PROTEINS: Structure, Function, and Genetics 30:49–60 (1998)

Feature-Extraction From Endopeptidase Cleavage Sites in Mitochondrial Targeting Peptides

Gisbert Schneider,[1] Sara Sjöling,[2] Erik Wallin,[2] Paul Wrede,[1] Elzbieta Glaser,[2] and Gunnar von Heijne[2]*
[1]*Freie Universität Berlin, Universitätsklinikum Benjamin Franklin, Institut für Medizinische/Technische Physik und Lasermedizin, AG Molekulare Bioinformatik, Berlin, Germany*
[2]*Department of Biochemistry, Stockholm University, Stockholm, Sweden*

ABSTRACT **Cleavage sites in nuclear-encoded mitochondrial protein targeting peptides (mTPs) from mammals, yeast, and plants have been analysed for characteristic physicochemical features using statistical methods, perceptrons, multilayer neural networks, and self-organizing feature maps. Three different sequence motifs were found, revealing loosely defined arginine motifs with Arg in positions −10, −3, and −2. A self-organizing feature map was able to cluster these three types of endopeptidase target sites but did not identify any species-specific characteristics in mTPs. Neural networks were used to define local sequence features around precursor cleavage sites. Proteins 30:49–60, 1998.**

Key words: kohonen network; mitochondrial processing peptidase (MPP); mitochondrial intermediate peptidase (MIP); neural network; protein import; sequence motif; mitochondrial targeting

INTRODUCTION

Most mitochondrial proteins are encoded by the nuclear genome.[1,2] After ribosomal synthesis these proteins are imported as precursors from the cytosol into mitochondria. The appropriate targeting signal directing a newly synthesized protein to a mitochondrion is encoded by an N-terminal elongation of the protein, the *prepeptide* or *mitochondrial targeting peptide* (mTP).[3–8] The targeting signal is cleaved off by mitochondrial endopeptidases immediately after import into the organelle.

At least two different signal peptidases exist in mitochondria. The mitochondrial processing peptidase (MPP) in yeast and mammals is located in the mitochondrial matrix. In *Neurospora crassa*, MPP is partially associated with the mitochondrial membrane, and the beta-subunit is identical to one of the subunits of the bc_1 complex of the respiratory chain.[9] In plants, MPP is entirely associated with the inner membrane,[10] and both the alpha and beta subunits are integrated into the bc_1 complex.[11–13] The second precursor-processing enzyme in the matrix is the mitochondrial intermediate peptidase (MIP), a thiol-dependent metallopeptidase.[14] So far, MIP activity has been found in several mammals, yeast, and *N. crassa*, but nothing is known about the occurrence of MIP in plants.

MPP is involved in the processing of most imported precursors, whereas MIP removes an additional stretch of eight residues from some of the precursor proteins subsequent to MPP cleavage.[15] Both MPP and MIP recognize a large variety of target sites with rather degenerate recognition patterns in the precursor sequences. It has been demonstrated that the number of arginines present in the targeting peptide influences binding to MPP.[16] Synthetic peptides with a greater number of arginines were shown to be better inhibitors of MPP than those with fewer positively charged residues, and deletion of single arginines at certain positions prevented processing. This observation in combination with NMR and circular dicroism (CD) studies of mTPs led to the suggestion of a higher-order recognition signal than just a linear sequence motif.[17–20]

Artificial neural networks are able to automatically extract characteristic features from sets of amino acid sequences sharing a certain structure or biological function.[21,22] These systems have some attractive properties which are useful for amino acid sequence analysis, for example, they process sequence information in an inherently parallel way, and they are suited for feature extraction from noisy data.[23] Here, two different types of networks have been employed for cleavage site sequence analysis[24]: Kohonen networks (self-organizing feature maps) performing an unsupervised classification of amino acid sequences, and multilayered feedforward networks trained in a supervised manner on the prediction of cleavage site positions in precursor sequences

Contract grant sponsor: Fonds der Chemischen Industrie; Contract grant sponsor: Boehringer Ingelheim Fonds; Contract grant sponsor: Swedish Natural Sciences Research Council; Contract grant sponsor: BMBF DETHEMO Project
Dr. Schneider is currently at F. Hoffmann-La Roche Ltd., PRPI-S, Bioinformatics, Bldg. 65/318, CH-4070 Basel, Switzerland
*Correspondence to: Gunnar von Heijne, Department of Biochemistry, Stockholm University, S-10691 Stockholm, Sweden. E-mail: gunnar@biokemi.su.se
Received 24 April 1997; Accepted 18 August 1997

('prediction networks'). In addition, we have performed secondary structure predictions to investigate local helix formation in mitochondrial targeting sequences, since mTPs have been shown to have a propensity to adopt amphiphilic α-helical structure by NMR experiments and theoretical considerations.[17,20,25–28] Our aims were to get a more detailed idea of what the structural MPP and MIP target site features are and to search for possible differences in cleavage site patterns between different groups of organisms. Identification of characteristic target site features in the sequences of mitochondrial precursor proteins may suggest models of enzyme action and could become useful in genome analysis.

MATERIALS AND METHODS

Protein Sequence Data

The OWL Data Base Rel. 26.0[29] was searched for mitochondrial protein precursors from mammals, yeast, *N. crassa*, and plants (search terms: "mito" && "precurs"). Only entries with cleavage-sites not denoted as "potential," "putative," or "by similarity" were considered. The resulting collection of precursor proteins contained over 700 sequences. From this set, entries were selected which had experimentally confirmed cleavage-sites, that is, confirmed N termini of the mature forms (by microsequencing) and known targeting peptide sequence. For this the original literature was checked. Homologous sequences were eliminated by all-against-all pairwise alignment with FASTA, version 1.7a3[30] (with PAM250 matrix, ktup = 2), and subsequent comparison to a structure/similarity threshold θ[31]:

$$\text{if } (l > 80) \text{ then } \theta = 25$$

$$\text{else if } (l < 10) \text{ then } \theta = 110$$

$$\text{else } \theta = 290.15 * \exp(-0.56 * \ln(l))$$

where l is the length of the alignment (overlap), and θ is the similarity threshold. Pairs of sequences with an alignment score above the threshold were considered as sequence homologues. As a result, groups of similar sequences were obtained. One member from each group was selected randomly to be included in the final dataset. The numbers of precursor sequences before ("raw data") and after elimination of homologues ("final data") are given in Table I. In total, 125 cleavage site regions from 122 mitochondrial protein precursors were used in this work (122 confirmed MPP and 3 confirmed MIP sites). Sequence windows encompassing positions [−15, +10] and [−10, +5] were analyzed. Residue positions in the targeting peptide were assigned negative numbers with the cleavage site located between positions −1 and +1.

The cleavage sites were classified according to the location of arginine residues in the targeting peptide portion. "R-10," "R-3," and "R-2" examples contain an arginine in either of the positions −10, −3, or −2. Cleavage sites with multiple arginines are denoted, for example, by "R-10/−2," if two arginines are present, one in −10 and the second in −2.

TABLE I. Numbers of Nuclear-Encoded Precursor Sequences of Mitochondrial Proteins Before (Raw Data) and After (Final Data) Elimination of Sequence Homologues

	Raw data	Final data
Mammalia	85	61
Yeast	42	31
N. crassa	16	16
Plants	18	14

The OWL-identifiers (SwissProt[32] or PIR format[33]) of the proteins included in the final set are as follows:

Mammalian proteins (N = 61): AATM_MOUSE ('R-2'), ACDL_RAT, ACDM_RAT, ACON_PIG ('R-10'), ADX1_BOVIN ('R-10/-3'), ATP0_BOVIN ('R-10/-2'), ATPB_HUMAN, ATPD_BOVIN ('R-3'), ATPG_BOVIN ('R-3'), ATPL_HUMAN ('R-10'), ATPM_BOVIN ('R-10'), ATPO_BOVIN ('R-10/-2'), ATPR_BOVIN ('R-10'), BDH_RAT ('R-3'), CAH5_HUMAN ('R-10/-2'), CAH5_MOUSE ('R-2'), CISY_PIG, COX4_BOVIN ('R-10'), COXA_HUMAN ('R-3'), COXJ_BOVIN ('R-10/-3/-2'), COXK_BOVIN ('R-10/-3'), COXM_BOVIN ('R-2'), COXO_BOVIN ('R-10/-2'), CP27_RAT, CPM1_BOVIN, CPN2_RAT ('R-3'), CPSM_RAT ('R-2'), CPT2_RAT ('R-3'), D3D2_RAT ('R-10/-2'), DHAM_HUMAN ('R-3'), DHAM_RAT ('R-10/-3'), DLDH_HUMAN ('R-3'), FUMH_RAT ('R-2'), GCSH_BOVIN ('R-10'), GPDM-_RAT, HEM6_HUMAN, HUMOTC ('R-10'), IATP-_BOVIN ('R-3'), IVD_HUMAN ('R-2'), JX0071 ('R-3'), KCRS_HUMAN, MDHM_MOUSE ('R-10/-3/-2'), MPCP_BOVIN, MTDC_MOUSE ('R-2'), MUTA_HUMAN ('R-2'), NIAM_BOVIN ('R-10/-2'), NNTM_BOVIN ('R-10'), NUAM_BOVIN ('R-2'), NUBM_BOVIN ('R-2'), NUCG_BOVIN ('R-10'), NUGM_BOVIN ('R-2'), NUHM_BOVIN ('R-10'), ODBB_BOVIN ('R-2/-3'), ODP2_HUMAN ('R-3'), ODPA_HUMAN ('R-2'), P60_HUMAN ('R-3'), PCCA_RAT ('R-10'), PMIP-_RAT ('R-2'), SODM_HUMAN ('R-2'), SUCA_RAT ('R-2'), THIL_RAT ('R-3').

Yeast proteins (N = 31): ATPA_YEAST, ATPB-_YEAST, ATPD_YEAST ('R-2'), ATPG_YEAST, COX4_YEAST ('R-10'), COX6_YEAST ('R-3'), COX8_YEAST ('R-2'), COXA_YEAST ('R-2'), CYPC_YEAST ('R-10/-2'), DLDH_YEAST ('R-10/-3'), HEMZ_YEAST ('R-10'), HS77_YEAST ('R-2'), IATP_YEAST ('R-3'), MDHM_YEAST ('R-10'), MPP2_YEAST, NDI1_YEAST, ODP2_YEAST ('R-10/-3'), ODPX_YEAST, RIM1_YEAST ('R-10'), RM02_YEAST ('R-2'), RM09_YEAST ('R-2'), RM13_YEAST, RM27_YEAST, RM32_YEAST,

RM36_YEAST, RM37_YEAST, RM41_YEAST, RT28_YEAST ('R-10'), SDH4_YEAST ('R-10'), SODM_YEAST ('R-2/-3'), SYH_YEAST.

***N. crassa* proteins** (N = 16): ACP_NEUCR ('R-3'), ATP9_NEUCR ('R-2'), COX4_NEUCR ('R-10/-2'), COX5_NEUCR ('R-2'), CY1_NEUCR, MPP1_NEUCR ('R-2'), MPP2_NEUCR ('R-2/-3'), NUAM_NEUCR ('R-10/-2'), NUBM_NEUCR ('R-10/-3'), NUEM_NEUCR ('R-2/-3'), NUFM_NEUCR ('R-3'), NUYM_NEUCR ('R-2/-3'), ODP2_NEUCR ('R-3'), SYLM_NEUCR ('R-2'), SYYM_NEUCR ('R-2/-3'), UCRI_NEUCR ('R-10').

Plant proteins (N = 14): A-MPP2_SOLTU ('R-10/-2'), ADT1_MAIZE, ATP2_NICPL, ATP3_IPOBA ('R-2'), ATPO_IPOBA, B-MPP1_SOLTU ('R-2'), B-MPP2_SOLTU ('R-3'), CY11_SOLTU, GLYM_PEA, MAON_SOLTU ('R-2/-3'), MDHM_CITVU ('R-3'), SODM_MAIZE ('R-2'), SODM_NICPL ('R-2'), UCRI_SOLTU ('R-3').

The above mTP sequence collection can be retrieved by anonymous FTP from ftp.biokemi.su.se/pub/gisbert (filename MTP.TXT).

In addition to mitochondrial protein precursors, a selection of 269 nonredundant eukaryotic cytoplasmic sequences (10,760 residues) collected by Nielsen and colleagues (unpublished data) was used in our analysis.

Statistical Analysis of Amino Acid Compositions

The statistical significance of differences in amino acid composition between the mTP collection (4,416 residues) and Nielsen's collection of eukaryotic cytoplasmic sequences was assessed by chi-square analysis. For each kind of residue, the number of occurrences in the two sequence collections were compared, and the statistical significance was calculated for one degree of freedom. To assess the significance of arginine enrichment in positions −2, −3, and −10 in the mTP collection, the number of arginine residues in each of these positions was compared to the number expected from the average frequency of arginine in the collection (12.5%) using chi-square analysis. Position-specific amino acid frequencies were compared between three collections of mTPs characterized by the presence of arginine in position −2, −3, or −10 (sequences with arginine in two or all three of these positions were included in all the relevant groups) by averaging the three pairwise linear correlation coefficients between the overall amino acid composition in position i in the first group, position j in the second, and position k in the third group.

Self-Organizing Feature Map

A Kohonen network[24,34,35] was employed to automatically classify the precursor cleavage sites according to their similarity. 124 sequence windows encompassing positions −10 to +5 relative to the processing site were encoded in terms of hydrophobicity[36] and volume[37] for each residue. This resulted in 30-dimensional numerical descriptions of the sequence data. The physicochemical property scales were normalized in [−1,1]. Here "similarity" between patterns means the euclidian distance in the 30-dimensional space. This data description was used to keep the dimension of the pattern space as low as possible. If the frequently applied unary encoding scheme ("distributed encoding") had been used a $15 \times 21 = 315$-dimensional space would have been spanned. The number of training patterns available was too small to ensure reliable network optimization for this large input space. The network used was of feedforward type with an input layer fed with the 30-dimensional patterns and a planar output layer of neurons.

In a Kohonen network each neuron is characterized by its position $\vec{r}$ and its weight vector $\vec{\mathbf{w}}$. The weight vectors have the same dimension as the training patterns. Starting from a random initialization the network was trained to perform a nonlinear mapping of the sequence space onto a two-dimensional feature map. During the training process patterns were selected randomly and presented to the network. The winner neuron was determined by finding the weight vector $\vec{\mathbf{w}}^{r_{win}}$ closest to the training pattern $\vec{\mathbf{v}}$ presented. After pattern presentation the weight vectors were adapted:

$$\vec{\mathbf{w}}^{r_k}(t+1) = \vec{\mathbf{w}}^{r_k}(t) + \left(\vec{\mathbf{v}}^{r} - \vec{\mathbf{w}}^{r_k}(t)\right)\eta e^{-1/2r^2(\vec{\mathbf{r}}^{r_k} - \vec{\mathbf{r}}^{r_{win}})^2}.$$

The learning rate η and the update radius r were diminished during the training process according to

$$\eta(t) = \frac{\eta_0}{1 + (t/\tau)} \qquad \text{and} \qquad r(t) = \frac{r_0}{1 + (t/\tau)}$$

with τ being proportional to the number of training-patterns, $\eta_0 = 0.7$, and $r_0 = 3.5$. Each pattern was presented approximately 10 times.

Neural Network Architecture and Supervised Training

Simple perceptrons[38] and three-layered feedforward networks[39] were trained on the classification of cleavage sites (positive examples) and noncleavage sites (negative examples) by a (1,100) evolutionary strategy (notation according to Rechenberg[40]). In contrast to the commonly used "backpropagation of error" gradient descent technique[39] network training by an evolutionary strategy involves an adaptive random search. The training method and network

architectures were identical to previous studies.[22,23,41,42] Pattern classification by neural networks was performed for six different datasets, consisting of cleavage site sequences with an arginine in either of the positions −2, −3, or −10 (positive examples), and negative examples stemming from either the targeting peptides or from cytoplasmic sequences. Sequence windows encompassing positions −10 to +5 were used as cleavage site examples (positive examples). If the negative examples were taken from the mTPs, all other possible sequence windows covering 15 residues in the region [−15, +10] were treated as noncleavage sites. Negative examples from cytoplasmic proteins were taken from the 40 most N-terminal residue positions. All sequence windows were encoded by two physicochemical values per residue, hydrophobicity[36] and volume.[37] These scales were found to be useful for feature extraction from MPP target sites in previous experiments.[43] The properties were normalized in [−1,1].

For network training the data were randomly split into a training set (70%) and a test set (30%). Network weights were initialized in [−1,1]. Each experiment was repeated 10 times with random weight initialization for statistics. Ten times cross-validation was performed to estimate the average generalization ability of the features extracted. Cleavage site examples were assigned the target value 1, noncleavage sites the value 0. The mean-square-error (MSE) calculated from the differences between the output values of the network and the target values served as the quality function (error function) for network training:

$$\mathrm{MSE} = \frac{1}{N}\sum_{p=1}^{N}(\mathrm{net}_p - \mathrm{target}_p)^2$$

where N is the number of patterns in the data set, net_p is the network output for the pattern p, and target_p is its target value.

A sigmoidal function was used as the neuron transfer function:

$$\mathrm{Sigm}(x) = \frac{1}{1 + e^{-x}}$$

where x is the neuron input. With this neuron transfer function the overall transformation function of a perceptron is

$$f\left(\vec{\xi}\right) = \mathrm{Sigm}\left(\sum_{i=1}^{30}\xi_i w_i - \vartheta\right)$$

with ϑ, output neuron bias, $\vec{\mathbf{w}}$, weight vector, and $\vec{\xi}$, training pattern. The three-layered networks had five hidden neurons giving the nonlinear overall network transformation:

$$f\left(\vec{\xi}\right) = \mathrm{Sigm}\left[\sum_{j=1}^{5}\hat{w}_j\,\mathrm{Sigm}\left(\sum_{i=1}^{30} w_i^j\xi_i - \hat{\vartheta}_j\right) - \vartheta\right]$$

with $\vec{w}$ = weights connecting the input layer to the hidden layer (input-to-hidden weights), $\vec{\hat{w}}$ = weights connecting the hidden layer with the single output neuron (hidden-to-output weights), ϑ = output neuron bias, and $\vec{\hat{\vartheta}}$ = bias values of the hidden layer neurons. The real-coded output values were converted to binary values using a step function with the threshold value 0.5.

To determine the prediction accuracy of a network with the training and test data the correlation coefficient according to Matthews[44] was calculated:

$$cc = \frac{(PN) - (UO)}{\sqrt{(N + U)(N + O)(P + U)(P + O)}}$$

where P = number of correctly predicted cleavage sites, N = number of correctly predicted noncleavage sites, U = number of missed cleavage sites (underprediction), and O = number of additional cleavage sites predicted (overprediction). The values of cc range between −1 and 1. A value of 1 indicates perfect prediction.

The features extracted by the perceptrons were analyzed by plots similar to Hinton diagrams, which are graphical representations of the weight vector of a trained network.[39,45] For this purpose the weights were normalized in [−1,1].

RESULTS AND DISCUSSION

Mitochondrial targeting peptides were analyzed for characteristic sequence features by means of simple statistical analysis of residue frequencies, classification by a self-organizing feature map (Kohonen network), and by multilayer neural network systems. The networks were also tested for their applicability for MPP and MIP cleavage site prediction. Finally, secondary structure predictions were performed to get an idea of potential structural mTP features.

Analysis of Mean Length and Amino Acid Composition of Mitochondrial Targeting Sequences

The average length of the mitochondrial targeting peptides was determined separately for mammalian, yeast, plant, and *N. crassa* precursor proteins (Table II). The yeast mTPs are most variable in length, varying between nine residues in yeast mitochondrial leucyl-tRNA synthetase precursor (SYLM_YEAST) and 121 residues in yeast mitochondrial ribonuclease P precursor (RPM2_YEAST).

As judged by chi-square analysis (see Materials and Methods section), the full set of mTPs contains

TABLE II. Mean Length of mTPs (Number of Residues) and Standard Deviations

	Mean length	Min.	Max.
Mammalia	33.26 ± 11.33	16	68
Yeast	31.61 ± 22.39	9	121
N. crassa	32.94 ± 13.59	8	70
Plants	39.71 ± 18.77	18	77

min., minimal length observed; max., maximal length observed.

significantly fewer aspartates, glutamates, lysines, and isoleucines than cytosolic proteins, whereas it is significantly enriched in positively charged arginine residues as well as alanines, leucines, and serines ($P < 10^{-4}$; data not shown). This substantiates previous findings based on much smaller data sets.[18,46,47] On a species-specific basis, alanine residues seem to be over-represented in mammalian, plant, and especially in *N. crassa* mTPs, but not in yeast mTPs (Fig. 1). The content of Ser residues is higher in plant mTPs (16.5%) than in any other species (mammalian 9%; yeast 12.7%; *N. crassa* 8.9%).

The distribution of Arg in the [−15,+10] cleavage site regions was analyzed in more detail since positively charged residues were reported to be an important feature of mTPs and drastically affect MPP-binding.[16] In the full mTP collection, arginines are significantly enriched in positions −2, −3, and −10 ($P < 0.005$ by chi-square analysis; data not shown). This enrichment is seen in mTPs from all groups of organisms (Fig. 2), except that there are rather few arginines in position −10 in the plant sequences. The positions downstream of the cleavage site have a rather low content of arginines.

As shown in the next section, mTPs can be classified as belonging to one of three groups characterized, respectively, by an arginine in position −2, −3, or −10. In an attempt to detect further residue patterns around the MPP cleavage site, we looked for correlations in amino acid composition between different positions in the different classes. Two positions relative to the diagnostic arginine were found to have similar amino acid compositions in all three classes (mean pairwise linear correlation coefficients = 0.8 in both cases), namely those corresponding to positions −3 and +2 in the 'R-2' class (i.e., positions −4 and +1 in the 'R-3' class and positions −11 and −7 in the 'R-10' class). Position −3 in the 'R-2' class and the equivalent positions in the other classes are dominated by valine (21%), alanine (16%), and serine (12%), while position +2 and its equivalents are dominated by serine (35%) and alanine (20%). On the other hand, there is also a clear distinction between the three classes, in that position +1 in the 'R-2' class contains many different residues, while the corresponding positions −1 in the 'R-3' class and −8 in the R-10 class are dominated, respectively, by tyrosine (42%) and phenylalanine (43%) (Fig. 3).

Feature Extraction From Cleavage Sites by a Kohonen Network

The [−10,+5] cleavage sites served as the basis for sequence classification by a self-organizing feature map (Kohonen network). Each residue position was encoded by the respective hydrophobicity[36] and volume[37] values, resulting in a 30-dimensional input space (4 × 4). Kohonen networks were trained with random selections of only 75% of the data. The resulting clustering of data points was very similar in each training round. Kohonen networks were also trained with sequence data encompassing residue positions −9 to +5, and positions −15 to +10. The classification results were comparable to those obtained with the [−10,+5] sequence windows (data not shown).

To analyze the sequences clustered by each neuron in more detail we investigated them for common motifs. Each group of sequences shown in Figure 4 was represented by one neuron of the network, that is, the members of each group share a certain physicochemical motif. The clustering reveals some obvious residue patterns: the group of neurons in the lower left corner of the map have the 'R-2' consensus motif, and the group of neurons in the upper left share the 'R-3' consensus motif (terminology according to (Gavel and von Heijne[48]). 'R-10' sites (representing potential MIP target sites) are clustered in the upper right part of the neuron map. The majority of the sequences in the upper right of the feature map also have a large hydrophobic residue in position −8. Cleavage sites without an arginine in −2, −3, or −10 are mainly found in neuron (4/1). The topology of the feature map suggests that the sequences in neuron (4/1) are similar to 'R-2' or 'R-10' sites since no 'R-3' example is clustered in any of the adjacent neurons. Based on the classification of the Kohonen network and the statistical analysis reported in the previous section, the patterns around the 'R-2', 'R-3', and 'R-10' MPP cleavage sites can be specified as follows:

'R-2': (V|A|S)-R-X↓X-(S|A)

'R-3': (V|A|S)-R-X-(Y|L)↓(S|A)

'R-10': (V|A|S)-R-X↓(F|L)-(S|A)

where X denotes an arbitrary amino acid residue and the arrow specifies the endopeptidase processing site. A strong similarity between the three patterns is observed if the 'R-3' pattern is shifted by one position to the right. It is possible that some or all 'R-3' processing sites do not result from direct MPP cleavage, rather they might be formed by 'R-2' MPP cleavage and subsequent removal of an additional hydrophobic residue by an unknown proteolytic activity.

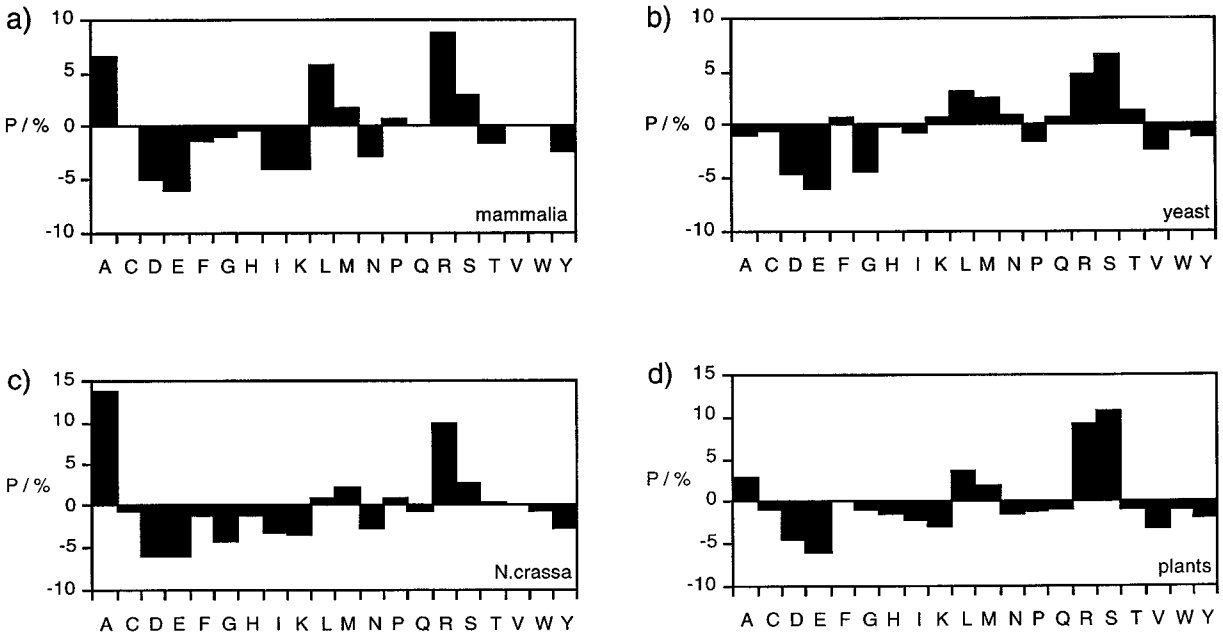

Fig. 1. Difference in amino acid frequencies (in percent) between mitochondrial targeting peptides and cytosolic protein sequences. Positive values indicate overrepresentation in the mTPs.

The Kohonen feature map does not reveal a species-specific grouping of cleavage site examples (Fig. 4), except that only one plant sequence is found in the 'R-10' part of the map. Rather, the network has classified the data according to the cleavage site pattern across all organisms. Thus, the overall sequence requirements for MPP and MIP cleavage seem to be universal. This interpretation is, however, quite crude, and more subtle species-specific differences may exist. For this reason, Kohonen feature maps were trained on only the 'R-2', only the 'R-3', or only the 'R-10' sequences, but again no clustering according to species was observed (data not shown).

From this analysis, we conclude that three major classes of cleavage-site motifs, 'R-2', 'R-3', and 'R-10', exist in mTPs. The Kohonen networks were unable to extract significant differences between mTPs from the different taxonomic groups, suggesting that the features important for recognition and cleavage of mTPs have been conserved in evolution. The only qualification to these statements is that the existence of 'R-10' cleavage sites in plant mTPs is uncertain.

Supervised Training of Neural Networks for Prediction of Cleavage Sites

Fully connected feed-forward neural networks without hidden layer neurons (perceptron) and networks containing one hidden layer with five neurons were trained on the prediction of MPP (Arg in −3 and/or −2) and potential MIP (Arg in −10) cleavage sites, again using hydrophobicity and volume as the residue descriptors. Three sets of cleavage site data (positive examples) were prepared based on the results obtained from the self-organizing feature map: 'R-3' cleavage sites, 'R-2' cleavage sites, and 'R-10' cleavage sites (as detailed in Fig. 4). Networks were also trained using the full dataset. Not surprisingly, their prediction accuracy was worse than the accuracy of the networks trained with separate data sets (results not shown). Negative training patterns, that is, noncleavage sites, were taken from either N-terminal parts of cytoplasmic proteins (first 40 residues) or from sequence regions located around the cleavage sites (within a [−15,+10] window).

The results of neural network training with negative data stemming from cleavage site regions are given in Table III. Simple perceptrons lacking a hidden layer were already able to correctly classify nearly all training set examples, as indicated by correlation coefficients around 0.99 (Table III). The resulting MSE values are low without significant variation in different training runs. The extremely low standard deviation of 0.01 of the training correlation coefficient reflects reproducible convergence. Test set prediction of the perceptrons was best for 'R-3' examples (cc = 0.9), 'R-2' examples were nearly as reliably predicted (cc = 0.86), but the 'R-10' examples were comparably poorly recognized (cc = 0.67), indicating either larger diversity of the training patterns or that the perceptrons extracted an ambiguous sequence feature.

The networks containing hidden layer units perfectly reclassified the training data in every independent training run (Table III). The final MSE values

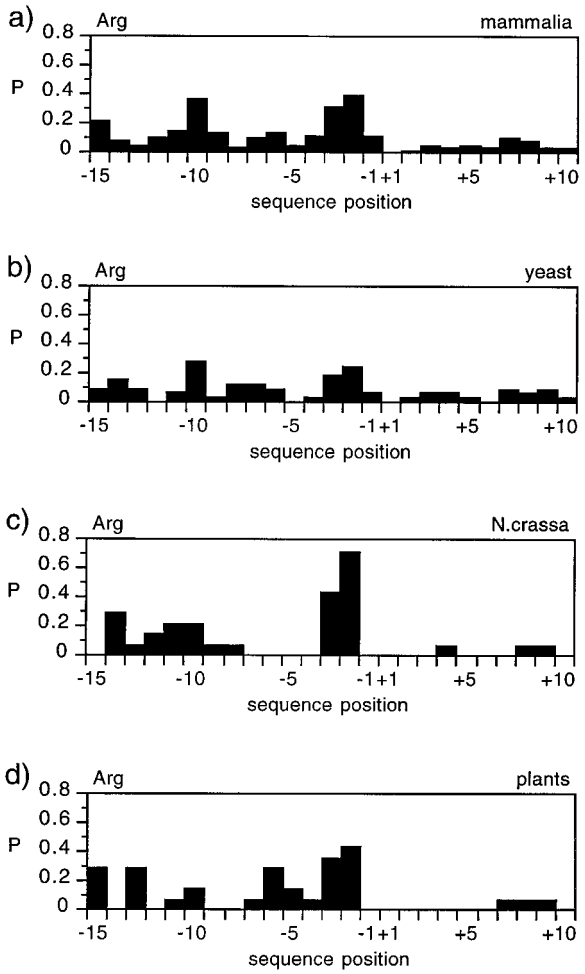

Fig. 2. Positional frequencies of arginine residues in the [−15,+10] cleavage site region of mitochondrial protein precursors.

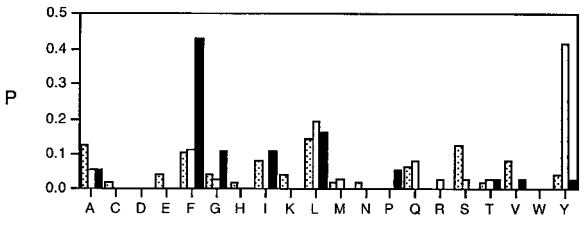

Fig. 3. Distributions of amino acid residues in set 'R-2' position +1 (*gray bars*), set 'R-3' position −1 (*white bars*), and set 'R-10' position −8 (*black bars*).

obtained were very low, and the 'R-3' network even gave a value very close to zero for the training data. These observations also indicate reliable convergence of the training process. The training results for the 'R-10' data are slightly better than those obtained with the perceptrons, that is, lower MSE values were obtained with the networks containing a hidden layer. Test set prediction accuracy (generalization) dropped for the 'R-3' and 'R-2' networks ($cc = 0.76$, and $cc = 0.73$), which might be explained by an "overlearning" effect, that is, the networks might have been too large for this task. However, test data prediction increased for the 'R-10' data compared to the perceptron results. An average correlation coefficient of 0.74 was calculated in this case from ten times cross-validation (Table III).

To get an idea of the physicochemical cleavage site features extracted by the networks, graphical representations of the perceptron weights were made (Fig. 5). In these plots each small square stands for one weight value, where light shading indicates a large value and dark shading means a low weight value. The extracted feature is represented by the entire plot, and it is risky to discuss the meaning of individual weights. The following interpretations, therefore, must be treated with care.

Residue positions −10, −9, and −8 were assigned extreme weight values by the 'R-10' network (Fig. 5a). A low hydrophobicity and large volume at −10 is in good accordance with the overrepresentation of arginines in this residue position. Position −9 reveals a preference for small residues, position −8 for large residues. The 'R-3' perceptron weights have extreme values at positions −10, −3, −1, and +1 (Fig. 5b). Again, a low hydrophobicity and large volume reflects the presence of arginines. The 'R-3' motif is compatible with a large hydrophobic residue at position −1, and small residues in +1, Figure 5b. The feature extracted by the 'R-2' perceptron mainly contains a nonhydrophobic residue in −9, a nonhydrophobic large residue (i.e., Arg) in −2, and a nonhydrophobic small residue in +2 (Fig. 5c). It must be stressed, however, that we cannot be sure that the mTP cleavage-site fragments used for network training were properly split up between the 'R-2', 'R-3', and 'R-10' groups, since this grouping was based only on the clustering in the Kohonen network.

Results of network training employing negative examples from cytoplasmic proteins are listed in Table IV. Weight optimization converged reliably as indicated by the low mse-values obtained from cross-validation. However, the correlation coefficients are significantly larger for reclassification than for generalization. The latter are quite small, ranging from 0.13 to 0.25. This means that the networks did not extract generalizing features that are useful for a distinction between actual mTP cleavage sites and similar patterns occurring in cytosolic sequences. Therefore, we did not analyse the extracted features in detail.

Prediction of Precursor Cleavage Sites by Neural Networks

The test set correlation coefficients suggest that the neural networks trained on the prediction of mTP cleavage sites are not accurate enough to be used as a reliable prediction method. This is also

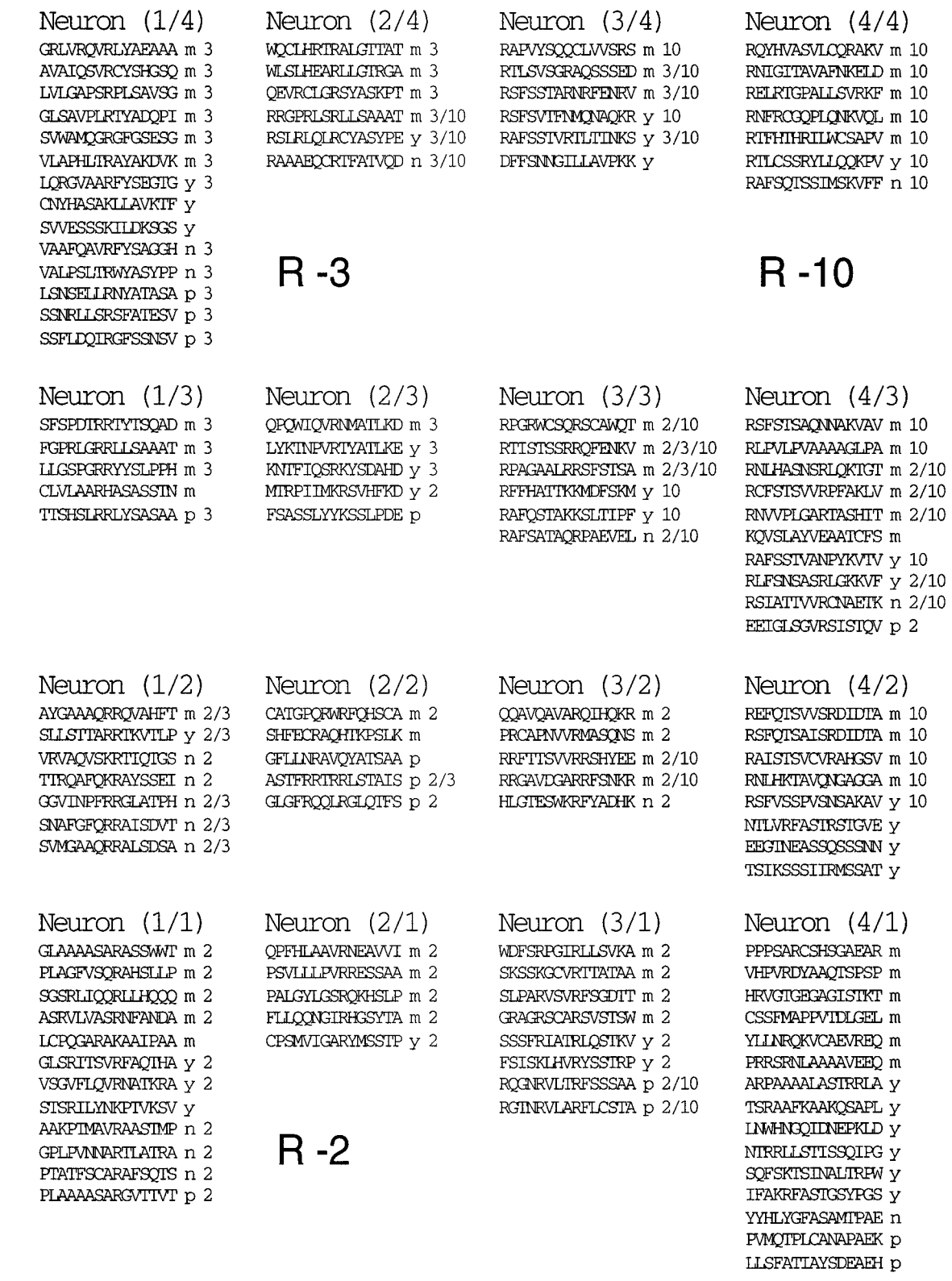

Fig. 4. Feature map formed by a trained (4×4) Kohonen network. The $[-10,+5]$ cleavage site sequences clustered by each neuron are shown. The predominant regions of the map are indicated by 'R-10' (upper right), 'R-3' (upper left), and 'R-2' (lower left). Beside each sequence its origin is denoted by *m* (mammal), *y* (yeast), *n* (*Neurospora*), and *p* (plant). Numbers indicate whether the sequence contains an arginine in either of the positions -10 ('10'), -3 ('3'), or -2 ('2').

TABLE III. Results of Supervised Neural Network Training on mTP Cleavage Sites, Negative Examples From the Cleavage Site Region

Network	$MSE_{train} \times 10^{-3}$	$MSE_{test} \times 10^{-3}$	cc_{train}	cc_{test}
R-3, perceptron	0.52 ± 0.16	20.37 ± 16.60	0.99 ± 0.01	0.90 ± 0.07
R-2, perceptron	2.42 ± 1.20	14.24 ± 2.57	0.99 ± 0.01	0.86 ± 0.03
R-10, perceptron	2.57 ± 1.58	30.70 ± 4.59	0.98 ± 0.01	0.67 ± 0.05
R-3, 5 hidden units	0.00 ± 0.0	49.60 ± 21.88	1.00 ± 0.0	0.76 ± 0.10
R-2, 5 hidden units	0.24 ± 0.49	29.66 ± 12.95	1.00 ± 0.01	0.73 ± 0.11
R-10, 5 hidden units	1.03 ± 1.99	25.30 ± 8.11	0.99 ± 0.02	0.74 ± 0.09

MSE, mean-square error of the network output; *cc*, Matthew's correlation coefficient.

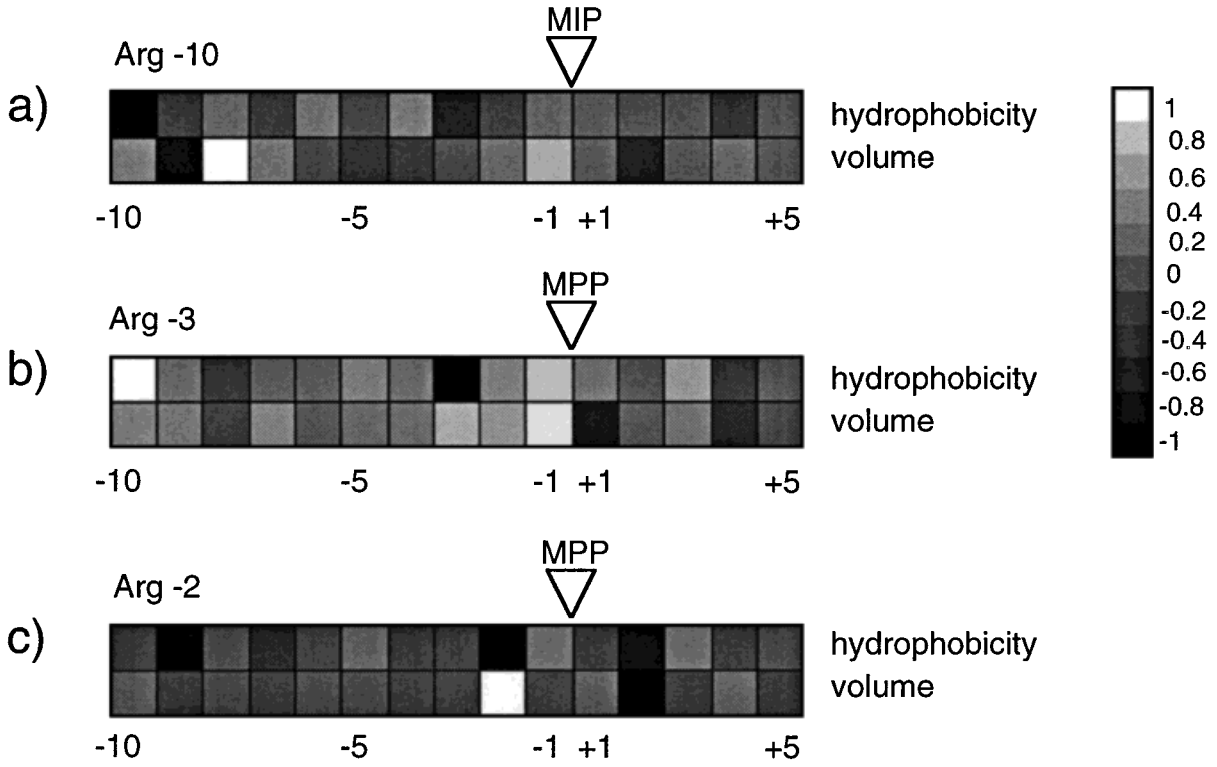

Fig. 5. Graphical representation of perceptron weights for different cleavage site motifs in mitochondrial protein precursors. The perceptrons were trained with cleavage site examples and noncleavage site examples stemming from regions around the actual cleavage sites. Sequence windows encompassing positions −10 to +5 around the processing site (white arrowhead) were analyzed. **a:** 'R-10' network weights. **b:** 'R-3' network weights. **c:** 'R-2' network weights. All weight values were normalized in [−1,1]. Extreme values indicate important sequence positions (light shading represents a large value for the parameter in question, dark shading a low value). The hydrophobicity scale of Engelman and coworkers[36] and the volume scale of Harpaz and coworkers[37] were used for feature extraction. Relative residue positions are given below each plot.

clear from a typical example of potential cleavage sites predicted in an mTP (mouse malate dehydrogenase) with experimentally confirmed two-step processing by MPP and MIP (Fig. 6). Besides the correct cleavage sites, several additional potential target sites for both MPP and MIP are predicted. Similar prediction patterns were obtained for other sequences (data not shown). In total, the 'R-2' network predicted 496 out of 8448 sites tested (6%) as potential MPP cleavage sites, and it correctly identified all of the 36 true cleavage sites. The corresponding figures for the 'R-3' and 'R-10' networks were 772 predicted out of 8448 tested sites (9%) with 27 true sites out of 37 (73%) correctly identified, and 664 predicted out of 8448 tested sites (7%) with 47 true sites out of 50 (94%) correctly identified.

We conclude that these networks on their own cannot be used to efficiently predict mTP cleavage sites, though they do indicate a rather limited number of possible sites.

Secondary Structure Prediction of Mitochondrial Targeting Peptides

The secondary structure of mTPs have been found to be important both for import and cleavage.[20] Two secondary structure prediction methods were used to

TABLE IV. Results of Supervised Neural Network Training on mTP Cleavage Sites, Negative Examples from Cytoplasmic Sequences

Network	$MSE_{train} \times 10^{-3}$	$MSE_{test} \times 10^{-3}$	cc_{train}	cc_{test}
R-3, perceptron	1.107 ± 0.446	2.605 ± 0.887	0.758 ± 0.265	0.235 ± 0.181
R-2, perceptron	4.341 ± 1.029	6.314 ± 1.375	0.543 ± 0.133	0.251 ± 0.103
R-10, perceptron	2.555 ± 0.522	4.312 ± 0.777	0.174 ± 0.129	0.174 ± 0.159
R-3, 5 hidden units	0.871 ± 0.452	3.662 ± 0.921	0.808 ± 0.139	0.135 ± 0.122
R-2, 5 hidden units	3.014 ± 0.585	7.356 ± 1.408	0.729 ± 0.045	0.167 ± 0.11
R-10, 5 hidden units	1.452 ± 0.662	5.078 ± 1.649	0.775 ± 0.11	0.196 ± 0.171

MSE, mean-square error of the network output; *cc*, Matthew's correlation coefficient.

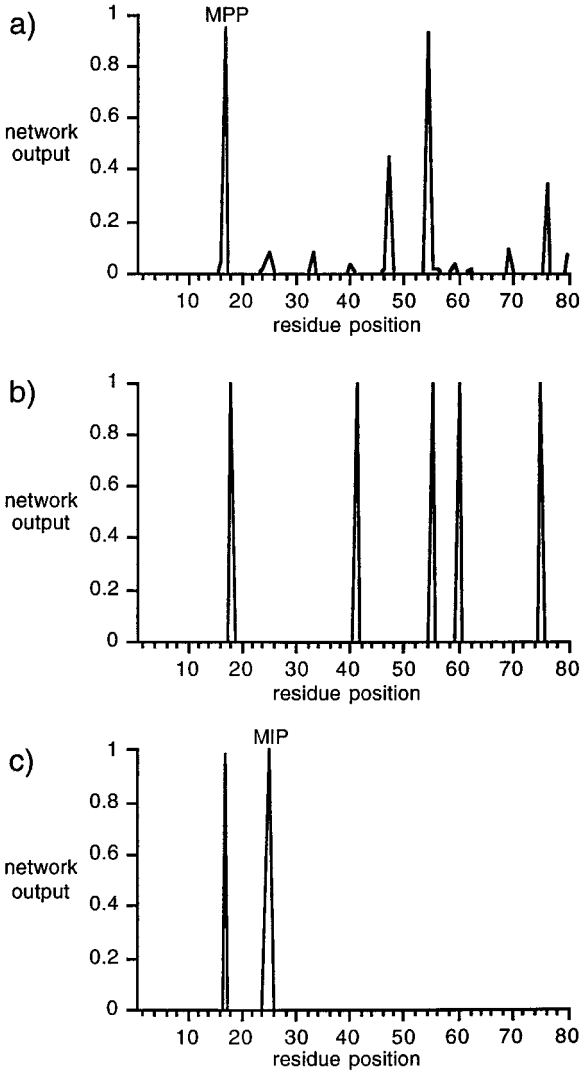

Fig. 6. Neural network predictions of MPP and MIP cleavage sites in the N-terminal part of the mouse malate dehydrogenase precursor (MDHM). The experimentally determined MPP and MIP processing sites are marked. **a:** 'R-2' network output. **b:** 'R-3' network output. **c:** 'R-10' network output.

get an idea of potential local structural preferences in mTPs and the MPP and MIP cleavage site regions: the PhD method of Rost and Sander[49,50] and the AGADIR program of Munoz and Serrano.[51,52] While the PhD method is based on neural networks trained on protein secondary structure prediction, AGADIR implements an algorithm based on the helix-coil transition theory aiming at the calculation of helical propensities in isolated peptides lacking significant tertiary interactions.

Overall, we observed a predominance of helix predictions (44% of the mTP residues were predicted as helical, 9% as extended, and 47% as coil or uncertain by the Rost/Sander method), but there was no clear correlation between the predicted secondary structure and the location relative to the 'R-2', 'R-3', or 'R-10' cleavage sites using either of the two prediction methods (data not shown). We conclude that mTPs have a high potential for helix formation but that secondary structure predictions do not appear to provide information that is useful in the identification of mTP cleavage sites.

CONCLUSIONS

The main result in this work is that an unsupervised Kohonen network classifies mTPs from a range of organisms according to three distinct cleavage-site motifs: 'R-2', 'R-3', and 'R-10'. Similar motifs were previously suggested based on more traditional statistical analysis[48,53] and experimental data,[54–56] and the concordance between the different approaches strengthens the hypothesis that the three motifs represent the major cleavage site motifs in mTPs. The Kohonen network failed to detect significant differences between mTPs from different groups of organisms, suggesting that MPP and MIP substrate specificity has been largely conserved throughout evolution. We note, however, that there was no clear indication of 'R-10' sites cleaved by MIP in the plant sequences, suggesting that MIP may not exist in plant mitochondria. Indeed, no candidate MIP gene has so far been found in any plant genome.

Although supervised training of feedforward neural networks allowed mTP cleavage sites to be recognized with high reliability, overprediction turned out to be a serious problem, preventing the use of these networks as cleavage site predictors. Secondary structure predictions did not reveal sufficiently distinct structural patterns in the mTPs to be of any use for prediction purposes. They do, however, strongly suggest that mTPs have a high helical

potential, consistent with the postulated importance of amphiphilic helices in mTPs.[28,46,57] Binary word encoding rather than encoding using a small number of physicochemical parameters may be one way to improve prediction performance,[58] although this requires a significantly larger training set than used here.

ACKNOWLEDGMENTS

We thank Petra Schneider, Johannes Schuchhardt, and Arne Elofson for valuable advice and discussions. This work was supported by a grant from the Fonds der Chemischen Industrie and a Boehringer Ingelheim Fonds fellowship (to G.S.), the BMBF DETHEMO project (to P.W.), and the Swedish Natural Sciences Research Council (to G.v.H. and E.G.).

REFERENCES

1. Attardi, G., Schatz, G. Biogenesis of mitochondria: Assembly of the mitochondrial membrane system. Annu. Rev. Cell. Biol. 4:289–333, 1988.
2. Hartl, F.-U., Neupert, W. Protein sorting to mitochondria: Evolutionary conservations of folding and assembly. Science 247:930–938, 1990.
3. Glick, B., Schatz, G. Import of proteins into mitochondria. Annu. Rev. Genet. 25:21–44, 1991.
4. Glick, B.S. Can Hsp70 proteins act as force-generating motors? Cell 80:11–14, 1995.
5. Neupert, W., Hartl, F.-U., Craig, E.A., Pfanner, N. How do polypeptides cross the mitochondrial membranes? Cell 63:447–450, 1990.
6. Pfanner, N., Meijer, M. Protein sorting: Pulling in the proteins. Curr. Biol. 5:132–135, 1995.
7. Schatz, G. The protein import machinery of mitochondria. Protein Sci. 2:141–146, 1993.
8. von Heijne, G. Design of protein targeting signals and membrane protein engineering. In: "Concepts in Protein Engineering and Design." Wrede, P., Schneider, G. (eds.). Berlin: Walter de Gruyter Verlag, 1994:263–279.
9. Schulte, U., Arretz, M., Schneider, H., Tropschug, M., Wachter, E., Neupert, W., Weiss, H. A family of mitochondrial proteins involved in bioenergetics and biogenesis. Nature 339:147–149, 1989.
10. Eriksson, A.C., Glaser, E. Mitochondrial processing proteinase: A general processing proteinase of spinach leaf mitochondria is a membrane-bound enzyme. Biochim. Biophys. Acta 1140:208–214, 1992.
11. Braun, H.P., Emmermann, M., Kruft, V., Schmitz, U.K. The general mitochondrial processing peptidase from potato is an integral part of the cytochrome bc1 complex of the respiratory chain. EMBO J. 11:3219–3227, 1992.
12. Eriksson, A.C., Sjöling, S., Glaser, E. A general processing proteinase of spinach leaf mitochondria is associated with the bc1 complex of the respiratory chain. In: "Plant Mitochondria." Brennicke, A., Kuck, U. (eds.). Weinheim: VCH Verlagsgesellschaft, 1993:233–241.
13. Eriksson, A.C., Sjöling, S., Glaser, E. The ubiquinol cytochrome c oxidoreduxtase of spinach leaf mitochondria is involved in both respiration and protein processing. Biochim. Biophys. Acta 1186:221–231, 1994.
14. Isaya, G., Kalousek, F. The mitochondrial intermediate peptidase. In: "Signal Peptidases." von Heijne, G. (ed.). Austin: RG Landes, 1994:87–103.
15. Kalousek, F., Hendrick, J.P., Rosenberg, L.E. Two mitochondrial matrix proteases act sequentially in the processing of mammalian matrix enzymes. Proc. Natl. Acad. Sci. U.S.A. 85:7536–7540, 1988.
16. Ou, W., Kumamoto, T., Mihara, K., Kitada, S., Niidome, T., Ito, A., Omura, T. Structural requirement for recognition of the precursor proteins by the mitochondrial processing peptidase. J. Biol. Chem. 269:24673–24678, 1994.
17. Thornton, K., Wang, Y., Weiner, H., Gorenstein, D.G. Import, processing, and two-dimensional NMR structure of a linker-deleted signal peptide of rat liver mitochondrial aldehyde dehydrogenase. J. Biol. Chem. 268:19906–19914, 1993.
18. von Heijne, G. Protein targeting signals. Curr. Opin. Cell Biol. 2:604–608, 1990.
19. Sjöling, S., Eriksson, A., Glaser, E. A helical element in the C-terminal domain of the *N. plumbaginifolia* $F_1\beta$ presequence is important for recognition by the mitochondrial processing peptidase. J. Biol. Chem. 269:32059–32062, 1994.
20. Waltner, M., Weiner, H. Conversion of a nonprocessed mitochondrial precursor protein into one that is processed by the mitochondrial processing peptidase. J. Biol. Chem. 270:26311–26317, 1995.
21. Hirst, J.D., Sternberg, M.J.E. Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. Biochemistry 31:7211–7218, 1991.
22. Schneider, G., Schuchhardt, J., Wrede, P. Artificial neural networks and simulated molecular evolution are potential tools for sequence-oriented protein design. Comput. Appl. Biosci. 10:635–645, 1994.
23. Schneider, G., Schuchhardt, J., Wrede, P. Development of simple fitness landscapes for peptides by artificial neural filter systems. Biol. Cybernet. 73:245–254, 1995.
24. Hertz, J.A., Palmer, R.G., Krogh, A.S. "Introduction to the Theory of Neural Computation." Redwood City, CA: Addison-Wesley, 1991.
25. Endo, T., Mitsui, S., Nakai, M., Roise, D. Binding of mitochondrial presequences to yeast cytosolic heat shock protein 70 depends on the amphiphilicity of the presequence. J. Biol. Chem. 271:4161–4167, 1996.
26. Hammen, P.K., Gorenstein, D.G., Weiner, H. Structure of the signal sequences for two mitochondrial matrix proteins that are not proteolytically processed upon import. Biochemistry 33:8610–8617, 1994.
27. Sjöling, S., Waltner, M., Kalousek, F., Glaser, E., Weiner, H. Processing comparison between the membrane-bound spinach leaf mitochondrial processing peptidase: MPP. integrated into the cytochrome bc1 complex and the soluble rat liver matrix MPP. Eur. J. Biochem. 242:114–121, 1996.
28. von Heijne, G. Mitochondrial targeting sequences may form amphiphilic helices. EMBO J. 5:1335–1342, 1986.
29. Bleasby, A.J., Akrigg, D., Attwood, T.K. OWL: A non-redundant composite protein sequence database. Nucleic Acids Res. 22:3574–3577, 1994.
30. Pearson, W.R., Lipman, D.L. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85:2444–2448, 1988.
31. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
32. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank. Nucleic Acids Res. 20:2019–2022, 1992.
33. Barker, W.C., George, D.G., Mewes, H.W., Tsugita, A. The PIR-International protein sequence database. Nucleic Acids Res. 20:2023–2026, 1992.
34. Kohonen, T. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43:59–69, 1982.
35. Kohonen, T. "Self-Organization and Associative Memory." 2nd edit. Berlin: Springer-Verlag, 1988.
36. Engelman, D.A., Steitz, T.A., Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu. Rev. Biophys. Biophys. Chem. 15:321–353, 1986.
37. Harpaz, Y., Gerstein, M., Chothia, C. Volume changes on protein folding. Structure 2:641–649, 1994.
38. Minsky, M., Papert, S. "Perceptrons." Cambridge, MA: MIT Press, 1969.
39. Rumelhart, D.E., McClelland, J.L., The PDB Research

Group. "Parallel Distributed Processing." Cambridge, MA: MIT Press, 1986.
40. Rechenberg, I. "Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution." Stuttgart: Frommann-Holzboog, 1973.
41. Schneider, G., Wrede, P. Development of artificial neural networks for pattern recognition in protein sequences. J. Mol. Evol. 36:586–595, 1993.
42. Schneider, G., Schuchhardt, J., Wrede, P. Evolutionary optimization in multimodal search space. Biol. Cybern. 74:203–207, 1996.
43. Schneider, G., Schuchhardt, J., Wrede, P. Peptide design in machina: Development of artificial mitochondrial protein precursor cleavage sites by simulated molecular evolution. Biophys. J. 68:434–447, 1995.
44. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta. 405:442–451, 1975.
45. Schneider, G., Röhlk, S., Wrede, P. Analysis of cleavage site patterns in protein precursor sequences with a Perceptron-type neural network. Biochem. Biophys. Res. Commun. 194:951–959, 1993.
46. Roise, D., Schatz, G. Mitochondrial presequences. J. Biol. Chem. 263:4509–4511, 1988.
47. von Heijne, G., Steppuhn, J., Herrman, R.G. Domain structure of mitochondrial and chloroplast targeting peptides. Eur. J. Biochem. 180:535–545, 1989.
48. Gavel, Y., von Heijne, G. Cleavage site motifs in mitochondrial targeting peptides. Protein. Eng. 4:33–37, 1990.
49. Rost, B., Sander, C. Prediction of protein structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.
50. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
51. Munoz, V., Serrano, L. Elucidating the folding problem of helical peptides using empirical parameters. Nature Struct. Biol. 1:399–409, 1994.
52. Munoz, V., Serrano, L. Helix Design, prediction and stability. Curr. Opin. Biotech. 6:382–386, 1995.
53. Hendrick, J.P., Hodges, P.E., Rosenberg, L.E. 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. U.S.A. 86:4056–4060, 1989.
54. Niidome, T., Kitada, S., Shimokata, K., Ogishima, T., Ito, A. Arginine residues in the extension peptide are required for cleavage of a precursor by mitochondrial processing peptidase. J. Biol. Chem. 269:24719–24722, 1994.
55. Ogishima, T., Niidome, T., Shimokata, K., Kitada, S., Ito, A. Analysis of elements in the substrate required for processing by mitochondrial processing peptidase. J. Biol. Chem. 270:30322–30326, 1995.
56. Arretz, M., Schneider, H., Guiard, B., Brunner, M., Neupert, W. Characterization of the mitochondrial processing protease of *Neurospora crassa*. J. Biol. Chem. 269:4959–4967, 1994.
57. Bedwell, D.M., Strobel, S.A., Yun, K., Jongeward, G.D., Emr, S.D. Sequence and structural requirements of a mitochondrial protein import signal defined by saturation cassette mutagenesis. Mol. Cell. Biol. 9:1014–1025, 1989.
58. Kawabata, T., Doi, J. Improvement of protein secondary structure prediction using binary word encoding. Proteins 27:36–46, 1997.