Peptide Design *in Machina*: Development of Artificial Mitochondrial Protein Precursor Cleavage Sites by Simulated Molecular Evolution

Gisbert Schneider, Johannes Schuchhardt, and Paul Wrede
Freie Universität Berlin, Institut für Medizinische/Technische Physik und Lasermedizin, AG Molekulare Bioinformatik, D-12207 Berlin, Germany

ABSTRACT Artificial neural networks were used for extraction of characteristic physicochemical features from mitochondrial matrix metalloprotease target sequences. The amino acid properties hydrophobicity and volume were used for sequence encoding. A window of 12 residues was employed, encompassing positions −7 to +5 of precursors with cleavage sites. Two sets of noncleavage site examples were selected for network training which was performed by an evolution strategy. The weight vectors of the optimized networks were visualized and interpreted by Hinton diagrams. A neural filter system consisting of 13 perceptron-type networks accurately classified the data. It served as the fitness function in a simulated molecular evolution procedure for sequence-oriented de novo design of idealized cleavage sites. A detailed description of the strategy is given. Several putative high-quality cleavage sites were obtained revealing the critical nature of the residues in the positions −2 and −5. Charged residues seem to have a major influence on cleavage site function.

INTRODUCTION

Sequence-oriented design of functional peptides and proteins with unique native-like structures is a challenging goal in protein science (Sander, 1991; Richardson et al., 1992). Advances in recombinant DNA technology and heterologous expression systems have made it possible to begin considering optimization of protein function by engineering new amino acid sequences (Wrede and Schneider, 1994). In a recent publication we have presented a method for generating rational choices of amino acid sequences that can be used for a given peptide or protein function (Schneider and Wrede, 1994). The method is based on the representation of the fitness function by artificial neural networks and an evolution strategy for sequence optimization. We have called it simulated molecular evolution (SME) to stress that its sequence development procedure takes into account simple models and phenomena of the natural evolution of proteins. In the current work, a more detailed mathematical description is given, together with another application. We have employed simple neural networks, perceptrons (Rosenblatt, 1962; Minsky and Papert, 1988), for the extraction of relevant sequence features ("design rules") from matrix targeting signals of nuclear-encoded mitochondrial precursor sequences from *Neurospora crassa* (Hartl and Neupert, 1990; Pfanner and Neupert, 1990). Cleavage sites of mitochondrial matrix metallopeptidase target sequences were analyzed (Miura et al., 1986; Arretz et al., 1991). This enzyme is analogous to MPP+PEP in yeast (Hawlitschek et al., 1988; Schatz, 1993), and has functions comparable to signal peptidase. Its natural target sites are at the junction between a matrix targeting signal and the mature region of a nuclear-encoded mitochondrial protein. Sets of noncleavage site sequences were also selected as negative examples in the training process. The optimized network systems were used as the fitness function in the SME design procedure.

Several successful experiments predominantly based on the analysis of the distribution of amino acid residues and their physicochemical properties had already been performed to deduce rules for matrix metallopeptidase target sequences. Mainly elegant statistical analyses (von Heijne, 1986; Hendrick et al., 1989; Gavel and von Heijne, 1990; Arretz et al., 1991) and an approach involving logic-based machine learning (King and Sternberg, 1990; Schneider and Wrede, 1993a) have revealed several striking features. Since matrix metallopeptidase target sequences lack a common sequence motif (von Heijne, 1986; von Heijne et al., 1989; Hendrick et al., 1989) the description of amino acid sequences in terms of physicochemical residue properties is promising for feature extraction by artificial neural networks (Schneider and Wrede, 1993b, c). The distribution of positively charged residues has been identified as being important for precursor targeting and processing (Horwich et al., 1985; 1986; von Heijne, 1986; Gavel and von Heijne, 1990). Our major aims were to test whether simple neural networks can extract the already known cleavage site features from a small set of data and to analyze them for hitherto unknown additional features.

Received for publication 2 June 1994 and in final form 2 November 1994.

Address reprint requests to Paul Wrede, Freie Universität Berlin, Institut für Medizinische/Technische Physik und Lasermedizin, AG Molekulare Bioinformatik, Krahmerstrasse 6–10, D-12207 Berlin, Germany. Tel.: 030-798-4158; Fax: 030-834-4004.

This paper is dedicated to Alexander Rich ("Alex in Wonderland").

0006-3495/95/02/434/14 $2.00

MATERIALS AND METHODS

Sequence data

Eleven precursor sequences of nuclear-encoded mitochondrial proteins from *N. crassa* were used in the analysis (Table 1). In these sequences a dozen matrix metallopeptidase cleavage sites were identified by in vitro experiments with purified enzyme and verified by subsequent N-terminal sequencing (Arretz et al., 1991; W. Neupert, personal communication). For testing the neural networks several sequences were selected from the

TABLE 1 Cleavage site sequences used for filter development and feature extraction

No.	Name	Sequence
1	m107.Mx, IM PEP, 52 kDa PEP	...INPFRRG ↓ LATPH...
2	m46.IM, Mx side COX IV	...ATTVVRC ↓ NAETK...
3	m134.IM Cox V	...PTMAVRA ↓ ASTMP...
4	m168.Mx Leu-5, leucyl-tRNA synthetase	...ESWKRFY ↓ ADHKL...
5	m298.IM ATPase F1 β subunit	...APALSRF ↓ ASSAG...
6	m204.IM periph. Mx-side complex I, 49 kDa	...ASALRRY ↓ AEPSY...
7	m211.Mx NADH-DH 40 kDa subunit	...FGFQRRA ↓ ISDVT...
8	m297.IM cytochrome *c*1	...FKFAKRS ↓ ASTNS...
9	m12.IM proteolipid (1st site)	...AQVSKRT ↓ IQTGS...
10	m12.IM proteolipid (2nd site)	...QAFQKRA ↓ YSSEI...
11	m26.IM, Rieske 2Fe-2S protein	...PARAVRA ↓ LTTST...
12	m111.Mx cyclophilin	...TFSCARA ↓ FSQTS...

Courtesy of M. Arretz, München. The arrows indicate experimentally determined matrix metallopeptidase target sites. Mx, matrix, IM_1 intermembrane.

PIR-International database, Rel. 35 (Barker et al., 1992): two cytosolic proteins from *N. crassa* (PIR1 SYNCLC, PIR1 CSNCC), two additional nuclear-encoded mitochondrial precursors from *N. crassa* (PIR1 LWNCA, PIR2 S17192), and an *Escherichia coli* periplasmic protein precursor sequence (PIR1 HQECSN).

For network training, the cleavage site regions (positive examples) were restricted to sequence windows covering the positions −7 to +5 relative to the processing site, which is between −1 and +1. Position +1 indicates the N-terminal amino acid of the mature protein. This window size and orientation has been shown to provide characteristic cleavage site features (Schneider and Wrede, 1993b). Two different sets of negative examples were compiled. Set A contained 48 examples randomly selected from the whole precursor sequences (4 from each sequence). Set B consisted of 48 negative examples from the regions adjacent to the cleavage sites (4 from each sequence). As a result, every training set consisted of 11 positive examples and 44 negative examples from either set A or set B. Every test set contained 1 positive example and 4 negative examples from the same precursor sequence. Since complete cross-validation was performed for network training, 12 training and 12 test sets were compiled using the negative examples of set A and set B. This resulted in a total of 24 training and 24 test sets. The complete sequence data are available from the authors on request.

Neural network architecture and training

Due to the very limited number of precursor sequences with confirmed matrix metallopeptidase target sequences available for a single organism, complete cross-validation was performed for network training and testing. Twelve neural networks containing two layers of units (Fig. 1 *A*) were optimized per training set. For sequence encoding every sequence window ($\vec{r}$) of 12 residues was translated into a 24-dimensional input vector ($\vec{x}$) in the input layer of the network, which was built up in turn from 24 linear fan-out units. Every input unit codes for a physicochemical property value $x_{i,\alpha}$. The property scales for hydrophobicity, $\alpha = 1$, (Engelman et al., 1986) and side-chain volume, $\alpha = 2$, (Zamyatnin, 1972) were used. These scales are not substantially correlated, as is indicated by the correlation index of

$$r = \frac{\sum_{i=1}^{20}(x_{i,1} - \bar{x}_1)(x_{i,2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{20}(x_{i,1} - \bar{x}_1)^2 \sum_{i=1}^{20}(x_{i,2} - \bar{x}_2)^2}} = -0.003.$$

Every scale was normalized to a (−1, 1) interval to facilitate weight optimization. The single unbiased unit of the second network layer (output unit) limits the network response to real values between 0 and 1. The common sigmoidal Fermi function $F(\text{unit}_{in})$ was employed as the transfer function (squashing function), and the overall sequence transformation of a network is given by

$$\text{output} = F\left(\sum_{\alpha=1}^{2}\sum_{i=1}^{12} x_{i,\alpha}\, w_{i,\alpha}\right); \qquad F(\text{unit}_{in}) = \frac{1}{1 + e^{-\text{unit}_{in}}}.$$

Due to its simple architecture, the performance of every single network is limited to linear separation of the input space (Minsky and Papert, 1988). The training task was to separate positive from negative sequence examples, i.e., to find features based on physicochemical properties that facilitate correct classification of the training data. Ideally, every positive example (cleavage site) should lead to a network response of 1, every negative example (noncleavage site) to a response of 0. The mean-square error (MSE),

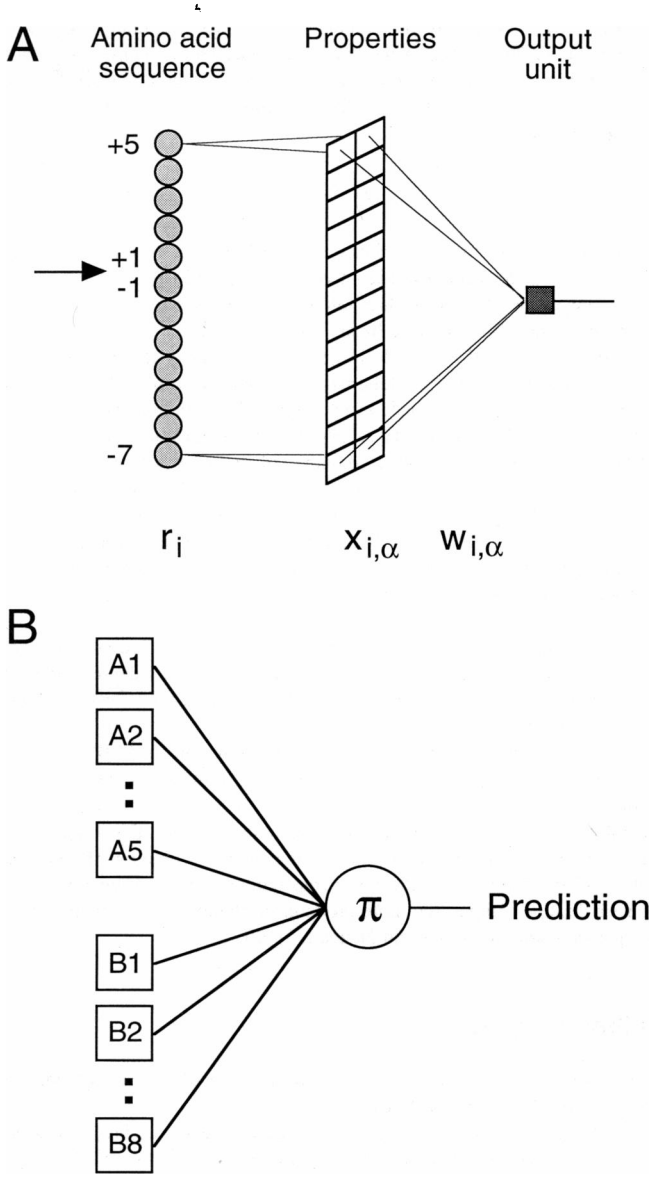

FIGURE 1 (*A*) Scheme of the network architecture used for feature extraction from matrix metallopeptidase cleavage sites. Amino acid sequences of length 12 were investigated covering the relative precursor positions −7 to +5. The arrow indicates the processing site, and the mature protein starts at position +1. The flow of information is from left to right. For clarity, only a few network connections are drawn. Each of the columns of the layer "properties" stands for a single amino acid property (hydrophobicity and volume). (*B*) Combination of networks by a linear unbiased π-unit. Architectures of A1–A5 and B1–B8 as given in *A*. The prediction system used was built up from 13 network modules.

served as an estimate of the learning success during training which was stopped after 20 learning cycles. This procedure had turned out to be advantageous in preliminary experiments. A simple measure of classification quality, Q, was employed to calculate the proportion of correct predictions during the training phase:

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

where P is the number of correctly predicted positive examples, N is the number of correctly predicted negative examples, and T is the total number of examples investigated. A threshold value of 0.5 was used to convert the continuous network output to binary values.

For weight optimization a (1,100)-evolution strategy (Rechenberg, 1973) was used. This is a simple evolutionary algorithm performing a local random search which has previously been applied to neural network training (Lohmann, 1992; Lohmann et al., 1994; Schneider and Wrede, 1992, 1993b, 1994). All weights were initialized with random values between -1 and 1, and the resulting weight vector was declared as "parent" of the first optimization cycle ("generation"). One generation consisted of three successive steps: 1) generate 100 new Gaussian-distributed weight vectors ("offspring") around the parent; 2) calculate MSE for every new weight vector; 3) declare the weight vector leading to the lowest MSE as parent for the next generation.

In the first generation the standard deviation of the Gaussian was set to 1. The standard deviation itself was subjected to an adaptive process (Rechenberg, 1973; Davidor and Schwefel, 1992; Bäck and Schwefel, 1993) as described below.

The final filter system used as the fitness function for SME was constructed by combining those of the 24 networks trained which were able to correctly classify both training and test data. The network modules were combined by feeding their output to a single linear π-unit (Fig. 1 *B*), which is a continuous implementation of the logical AND function (Rumelhart et al., 1986). We interpret the continuous output in the following way: the larger the filter output, the more pronounced the matrix metallopeptidase sequence feature (Pfanner and Neupert, 1990).

Filter analysis

Three tests were performed to evaluate the accuracy of the filter system and to interpret the features extracted.

1. Graphical representations of the weight vectors were investigated to analyze the sequence features found by the individual networks. This procedure is similar to the use of Hinton diagrams (Qian and Sejnowsky, 1988; Rumelhart et al., 1986; Holbrook et al., 1993; Schneider et al., 1993). The weights were normalized to a $(-1, 1)$ interval and their values were interpreted as the relative importance of sequence positions and the corresponding physicochemical properties. Independent analysis of the results for side-chain volume (Zamyatnin, 1972) and for hydrophobicity (Engelman et al., 1986) is possible because these scales are orthogonal. All the network modules that were combined to form the final matrix metallopeptidase filter system for SME were treated separately. In addition, an averaged weight vector was analyzed.

2. The predictive quality of the final filter was determined by scanning the N-terminal parts of the precursor sequences up to position 75 using the sliding window technique. Further, a comparison between our filter system for the *E. coli* signal peptidase I cleavage site (Schneider and Wrede, 1994) and the filter for matrix metallopeptidase target sequences was performed by applying the networks to a periplasmic protein precursor of *E. coli* and a precursor sequence of an *N. crassa* mitochondrial protein. None of these sequences was part of the data used for network training. Four prediction qualities were calculated: Q (see above), Q_{over}, Q_{under}, and the correlation coefficient Q_{corr} (Mathews, 1975). P is the number of positive correct predictions, N is the number of negative correct predictions, O is the number of false-positives (overpredicted), and U the number of false-negatives (underpredicted). To focus on the smallest space containing all positive examples we have chosen a threshold value of 0.2 for classification of the filter output. This value defines the maximum threshold leading to 100% positive-correct predictions with $U = 0$ and $Q_{under} = 1$ (see Results). Of course, P, N, O, and U and, therefore, the prediction accuracy defined by Q, Q_{over}, Q_{under}, and Q_{corr}, depend on the threshold value used.

$$Q_{over} = \frac{P}{P + O}; \qquad Q_{under} = \frac{P}{P + U};$$

$$Q_{corr} = \frac{(P \times N) - (U \times O)}{\sqrt{(N + U)(N + O)(P + U)(P + O)}}.$$

To visualize the filter outputs for artificial random sequences a histogram was calculated. For this, 10^6 random sequences were generated and evaluated by the filter system. We estimated that the error of a histogram entry containing N filter responses was $\sqrt{N}$.

3. Surface plots of the fitness landscape of the SME filter were obtained by calculating the filter response to continuous input values. The starting point was the "optimal" cleavage site peptide, i.e., the sequence of the peptide leading to the maximum filter output of all the sequences investigated. The volume and hydrophobicity values of selected sequence positions were systematically varied between -1 and 1. All the other amino acids of the peptide were kept fixed. Geometrically this results in a two-dimensional intersection of the 24-dimensional fitness landscape.

Simulated molecular evolution

In general, we applied the SME technique as described by Schneider and Wrede (1994). In the current work, a more detailed mathematical description is given, where SME is regarded as a biologically motivated algorithm for stochastic search in sequence space.

Metric in sequence space

A sequence of amino acids $r_1 \cdots r_{12}$ (length of the sequence windows investigated) is mapped into an intermediate space on which a metric has been defined based on meaningful amino acid distance matrices. We assume that the intermediate space reflects at least partially properties of the phenotype space, e.g., the protein structure and function (Ebeling et al., 1990; Fontana et al., 1993):

amino acid sequence → intermediate space → protein function.

Several distance matrices were selected from the literature: the Feng matrix (Feng et al., 1985), the Risler matrix (Risler et al., 1988), the Grantham matrix (Grantham, 1974), and the Myata matrix (Myata et al., 1979). In addition, a "context matrix" based on the similarity of amino acids according to their physicochemical properties was defined. To obtain this matrix, each amino acid was assigned a pair of physicochemical properties. We used the properties volume (V) (Zamyatnin, 1972) and hydrophobicity (H) (Engelman et al., 1986).

$$r_1 \rightarrow (V(r_1), H(r_1))$$
$$\vdots$$
$$r_{12} \rightarrow (V(r_{12}), H(r_{12}))$$

This is motivated by experimental findings which strongly suggest that functional properties of signal sequence cleavage sites depend almost continuously on their physicochemical properties (Kaiser et al., 1987; Bird et al., 1990; Hendrick et al., 1989; Gavel and von Heijne, 1990; Pfanner and Neupert, 1990). The continuous (ordered) character of the physicochemical properties allows one to define distances between pairs of amino acids, r and r', in various ways. For the construction of our context matrix we have used the Euclidian distance

$$d(r,r') = \sqrt{(V(r) - V(r'))^2 + (H(r) - H(r'))^2}.$$

Stochastic search algorithm

The search performed by SME takes into account several biological observations concerning the evolution of proteins. 1) Usually, evolution tends

to proceed in small steps on the molecular level; while large steps are possible within a generation, the resulting variants rarely survive (Myata et al., 1979). 2) The probability of mutation is not identical at each site of the amino acid sequence; there are more variable and less variable regions. 3) Which region should be kept constant or variable has to be learned during the process of evolution (meta-evolution).

What is meant by the terms "small step" and "large step" needs to be clarified. Looking at an amino acid sequence without any further information one might imagine that a small step might be a single amino acid substitution. Depending on its context, however, this can result in a drastic change, a moderate change, or even no significant change at all in the protein's function (Dayhoff and Eck, 1968; Kimura, 1983). Since to date there is no general method for predicting the structural alterations induced by a single substitution, we hope that an adequate description of the context in terms of amino acid properties (in the current case the physicochemical properties hydrophobicity and volume) may at least give a hint as to how significant the alteration will be.

The suitability of such a description and of the assumption that small steps lead only to small structural changes depends as well on the context in which the substitution occurs. A very simple assumption is that, e.g., side chain polarity might be an important criterion at one protein site, whereas side chain volume is important at another. Further, an amino acid might well play different roles in different environments. The choice of different distance matrices is intended to take this context-boundedness into account and to reflect a correlation between the intermediate (artificial) world and the functional (real) world.

SME performs a kind of stochastic search which incorporates the preference for small steps (George et al., 1990). Evolution is regarded as an optimization process (Parker and Maynard Smith, 1990; Fontana et al., 1993). Following the theory of evolution strategies (Rechenberg, 1973) we assume a Gaussian distribution for the transition probability $r \rightarrow r'$. Altering the common evolution strategy as proposed by Rechenberg (1973) we reduce the search space from the continuous multidimensional hypercube to a discrete lattice of the same dimension. Its nodes are given by the possible amino acid sequences (20^{12} for matrix metallopeptidase cleavage sites). The nodes of the lattice are not evenly spaced; the distances differ according to the distance matrix chosen, e.g., a physicochemical distance.

Beginning with a random initial sequence, λ successors (offspring) are generated as follows. For each of the 20 possible amino acids Ala, $\cdots$, Trp a ranking procedure is performed according to the distance matrix selected. Starting with alanine, e.g., we find serine next to it followed by cysteine, and so on. According to the context matrix the amino acid which is farthest from alanine is arginine. The distance values are normalized to a (0, 1) interval. To develop the offspring a Gaussian-distributed random vector, $\zeta_1 \cdots \zeta_{12}$, is generated using the Box-Muller formula

$$\xi_k = \sigma \sqrt{-2 \ln(i)} \sin(2\pi j),$$

where i and j are random numbers equally distributed between 0 and 1. ζ_k determines the new residue r_{new} at position k. How this is done is shown schematically in Fig. 2. Here, alanine is the old r_{old} residue at position k and the other 19 amino acids are positioned along the distance axis according to the ranking procedure described above. The amino acid given by ζ_k is selected as r_{new}. In Fig. 2 the probability of finding cysteine as r_{new} is given by the area shaded in gray. As illustrated, the selection probability for a given substitution depends on the variance σ_k^2 of the Gaussian distribution generated. With the small variance shown a switch from alanine to cysteine is extremely improbable, whereas with a sufficiently large variance there is a rather high probability, $p_{A \rightarrow C}$, for this substitution

$$P_{A \rightarrow C} = \int_{x_{left}}^{x_{right}} \frac{\exp(-z^2/2\sigma^2)}{\sqrt{2\pi\sigma^2}} dz.$$

Since the mutation probability is not the same for all positions in a sequence, an independent variance σ_k^2 has been used. In SME, all the variances are also subjected to an adaptive evolution:

$$\sigma_{new} = \sigma_{alt} \, |1 + \xi|,$$

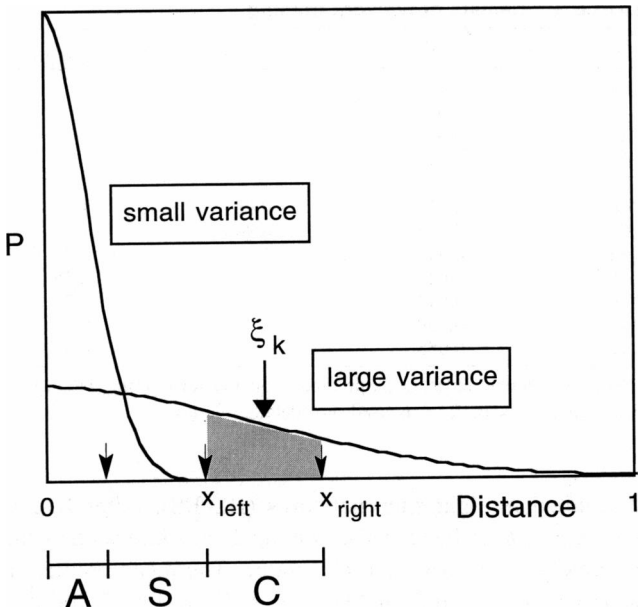

FIGURE 2 Ranking of amino acids for substitutions in simulated molecular evolution. In the example shown the residue at position k to be substituted is alanine, which is located at the origin. The next neighbor is serine followed by cysteine. They are positioned according to their amino acid distance to alanine. The substitution probability p for Ala $\rightarrow$ Cys is given by the shaded area. If a very small variance σ_k^2 is used the probability of the Ala$\rightarrow$Cys mutation is vanishingly small. ζ_k is a Gaussian random number defining the new residue at sequence position k.

where ξ is a Gaussian-distributed random number with variance 1. The initial value of the variance was deliberately chosen to be $\sigma_0 = 1$ in all experiments.

The design experiments with SME were terminated after 200 generations. The number of offspring was $\lambda = 500$ per generation. For a comparison of distance matrices 100 generations and different experiments with $\lambda = 50$, $\lambda = 100$ and $\lambda = 200$ were performed.

RESULTS AND DISCUSSION

Development of a neural filter system for mitochondrial matrix metallopeptidase target sequences

In total, 24 perceptrons were trained for the recognition of characteristic cleavage site features of *N. crassa* matrix metallopeptidase target sequences. Complete cross-validation was applied to either of two training sets, A and B, resulting in 12 networks each. Set A contained randomly selected negative examples, set B contained negative examples taken from the regions adjacent to the cleavage sites. The sets had the positive cleavage site examples in common. A (1, 100) evolution strategy with adaptive step size control (Rechenberg, 1973) served as a training technique, and optimization was performed for a fixed number of 20 cycles ("generations"). The training results are summarized in Table 2. Set A yielded five filters (A1–A5) correctly covering all 55 training examples (11 positive, 44 negative) and the 5 test examples (1 positive, 4 negative) where $Q_{train} = Q_{test} = 1$. Set

TABLE 2 Results of network training

A	MSE	Q_{train}	Q_{test}	B	MSE	Q_{train}	Q_{test}
A1	0.065	1	1	B1	0.028	1	1
A2	0.092	1	1	B2	0.021	1	1
A3	0.074	1	1	B3	0.015	1	1
A4	0.093	1	1	B4	0.013	1	1
A5	0.064	1	1	B5	0.026	1	1
A6	0.153	0.98	1	B6	0.032	1	1
A7	0.155	0.98	1	B7	0.011	1	1
A8	0.015	1	0.75	B8	0.037	1	1
A9	0.042	1	0.75	B9	0.016	1	0.75
A10	0.040	1	0.75	B10	0.003	1	0.75
A11	0.039	1	0.75	B11	0.012	1	0.75
A12	0.058	1	0.75	B12	10^{-4}	1	0.50

MSE, mean square error; Q_{train}, prediction quality for training data; Q_{test}: prediction quality for test data. A: set A of negative examples was used; B: set B of negative examples was used. For details, see text.

B resulted in eight such networks (B1–B8). Filter B12 is apparently specialized on the training data as indicated by the extremely small MSE and Q_{test} value. The mean prediction accuracy Q_{test} of all 24 networks is 81%. In Fig. 3, training protocols of the 13 perceptrons combined to build up the final matrix metallopeptidase filter are shown. From training set B cleavage site features were extracted more quickly than from set A, as indicated by the course of the Q values during network optimization. Set B also led to lower MSEs. It may be that characteristic cleavage site features are more pronounced when cleavage sites are compared to sequence examples stemming from the region around the actual processing site (set B).

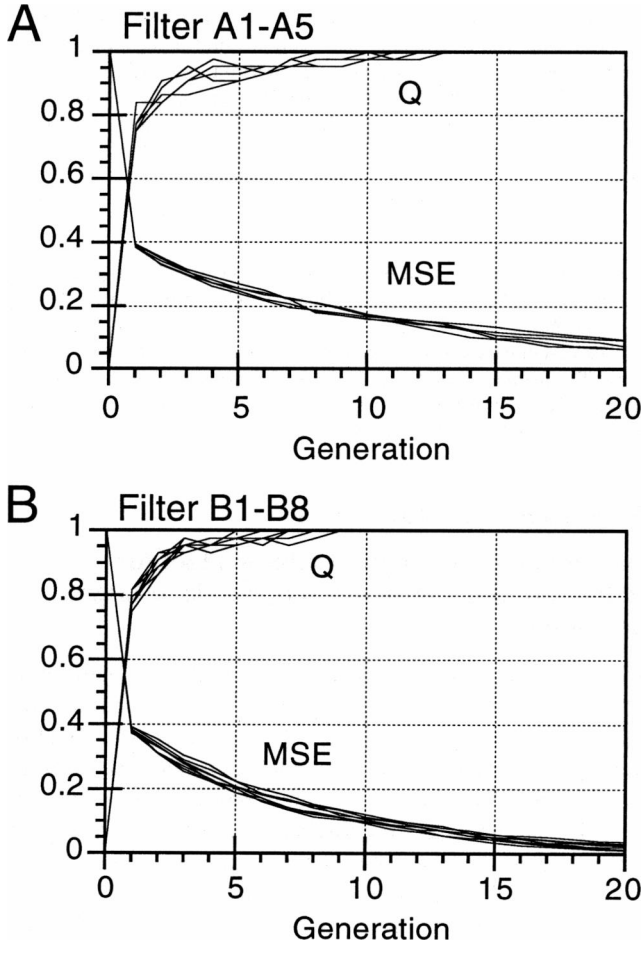

FIGURE 3 Course of MSE and Q of the neural networks during network training. A training phase of 20 generations and $\lambda = 100$ is presented. Two different training sets were applied: (*A*) set A with negative training examples taken from the whole precursor sequences; (*B*) set B with negative examples taken from the vicinity of the matrix metallopeptidase cleavage site.

To get an idea of the features extracted, graphical representations of the networks' weights have been interpreted in such a way that very large or very small values indicate important sequence positions and amino acid properties (Fig. 4). Position −2 seems to be of major importance for the separation of cleavage sites from noncleavage sites. This observation is independent of the training data used (either set A or set B, see Materials and Methods) as indicated by extreme weight values for both properties hydrophobicity and volume (Fig. 4, *A* and *B*): large non-hydrophobic residues are predominant in −2. This feature appears to be equally well pronounced in Fig. 4 *B* and in Fig. 4 *A*. Position −2 has been identified as very important for matrix metallopeptidase cleavage by several statistical analyses (Hendrick et al., 1989; Gavel and von Heijne, 1990; Arretz et al., 1991). We were able to substantiate this finding.

The positions +1, −1 and −5 seem to be slightly preferred by hydrophobic amino acids (Fig. 4, *A* and *B*). An additional slight preference for large residues is also found in position −5 (Fig. 4 *B*). A surprising finding is the sequence [small–large–small] and [hydrohobic–non-hydrophobic–hydrophobic] for the positions −1, −2, −3, which is characteristic for eubacterial and eukaryotic signal peptidase cleavage sites (Perlman and Halvorson, 1983; von Heijne, 1983; Schneider and Wrede, 1993a). Whether it is also important for the catalytic activity of mitochondrial matrix metallopeptidase cannot be decided on the basis of the current analysis. Only site-directed mutagenesis studies can determine which residues are important for actual processing. Since only mitochondrial inner membrane protease 1 and 2 are homologous to signal peptidase (Nunnari et al., 1993; Schneider et al.,

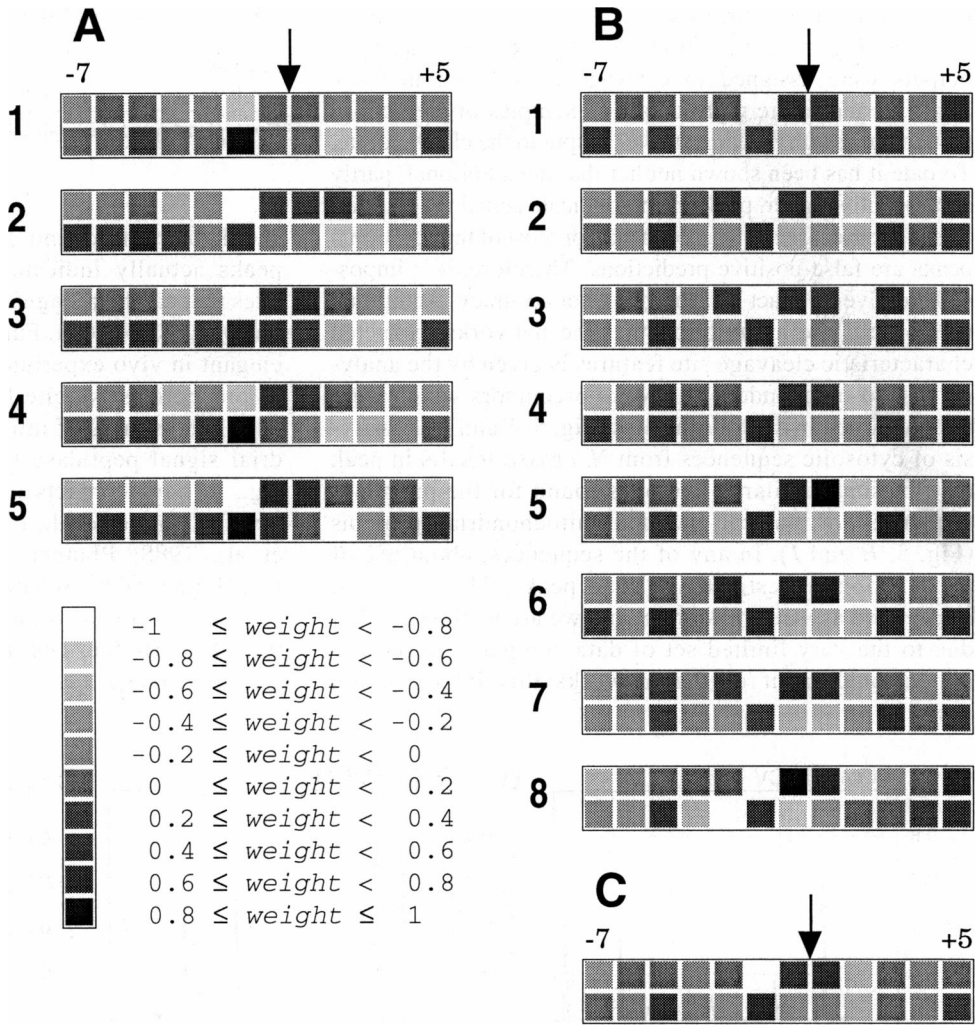

FIGURE 4 Graphic representation of the weight vectors of the optimized networks A1–A5 and B1–B8 leading to the highest prediction accuracy. Weight values are shown in a gray scale code. The top lines correspond to the amino acid property "hydrophobicity," the bottom lines to the property "volume." Numbers indicate the sequence positions and the arrow the matrix metallopeptidase cleavage site. (*A*) Negative training examples were taken from the whole precursor sequences (set A); (*B*) Negative examples were taken from the vicinity of the matrix metallopeptidase cleavage site (set B). (*C*) Diagram of the average weights of A1–A5 and B1–B8.

1991), there is no reason to expect that the cleavage site specificity is similar for the non-homologous matrix metallopeptidase. It has been hypothesized that signal peptidases belong to a new class of proteases (Arretz et al., 1991; Dalbey and von Heijne, 1992). Certainly, a striking difference between homologous signal peptidase cleavage sites and matrix metallopeptidase cleavage sites is the existence of a charged amino acid at position −2 in the latter. This residue usually is an arginine (von Heijne, 1986; Hendrick et al., 1989; von Heijne et al., 1989). Further, the existence of a positive charge in the C-terminal region of secretory signals results in a loss of activity and the cleavage site is not recognized and hydrolyzed by the eubacterial or eukaryotic enzyme (Kaiser et al., 1987; Bird et al., 1990; Laforet and Kendall, 1991). We are presently performing site-directed mutagenesis studies to evaluate our findings.

In Fig. 4 *C*, the weight diagram of the average weight values calculated from Fig. 4, *A* and *B*, is shown, substantiating a possible predominance of the positions −5, −2, −1, and +1. However, the prediction accuracy of the single "average" perceptron (Fig. 4 *C*) is smaller (Q = 73%) than the accuracy of the multi-modular network (Q = 99%, see below).The networks A1–A5 and B1–B8 were combined by a linear π-unit, i.e., by unbiased multiplication of their individual output values (Rumelhart et al., 1986). This multi-modular filter consisting of 13 perceptrons was used as the fitness function for SME. To test its prediction accuracy, the N-terminal parts of the 11 sequences used for feature extraction and several additional sequences were subjected to cleavage site analysis. In Table 3 the quality indices calculated from the prediction results for the 11 *N. crassa* sequences (Table 1) are listed. The filter is able to separate positive from negative examples with high accuracy (Q = 99%, Q_{corr} = 0.63) and no cleavage site found by the in vitro experiments is missed (Q_{under} = 1). However, a remarkable number of additional positive predictions is made (Q_{over} = 0.41). Whether these filter outputs indicate putative cleavage sites that can be recognized and cleaved by matrix metallopeptidase is unclear and cannot be decided on the basis of our results alone. It has been reported that about 20% of naturally occurring random sequences are able to function as N-terminal mitochondrial transit peptides (Schatz, 1993). Further, some precursor sequences are known to contain more than just one matrix metallopeptidase processing site

(Sztul et al., 1987; Arretz et al., 1991). Fig. 5 *D* gives an example. Fig. 5, *A, C,* and *D* show that the maximal filter outputs were assigned to cleavage sites found in vitro. Fig. 5, *B* and *E,* are representative examples of predictions that do not assign the largest filter output to the cleavage site. To date it has been shown neither that the additional, partly significantly higher, peaks do not indicate actual matrix metallopeptidase target sites, nor that all or most of the additional peaks are false-positive predictions. Therefore, it is impossible to give an exact measure for filter accuracy. Additional support for the assumption that the networks extracted characteristic cleavage site features is given by the analysis of two independent *N. crassa* precursors which were not contained in the training sets (Fig. 5, *F* and *G*). Analysis of cytosolic sequences from *N. crassa* results in peak distributions similar to the ones found for the precursor sequences of nuclear-encoded mitochondrial proteins (Fig. 5, *H* and *I*). In any of the sequences, about 6% of the positions investigated lead to peaks. This might be interpreted as a network error, and we are well aware that due to the very limited set of data compared to the 24-dimensional input of the networks this interpretation should be kept in mind. It is, however, very likely that the peaks actually indicate matrix metallopeptidase target sites that can be recognized by the enzyme because of the reasons given above. Further, it has been concluded from elegant in vivo experiments that the information content of prepeptides is quite low and that there must exist an additional source of information directing the mitochondrial signal peptidase to the appropriate cleavage site, e.g., tertiary contacts or local secondary structure formation (Nguyen et al., 1987; Vassarotti et al., 1987; Gavel et al., 1988; Pfanner and Neupert, 1990; Gavel and von Heijne, 1990; Schatz, 1993). Mitochondrial targeting sequences show a strong tendency to adopt helical structures (von Heijne, 1986), which might also play a part in signal cleavage.

TABLE 3 Prediction qualities of the combined network

Q	Q_{over}	Q_{under}	Q_{corr}
0.99	0.41	1	0.63

For explanation of the quality indices, see text.

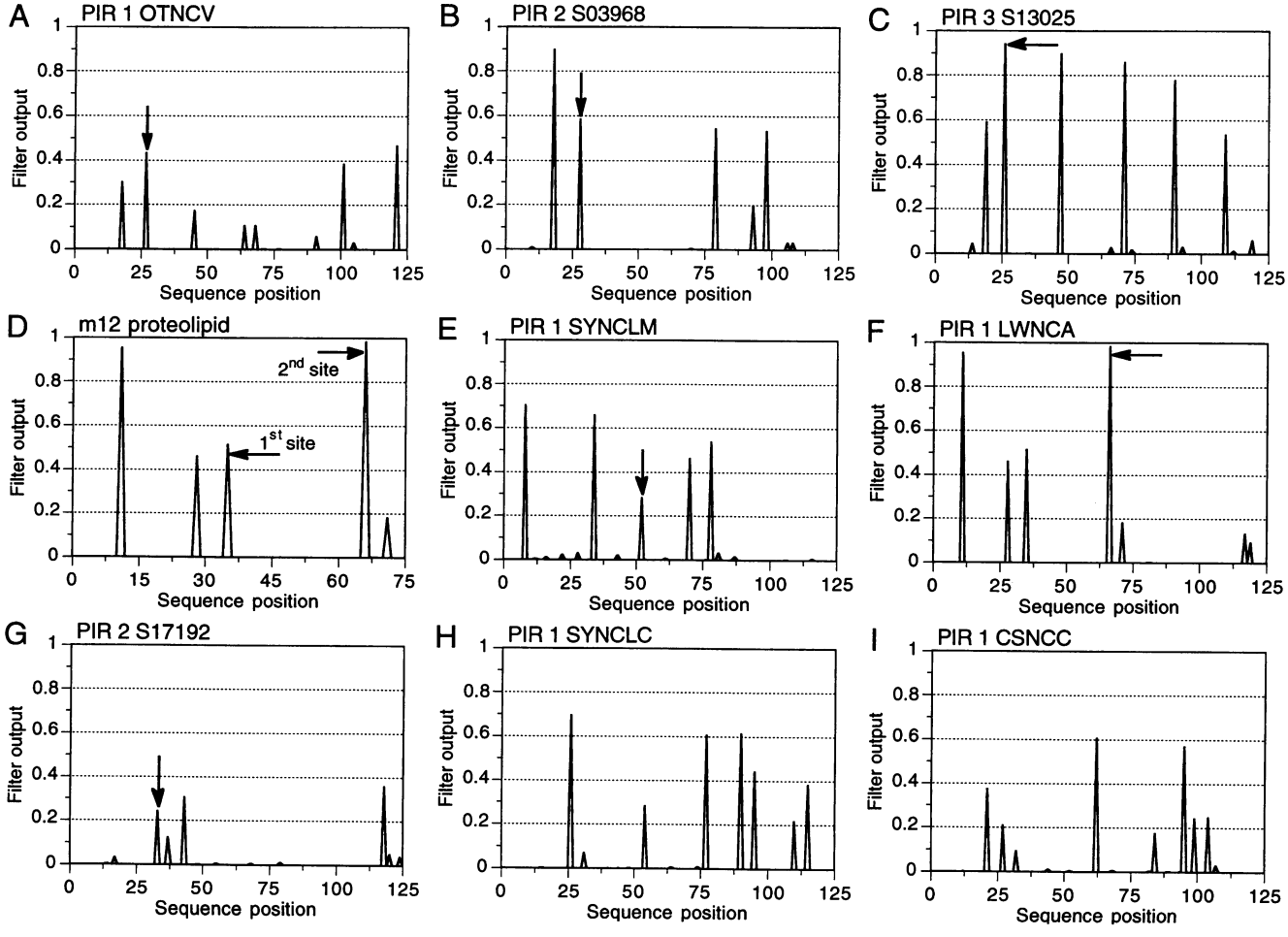

FIGURE 5 Prediction results using the combined filter system consisting of 13 network modules (cf. Fig. 1 *B*). (*A–E*) Training sequences; (*F, G*) independent test sequences; (*H, I*) cytosolic sequences of *N. crassa*. The PIR-IDs are given above the plots. Sequence positions start at the N-termini of the precursors, and the matrix metallopeptidase cleavage sites are indicated by arrows.

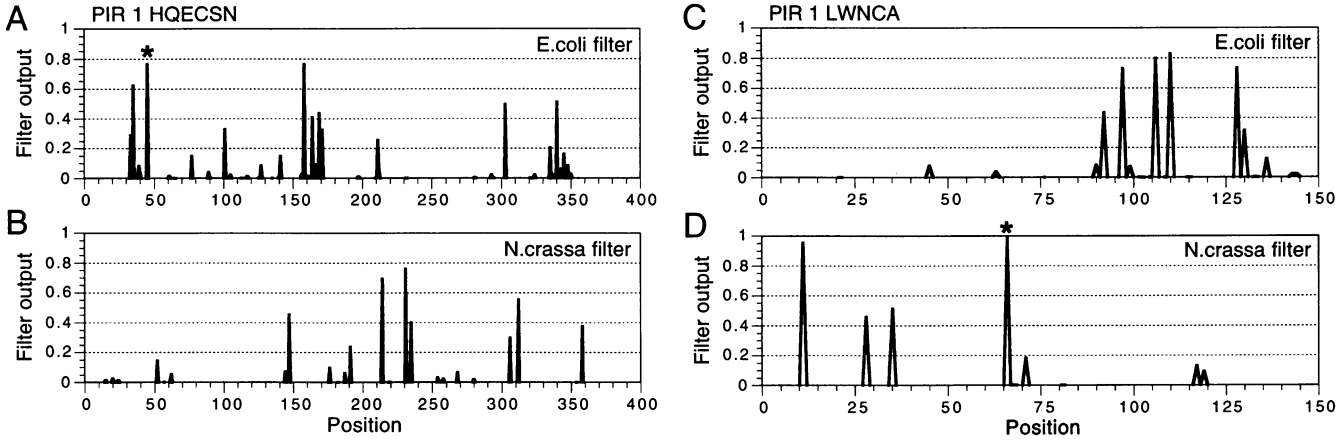

FIGURE 6 Prediction results of two neural networks trained on the recognition of *E. coli* signal peptidase I cleavage sites and on the recognition of *N. crassa* mitochondrial matrix metallopeptidase target sites. The upper plots (*A* and *C*) show the network output of the *E. coli* filter (Schneider and Wrede, 1994), the lower two plots (*B* and *D*) give the corresponding output of the *N. crassa* filter. Output values are determined as described in Fig. 5. (*A, B*) Prediction for *E. coli* hydrogenase (EC 1.18.99.1) (NiFe) small chain precursor; (*C, D*) prediction for *N. crassa* H^+-transporting ATP synthase (EC 3.6.1.34) lipid-binding precursor. The PIR-IDs are given above the plots.

Fig. 6 demonstrates that cleavage site features of matrix metallopeptidase target sequences are not present in the N-terminal parts of precursors of eubacterial periplasmic proteins (Fig. 6, *A* and *B*) and that features of eubacterial signal peptidase I are not found in precursor sequences of nuclear-encoded mitochondrial proteins (Fig. 6, *C* and *D*). For our analysis, the filter described in the present paper and a neural network for eubacterial signal peptidase I cleavage sites was used (Schneider and Wrede, 1994). The predictions of the two filters are uncorrelated with respect to the two example sequences investigated, as indicated by a correlation coefficient of $r = -0.03$ (see Materials and Methods). This means that there is no positive prediction made by the two filters at the same time and that they can be regarded as being independent. This observation is in accordance with theories on the evolutionary development of mitochondrial targeting sequences (Pfanner and Neupert, 1990; Schneider et al., 1992; Schneider and Wrede, 1993a). A eukaryotic cell must be able to differentiate with absolute accuracy between targeting sequences directing protein precursors either to mitochondria or to a secretory route. Therefore, it is reasonable to assume that the cell has evolved orthogonal signals and corresponding decoding proteins. This hypothesis is supported by our results.

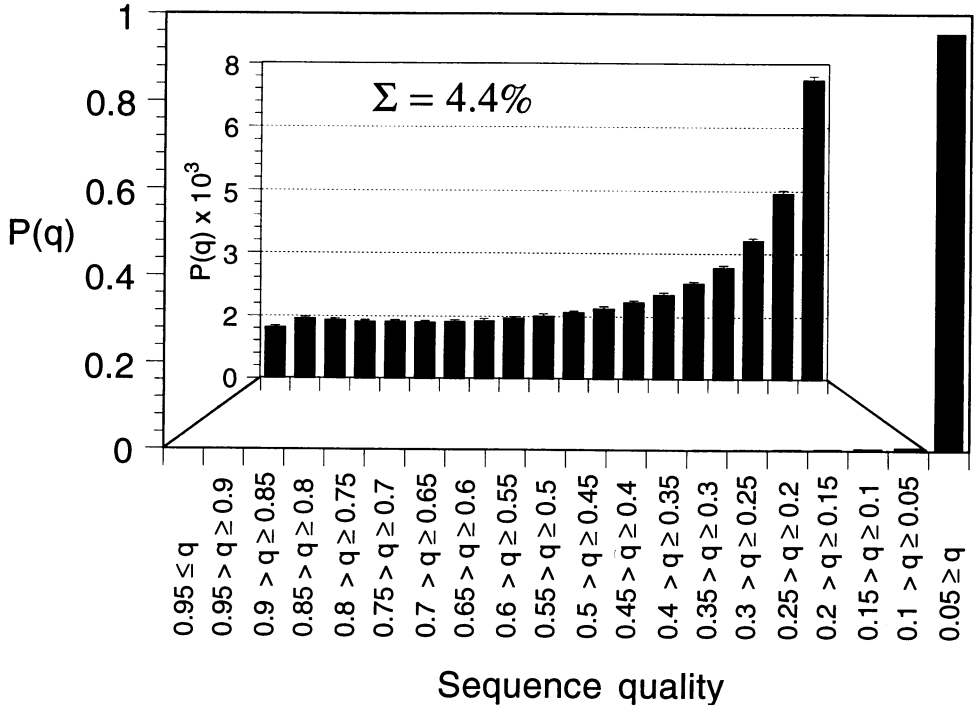

FIGURE 7 Histogram giving the probability $P(q)$ of sequence quality q calculated from the neural network output values for 10^6 random sequences. The network output values are interpreted as sequence quality. The relative frequencies for $q > 0.05$ are shown enlarged in the center of the plot. Note: most of the values of p are very small, except for $0.05 \geq q$.

The histogram in Fig. 7 shows the probability distribution of network outputs when the filter is applied to artificial random sequences. Output values have been interpreted as sequence quality, q, in SME. The vast majority of random sequences (96.6%) are predicted as being noncleavage sites. Only 4.4% lead to a network output (sequence quality) above 0.05 and less than 3% are assigned a quality above 0.2. Compared to the 20% of random sequences expected to function as transit peptides (Schatz, 1993) the filter is rather limiting. However, how "function" is defined is often arguable, and a stringent set of criteria might indicate that <20% of random sequences would provide efficient and cleavable signals. Nonetheless, it must be stressed that the 20% value was obtained from analyzing real sequences, whereas our random sequences were artificially generated strings of amino acids stemming from a pool of 20^{12} possible sequences. This set is likely to be significantly larger than the one covering all naturally existent 12-residue sequences.

Design of idealized matrix metallopeptidase cleavage sites

The most direct approach for obtaining high-quality sequences on the computer is to start from a deliberately chosen random sequence and to perform successively 19 permutations per position keeping all the other amino acids fixed. This method is a discrete analog to Seidel iteration (Mathews, 1992). The peptide RDWRGRM ↓ GGGDW (arrow indicates the cleavage site) was found to be the best sequence, with $q = 0.999962$. A discussion of its amino acid motifs, physicochemical features, and quality is given below. This idealized sequence served as the starting point for the calculation of reduced fitness landscapes of the sequence space (Fig. 8).

Fig. 8 *A* shows the fitness function (network output) at position −2, Fig. 8 *B* at position −3. At position −2, two clearly separated areas are defined by the fitness function. Large, non-hydrophobic side chains seem to be preferred here. This gives further support for the above interpretation of the weight diagrams (Fig. 4). In contrast, position −3 has a slight tendency for small, hydrophobic residues, and fitness at this position does not vary significantly (Fig. 4 *B*).

The locations of the 20 amino acids in the fitness landscape of position −2 are shown in Fig. 9. Polar and charged residues including tyrosine and tryptophan are clearly separated from hydrophobic, small side chains. Arginine, which is most frequently found in natural cleavage sites at −2 (Hendrick et al., 1989; Arretz et al., 1991), is located in the fitness optimum. Lysine, which is also positively charged, is located next to arginine. It is clear from Fig. 9 that extrapolation of the fitness function leads to a reasonable result, which corresponds to the statistically found optimum. Uncharged, small amino acids like glycine or cysteine are not found at −2 in natural matrix

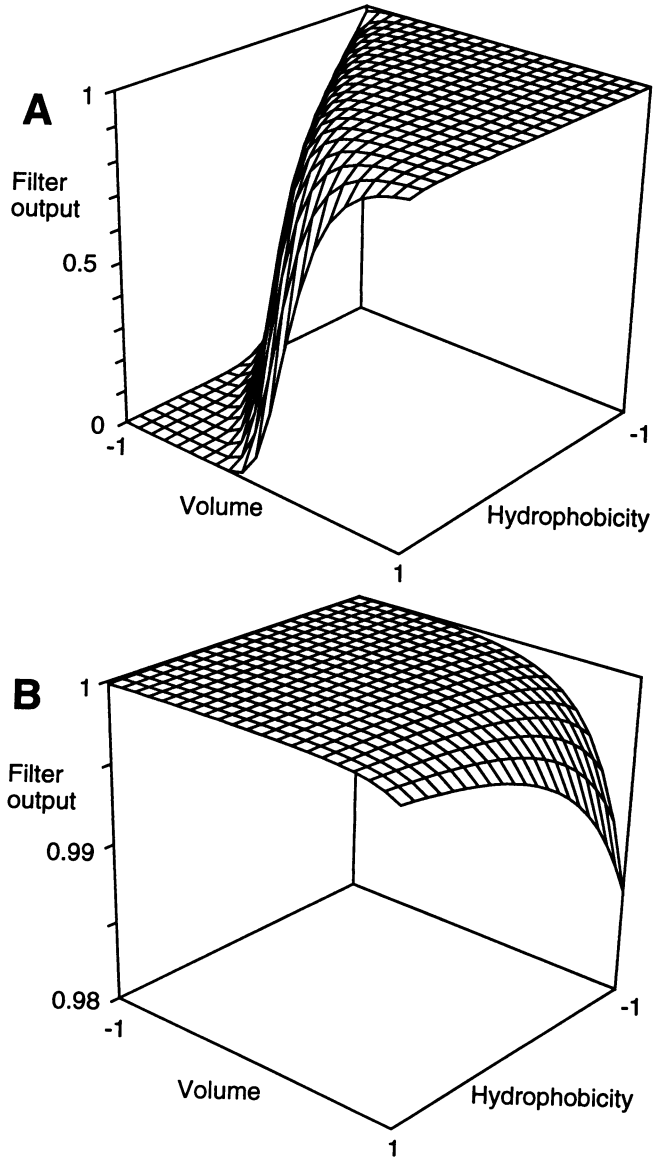

FIGURE 8 Two-dimensional intersections of the fitness landscapes for the selected positions −2 and −3 of matrix metallopeptidase target sites. (*A*) A strong dependence of the filter output on the input values (volume and hydrophobicity) between 0 and 1 is found at position −2. (*B*) At position −3 only a weak dependence between 0.98 and 1 is observed. The amino acid residues at all other sequence positions were fixed optimally.

metallopeptidase target sites (Hendrick et al., 1989; Gavel and von Heijne, 1990). Indeed, these amino acids result in the loss of biological function (Horwich et al., 1986; Sztul et al., 1987), which is correctly predicted by the neural network system. A detailed analysis of several substitutions of the −2 Arg in the human ornithine transcarbamylase precursor revealed that R → K results in a loss of cleavage site function of about 20%, R → N of 80%, and R→G of 100% (Horwich et al., 1987). This is qualitatively predicted by the network system for *N. crassa* sequences as well (Fig. 9). In contrast to −2, position −3 appears to be more versatile, i.e., more amino

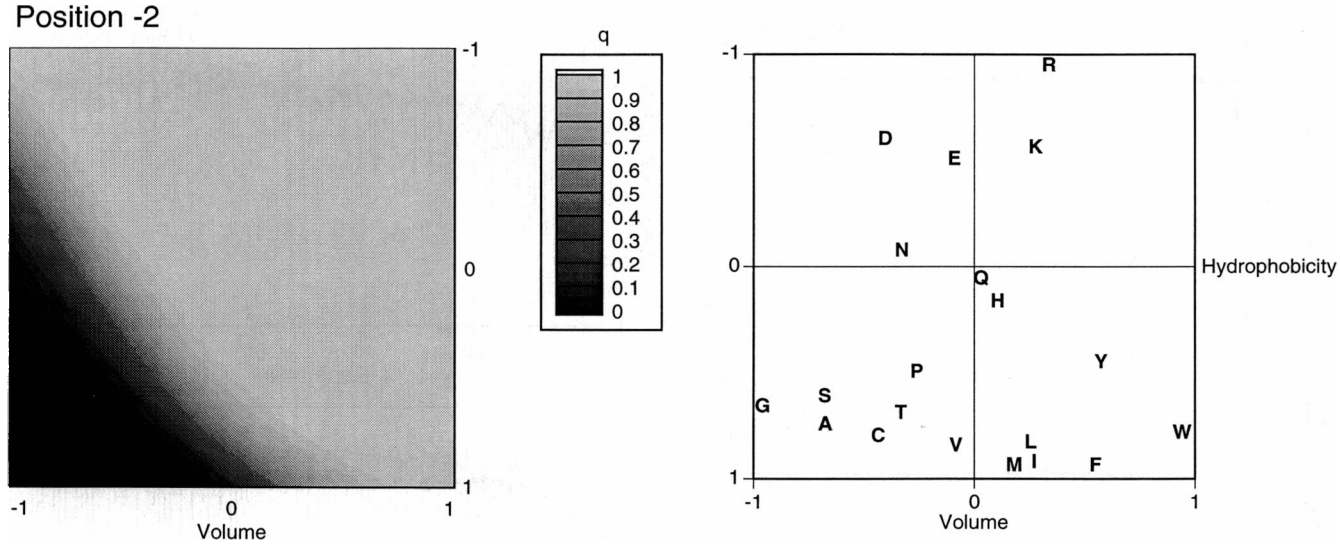

FIGURE 9 Dependence of neural network outputs (sequence quality) on physicochemical input values at position −2 of matrix metallopeptidase target sites. (*Left*) Density plot of sequence fitness. The dark area indicates a low-quality region, the light area corresponds to input values for hydrophobicity and volume which lead to high sequence quality. The amino acid residues at all other sequence positions were fixed optimally. (*Right*) The corresponding distribution of the 20 amino acids in the plane spanned by the two properties "hydrophobicity" and "volume." Here the hydrophobicity scale of Engelman et al. (1986) and the volume scale of Zamyatnin (1972) were used.

acids lead to an equivalent overall sequence quality (Fig. 8 *B*). Only the charged arginine and lysine are slightly less preferred here, although in natural cleavage sites of matrix metallopeptidase target sequences an arginine residue in −3 may exist, the "R −3 group" (Gavel and von Heijne, 1990). According to our neural filter the corresponding change in sequence quality is fairly small. This also indicates that the networks have extracted biologically relevant cleavage site features.

Convergence of the optimization process depends on the amino acid distance matrix used as a lookup-table for simulated mutations. Two distance maps have been compared: the Feng matrix (Feng et al., 1985) and the Risler matrix (Risler et al., 1988), with various numbers of offspring per generation being employed. Fig. 10 shows the statistics from 30 runs. With an increasing number of variants per generation the average sequence quality converges faster and shows smaller deviations. This holds for both the Feng matrix (left side of Fig. 10, *A–C*) and the Risler matrix (right side of Fig. 10, *A–C*). The process of convergence seems to be more irregular for the Risler matrix than for the Feng matrix (Fig. 10, *A–C*). We conclude that the two matrices compared here differ in their applicability for SME, with the Feng matrix more accurately reflecting amino acid distances in the fitness landscape given by our matrix metallopeptidase filter system. This should not be surprising, since the Feng matrix is based on the genetic and physicochemical distance of amino acids, whereas the Risler matrix is derived from the analysis of three-dimensional protein structures. Since the neural networks have been trained on the recognition of physicochemical sequence patterns it is to be expected that any matrix calculated from physicochemical properties will be preferred here.

Four additional amino acid distance matrices were tested. Their influence on the optimization process has not been analyzed in detail yet. However, they all led to very similar best sequences which can be regarded as equal in terms of their quality as the q values do not differ until the fifth decimal place. The arrows indicate the cleavage site and the sequence quality is given in parentheses:

Context matrix `RDWRGRM ↓ GGGDW (0.999962)`

Feng matrix (Feng et al., 1985) `RDWRGRM ↓ GGGDW (0.999962)`

Risler matrix (Risler et al., 1988) `RDWRCRC ↓ GGGDW (0.999955)`

Myata matrix (Myata et al., 1979) `RDWDNRC ↓ GGGDW (0.999952)`

The amino acids in the sequences obtained differ at the positions −3 and −1, the sequence obtained with the Myata matrix also differs at the −4 position. The putative cleavage site peptides are characterized by a high content of positively charged residues at positions −2, −4, and −7, and three glycine residues at positions +1, +2, and +3. Surprisingly, aspartic acid has been selected at position −6, although the presequences are known to have an almost complete lack of acidic amino acids (Arretz et al., 1991). Most of the residues of the "idealized" cleavage sites do not occur at the corresponding position in the data used for network training (Table 1). An "evolutionary history" of a design run using the Feng matrix is shown in Table 4. Already after the first generation sequences of very high quality were generated. The positions −2 and −5 are conserved early during SME

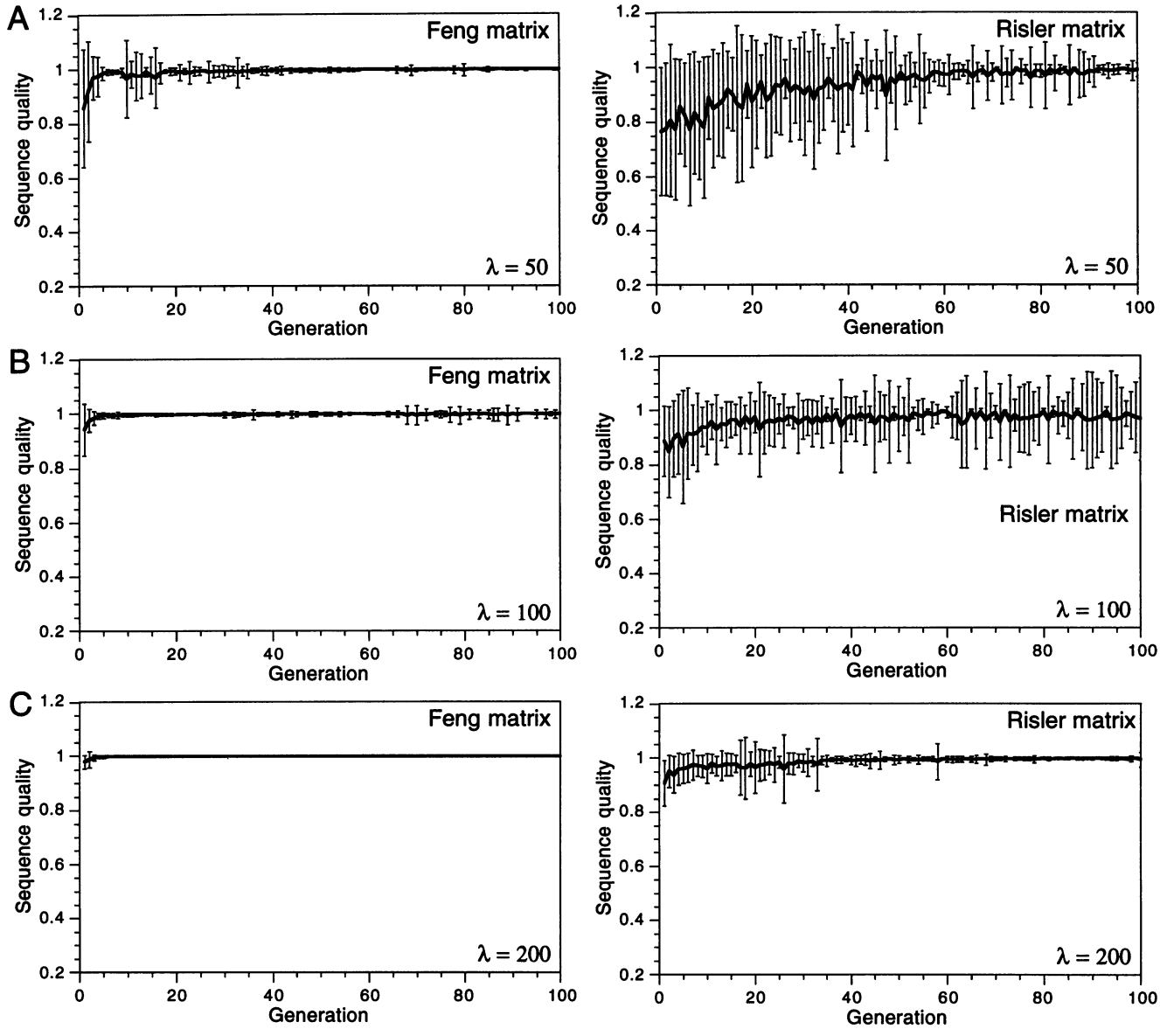

FIGURE 10 Sequence optimization performed with two different amino acid distance matrices. The course of mean sequence quality is averaged over 30 optimization runs (*thick line*). The bars indicate the standard deviation. Three different numbers of offspring per generation (λ) were used. The plots in the left column show the results using the Feng matrix (Feng et al., 1985), the right column shows the results for the Risler matrix (Risler et al., 1988). *A:* λ = 50; *B:* λ = 100; *C:* λ = 200.

(after the second and third generation, respectively) and, therefore, seem to be less variable than all the other positions investigated. The fact that high-quality sequences were obtained from random guesses in the 0th generation (Table 4) reflects the low specificity of the neural filter system (Fig. 7) as well as the usefulness of the SME optimization technique.

Whether the idealized cleavage sites have biological activity is currently being tested by an in vitro protease system (Wrede et al., in preparation). Only the results of these experiments will allow us to judge the applicability of the SME approach. Nonetheless, independently of the applicability of the sequences to the real world, de novo design of mitochondrial matrix metallopeptidase target sequences is a second example of the usefulness of the SME procedure.

Sequence windows taken from naturally occurring amino acid sequences can be regarded as multifunctional units, i.e., more than just a single function is encoded by this stretch of amino acid residues. What we have tried to do is to optimize one very special feature to obtain a set of highly specialized monofunctional sequences by extrapolating a rather simple fitness function. That attempts at designing specialized sequences are very promising has already been shown for quite a large number of examples (Sander, 1991; Thornton, 1992; Richardson et al., 1992). However, it seems clear that many functional features cannot be optimized by simply extrapolating selected physicochemical amino acid properties. Nonetheless, for targeting sequences at least, extrapolation does make sense (Bird et al., 1990; Schneider et al., 1994)

TABLE 4 Evolutionary history of an SME design run using the Feng matrix

Generation	Sequence (−7 … +5)	Quality	Generation	Sequence (−7 … +5)	Quality
			40	D G V↓G AD	0.99993992
0	RASDGRV↓ISVNH	0.99353999	41	R G V↓G AD	0.99994743
1	REWDSKV↓IDVNH	0.99614215	42	K G V↓G AD	0.99994773
2	RDFDSRV↓LDATR	0.99954051	44	R G V↓G AD	0.99994743
3	REWDG V↓VDPAK	0.99954182	45	K G V↓G AD	0.99994773
4	KD RS A↓VDPSR	0.99955058	46	R G V↓G AD	0.99994743
5	RE RP A↓VGPGF	0.99984252	47	K G V↓G AD	0.99994773
6	RD KA A↓IGPVY	0.99953765	48	E G V↓G AD	0.99994528
7	RD RP S↓IGAVF	0.99944466	49	L G V↓G AD	0.99994117
8	KD RT T↓VGATI	0.99937135	50	E G V↓G AD	0.99994528
9	RN RS T↓VGATW	0.99983269	52	H G V↓G AD	0.99994445
10	KD RS T↓IGASF	0.99964583	53	R G V↓G AD	0.99994743
11	KD RG T↓VGAGY	0.99977785	54	K G V↓G AD	0.99994773
12	RA RS T↓VSAGY	0.99973744	55	Q S V↓G AN	0.99992752
13	RG RP A↓MSASW	0.99977535	56	Q S V↓G AD	0.99994314
14	RA RA A↓MSAAY	0.99974829	57	F S V↓G AN	0.99991363
15	RF KA A↓VGAAF	0.99972999	58	E S V↓G AE	0.99993098
16	RF KA A↓VGAAY	0.99972761	59	P S V↓G AD	0.99994093
17	RS RD A↓AGAAF	0.99985790	60	D S V↓G AD	0.99994588
18	RS RQ A↓AGAAL	0.99979407	61	N S V↓G AD	0.99994421
19	KS KQ R↓AGAAW	0.99972934	63	H S V↓G AE	0.99993008
20	RS RQ A↓AGAAW	0.99989468	64	N S V↓G AD	0.99994421
22	RP KQ A↓GGATY	0.99973691	66	D S V↓G A	0.99994588
23	KA RQ A↓GGATW	0.99979132	70	E S V↓G A	0.99994475
24	RA D V↓GAATW	0.99977458	71	D S V↓G A	0.99994588
25	KT E V↓GGATW	0.99967569	77	E S V↓G S	0.99994701
26	RE G V↓G ATI	0.99980050	78	D S V↓G S	0.99994791
27	V G V↓G AAY	0.99989653	80	E S V↓G S	0.99994701
29	V A V↓G AGF	0.99988669	81	D S V↓G S	0.99994791
30	E G V↓G AGW	0.99993312	83	N S V↓G S	0.99994630
31	K S V↓G AG	0.99989039	84	E S V↓G S	0.99994701
32	A N V↓G AS	0.99990880	85	V S V↓G S	0.99994135
33	E S V↓G AG	0.99991798	90	V S A↓G G	0.99994737
34	Q G V↓G AS	0.99991983	91	V S A↓G	0.99994737
35	Q A V↓G AS	0.99991393	92	E S A↓A	0.99995416
36	N V V↓G AN	0.99990249	95	E S A↓G	0.99995583
37	N T V↓G AN	0.99992204	104	E G V↓	0.99995857
38	N A V↓G AG	0.99992895	105	D V↓	0.99995869
39	S A V↓G AG	0.99992269	148	RDWRGRM↓GGGDW	0.99996209

500 variants were generated per generation. The best sequences ("parents") of any generation are listed. Conserved residues are indicated by a blank, and the matrix metallopeptidase target sites are indicated by an arrow. The final sequence is given completely.

and our designed *E. coli* sequences tested so far show biological activity (unpublished results). For other applications new target models represented by a corresponding fitness function must be developed, and the many different types of artificial neural networks certainly are potential additional tools for this purpose.

The separation task performed by the perceptron network could, of course, have been achieved by other clustering algorithms as well. A special attraction of the network approach, however, is its simplicity and the possibility of interpreting the weight diagrams in a biochemically meaningful way. A further advantage is the fact that even interactions between residues some distance apart on the sequence can be taken into consideration. In this general sense this is only possible with larger networks containing at least one hidden layer of units. In such networks even higher-order correlations between the residues can be calculated. Another point of interest is the selection of encoding parameters. Here we have employed two special physicochemical properties. The scales chosen to rank amino acids (Engelman et al., 1986; Zamyatnin et al., 1972) are quite well suited for this particular application. The use of another scale might well have changed the outcome. We have tested several physicochemical properties with our network but so far only the two specific scales for hydrophobicity and volume selected here led to successful clustering of both training and test data. They are also orthogonal, which facilitates interpretation of the weight diagrams. Recently, new scales for volume have been published by Harpaz and co-workers (1994). Along with hydrophobicity scales resulting from principal component analysis of 38 different hydrophobicity scales (Cornette et al., 1987) this seems to be a good first choice for sequence encoding. A description of amino acid sequences in terms of physicochemical residue properties will not always be appropriate. If one is interested, e.g., in the analysis of structural features quite different descriptive parameters will be required. In a recent publication we have applied an automatic procedure for identifying useful property scales for sequence encoding (Lohmann et al., 1994). In this work networks were trained on the recognition of membrane-spanning sequences

and the input property matrix of the system was subjected to systematic optimization. This resulted in a neural filter which is not fully connected. A detailed description of the method and its possible applications to sequence analysis has been submitted for publication. This filter system is also a potential fitness function for SME.

One may differ as to the significance of optimizing mitochondrial matrix metalloprotease cleavage sites since several laboratories have shown that efficient processing can occur with many residues at the positions investigated here (Pugsley, 1989). Designing a 12-residue peptide is likely to be more easily achieved, of course, by a random mutagenesis approach, e.g., a peptide library. But if longer sequences are to be investigated, for example sequence windows covering 30 residues, the number of variants to be generated and tested will be 20^{30}. This number is far too large for an exhaustive search. Here, only good guesses or some kind of rational design are likely to be successful within a reasonable period of time. The idea behind SME is to generate a first choice of useful sequences by a systematic search in sequence space. These can be directly tested for their function, or serve as a starting point for random or site-directed mutagenesis studies. We are now working on the development of filter systems for small structural units of proteins, and the analysis of mitochondrial matrix metalloprotease target sites should be regarded as a step toward analysis of locally encoded structural features using neural networks and SME.

We thank Gerhard Müller, Walter Neupert, Georg Büldt, Ingo Rechenberg, Werner Ebeling, Oliver Grammel, and Peter Germain for stimulating discussions and encouragement. We are also grateful to Peter Germain for careful editing. Michael Arretz kindly supplied us with the *N. crassa* sequence data containing experimentally confirmed cleavage sites.
This work was supported by the BMFT (DETHEMO project) and the FCI.

REFERENCES

Arretz, M., H. Schneider, U. Wienhues, and W. Neupert. 1991. Processing of mitochondrial precursor proteins. *Biomed. Biochim. Acta.* 50:403–412.

Bäck, T., and H. P. Schwefel. 1993. An overview of evolutionary algorithms for parameter optimization. *Evol. Comput.* 1:1–24.

Barker, W. C., D. G. George, H. W. Mewes, and A. Tsugita. 1992. The PIR-International protein sequence database. *Nucleic Acids Res.* 20:2023–2026.

Bird, P., M. J. Gething, and J. Sambrook. 1990. The functional efficiency of a mammalian signal peptide is directly related to its hydrophobicity. *J. Biol. Chem.* 265:8420–8425.

Cornette, J. L., K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzofsky, and C. DeLisi. 1987. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* 195:659–685.

Dalbey, R. E., and G. von Heijne. 1992. Signal peptidases in prokaryotes and eukaryotes—a new protease family. *Trends Biochem. Sci.* 17: 474–478.

Davidor, Y., and H. P. Schwefel. 1992. An introduction to adaptive optimization algorithms based on principles of natural evolution. *In* Dynamic, Genetic, and Chaotic Programming. B. Souçek and the IRIS Group, editors. John Wiley, New York. 183–202.

Dayhoff, M. O., and R. V. Eck. 1968. A model of evolutionary change in proteins. *In* Atlas of Protein Sequence and Structure. M. O. Dayhoff, editor. National Biomedical Research Foundation, Washington, DC. 345–352.

Ebeling, W., A. Engel, and R. Feistel. 1990. Physik der Evolutionsprozesse. Akademie-Verlag, Berlin.

Engelman, D. A., T. A. Steitz, and A. Goldman. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* 15:321–353.

Feng, D. F., M. S. Johnson, and R. F. Doolittle. 1985. Aligning amino acid sequences: comparison of commonly used methods. *J. Mol. Evol.* 21: 112–125.

Fontana, W., P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster. 1993. RNA folding and combinatory landscapes. *Phys. Rev. E.* 47: 2083–2099.

Gavel, Y., L. Nilsson, and G. von Heijne. 1988. Mitochondrial targeting sequences—why "non-amphiphilic" peptides may still be amphiphilic. *FEBS Lett.* 235:173–177.

Gavel, Y., and G. von Heijne. 1990. Cleavage site motifs in mitochondrial targeting peptides. *Protein Eng.* 4:33–37.

George, D. G., W. C. Barker, and L. T. Hunt. 1990. Mutation data matrix and its uses. *Methods Enzymol.* 183:333–351.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science.* 185:862–864.

Harpaz, Y., M. Gerstein, and C. Chothia. 1994. Volume changes on protein folding. *Structure.* 2:641–649.

Hartl, F. U., and W. Neupert. 1990. Protein sorting to mitochondria—evolutionary conservations of folding and assembly. *Science.* 247: 930–938.

Hawlitschek, G., H. Schneider, B. Schmidt, M. Tropschug, F. U. Hartl, and W. Neupert. 1988. Mitochondrial protein import: identification of processing peptidase of PEP, a processing enhancing protein. *Cell.* 53:795–806.

Hendrick, J. P., P. E. Hodges, and L. E. Rosenberg. 1989. Survey of amino-terminal proteolytic cleavage sites in mitochondrial precursor proteins: leader peptides cleaved by two matrix proteases share a three-amino acid motif. *Proc. Natl. Acad. Sci. USA.* 86:4056–4060.

Holbrook, S. R., S. M. Muskal, and S. H. Kim. 1993. Predicting protein structural features with artificial neural networks. *In* Artificial Intelligence and Molecular Biology. L. Hunter, editor. AAAI Press, Menlo Park, CA/MIT Press, Cambridge, MA. 161–194.

Horwich, A. L., F. Kalousek, W. A. Fenton, K. Furtak, R. A. Pollock, and L. E. Rosenberg. 1987. The ornithine transcarbamylase leader peptide directs mitochondrial import through both its midportion structure and net positive charge. *J. Cell Biol.* 105:669–677.

Horwich, A. L., F. Kalousek, W. A. Fenton, R. A. Pollock, and L. E. Rosenberg. 1986. Targeting of pre-ornithine transcarbamylase to mitochondria: definition of critical regions and residues in the leader peptide. *Cell.* 44:451–459.

Horwich, A. L., F. Kalousek, and L. E. Rosenberg. 1985. Arginine in the leader peptide is required for both import and proteolytic cleavage of a mitochondrial precursor. *Proc. Natl. Acad. Sci. USA.* 82:4930–4933.

Kaiser, C. A., D. Preuss, P. Grisafi, and D. Botstein. 1987. Many random sequences functionally replace the secretion signal sequence of yeast invertase. *Science.* 235:312–317.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. University Press, Cambridge.

King, R. D., and M. J. E. Sternberg. 1990. Machine learning approach for the prediction of protein secondary structure. *J. Mol. Biol.* 216:441–457.

Laforet, G. A., and D. A. Kendall. 1991. Functional limits of conformation, hydrophobicity, and steric constraints in prokaryotic signal peptide cleavage regions. *J. Biol. Chem.* 266:1326–1334.

Lohmann, R. 1992. Structure evolution in neural systems. *In* Dynamic, Genetic, and Chaotic Programming. B. Souçek and the IRIS Group, editors. John Wiley, New York. 395–413.

Lohmann, R., G. Schneider, D. Behrens, and P. Wrede. 1994. A neural network model for the prediction of membrane-spanning amino acid sequences. *Protein Sci.* 3:1597–1601.

Mathews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* 405:442–451.

Mathews, J. H. 1992. Numerical Methods for Mathematics, Science, and Engineering. Prentice-Hall, Englewood Cliffs, NJ.

Minsky, M. L., and S. Papert. 1988. Perceptrons. MIT Press, Cambridge, MA.

Miura, S., Y. Amaya, and M. Mori. 1986. A metalloprotease involved in the processing of mitochondrial precursor proteins. *Biochem. Biophys. Res. Commun.* 134:1151–1159.

Myata, T., S. Miyazawa, and T. Yasunaga. 1979. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* 12:219–236.

Nguyen, M., C. Argan, W. P. Sheffield, A. W. Bell, D. Shields, and G. C. Shore. 1987. A signal sequence domain essential for processing, but not import, of mitochondrial pre-ornithine carbamyl transferase. *J. Cell Biol.* 104:1193–1198.

Nunnari, J., T. D. Fox and P. Walter. 1993. A mitochondrial protease with two catalytic subunits of nonoverlapping specificities. *Science.* 262:1997–2004.

Parker, G. A., and J. Maynard Smith. 1990. Optimality theory in evolutionary biology. *Nature.* 348:27–33.

Perlman, D., and H. A. Halvorson. 1983. A putative signal peptidase recognition site and sequence in eukaryotic and prokaryotic signal peptides. *J. Mol. Biol.* 167:391–490.

Pfanner, N., and W. Neupert. 1990. The mitochondrial protein import apparatus. *Annu. Rev. Biochem.* 59:331–353.

Pugsley, A. P. 1989. Protein Targeting. Academic Press, San Diego.

Qian, N., and T. J. Sejnowski. 1988. Predicting the secondary structure of globular 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.

Richardson, J. S., D. C. Richardson, N. B. Tweedy, K. M. Gernert, T. P. Quinn, M. H. Hecht, B. W. Erickson, Y. Yan, R. D. McClain, M. E. Donlan, and M. C. Surles. 1992. Looking at proteins: representations, folding, packing, and design. *Biophys. J.* 63:1186–1209.

Risler, J. L., M. O. Delorme, H. Delacroix, and A. Henaut. 1988. Amino acid substitutions in structurally related proteins: a pattern recognition approach. Determination of a new and efficient scoring matrix. *J. Mol. Biol.* 204:1019–1029.

Rosenblatt, F. 1962. Principles of Neurodynamics. Spartan Books, New York.

Rumelhart, D. E., J. L. McClelland, and The PDP Research Group (editors). 1986. Parallel Distributed Processing. MIT Press, Cambridge, MA.

Sander, C. 1991. De novo design of proteins. *Curr. Opin. Struct. Biol.* 1:630–638.

Schatz, G. 1993. The protein import machinery of mitochondria. *Protein Sci.* 2:141–146.

Schneider, A., M. Behrens, P. Scherer, E. Pratje, G. Michaelis, and G. Schatz. 1991. Inner membrane protease I, an enzyme mediating intramitochondrial protein sorting in yeast. *EMBO J.* 10:247–254.

Schneider, G., H. Christmann, and P. Wrede. 1992. Protein targeting in the view of endocytobiology. *Endocytobiosis Cell Res.* 9:83–101.

Schneider, G., S. Röhlk, and P. Wrede. 1993. Analysis of signal peptidase cleavage sites with a perceptron-type neural network. *Biochem. Biophys. Res. Commun.* 194:951–959.

Schneider, G., J. Schuchhardt, and P. Wrede. 1994. Artificial neural networks and simulated molecular evolution are potential tools for sequence-oriented protein design. *Comput. Appl. Biosci.* 10:635–645.

Schneider, G., and P. Wrede. 1992. Modular feature extraction in amino acid sequences by artificial neural networks: analog model for symbiogenous constraints. *Endocytobiosis Cell Res.* 9:1–12.

Schneider, G., and P. Wrede. 1993a. Signal analysis in protein targeting sequences. *Protein Seq. Data Anal.* 5:227–236.

Schneider, G., and P. Wrede. 1993b. Development of artificial neural filters for pattern recognition in protein sequences. *J. Mol. Evol.* 36:586–595.

Schneider, G., and P. Wrede. 1993c. Prediction of the secondary structure of proteins from the amino acid sequence with artificial neural networks. *Angew. Chemie Int. Ed. Engl.* 32:1141–1143.

Schneider, G., and P. Wrede. 1994. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. *Biophys. J.* 66:335–344.

Sztul, E. S., J. P. Hendrick, J. P. Kraus, D. Wall, F. Kalousek, and L. E. Rosenberg. 1987. Import of rat ornithine transcarbamylase precursor into mitochondria: two-step processing of the leader peptide. *J. Cell Biol.* 105:2631–2639.

Thornton, J. 1992. Lessons from analyzing protein structures. *Curr. Opin. Struct. Biol.* 2:888–894.

Vassarotti, A., W. J. Chen, C. Smagula, and M. G. Douglas. 1987. Sequences distal to the mitochondrial targeting sequence are necessary for the maturation of the F_1-ATPase β-subunit precursor in mitochondria. *J. Biol. Chem.* 262:411–418.

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

von Heijne, G. 1986. Mitochondrial targeting sequences may form amphiphilic helices. *EMBO J.* 5:1335–1342.

von Heijne, G., J. Steppuhn, and R. G. Herrmann. 1989. Domain structure of mitochondrial and chloroplast targeting peptides. *Eur. J. Biochem.* 180:535–545.

Wrede, P., and G. Schneider (editors). 1994. Concepts in Protein Engineering and Design. Walter de Gruyter, Berlin/New York.

Zamyatnin, A. A. 1972. Protein volume in solution. *Prog. Biophys. Mol. Biol.* 24:107–123.