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

Gisbert Schneider and Paul Wrede

Freie Universität Berlin, Institut für Experimentalphysik, AG Biophysik, Arnimallee 14, D-14195 Berlin, Germany; and Laser-Medizin-Zentrum, Institut an der Freien Universität Berlin, Krahmerstraße 6–10, D-12207 Berlin, Germany

ABSTRACT   A method for the rational design of locally encoded amino acid sequence features using artificial neural networks and a technique for simulating molecular evolution has been developed. De novo in machine design of *Escherichia coli* leader peptidase (SP1) cleavage sites serves as an example application. A modular neural network system that employs sequence descriptions in terms of physicochemical properties has been trained on the recognition of characteristic cleavage site features. It is used for sequence qualification in the design cycle, representing the sequence fitness function. Starting from a random sequence several cleavage site sequences were generated by a simulated molecular evolution technique. It is based on a simple genetic algorithm that takes the quality values calculated by the artificial neural network as a heuristic for inductive sequence optimization. Simulated in vivo mutation and selection allows the identification of predominant sequence positions in *Escherichia coli* signal peptide cleavage site regions (positions −2 and −6). Various amino acid distance maps are used to define metrics for the step size of mutations. Position-specific mutability values indicate sequence positions exposed to high or low selection pressure in the simulations. The use of several distance maps leads to different courses of optimization and to various idealized sequences. It is concluded that amino acid distances are context dependent. Furthermore, a method for identification of local optima during sequence optimization is presented.

INTRODUCTION

Genetic algorithms can be applied to systematic optimization of amino acid sequences in large search spaces (Dandekar and Argos, 1992). They are well suited for training of artificial neural networks specialized on pattern recognition in protein sequences (Schneider and Wrede, 1993a). Simulated molecular evolution (SME) of amino acid sequences is a new technique for rational sequence-oriented protein design employing an amino acid sequence generating procedure that is tightly coupled to a selection mechanism represented by an artificial neural network. Standard evolutionary algorithms (Rechenberg, 1973; Goldberg, 1989; Holland, 1992; Koza, 1992) have been used for both neural network training and sequence optimization by SME. In the protein design cycle the network system provides the sequence fitness function to be used for an evaluation of sequence features, and simulated amino acid mutations are applied to come up with new sequences (Fig. 1).

The trained neural filter must be able to recognize desired features in an amino acid sequence and calculate a real-coded quality value to be used as a measure for sequence fitness in sequence optimization. Because multilayer feedforward networks ("neural filters") can be regarded as universal function estimators (Hornik et al., 1989) they are a method of choice for feature extraction in amino acid sequences and seem to be well suited for representing sequence-function or sequence-structure relations. Further, artificial neural networks process sequence information in an inherently parallel way and are able to extract essential sequence features by distinguishing relevant from irrelevant sequence information (Holley and Karplus, 1991; Hirst and Sternberg, 1992). For these unique advantages we selected artificial neural networks for the development of accurate filter systems (Fig. 1).

Starting from a random sequence any feature can, in principle, be designed by SME, provided a reliable filter system is available. Because of the networks' parallel way of information processing the effects of simulated mutations are always calculated in a parallel, context-dependent manner. Therefore, an amino acid sequence will be optimized as a whole block rather than separate and independent optimization of single positions. Instead of subsequent position-specific design all residues under investigation will be optimized in parallel, taking into account their specific interactions that are represented by the connection weights of the neural filter system.

Amino acid sequences can be engineered and designed by in vivo selection (Wells and Lowman, 1992), but a major disadvantage of this technique is that laborious screening procedures for the "optimal" sequence in many cases is required. In contrast, any "rational" approach will build up a certain sequence or structure with desired properties and function from a model based on theoretical principles, and no exhaustive experimental screening will be needed (Richardson et al., 1992). However, a "rational" approach requires perfect knowledge of the essential structural features that are responsible for a certain protein or peptide function. As only a limited number of highly resolved protein tertiary structures are known at present the application
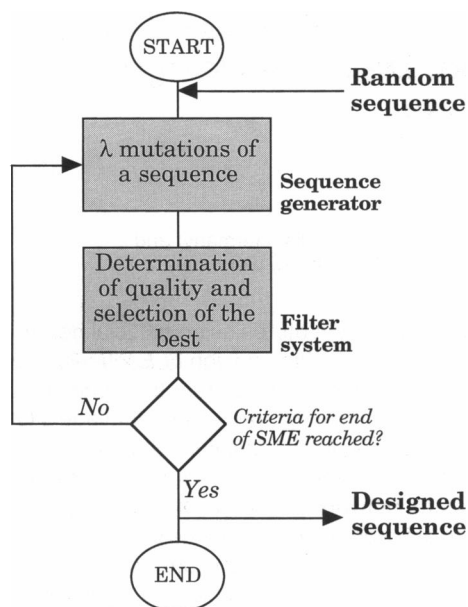
FIGURE 1   Scheme of a protein design cycle using SME for systematic sequence optimization. Starting from a random sequence, mutational changes are evaluated by a neural filter system. λ: Number of generated new sequences. Optimization stops after a certain number of simulated generations (200 in the experiments) or when a sequence having the quality value 1 is designed.

of "rational" methods starting from a three-dimensional model of the protein still is rather limited. Significant progress, however, has been made in de novo designing protein structures during the last years (Sander, 1991).

Simulated in vivo selection using a standard genetic algorithm, the (1, λ) Evolution Strategy (Rechenberg, 1973; Schneider and Wrede, 1993a), is applied to generate and test variant amino acid sequences where the neural network transformation function is employed as a supervising system turning the blind, "irrational" search into a systematic, "rational" design approach. Sequence-oriented de novo design of cleavage site regions for the *Escherichia coli* (*E. coli*) leader peptidase (SP1) (von Heijne et al., 1988) serves as an example application for the new SME technique.

MATERIALS AND METHODS

Data

Twenty-four sequences of *E. coli* periplasmic protein precursors with known leader peptidase cleavage sites were taken from the SwissProt database (Release 20) for protein filter induction (Schneider and Wrede, 1992, 1993a). They were split into a training set of 17 sequences and a test set of 7 sequences. The data were restricted to strings of 12 residues each covering the positions −10 to +2 relative to the leader peptidase processing site (see Fig. 4) (Schneider and Wrede, 1992, 1993a, b). For every positive cleavage site example four negative examples were randomly selected from the corresponding precursor sequence. Thus, the training set that was used for neural network training consisted of 85 examples, and the test set for evaluation of the generalization ability of the networks covered 35 examples.

The 17 precursor sequences of the training set were: Glucose-1-phosphatase, L-arabinose-binding protein, Lysine-arginine-ornithine-binding protein, L-asparaginase II, Peptidyl-prolyl *cis-trans*-isomerase,

D-galactose-binding protein, Gamma-glutamyltranspeptidase, Glutamine-binding protein, Leu/Ile/Val-binding protein, Maltose-binding protein, Penicillin-insensitive murein endopeptidase, Penicillin acylase, Phosphate-repressible phosphate-binding protein, Glycine-betaine-binding protein, Alkaline phosphatase, pH 2.5 acid phosphatase.

The 7 precursor sequences of the test set were: Ribonuclease I, Protease III, D-ribose-binding periplasmic protein, Sulfate-binding protein, Periplasmic trehalase, Glycerol-3-phosphate-binding protein, UDP-sugar hydrolase.

A complete list of the training and test sequences including all negative and positive examples is available from the authors on request.

Network architecture and training

Three-layered feedforward networks were used for feature extraction from the sequence data (Fig. 2). The hidden units and the single output unit used a sigmoidal transfer function (squashing function) $S(unit_{in})$, and all units were equally biased:

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

where

$$unit_{in} = \sum_{l,m} X_{l,m} w_{l,m}$$

$X_{l,m}$ is a physicochemical property value of an amino acid (Fig. 2). A sequence description in terms of physicochemical properties is a very helpful data representation for recognition and prediction of leader peptidase cleavage sites by artificial neural networks (Schneider and Wrede, 1992, 1993a, b). The normalized property scales hydrophobicity (Engelman et al., 1986), hydrophilicity (Hopp and Woods, 1981), polarity (Jones, 1975), and side-chain volume (Zamyatnin, 1972) were selected to build up the network input matrix $X_{l,m}$ (Fig. 2).

For determination of the network weights $w_{k,l,m}$ and $w_{j,k}$ (Fig. 2) a (1, 100) evolution strategy with adaptive control of the learning stepsize was used (Schneider and Wrede, 1993a). Two supervising functions were applied:

1. The least-mean-square error of the network output (E_{LMS}) that had to be minimized (the desired output value for a positive example $P_{i_{out}}$ was 1,
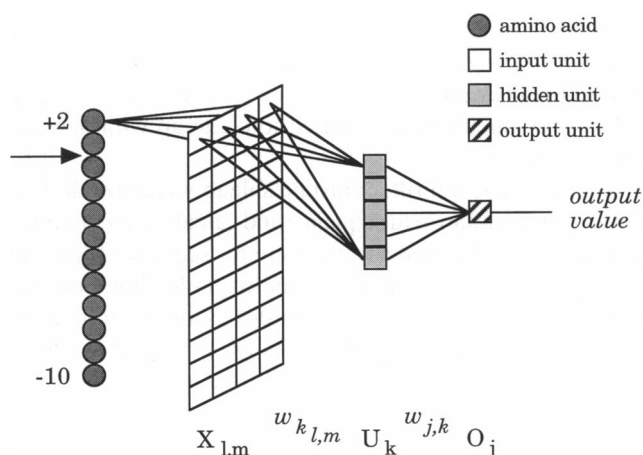

FIGURE 2   Model of the artificial neural network architecture used for the development of leader peptidase cleavage site filters. The arrow indicates the processing site, the numbers indicate sequence positions. A 12 × 4 input matrix consisting of physicochemical property values has been selected for sequence encoding. For clarity, only some network connections are drawn. An evolution strategy was applied to network training. The final filter system consists of three independent network modules.

and for a negative example $N_{j_{out}}$ the desired output was 0):

$$E_{LMS} = \frac{\sum_{i=1}^{pos}(1.0 - P_{i_{out}})^2 + \sum_{j=1}^{neg}(N_{j_{out}})^2}{(n + m)} \rightarrow \text{Min.}$$

where *pos* is the number of positive examples and *neg* is the number of negative examples in the training set.

2. the prediction accuracy Q = P + N (*P*: probability of positive correct prediction, *N*: probability of negative correct prediction), which must reach a value of 1.

Training was stopped when the prediction accuracy in the training set reached a value of 1 or when 200 learning cycles had passed. To determine an ideal network architecture the number of hidden units was systematically changed between 1 and 12 units, and the networks with the best prediction results and lowest E_{LMS} were combined to form a modular multinetwork system. The output values of the single systems were multiplied for this purpose. Sequence quality was calculated according to the following equation, where Out_1, Out_2, and Out_3 are the outputs of the single networks:

$$\text{Quality} = Out_1 \times Out_2 \times Out_3$$

A single network output Out_n is defined by the network transformation function; $S(x)$ is the sigmoidal transfer function (activity function) given above:

$$Out_n = S\left(\sum_{j,k} w_{j,k} S\left(\sum_{l,m} W_{k_{l,m}} X_{l,m}\right)\right)$$

Further details of the special training technique can be found in a previous publication on the development of artificial neural networks for pattern recognition in amino acid sequences (Schneider and Wrede, 1993a).

Simulated molecular evolution

For the formation of amino acid sequences to be qualified by the neural network filter an evolutionary algorithm has been developed that is based on a simple (1, λ) evolution strategy (Rechenberg, 1973; Davidor and Schwefel, 1992). In every simulated generation (Fig. 1) a 12-residue sequence (parent sequence) is mutated λ times leading to an offspring of λ sequences. The best of the offspring according to the neural filter is selected as parent sequence for the next generation. A total of 200 generations (optimization cycles) was allowed, and λ was 500 in all experiments.

To define large and small mutations and to fulfill the requirement for strong causality, which is an essential prerequisite for any systematic optimization (i.e., small changes in a sequence will lead to only small changes of its quality), the λ mutations had to occur Gaussian-distributed around the parent sequence (Fig. 3). Five different amino acid distance maps were employed; three were taken from the literature (Grantham, 1974; Myata et al., 1979; Risler et al., 1988), one was a random matrix, and one has been calculated using the four physicochemical properties that were used for network training as describing parameters ("Context-matrix", Table 1). The Euclidian distance between the four selected physicochemical parameters of the 20 amino acids was used to calculate the distance values. All distance maps were normalized to obtain comparable values between 0 and 1 (Fig. 3).

Furthermore, every sequence position was allowed to adapt its mutability σ (standard deviation of the Gaussian distribution of mutations) to facilitate convergence (Fig. 3). This was achieved by optimizing σ itself by a (1, 500) evolution strategy. The initial value was 1 at every position. The average mutability ("mean step") was calculated by summing up all individual σ values and dividing that by the number of sequence positions to be designed (12 in the example application). This resulted in a mutation rule that is the same for every sequence position:

$$d = G \times \sigma$$

$$R_{new} = F(d, R_{old})$$

The new residue (R_{new}) is a function of the old residue (R_{old}), the position-specific mutability σ, a Gaussian-distributed random number G,
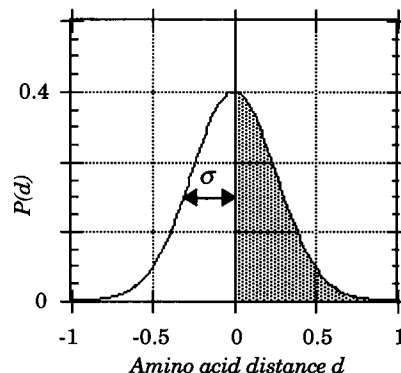
FIGURE 3   Scheme for a simulated mutation of an amino acid by SME. Starting from the residue to be mutated (distance d = 0) the order of amino acids along the distance axis is given by the selected distance matrix. With decreasing probability, *P(d)*, an amino acid is mutated to another one spaced further apart. A Gaussian distribution determines the mutation probability. Its variance σ can be interpreted as a position-specific mutability. Because positive distance values were used, only the gray-shaded part of the Gaussian was calculated for mutation.

and the employed amino acid distance metric, which determines the residue selection function $F(d, R_{old})$. Gaussian-distributed random numbers were generated using two equally distributed random real numbers *i* and *j* having values between 0 and 1:

$$G = \sigma \sqrt{2.0 \ln(i)} \sin(\pi j)$$

To decide whether sequence optimization by SME has putative local optima and to know whether a designed sequence reached a quality optimum, a one-dimensional plot of the search space is calculated. Fig. 4 gives a calculation scheme for a peptide consisting of only three residues, R1, R2, and R3. In steps of 0.01 the amino acid distance from the original residues R1, R2, and R3 is systematically increased, and every amino acid is changed according to an amino acid distance map. The corresponding quality of the new sequence is determined by the neural filter system. This results in a well-defined diagonal line of sight through the search space giving the sequence quality as a function of distance from the original sequence.

All experiments were performed on a PC running under DOS. The programs are implemented in the programming language Modula2.

RESULTS

Development of artificial neural network filters

Twelve artificial neural feedforward networks were trained by a (1, 100) Evolution Strategy (Rechenberg, 1973; Schneider and Wrede, 1993a), a simple genetic algorithm mainly based on a repetitive mutation and selection scheme for the networks' weight values, on the recognition of leader peptidase cleavage sites in *E. coli* periplasmic protein precursor sequences. Three different filters extracted relevant cleavage site features leading to 100% correct classification of both the training sequences and the independent test data (Table 2). The training protocols of the three networks show a decrease of the networks' least-mean-square error (E_{LMS}) and an increase of their prediction accuracies (Q) during network optimization (Fig. 5). Both functions were used as supervising functions for network training by the evolutionary algorithm (see methods). Network training was stopped when Q reached the value 1, i.e., when all training patterns

TABLE 1 The Context matrix giving the normalized amino acid distances

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	0.112	0.819	0.827	0.54	0.208	0.696	0.407	0.891	0.406	0.379	0.318	0.191	0.372	1	0.094	0.22	0.273	0.739	0.552
C	0.114	0	0.847	0.838	0.437	0.32	0.66	0.304	0.887	0.301	0.277	0.324	0.157	0.341	1	0.176	0.233	0.167	0.639	0.457
D	0.729	0.742	0	0.124	0.924	0.697	0.435	0.847	0.249	0.841	0.819	0.56	0.657	0.584	0.295	0.667	0.649	0.797	1	0.836
E	0.79	0.788	0.133	0	0.932	0.779	0.406	0.86	0.143	0.854	0.83	0.599	0.688	0.598	0.234	0.726	0.682	0.824	1	0.837
F	0.508	0.405	0.977	0.918	0	0.69	0.663	0.128	0.903	0.131	0.169	0.541	0.42	0.459		0.548	0.499	0.252	0.207	0.179
G	0.206	0.312	0.776	0.807	0.727	0	0.769	0.592	0.894	0.591	0.557	0.381	0.323	0.467	1	0.158	0.272	0.464	0.923	0.728
H	0.896	0.836	0.629	0.547	0.907	1	0	0.848	0.566	0.842	0.825	0.754	0.777	0.716	0.697	0.865	0.834	0.831	0.981	0.821
I	0.403	0.296	0.942	0.891	0.134	0.592	0.652	0.000	0.892	0.013	0.057	0.457	0.311	0.383		0.443	0.396	0.133	0.339	0.213
K	0.889	0.871	0.279	0.149	0.957	0.9	0.438	0.899	0	0.892	0.871	0.667	0.757	0.639	0.154	0.825	0.759	0.882	1	0.848
L	0.405	0.296	0.944	0.892	0.139	0.596	0.653	0.013	0.893	0	0.062	0.452	0.309	0.376	1	0.443	0.397	0.133	0.341	0.205
M	0.383	0.276	0.932	0.879	0.182	0.569	0.648	0.058	0.884	0.062	0	0.447	0.285	0.372	1	0.417	0.358	0.12	0.391	0.255
N	0.424	0.425	0.838	0.835	0.766	0.512	0.78	0.615	0.891	0.603	0.588	0	0.266	0.175		0.361	0.368	0.503	0.945	0.641
P	0.22	0.179	0.852	0.831	0.515	0.376	0.696	0.363	0.875	0.357	0.326	0.231	0	0.228		0.196	0.161	0.244	0.72	0.481
Q	0.512	0.462	0.903	0.861	0.671	0.648	0.765	0.532	0.881	0.518	0.505	0.181	0.272	0	1	0.461	0.389	0.464	0.831	0.522
R	0.919	0.905	0.305	0.225	0.977	0.928	0.498	0.929	0.141	0.92	0.908	0.69	0.796	0.668	0	0.86	0.808	0.914	1	0.859
S	0.1	0.185	0.801	0.812	0.622	0.17	0.718	0.478	0.883	0.474	0.44	0.289	0.181	0.358	1	0	0.174	0.342	0.827	0.615
T	0.251	0.261	0.83	0.812	0.604	0.312	0.737	0.455	0.866	0.453	0.403	0.315	0.159	0.322	1	0.185	0	0.345	0.816	0.596
V	0.275	0.165	0.9	0.867	0.269	0.471	0.649	0.135	0.889	0.134	0.12	0.38	0.212	0.339		0.322	0.305	0	0.472	0.31
W	0.658	0.56	1	0.931	0.196	0.829	0.678	0.305	0.892	0.304	0.344	0.631	0.555	0.538	0.968	0.689	0.638	0.418	0	0.204
Y	0.587	0.478	1	0.932	0.202	0.782	0.678	0.230	0.904	0.219	0.268	0.512	0.444	0.404	0.995	0.612	0.557	0.328	0.244	0

The residue properties hydrophobicity, hydrophilicity, polarity, and side-chain volume were employed as describing parameters. The matrix was calculated using the Euklidian distance measure.
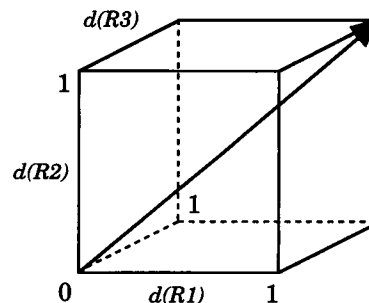

FIGURE 4   A line of sight through a three-dimensional search space (sequence space) to identify sequence optima. The quality along the resultant (*thick arrow*) is determined by an artificial neural filter. R1, R2, and R3 indicate three residues of a tripeptide; $d(R1)$, $d(R2)$, and $d(R3)$ are distance values. In the design experiments, a 12-dimensional sequence space was investigated.

TABLE 2 Results of neural network training by evolutionary computing

No. of Hidden Neurons	Q(train)	Q(test)	$E_{LMS} \times 10^{-1}$
1	0.8	0.8	2.30
2	1.0	1.0	1.28
3	1.0	0.89	0.72
4	1.0	1.0	1.19
5	1.0	0.94	1.32
6	1.0	0.97	8.24
7	1.0	0.97	0.67
8	1.0	0.97	1.06
9	1.0	0.97	0.67
10	1.0	0.94	1.20
11	1.0	1.0	0.76
12	1.0	0.94	0.79

The prediction accuracy of trained cleavage-site filters in the training set (Q(train)) and the test set (Q(test)) are given. E_{LMS} is the least-mean-square error of the networks. The architectures with 2, 4, and 11 hidden layer neurons were combined for the final filter system.

are correctly classified. With an increasing number of network connection weights more time ("Generations") was needed for their optimization (Fig. 5).

The three successful network architectures employed 2, 4, and 11 hidden layer units. To reduce overprediction, i.e., false positive predictions of cleavage sites, and to allow large network output values only for correct predictions the networks were combined to form a single network system by multiplying their output values (Schneider and Wrede, 1992). This filter system consisting of three modules was used for sequence classification and calculation of cleavage site quality for sequence design in the SME approach. It represented the sequence fitness function for systematic optimization of a random sequence.

Surprisingly, the best network architectures, i.e., the networks with the highest prediction accuracy, do not have the lowest E_{LMS} values of all 12 trained network architectures, although 100% correct data classification can be achieved (Table 2). It is assumed that the networks having the lowest E_{LMS} specialized on the training data rather than extracting generalizing features and, therefore, overlearning occurred.
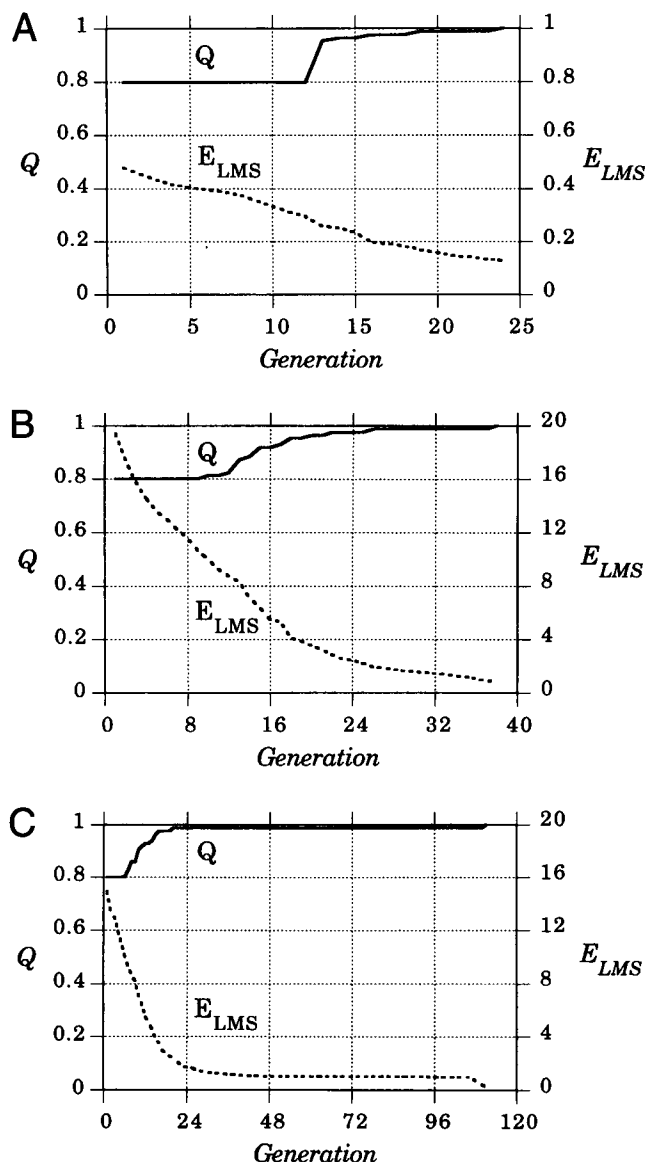
FIGURE 5   Protocols of neural network training. Two supervising functions of the genetic algorithm used are shown (E_{LMS}: least-mean-square error; Q: prediction accuracy in the training set). (*A*) Two hidden network units; (*B*) Four hidden network units; and (*C*) 11 hidden network units.

The prediction results of the resulting three-module network are shown in Fig. 6. In most cases the highest output values are assigned to the correct cleavage sites. This means that it is reasonable to design an idealized cleavage site region by searching for a sequence of highest quality (network output value). Nonetheless, it must be stressed that there are some completely false predictions in the independent test set sequences (Fig. 6). This observation is interpreted as a network error indicating that the filter system has to be further improved. Whether the false predictions have a biological background cannot be decided from these results alone. However, the obtained filter system for leader peptidase cleavage sites seems to be well suited for sequence quali-

fication and thus to serve as a sufficiently reliable filter for sequence design to demonstrate the SME idea.

Sequence design

The resulting neural filter system was used as selection operator in the design of the 12-residue cleavage site region of leader peptidase. Successful optimization by evolutionary algorithms requires a causal connection between the parameters to be optimized (the amino acid sequence) and the corresponding quality value (network output). Therefore, the SME technique demands for a definition of large and small mutations in terms of an amino acid distance map (see *Materials and Methods*). To find the appropriate metric for leader-peptidase cleavage sites five different amino acid distance matrices were tested for their applicabilty. A maximum of 200 generations with (500 × 12) point mutations per generation was specified for a SME-design run. Thus, a total of 1,200,000 mutations were generated leading to 100,000 different sequences per distance matrix. The best sequence of a generation was declared as the "parent" sequence for the next generation. Every design experiment was repeated five times leading to 25 idealized cleavage sites with slightly different amino acid sequences. The corresponding sequence qualities are given in parentheses, the cleavage sites are indicated by an arrow.

Context matrix (Table 1):
  1 FFFFGWYGWA ↓ RE (0.89863014)
  2 FWMFGWWGWA ↓ RG (0.89854997)
  3 FIFFGWYGWA ↓ RE (0.89852982)
  4 FFMFGWYGWG ↓ NE (0.89848864)
  5 FFMFGWYGWG ↓ NE (0.89848864)
Grantham matrix (Grantham, 1974):
  1 FFFWGWWGWA ↓ RE (0.89861321)
  2 FWMWGWYGWA ↓ RE (0.89858841)
  3 FWMWGWLGWA ↓ RE (0.89858329)
  4 FCFWGWWGWV ↓ RK (0.89826369)
  5 LFTFGYNGWW ↓ QD (0.89733797)
Myata matrix (Myata et al., 1979):
  1 FWMFGWVGWV ↓ RE (0.89856272)
  2 FFMFGWYGWV ↓ RE (0.89854317)
  3 IWMWGWWGWV ↓ RE (0.89848840)
  4 FWFFGWNGWG ↓ RK (0.89844328)
  5 FFFWGWQGWG ↓ RK (0.89837473)
Risler matrix (Risler et al., 1988):
  1 IWIWGWYGWC ↓ RK (0.89847743)
  2 FIMWGYYGWC ↓ RR (0.89812397)
  3 FFMWGYCGFC ↓ RE (0.89797747)
  4 FWIWSYYCWC ↓ RK (0.89728862)
  5 VIMWGYNGFG ↓ RK (0.89672505)
Random matrix:
  1 FWFTNWLQWF ↓ RT (0.88303786)
  2 ITVLCFWQIA ↓ DD (0.82229781)
  3 MLCGGLYVHM ↓ EM (0.74732530)
  4 GISVLWCVHY ↓ YK (0.59404093)
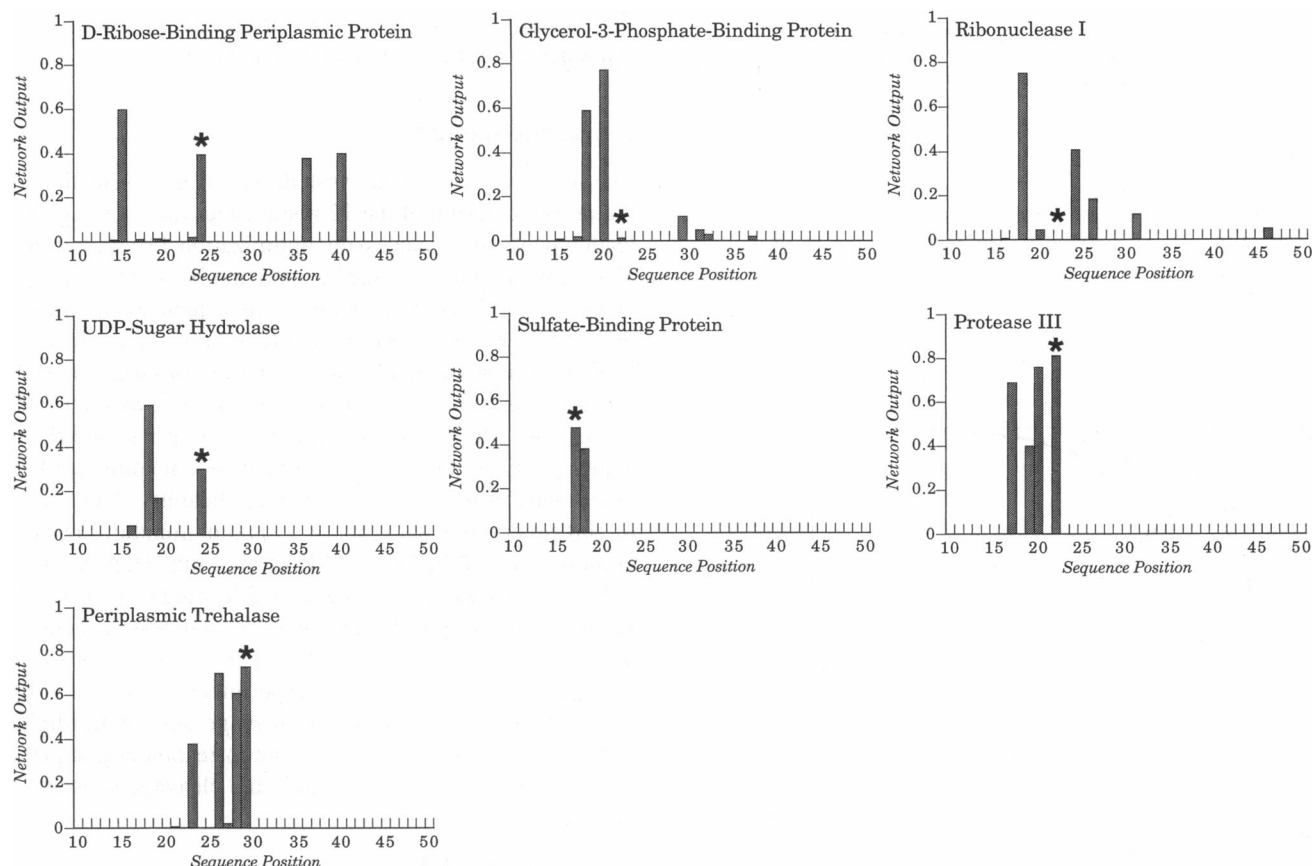 5 FFTRCAVSNI ↓ RG (0.29320418)

FIGURE 6   Prediction results in the NH$_2$-terminal parts of 7 independent test set sequences. The output values of the trained neural network system (3 combined filter modules) is plotted versus sequence position. The asterisks indicate the leader peptidase cleavage site as specified in the SwissProt Database.

None of the runs led to sequences of qualities near 1.0. Instead, qualities around 0.89 seem to represent good cleavage site sequences. This observation is the result of the architecture of the neural filter system, which multiplies the output values of the three single networks to obtain the final quality value.

The design protocols for the best results are given for every distance matrix in Fig. 7. The Context matrix for signal peptidase cleavage sites (Table 1), the Grantham matrix, the Myata matrix, and the Risler matrix led to a significant increase of quality with decreasing average mutability ("mean step"). Optimization converged by application of these amino acid distance metrics (Fig. 7, *A–D*). In contrast, a random distance matrix led to an increase of the mean mutability during optimization, and the sequence quality did not converge at all toward a maximum compared with the other distance metrics (Fig. 7 *E*). The random matrix failed completely in systematic sequence optimization, as expected.

An important conclusion can be drawn that rational design by SME is possible only if appropriate metrics for amino acid distances will be selected to define the appropriate residue selection function: F(d, R$_{old}$) (see *Materials and Methods*). The context-specific metric led to the best sequence within the 200 allowed optimization cycles, i.e., the sequence of highest quality according to the neural net-

work. The resulting leader peptidase cleavage site region FFFFGWYGWA ↓ RE (Quality = 0.8986; the arrow indicates the processing site) can be regarded as an idealized sequence representing the "optimal" amino acid motif. A systematic permutation of the 12-residue window resulted in the identical sequence supporting this SME result. No sequence with a higher quality according to the specified neural filter system is possible employing the common 20 amino acid residues. A different "optimal" sequence might be obtained using a different filter system.

Identification of important sequence positions

Adaptive control of mutabilities (step sizes of position-specific mutations) can be used for the identification of predominant positions in the designed sequence motif. Table 3 presents a representative protocol of the "simulated evolutionary history" for the best designed sequence (cp. Fig. 7 *A*): in the first generation already the final mutations occur at position −2 (Trp), at position −6 (Gly) in generation 10. The two amino acid residues are conserved during all following optimization cycles. These positions can therefore be assumed to be of major importance for signal peptide function. Gly at −6 and Trp at −2 are found in all designed sequences,
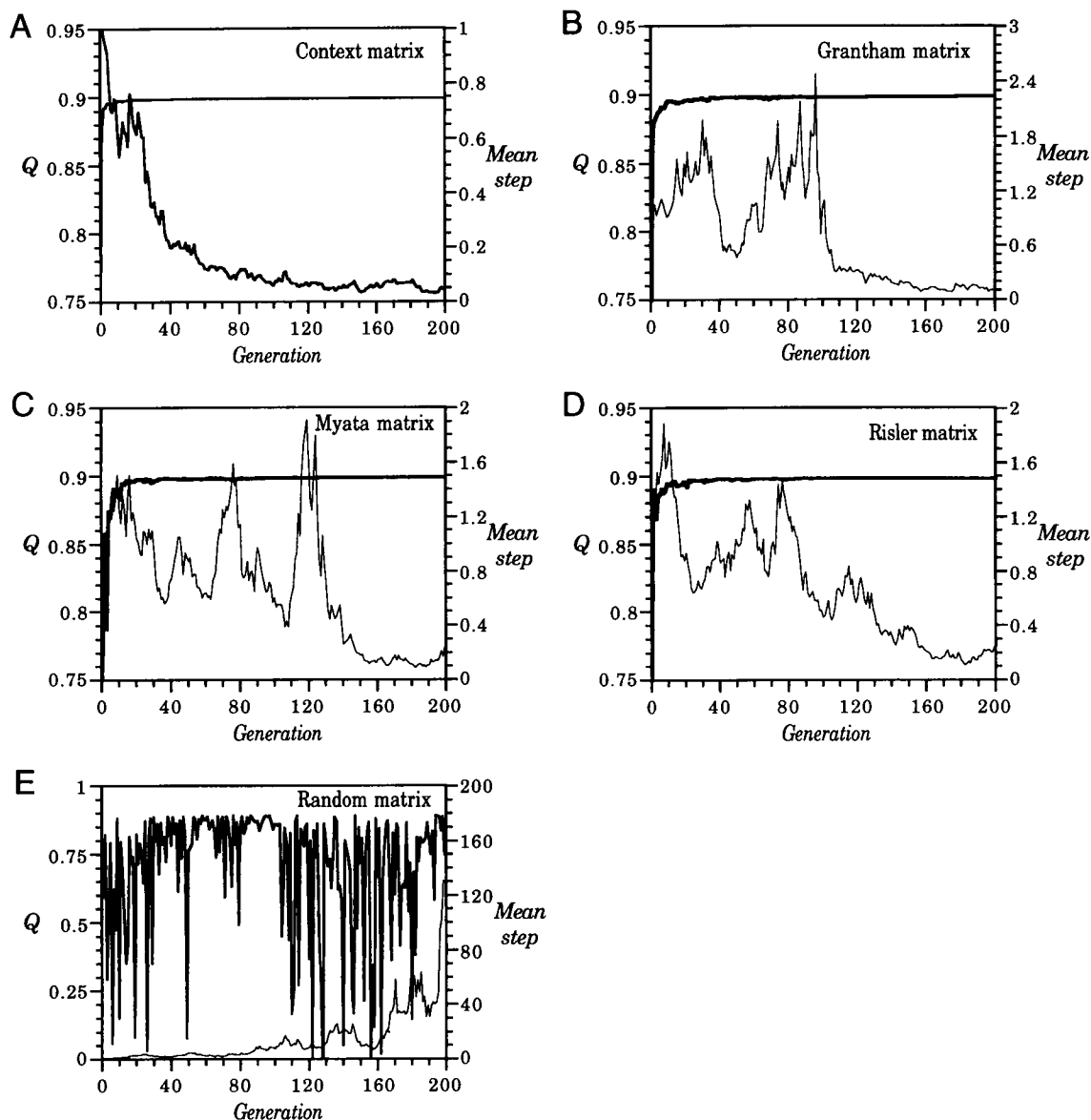
FIGURE 7   Protocols of the design experiments leading to the best sequences by application of five different amino acid distance matrices. Sequence qualities (Q) are drawn in thick lines, and thin lines indicate the average mutability ("Mean step").

regardless of the distance metric used for mutation. These amino acids seem to ideally represent appropriate residues.

In contrast, the residues forming the hydrophobic core of the signal peptide, which can be identified to begin at position −7 (Table 3), are conserved later during design. Although Phe is established at −10 to −7 in the final sequence, Ile, Met, or Trp seem to be suited equally well. Charged residues are found in the mature part of the cleavage site region (positions +1, +2) early during the SME already (Table 3). Surprisingly, a positive charge (Arg) and a negative charge (Glu) are located in adjacent positions in all best final sequences. It is likely that polar residues are required there but a net charge is unfavorable.

The "−3, −1 rule" (von Heijne, 1983; Perlman and Halvorson, 1983) describing a characteristic cleavage site motif in

E. coli precursors implies the positions −3 and −1 as being mainly occupied by small and hydrophobic (neutral) residues. This pattern is formed in two steps during generations 15 and 52 (Table 3). In the SME protocol, position −1 is significantly less variable than position −3. Indeed, at both positions small and hydrophobic residues are fixed early during simulated evolution. Position −4 does not seem to contribute much to the cleavage site signal inasmuch as no conservation occurred and many different residues are equally well suited (Table 3).

This one "sequence evolution" example has an initial quality value of 0.7565 for a "random" sequence already. This "random" sequence has been selected for the onset of optimization from a total of 500 "random" sequences produced in the 0th generation. The one of highest quality was used.

TABLE 3 The sequences and corresponding qualities of the SME-design run leading to the best cleavage-site region ("evolutionary history").

Cycle	Sequence	Quality
0	F I C L T M G Y I C ↓ R C	0.7565
1	F L T L P V S M W F ↓ E Q	0.8764
2	W I T F T I A M I ↓ E T	0.8920
3	W M I F S Y W G F ↓ R A	0.8929
4	F F I L S F W A I ↓ K T	0.8958
5	M I M M A F W S V ↓ R P	0.8965
6	V F L F P W W S M ↓ R S	0.8951
7	M F I F S F F A V ↓ R N	0.8955
8	M F L F A W F P C ↓ E E	0.8975
9	I W M L S W F G C ↓ Q D	0.8976
10	V F F I G L F G C ↓ R E	0.8971
11	V L I L Y I G C ↓ R E	0.8975
12	M I F M F I G A ↓ R D	0.8973
13	L M I I W M G A ↓ N D	0.8970
14	L L M F I G S ↓ Q D	0.8977
15	I I M F F G A ↓ R D	0.8978
16	M L M W F S ↓ R E	0.8977
17	L Y M W V A ↓ R E	0.8980
18	I F I W C C ↓ K E	0.8973
19	F F M F C A ↓ Q E	0.8980
20	W I F A A ↓ R E	0.8983
21	W M F A S ↓ K E	0.8981
22	W V F C A ↓ R D	0.8982
23	F M F C C ↓ D E	0.8981
24	F I I C C ↓ R D	0.8980
25	F F L A A ↓ D D	0.8977
26	F I M C A ↓ R D	0.8980
27	F F F C A ↓ D	0.8983
28	F F C A ↓ D	0.8983
29	F I C A ↓ D	0.8982
30	F F C A ↓ D	0.8983
31	F C A ↓ E	0.8984
32	F C A ↓ E	0.8984
33	W C A ↓ E	0.8984
34	W C A ↓ E	0.8984
35	W C A ↓ E	0.8984
36	W C A ↓ E	0.8984
37	W C A ↓ E	0.8984
38	W C A ↓ E	0.8984
39	W C A ↓ E	0.8984
40	W C A ↓ E	0.8984
41	W C A ↓ E	0.8984
42	W C A ↓ K	0.8984
43	W A A ↓ K	0.8984
44	W A A ↓ K	0.8984
45	W A A ↓ K	0.8984
46	W A A ↓ K	0.8984
47	W A A ↓ K	0.8984
48	W A A ↓ K	0.8984
49	W A A ↓ K	0.8984
50	W A A ↓ K	0.8984
51	F Y S ↓ E	0.8984
52	F F F F G W Y G W A ↓ R E	0.8986

The arrows indicate the leader peptidase cleavage site. Conserved residues in the sequences are not shown. At position −2 a Trp is conserved in the second generation already. After 52 generations (cycles) of the simulated evolution no further change occurred until the SME-design run was stopped after the 200th generation. The final sequence is given completely.

A rapid increase of quality during the first generations and very slow improvement later on is typical of an SME run.

Identification of putative local optima

The best designed sequences were used to define the starting points for views through the corresponding search space (se-quence space) (Fig. 8). The only amino acid distance metric not leading to putative local optima along the line of sight through the sequence-space is the Myata matrix (Myata et al., 1979) (Fig. 8 C). In all other cases the sequence space has several possible local optima (Fig. 8, A, B, D, and E). Using a distance matrix based on three-dimensional residue relationships a higher number of local optima are found compared with property-based distance metrics: the Risler-matrix (Risler et al., 1988) leads to a multimodal search space (Fig. 8 D).

DISCUSSION

Development of automatic extraction and classification routines is of great importance for structural and functional protein analysis (Eigen et al., 1988; Thornton, 1992). A new approach to this need and to the rational design of amino acid sequences is given by the SME technique. Several leader peptidase cleavage site sequences with high quality according to an artificial neural network have been designed de novo employing this method. Whether these sequences are biologically active is currently being tested in an in vivo expression and secretion system (P. Wrede, U. Hahn and G. Schneider, manuscript in preparation). The SME results, however, led to a deeper insight into the architecture of leader peptidase cleavage sites: two positions of predominant importance were identified (positions −2 and −6). At −6 the amino acid glycine was selected as being ideal. Similar observations were made by in vivo studies of designed prokaryotic cleavage sites (Laforet and Kendall, 1991). Position −2 needs a big residue, such as tryptophan, to allow the "−3, −1 rule" to develop (Table 3). These findings are also supported by other theoretical considerations by us (Schneider and Wrede, 1993b; Schneider et al., 1993).

The failure of any successful design using a random amino acid distance matrix (Figs. 7 E and 8 E) clearly demonstrates that amino acid distances are important parameters for simulating sequence evolution. Selecting an appropriate matrix is a crucial step for SME design success, because both the evolutionary optimization procedure employed and the design of a certain sequence feature (e.g., leader peptidase cleavage sites) will be successful only if there is a causal connection between the mutation distance metric and the fitness function used. The Myata-matrix (Myata et al., 1979) did not lead to local optima (Fig. 8 C) and seems, therefore, to be well suited for the design of leader peptidase cleavage sites.

Surprisingly, the obtained designed sequences are "suboptimal" compared with those generated using the Context matrix. High quality sequences performing a special task might be desired, but such "ideal" sequences bear the danger of too much specialization for an organism. Having "suboptimal" sequences still allows a cell to fulfill the desired task (e.g., precursor cleavage), but it also allows it to adapt to new situations, which is more difficult with highly specialized sequences. Thus, it should not be surprising if the DNA triplet code can be shown to be an ideal general representation of amino acids for SME. First results using the distance ma-
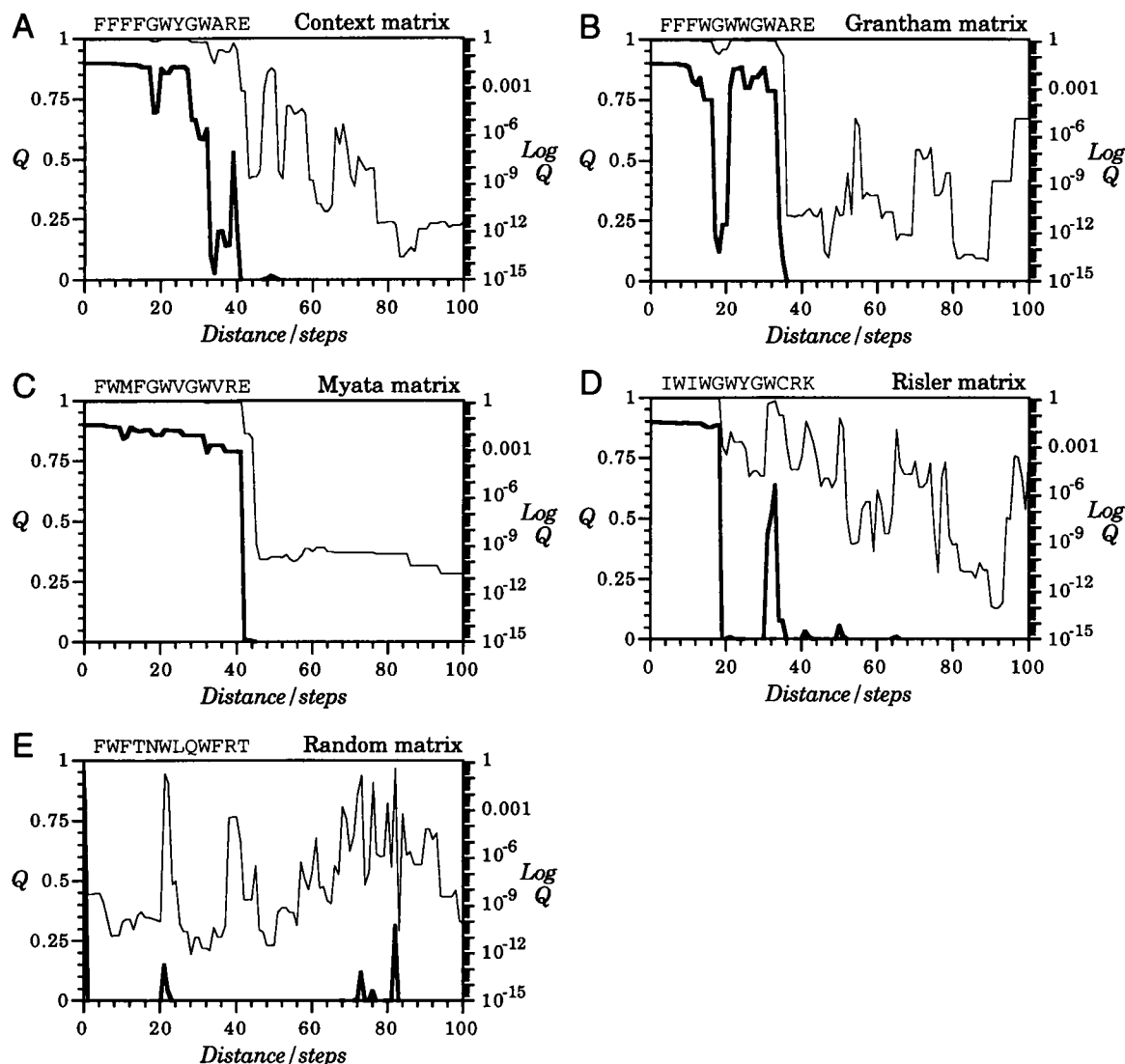
FIGURE 8   One-dimensional views of the search spaces using different amino acid distance matrices. The initial sequences (Distance = 0) are given above the plots. Thick lines: linear quality (Q); thin lines: logarithmic quality (Log Q).

trix of Feng et al. (1985) support this idea (Schneider and Wrede, submitted for publication).

Furthermore, the idealized cleavage site sequence FFFFGWYGWA ↓ RE contains both a Phe-rich hydrophobic region and a contrarily charged N-terminus of the mature protein (Arg-Glu) that might cause an impaired function compared to a wild type cleavage site in a biological assay. We think that the obtained sequence ideally represents a leader peptidase cleavage site motif. But no further functional feature, e.g., for the control of enzyme processing kinetics, has yet been taken into account during the design run. Only the biological test will help to evaluate the applicability of the theoretical SME approach.

Whether SME can be used for any design task will have to be proven in the future, too. A major limiting factor is the development of appropriate filter systems (Fig. 1). As soon as another reliable filter is available the SME technique will be applied and tested again. At present we develop neural

networks for the recognition and qualification of signal peptidase cleavage sites in mitochondrial protein precursors and for analysis and design of transmembrane segments of integral membrane proteins.

We are well aware that the number of sequences used for network training and testing is rather small (24 cleavage site sequences in total). Therefore, the networks are no ideal prediction systems still (Fig. 6), and the obtained idealized cleavage site sequence certainly does not represent the generally optimal sequence. Nonetheless, it could be shown that the search strategy employed is, in principle, able to find the ideal sequence with regard to the fitness function. A complete cross validation test, which is a useful method for a more reliable estimation of the neural network's generalization ability in future experiments, must be performed. A further disadvantage of the SME method is the limitation to the design of locally encoded sequence features. As long as artificial neural networks are restricted to analyzing only se-

quence sections using the "sliding window" technique no neural filter system can be constructed that is able to focus on globally encoded features. Despite these general limitations of the approach it could be clearly demonstrated that:

- using the SME technique amino acid sequences representing a desired feature can be generated de novo without knowledge of corresponding three-dimensional structures;
- a genetic algorithm like the $(1, \lambda)$ - Evolution Strategy is a useful method for finding the global optimum in a vast sequence space;
- artificial neural networks are well suited for estimating the quality of an amino acid sequence feature in terms of real values between 0 (the desired functional or structural feature is not represented by the sequence) and 1 (the feature is ideally represented);
- amino acid distances are context-dependent, i.e., that an amino acid residue may perform different tasks in different environments given by the protein structure;
- predominant sequence positions responsible for the manifestation of a desired sequence motif can easily be identified by looking at the development of the position-specific mutability values during the SME run (Table 3); and
- artificial neural networks can be used for the parallel design of protein sequence positions.
- Furthermore, a combination of the new method for visualizing the search space (Fig. 8) with evolutionary optimization algorithms might result in a useful strategy for sequence optimization leading to convergence at the global optimum in sequence space with high reliability.

REFERENCES

Dandekar, T., and P. Argos. 1992. Potential of genetic algorithms in protein folding and protein engineering. *Protein Eng.* 5:637–645.

Davidor, Y., and Schwefel, H. P. 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. Wiley & Sons U. S. A., New York. 183–203.

Eigen, M., R. Winkler-Oswatitsch, and A. Dress. 1988. Statistical geometry in sequence space: a method of quantitative comparative sequence analysis. *Proc. Natl. Acad. Sci. USA.* 85:5913–5917.

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.

Goldberg, D. E. 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Redwood City, CA. 412 pp.

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

Hirst, J. D., and M. J. E. Sternberg. 1992. Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry* 31:7211–7218.

Holland, J. H. 1992. Adaptation in natural and artificial systems. 2nd ed. MIT Press, Cambridge, MA. 211 pp.

Holley, H. L., and M. Karplus. 1991. Neural networks for protein structure prediction. *Methods Enzymol.* 202:204–224.

Hopp, T. P., and K. R. Woods. 1981. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA.* 78: 3824–3828.

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

Jones D. D. 1975. Amino acid properties and side chain orientation in proteins: a cross correlation approach. *J. Theor. Biol.* 50:167–183.

Koza, J. R. 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA. 819 pp.

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.

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

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

Rechenberg, I. 1973. Evolutionsstrategie - Optimierung Technischer Systeme nach Prinzipien der Biologischen Evolution. Frommann-Holzboog, Stuttgart.

Richardson, J. S., D. C. Richardson, N. B. Tweedy, et al. 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.

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

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

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

Schneider, G., and P. Wrede. 1993b. Signal analysis of protein targeting sequences. *Protein Sequences & Data Anal.* 5:227–236.

Schneider, G., S. Röhlk, and P. Wrede. 1993. Analysis of cleavage-site patterns in protein precursor sequences with a Perceptron-type neural network. *Biochem. Biophys. Res. Comm.* 194:951–959.

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

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

von Heijne, G., W. Wickner, and R. E. Dalbey. 1988. The cytoplasmic domain of *Escherichia coli* leader peptidase is a "translocation poison" sequence. *Proc. Natl. Acad. Sci USA.* 85:3363–3366.

Wells, J. A., and H. B. Lowman. 1992. Rapid evolution of peptide and protein binding properties in vitro. *Curr. Opin. Struct. Biol.* 2:597–604.

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