J Mol Evol (1993) 36:586–595

Journal
of Molecular
Evolution

Development of Artificial Neural Filters for Pattern Recognition in Protein Sequences

Gisbert Schneider[1,2] and Paul Wrede[1]

[1]Freie Universität Berlin, Fachbereich Physik, AG Biophysik, Arnimallee 14, D-1000 Berlin 33
[2]Technische Universität Berlin, Fachgebiet Bionik & Evolutionstechnik, Ackerstraße 71-76,
D-1000 Berlin 65, Federal Republic of Germany

Summary. Four different artificial neural network architectures have been tested for their suitability to extract and predict sequence features. For optimization of the network weights an evolutionary computing method has been applied. The networks have feedforward architecture and provide adaptive neural filter systems for pattern recognition in primary structures and sequence classification. The recognition and prediction of signal peptidase cleavage sites of *E. coli* periplasmic protein precursors serves as an example for filter development. The primary structures are represented by seven physicochemical residue properties. This amino acid description provides the feature space for network optimization. The properties hydrophobicity, hydrophilicity, side-chain volume, and polarity allowed an accurate classification of the data. A three-layer network architecture reached a learning success of 100%; the highest prediction accuracy in an independent test set of sequences was 97%. This network architecture appears to be most suited for the analysis of *E. coli* signal peptidase cleavage sites. Further suggestions about the design and future applications of artificial neural networks for protein sequence analysis are made.

Key words: Evolution strategy — Feature extraction — Filter induction — Neural network — Prediction — Signal peptidase

A reliable prediction of a protein's structure and function from its amino acid sequence with high accuracy is still not possible in many cases (Fasman 1989). Most successful is the prediction of transmembrane regions of membrane proteins which is based on the recognition of hydrophobic amino acid stretches in the protein primary structure (Kyte and Doolittle 1982; Jähnig 1990). The prediction of secondary structure reaches an overall accuracy of up to 64% with the application of artificial neural networks (Qian and Sejnowski 1988; Bohr et al. 1988; Holley and Karplus 1989) or hybrid statistical and neural systems (Stolorz et al. 1992; Zhang et al. 1992). First attempts to predict tertiary structure of protein backbones have been made (Bohr et al. 1990). Such neural filter systems for structure prediction are regarded as useful since neural networks allow both pattern recognition and sequence classification by a single system (Hirst and Sternberg 1992). We investigated several network architectures for their suitability as protein sequence filters and present a method for the construction of artificial neural networks for the prediction of *E. coli* signal peptidase cleavage sites in precursor sequences of periplasmic proteins.

A neural network must manage two tasks: feature extraction from the sequence data and classification of the sequence examples according to the extracted feature. In most neural network applications protein sequences are represented by binary numbers coding for the amino acids (Hirst and Sternberg 1992). In contrast, our approach is based on the extraction of sequence features from a nu-

Offprint requests to: P. Wrede

merical description of the primary structures by amino acid property values (real numbers). Seven properties were used, providing the feature space: hydrophobicity, hydrophilicity, polarity, surface area, volume, bulkiness, and refractivity of the residues. Compared to binary scales, the main advantage of the sequence description by real coded amino acid property scales is the generation of a feature space which is based on the physicochemical similarity of amino acids. This allows the development of a prediction system which is not based on the analogy of character strings but on the similarity of primary structures with regard to their physicochemical properties. Therefore, features basing on chemical theory can be extracted. These features are used for the classification of sequences. In our approach, the classification of the sequence examples by the neural filters was restricted to a binary decision: A given example obtains a cleavage site, or it does not. This imitates the biological signal recognition: A precursor sequence is processed by signal peptidase at a particular cleavage site, or it is not. For prediction of the cleavage sites the sequences are scanned by the neural filters, analyzing a sequence window of 13 residues at a time.

The development of neural filters is separated into a training phase and a test phase. During the filter training optimal values for the network's connection weights are determined. Since these values cannot be numerically calculated, their determination is an optimization problem. An evolution strategy has been used for this purpose (Rechenberg 1973). In the test phase the obtained optimized filters are evaluated with regard to their prediction accuracy by application to a test set of sequences which is distinct from the training set.

We chose the prediction of signal peptidase cleavage sits as example for several reasons:

1. The cleavage-site regions of the different signal peptides show no conserved stretches of amino acids (Perlman and Halvorson 1983; von Heijne 1983). This fact demonstrates that a sequence description based on the similarity of the amino acid residues may be useful for a successful feature extraction.
2. The signal for cleavage-site recognition by signal peptidase appears to be locally encoded (Laforet and Kendall 1991; Schneider and Wrede 1993). Since long-range interactions between residues cannot be taken into consideration by the chosen network architectures, the accurate prediction of a locally encoded signal is likely.
3. Another prediction method for signal peptidase cleavage sites which is based on a statistical approach is available (von Heijne 1986). The neural filters can be compared to this statistical method.

Table 1. The names of the *E. coli* periplasmic proteins used for the filter induction and the prediction experiments

Training-set sequences	Test-set sequences
Glucose-1-phosphatase precursor	Sulfate-binding protein precursor
L-arabinose-binding protein precursor	Periplasmic trehalase precursor
Lysine-arginine-ornithine-binding protein precursor	Glycerol-3-phosphate-binding protein precursor
L-asparaginase II precursor	UDP-sugar hydrolase precursor
Peptidyl-prolyl *cis-trans* isomerase precursor	Protease III precursor
D-galactose-binding protein precursor	D-ribose-binding periplasmic protein precursor
Gamma-glutamyltranspeptidase precursor	Ribonuclease I precursor
Glutamine-binding protein precursor	
Leu/Ile/Val-binding protein precursor	
Leucine-specific binding protein precursor	
Maltose-binding protein precursor	
Pennicillin-insensitive murein endopeptidase precursor	
Pennicillin acylase precursor	
Periplasmic phosphate-binding protein precursor	
pH 2.5 acid phosphatase precursor	
Alkaline phosphatase precursor	
Periplasmic glycine betaine-binding protein precursor	

Methods

Data. All protein sequences were collected from the SwissProt Database, release 18 (IntelliGenetics, Inc.). Twenty-four precursor sequences of *E. coli* periplasmic proteins with experimentally confirmed signal peptidase cleavage sites were found. These data were split into a training set of 17 sequences and a test set of 7 sequences (Table 1) to give a random 7:3 distribution between training and test examples. The neural-network filters were applied to these precursor sequences for cleavage-site prediction.

For training of the neural networks the sequences were restricted to strings of 13 residues in length. Both training and test set consist of positive and negative examples. The positive examples cover the positions -10 to $+3$ (Fig. 1). Two different training sets for the feature extraction were used. They have in common the positive examples. The negative examples are randomly chosen strings (13 residues each) from the precursor sequences in set 1 and 13 residue strings from the cleavage site region (-12 to $+5$) in training set 2 (Fig. 1). The idea is that different filters can be obtained by the use of different training sets: set 1 allows the extraction of general cleavage-site features, and the corresponding neural filters are thought to discriminate between cleavage sites and noncleavage sites. In contrast, set 2 is thought to allow the construction of filters which are special-

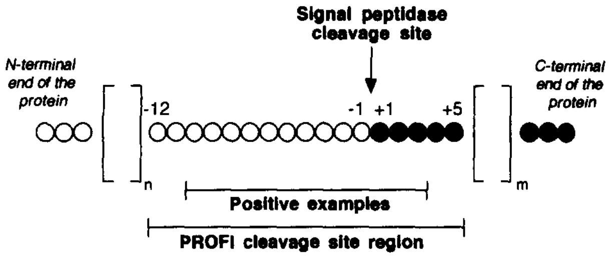
Signal peptidase cleavage site
N-terminal end of the protein
C-terminal end of the protein
-12
-1 +1
+5
n
m
Positive examples
PROFI cleavage site region

Fig. 1. The data preparation. The targeting sequence positions have negative numbers from the cleavage site to the N-terminal end of the protein. The positions in the mature part of the protein are indicated by positive numbers. The positive examples of the training and test sets cover the positions −10 to +3. The negative examples are randomly selected sequence strings (length 13) from the targeting sequence region −12 to +5 (training-set 2) or from the whole precursors (training-set 1).

ized on the detection of the exact cleavage-site position. The training sets consist of 17 positive examples and 68 negative examples; the test sets consist of seven positive and 28 negative examples for the filter induction experiments. Thus, the fraction of positive examples is 20% in all sequence sets.

We are aware that the small number of test-set cleavage-site examples is not sufficient for a reliable statistical evaluation of the prediction results. The training set, too, is small compared to other training sets for neural networks (Qian and Sejnowski 1988; Bohr et al. 1988; Holley and Karplus 1989). The idea was to show that our training technique is successful even with few training data, since for many biochemical prediction problems only few data are available.

The Network Architectures. Four different networks architectures have been used for the development of cleavage-site filters (Fig. 2). They were named "PROFI 1–4" (*PRO*tein *F*ilter *I*nduction), since the training technique is based on a top–down induction strategy. (See below.) All PROFI networks have feedforward architecture: The flow of information is unidirectional from the input layer toward the output-layer unit. The four networks differ in the number of hidden layers and hidden-layer units (Fig. 2). Further, the number of parallel amino acid representations is increased from PROFI 2 (a single physicochemical property) to PROFI 3 (four physicochemical properties). Thus, the size of the input layer is 13 units for PROFI 1 and PROFI 2; it is 52 (13 × 4) in the PROFI 2 and PROFI 4 systems, respectively. The neural filters calculate the boolean values TRUE or FALSE as prediction result with regard to the peptide bond between the relative window positions 10 and 11 (Fig. 1).

The *input-layer units* convert an amino-acid character (single-letter code) to a numerical property value (Table 2). For the experiments, these values were normalized to obtain comparable property scales between −1.0 and 1.0. No further calculations are performed by an input-layer unit ("fan-out" unit). In contrast, the output of a *hidden-layer unit* is determined by its transfer function F(*unit*$_{in}$):

$$\mathrm{F}(unit_{in}) = \frac{1}{1 + e^{-unit_{in}}}$$

This is the common sigmoid function (Fermi function) which limits the output of a unit to values $0.0 < \mathrm{F}(unit_{in}) < 1.0$. The term *unit*$_{in}$ is the total input of a unit *i* (the weighted input sum), w_{ij} stands for the connection weights, ξ_{ij} is the output of a unit of the previous layer:

$$unit_{in} = \sum_{j=1}^{n} w_{ij}\xi_{ij}$$

All hidden layer units use the same sigmoid transfer function. The single *output-layer unit* uses the step function $\Theta(x)$:

$$\Theta(x) = \Theta[F(unit_{in})] = \begin{cases} \text{TRUE} & \text{if } F(unit_{in}) \geq \text{threshold} \\ \text{FALSE} & \text{otherwise} \end{cases}$$

The threshold was 0.5 in all experiments. With the help of this unit step function, a binary sequence classification is achieved: A sequence example will be regarded as a cleavage-site example if the output "TRUE" is calculated. It will be regarded as a "noncleavage-site" example otherwise.

The four PROFI systems are now described in more detail. Since neural networks can be regarded as "function estimators" (Kosko 1992), the transformation functions which are calculated by the different PROFI networks are given. The used variable names are explained in Fig. 2.

- *PROFI 1:* This network is a Perceptron system (Minsky and Papert 1988). It consists of two layers, an input layer and an output layer with the single classifying unit. PROFI 1 is limited to linear separation between positive and negative training-set examples because of the use of the transfer function *F*(*unit*$_{in}$). In contrast to the classical Perceptron (Rosenblatt 1962), the output unit has a fixed threshold value of 0.5 instead of a calculated value. The transformation of an input pattern $\{x_k\}$ is given by:

$$\text{output} = \theta[F(unit_{in})] = \theta\left[F\left(\sum_{j} w_{ij}x_k\right)\right]$$

  With PROFI 1 no feature extraction employing higher-order correlations between the different sequence positions and residue properties is performed, since no hidden layer is present. Only first-order correlations between the different window positions, expressed as connection weights $\{w_{ij}\}$, are taken into consideration for classification.
- *PROFI 2:* It has been proved that one hidden layer is sufficient to approximate any continuous function (Cybenko 1989; Hornik et al. 1989). The PROFI 2 architecture can therefore be regarded as a "minimal architecture" for the development of protein sequence filters. The hidden layer allows the extraction of features such as "contrast" for sequence classification. As with PROFI 1, the PROFI 2 networks use a single amino acid property in the input layer. Thus, the sequence classification is based only on the features which can be extracted from a single property description of the protein sequences. The output of the three-layer network PROFI 2 is given by:

$$\text{output} = \theta\left\{F\left[\sum_{j} w_{ij}F\left(\sum_{k} w_{jk}x_k\right)\right]\right\}$$

  To determine the optimal number of hidden-layer units the prediction qualities of optimized PROFI 2 filters with one to 13 hidden units were measured.

Table 2. The values of the amino-acid properties used for the description of the protein sequences

Amino acid	Hydrophobicity[a]	Volume[b]	Surface area[c]	Hydrophilicity[d]	Bulkiness[e]	Refractivity[f]	Polarity[g]
A	1.6	88.6	115	−0.5	11.50	4.34	0.00
R	−12.3	173.4	225	3.0	14.28	26.66	52.00
N	−4.8	117.7	160	0.2	11.68	12.00	49.70
D	−9.2	111.1	150	3.0	12.82	13.28	3.38
C	2.0	108.5	135	−1.0	13.46	35.77	1.48
Q	−4.1	143.9	180	0.2	13.57	17.26	49.90
E	−8.2	138.4	190	3.0	14.45	17.56	3.53
G	1.0	60.1	75	0.0	3.40	0.00	0.00
H	−3.0	153.2	195	−0.5	13.69	21.81	51.60
I	3.1	166.7	175	−1.8	21.40	19.06	0.13
L	2.8	166.7	170	−1.8	21.40	18.78	0.13
K	−8.8	168.6	200	3.0	15.71	21.29	49.50
M	3.4	162.9	185	−1.3	16.25	21.64	1.43
F	3.7	189.9	210	−2.5	19.80	29.40	0.35
P	−0.2	122.7	145	0.0	17.43	10.93	1.58
S	0.6	89.0	115	0.3	9.47	6.35	1.67
T	1.2	116.1	140	−0.4	15.77	11.01	1.66
W	1.9	227.8	255	−3.4	21.67	42.53	2.10
Y	−0.7	193.6	230	−2.3	18.03	31.53	1.61
V	2.6	140.0	155	−1.5	21.57	13.92	0.13

[a] From (Engelman et al. 1986)
[b] From (Zamyatnin 1972)
[c] From (Chothia 1975)
[d] From (Hopp and Woods 1981)
[e] From (Jones 1975)

- *PROFI 3:* In contrast to PROFI 2, which employs a sequence description by a single amino acid property, the PROFI 3 architecture uses four amino acid properties for the numerical description of the protein primary structures. The selection of these properties is based on the results of the PROFI 2 experiments: The properties which led to the highest prediction qualities of PROFI 2 filters have been chosen. The use of four properties allows the extraction of complex features, such as "contrast between two amino acid properties." The number of hidden-layer units was changed systematically between one and 13 to determine an optimal network architecture. The output of a PROFI 3 three-layer network is given by:

$$\text{output} = \theta\left\{F\left[\sum_{j} w_{ij}F\left(\sum_{k,l} w_{jk,l}x_{k,l}\right)\right]\right\}$$

- *PROFI 4:* Compared to PROFI 3 a second hidden layer is introduced into this network. This provides additional variables (network weights) for the description of an input–output transformation function. The number of hidden-layer units is the same for both hidden layers in all experiments. As for PROFI 2 and 3, the number of hidden-layer units was systematically altered between one and 13 to determine an optimal network architecture. The output of the four-layer network PROFI 4 is given by:

$$\text{output} = \theta\left(F\left\{\sum_{j} w_{ij}F\left[\sum_{k} w_{jk}F\left(\sum_{l,m} w_{kl,m}x_{l,m}\right)\right]\right\}\right)$$

The Training Technique. The goal was to adjust all network weights in such a way that the output unit calculated an output value above or equal to a threshold in the case a positive example (a cleavage-site region) was presented at the filter input. This leads to the final binary output "TRUE." A value below the threshold must be calculated in case of negative examples ("FALSE"). The threshold of the output-layer unit was 0.5 in all experiments. To measure the learning success during the training phase and to determine the actual quality of a filter, a quality index Q was calculated. We defined the prediction accuracy of a filter as its quality: Q is the sum of positive (P) and negative correct predictions (N) divided by the total number of examples (T) in the training set:

$$Q = \frac{P + N}{T}$$

If all examples are classified in such a way that positive examples produce a filter output $\geqslant 0.5$ and negative examples produce a value < 0.5, the value of Q will be 1.0 (100% correct classification). This quality function serves as a simple heuristic to separate the feature space into regions of high and low quality. The point of highest quality provides the weight values $\{w\}$ for the filter with the highest prediction accuracy.

A (μ,λ) evolution strategy with adaptive stepsize control has been used for training (Rechenberg 1973). This top–down search strategy includes repeated generate and test cycles ("generations" or "learning cycles") for a systematic generation and test of variable values. The alteration of the values is achieved by a mutation procedure which is done by adding Gaussian distributed random numbers to the old parameter values. The values leading to the highest filter quality are selected for the next learning cycle. This stepwise learning of examples can be regarded as an inductive process which leads to the "proof" of the generated set of variable values. The number of mutations of a weight per generation (λ) was 500; the number of parents (μ) (selection of the best) was one per generation. Initially, all network weights were random numbers between − 1.0 and 1.0. One generate and test cycle consists of (1) λ mutations of the parental value, (2) quality determination of the offspring (calculation of Q), and (3) selection of the mutation with the highest quality as the new

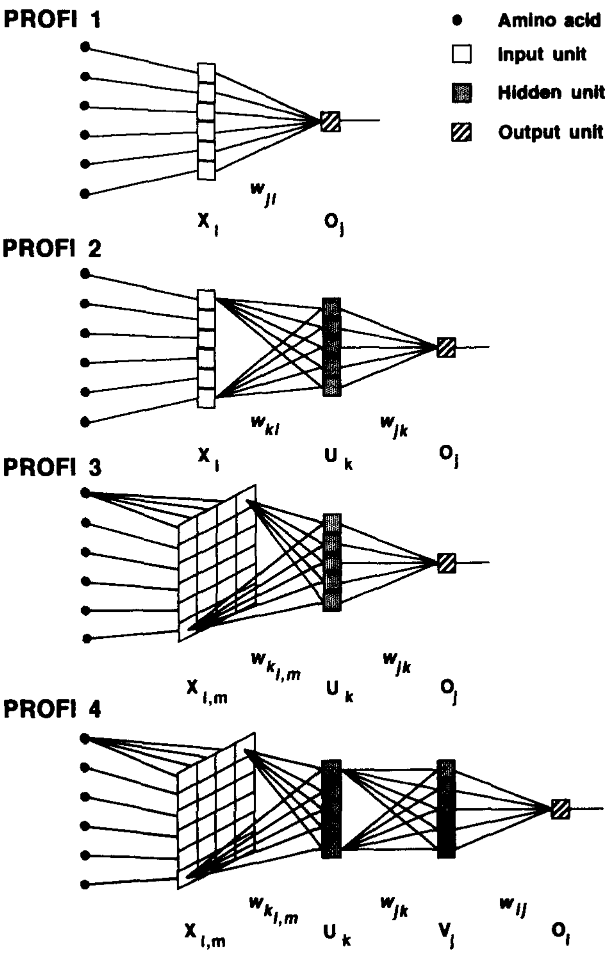

Fig. 2. The PROFI network architectures. For clarity only some connections between the units are shown for PROFI 2–4. For details, see text.

parent value (μ). The number of training cycles was limited to 1,000 for every experiment.

The offspring of a generation (w_{new}) show a Gaussian distribution (*G:* Gaussian-distributed random number) around the parental value (w_{old}), which means that most mutations are not significantly different from the parental value:

$$w_{new} = w_{old} + \delta G$$

These small changes per generation allow a stepwise hill-climbing toward the quality optimum. To avoid getting stuck in "local quality maxima" of the quality function the learning rate δ (stepsize) was set at its initial value of 0.3 every 100 generations. The initial value of 0.3 has been empirically chosen, for it leads to a fast initial learning progress (rapid increase of quality per generation). Besides these fixed changes, the stepsize δ itself was a free parameter during the whole training phase. It was mutated like the weight values, and the stepsize leading to the best weights of a learning cycle was selected as the parent value for the next cycle. There certainly are local maxima (or minima) in some optimization problems (McInerny et al. 1989), though it is not known whether the quality function used in the PROFI approach actually shows such maxima. The filter induction and the prediction experiments were performed on a PC running under DOS (80486 processor). All PROFI systems are implemented in Modula 2.

The Prediction Method

To determine the overall prediction quality of the induced filters the precursor sequences of the training and the independent test set were scanned. Starting with the N-terminal end of the protein sequence (position *n*, $n = 1$) a window of 13 residues is analyzed at a time. In the case in which the filter produces the output "TRUE," the corresponding sequence position will be regarded as a cleavage site. The filter moves to the next sequence position (position $n + 1$) and the new filter output is calculated. This procedure is repeated until the C-terminus of the precursor sequence is reached.

To determine the prediction power and accuracy of the PROFI system single filters and combinations of filters were used. In the latter case all combined filters must produce the output "TRUE" for the same sequence position to get a positive cleavage-site prediction (logical AND connection). This method was thought to produce more reliable results. Three quality indices (Schulz and Schirmer 1979) were calculated to evaluate and compare the prediction results of the PROFI method and the statistical approach:

$$Q1 = p + n \qquad Q2 = \frac{p}{p + o} \qquad Q3 = \frac{p}{p + u}$$

p: Number of correctly predicted cleavage sites
n: Number of correctly predicted "noncleavage sites"
o: Number of incorrectly predicted "noncleavage sites" (overprediction)
u: Number of incorrectly predicted cleavage sites (underprediction)

*Q*1 provides a measure for the ability of a filter to discriminate between cleavage sites (*p*) and noncleavage sites (*n*), although it is dominated by the negative correct prediction. *Q*2 and *Q*3 allow one to determine the degree of overprediction and underprediction, respectively.

In contrast to the PROFI method, the prediction of cleavage sites with the statistical method uses a sequences window of 12 residues (von Heijne 1986). It is mainly based on the detection of the "−1,−3 rule" (von Heijne 1983; Perlman and Halvorson 1983) in the cleavage-site region.

Results

Filter Induction Without Hidden Layers (PROFI 1)

For the seven property scales seven PROFI 1 filters were obtained. Training set 1 was used. It is striking that some amino acid properties are useful for the analysis of signal peptidase cleavage sites, and some are not. Hydrophobicity, polarity, and surface area appear to be essential residue properties for an accurate description of the sequence data. These filters show the highest training-set qualities (Table 3). The test-set qualities (test set 1) are rather poor; only the hydrophobicity filter reaches an classification accuracy of 91% with the test-set examples.

Table 3. The prediction quality $Q1$ of the different PROFI 1 filters which were applied to training-set 1 and test-set 1

	Quality $Q1$ (%)	
Property	Training-set 1	Test-set 1
Hydrophobicity	96	91
Volume	91	83
Surface area	95	86
Hydrophilicity	88	74
Polarity	92	86
Bulkiness	80	80
Refractivity	81	80

Filter Induction with One Hidden Layer (PROFI 2 and PROFI 3)

The induction experiments with training set 1 led to a total of seven PROFI 2 filters for the *E. coli* signal peptidase cleavage site of periplasmic proteins. The number of hidden-layer units was 10 in all PROFI 2 experiments. (See below and Fig. 3A.) Each of the filters makes use of a different chemophysical property of the amino acids for the description of the protein sequences (Table 2). The learning success (filter quality in training set 1) was high for all properties (Table 4), indicating that the sequence descriptions reveal striking patterns that allow a sequence classification. The prediction quality in the independent test set is about the same as the training-set quality only for the property hydrophobicity. All other filters show an overfitting to the training data (Table 4), especially the filters employing the refractivity-and-bulkiness scale. These filters are specialized on the training sequences. The surface-area filter shows the poorest learning success (81%), which corresponds with its test-set quality (80%). From the PROFI 2 results the properties hydrophobicity, volume, hydrophilicity, and polarity were selected as input-layer properties for the PROFI 3 system because of their high learning success.

The optimization of the hydrophobicity filter served as an example for the experimental determination of the optimal number of hidden-layer units for PROFI 2. Training set 1 and test set 1 were used. The filter quality after 100 generations was plotted against the number of hidden-layer units (Fig. 3A). It turned out that four or 10 units lead to the highest test-set quality (97%) and therefore allow the detection of general cleavage-site patterns. Different numbers of hidden-layer units lead to a specialization of the filter on the training-set examples as indicated by the much higher training-set quality compared to the test-set quality (Fig. 3A). This observation of overfitting (overlearning) is striking with six or seven hidden-layer units. As a result of these investigations, we used a hidden layer with 10 units for the development of neural filters with PROFI 2.

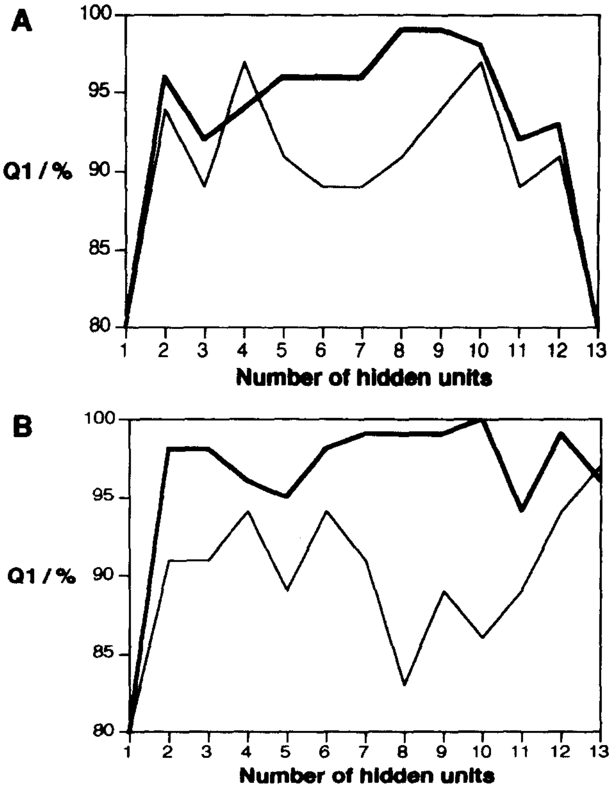

Fig. 3. The selection of the optimal number of hidden-layer units for the PROFI 2 network which employs the amino acid property "hydrophobicity" in the input layer (A) and for the PROFI 3 network (B). The training-set qualities (reclassification) are drawn in *thick lines;* the *thin lines* show the filter qualities for the test set (classification). Training-set 1 and test-set 1 were used in both experiments.

Table 4. The prediction quality $Q1$ of the different PROFI 2 filters which were applied to training-set 1 and test-set 1

	Quality $Q1$ (%)	
Property	Training-set 1	Test-set 1
Hydrophobicity	98	97
Volume	95	83
Surface area	81	80
Hydrophilicity	93	86
Polarity	95	89
Bulkiness	89	80
Refractivity	91	80

On the other hand, PROFI 3 filters show the highest test-set qualities with 4, 6, 12, or 13 hidden-layer units. The correct classification reaches 97% with 13 units (Fig. 3B). These four filters were used in the prediction experiments. Additional PROFI 3 filters for prediction were obtained by performing an filter induction with training set 2. The four selected filter architectures were used. They all led to

Table 5. The prediction quality $Q1$ of the four selected PROFI 3 filters which were applied to training-set 2 and test-set 2

	Quality $Q1$ (%)	
Filter	Training-set 2	Test-set 2
4 hidden units	99	83
6 hidden units	99	91
12 hidden units	99	91
13 hidden units	99	86

Table 6. The prediction quality $Q1$ of the different PROFI 4 filters which were applied to training-set 1 and test-set 1

No.	Quality $Q1$ (%)	
hidden units	Training-set 1	Test-set 1
1	80	80
2	80	43
3	95	26
4	93	60
5	94	14
6	96	71
7	94	51
8	99	29
9	92	25
10	93	37
11	99	57
12	99	57
13	98	49

a learning success of 99% (Table 5). Here, the filters with 6 and 12 hidden units allowed the most accurate classification of the test set 2 examples (91%).

Filter Induction with Two Hidden Layers (PROFI 4)

Two hidden layers led to high learning success in most cases, but the neural filters could not accurately classify the test-set examples (Table 6). The high training-set qualities up to 99% clearly show that the feature extraction was successful. The low test-set qualities indicate that only special training-set features were found. Thus, PROFI 4 filters were not used for the prediction experiments. The PROFI 4 architecture leads to filters which learn the training-set examples "by heart." The development of generalizing sequence filters for the prediction of signal peptidase cleavage sites is not possible with these systems. We think that the PROFI 4 architecture offered too many variables for the description of cleavage sites. This may be further support for the idea that signal peptidase cleavage sites are characterized by locally encoded signals since only a limited number of variable network parameters (PROFI 3 systems) allow the extraction of general cleavage-site features.

The Prediction Experiments

The 17 training-set and the seven test-set precursor sequences were scanned for cleavage-site prediction to determine the reclassification and classification power of the neural filters. Only PROFI 3 filters were used because the PROFI 3 architecture led to the highest filter qualities when applied to the corresponding 28 examples of the test sets 1 and 2 (Fig. 3B, Table 5). Three different prediction runs were performed with the different sets of neural filters:

1. Prediction with the best filters which were optimized with training-set 1 (Table 7)
2. Prediction with the best filters which were optimized with training-set 2 (Table 8)
3. Prediction with combinations of the best single filters from experiments 1 and 2 (Table 9)

A general observation is that all filters show a higher reclassification quality (training-set prediction) than the corresponding classification quality (test-set prediction) for all three quality functions $Q1$, $Q2$, and $Q3$. Further, the qualities $Q1$ and $Q2$ are higher for the filter combinations than for the single filters, whereas $Q3$ is decreased. The reason is that single filters lead to many more "TRUE" predictions for a precursor sequence than combinations of filters do. As a consequence, the number of incorrectly predicted positions is high (low $Q1$ values) and a lot of overprediction occurs (low $Q2$ values). But in most cases a cleavage site is correctly predicted and only some cleavage sites are missed. This is indicated by high $Q3$ values, which can be regarded as a measure for underprediction. In one case a $Q3$ value of 100% is reached by a single filter (Table 7).

In general, no great differences in the prediction accuracy between the filters from training-set 1 and training-set 2 can be observed. This is remarkable since the different filter types are thought to employ different features for sequence classification. The best prediction can be achieved with the combinations of all four single filters which allow a highly accurate discrimination between cleavage sites and noncleavage sites ($Q1$) although not all cleavage sites are recognized ($Q3$) (Tables 7, 8).

The best single filters from training-set 1 have 12 and 13 hidden-layer units; the best filter from training-set 2 has 12 hidden-layer units. These three neural filters were selected because of their high classification qualities. Combinations of them were used in the third prediction experiment (Table 9). Again, a striking increase of prediction accuracy

Table 7. The prediction results of the PROFI 3 filters which were optimized with training-set 1: AND is the logical AND of the output values

	Training data			Test data		
Filter	*Q*1	*Q*2	*Q*3	*Q*1	*Q*2	*Q*3
4 hidden units	0.891	0.022	0.882	0.897	0.014	0.714
6 hidden units	0.944	0.042	0.941	0.984	0.026	0.714
12 hidden units	0.877	0.021	0.941	0.901	0.017	0.857
13 hidden units	0.846	0.018	1.000	0.851	0.012	0.857
4 AND 6	0.972	0.070	0.882	0.972	0.037	0.571
4 AND 12	0.960	0.050	0.824	0.968	0.033	0.571
4 AND 13	0.936	0.037	0.882	0.943	0.024	0.714
6 AND 12	0.974	0.090	0.882	0.985	0.061	0.571
6 AND 13	0.977	0.088	0.941	0.974	0.039	0.571
12 AND 13	0.948	0.045	0.941	0.953	0.029	0.714
4 AND 6 AND 12	0.988	0.121	0.824	0.990	0.067	0.429
6 AND 12 AND 13	0.991	0.150	0.824	0.990	0.065	0.429
4 AND 6 AND 12 AND 13	0.994	0.170	0.824	0.994	0.091	0.429

Table 8. The prediction results of the PROFI 3 filters which were optimized with training-set 2: AND is the logical AND of the output values

	Training data			Test data		
Filter	*Q*1	*Q*2	*Q*3	*Q*1	*Q*2	*Q*3
4 hidden units	0.895	0.017	0.882	0.865	0.009	0.571
6 hidden units	0.701	0.008	0.882	0.700	0.005	0.714
12 hidden units	0.916	0.028	0.882	0.918	0.021	0.857
13 hidden units	0.903	0.024	0.882	0.910	0.016	0.714
4 AND 6	0.895	0.023	0.882	0.901	0.012	0.571
4 AND 12	0.966	0.060	0.882	0.966	0.031	0.571
4 AND 13	0.964	0.058	0.882	0.963	0.029	0.571
6 AND 12	0.955	0.048	0.882	0.958	0.032	0.714
6 AND 13	0.949	0.044	0.882	0.953	0.024	0.571
12 AND 13	0.977	0.083	0.882	0.979	0.048	0.571
4 AND 6 AND 12	0.973	0.073	0.882	0.974	0.040	0.571
6 AND 12 AND 13	0.985	0.109	0.882	0.985	0.063	0.571
4 AND 6 AND 12 AND 13	0.991	0.150	0.882	0.990	0.083	0.571

Table 9. The prediction results of the selected best neural filters[a]

	Training data			Test data			Small test data		
Filter	*Q*1	*Q*2	*Q*3	*Q*1	*Q*2	*Q*3	*Q*1	*Q*2	*Q*3
12* AND 12	0.982	0.099	0.882	0.984	0.071	0.714	0.999	0.294	0.714
12* AND 13	0.976	0.081	0.882	0.977	0.063	0.857	0.999	0.333	0.857
12* AND 13 AND 12	0.992	0.165	0.882	0.993	0.125	0.714	0.999	0.417	0.714

[a] The column "Filter" gives the numbers of hidden-layer units. The asterisk indicates the filter which was optimized with training-set 2; the other two filters were optimized with training-set 1. AND is the logical AND of the output values.

could be achieved and underprediction was reduced: *Q*3 reached 85.7%, which means that most of the cleavage sites were correctly predicted. Overprediction was reduced, too (*Q*2). If only the first 50 residues of the precursor sequences were scanned ("small test data"), a *Q*1 quality of 99.9% could be obtained (Table 9). These filters allowed a discrimination between a cleavage site and any other noncleavage site with nearly absolute certainty.

Discussion

It could be shown that a simple neural filter system can predict signal peptidase cleavage sites with high accuracy. We are well aware that the small data set does not allow a general evaluation. Rather, the development of the special filter architecture used for this prediction task was our intention. The "optimal" filter consists of an input layer which employs at least four physicochemical amino acid properties for the sequence description (hydrophobicity, hydrophilicity, polarity, and volume), one hidden layer for the feature extraction, and a single output layer for classification. Here, we trained the systems to produce a binary output (TRUE or FALSE). The use of linear output values leads to similar results (data not shown). We conclude that the use of two hidden layers of the same size was inadequate for the filter development. These filters could not extract general features from the data. They were rather specialized on the training-set examples. Whether this architecture may be useful for application to different prediction problems is unclear. This could be the case when large sequence windows (more than 20 residues) are employed as input patterns or the number of parallel amino acid properties in the input layer is drastically increased.

From the prediction experiments it is clear that the selection of the training examples is another important step for the whole filter development: Different training sets allow the extraction of different sequence features which can be equally valid for the protein structure or function under investigation. The combination of only two different filters already leads to a striking increase of prediction accuracy compared to filters which utilize only a single feature for sequence classification. Thus, we recommend the use of at least two different training sets for further filter induction experiments.

A statistical method for the prediction of eubacterial signal peptidase cleavage sites is known from literature (von Heijne 1986). It is reported to have an accuracy of 70–80%. Our neural filter approach leads to higher-quality values around 90%. We are aware that this result is not representative, since only seven test-set sequences were investigated. In contrast to von Heijne (1986), we have restricted the networks to focus on *E. coli* sequences to obtain species-specific filters. Thus, we obtained less sequences with known cleavage sites. Nevertheless, the neural filters for signal peptidase cleavage sites provide at least a second method for the prediction of cleavage sites which is independent from the statistical approach. Since the neural networks store the extracted sequence features in a distributed, nonsymbolic way it is impossible to give explicite cleavage site features such as the "$-1,-3$ rule" (von Heijne 1983, 1986; Perlman and Halvorson 1983) or sequence "descriptors" which can be obtained by the use of symbolic methods (Gascuel and Danchin 1986; Schneider and Wrede 1993). Thus, the obtained filters must be regarded as a "black box" prediction system.

We conclude that the PROFI method provides a first simple system for the development of neural sequence filters employing physicochemical amino acid properties, although its general applicability remains to be tested. The use of residue properties was helpful for the analysis and prediction of cleavage sites. It is not guaranteed that this holds for any prediction task—e.g., secondary structure prediction. It should be interesting to apply the PROFI system to this problem since the neural network methods which were first published (Quian and Sejnowski 1988; Holley and Karplus 1989) did not use any additional sequence information besides the sequence character code. These systems are inferior to classical prediction methods like SIMPA (Levin et al. 1988) and PROMIS (King and Sternberg 1990). It is not proven that a neural network can perform prediction tasks which, in principle, cannot be solved otherwise. Nevertheless, the use of additional information such as physicochemical amino acid properties in the cleavage-site prediction task might play a key role in the development of further neural networks for protein sequence analysis (Hirst and Sternberg 1992).

It must be stressed that an optimization of the empirical network parameters—e.g., the size of the input layer, the transfer function itself, or the ratio between positive and negative examples—must be done for every new application. A systematic approach for this task has already been proposed (Lohmann 1992). Unfortunately, a neural filter basing on the PROFI architecture cannot take into consideration amino acid interactions from residues which are spaced far apart on the protein sequence. Thus, it is limited to locally encoded protein functions and structures in its present state. The next steps will be the development of filters for the prediction of membrane protein topology and the prediction of secondary structures from the amino acid sequence. The first experiments with these new prediction systems already are very promising.

Acknowledgments. We thank Reinhard Lohmann for support and his advice and Georg Büldt and Ingo Rechenberg for helpful discussions and encouragement. We are thankful to the Fonds der Chemischen Industrie (FCI) and the Deutsche Forschungsgemeinschaft (DFG, SfB 312) for financial support.

References

Bohr H, Bohr J, Brunak S, Cotterill RMJ, Lautrup B, Norskov L, Olsen O, Petersen SB (1988) Protein secondary structure and homology by neural networks. FEBS Lett 241:223–228

Bohr H, Bohr S, Brunak S, Cotterill RMJ, Fredholm H, Lautrup B, Petersen SB (1990) A novel approach to prediction of the

3-dimensional structures of protein backbones byneural networks. FEBS Lett 261:43–46

Cybenko G (1989) Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems 2:303–314

Chothia C (1975) The nature of accessible and buried surfaces in proteins. J Mol Biol 105:1–14

Engelman DA, Steitz TA, Goldman A (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biophys Chem 15:321–353

Fasman G (ed) (1989) Prediction of protein structure and the principles of protein conformation. Plenum Press, New York

Gascuel O, Danchin A (1986) Protein export in prokaryotes and eukaryotes: indications of a difference in the mechanism of exportation. J Mol Evol 24:130–142

Hirst JD, Sternberg MJE (1992) Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. Biochemistry 31:7211–7218

Holley LH, Karplus M (1989) Protein secondary structure prediction with a neural network. Proc Natl Acad Sci 86:152–156

Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 78:3824–3828

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Networks 2:359–366

IntelliGenetics Inc, 700 East Camino Real, Mountain View, CA 94040, USA

Jähnig F (1990) Structure predictions of membrane proteins are not that bad. Trends Biochem Sci 15:93–95

Jones DD (1975) Amino acid properties and side chain orientation in proteins: A cross correlation approach. J Theor Biol 50:167–183

King RD, Sternberg MJE (1990) Protein secondary structure prediction: a machine learning approach. J Mol Biol 216:441–457

Kosko B (1992) Neural networks and fuzzy systems. Prentice-Hall International, London

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Laforet GA, Kendall DA (1991) Functional limits of conformation, hydrophobicity, and steric constraints in prokaryotic signal peptide cleavage regions. J Biol Chem 266:1326–1334

Levin JM, Garnier J, Biou V, Gibrat JF, Robson B (1988) Secondary structure prediction: combination of three different methods. Protein Engineering 2:185–191

Lohmann R (1992) Structure evolution in neural systems. In: (Soucek B, the Iris Group (eds) Dynamic, genetic and chaotic programming. Wiley & Sons, New York

McInerny JM, Haines KG, Biafore S, Hecht-Nielsen R (1989) Back propagation error surfaces can have local minima. In: International Joint Conference on Neural Networks (Washington 1989), Vol II. IEEE, New York, p 627

Minsky M, Papert S (1988) Perceptrons. MIT Press, Cambridge, MA

Perlman D, Halvorson HA (1983) A putative signal peptidase recognition site and sequence in eukaryotic and prokaryotic signal peptides. J Mol Biol 167:391–409

Qian N, Sejnowski TJ (1988) Predicting the secondary structure of globlar proteins using neural network models. J Mol Biol 202:865–884

Rechenberg I (1973) Evolutionsstrategie—Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. Frommann-Holzboog, Stuttgart

Rosenblatt F (1962) Principles of neurodynamics. Spartan, New York

Schneider G, Wrede P (1993) Analysis of protein targeting sequence features. Protein Seq Data Anal (in press)

Schulz GE, Schirmer RH (1979) Principles of protein structure. Springer-Verlag, Heidelberg

Stolorz P, Lapedes A, Xia Y (1992) Predicting protein secondary structure using neural net and statistical methods. J Mol Biol 225:363–377

von Heijne G (1983) Patterns of amino acids near signal-sequence cleavage sites. Eur J Biochem 133:17–21

von Heijne G (1986) A new method for predicting signal sequence cleavage sites. Nucleic Acids Res 14:4683–4690

Zamyatnin AA (1972) Protein volume in solution. Prog Biophys Mol Biol 24:107–123

Zhang X, Mesirov JP, Waltz DL (1992) Hybrid system for protein secondary structure prediction. J Mol Biol 225:1049–1063

Received July 28, 1992/Revised November 2, 1992