

G4.1 A neural network for prediction of protein secondary structure

Burkhard Rost

Abstract

Currently, the prediction of a three-dimensional protein structure from a protein sequence poses insurmountable difficulties. As an intermediate step, a much simpler task has been pursued extensively: predicting one-dimensional strings of secondary structure. Here, a composite neural network is described which predicts three secondary-structure states (helix, strand, loop). The network system comprises two levels of feedforward networks (one hidden layer each) and a final jury decision over differently trained networks. Training is done by an adaptive-like backpropagation. An important key feature of the system is that the input is not only the sequence of one protein but the profile of a set of sequences from proteins which have the same three-dimensional structure. The combination of the problem-specific topology and the preprocessing of the input improve prediction accuracy from 62% to 72%. Furthermore, the specific topology and training procedure successfully correct for shortcomings of both simpler neural network and classical methods. Over the last few years, the network system has been the best automatic predictor in a very competitive area of research.

G4.1.1 Introduction to protein structure prediction

G4.1.1.1 Protein folding

Proteins are formed by joining amino acids into a long stretched chain, the protein sequence. They differ in length (from 30 to 30000 amino acids) and in the arrangement of the amino acids (called residues, when joined in proteins). In water, the chain folds into a unique three-dimensional structure. The main driving force for folding is the need to pack residues for which a contact with water is energetically unfavorable (hydrophobic residues) into the interior of the molecule. This is only possible if the protein forms regular patterns of a macroscopic substructure called secondary structure (figure G4.1.1); for an introduction see Brändén and Tooze (1991).

G4.1.1.2 Sequence-structure gap

Today the sequence is known for more than 40000 proteins (Bairoch and Boeckmann 1992), but the three-dimensional structures for only 3000 have been determined by crystallography (Bernstein *et al.* 1977). Large-scale gene sequencing projects increase this sequence-structure gap further (Oliver *et al.* 1992).

G4.1.1.3 Protein structure prediction

Protein three-dimensional structure determines protein function. It is well established that the three-dimensional structure is uniquely determined by the sequence (Anfinsen 1973). Thus, in principle, three-dimensional structure could be predicted from first principles. Unfortunately, the CPU time required is many orders of magnitude beyond today's scope (van Gunsteren 1993, Yun-yu *et al.* 1993). However, it is of practical importance to know the three-dimensional structure, for example, for rational drug design.

G4.1.1.4 Protein structure prediction by alignment

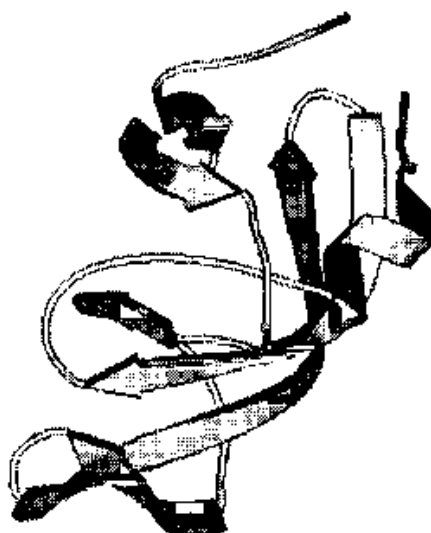
The evolutionary pressure conserves protein function. Thus, protein structure is more conserved than sequence. Evolution has created pairs of proteins which have similar structure but only 25% identical residues (Sander and Schneider 1991). Therefore, three-dimensional structure can be predicted accurately by homology if a protein with sufficient sequence identity and known three-dimensional structure is found in the databank. Homology modeling reduces the sequence-structure gap by about 10 000 proteins (Sander and Schneider 1993, Roat and Sander 1994J).

G4.1.1.5 Drastic simplification of the prediction problem

If homology modeling is not applicable, that is, for about 30 000 of the known sequences, the prediction problem has to be simplified. An extreme simplification is the prediction of one-dimensional strings of secondary-structure assignment (figure G4.1.1). One tool that has been applied to various aspects of the protein structure prediction problem is the artificial neural network (ANN) (McGregor *et al* 1989, Bengio and Pouliot 1990, Bohr *et al* 1990, Bossa and Pascarella 1990, Holbrook *et al* 1990, Kneller *et al* 1990, Petersen *et al* 1990, Brunak 1991, Friedrichs *et al* 1991, First and Sternberg 1991, Bühm *et al* 1992, Ferrin and Ferrara 1992b, Ferrin and Ferrara 1992a, Frishman and Argos 1992, Goldstein *et al* 1992a,

P	PPF		
Q	QQQV		
I	IFQVI	E	E
T	TSIVR	E	E
L	LLSTL	E	E
N	NNQED	E	E
Q	QQQAK		
R	RRRPO		
P	PPPPP		
L	VVTRF	E	E
V	VVLI	E	E
T	TTKEX	E	E
I	AALIV	E	E
K	HYKKF	E	E
I	IIIV		
G	GGNGG		
G	GGGTG		
Q	QQRRR		
L	PPLMN	E	E
K	VVTRV	E	E
E	EESK	E	E
A	VVGLG	E	E
L	LLILL	E	E
L	LLLVV	E	E
D	DDDDD		
T	TTTTT		
O	GGGGG		
A	AAAAA		
D	DDDDD		
D	DDARE		
T	SSTTV		
V	IIIVV	E	E
L	VVIVL	E	E

(a)



(b)

Figure G4.1.1. Structural representation of HIV-1 protease with PDD (a databank of proteins with known three-dimensional structure) code IIIHP (Bernstein *et al* 1977) in one and three dimensions. (a) Amino acids for the first 33 residues (one letter code, first column); alignment of five proteins with the same three-dimensional structure as HIV-1 protease (second column); secondary structure computed from three-dimensional structure using the program DSSP (dictionary of secondary structures of proteins, a program that computes secondary-structure segments from three-dimensional coordinates, Kaboch and Sander 1983a), H, strand = 1, rest = blank (third column); and a typical prediction by the neural network program (Roat and Sander 1994b) for secondary structure (in italics, fourth column). (b) The protein chain in three-dimensions is plotted schematically as a ribbon. Strands are indicated by arrows; the short helix is on the right towards the end of the protein. Graph by Christos Ouzounis (European Molecular Biology Laboratory) using the program MOLSCRIPT (Kraulis 1991).

1992b, Hayward and Collins 1992, Muskal and Kim 1992, Pancoska *et al* 1992, Xin *et al* 1992, Andrade *et al* 1993, Dubchak *et al* 1993, Fariselli *et al* 1993, Ferrán and Pflugfelder 1993, Maclin and Shavlik 1993, Metfessel *et al* 1993, Presnell and Cohen 1993, Rost and Sander 1993c, Rost and Sander 1993a, Sasagawa and Tajima 1993, Tshoumarchenko *et al* 1993, Dombi and Lawrence 1994, Radomski *et al* 1994, Rost and Sander 1994a, 1994c, Tolstrup *et al* 1994).

G4.1.2 Design process

G4.1.2.1 Motivation for a neural network solution

Even the simplified task of predicting secondary structure is a difficult problem. Thus, secondary-structure prediction became a playground to apply any fancy new pattern classification techniques, for example, neural networks (Bohr *et al* 1988, Qian and Sejnowski 1988, Holley and Karplus 1989). The hope was that neural networks could use higher-order correlation in the data. However, this failed—neural networks with and without a hidden layer were equally accurate (Holley and Karplus 1989). The motivation to try again was twofold: first, evolutionary records provide a rich resource of structural information which should contain higher orders of correlation; and second, some disadvantages of both neural network and non-neural network predictions should be correctable by alternatives to *backpropagation training* (Stolorz *et al* 1992) or composite neural networks. G4.1.2

G4.1.2.2 General description of the neural function

The task is to classify residues from a protein into three secondary-structure types. A window of a adjacent residues is taken from a protein sequence and input to the network. The output consists of three units for the secondary structure of the residue in the center of the input window. The window is shifted through the whole protein, such that a protein with R residues provides R classification examples.

G4.1.2.3 Topology

Helices extend over at least four residues; the average length of a helix is typically some ten residues. A simple neural network as described in the previous paragraph does not capture the correlation between secondary-structure states of adjacent residues. Thus, for example, the average length of a predicted helix is about four instead of ten residues. Correlations between adjacent residues can be introduced by using a second level of structure-to-structure neural network (figure G4.1.2). Such a second level of neural network improves overall prediction accuracy only marginally (Qian and Sejnowski 1988), but the average length of predicted secondary-structure segments is more similar to observed averages than for the first-level sequence-to-structure neural network (Rost and Sander 1992, 1993b, 1994b). A further difficulty with a simple neural network is that different training procedures result in different predictions. Which one to take? A simple solution is to compute an arithmetic average over differently trained networks (jury decision or committee machine, Hansen and Salamon 1990). Such a third level improves overall accuracy and tends to combine the advantages of differently trained networks.

G4.1.3 Training methods

G4.1.3.1 Balanced training

Neural networks trained by backpropagation (Rumelhart *et al* 1986) in an on-line mode (updated for each training pattern) typically result in a three-state accuracy of around 62% (Rost and Sander 1993b, Rost *et al* 1993). The accuracy is very unbalanced between the three secondary-structure types (helix 56%, strand 41%, loop 76%). This reflects the typical distribution of secondary structure in the data set: 32% helix, 21% strand, 47% loop (Rost and Sander 1992, Rost and Sander 1994a). A simple way to balance the prediction and thus to more accurately predict the most abundant class of strand is an adaptive-like training: instead of choosing the training samples at random from all examples, now at each time step an example is chosen at random from each of the three classes (helix, strand, loop):

$$\Delta W_{ij}(t+1) = -\epsilon \frac{\partial E_{\text{min}}(t)}{\partial W_{ij}} + \alpha \Delta W_{ij}(t-1)$$

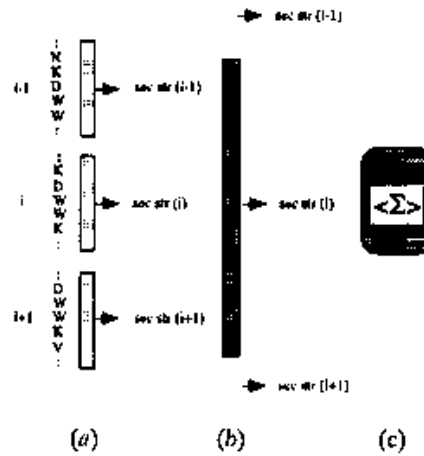


Figure G4.1.2. Three-level system for prediction of secondary structure. (a) First level, sequence-to-structure network: a window of $n = 13$ adjacent residues is shifted through all proteins. For each window the task of the network is to predict the secondary structure state of the central residue (D, W, W). Neural network: unidirectional connections. Number of units (see figure G4.1.4): $N_1 = 536$, $N_2 = 15$, $N_3 = 3$. (b) Second level, structure-to-structure network: a window of $n = 17$ adjacent residues is shifted through all proteins. Again the task is to predict the secondary structure for the central residue. But now the input are the output values (i.e. the predictions) of the first-level network (as shown, the second level predicts the secondary structure for W at position i). Neural network: unidirectional connections. Number of units (see figure G4.1.3): $N_1 = 627$, $N_2 = 15$, $N_3 = 3$. (c) Third level, jury decision: the output from differently trained networks (figure G4.1.4) for the same sequence position is summed. The secondary-structure prediction for residue W at sequence position i is assigned to the unit with the maximal sum.

with the learning rate ϵ (set to 0.05), the momentum term ν (set to 0.2), the algorithmic time t , and error E_{sum} :

$$E_{\text{sum}} = \sum_{\mu=1}^3 \sum_{k=1}^3 (a_k^{\mu} - d_k^{\mu})^2 \quad (\text{G4.1.1})$$

where a_k^{μ} is the value of output unit k (helix, $k = 1$; strand, $k = 2$; loop, $k = 3$) for pattern μ , and d_k^{μ} the desired value for unit k (e.g. for $k = 1$ and $\mu = 1$, i.e. the first output unit of the helix example: $d = 1$ if the central residue of pattern μ is in helix, and $= 0$ otherwise). The three patterns μ are chosen such that, for example, $\mu = 1$ represents a helix, $\mu = 2$ a strand, and $\mu = 3$ a loop. Training is stopped when the accuracy has reached 76%. This empirical value reflects a flat curve for overtraining; that is, stopping at values of 76–85% resulted in only marginal differences in terms of generalization. Such a training results in a more balanced prediction accuracy (helix 59%, strand 58%, loop 61%).

G4.1.3.2 Training and testing set

To evaluate the *generalization performance*, multifold cross-validation experiments have to be performed: the data set containing 126 proteins is split into seven partitions 108 + 18 proteins. The 108 are used for training, the 18 for testing. This is repeated seven times (i.e. seven neural networks are trained independently) until each protein has been used once for testing. Two problem-specific constraints are imposed on the data set. First, sequence similarity between any two proteins used has to be lower than 25% (Sander and Schneider 1991), as above 25% sequence identity homology modeling is applicable and is clearly superior to any *ab initio* prediction; Rost *et al.* (1994b). Second, the size of the set should be sufficiently large as prediction accuracy differs between proteins (Rost and Sander 1993a, Rost *et al.* 1993). Sets are taken from PDB, the databank of known three-dimensional structures (Bernstein *et al.*

1977). Currently, there are more than 200 unique proteins of known three-dimensional structure with more than 60000 residues (i.e. patterns) in total (Hobohm and Sander 1994). Secondary structure can be compiled automatically from three-dimensional structure and is stored in databases such as DSSP (Kabsch and Sander 1983a) or HSSP (a database of the homology-derived structures of proteins, Sander and Schneider 1993).

G4.1.4 Input preprocessing

G4.1.4.1 Input coding, single sequences

Each residue is coded by 20 input units for 20 different amino acids. Binary coding (19 units = 0; one unit = 1) is as good as or better than any alternative coding scheme (Cherkauer and Shavlik 1993, Rost 1993, Rost and Sander 1993b, Maza 1994). To allow the first and last residues of a protein to be used as the central residue in a window, an additional 21st input unit is used as a spacer.

G4.1.4.2 Input coding, multiple alignment profiles

The elaborated neural network system described so far is still limited to a performance accuracy of about 65%. The input information is not sufficient. As stated above, naturally evolved proteins can exchange about 75% of their residues without changing the three-dimensional structure. Such evolutionary information is highly specific for three-dimensional structure (figure G4.1.3) and can thus be used for prediction (Dickerson *et al* 1976, Maxfield and Scheraga 1979, Zvelebi *et al* 1987). Profiles of evolutionary exchanges are taken from HSSP, a database of homology-derived predictions (Sander and Schneider 1993).

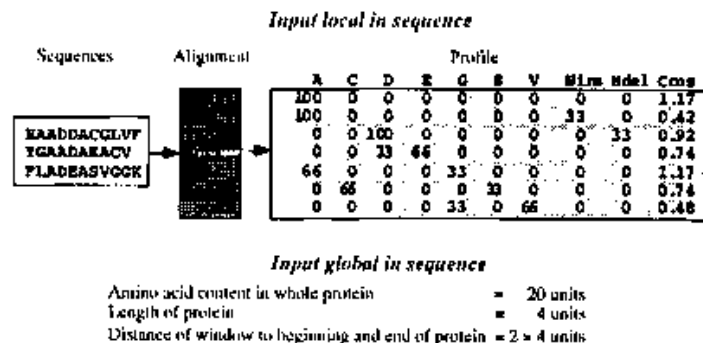


Figure G4.1.3. Preprocessing input data. First, a protein is taken from PDB (Bernstein *et al* 1977), then proteins with similar sequence are searched in SWISSPROT (a databank of known protein sequences, Bairoch and Boeckmann 1992). For naturally evolved proteins it is possible to select proteins of homologous three-dimensional structure purely on the basis of sequence identity (Sander and Schneider 1991). Homologues (three here) are aligned with the alignment program MAXHOM (Sander and Schneider 1991). At each residue position the occurrence (percentage) of each amino acid (given in one-letter code) is compiled along with the number of insertions (Ins) and deletions (Del) necessary to render an optimal alignment. Such a profile is fed as input into the neural network, instead of just the sequence of the first protein. Acids E and D are mutually more similar in terms of their biochemical properties than E and C. The conservation weight (Cons) reflects the degree of similarity of the residues found at a particular position of the alignment (Rost and Sander 1993b). In addition to the information locally available from, for example, 13 adjacent residues, global information can be compiled, such as the content of each amino acid in the whole protein, the length of the protein, or the distance of the window from the beginning and end of the protein.

G4.1.4.3 Further preprocessing of input

Alignments of homologous proteins contain further details (figure G4.1.3). First, the more insertions and deletions necessary to render an optimal alignment the more likely this region occurs in a loop. Second, consecutive stretches of high conservation of physicochemical properties of exchanged amino acids often indicate the presence of either a helix or a strand. Third, the amino acid composition of the whole protein is specific for certain types of proteins (e.g. all-helical proteins). Information about the protein class (e.g. all-helical) can improve prediction accuracy further (Knoller *et al.* 1990); however, in practice this marginal gain is lost by the inaccuracy in predicting the class (Rost and Sander 1993c).

G4.1.5 Output Interpretation

G4.1.5.1 Jury decision over various neural networks

The final output of the composite neural network is an arithmetic average over 12 second-level structure-to-structure neural networks (figure G4.1.2) which differ both in the training method and the input preprocessing (figure G4.1.4).

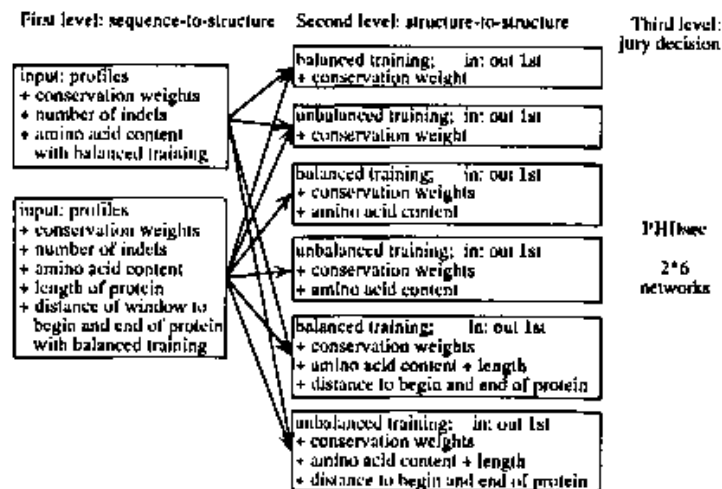


Figure G4.1.4. Generating different networks for jury decision. The final prediction of the composite neural network is an arithmetic average (jury decision) over 12 different neural networks. The neural networks differ in training procedure (unbalanced and balanced training (see section G4.1.3)) and different preprocessing of the evolutionary information (see section G4.1.4), both in the first- and second-level neural networks (figure G4.1.2).

G4.1.5.2 Output to prediction

The final prediction is derived by a winner-take-all decision, that is, the unit with the largest sum after the jury decision is chosen as the neural network prediction. An additional filtering is applied: helices shorter than three and strands shorter than two residues are elongated or interpreted as loops, depending on the strength of the prediction. The final composite neural network using evolutionary information as input—dubbed PHDsec, a profile neural network system from Heidelberg, Germany, for prediction of secondary structure—has an expected overall accuracy greater than 72% (Rost and Sander 1994b).

G4.1.5.3 Reliability index

The strength of the prediction correlates with prediction accuracy. An empirically reasonable index for the reliability of the prediction is

$$RI = INT(10 \times (a_{max} - a_{next})) \quad (G4.1.2)$$

where a_{max} is the output value of the output unit with highest value and a_{next} that of the unit with the next highest value. The factor 10 normalizes RI to integer values from 0 to 9.

G4.1.6 Comparison with traditional methods

G4.1.6.1 Neural network versus traditional predictions of secondary structure

Prediction accuracy, direct comparison from literature. Predictions of neural networks have been reported to yield a three-state prediction accuracy of better than 66% (Zhang *et al.* 1992). This is comparable to non-neural network methods (Biou *et al.* 1988, Munson *et al.* 1994) as shown in table G4.1.1. Predictions using multiple alignment information as input are, in general, significantly more accurate than those using single sequences only (table G4.1.2). For most methods the comparisons are problematic, as results are based on different evaluation sets, and most data sets used were too small or contained proteins of significant pairwise sequence identity (see table G4.1.3). For example, a simple neural network, if evaluated on 126 unique proteins, scores at some 62% accuracy (Rost and Sander 1993b), and at greater than 64% if evaluated on 15 proteins with homologies to the training set (Qian and Sejnowski 1988). For an appropriate comparison the accuracy has to be evaluated on identical, sufficiently large, and unique data sets.

Prediction accuracy, identical data sets. Laborious comparisons based on identical data sets have revealed two results. First, the composite neural network PHDsec is clearly superior to any other prediction method published so far. Second, comparisons have to be based on identical data sets; for example, for a 'favorable' data set (such as used by Levin *et al.* 1994) prediction accuracy PHDsec had an accuracy of about 75% (see also the comparison between Biou *et al.* 1988 in table G4.1.2 and in table G4.1.3).

G4.1.6.2 Specific improvements of the network system PHDsec

Improvements on the network side. The composite neural network improves performance in three ways (Rost and Sander 1994b). First, balanced training (see section G4.1.3) yields more accurate strand G4.1.3 predictions than most traditional methods (exception Gascuel and Colnard 1988). Second, the second-level structure-to-structure neural network (figure G4.1.2) results in more protein-like predictions than most G4.1.5 published traditional methods. Third, the final jury average (see section G4.1.5) improves overall accuracy by about one to two percentage points, and finds a compromise between unbalanced (overall more accurate) and balanced (strands more accurate) neural networks. The latter improvement is comparable to classical 'joint prediction methods' (Biou *et al.* 1988, Nishikawa and Noguchi 1991, Viswanadhan *et al.* 1991).

Improvements by using biological information. Using only profiles as input improves prediction accuracy by more than five percentage points (table G4.1.2). The composite neural network successfully uses further important input information. For all steps of adding relevant input information, the composite neural network has, so far, outperformed traditional methods (table G4.1.2).

G4.1.6.3 Practical impact of the neural network system PHDsec

How good is the prediction for a protein of unknown three-dimensional structure? Prediction accuracy varies with the protein, thus the expected prediction accuracy of PHDsec is $72 \pm 9\%$ (one standard deviation). This implies that users cannot deduce from the prediction whether it is 45% or 95% correct. Here, the definition of a reliability index (equation (G4.1.2)) proves to be of immense practical importance as it correlates with prediction accuracy; that is, residues predicted with higher reliability are on average predicted more accurately. Comparable indices exist for traditional methods but the composite neural network is significantly more accurate: half the residues are predicted at an expected accuracy of 88% (Rost and Sander 1994b).

How can the neural network predictions be obtained? Predictions from the composite neural network system PHDsec are available via a fully automatic prediction service (Rost *et al.* 1994a). The user sends a sequence or an alignment and the prediction is returned. (Send the word 'help' by electronic mail

Table G4.1.1. Secondary structure prediction accuracy (from the literature). Methods are abbreviated as in the reference list ('Rost and Sander 1993'—reference) is a simple neural network used as reference point for the performance on a large unique data set). All methods given use single sequences as input. Abbreviations used: 'accuracy', percentage of correctly predicted residues in three states; 'number of proteins', number of proteins used for evaluation; 'unique set', a set allowing for pairwise sequence identity greater than 25% is dubbed 'not unique'. For more recent methods more than 100 proteins is a sufficiently large data set. KS, Kabach and Sander (1983b); subKS, subset of KS; QS, Qian and Sejnowski (1988) (unfortunately this completely inadequate set allowing for pairwise identities greater than 50% is widely used); subQS, subset of QS; RS, globular proteins of Rost and Sander (1993b).

Method	Accuracy	Number of proteins	Unique set?
Non-neural network predictions			
Asai <i>et al.</i> (1993)	66.0	120	?
Bisio <i>et al.</i> (1988)	65.5	62 ^{RS}	yes
Carroll <i>et al.</i> (1991)	61.0	53 ^{subKS}	yes
Casacel and Giolandi (1988)	58.7	62 ^{RS}	yes
Geourjon and Deléage (1994)	69.0	239	no
King and Sternberg (1990)	60.0	18	yes
Leng <i>et al.</i> (1994)	68.2	74	no
Mitsun <i>et al.</i> (1994)	65.9	67	no
Nishikawa and Noguchi (1991)	64.8	27	yes
Salzberg and Cost (1992)	65.1	128	no
Viswanathan <i>et al.</i> (1991)	64.0	45	?
Yi and Lander (1993)	68.0	110	no
Neural network predictions			
Fariselli <i>et al.</i> (1993)	64.0	62	yes
Fogelman-Soulé and Mejia (1990)	58.8	62 ^{RS}	yes
Hulley and Karpur (1989)	63.2	14 ^{subQS}	yes
Kneller <i>et al.</i> (1990)	65.0	103 ^{QS}	no
Maclin and Shavlik (1993)	63.4	106 ^{QS}	no
Qian and Sejnowski (1988)	64.3	14 ^{subQS}	no
Rost and Sander (1994) reference	62.1	126 ^{RS}	yes
Sasagawa and Tajima (1993)	60.1	29	yes
Stoloz <i>et al.</i> (1992)	64.4	14 ^{subQS}	no
Zhang <i>et al.</i> (1992)	63.1	107	no
Zhang <i>et al.</i> (1992)	66.4	107	no

Table G4.1.2. Prediction accuracy for alignment-based methods (from the literature): all methods given use multiple alignments as input and are evaluated on unique data sets. Only the PHDx methods use neural networks. The following abbreviations indicate different stages of input preprocessing (section G4.1.4): PHD0, alignment profiles; PHD1, PHD0+ conservation weight; PHD2, PHD1 + insertions and deletions; PHDsec, PHD2 + amino acid content. The following data sets are labeled to indicate identical sets: LPAG, Levin *et al.* (1993); RS, Rost and Sander (1993b); and superRS, a super set of RS = RS + RS2 (Rost and Sander 1994b). Further abbreviations used as in table G4.1.1.

Method	Accuracy	Number of proteins
Rost and Sander (1994) reference	62.1	126 ^{RS}
Bowcott <i>et al.</i> (1993)	64.0	31
Levin <i>et al.</i> (1993)	68.5	60 ^{LPAG}
Rost and Sander (1994)—PHD0	69.7	126 ^{RS}
Rost and Sander (1994)—PHD1	70.8	126 ^{RS}
Rost and Sander (1994)—PHD2	71.41	26 ^{RS}
Rost and Sander (1994)—PHDsec	72.1	250 ^{superRS}
Wako and Blundell (1994)	69.0	13
Zweibel <i>et al.</i> (1987)	66.1	11

Table G4.1.3. PHDsec versus other methods evaluated on identical data sets: abbreviations used as in table G4.1.1 and table G4.1.2. For comparison results on set RS are given.

Method	Accuracy	Number of proteins
Chou and Fasman (1974)	49	62 ^{KS}
Gasuel and Colnard (1988)	58.7	62 ^{KS}
Rost and Sander (1994)—PHD1	72.5	62 ^{KS}
Rost and Sander (1994)—PHD1	70.81	26 ^{KS}
Levin <i>et al.</i> 1993	68.5	60 ^{L₁WQ}
Rost and Sander (1994)—PHD2	74.6	60 ^{L₁WQ}
Rost and Sander (1994)—PHD2	71.4	126 ^{KS}
Gibrat <i>et al.</i> (1987)	58.9	124 ^{RS2}
Biou <i>et al.</i> (1988)	60.9	124 ^{RS2}
Rost and Sander (1994)—PHDsec	72.5	124 ^{RS2}
Rost and Sander (1994)—PHDsec	71.6	126 ^{KS}

to the internet address 'PredictProtein@EMBL-Heidelberg.de', or use the WWW site 'http://www.embl-heidelberg.de/predictprotein/predictprotein.html'.) Both improved prediction accuracy and rigorous testing procedures have led to about 100 prediction requests per day.

G4.1.7 Conclusions

Neural networks can easily be tailored to the problem. The three improvements on the network side (see above) illustrate that a deeper understanding of the stochastic behavior of the 'black-box pattern classifier neural network' can be used to avoid problem specific disadvantages of a simple neural network.

Highest gain from preprocessing input data by biological expertise. It is not enough to tailor the composite network system to the problem. Instead, the most significant improvement of the prediction accuracy stems from the incorporation of biological knowledge (evolutionary information).

Composite system superior to any other prediction method. Often neural networks are shown to be the second-best solution of a problem. The composite Neural network described here, today, is clearly better than any other prediction method. Further improvements of the method appear possible. Thus, the neural network for secondary-structure prediction is likely to remain one of the best tools in a very competitive field of research.

Appropriate evaluation and availability of methods is the key to applications. Most methods developed in the field of 'biocomputing' rely upon time-consuming literature searches (step 1), appropriate testing procedures (step 2) and making the program available (step 3). However, theoretical tools for the prediction of protein structure can influence research in molecular biology only if these simplifications are avoided.

Perspectives for the future? The goal is to predict protein three-dimensional structure. The explosion of protein databases may bring this goal in reach in the near future. Neural networks have a fair chance to be part of a hybrid system that will first predict three-dimensional structure. But even if one heads for less ambitious projects, there are many problems for which sufficiently tested, available neural network solutions would be highly welcomed by experimentalists.

References

- Andrade M A, Chacón P, Merelo J J and Morán F 1993 Evolution of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network *Protein Eng.* **6** 383–90
- Aufusner C B 1973 Principles that govern the folding of protein chains *Science* **181** 223–30
- Barooh A and Boeckmann B 1992 The SWISS-PROT protein sequence data bank *Nucleic Acids Res.* **20** 2019–22
- Bengio Y and Pouhot Y 1990 Efficient recognition of immunoglobulin domains from amino acid sequences using a neural network *Comput. Appl. Biol. Sci.* **6** 319–24
- Berstein F C, Koetzle T F, Williams G J B, Meyer J F, 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** 535–42
- Biou V, Gibrat J F, Levin J M, Robson B and Garnier J 1988 Secondary structure prediction: combination of three different methods *Protein Eng.* **2** 185–91

- Böhan G, Muhr R and Jaenicke R 1992 Quantitative analysis of protein far UV circular dichroism spectra by neural networks *Prot. Eng.* **5** 191-5
- Bohr H, Bohr J, Brunak S, Cotterill R M J, Lautrup B, Nørskov L, Olsen O H and Petersen S B 1988 Protein secondary structure and homology by neural networks *FEBS Lett.* **241** 223-8
- Bohr H, Bohr J, Brunak S, 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** 43-6
- Bossa F and Pascarella S 1990 PRONET: a microcomputer program for predicting the secondary structure of proteins with a neural network *Comput. Appl. Biol. Sci.* **5** 319-20
- Brändén C and Tooze J 1991 *Introduction to Protein Structure* (New York: Garland)
- Brunak S 1991 Non-linearities in training sets identified by inspecting the order in which neural networks learn *Neural Networks From Biology to High Energy Physics* ed O Benhar, C Bosio P Del Giudice and E Tabet (Italy: Elba) pp 277-88
- Cherkauer K J and Shavlik J W 1993 Protein Structure Prediction: Selecting Salient Features from Large Candidate Pools *Proc. First Int. Conf. on Intelligent Systems for Molecular Biology* (Bethesda, MD: AAAI Press) in press
- Dickerson R E, Timkovich R and Almasy R J 1976 The cytochrome fold and the evolution of bacterial energy metabolism *J. Mol. Biol.* **100** 473-91
- Domhi G W and Lawrence J 1994 Analysis of protein transmembrane helical regions by a neural network *Prot. Sci.* **3** 557-66
- Dubchak I, Holbrook S R and Kim S-II 1993 Prediction of protein folding class from amino acid composition *Prot.: Struct. Func. Gen.* **16** 79-91
- Fariselli P, Compiani M and Casadio R 1993 Predicting secondary structures of membrane proteins with neural networks *Eurp. Biophys. J.* **22** 41-51
- Ferrán E and Ferrara P 1992a Clustering proteins into families using artificial neural networks *Comput. Appl. Biol. Sci.* **8** 39-44
- 1992b A neural network dynamics that resembles protein evolution *Physica* **185A** 395-401
- Ferrán E A and Pflugfelder B 1993 A hybrid method to cluster protein sequences based on statistics and artificial neural networks *Comput. Appl. Biol. Sci.* **9** 671-80
- Friedrichs, M S, Goldstein R A and Wolynes P G 1991 generalized protein tertiary structure recognition using associative memory Hamiltonians *J. Mol. Biol.* **222** 1013-34
- Frisman D and Argos P 1992 Recognition of distantly related protein sequences using conserved motifs and neural networks *J. Mol. Biol.* **228** 951-62
- Gascuel O and Golmard J L 1988 A simple method for predicting the secondary structure of globular proteins: implications and accuracy *Comput. Appl. Biol. Sci.* **4** 357-65
- Goldstein R A, Luthey-Schulten Z A and Wolynes P G 1992a Optimal protein-folding codes from spin-glass theory *Proc. Natl Acad. Sci.* **89** 4918-22
- 1992b Protein tertiary structure recognition using optimized Hamiltonians with local interactions *Proc. Natl Acad. Sci.* **89** 9029-33
- Hansen L K and Salamon P 1990 Neural Network Ensembles *IEEE Trans. Patt. Anal. Machine Intell.* **12** 993-1001
- Hayward S and Collins J F 1992 Limits on α -helix prediction with neural network models *Proteins* **14** 372-81
- Hirst J D and Sternberg M J E 1991 Prediction of ATP-binding motifs: a comparison of a perceptron-type neural network and a consensus sequence method *Prot. Eng.* **4** 615-23
- Hohohm U and Sander C 1994 Enlarged representative set of protein structures *Prot. Sci.* **3** 522-4
- Holbrook S R, Muskal S M and Kim S-II 1990 Predicting surface exposure of amino acids from protein sequences *Prot. Eng.* **3** 659-65
- Holley H L and Karplus M 1989 Protein secondary structure prediction with a neural network *Proc. Natl Acad. Sci.* **86** 152-6
- Kabsch W and Sander C 1983a Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features *Biopolymers* **22** 2577-637
- 1983b how good are predictions of protein secondary structure? *FEBS Lett.* **155** 179-82
- Kneller D G, Cohen F E and Langridge R 1990 Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network *J. Mol. Biol.* **214** 171-82
- Kraulis P 1991 MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures *J. Appl. Crystallogr.* **24** 946-50
- Levin J M, Pascarella S, Argos P and Garnier J 1993 Quantification of secondary structure prediction improvement using multiple alignments *Prot. Eng.* **6** 849-54
- Maclin R and Shavlik J W 1993 Using knowledge-based neural networks to improve algorithms: refining the Chou-fasman algorithm for protein folding *Machine Learning* **11** 195-215
- Muxfield F R and Scheraga H A 1979 Improvements in the prediction of protein topography by reduction of statistical errors *Biochemistry* **18** 697-704
- Muza M d I 1994 Generate, test, and explain: synthesizing regularity exposing attributes in large protein databases *27th Hawaii Int. Conf. on System Sciences* ed L Hunter (Wailea, Hawaii: IEEE Society Press) pp 123-32

- McGregor M J, Flores T P and Sternberg M J E 1989 Prediction of α -turns in proteins using neural networks *Prot. Eng.* **2** 521-6
- Metfessel B A, Saurugger P N, Connelly D P and Rich S S 1993 Cross-validation of protein structural class prediction using statistical clustering and neural networks *Proc. Sci.* **2** 1171-82
- Munson P J, Di Francesco V and Porrelli R 1994 Prediction of protein secondary structure using linear and quadratic logistic models with penalized maximum likelihood estimation *27th Hawaii Int. Conf. on System Sciences* ed L Hunter (Wailea, HI: IEEE Computer Society Press) pp 375-84
- Muskal S M and Kim S-II 1992 Predicting protein secondary structure content. A tandem neural network approach *J. Mol. Biol.* **225** 713-27
- Nishikawa K and Noguchi T 1991 Predicting protein secondary structure based on amino acid sequences *Meth. Enz.* **202** 31-44
- Oliver S *et al* 1992 The complete DNA sequence of yeast chromosome III *Nature* **357** 38-46
- Pancoska P, Blazek M and Keiderling T A 1992 Relationships between secondary structure fractions for globular proteins. Neural network analyses of crystallographic Data sets *Biochemistry* **31** 10250-7
- Petersen S B, Bohr H, Bohr J, Brunak S, Cotterill R M J, Fredholm H and Laurup B 1990 Training neural networks to analyse biological sequences *TIBTECH* **8** 304-8
- Premell S R and Cohen F E 1993 Artificial Neural Networks for Pattern Recognition in Biochemical Sequences *Ann. Rev. Biophys. Biomol. Struct.* **22** 283-98
- Qian N and Sejnowski T J 1988 Predicting the secondary structure of globular proteins using neural network models *J. Mol. Biol.* **202** 865-84
- Radomski J P, van Halbeek H and Meyer B 1994 Neural network-based recognition of oligosaccharide 1H-NMR spectra *Nature Struct. Biol.* **1** 217-8
- Rost B 1993 Neural networks and evolution—advanced prediction of protein secondary structure *Doctoral Thesis* Department of Physics and Astronomy, University of Heidelberg, Germany
- Rost B and Sander C 1992 Exercising Multi-layered Networks on Protein Secondary Structure *Neural Networks: From Biology to High Energy Physics* ed O Benhar, S Brunak, P DelGiudice and M Grandolfo (Italy: Elba) *Int. J. Neural Systems* 209-20
- 1993a Improved prediction of protein secondary structure by use of sequence profiles and neural networks *Proc. Natl Acad. Sci.* **90** 7558-62
- 1993b Prediction of protein secondary structure at better than 70% accuracy *J. Mol. Biol.* **232** 584-99
- 1993c Secondary structure prediction of all-helical proteins in two states *Prot. Eng.* **6** 831-6
- 1994a 1D secondary structure prediction through evolutionary profiles *Prot. Struct. Distance Analysis* ed H Bohr and S Brunak (Amsterdam, Oxford, Washington: IOS Press) pp 257-76
- 1994b Combining evolutionary information and neural networks to predict protein secondary structure *Proteins* **19** 55-72
- 1994c Conservation and prediction of solvent accessibility in protein families *Proteins* **20** 216-26
- 1994d
- Rost B, Sander C and Schneider R 1993 Progress in protein structure prediction? *Trends in Biochem. Sci.* **18** 120-3
- 1994a PHD—an automatic server for protein secondary structure prediction *Comput. Appl. Biol. Sci.* **10** 53-60
- 1994b Redefining the goals of protein secondary structure prediction *J. Mol. Biol.* **235** 13-26
- Rumelhart D E, Hinton G E and Williams R J 1986 Learning representations by back-propagating error *Nature* **323** 533-6
- Sander C and Schneider R 1991 Database of homology-derived structures and the structural meaning of sequence alignment *Proteins* **9** 56-68
- 1993 The HSSP data base of protein structure-sequence alignment *Nucleic Acids Res.* **21** 3105-9
- Sasagawa F and Tajima K 1993 Prediction of protein secondary structures by a neural network *Comput. Appl. Biol. Sci.* **9** 147-52
- Stolorz P, Lapedes A and Xia Y 1992 Predicting protein secondary structure using neural net and statistical methods *J. Mol. Biol.* **225** 363-77
- Tchoumatchenko I, Vissutsky F and Ganasia J-G 1993 *How to Make Explicit A Neural Network Trained to Predict Proteins Secondary Structure* ACASA, LAFORIA-CNRS, Université Paris VI, 4 Place Jussieu, 75 252 Paris, CEDEX 05, France
- Tolstrup N, Toftgård J, Engelbrecht J and Brunak S 1994 Neural network model of the genetic code is strongly correlated to the GES scale of amino acid transfer free energies *J. Mol. Biol.* submitted
- van Gunsteren W F 1993 Molecular dynamics studies of proteins *Current Opinion in Struct. Biol.* **3** 167-74
- Viswanadhan V N, Denckla B and Weinstein J N 1991 New Joint Prediction Algorithm (Q7-JASEP) Improves the Prediction of Protein Secondary Structure *Biochemistry* **30** 11 164-72
- Xin Y, Carmeli T T, Liebanan M N and Wilcox G L 1992 Use of the backpropagation neural network algorithm for prediction of protein folding patterns *Second Int. Conf. on Bioinformatics, Supercomputing and Complex Genome Analysis* ed H A Liu, J W Fickett, C R Cantor and R J Robbins (St Petersburg Beach, FL: World Scientific) pp 360-76

- Yun-yu S, Mark A E, Cun-xin W, Fuhua H, Berendsen H J and van Gunsteren W F 1993 Can the stability of protein mutants be predicted by free energy calculations? *Prot. Eng.* **6** 289-95
- Zhang X, Mesirov J P and Waltz D L 1992 Hybrid system for protein secondary structure prediction *J. Mol. Biol.* **225** 1049-63
- Zvelebil M J, Barton G J, Taylor W R and Sternberg M J E 1987 Prediction of protein secondary structure and active sites using alignment of homologous sequences *J. Mol. Biol.* **195** 957-61