International Journal of Neural Systems, Vol. 3 (Supp. 1992) 209–220
Proceedings of the Neural Networks: From Biology to High Energy Physics

EXERCISING MULTI–LAYERED NETWORKS ON PROTEIN SECONDARY STRUCTURE

Burkhard Rost*

and

Chris Sander

The quality of a multi–layered network predicting the secondary structure of proteins is improved substantially by: (i) using information about evolutionarily conserved amino acids (increase of overall accuracy by six percentage points), (ii) balancing the training dynamics (increase of accuracy for strand), and (iii) combining uncorrelated networks in a jury (increase two percentage points). In addition, appending a second level structure–to–structure network results in better reproduction of the length of secondary structure segments.

1. Validity of Protein Secondary Structure Prediction

Proteins perform important tasks in organisms, such as catalysis of biochemical reactions, recognition and transmission of signals, transportation of materials, and replication of the genetic information. A protein is determined by the sequence of amino acids (primary structure). The average length is about 300 polymer units. Exposed to a certain environment (solvent) the linear chain folds into a unique compact three dimensional (tertiary) structure which is crucial for the protein's function. Although recently proteins (called chaperones[1]) have been discovered that appear to assist folding in vivo, the basic credo that the tertiary structure is uniquely determined by the primary one[2] remains valid.[3] Due to improvements in DNA sequencing technology the number of known sequences is increasing explosively (from 25.000 in 1991 to 36.000 in 1992[4]). The number of proteins with known tertiary structure is 50 times smaller. Currently there are some 700 entries in the Protein Data Bank (PDB[5]). The difference stems from the difficulties in determining the three dimensional structure which is done typically by crystallography, or nuclear magnetic resonance. As a protein's function depends on its tertiary structure, it is important to predict the tertiary from the primary structure. A typical reason could be the intended design of a drug. However, the prediction of tertiary structures has proved to be difficult. An intermediate step can be the prediction of the secondary structure (a reduced description of the tertiary structure). The most important two regular secondary structure elements are α–helices and β–strands (Fig. 1). These

*Protein Design Group, European Molecular Biology Laboratory, Meyerhofstr. 1, 6900 Heidelberg, FRG.

elements are determined by their hydrogen bonding pattern (DSSP: Dictionary of Secondary Structure of Proteins[6]).

The most promising approach to structure prediction is the comparison with homologous proteins of known structure.[7] Particular amino acids at particular positions within a protein can be exchanged by experiment — or are exchanged by evolution — without altering the function. The evolutionary mutations which conserved function indicate that two naturally occurring proteins longer than 80 residues share the same principle tertiary structure if their mutual sequence homology is beyond 30%. A length–dependent cut–off for 'structural homology' is provided by HSSP (Homology–derived Secondary Structure of Proteins[8]). The current data bank of 700 known tertiary structures contains only about 150 protein chains with mutually less than 30% sequence identity.[9] The joint European project to sequence a complete yeast chromosome III yielded at least 182 new protein sequences, 42% of which have known function, and only 14% are homologous to proteins with known structure.[10] However, homologous proteins can differ in details which might be important for certain tasks. Moreover, the engineered exchange of one single amino acid can change the tertiary structure completely. The prediction tools are rather poor for non–artificial sequences which do not posses sufficient sequence homology to those of known structure. The further improvement of the prediction of secondary structure, is an intermediate step on the way to accurate prediction of protein tertiary structure.

2. Choosing a Database and Assessing the Quality of Predictions

Two difficulties in assessing the quality of a certain prediction algorithm can be singled out: (i) given an algorithm tested on known proteins, how reliably does it predict the secondary structure of proteins of unknown tertiary structure, and (ii) given a particular prediction on a known protein, how does one best measure its quality. (i) The first problem is very difficult to solve. This is in part because the current data bank of protein 3D structures (PDB) does not exhaustively represent the full potential variety of existing polypeptides. For example, globular soluble proteins like immunglobulins, or virusbarrels, are highly over–represented in PDB, whereas membrane proteins are under–represented. One way to address this problem, is to pick a set of protein chains without mutual homologies. (ii) Once the data set has been chosen, the problem is to define a measure for evaluating the quality of a particular prediction. For predictions of the three states α–helix (α), β–strand (β), and coil (c), the percentage of overall accuracy is measured by Q_3J(3 states):

$$Q_3 = \frac{\sum_{i=1}^{3} \text{pred}_i}{\text{total number of residues}}, \tag{1}$$

with predi being the number of residues correctly predicted to be in structure i. A simple alternative is to sum over per class percentages:

$$Q = \frac{1}{3}\sum_{i}^{3} Q_i \qquad = \frac{1}{3}\sum_{i}^{3} \frac{\text{pred}_i}{\text{obs}_i}, \tag{2}$$

where obs_i is the number of residues observed to be in structure i. Q_3 differs from Q_3 in that it tends to be lower if one class is predicted with lower accuracy.

These quotients for single residue accuracy do not fully capture the quality of a prediction method.[11] Suppose the following two predictions were to be compared:

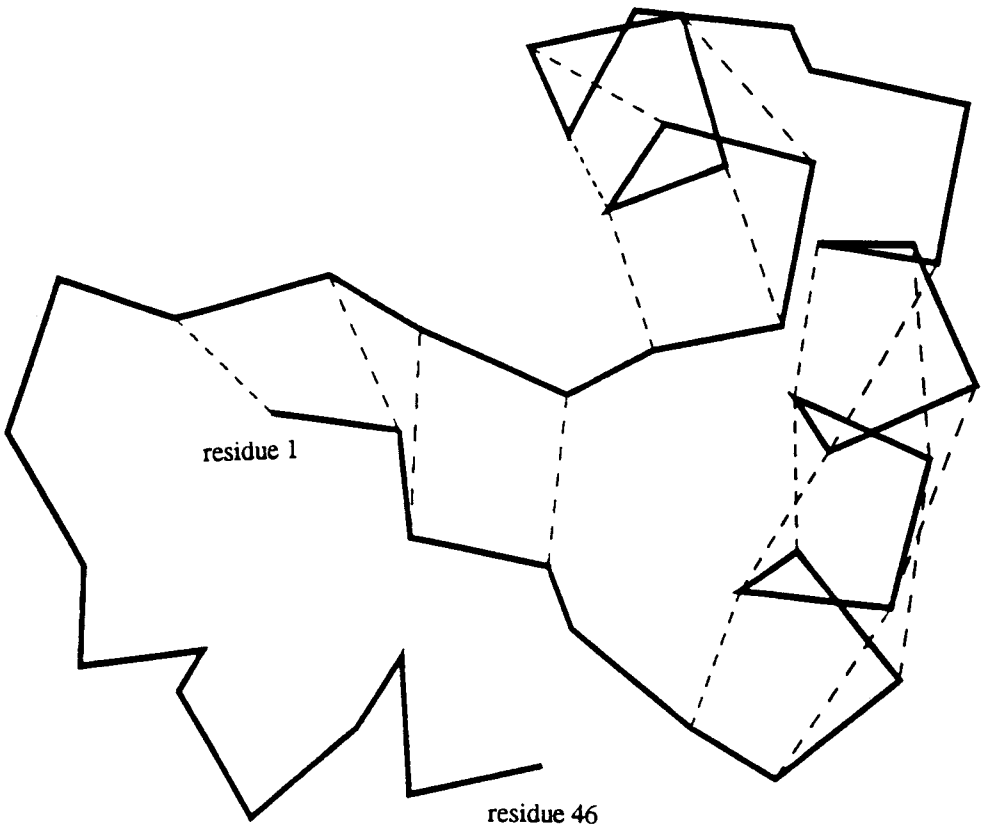

Fig. 1. Structure of crambin. Trace of the main chain C_α atoms of crambin (Abbyssinian Cabbage seed) is shown. The protein has 46 amino acids. The dotted lines connect C_α atoms of amino acids that are partners in hydrogen bonds. The first (left) four bonds connect two strands. Two helices with 12 (far right), and 10 (centre) amino acids are on the right.

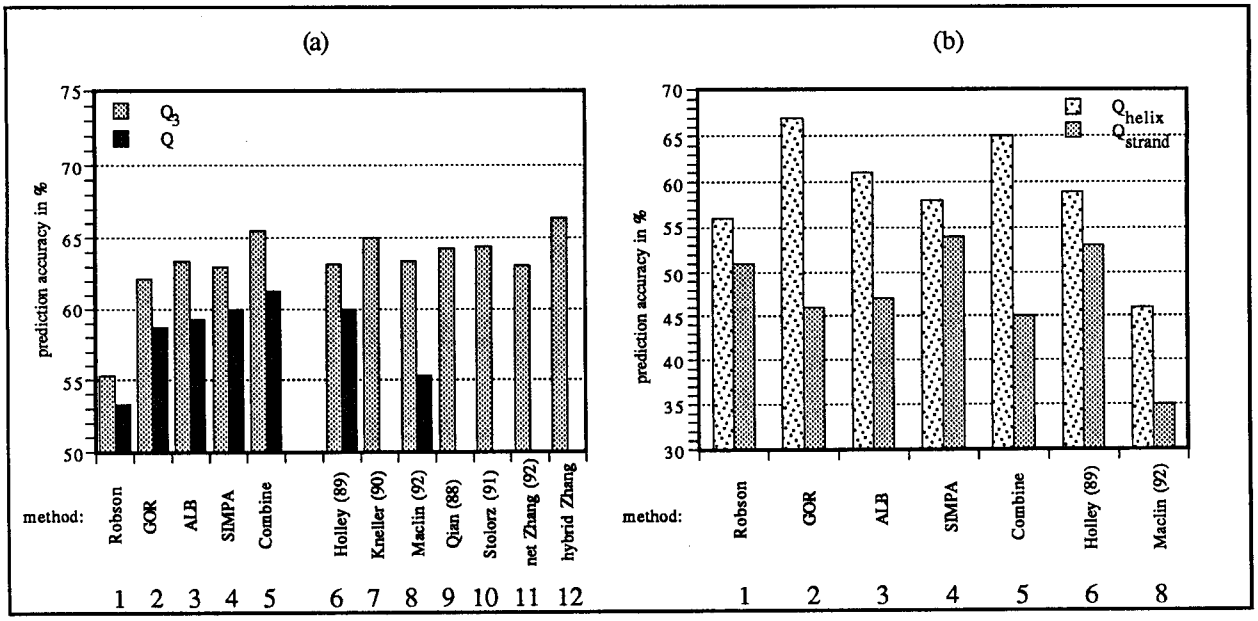

Fig. 2. Performance accuracy for previous prediction methods. Predictions by previously published methods with the following abbreviations (for the citations see 12 and 13 in the reference list): (1) Robson: Garnier, 1978; (2) GOR: Garnier, 87; (3) ALB: Ptitsyn, 83; (4) SIMPA: Levin, 88a; (5) Combine: Levin 88b, Garnier, 91; (6) — (10) quoted according to literature; (11) net Zhang: multi-layered network of Zhang, 92; (12) hybrid Zhang: "hybrid" system combining three different prediction methods, Zhang, 92. Note that only (1)–(4) where tested on the same set of 150 non-homologous chains used in this work (Altenberg, unpublished). For (5) — (12) the results are taken from the original articles, the sets of proteins used differ from the one we used (higher degree of pairwise homology). (6), (7), (9), (10) have used only one test set. Thus the latter cannot adequately be compared to the rest. (a) gives the overall percentages as defined by Eqs. 1, and 2, and (b) explicitly illustrates the per class percentages for helix and strand, which are not available for all methods.

obseved: $\alpha\alpha\alpha\alpha\alpha\alpha\alpha\alpha\alpha ccc$

prediction 1: $c\alpha\alpha\alpha c\alpha\alpha\alpha c\alpha ccc$

prediction 2: $cc\alpha\alpha\alpha\alpha\alpha\alpha\alpha\alpha\alpha c$

Although prediction 1 results in a higher Q than 2, the latter predicts better the long helix and is therefore more valuable in practice. This example illustrates that quantities like the number of predicted secondary elements, their average length, and the length distribution are important quantities to use in evaluation.

3. Two Levels of Networks and a Balanced Prediction

3.1. *First level: Sequence-to-structure — inferior to classical methods*

The first extensive analysis on the possibility of predicting secondary structure with a multi-layered perceptron was published by Qian and Sejnowski.[12] They reported a three state success quotient Q_3 of 64.3%. Further work more or less confirmed these results.[12] The emerging picture is that layered networks perform as well as (or better than) 'classical' methods[13] (Fig. 2). However, most network approaches shared two conceptual difficulties: (i) Only one single testing set was taken to evaluate the performance accuracy. The network parameters were optimized for this particular set (although this is not true for either Maclin *et al.*, or for Zhang *et al.*[12]). (ii) Between training and testing sets there were significant homologies (up to 46% for Zhang *et al.*).

Figure 3 shows the architecture of a typical feed-forward network for predicting secondary structure. From the amino acid sequence of a protein a window of length w (here 13) is cut out, and fed as the input into the net. Each amino acid is encoded by a 21 component binary vector (20 for the different amino acids, an additional one as a spacer to allow a window extending beyond the ends of a protein). The size of the output layer is determined by the number of secondary structure classes to be distinguished (here three: helix, strand, coil). The output of the network is given by three real numbers between 0 and 1. The actual prediction can be assigned by choosing the maximum of the three units. Training is done by back-propagation[14]:

$$E(\{J_1\},\{J_2\}) = \sum_{\nu=1}^{N_{\text{samples}}} \sum_{i=1}^{N_2+1} (s_i^{2,\nu} - d_i^{\nu})^2 , \tag{3}$$

ν is the index running over all N_{samples} samples, $s_i^{2,\nu}$ is the output of the network for output unit i and sample ν, d_i^{ν} the DSSP related secondary structure, and N_2 the number of output units. Surpressing the indices for the units, the output of the network is given by

$$s_2 = f\{J_2 f\{J_1 s_0\}\}$$

where s_0 is the input and f the usual sigmoid function. The learning dynamic is given by

$$J(t+1) = J(t) - \varepsilon \frac{\partial E_\nu(t)}{\partial J(t)} + \eta \frac{\partial E_\nu(t-1)}{\partial J(t-1)} , \tag{4}$$

where t is a discrete 'time' step, ε the step–width for each change, and h a inertia term. The sample index ν indicates that ΔJ is computed for the error on one pattern ν for each iteration.

After seven–fold cross validation (i.e. training seven different networks on seven distinct partitions of the data bank into testing and training set, and then averaging over the seven generalization errors) the average for such a network is $Q_3 = 61.7\%$, and $Q = 57.3\%$. Three points should be stressed: (i) The difference between performance accuracy of the best and worst performing of the seven test sets is more than five percentage points (Fig. 5(a)). This proves that choosing only one single test set is not sufficient to evaluate the generalization ability of this type of network. (ii) The network is clearly inferior in predicting strands which results in a Q being four percentage points lower than Q_3 (Fig. 5(b)). (iii) The sequence–to–structure network turns out to predict secondary structure elements with half the average length of those observed (for detailed distribution see Fig. 5).

3.2. *Second level: Structure–to–structure — improved segment reproduction*

The shortcoming of predicting too short segments can be corrected by using a second network that learns to classify consecutive strings of secondary structure obtained by the first net: (i) The first net (sequence–to–structure) generates the output from the segments of amino acids ($21 * w$ input units). It learns to classify consecutive segments of residues. (ii) The second net (structure–to–structure) computes its output from the consecutive segments of output signals received from the first net ($(3 + 1) * w$ units, the additional is again used as a spacer). The average performance for the combination of first and second net is $Q_3 = 62.6\%$ compared to 61.7% for the first alone. This marginal increase in Q_3 might have been the reason why previous authors, e.g., Qian and Sejnowski, did not emphasize the benefit of the cascade. The most relevant improvement achieved by the second network is the prediction of longer elements with average lengths: first net: $\langle L_\alpha \rangle = 4.2$, $\langle L_\beta \rangle = 2.9$; combination of first and second net: $\langle L_\alpha \rangle = 6.2$, $\langle L_\beta \rangle = 3.8$; observed: $\langle L_\alpha \rangle = 9.0$, $\langle L_\beta \rangle = 5.1$.

3.3. *Balanced prediction: Best method for strands*

Usually, the succession of training samples is chosen at random. Since the data used contains two times more coil than strand residues, the latter are learned roughly half as often (data bank: about 43% coil, 35% helix, 22% strand). An alternative is to present at each time step one helix, one strand, and one coil residue (each of these chosen at random from the stack of all segments). Such a procedure implies that a particular segment with the central residue being in a strand is presented about twice as often during the training than a particular coil segment. The success rate of the balance net is $Q = 59.7\%$ compared to 58.2% for the unbalanced net. The increase stems mainly from improved helix and strand prediction: $Q_\alpha = 58\%$ (unbalanced: 56), $Q_\beta = 59\%$ (unbalanced: 41% !) (Fig. 5(b)). Thus to a certain degree, the bads prediction of strands is simply an artifact of the prediction method applied. An inferior performance for strands might be due to the particular way in which the information contained in the data bank is extracted. It should be added, that the inferior performance for strands is partly due to the fact that the secondary structure of sheets is less local than

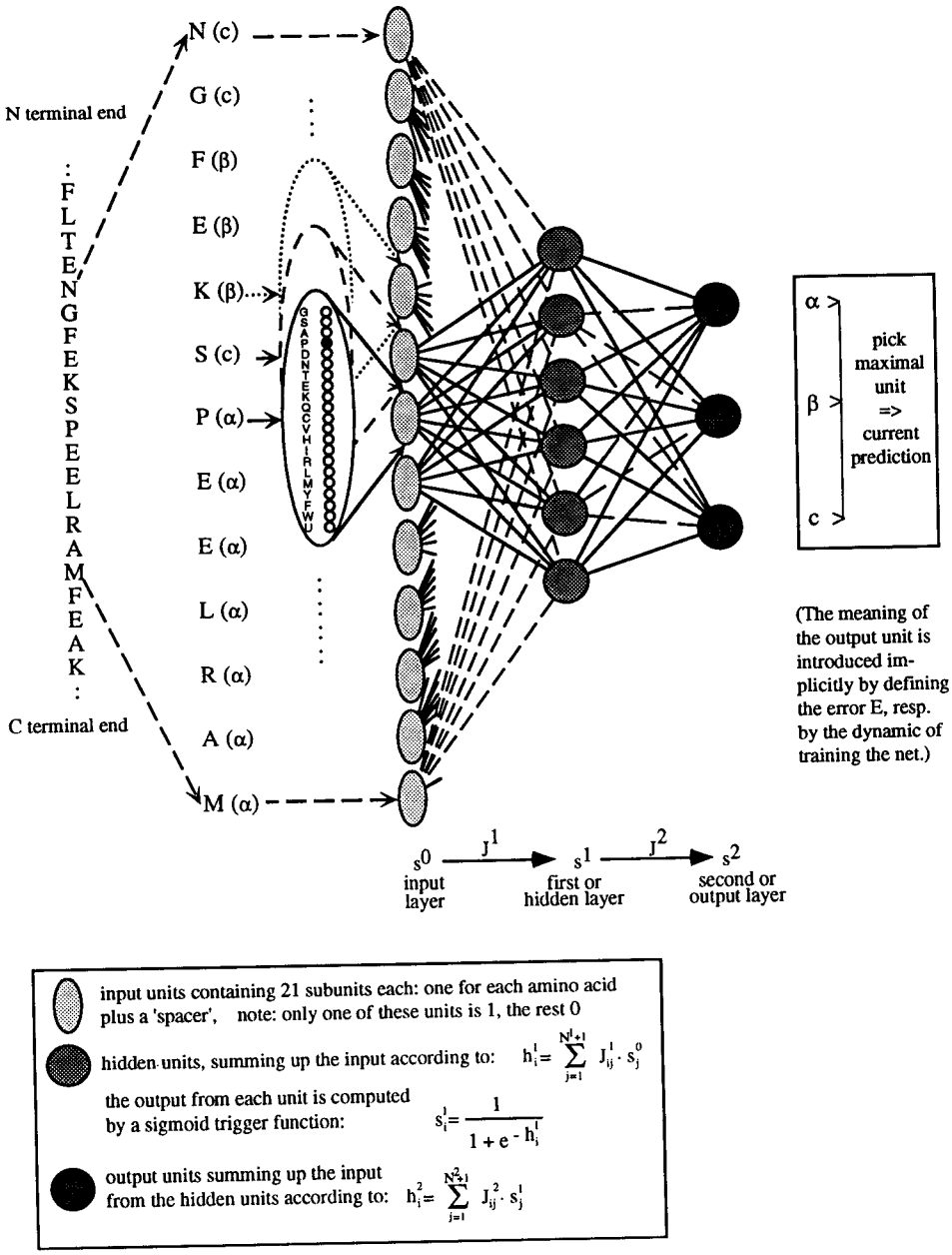

Fig. 3. Layered network for prediction. $w = 13$ successive residues are taken from a protein (here: 1rhd Rhodanese). The central Proline is known to be in a helix (the others from the first N to M: $c\,c\,\beta\,\beta\,\beta\,c\,\alpha\,\alpha\,\alpha\,\alpha\,\alpha\,\alpha\,\alpha$). Each residue contributes 21 units to the input layer. All (273) input units have a connection to each of the six hidden units. The prediction is assigned by simply taking the output unit with maximal value.

that of helices (Fig. 1).

4. Evolution Supplies Important Information

Families of homologous proteins reveal that the tertiary structure is evolutionarily more conserved than the sequence. A straightforward idea is to use this evidence by feeding —

instead of a particular residue — a profile of frequencies into a network. The frequency is given by 20 entries for the number of occurrences of each of the 20 amino acids at that position in the alignment of a protein family. Thus, the single input unit becomes a real number between 0 and 1. (Alternatively, the real number is coded by 4 binary numbers.) Otherwise, the network is the same as described above. The profile entries where not weighted by, e.g., the number of alignments. All alignments with more than 30% identical residues were included. The profiles were taken from the HSSP files.[15]

The information introduced by using family profiles proved to be highly important for secondary structure prediction. The learning is speeded up and the generalization substantially improved. For the second network Q reaches 66.8%, $Q_3 = 67.4\%$ (Figs. 5(a) and (b)).

5. Third Level: A Jury Decision — Best Current Prediction

Training the network is a walk through a relatively complex space. The gradient descent is sensitive to minor changes: it is for instance important how the initial junctions are chosen, and how the parameters for the dynamic are adjusted (such as step width ε, inertia parameter η, definition of the error E, slope and form of sigmoid decision function). A particular realization of the classification task is associated with a particular error. This error can be regarded, partly, as a random noise. Combining N_χ different architectures results in a reduction of the noise provided the networks are not completely correlated. The simplest way to combine independent networks is to compute an arithmetic average ('jury decision'):

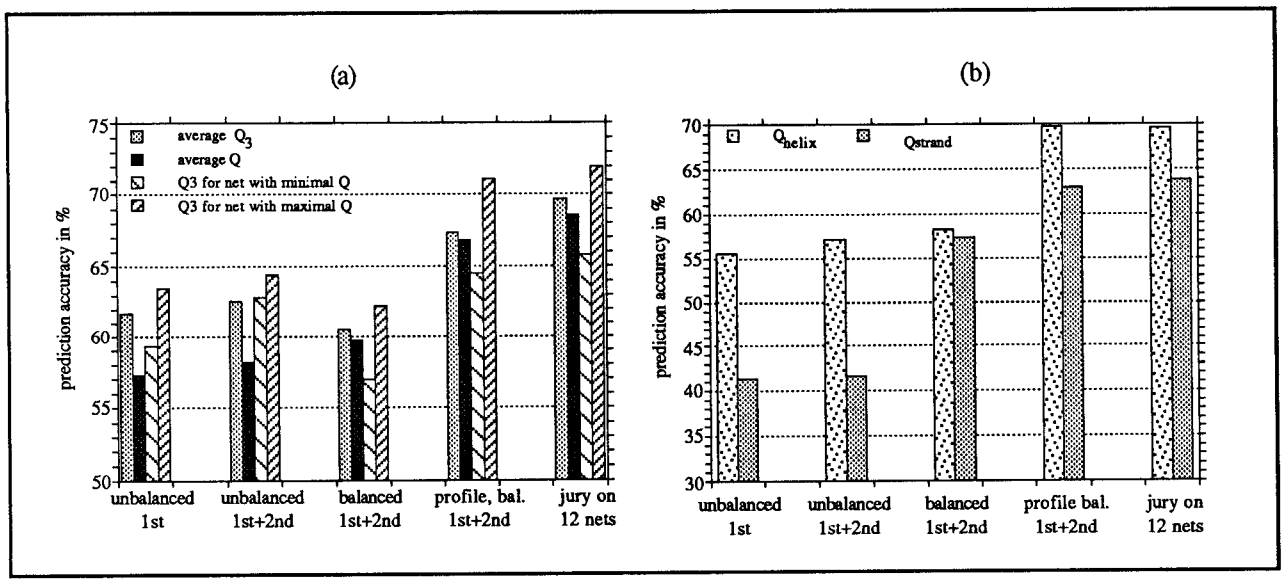

Fig. 4. Performance of various networks. The performance of the following networks is given: first level, unbalanced (described in 3.1), first and second level, unbalanced (3.2), first and second level, balanced (3.3), first and second level, balanced profile (4.), and a jury on 12 different networks (5.). (a) the overall performance as defined by Eqs. 1 and 2. Q_3 is given for those of the seven test sets with minimum, res. maximum Q, (b) per class percentages.

$$\langle s_i \rangle_\chi = \frac{1}{N_\chi} \sum_{a=1}^{N_\chi} s_i^a \qquad \forall i \tag{5}$$

s_i^a is the output of architecture a for unit i. For artificial problems the pay–off of a jury decision has been shown.[16] Different architectures were obtained by simply taking all networks that were trained when examining the influence of various changes: one and two levels, balanced and unbalanced networks and those with and without profiles.

Two questions arise: (i) Is the jury capable of 'using the best' of each particular net, i.e. is it better than the best single network? (ii) Is it advisable to extract the information contained in a particular set of chains by applying a jury decision on sub–partitions of that set? Or is it preferable to use from the beginning all available information for the training procedure? The analysis shows that the jury usually outperforms the best single network (i). However, it turns out to be advantageous to train from the beginning with the full training set available instead of splitting it and then taking the jury over the different partitions (ii). The gain by the jury is inversely proportional to the correlation between different architectures. This is fairly well to be seen by the following result:

$$\begin{aligned} \text{jury on 7 profile nets (each } Q > 65\%) \qquad &= 66.8\% \\ \text{jury on 7 profile + 2 non–profile nets (the latter with } Q \sim 60\%) \qquad &= 67.5\%\,. \end{aligned}$$

The two 'bad' non–profile nets provided valid information, probably, because their error was rather uncorrelated to the one made by the profile nets.

Further attempts to increase the performance by (i) training the jury, and (ii) trying to choose an optimal way to weight different architectures by using correlation or entropy as weight factors have not been successful yet. The simple linear average is not the best possible solution for combining the different architectures, but it proves to be hard to do better when only using the information given from the training set. The currently best result was obtained by a jury on 12 different networks: $Q = 68.6\%$, $Q_3 = 69.7\%$ (Figs. 4(a) and (b)). (The correlation coefficients as defined by Matthews[17]: $c_\alpha = 0.58$, $c_\beta = 0.50$, and $c_c = 0.50$.)

6. Summary and Perspectives

Secondary structure can be predicted by multi–layered networks, as was shown by 1988. The performance of initially published networks is worse than reported when a seven–fold cross–validation is performed on protein chains with a low level of pairwise homology. That was the reason for the skepticism of those familiar with classical prediction algorithms, and was probably part of the reason why networks did not penetrate into biological laboratories as everyday prediction tools.

However, the combination of relevant biological information (profiles of evolutionary conservation) and technical tricks (balanced training, second level network, jury) resulted in a network system that is the best current prediction tool. This network prediction outperforms previous methods in that (i) the overall accuracy is at least three percentage points beyond the second best method (Figs. 2(a) and 4(a)), (ii) the prediction for strand is superior (Figs. 2(b) and

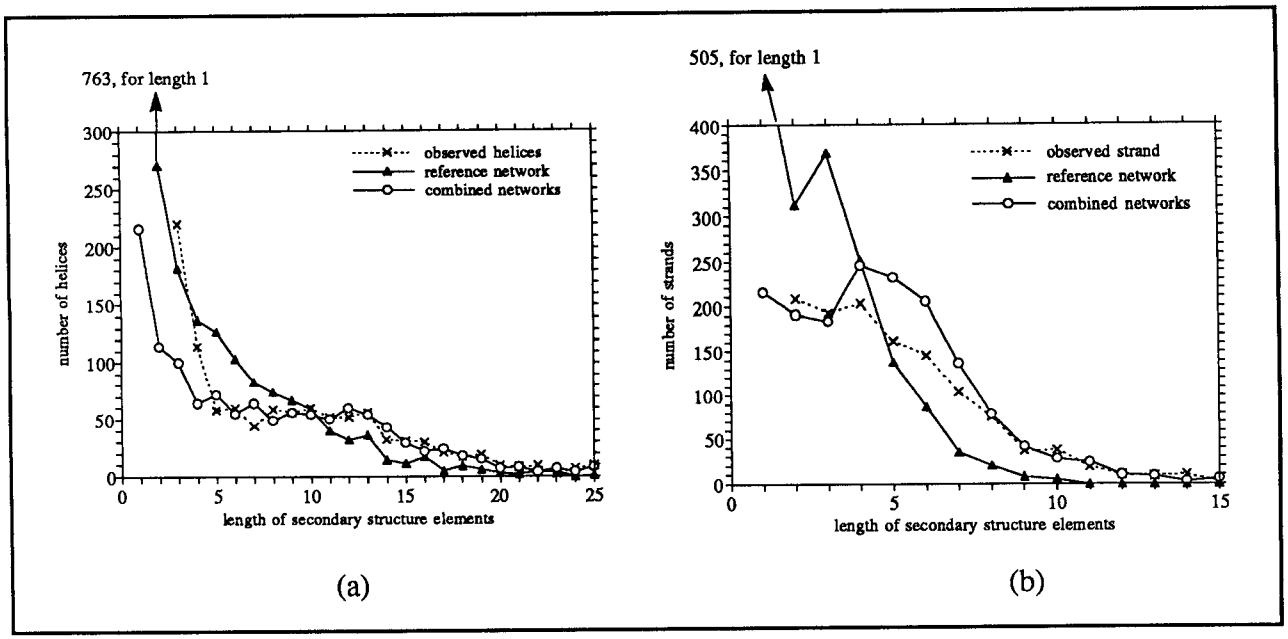

Fig. 5. Length of secondary structure segments. (a) length distribution of all predicted, res. observed helices, (b) the distribution for strands. The randomly trained first level network predicts too many short and too few long segments for both structure types. The ultimate network (here the jury on 9 different networks) better reproduces the observed distribution.

4(b)), and (iii) the extent of secondary structure segments has a more protein–like appearance (Fig. 5).

We intend to test the method on the sequences of proteins whose structure is currently being determined by crystallography or n.m.r. (nuclear magnetic resonance). Predictions can be requested by sending a protein sequence in one letter code to the EMBL file server: PredictProtein@ EMBL–Heidelberg.de.

The explosive development in large scale sequencing projects, soon, will result in an enormous flood of data on proteins. Managing making use of such an amount of information will be difficult. For some of the problems multi–layered networks or similar approaches might be of help.

The project presented here profited from an environment of vivid inter–disciplinary research. For support and valuable suggestions we should like to thank our colleagues R. Schneider and G. Vriend. A similar stimulating environment, for one of us, was the worth and importance of the workshop held in Elba. It was there that crucial encouragement for this work was given last year (thanks to S. Brunak, F. Fogelman–Soulie, and S. Solla). Thanks, as well, to the organizers, in particular to O. Benhar and P. del Giudice, respectively the secretaries (B. Ceccarelli, P. Di Ciaccio, G. Monteleone) who enabled once again the communication between such widespread topics. We look forward to the next opportunity to return from that island, with the back pack full of stimulation and ideas.

References

1. T. J. P. Hubbard and C. Sander, "The role of heat–shock and chaperone proteins in protein folding: Possible molecular mechanisms," *Protein Eng.* **4**, 711–717 (1991).

2. C. B. Anfinsen, C. J. Epstein and R. F. Goldberger, "The genetic control of tertiary protein structure: Studies with model systems," in *Cold Spring Harbour Symp. Quant. Biol.* **28**, 439–49 (1963).
3. Jonathan J. Ewbank and Thomas E. Creighton, "Protein folding by stages," *Current Biol.* **2**, 347–49 (1992).
4. A. Bairoch and B. Boeckmann, "The SWISS–PROT protein sequence data bank," *Nucl. Acids Res.* **19**, 2247–2250 (1991).
5. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, "The Protein Data Bank: A computer based archival file for macromolecular structures," *J. Mol. Biol.* **112**, 535–42 (1977).
6. W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features," *Biopolymers* **22**, 2577–2637 (1983).
7. J. Overington, M. S. Johnson, A. Sali and T. L. Blundell, "Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structure prediction," *Proc. R. Soc. Lond. B*, **241**: 132–145 (1990); W. R. Taylor and C. A. Orengo, "Protein structure alignment," *J. Mol. Biol.* **288**, 1–22 (1989); G. Vriend and C. Sander, "Detection of common three dimensional substructures in proteins," *Proteins* **11**, 52–58 (1991) and Ref. 8.
8. R. Schneider and C. Sander, "Database of homology–derived structures and the structural meaning of sequence alignment," *Proteins* **9**, 56–68 (1991).
9. U. Hobohm, M. Scharf. R. Schneider C. Sander, "Selection of representative protein data sets," *Protein Science* **1**, 409–417 (1992).
10. S. Oliver *et al.*, "The complete DNA sequence of yeast chromosome III," *Nature* **357**, 38–46 (1992); P. Bork, C. Ouzounis, C. Sander, M. Scharf and R. Schneider E. Sonnhammer "What's in a genome?" *Nature* **358**, 287 (1992).
11. J. M. Thornton, T. P. Flores, D. T. Jones and M. B. Swindells, "Prediction of progress at last," *Nature* **354**, 105–06 (1991).
12. N. Qian and Terrence J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *J. Mol. Biol.* **202**, 865–884 (1988); H. Bohr, J. Bohr, S. Brunak, R. M. J. Cotterill, B. Lautrup, Leif Nørskov, O. H. Olsen and S. B. Petersen, "Protein secondary structure and homology by neural networks," *FEBS Lett* **241**, 223–28 (1988); H. L. Holley and M. Karplus, "Protein secondary structure prediction with a neural network," *Proc. Natl. Acad. Sci. USA* **86**, 152–56 (1989); F. Bossa and S. Pascarella, "PRONET: A microcomputer program for predicting the secondary structure of proteins with a neural network," *CABIOS* **5**, 319–320 (1990); D. G. Kneller, F. E. Cohen and R. Longridge, "Improvements in protein secondary structure prediction by an enhanced neural network," *J. Mol. Biol.* **214**, 171–182 (1990); R. Maclin and J. W. Shavlik, "Refining algorithms with knowledge–based neural networks: improving the Chou–Fasman algorithm for protein folding," in *Computational Learning Theory and Natural Learning Systems*, S. Hanson, G. Drostal and R. Rivest (eds.) (MIT Press Massachusets, 1992); P. Stolorz, A. Lapedes and Y. Xia, "Predicting protein secondary structure using neural net and statistical methods," *J. Mol. Biol.* **225**, 363–377 (1992); X. Zhang, J. P. Mesirov and D. L. Waltz, "Hybrid system for protein secondary structure prediction," *J. Mol. Biol.* **225**, 1049–1063 (1992).
13. J. Garnier, D. J. Oguthorpe and B. Robson, "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins," *J. Mol. Biol.*

120, 97–120 (1978); O. B. Ptitsyn and A. V. Finkelstein "Theory of protein secondary structure and algorithm of its prediction," *Biopolymers* **22**, 15–25 (1983); J. M. Levin, J. Garnier and B. Robson, "An algorithm for secondary structure determination in proteins based on sequence similarity," *FEBS Letters* **205**, 303–308 (1986); J.–F. Gibrat, J. Garnier and B. Robson, "Further developments of protein secondary structure prediction using information theory, new parameters and consideration of residue pairs," *J. Mol. Biol.* **198**, 425–443 (1987); J. M. Levin and J. Garnier, "Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool," *Biochimica et Biophysica Acta* **955**, 283–95 (1988); J. M. Levin, J. Garnier, V. Biou, J. F. Gibrat and B. Robson, "Secondary structure prediction: Combination of three different methods," *Protein Engineering* **2**, 185–191 (1988); Michael J. E. Sternberg and Ross D. King, "Machine learning approach for the prediction of protein secondary structure," *J. Mol. Biol.* **216**, 441–457 (1990); Jean Garnier and Jonathan M. Levin, "The protein structure code: What is its present status?" *Cabios* **7**, 133–142 (1991).

14. D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning representations by backpropagating errors," *Nature* **323**, 533–536 (1986).
15. See Ref. 8.
16. L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Patt. Anal. and Machine Intell.* **12**, 993–1001 (1990); W. P. Lincoln and J. Skrzypek, "Synergy of clustering multiple backpropagation networks," in *Neural Information Processing Systems 2*, David S. Touretzky (ed.) (Morgan Kaufmann, San Mateo, CA., 1990) 650–657.
17. B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta* **405**, 442–451 (1975).