0 to 32 Å. The distance matrix representation has some advantages over displaying the structures on the computer screen: It is rotation invariant so that the structures do not have to be aligned. It can show both local small scale details (close to

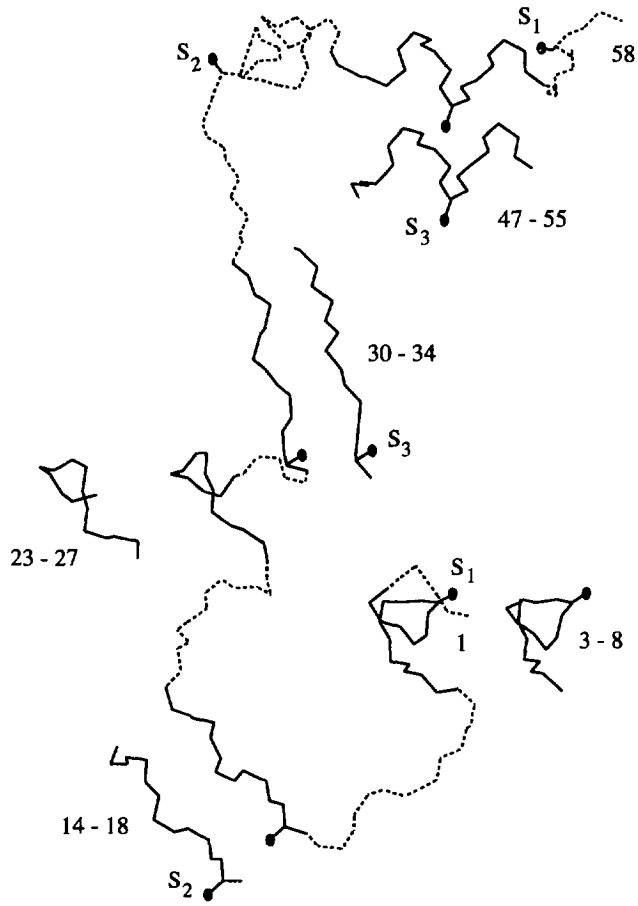


Figure 6. Fragments of the BPTI backbone from X-ray data aligned with the predicted structure. Solid line shows the correctly predicted regions. The numbers refer to the amino acid sequence positions. The three sulphur bridges S_1-S_3 are also shown.

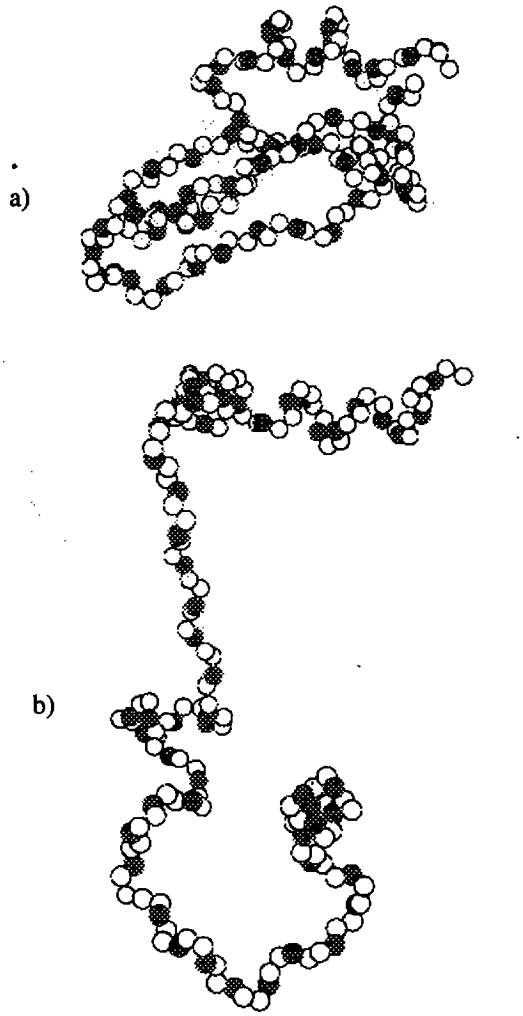


Figure 7. The native structure (a) and the predicted structure (b) of BPTI.

the diagonal) and the overall conformation in the large (the overall pattern). The scale depends on from how far you are looking at the picture. It is insensitive to few big errors in dihedral angles. Inspecting visually the three dimensional structures on the computer screen can be very difficult since although most of the predicted angles are relatively close to correct ones, there are few angles that are predicted with almost the maximum error, 180 degrees. This distorts the overall structure and the corresponding residues may end up in completely different places and rotations.

In Figure 8 is the distance matrix of BPTI. The two sides close to the diagonal match rather nicely. The α helices which are shown in light colour are predicted correctly otherwise but there is an extra piece of helix in the lower right part. Only the very start of the β sheet structure (light strand

Figure 8. The distance matrix of BPTI.

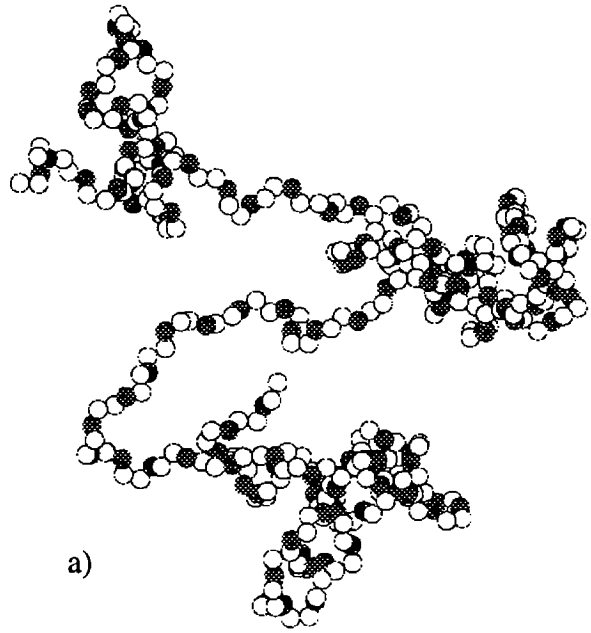
perpendicular to the diagonal) is predicted. One could also imagine to see parts of the two spherical "eyes" on both sides of the β sheet. The overall structure is elongated, thus the right upper corner is black which tells that the residues far from each other in the sequence are also far from each other in the three dimensional structure.

Prediction of the ϕ and ψ angles of LZ

Human Lysozyme (1LZ1) is an example of a bit longer protein chain. It has 130 residues. Again the overall structure of the prediction seems to be nonglobular, Figure 9a. It seems to be put together from smaller regions that are connected by long strands of random coil. The distance matrix in Figure 10 shows the same interesting features. The closely packed regions have good correspondence to the native protein. Specially the upper left corner of the distance matrix, which shows the start of the protein chain, makes almost an exact match. α helices are correctly identified. The β- structures in the middle of the matrix seem to have some relation to the native side.

Prediction of the ϕ and ψ angles of PPT

Avian Pancreatic Polypeptide (1PPT) is an example of a small protein. It has only 36 residues. The native structure of 1PPT is a fold of one long α helix and a shorter strand of β sheet. This can also be seen from the lower left part of the distance matrix, Figure 11. The prediction has correctly produced the helix and the β sheet but their relative



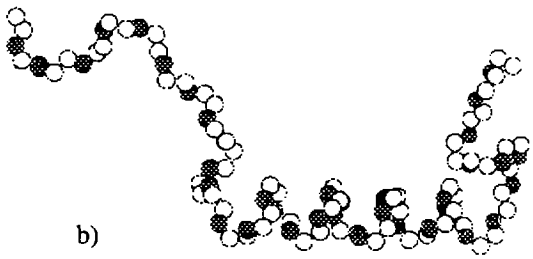


Figure 9. The predicted structure of 1LZ1 (a) and 1PPT (b).

positions are wrong. The secondary structures are correct but their packing is not. This can also be seen in Figure 9b.

The user interface

The arrangements of the elements in the graphical user interface for the protein tertiary structure prediction system is shown in Figure 12. This version is build using the OpenWindows Graphical User Interface Design Editor, GUIDE (SunSoft 1991).

The **Protein structures** list shows the structures that are available on the computer system. Each structure is in a separate file in PDB-format. The system uses only the SEQRES and ATOM records. The training set for the neural network is created by copying structures from the **Protein structures** list to the **Training set** list. This is done with

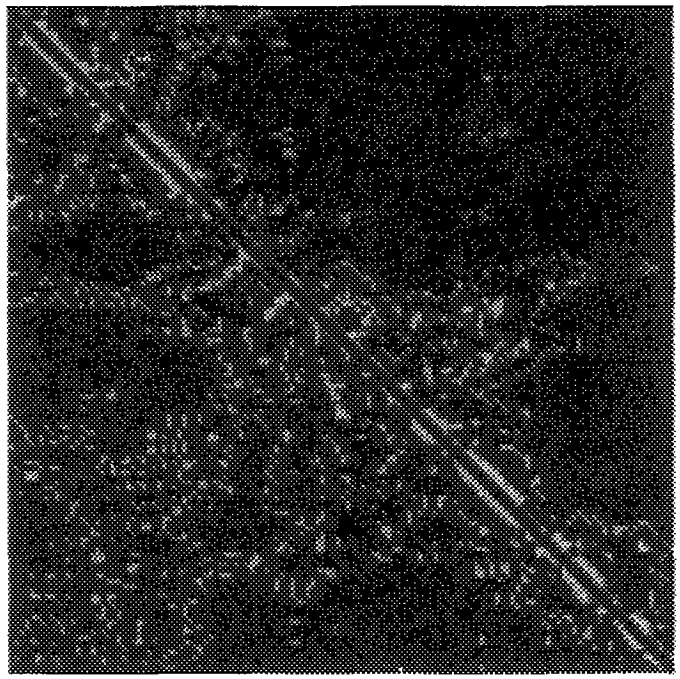


Figure 10. The distance matrix of 1LZ1.

the **Insert** button. First a structure name is highlighted by clicking the mouse button on that entry and then clicking on the **Insert** button. The system will copy the structure name on the **Training set** list. If a wrong structure is accidentally inserted into the training set, it can be removed by highlighting the structure name and clicking the **Delete** button. The PDB dataset is first loaded by giving its location on the computer file system (*i.e.* the directory containing the files) on the line next to the **Load** button and then clicking the button. **Training set** list shows the names of the structures that will be used when training the neural network. The list can be saved for a later use with the **Save**

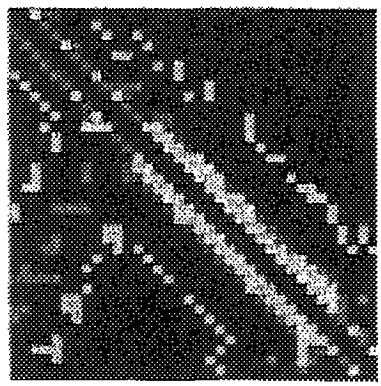


Figure 11. The distance matrix of 1PPT.

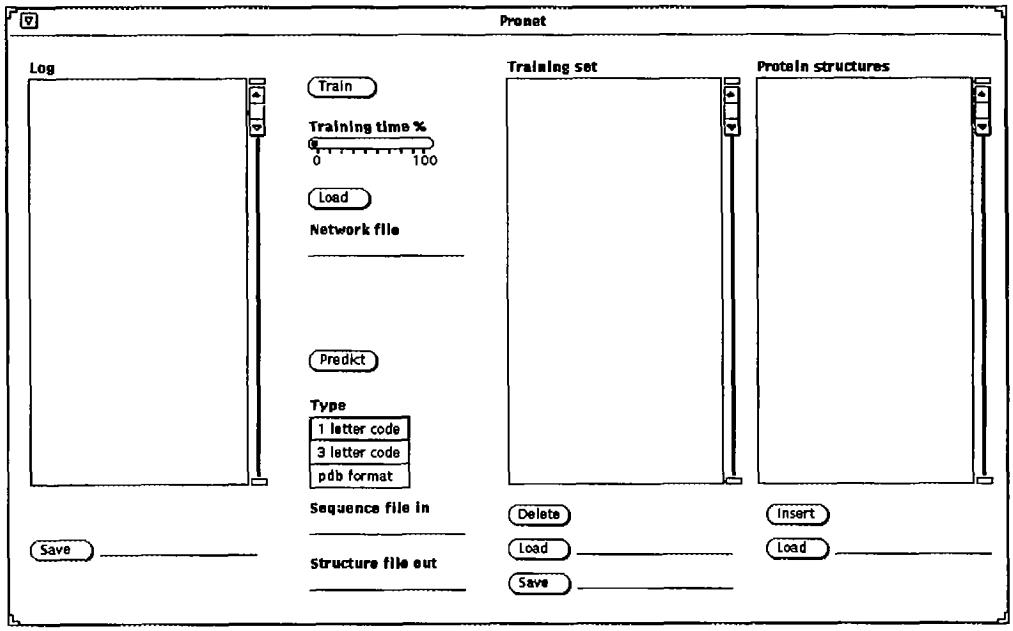


Figure 12. The graphical user interface.

button. The saved set can be retrieved from the saved file with the **Load** button. In this way several different training sets can easily be maintained for *e.g.* predicting the structures in a known protein family.

When the training set has been created the network can be trained simply by clicking the **Train** button. Since the training may take longer than one is willing to wait at the keyboard there is a gauge that shows how far the training has proceeded. The trained network is saved into a file whose name must be given on the line below the **Network file** heading. In case the network has been trained before with the same training set there is no sense to train the network again. Thus the network can be retrieved from a previously saved file with the **Load** button. If the network is already trained or loaded from a file before the **train** button is clicked the training is continued from this old state.

A new structure can be generated by first giving the name of the file containing the sequence information. The file may be in either PDB-format or written with the 1-letter code or the 3-letter code. The type must be indicated by clicking the appropriate button. The name of the prediction result which is always in the PDB-format (with only SEQRES and ATOM records) must also be given. The structure is generated by clicking the **Predict** button. The prediction phase is very fast compared to the training phase, thus there is no gauge to show the training time.

Everything that is done is echoed on the **Log** window. Every time a button is pressed or a file name is given the log is updated. You can also make your own notes on the log. The log can be saved into a file for later reference with the **Save** button.

Discussion and future development

We recall that Holley and Karplus and Qian and Sejnowski have attempted a coding based on physicochemical properties of amino acids They were not successful in improving over the simple method, using singular coding. Perhaps this has blurred the network visibility and the identity of residues is no longer visible to the network. On the other hand we know that the stereochemical structure is of importance especially in the short range interactions *e.g.* in forming the hydrogen bonds, as has been shown by Presta and Rose (1988). Another feasible explanation might be that the performance of the network is not limited by the networks ability to extract the information from the training set whatever the input coding is but rather by some other factor inherent in the approach or in the training set.

Using a neural network methods for predicting the tertiary structure of a globular protein has its limitations. Some of these are inherent to the neural network approach itself and some raise from the nature of the application. The main limitation is probably the number of currently available protein structure data. Rooman and Wodak (1988) show that some 1500 proteins with average length of 180 residues would be needed in their approach to obtain good results in structural prediction. It will take several years to collect the required structural data and even then there is no guarantee that the new proteins contain the missing infor-

mation of the sequence-structure mapping.

Also for a neural network there is always a trade-off between the memorization and the ability of generalization. The size of the network plays a major role. As a rule of the thumb (which also has theoretical foundations as described by Baum and Haussler 1989), in order to get a good degree of generalization, there should be about 10 times as many training examples in the training set as there are weights in the network. In practice this rule holds for every fitting of parameters, if they are nearly linearly independent. The smallest network we are currently using has 2540 weights (154 input nodes, 10 hidden nodes and 100 output nodes). Thus there should be about 25000 training examples. The training set with its 7000 examples is only one third of the theoretical minimum.

We conclude that the back propagation neural network has had limited success in predicting the protein backbone dihedral angles. Prediction can be quite close to the correct one over a short sequence of residues. For example α helices that are energetically the most favourable structures are often predicted in their correct places. By now it has become apparent that attempt to reproduce the whole three dimensional structure from the backbone dihedral angles will be too sensitive to errors in them. What is needed is a representation that is more robust and contains more redundancy. In this respect the distance matrix representation seems to be promising.

As a bottom line we would like to emphasize that the neural network (or in a broader context the artificial intelligence) approach is not per se unsuitable for the prediction of molecular phenomena. Our research is only taking its first steps and it may have a long way to go before it will prove its feasibility and later yield significant results.

References

Baum, E. B., and Haussler, D. 1989. What Size Net Gives Valid Generalization?. *Advances in neural information processing systems I.* , Touretzky, D.S. ed., Morgan Kaufmann Publishers, San Mateo, CA.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112: 535-542.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Lautrup, B., Norskov, L., Olsen, O.H., and Petersen, S.B. 1988. Protein secondary structure and homology by neural networks. *FEBS Lett.* 241: 223-228.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Fredholm, H., Lautrup, B., and Petersen, S.B. 1990. A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Lett.* 261: 43-46.

Cantor, C.R., and Schimmel, P.R. 1980. *Biophysical Chemistry, Part 1: The conformation of biological macromolecules.* W.H. Freeman and Company.

Clementi, E. ed. 1991. *MOTECC Modern Techniques in Computational Chemistry.* ESCOM, Leiden.

Friedrics, M.S., and Wolynes, P.G. 1989. Toward Protein Tertiary Structure Recognition by Means of Associative Memory Hamiltonians. *Science* 246: 371-373.

Holley, L.H., and Karplus, G. 1989. Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. U.S.A.* 86: 152-156.

Kabsch, W., and Sander, C. 1983. How good are predictions of protein secondary structure?. *FEBS Lett.* 155: 179-182.

Lambert, M.H., and Scheraga, H.A. 1989a. Pattern Recognition in the Prediction of Protein Structure. I. Tripeptide Conformational Probabilities Calculated from the Amino Acid Sequence. *Journal of Computational Chemistry* 10: 770-797.

Lambert, M.H., and Scheraga, H.A. 1989b. Pattern Recognition in the Prediction of Protein Structure. II. Chain Conformation from a Probability-Directed Search Procedure. *Journal of Computational Chemistry* 10: 798-816.

Lambert, M.H., and Scheraga, H.A. 1989c. Pattern Recognition in the Prediction of Protein Structure. III. An Importance-Sampling Minimization Procedure. *Journal of Computational Chemistry* 10: 817-831.

Mejia, C., and Fogelman-Soulie, F., 1990. Incorporating knowledge in multilayer networks: The example of proteins secondary structure prediction. Rapport de recherche, Laboratoire de Recherche en Informatique, Univ. de Paris-sud.

Momany, F.A., McGuire, R.F., Burgess, A.W., and Scheraga, H.A. 1975. Energy Parameters in Polypeptides. VII. Geometric Parameters, Partial Atomic Charges, Nonbonded Interactions, Hydrogen Bond Interactions, and Intrinsic Torsional Potentials for the Naturally Occurring Amino Acids. *The Journal of Physical Chemistry* 79: 2361-2381.

Presta, L.G., and Rose, G.D. 1988. Helix Signals in Proteins. *Science* 240: 1632-1641.

Qian, N., and Sejnowski, T.J. 1988. Predicting the Secondary Structure of Globular Proteins Using Neural Nelwork Models. *J. of Mol. Biol.* 202: 865-884.

Rooman, M.J., and Wodak. S.J. 1988. Identification of predictive sequence motifs limited by protein structure data base size. *Nature* 335: 45-49.

SunSoft 1991. OpenWindows Developer's GUIDE 3.0, User Guide. SunSoft, Mountain View, CA.

Wilcox, G.L., and Poliac, M.O. 1989. Generalization of Protein Structure From Sequence Using A Large Scale Back propagation Network. Proceedings of the International Joint Conference on Neural Networks, IJCNN 1989, Washington, D.C.