

# COVER PAGE

## Neural networks predict protein structure: hype or hit?

**Burkhard Rost**<sup>1, 2, \*</sup>

1 CUBIC, Dept. of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168<sup>th</sup> Street  
BB217, New York, NY 10032, USA, [rost@columbia.edu](mailto:rost@columbia.edu)

2 Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion,  
1150 St. Nicholas Avenue, New York, NY 10032, USA

\* Corresponding author: [rost@columbia.edu](mailto:rost@columbia.edu), <http://cubic.bioc.columbia.edu/>  
Tel: +1-212-305-3773, fax: +1-212-305-7932

in:

### **Artificial Intelligence and Heuristic Methods in Bioinformatics**

**Paolo Frasconi and RonShamir (eds.)**

**IOS Press, Amsterdam  
2003**

**ISBN 1-58603-294-1  
pp. 34-50**

# Neural networks predict protein structure: hype or hit?

Burkhard Rost<sup>1,2,\*</sup>

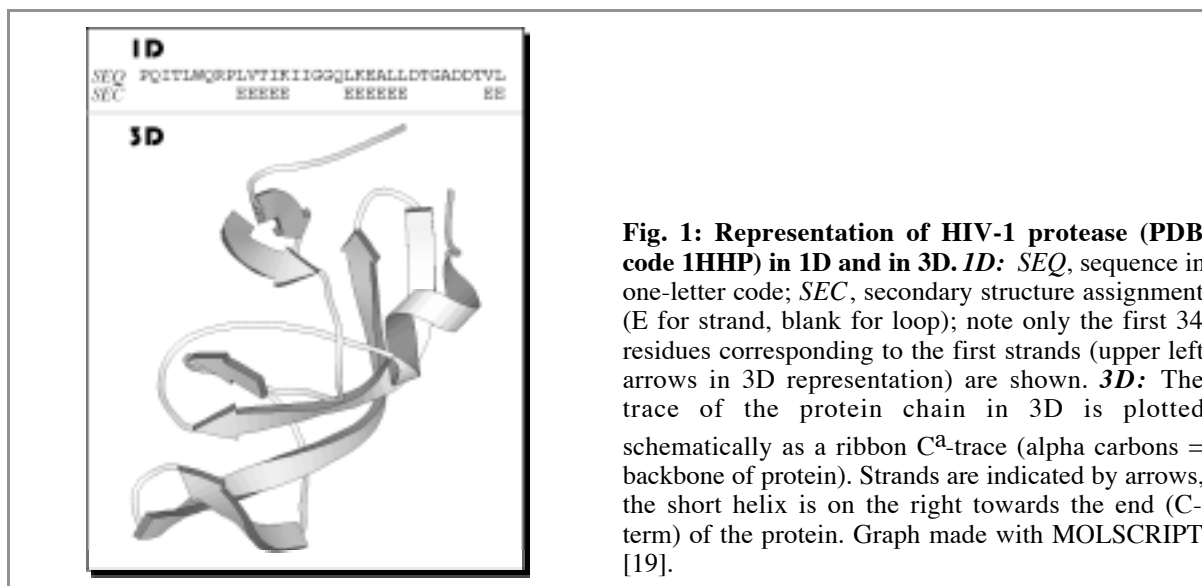
*CUBIC, Dept. Biochemistry and Molecular Biophysics, Columbia University, 650 West 168<sup>th</sup> Street BB217, New York, NY 10032, USA, rost@columbia.edu*

**Abstract.** Neural networks have been applied to many pattern classification problems. Here, I review applications to the problem of predicting protein structure from protein sequence. Initially, many methods were apparently designed by researchers who just wanted a real-life application for their gadget. However, the competitiveness of the field separated the wheat from the chaff. Meanwhile, several neural network-based methods have contributed significantly to advancing the field of bio-informatics, and some are clearly influencing molecular biology. Today, a plethora of network methods is used in everyday sequence analysis, and an increasing number of applications explore very novel problems.

## The protein structure prediction problem

*Proteins constitute life's machinery.* The first bacterial genome was sequenced in 1995 [1]; the first mono-cellular eukaryote (*Saccharomyces cerevisiae*, yeast) followed in 1996 [2]. Meanwhile, we know the entire proteomes (all proteins in a genome) of various multi-cellular eukaryotes: *Drosophila melanogaster* (fly) [3], *Caenorhabditis elegans* (worm) [4], the plant *Arabidopsis thaliana* [5-8]. The first drafts of the human genome [9, 10] have also been completed, however, one year later, we still do not know all human proteins [11]. Overall, more than 60 entire organisms have been sequenced over the last eight years. This avalanche of entirely sequenced organisms is exciting for biology because the genomes contain the blueprint for all parts of life's machinery. The machinery itself consists of proteins that perform most important tasks in organisms (catalysis of biochemical reactions, transport of nutrients, recognition, and transmission of signals). Proteins are formed by joining 20 different amino acids (dubbed residues, when joined in proteins) into a stretched chain. In water, many proteins fold into unique three-dimensional (3D) structures. The main driving force is the need to pack residues for which a contact with water is energetically unfavourable (hydrophobic residues) into the interior of the molecule. This appears possible through the formation of a macroscopic substructure called secondary structure (Fig. 1; for an introduction into protein structure, see: [12]; for principles of folding, see: [13]).

*Sequence determines structure determines function.* The world of proteins is governed by shape: interactions between proteins are mediated by the 'key-hole' principle, i.e., two proteins interact when they fit to one another like a key into a hole. Thus, protein structure determines protein function. What determines structure? All information about the native structure of a protein is coded in the amino acid sequence, plus its native solution environment [14]. Can we decipher the code, i.e., can we predict 3D structure from sequence, or in other words: Can we unboil the egg [15]? In principle, the code could be deciphered from physico-chemical principles using, e.g., molecular dynamics [16-18]. In

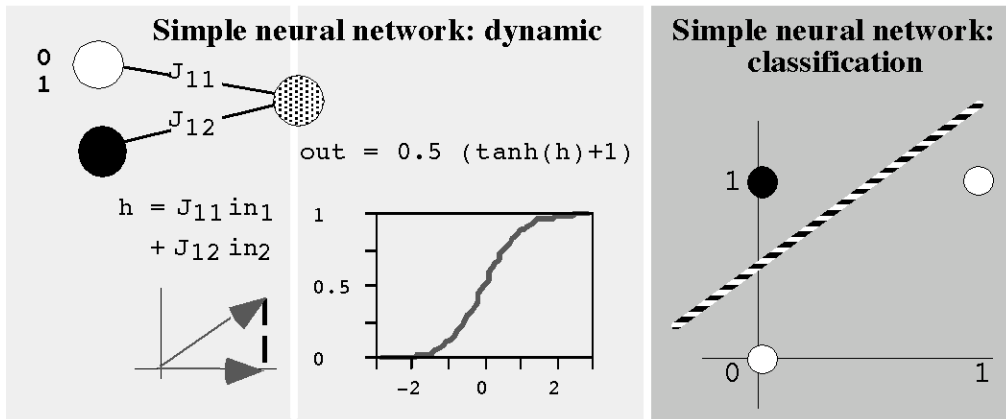


practice, such approaches are frustrated by principle obstacles [20, 21]. Furthermore, the last decade has unravelled that possibly most proteins do not adopt their native 3D structure in vitro, rather they need the cellular machinery to correctly fold in vivo [22-33].

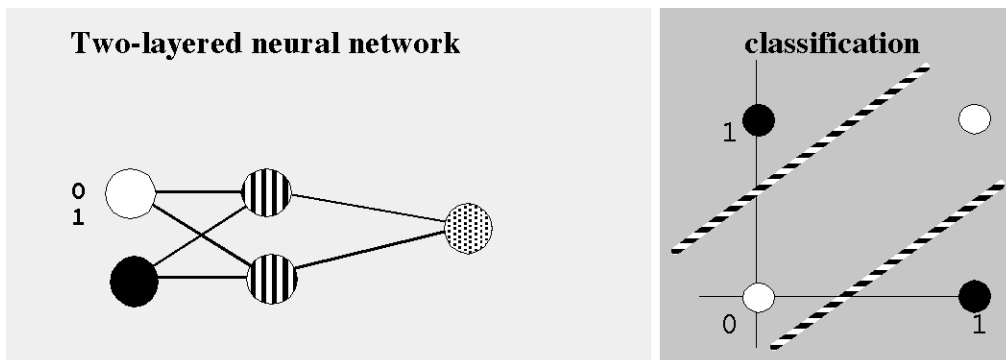
*State-of-the-art in protein structure prediction.* For over 40 years, there has been an ardent search for methods predicting protein structure from sequence (reviews: [34, 35, 20, 21, 36, 37]; books: [38-40]). Many methods were found which looked initially very promising - but always the hope has been dashed [41]. How well do we do in practice? The following results stand out after four experiments initiated by John Moult (CARB, Washington) to explore the accuracy of structure prediction [42-45]. (1) The goal to predict structure from sequence has not been reached, yet. However, most recently approaches that assemble fragments have scored considerable successes [46, 37, 47]. (2) Comparative modelling enables rather accurate predictions of 3D structure for proteins that have significant sequence similarity to proteins of known structure. This technique increases the number of known structures effectively by a factor of 5-30 [35, 20, 48, 36, 49, 50]. (3) Methods addressed at solving simplified structure prediction problems, such as predictions of secondary structure, solvent accessibility, and inter-residue distances [35, 20, 51] have become significantly more accurate, and useful by using the information contained in growing sequence databases.

*How can neural networks predict protein structure?* In practice, the only successful attempts at predicting aspects of protein structure are based on an analysis of common features extracted from proteins of known structures. Neural networks comprise a particular tool for pattern classification (Fig. 2) that – along with many others – has been applied to the problem of protein structure prediction [52-58]. We could write the history of neural network applications in four chapters. (1) Initially, researchers applied black boxes, and searched improvements through optimising the internal free parameters (training speed, network architecture). These early applications were often not evaluated on representative data sets, and thus were scarcely used by biologists. (2) Later, researchers have opened the black box by extracting, or implementing rules, by carving specific knowledge into the networks, and by using networks to detect errors or outliers in data bases. (3) Then, the combination of neural networks with evolutionary information unleashed the full potential of the tool and thus established networks in the bioinformatics community. Incidentally, the successful prediction of protein secondary structure was one of the first examples for applications of neural networks in which these significantly exceeded the performance of

**Fig. 2: Principles about simple feed-forward neural networks**



**Simple neural network:** The simplest layered feed-forward neural network consists of a layer of input units (here two), and a layer of output unit(s) (here one). Signals are transmitted from input to output layer (feed-forward) via the connections ( $J$ 's). The network dynamic consists of a linear and a non-linear step. (1) The value of each input unit (example: 0 for unit 1; 1 for unit 2) is multiplied with the strength of the connection; the products sum to a local field ( $h$ ) representing the signal that arrives at the output unit. The multiplication represents a projection of the input vector onto the vector of the connections. (2) The final output is determined by applying a sigmoid function (shown is the hyperbolic tangent) to the local field. The result is that the output is constrained to values between 0 and 1. On the right hand side the potential of such a network is illustrated: the open, and the dark circles are separated by a line.



**Two-layered neural network:** Two open and two dark circles can obviously not be separated by a single straight line. Two lines would enable the separation, but how can a neural network introduce two lines? The simple trick is the introduction of a layer of hidden layers (hidden as neither input nor output). The dynamics of such a network are identical to the simple network without hidden layer.

**Training a neural network:** How can particular pattern classification problems be implemented? The input is fixed by the pattern, as well is the desired output. The output for a given set of connections is uniquely determined by the dynamics of the network described above. The actual network error can be written as:

$$E = (\text{output} - \text{desired})^2$$

The free variables that contain the potential of the network to learn a given problem are the connections between the layers of units. The simplest way to reduce the network error is by changing the connections according to the derivative of the error with respect to the connections, i.e., by a gradient descent that assures to move downhill in the error-landscape:

$$\Delta J = - \frac{\partial E}{\partial J}$$

This is often referred to as back-propagating the error through the neural network [59, 60]. To avoid being trapped in local minima, in practise, the actual training is typically performed by a variant of this algorithm that permits up-hill moves (conjugate gradient descent [61, 62, 57]).

**Generalisation ability:** With enough hidden units neural networks can learn to separate any set of patterns. Typical applications require to extract particular features (underlying rules) present in the patterns rather than to learn the known examples 'by heart'. A successful extraction of such features permits the network to generalise, i.e., to also correctly classify patterns that have not been learned explicitly. Generalisation requires a balance between the number of training examples (enough to enable feature extraction), and the number of connections (enough to separate patterns). As a rule-of-thumb the number of connections should be an order of magnitude lower than the number of patterns to avoid over-fitting the training data (this learning-by-heart of the training set is also referred to as 'over-training').

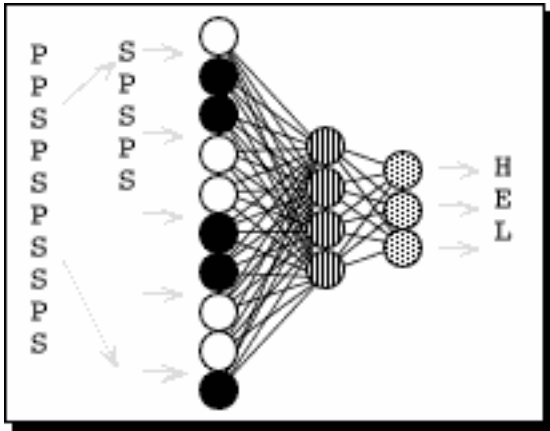
all other systems and even of experts [63]. (4) The last five years of network applications combined many of the lessons learned in the first decade of the tools' history: non-network-experts applied these methods as a standard element in a tool box, and network-experts developed new architectures tailored to particular problems. Here, I focused on discussing but a few representative applications of neural networks (for an excellent review of network applications: [58]).

## Deus ex machina?

*No improvement in secondary structure prediction by black boxes.* Secondary structure prediction methods distinguish between helix (H), strand (E: for extended structure), and other (L for loop). Some stretches of sequence show a particular preference to be in one of these three states. The prediction task is to classify  $w$  adjacent residues as either H, E, or L (Fig. 3). Simple neural networks reached values around 60% accuracy (percentage of residues predicted correctly in any of the three states HEL) [64-66]. This was similar to the best methods 30 years of research had resulted in by the end of the 80's [67, 61, 68, 69, 35, 70, 51]. Attempts to improve performance by changing the network details failed [71, 53]. In contrast, combining neural networks with other methods succeeded to some extent [72].

*Prediction of functional class.* Methods predicting functional similarities between proteins have been based (i) on multiple feed-forward networks [73] using proteins of similar sequences as input, (ii) on simple feed-forward networks using different amino acid features as input [74, 75, 58], and (iii) on Kohonen maps [54] using the frequency with which any of the 20\*20 possible residue pairs occurs in the sequence [76-79], or using the information extracted from database annotations [80, 81]. While feed-forward networks are useful to learn a classification into known features (e.g. types of secondary structure), Kohonen maps have been applied to render a general classification scheme of proteins (e.g. A and B, are similar, and A is more similar to C, than B). Such a classification is a priori not evident and by itself an area of controversy in active research, e.g., attempting to answer questions like: Are we more similar to an orang-utan than to a pig?. One hope guiding such analyses is to end up with similarities between proteins that might help to learn about details in bio-chemical reaction pathways. The neural network-based automatic annotation system [80, 81] has already been applied successfully to genome analysis [82].

*Prediction of surface exposure, and function-specific motifs.* When attempting to arrange secondary structure segments in 3D, one needs to know to which extent a particular residue is exposed to solvent. Neural networks were used to classify amino acid residues as either buried or exposed [83]. Often protein function is associated with relatively short (5-10 residues) sequence motifs (unique pattern of adjacent amino acids). Examples for motifs found by neural networks include: (i) sequence motifs that reveal binding of energy storage molecules [52], (ii) sequence motifs specific for particular proteins, e.g. the immunoglobulins [84], and (iii) signal peptide motifs in sequences [85, 86]. The group of



**Fig. 3: Simple neural network for secondary structure prediction.** For simplification the protein sequence given consists of two amino acid types (S and P). The protein sequence is translated into patterns by shifting a window of  $w$  adjacent residues (shown  $w = 5$ ; typical values in practice are  $w = 13-21$ ) through the protein. The output of the network is uniquely determined (Fig. 2). Suppose the output would be: 0.2, 0.4, 0.5 for the three output states (H, E, L). For known examples the desired output is also known (1, 0, 0 if the central residue is in a helix). Consequently, the network error is given by the difference between actual network output and desired output. The only free variables are the connections. Training or learning means changing the connections

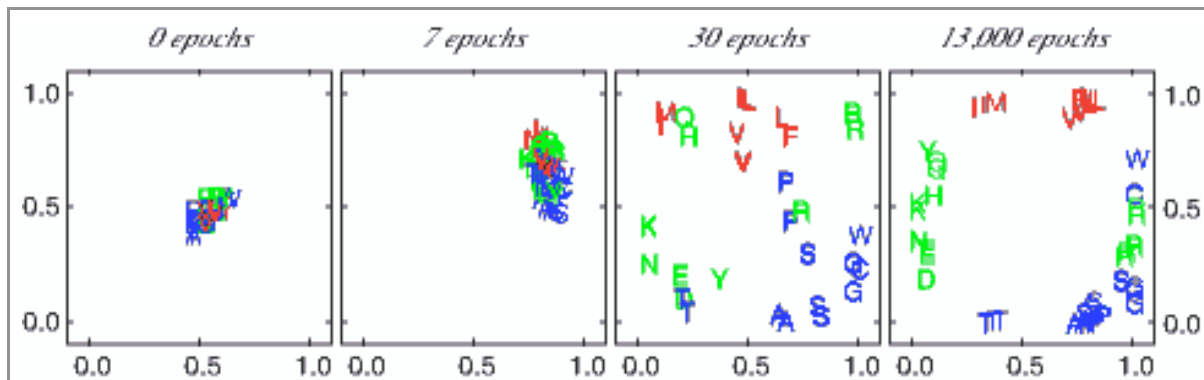
such that the error decreases for the given examples. A training set typically comprises some 30,000 examples. If training is successful, the patterns are correctly classified. But how can new patterns be classified correctly? The hope is that the network succeeds in extracting general rules by the classification of the training patterns. The generalisation ability is checked by another set of test samples for which the mapping of sequence window to secondary structure is also known. Sufficient testing is crucial and requires (1) to remove any significant sequence similarity between test and training set, and (2) to evaluate the expected prediction accuracy on a sufficient number of test proteins (rule of thumb:  $> 100$ ).

Søren Brunak (Copenhagen) has developed two methods of particular practical impact: (i) a system of neural networks predicting signal peptides and cleavage signals [87], (ii) and a combination of rules, and networks predicting glycosilation sites [88].

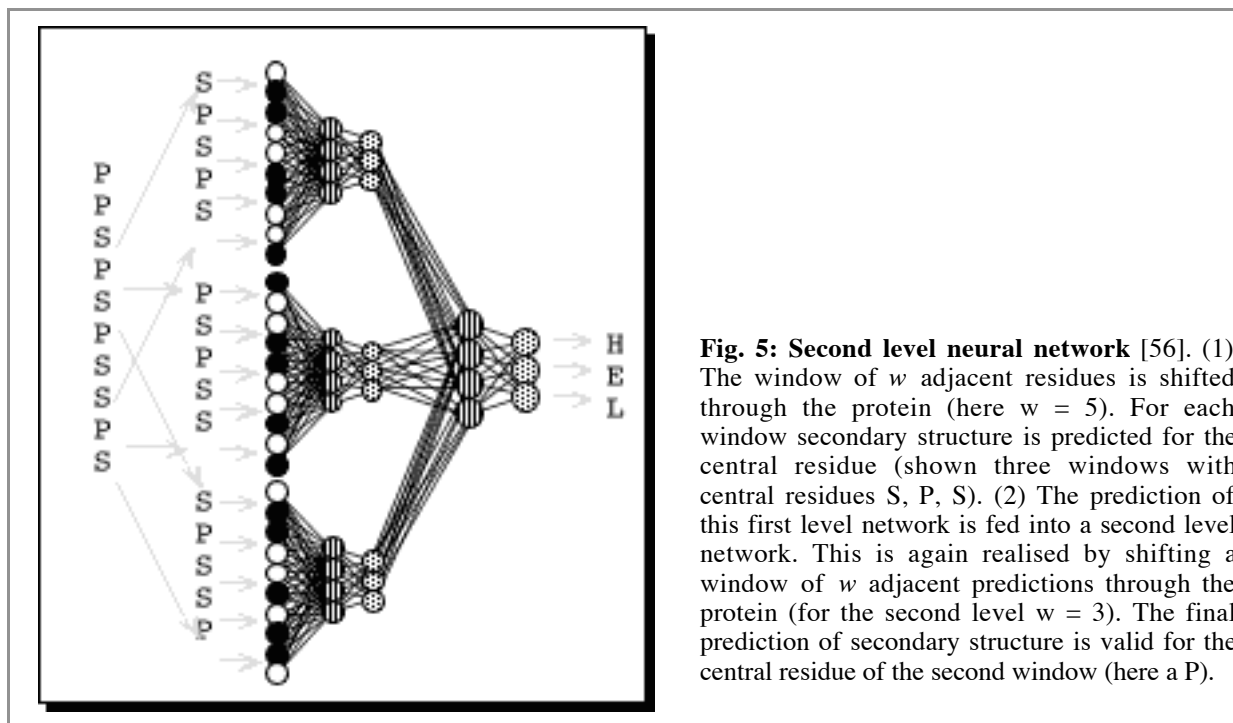
## Opening the black box

*Extracting rules from, and implementing rules into neural networks.* Genome sequences do contain some information about protein structure [89]. A prerequisite to uncover this result was to learn the genetic code by a neural network, i.e., the mapping between the four-letter alphabet of the nucleic acids (DNA), and the 20-letter alphabet of the amino acids (proteins). Analysing the rules learned by the network suggested evidence for a particular scenario for the evolution of the genetic code (Fig. 4). In a similar attempt to extract rules by specific modulation of the training procedure Tchoumatchenko, Vissotsky and Ganascia extracted more complicated rules from networks that learned to predict secondary structure than were available by statistical analysis [90]. Unfortunately, that attempt did not improve performance. Maclin and Shavlik explored the opposite approach by incorporating expert rules into a neural network and thus improved performance over simple statistical devices [91, 92]. All these approaches proved that neural networks are not black boxes, but can become as 'transparent' as rule-based systems. The problem often has been to make use of the complex rules extracted.

*Carving biology into neural networks.* Two problems are common to most secondary structure prediction methods (including simple networks, Fig. 3): (1) strands are predicted at almost random levels of accuracy, (2) and predicted secondary structure segments are too short [61, 68]. The common explanation for the first problem was that strands are stabilised by long-range interactions not visible in a segment of 13-21 residues. The training dynamics of neural networks revealed that networks learned to classify helix, and loop ten times faster than strand [56]. Consequently, the idea was to simply increase the frequency in presenting strand residues during training. This change of the training dynamics



**Fig. 4: Learning the genetic code.** The four-letter nucleic acid code from the genomes is translated into a 20-letter amino acid code from proteins. Three nucleic acids (dubbed one codon) code for one amino acid. This implies that the four nucleic acid can code for  $4 \times 4 \times 4 = 64$  amino acids, i.e., the code is redundant: some amino acids are coded for by more than one codon, and three codons are used for stop-signals during the translation procedure. The minimal network that learned the genetic code had two hidden units [93]. The four graphs represent the connections between the 20 input and the two hidden units. (1) The untrained network with randomly assigned weights locates all 61 points near the centre of the square. (2) After seven training epochs the points have moved into a transient local minimum, where the activities of the intermediate units are close to one and the activities of all the output units are close to zero. (3) At 30 epochs the groups have started to segregate, but are still mixed. (4) Finally at 13,000 epochs the network groups the 61 codons at the edge of the circular region. After the four epochs shown the number of correctly classified codons was 2, 6, 26 and 61, respectively. The final grouping separates hydrophobic residues (top: IMVPF) from hydrophilic (centre right and left: YQHKNE DR), and others (lower right: TSAGPCW). The figure is taken from [93].



**Fig. 5: Second level neural network** [56]. (1) The window of  $w$  adjacent residues is shifted through the protein (here  $w = 5$ ). For each window secondary structure is predicted for the central residue (shown three windows with central residues S, P, S). (2) The prediction of this first level network is fed into a second level network. This is again realised by shifting a window of  $w$  adjacent predictions through the protein (for the second level  $w = 3$ ). The final prediction of secondary structure is valid for the central residue of the second window (here a P).

improved strand accuracy significantly, indicating that the inferior prediction of strand did NOT result primarily from long-range interactions, but from technical problems. The second

problem of predicting too short segments originates from the fact that the sliding window (Fig. 3) erases the correlation between adjacent residues. This shortcoming was corrected by introducing a second level network [56] (Fig. 5). Such a network system learned correlations between adjacent residues. These examples illustrate that neural networks can easily be tailored to particular problems.

*Detecting database errors during training.* Neural networks generalise by extracting the underlying physico-chemical principles from the training data. Obviously, this requires a correct training set. Søren Brunak has pioneered the idea to unravel errors in the training set by monitoring samples that could not be learned even when the networks were trained until over-fitting the data [94-97, 87, 98, 99]. This technique has not only been used successfully to identify errors, and inconsistencies in public databases, but also to improve the performance of the networks.

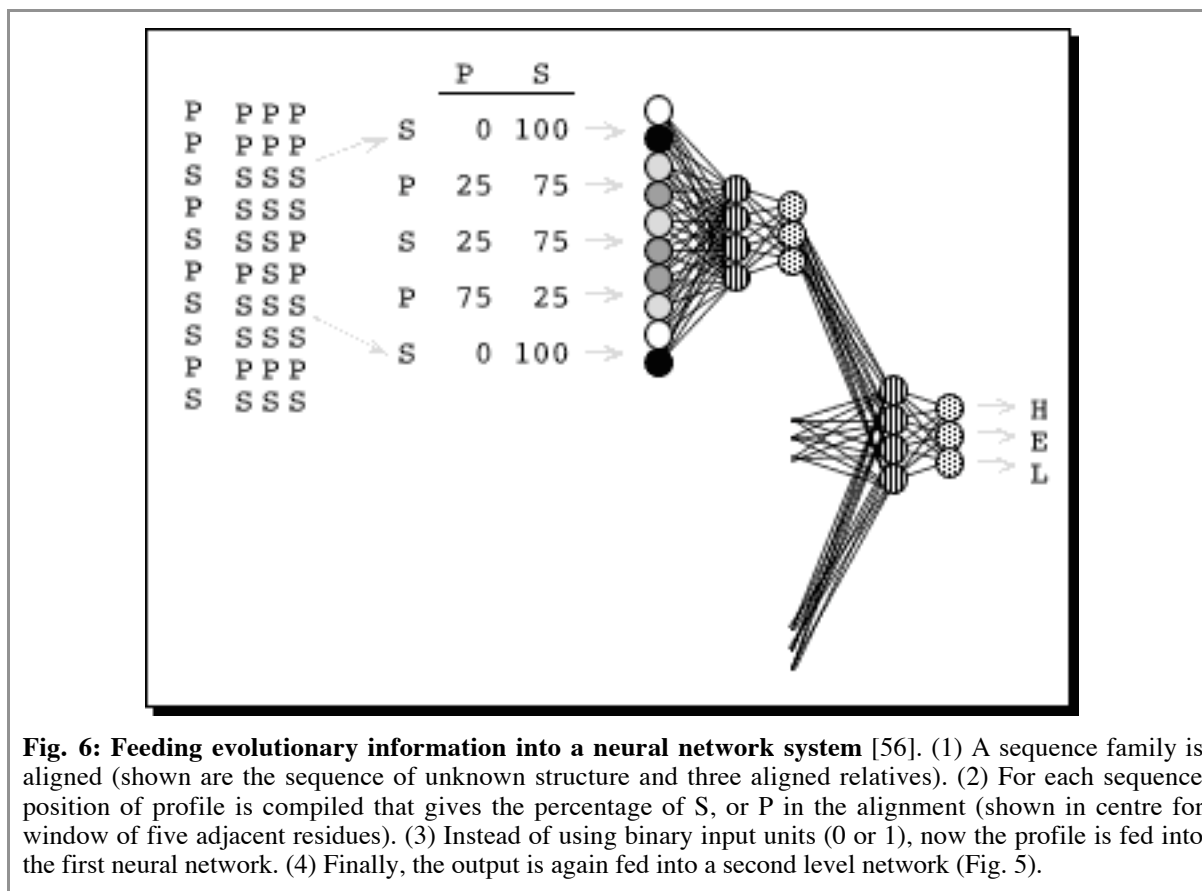
## Learning from evolution to predict protein structure

*Long-range information in multiple sequence alignments.* Some residue substitutions do not alter protein structure. However, not every amino acid can be replaced by any other. On the contrary, one evolutionary step (exchange of one residue) can destabilise a structure. Thus, residue substitution patterns observed in protein families are highly specific for particular details of protein structure and function, i.e. they contain more information about structure than do single sequences. Furthermore, multiple alignments of sequence families implicitly also carry information about interactions between residues separated by more than  $w$  residues in sequence. We can profit from this evolutionary information for structure prediction in the following way [56]. (1) A sequence of unknown structure  $U$  is aligned against a database of known sequences. (2) Proteins with significant sequence identity to  $U$  are retrieved. (3) For each sequence position the profile of residue exchanges in the final multiple alignment is compiled, and fed into a network (Fig. 6).

*Significant improvement of secondary structure prediction.* Using evolutionary information has improved secondary structure prediction accuracy from 65% to over 70% [100], or even over 72% [56]. Such a profile based neural network system was the first method to surpass the magic line of 70% accuracy, and has proven to remain the most accurate method for almost a decade. Today's best methods still use the same idea of feeding evolutionary information into neural networks. The latest improvement to levels above 76% accuracy mainly resulted from larger databases and more sensitive search methods retrieving similar proteins from these larger databases [101-105, 51, 106].

*Evolutionary information improved accuracy in predicting solvent accessibility.* Solvent accessibility at each position of the protein structure is evolutionarily conserved within sequence families. This fact has been used to develop another neural network method for predicting accessibility from multiple alignment information [56]. The final network system is clearly more accurate than methods not using alignment information, and has been established to be more accurate than other prediction methods. More recently, several groups have attempted to refine the concept by training networks to predict different implementations of the output state for solvent accessibility [88, 101], by focussing on particular protein families [107-110], or by predicting the number of contact partners for each amino acid [111-114].

*Predicting transmembrane helices by combining networks with dynamic programming.* The neural network system designed to predict secondary structure for water-soluble proteins failed in predicting helices inserted into the lipid-bilayer of membranes. However, the networks have been re-trained to also predict transmembrane helices [56]. Again information from multiple alignments improved prediction accuracy significantly [56, 35, 20]. The problem of predicting transmembrane helices is ideal to incorporate additional



**Fig. 6: Feeding evolutionary information into a neural network system** [56]. (1) A sequence family is aligned (shown are the sequence of unknown structure and three aligned relatives). (2) For each sequence position of profile is compiled that gives the percentage of S, or P in the alignment (shown in centre for window of five adjacent residues). (3) Instead of using binary input units (0 or 1), now the profile is fed into the first neural network. (4) Finally, the output is again fed into a second level network (Fig. 5).

globular information into the prediction method. The principal idea is to regard the neural network prediction as an energy landscape and to search the best path through this landscape given that transmembrane helices are constrained to a minimal and a maximal length. The final system has achieved a significantly higher accuracy than the simple neural network-based system, and has been applied to analysing entire genomes [115-117, 49, 33].

## Neural network architectures as simple and plastic tools

*Avalanche of applications with standard architectures.* The first neural networks were applied to protein structure prediction in 1988 [64, 65]. A decade later, networks have become a standard method that is tested on problems in bioinformatics (Table 1). Particular recent examples are networks that discover motifs on 3D structures [118], distinguish good from bad drug targets [119], predict binding motifs [120, 121, 110, 122, 123], biological activity [124, 125], post-translational modifications [97, 126, 87, 88, 127, 99, 128-130], particular protein types [131], domains [132] folding rates [133], disordered proteins [134-137], and even substitution matrices [138]. The major strength of networks for many of these applications appears to be that they can readily be adapted to particular problems (Table 1). Two extreme examples of neural network applications are to replace first order statistics when averaging over a variety of scores, i.e. networks with very few input units that use results from a variety of threading methods to identify remote similarities between proteins [139, 140]. The opposite extreme is to generate thousand different networks each specialised on some aspects of the secondary structure prediction problem and to then statistically average over the outputs of these networks [103].

**Table 1: Applications of neural networks <sup>Δ</sup>**

<i>Quote</i>	<i>Application</i>	<i>ali?</i>	<i>+?</i>
<b>1988</b>			
Bohr & [64]	1D sec	no	no
Qian & Sejnowski	1D sec	no	no
<b>1989</b>			
Holley & Karplus [66]	1D sec	no	no
McGregor & [141]	1D beta-turns	no	yes
<b>1990</b>			
Bengio & Pouliot [84]	families	-	-
Bohr & [142]	2D	no	-
Bossa & Pascarella [143]	1D sec	no	no
Holbrook & [83]	1D acc	no	no
Kneller & [144]	1D sec	no	no
Muskal & [145]	2D, cysteine	no	-
<b>1991</b>			
Hirst & Sternberg [52]	binding ATP	no	-
Ladunga & [85]	signal peptide	no	yes
<b>1992</b>			
Ferran & Ferrara [77-79]	families	no	-
Frishman & Argos [73]	families	no	-
Hayward & [146]	1D sec	no	no
Maclin & Shavlik [91, 147, 92]	1D sec	-	-
Muskal & Kim [148]	1D sec	no	no
<b>1993</b>			
Andrade & [149]	1D sec	no	yes
Dubchack & [150]	families	no	-
Fariselli & [151]	1D htm	no	no
Metfessel & [152]	families	no	-
Reczko & [153]	1D sec	no	no
Rost & [61, 154-157, 56]	1D sec	yes	ye
Sasagawa & [158]	1D sec	no	no
Schneider & Wrede [86]	families	-	-
Tchoumatchenko & [90]	no	no	-
<b>1994</b>			
Casadio & [159]	1D htm	no	no
Dombi & [160]	1D htm	no	no
Rost & Sander [161, 56]	1D acc	yes	yes
<b>1995</b>			
Barlow [162]	1D sec	no	no
Casadio & [163]	folding	no	-
Chandonia & [164]	1D sec	no	no
Grossman & [165]	folding,	no	-
Hansen & [166, 88]	binding sugar	no	yes
Milik & [167]	packing	no	-
Rost & [115, 56, 116, 117]	1D htm	yes	yes
<b>1996</b>			
Brunak & Engelbrecht [89]	1D sec / gene	no	no
Casadio & [168]	1D htm	no	no
Fariselli & [169]	1D htm	no	no
Hanke & [170]	families/motifs	no	-
Riis & Krogh [100]	1D sec	yes	yes
Sun & [171, 172]	1D sec	no	-
Wu & [75, 58]	families	-	-
<b>1997</b>			
Aloy & [173]	1D htm	no	?
Andrade & [80]	families	-	-
Asogawa [174]	2D, beta-sheet	no	yes
Dopazo & Carazo [175]	families	-	-
Dosztanyi & [111]	2D	no	-
Dubchack & [176]	families	no	-
Fetrow & [118]	motif discovery	-	-
Gulukota & [120, 122]	binding motifs	-	-
Kawabata & Doi [177]	1D sec	no	no
Lebeda & Olson [121]	binding motifs	no	yes
Lund & [178]	2D	no	yes
Nielsen & [87, 99]	signal peptides	no	yes
<b>1998</b>			
Arrigo & [179]	1D htm	yes	-
Chou & Elrod [180]	localisation	no	no
Diederichs & [181]	1D btm	no	-
Honeyman & [124]	binding sites	no	-
Reinhardt & Hubbard [182]	localisation	no	yes
Wrede & [125]	binding specificity	no	-
<b>1999</b>			
Blom & [98]	motifs	no	yes
Casadio & [183]	1D sec	yes	yes
Emanuelsson & [184]	localisation	no	yes
Fariselli & [185]	2D, cysteine	yes	yes
Fariselli & [186]	2D	yes	yes
Gorodkin & [187]	2D	no	-
Guermeur & [188]	1D sec	no	?
Jones [139]	families	-	yes
Jones [189]	1D sec	yes	yes
Krogh & Riis [190]	1D sec	yes	yes
Pasquier & [191]	1D htm	no	-
Shepherd & [192]	1D beta-turns	no	yes
<b>2000</b>			
Baldi & [193]	2D, beta-sheet	no	yes
Cuff & Barton [101]	1D sec + acc	yes	yes
Emanuelsson & [194]	localisation	no	yes
Fariselli & [112]	2D	no	-
Frimurer & [119]	motif discovery	-	-
Gurvitz & [131]	families	-	-
Herrmann & [128]	motifs	no	-
Jacoboni & [195]	1D sec	no	-
Ouali & King [196]	1D sec	yes	yes
Petersen [103]	1D sec	yes	yes
Stahl & [110]	binding motifs	-	-
Workman & Stormo [197]	binding sites	no	yes
<b>2001</b>			
Babajide & [198]	2D potentials	no	-
Bohr & [199]	folding	no	-
Ding & Dubchack [200]	families	no	-
Fariselli & [113]	2D	no	-
Fariselli & [201]	2D	yes	yes
Iakoucheva [137]	disorder	no	-
Jacoboni & [202]	1D btm	yes	yes
Lin & [138]	families	-	-
Mlinsek & [123]	binding sites	no	-
Murvai & [132]	families	no	-
Pasquier & [203]	families	no	-
Pollastri & [114]	2D	no	yes
Zhou & Shan [204]	2D	no	-
Zhou & Zhou [133]	folding	-	-

## Abbreviations for Table 1:

**ali?:** information from sequence alignments used or not ('-' = not applicable); **+?:** improved performance over non-network methods ('-' = unclear or not applicable); **Applications:** 1D sec = secondary structure prediction, 1D acc = solvent accessibility prediction, 1D htm = transmembrane helix prediction, 1D btm = transmembrane strand prediction, 2D = residue-residue contact predictions.

---

*Recurrent networks: changing the architecture to improve performance.* Most applications used standard architectures of neural networks. More recently the groups of Pierre Baldi and Paolo Frasconi have introduced the concept of recurrent neural network into the field of protein structure prediction [205, 193, 206, 114, 105]. These networks handle more global information by feeding the output of the system back into the input and thus correlating information from regions outside of what is accessible through the sliding-window technique.

## Neural networks: hype or helpful?

*Structure prediction: work in progress...* To predict 3D structure from sequence is a task challenging enough to have occupied a generation of researchers. The bad news: We still cannot accurately predict 3D structure from sequence. The good news: we have come closer, and growing databases facilitate the task. A solution of the structure prediction problem would supposedly change experimental molecular biology more than any other theoretical method. We may witness such a break-through in the near future.

*Neural networks contributed to structure prediction.* Almost any imaginable algorithm has been applied to the secondary structure prediction problem. However, once researchers left the path of trying to optimise black-boxes it was through neural network applications that many break-throughs were achieved. For example, a neural network system for predicting various aspects of 1D structure based on evolutionary information is by far the most widely used prediction method [56]. Other network-based methods are unique, or superior in their field [79, 100, 81, 126, 88]. Furthermore, neural networks revealed data base errors, and principles underlying protein structures [95, 96, 93, 97, 56].

*Solving real problems requires real commitments.* In this brief review I covered only a few neural network applications to protein structure prediction. Many early neural network applications had no impact because theoreticians tended to spend more time on developing methods than on appropriately testing them, and frequently solved problems that were of no practical interest. Neural networks have surpassed these barriers, and have impacted structure prediction. The community of researchers understanding the biological relevance of problems and the principles of neural networks continues to grow.

*Will neural network constitute THE tool assisting molecular biology?* Sequencing the human genome took enormous human and financial resources. However, genome sequences as such have no impact on any problem of society. In order to benefit from genome sequences for health care, or understanding the necessity of bio-diversity, we need to attach knowledge about protein structure and function to the genes. Will neural network applications enable further advances? Protein structure prediction is not likely to be managed through a single genial discovery. Instead, it appears we need to direct an orchestra of different prediction methods such that they combine to a melody. I see the following possible contributions from neural networks to the harmony of the final prediction system. (1) Further improvement of predicting protein structure in 1D (secondary structure, solvent accessibility, and transmembrane helices). (2) Prediction of distances between residues in the final structure, sub-cellular localisation, active, or binding sites, particular functional motifs, and interfaces between proteins. (3) Prediction of any other feature associated with protein function. (4) Putting the predictions from various tools together may require an iteration of prediction

cycles, during which certainty about structure is gradually increased. For example, users may know that a particular set of residues is forming a cluster. Could we devise a neural network-based method that changes its prediction given this partial knowledge? Overall, there seem to be many tasks that could be tackled best by neural networks. We only have to do it, and do it with sufficient care and commitment to the growing field of bioinformatics.

**Acknowledgements:** Thanks to Jinfeng Liu (Columbia) for computer assistance, to Claus Andersen and Søren Brunak (CBS Copenhagen) for helpful comments on the manuscript, and to Henry Bigelow (Columbia) for critically reading the manuscript. Thanks to Paolo Frasconi (Firenze) and his team for organising the meeting that spun off this contribution. The work of BR was supported by the grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institute of Health. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

## References

- [1] R. D. Fleischmann, et al. and J. C. Venter, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269** (1995) 496-512.
- [2] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin and S. G. Oliver, Life with 6000 genes. *Science* **274** (1996) 546-567.
- [3] M. D. Adams, et al., The genome sequence of *Drosophila melanogaster*. *Science* **287** (2000) 2185-2195.
- [4] The *C. elegans* Sequencing Consortium, Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282** (1998) 2012-2018.
- [5] Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408** (2000) 796-815.
- [6] M. Salanoubat, et al., Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature* **408** (2000) 820-822.
- [7] S. Tabata, et al., Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature* **408** (2000) 823-826.
- [8] A. Theologis, et al., J. C. Venter and R. W. Davis, Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature* **408** (2000) 816-820.
- [9] The genome international sequencing consortium, Initial sequencing and analysis of the human genome. *Nature* **409** (2001) 860-921.
- [10] J. C. Venter, et al., The Human genome. *Science* **291** (2001) 1304-1351.
- [11] C. O'Donovan, R. Apweiler and A. Bairoch, The human proteomics initiative (HPI). *TIBTECH* **19** (2001) 178-181.
- [12] C. Brändén and J. Tooze, Introduction to Protein Structure. Garland Publ.: New York, London, 1991.
- [13] E. E. Lattman and G. D. Rose, Protein folding-what's the question? *Proc. Natl. Acad. Sci. U.S.A.* **90** (1993) 439-441.
- [14] C. B. Anfinsen, Principles that govern the folding of protein chains. *Science* **181** (1973) 223-230.
- [15] M. F. Perutz, "Unboiling" an egg. *Discovery* **May** (1940) reprint in Jaenicke, Rainer: Protein Folding. Amsterdam, New York: Elsevier, 1980, p. 1914.
- [16] M. Levitt and A. Warshel, Computer simulation of protein folding. *Nature* **253** (1975) 694-698.
- [17] A. T. Hagler and B. Honig, On the formation of protein tertiary structure on a computer. *Proc. Natl. Acad. Sci. U.S.A.* **75** (1978) 554-558.
- [18] W. F. van Gunsteren, Molecular dynamics studies of proteins. *Curr. Opin. Str. Biol.* **3** (1993) 167-174.
- [19] P. Kraulis, MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24** (1991) 946-950.
- [20] B. Rost and S. I. O'Donoghue, Sisyphus and prediction of protein structure. *CABIOS* **13** (1997) 345-356.
- [21] B. Rost, Protein structure prediction in 1D, 2D, and 3D. In: P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III and P. R. Schreiner (eds.), *The Encyclopaedia of Computational Chemistry*. John Wiley & Sons, Chichester, 1998, pp. 2242-2255.
- [22] T. J. P. Hubbard and C. Sander, The role of heat-shock and chaperone proteins in protein folding: possible molecular mechanisms. *Prot. Engin.* **4** (1991) 711-717.
- [23] A. Joachimiak, Capturing the misfolds: chaperone-peptide-binding motifs. *Nat. Struct. Biol.* **4** (1997) 430-434.
- [24] J. Martin and F. U. Hartl, Chaperone-assisted protein folding. *Curr. Opin. Str. Biol.* **7** (1997) 41-52.
- [25] W. J. Netzer and F. U. Hartl, Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* **388** (1997) 343-349.
- [26] R. J. Ellis, C. Dobson and U. Hartl, Sequence does specify protein conformation. *TIBS* **23** (1998) 468.
- [27] P. E. Wright and H. J. Dyson, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293** (1999) 321-331.

- [28] M. E. Gottesman and W. A. Hendrickson, Protein folding and unfolding by Escherichia coli chaperones and chaperonins. *Curr. Opin. Microbiol.* **3** (2000) 197-202.
- [29] C. R. Sanders and J. K. Nagy, Misfolding of membrane proteins in health and disease: the lady or the tiger? *Curr. Opin. Str. Biol.* **10** (2000) 438-442.
- [30] C. M. Dobson, The structural basis of protein folding and its links with human disease. *Philos Trans R Soc Lond B Biol Sci* **356** (2001) 133-145.
- [31] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner and Z. Obradovic, Intrinsically disordered protein. *J Mol Graph Model* **19** (2001) 26-59.
- [32] M. Fandrich, M. A. Fletcher and C. M. Dobson, Amyloid fibrils from muscle myoglobin. *Nature* **410** (2001) 165-166.
- [33] J. Liu, H. Tan and B. Rost, Eukaryotes full of loopy proteins? *J. Mol. Biol.* (2002) submitted.
- [34] G. J. Barton, Protein secondary structure prediction. *Curr. Opin. Str. Biol.* **5** (1995) 372-376.
- [35] B. Rost and C. Sander, Bridging the protein sequence-structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct.* **25** (1996) 113-136.
- [36] M. A. Marti-Renom, A. Stuart, A. Fiser, R. Sanchez, F. Melo and A. Sali, Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29** (2000) 291-325.
- [37] R. Bonneau and D. Baker, Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* **30** (2001) 173-189.
- [38] R. F. Doolittle, Computer methods for macromolecular sequence analysis. Academic Press: San Diego, 1996.
- [39] M. J. E. Sternberg, Protein structure prediction. Oxford Univ. Press: Oxford, 1996.
- [40] P. Baldi and S. Brunak, Bioinformatics: the machine learning approach. MIT Press: Cambridge, 2001.
- [41] B. Honig and F. E. Cohen, Adding backbone to protein folding: why proteins are polypeptides. *Folding & Design* **1** (1996) R17-R20.
- [42] CASP1, Special issue: First Meeting on Critical Assessment of Protein Structure prediction (CASP). *Proteins* **23** (1995).
- [43] CASP2, Special issue: Second Meeting on Critical Assessment of Protein Structure prediction (CASP). *Proteins Suppl.* **2** (1997).
- [44] CASP3, Special issue: Third Meeting on Critical Assessment of Protein Structure prediction (CASP). *Proteins Suppl.* **2** (1999).
- [45] CASP4WWW, Fourth meeting on the critical assessment of techniques for protein structure prediction. Prediction Center, Lawrence Livermore National Lab, 2000.
- [46] C. Bystroff, V. Thorsson and D. Baker, HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.* **301** (2000) 173-190.
- [47] A. M. Lesk, L. Lo Conte and T. J. P. Hubbard, Assessment of novel folds targets in CASP4: Predictions of three-dimensional structures, secondary structures, and interresidue contacts. *Proteins* (2001) in press.
- [48] S. A. Teichmann, C. Chothia and M. Gerstein, Advances in structural genomics. *Curr. Opin. Str. Biol.* **9** (1999) 390-399.
- [49] J. Liu and B. Rost, Comparing function and structure between entire proteomes. *Prot. Sci.* **10** (2001) 1970-1979.
- [50] J. Liu and B. Rost, Target space for structural genomics revisited. *Bioinformatics* (2002) submitted.
- [51] B. Rost, Protein secondary structure prediction continues to rise. *J. Struct. Biol.* **134** (2001) 204-218.
- [52] J. D. Hirst and M. J. E. Sternberg, Prediction of ATP-binding motifs a comparison of a perceptron-type neural network and a consensus sequence method. *Prot. Engin.* **4** (1991) 615-623.
- [53] S. R. Presnell and F. E. Cohen, Artificial neural networks for pattern recognition in biochemical sequences. *Annu. Rev. Biophys. Biomol. Struct.* **22** (1993) 283-298.
- [54] M. Arbib, The handbook of brain theory and neural networks. Bradford Books/The MIT Press: Cambridge, MA, 1995.
- [55] E. Fiesler and R. Beale, Handbook of Neural Computation. Oxford Univ. Press: New York, 1996.
- [56] B. Rost, PHD: predicting one-dimensional protein structure by profile based neural networks. *Meth. Enzymol.* **266** (1996) 525-539.
- [57] B. Rost, NN which predicts protein secondary structure. In: E. Fiesler and R. Beale (eds.), Handbook of Neural Computation. Oxford Univ. Press, New York, 1996, pp. G4.1.
- [58] C. H. Wu, Artificial neural networks for molecular sequence analysis. *Comput. Chem.* **21** (1997) 237-256.
- [59] B. Müller and J. Reinhardt, Neural Networks. Springer: Berlin, F.R.G., 1990.
- [60] J. A. Hertz, A. Krogh and R. G. Palmer, Introduction to the theory of neural computation. Addison-Wesley: Redwood City, C.A., U.S.A., 1991.
- [61] B. Rost and C. Sander, Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232** (1993) 584-599.
- [62] B. Rost and C. Sander, Protein structure prediction by neural networks. In: M. Arbib (eds.), The handbook of brain theory and neural networks. Bradford Books/The MIT Press, Cambridge, MA, 1995, pp. 772-775.
- [63] B. Rost and C. Sander, Jury returns on structure prediction. *Nature* **360** (1992) 540.
- [64] H. Bohr, J. Bohr, S. Brunak, R. M. J. Cotterill, B. Lautrup, L. Nørskov, O. H. Olsen and S. B. Petersen, Protein secondary structure and homology by neural networks. *FEBS Lett.* **241** (1988) 223-228.
- [65] N. Qian and T. J. Sejnowski, Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202** (1988) 865-884.
- [66] H. L. Holley and M. Karplus, Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. U.S.A.* **86** (1989) 152-156.

- [67] B. Rost and C. Sander, Exercising multi-layered networks on protein secondary structure. In: O. Benhar, S. Brunak, P. DelGiudice and M. Grandolfo (eds.), *Neural Networks: From Biology to High Energy Physics*. International Journal of Neural Systems, Elba, Italy, 1992, pp. 209-220.
- [68] B. Rost, C. Sander and R. Schneider, Progress in protein structure prediction? *TIBS* **18** (1993) 120-123.
- [69] B. Rost and C. Sander, Structure prediction of proteins - where are we now? *Curr. Opin. Biotech.* **5** (1994) 372-380.
- [70] B. Rost and C. Sander, Third generation prediction of secondary structure. *Meth. Mol. Biol.* **143** (2000) 71-95.
- [71] J. D. Hirst and M. J. E. Sternberg, Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochem.* **31** (1992) 615-623.
- [72] X. Zhang, J. P. Mesirov and D. L. Waltz, Hybrid system for protein secondary structure prediction. *J. Mol. Biol.* **225** (1992) 1049-1063.
- [73] D. Frishman and P. Argos, Recognition of distantly related protein sequences using conserved motifs and neural networks. *J. Mol. Biol.* **228** (1992) 951-962.
- [74] C. Wu, G. Whitson, J. McLarty, A. Ermongkonchai and T.-C. Chang, Protein classification artificial neural system. *Prot. Sci.* **1** (1992) 667-677.
- [75] C. H. Wu, S. Zhao, H.-L. Chen, C.-J. Lo and J. McLarty, Motif identification neural design for rapid and sensitive protein family search. *CABIOS* **12** (1996) 109-118.
- [76] E. A. Ferrán and P. Ferrara, Topological maps of protein sequences. *Biol. Cybern.* **65** (1991) 451-458.
- [77] E. Ferrán and P. Ferrara, Clustering proteins into families using artificial neural networks. *CABIOS* **8** (1992) 39-44.
- [78] E. Ferrán and P. Ferrara, A neural network dynamics that resembles protein evolution. *Physica A* **185** (1992) 395-401.
- [79] E. A. Ferrán and B. Pflugfelder, A hybrid method to cluster protein sequences based on statistics and artificial neural networks. *CABIOS* **9** (1993) 671-680.
- [80] M. A. Andrade, G. Casari, C. Sander and A. Valencia, Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biol. Cybern.* **76** (1997) 441-450.
- [81] M. A. Andrade and A. Valencia, Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. In: T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander and A. Valencia (eds.), *Fifth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Halkidiki, Greece, 1997, pp. 25-32.
- [82] C. Ouzounis, G. Casari, C. Sander, J. Tamames and A. Valencia, Computational comparisons of model genomes. *TIBTECH* **14** (1996) 280-285.
- [83] S. R. Holbrook, S. M. Muskal and S.-H. Kim, Predicting surface exposure of amino acids from protein sequence. *Prot. Engin.* **3** (1990) 659-665.
- [84] Y. Bengio and Y. Pouliot, Efficient recognition of immunoglobulin domains from amino acid sequences using a neural network. *CABIOS* **6** (1990) 319-324.
- [85] I. Ladunga, F. Czakó, I. Csabai and T. Geszti, Improving signal peptide prediction accuracy by simulated neural network. *CABIOS* **7** (1991) 485-487.
- [86] G. Schneider and P. Wrede, Development of artificial neural filters for pattern recognition in protein sequences. *J. Mol. Evol.* **36** (1993) 586-595.
- [87] H. Nielsen, J. Engelbrecht, S. Brunak and G. von Heijne, A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *International Journal of Neural Systems* **8** (1997) 581-599.
- [88] J. Hansen, O. Lund, N. Tolstrup, A. A. Gooley, K. L. Williams and S. Brunak, NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconjugate Journal* **15** (1998) 115-130.
- [89] S. Brunak and J. Engelbrecht, Protein structure and the sequential structure of mRNA:  $\alpha$ -helix and  $\beta$ -sheet signals at the nucleotide level. *Proteins* **25** (1996) 237-252.
- [90] I. Tchoumatchenko, F. Vissotsky and J.-G. Ganascia, 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, 1993.
- [91] R. Maclin and J. W. Shavlik, Refining Algorithms with Knowledge-Based Neural Networks: Improving the Chou-Fasman Algorithm for Protein Folding. In: S. Hanson, G. Drostal and R. Rivest (eds.), *Computational Learning Theory and Natural Learning Systems*. MIT Press, Cambridge, MA, 1992, pp.
- [92] R. Maclin and J. W. Shavlik, Using knowledge-based neural networks to improve algorithms: refining the Chou-Fasman algorithm for protein folding. *Machine Learning* **11** (1993) 195-215.
- [93] N. Tolstrup, J. Toftgård, J. Engelbrecht and S. Brunak, Neural network model of the genetic code is strongly correlated to the GES scale of amino acid transfer free energies. *J. Mol. Biol.* **243** (1994) 816-820.
- [94] S. Brunak, J. Engelbrecht and S. Knudsen, Neural network detects errors in the assignment of mRNA splice sites. *Nucl. Acids Res.* **18** (1990) 4797-4801.
- [95] S. Brunak, Non-linearities in training sets identified by inspecting the order in which neural networks learn. In: O. Benhar, C. Bosio, P. Del Giudice and E. Tabet (eds.), *Neural Networks From Biology to High Energy Physics*. ETS Editrice Pisa, Elba, Italy, 1991, pp. 277-288.
- [96] S. Brunak, J. Engelbrecht and S. Knudsen, Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* **220** (1991).
- [97] N. Blom, J. Hansen, D. Blaas and S. Brunak, Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. *Prot. Sci.* **5** (1996) 2203-2216.
- [98] N. Blom, S. Gammeltoft and S. Brunak, Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294** (1999) 1351-1362.

- [99] H. Nielsen, S. Brunak and G. von Heijne, Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Prot. Engin.* **12** (1999) 3-9.
- [100] S. K. Riis and A. Krogh, Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comp. Biol.* **3** (1996) 163-183.
- [101] J. A. Cuff and G. J. Barton, Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40** (2000) 502-511.
- [102] R. D. King, M. Ouali, A. T. Strong, A. Aly, A. Elmaghraby, M. Kantardzic and D. Page, Is it better to combine predictions? *Prot. Engin.* **13** (2000) 15-19.
- [103] T. N. Petersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G. P. Gippert and O. Lund, Prediction of protein secondary structure at 80% accuracy. *Proteins* **41** (2000) 17-20.
- [104] C. A. Andersen, H. Bohr and S. Brunak, Protein secondary structure: category assignment and predictability. *FEBS Lett.* **507** (2001) 6-10.
- [105] G. Pollastri, D. Przybylski, B. Rost and P. Baldi, Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* (2001) in press.
- [106] D. Przybylski and B. Rost, Alignments grow, secondary structure prediction improves. *Proteins* **46** (2002) 195-205.
- [107] F. J. Lebeda and M. A. Olson, Predicting differential antigen-antibody contact regions based on solvent accessibility. *J Protein Chem* **16** (1997) 607-618.
- [108] F. J. Lebeda and M. A. Olson, Predictions of secondary structure and solvent accessibility of the light chain of the clostridial neurotoxins. *J Nat Toxins* **7** (1998) 227-238.
- [109] F. J. Lebeda, T. C. Umland, M. Sax and M. A. Olson, Accuracy of secondary structure and solvent accessibility predictions for a clostridial neurotoxin C-fragment. *J Protein Chem* **17** (1998) 311-318.
- [110] M. Stahl, C. Taroni and G. Schneider, Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Prot. Engin.* **13** (2000) 83-88.
- [111] Z. Dosztanyi, A. Fiser and I. Simon, Stabilization centers in proteins: identification, characterization and predictions. *J. Mol. Biol.* **272** (1997) 597-612.
- [112] P. Fariselli and R. Casadio, Prediction of the number of residue contacts in proteins. *Ismb* **8** (2000) 146-151.
- [113] P. Fariselli and R. Casadio, RCNPRED: prediction of the residue co-ordination numbers in proteins. *Bioinformatics* **17** (2001) 202-204.
- [114] G. Pollastri, P. Baldi, P. Fariselli and R. Casadio, Improved prediction of the number of residue contacts in proteins by recurrent neural networks. *Bioinformatics* **17** (2001) S234-242.
- [115] B. Rost, R. Casadio, P. Fariselli and C. Sander, Prediction of helical transmembrane segments at 95% accuracy. *Prot. Sci.* **4** (1995) 521-533.
- [116] B. Rost, R. Casadio and P. Fariselli, Refining neural network predictions for helical transmembrane proteins by dynamic programming. In: D. States, P. Agarwal, T. Gaasterland, L. Hunter and R. F. Smith (eds.), Fourth International Conference on Intelligent Systems for Molecular Biology. Menlo Park, CA: AAAI Press, St. Louis, M.O., U.S.A., 1996, pp. 192-200.
- [117] B. Rost, R. Casadio and P. Fariselli, Topology prediction for helical transmembrane proteins at 86% accuracy. *Prot. Sci.* **5** (1996) 1704-1718.
- [118] J. S. Fetrow, M. J. Palumbo and G. Berg, Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins* **27** (1997) 249-271.
- [119] T. M. Frimurer, R. Bywater, L. Naerum, L. N. Lauritsen and S. Brunak, Improving the odds in discriminating "drug-like" from "non drug-like" compounds. *J Chem Inf Comput Sci* **40** (2000) 1315-1324.
- [120] K. Gulukota, J. Sidney, A. Sette and C. DeLisi, Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.* **267** (1997) 1258-1267.
- [121] F. J. Lebeda and M. A. Olson, Predicting differential antigen-antibody contact regions based on solvent accessibility. *J. Prot. Chem.* **16** (1997) 607-618.
- [122] K. Gulukota and C. DeLisi, Neural network method for predicting peptides that bind major histocompatibility complex molecules. *Meth. Mol. Biol.* **156** (2001) 201-209.
- [123] G. Mlinsek, M. Novic, M. Hodoscek and T. Solmajer, Prediction of enzyme binding: human thrombin inhibition study by quantum chemical and artificial intelligence methods based on X-ray structures. *J Chem Inf Comput Sci* **41** (2001) 1286-1294.
- [124] M. C. Honeyman, V. Brusica, N. L. Stone and L. C. Harrison, Neural network-based prediction of candidate T-cell epitopes. *Nat. Biotechnol.* **16** (1998) 966-969.
- [125] P. Wrede, O. Landt, S. Klages, A. Fatemi, U. Hahn and G. Schneider, Peptide design aided by neural networks: biological activity of artificial signal peptidase I cleavage sites. *Biochem.* **37** (1998) 3588-3593.
- [126] H. Nielsen, J. Engelbrecht, S. Brunak and G. von Heijne, Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Prot. Engin.* **10** (1997) 1-6.
- [127] R. Gupta, E. Jung, A. A. Gooley, K. L. Williams, S. Brunak and J. Hansen, Scanning the available Dictyostelium discoideum proteome for O-linked GlcNAc glycosylation sites using neural networks. *Glycobiology* **9** (1999) 1009-1022.
- [128] J. L. Herrmann, R. Delahay, A. Gallagher, B. Robertson and D. Young, Analysis of post-translational modification of mycobacterial proteins using a cassette expression system. *FEBS Lett.* **473** (2000) 358-362.
- [129] K. Nakai, Protein sorting signals and prediction of subcellular localization. *Adv Protein Chem* **54** (2000) 277-344.
- [130] K. Nakai, Review: prediction of in vivo fates of proteins in the era of genomics and proteomics. *J. Struct. Biol.* **134** (2001) 103-116.

- [131] A. Gurvitz, S. Langer, M. Piskacek, B. Hamilton, H. Ruis and A. Hartig, Predicting the function and subcellular location of *Caenorhabditis elegans* proteins similar to *Saccharomyces cerevisiae* beta-oxidation enzymes. *Yeast* **17** (2000) 188-200.
- [132] J. Murvai, K. Vlahovicek, C. Szepesvari and S. Pongor, Prediction of protein functional domains from sequences using artificial neural networks. *Genome Res.* **11** (2001) 1410-1417.
- [133] H. Zhou and Y. Zhou, Folding Rate Prediction Using Total Contact Distance. *Biophys. J.* **82** (2002) 458-463.
- [134] E. Garner, P. Cannon, P. Romero, Z. Obradovic and A. K. Dunker, Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. *Genome Inform.* **9** (1998) 201-214.
- [135] P. Romero, Z. Obradovic, C. Kissinger, J. E. Villafranca, E. Garner, S. Guilliot and A. K. Dunker, Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput.* **3** (1998) 437-448.
- [136] P. Romero, Z. Obradovic and A. K. Dunker, Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Lett.* **462** (1999) 363-367.
- [137] L. M. Iakoucheva, A. L. Kimzey, C. D. Masselon, J. E. Bruce, E. C. Garner, C. J. Brown, A. K. Dunker, R. D. Smith and E. J. Ackerman, Identification of intrinsic order and disorder in the DNA repair protein XPA. *Prot. Sci.* **10** (2001) 560-571.
- [138] K. Lin, A. C. May and W. R. Taylor, Amino acid substitution matrices from an artificial neural network model. *J. Comp. Biol.* **8** (2001) 471-481.
- [139] D. T. Jones, GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287** (1999) 797-815.
- [140] J. Lundstrom, L. Rychlewski, J. Bujnicki and A. Elofsson, Pcons: a neural-network-based consensus predictor that improves fold recognition. *Prot. Sci.* **10** (2001) 2354-2362.
- [141] M. J. McGregor, T. P. Flores and M. J. E. Sternberg, Prediction of beta-turns in proteins using neural networks. *Prot. Engin.* **2** (1989) 521-526.
- [142] H. Bohr, J. Bohr, S. Brunak, H. Fredholm, B. Laustrup and S. B. Petersen, A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Lett.* **261** (1990) 43-46.
- [143] F. Bossa and S. Pascarella, PRONET: a microcomputer program for predicting the secondary structure of proteins with a neural network. *CABIOS* **5** (1990) 319-320.
- [144] D. G. Kneller, F. E. Cohen and R. Langridge, Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network. *J. Mol. Biol.* **214** (1990) 171-182.
- [145] S. M. Muskal, S. R. Holbrook and S.-H. Kim, Prediction of the disulfide-bonding state of cysteine in proteins. *Prot. Engin.* **3** (1990) 667-672.
- [146] S. Hayward and J. F. Collins, Limits on  $\alpha$ -helix prediction with neural network models. *Proteins* **14** (1992) 372-381.
- [147] J. W. Shavlik, G. G. Towell and M. O. Noordewier, Using neural networks to refine existing biological knowledge. *Int. J. Genome Res.* **1** (1992) 81-107.
- [148] S. M. Muskal and S.-H. Kim, Predicting protein secondary structure content. A tandem neural network approach. *J. Mol. Biol.* **225** (1992) 713-727.
- [149] M. A. Andrade, P. Chacón, J. J. Merelo and F. Morán, Evaluation of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network. *Prot. Engin.* **6** (1993) 383-390.
- [150] I. Dubchak, S. R. Holbrook and S.-H. Kim, Prediction of protein folding class from amino acid composition. *Proteins* **16** (1993) 79-91.
- [151] P. Fariselli, M. Compiani and R. Casadio, Predicting secondary structures of membrane proteins with neural networks. *Eur. Biophys. J.* **22** (1993) 41-51.
- [152] B. A. Metfessel, P. N. Saurugger, D. P. Connelly and S. S. Rich, Cross-validation of protein structural class prediction using statistical clustering and neural networks. *Prot. Sci.* **2** (1993) 1171-1182.
- [153] M. Reczko, Protein secondary structure prediction with partially recurrent neural networks. In: (eds.), First International Workshop on Neural Networks Applied to Chemistry and Environmental Sciences. Gordon and Breach Science Publ., Lyon, France, 1993, pp. 153-159.
- [154] B. Rost and C. Sander, Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci. U.S.A.* **90** (1993) 7558-7562.
- [155] B. Rost and C. Sander, Secondary structure prediction of all-helical proteins in two states. *Prot. Engin.* **6** (1993) 831-836.
- [156] B. Rost and C. Sander, Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19** (1994) 55-72.
- [157] B. Rost, C. Sander and R. Schneider, PHD - an automatic server for protein secondary structure prediction. *CABIOS* **10** (1994) 53-60.
- [158] F. Sasagawa and K. Tajima, Prediction of protein secondary structures by a neural network. *CABIOS* **9** (1993) 147-152.
- [159] R. Casadio, P. Fariselli, C. Taroni and M. Compiani, A predictor of transmembrane  $\alpha$ -helix domains of proteins based on neural networks. *European Journal of Biophysics* (1994) submitted, 8/94.
- [160] G. W. Dombi and J. Lawrence, Analysis of protein transmembrane helical regions by a neural network. *Prot. Sci.* **3** (1994) 557-566.
- [161] B. Rost and C. Sander, Conservation and prediction of solvent accessibility in protein families. *Proteins* **20** (1994) 216-226.
- [162] T. W. Barlow, Feed-forward neural networks for secondary structure prediction. *J. Mol. Graph.* **13** (1995) 175-183.

- [163] R. Casadio, M. Compiani, P. Fariselli and F. Vivareli, Predicting free energy contributions to the conformational stability of folded proteins from the residue sequence with radial basis function networks. In: C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer and S. Wodak (eds.), Third International conference on Intelligent Systems for Molecular Biology (ISMB). AAAI Press, Cambridge, England, 1995, pp. 81-88.
- [164] J.-M. Chandonia and M. Karplus, Neural networks for secondary structure and structural class predictions. *Prot. Sci.* **4** (1995) 275-285.
- [165] T. Grossman, R. Farber and A. Lapedes, Neural net representations of empirical protein potentials. In: C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer and S. Wodak (eds.), Third International conference on Intelligent Systems for Molecular Biology (ISMB). AAAI Press, Cambridge, England, 1995, pp. 154-161.
- [166] J. E. Hansen, O. Lund, J. Engelbrecht, H. Bohr, J. O. Nielsen, J.-E. S. Hansen and S. Brunak, Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase. *Biochem. J.* **308** (1995) 801-813.
- [167] M. Milik, A. Kolinski and J. Skolnick, Neural network system for the evaluation of side-chain packing in protein structures. *Prot. Engin.* **8** (1995) 225-236.
- [168] R. Casadio, P. Fariselli, C. Taroni and M. Compiani, A predictor of transmembrane  $\alpha$ -helix domains of proteins based on neural networks. *European Journal of Biophysics* **24** (1996) 165-178.
- [169] P. Fariselli and R. Casadio, HTP: a neural network-based method for predicting the topology of helical transmembrane domains in proteins. *CABIOS* **12** (1996) 41-48.
- [170] J. Hanke, G. Beckmann, P. Bork and J. G. Reich, Self-organizing hierarchic networks for pattern recognition in protein sequence. *Prot. Sci.* **5** (1996) 72-82.
- [171] Z. R. Sun, C. T. Zhang, F. H. Wu and L. W. Peng, A vector projection method for predicting supersecondary motifs. *J. Prot. Chem.* **15** (1996) 721-729.
- [172] Z. Sun, X. Rao, L. Peng and D. Xu, Prediction of protein supersecondary structures based on the artificial neural network method. *Prot. Engin.* **10** (1997) 763-769.
- [173] P. Aloy, J. Cedano, B. Oliva, F. X. Avisel and E. Querol, 'TransMem': a neural network implemented in Excel spreadsheets for predicting transmembrane domains of proteins. *CABIOS* **13** (1997) 231-234.
- [174] M. Asogawa, Beta-sheet prediction using inter-strand residue pairs and refinement with Hopfield neural network. *Ismb* **5** (1997) 48-51.
- [175] J. Dopazo and J. M. Carazo, Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol.* **44** (1997) 226-233.
- [176] I. Dubchak, I. Muchnik and S.-H. Kim, Protein folding class predictor for SCOP: approach based on global descriptors. In: T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander and A. Valencia (eds.), Fifth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Halkidiki, Greece, 1997, pp. 104-107.
- [177] T. Kawabata and J. Doi, Improvement of protein secondary structure prediction using binary word encoding. *Proteins* **27** (1997) 36-46.
- [178] O. Lund, K. Frimand, J. Gorodkin, H. Bohr, J. Bohr, J. Hansen and S. Brunak, Protein distance constraints predicted by neural networks and probability density functions. *Prot. Engin.* **10** (1997) 1241-1248.
- [179] P. Arrigo, P. Fariselli and R. Casadio, Can functional regions of proteins be predicted from their coding sequences? The case study of G-protein coupled receptors. *Gene* **221** (1998) GC65-110.
- [180] K. C. Chou and D. W. Elrod, Using discriminant function for prediction of subcellular location of prokaryotic proteins. *Biochem Biophys Res Commun* **252** (1998) 63-68.
- [181] K. Diederichs, J. Freigang, S. Umhau, K. Zeth and J. Breed, Prediction by a neural network of outer membrane beta-strand protein topology. *Prot. Sci.* **7** (1998) 2413-2420.
- [182] A. Reinhardt and T. Hubbard, Using neural networks for prediction of the subcellular location of proteins. *Nucl. Acids Res.* **26** (1998) 2230-2235.
- [183] R. Casadio, M. Compiani, P. Fariselli and P. L. Martelli, A data base of minimally frustrated alpha helical segments extracted from proteins according to an entropy criterion. *Ismb* (1999) 68-76.
- [184] O. Emanuelsson, H. Nielsen and G. von Heijne, ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Prot. Sci.* **8** (1999) 978-984.
- [185] P. Fariselli, P. Riccobelli and R. Casadio, Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. *Proteins* **36** (1999) 340-346.
- [186] P. Fariselli and R. Casadio, A neural network based predictor of residue contacts in proteins. *Prot. Engin.* **12** (1999) 15-21.
- [187] J. Gorodkin, O. Lund, C. A. Andersen and S. Brunak, Using sequence motifs for enhanced neural network prediction of protein distance constraints. *Ismb* (1999) 95-105.
- [188] Y. Guermeur, C. Geourjon, P. Gallinari and G. Deleage, Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics* **15** (1999) 413-421.
- [189] D. T. Jones, Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292** (1999) 195-202.
- [190] A. Krogh and S. K. Riis, Hidden neural networks. *Neural Comput* **11** (1999) 541-563.
- [191] C. Pasquier and S. J. Hamodrakas, An hierarchical artificial neural network system for the classification of transmembrane proteins. *Prot. Engin.* **12** (1999) 631-634.
- [192] A. J. Shepherd, D. Gorse and J. M. Thornton, Prediction of the location and type of beta-turns in proteins using neural networks. *Prot. Sci.* **8** (1999) 1045-1055.

- [193] P. Baldi, G. Pollastri, C. A. Andersen and S. Brunak, Matching protein beta-sheet partners by feedforward and recurrent neural networks. *Ismb* **8** (2000) 25-36.
- [194] O. Emanuelsson, H. Nielsen, S. Brunak and G. von Heijne, Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300** (2000) 1005-1016.
- [195] I. Jacoboni, P. L. Martelli, P. Fariselli, M. Compiani and R. Casadio, Predictions of protein segments with the same aminoacid sequence and different secondary structure: A benchmark for predictive methods. *Proteins* **41** (2000) 535-544.
- [196] M. Ouali and R. D. King, Cascaded multiple classifiers for secondary structure prediction. *Prot.Sci.* **9** (2000) 1162-1176.
- [197] C. T. Workman and G. D. Stormo, ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.* (2000) 467-478.
- [198] A. Babajide, R. Farber, I. L. Hofacker, J. Inman, A. S. Lapedes and P. F. Stadler, Exploring protein sequence space using knowledge-based potentials. *J.Theor. Biol.* **212** (2001) 35-46.
- [199] H. G. Bohr, P. Rogen and K. J. Jalkanen, Applications of neural network prediction of conformational states for small peptides from spectra and of fold classes. *Comput. Chem.* **26** (2001) 65-77.
- [200] C. H. Ding and I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **17** (2001) 349-358.
- [201] P. Fariselli, O. Olmea, A. Valencia and R. Casadio, Prediction of contact maps with neural networks and correlated mutations. *Prot. Engin.* **14** (2001) 835-843.
- [202] I. Jacoboni, P. L. Martelli, P. Fariselli, V. De Pinto and R. Casadio, Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. *Prot. Sci.* **10** (2001) 779-787.
- [203] C. Pasquier, V. J. Promponas and S. J. Hamodrakas, PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide applications. *Proteins* **44** (2001) 361-369.
- [204] H. X. Zhou and Y. Shan, Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* **44** (2001) 336-343.
- [205] P. Baldi, S. Brunak, P. Frasconi, G. Soda and G. Pollastri, Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* **15** (1999) 937-946.
- [206] P. Baldi and G. Pollastri, Machine learning structural and functional proteomics. *IEEE Intelligent Systems* (2001) in press.