A neural network model for the prediction of membrane-spanning amino acid sequences

REINHARD LOHMANN,1 GISBERT SCHNEIDER,2 DIRK BEHRENS,2 AND PAUL WREDE2

1 Gesellschaft zur Förderung angewandter Informatik (GFAI), Rudower Chaussee 5, Gebäude 13.7, D-12484 Berlin, Germany
2 Freie Universität Berlin, Institut für Medizinische/Technische Physik und Lasermedizin (WE 19), AG Molekulare Bioinformatik, Krahmerstraße 6-10, D-12207 Berlin, Germany

(RECEIVED March 21, 1994; ACCEPTED June 29, 1994)

Abstract: The architecture and weights of an artificial neural network model that predicts putative transmembrane sequences have been developed and optimized by the algorithm of structure evolution. The resulting filter is able to classify membrane/nonmembrane transition regions in sequences of integral human membrane proteins with high accuracy. Similar results have been obtained for both training and test set data, indicating that the network has focused on general features of transmembrane sequences rather than specializing on the training data. Seven physicochemical amino acid properties have been used for sequence encoding. The predictions are compared to hydrophylicity plots.

Keywords: amino acid property; evolutionary algorithm; feature extraction; protein structure; sequence analysis

Prediction accuracy for membrane-spanning regions in amino acid sequences has still to be improved (Esposti et al., 1990; Jähnig, 1990). To date the most common prediction method is the analysis of hydrophobicity profiles (Kyte & Doolittle, 1982; Eisenberg et al., 1984; Fiser-Moore & Straud, 1984; Klein et al., 1985; Rose et al., 1985; Vogel et al., 1985; Engelman et al., 1986; Bangham, 1988). Despite their wide application, these techniques do not allow unambiguous assignment of transmembrane segments. The hydrophobic core region of signal peptides or helical structures, for example, can give rise to false-positive results (von Heijne, 1985; Fasman & Gilbert, 1990). Most techniques are not suitable for membrane proteins composed of β-sheets, such as porins (Paul & Rosenbusch, 1985).

As a step toward automatic feature extraction and prediction routines for membrane-spanning sequences, we have developed and optimized artificial neural network systems that focus on the recognition of endpoints of transmembrane sequences ("transition regions") by means of structure evolution (Lohmann, 1993). The choice of the correct filter architecture required for extraction of sequence features is the most critical step in any development of neural networks (Rumelhart et al., 1986). Therefore, both network architecture and network connection weights have been subjected as free parameters to the evolutionary optimization process of filter development. The range of different architectures was restricted only by defining the number of layers — 3 — and by operating the network in the usual "feedforward" scheme. The number of units ("neurons") in each layer remained variable, as well as the connectivity of each sigmoidal unit (Rumelhart et al., 1986).

Feedforward networks were already successfully applied to many prediction tasks concerning protein structure and function (Qian & Sejnowski, 1988; Holley & Karplus, 1991; Hirst & Sternberg, 1992; Rost & Sander, 1993; Schneider & Wrede, 1993a, 1993b, 1994). An attempt at applying a fully connected neural network to secondary structure prediction in sequences of integral membrane proteins has been successful (Bohr et al., 1988). Most approaches, however, do not focus on a systematical algorithmic optimization of network architectures.

The task is to develop an appropriate filter architecture representing the desired mathematical sequence transformation, which, for predicting transmembrane sequences, the relation between an amino acid sequence and the property "edge of membrane-spanning sequence."

Systematical simultaneous development of both architecture and connection weights of neural filter systems can be performed by a method termed "structure evolution" (Lohmann, 1992, 1993), an optimization strategy belonging to the class of evolutionary algorithms (Rechenberg, 1973; Holland, 1975; Schwefel, 1977; Goldberg, 1989; Koza, 1992; Bäck & Schwefel, 1993). Basic phenomena of biological evolution being modeled and implemented in evolutionary algorithms are inheritance, mutation (recombination), and selection. Structure evolution makes use of isolation and migration as additional phenomena observed in evolution. The procedure used for network optimization works as follows (Lohmann, 1990, 1993):
1. Generation of an initial set of network structures \( P \).
2. Assignment of \( P \) different structures to \( P \) separate populations.
3. Optimization of network weights in each population during isolation time.
4. Selection among the network architectures on the level of populations.
5. Mutation of the best architectures and replacing the worst.
6. Start next cycle at 3 with the partly new set of architectures.

The size of the corresponding search space for architectures is estimated to extend over over \( 10^{20} \) variants. During the process of network optimization, 655,213 different filter structures were generated and tested. A thorough description of the mutation procedures for network structures is in preparation.

Of great importance for successful feature extraction from amino acid sequences is their encoding scheme. A number of context-dependent physicochemical amino acid properties appear to be a method of choice when the formal binary representation (distributed representation) of amino acids leads to insufficient results (Schneider & Wrede, 1993b). For this reason, a set of 7 property scales normalized to \((-1, 1)\) intervals was selected for data representation: (I) hydrophobicity (Engelman et al., 1986); (II) hydrophilicity (Hopp & Woods, 1981); (III) polarity (Jones, 1975); (IV) volume (Zamyatin, 1972); (V) surface area (Chothia, 1975); (VI) bulkiness (Jones, 1975); (VII) refractivity (Jones, 1975). These properties are not orthogonal, as indicated by their correlation coefficients (Table 1) \( (x_1 \) and \( x_2 \) are 2 different property scales)

\[
\frac{\sum_{i=1}^{20} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{20} (x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{20} (x_{i2} - \bar{x}_2)^2}}.
\]

In total, 47 human integral membrane proteins served as data for the experiments. Proteins with transmembrane sequences not indicated as "putative" were selected from the SwissProt database (1993, release 20, EMBL Data Library, D-6900 Heidelberg, Germany, and A. Bairoch, Departement de Biochimie Medicale, Centre Medical Universitaire, 1211 Geneva 4, Switzerland; distributed by IntelliGenetics Inc., 700 East Camino Real, Mountain View, California 94040). These transmembrane sequences were identified by protease digestion mainly as described in the original literature. Because this method does not allow unambiguous assignment of the endpoints of transmembrane sequences, the data are rather noisy. This set of proteins was split into 36 training and 11 independent test sequences containing 114 and 38 positive input patterns, respectively. The ratio between positive (transition regions) and negative examples (nontransition regions) was 1:4. Any sequence window covering 13 amino acids without a membrane/nonmembrane transition between the relative window positions 6 and 7 was considered a negative example (Fig. 1A). The test patterns were not used during the training and optimization phase of the neural network system. They served for an evaluation of the filter's generalization ability in the prediction experiments. The training set was split into 3 subsets of unique size (Lohmann, 1993). None of the known 3D structures of membrane proteins was used for network training.

The final network architecture is shown in Figure 1B. The overall prediction accuracy, i.e., the fraction of correctly predicted positive and negative examples, is high for both training (92.2%) and test data (92.3%). The values demonstrate that characteristic features of transmembrane transition regions have been extracted from the sequence data. We suppose that this promising result is due to (1) a description of amino acid sequences in terms of physicochemical properties, and (2) the inductive training technique being part of structure evolution. There are, however, many false-positive predictions found in a sequence (see prediction results), and the high percentages given above only indicate extraction of a feature present in the limited set of data used.

As a result of network development, the property "bulkiness" turned out to be useless (Fig. 1B). Either this property scale does not code for important information in transmembrane sequences, or the information is contained in the other 6 property scales as indicated by their strong correlations (Table 1). We conclude that features of membrane transition regions extracted during network development seem to be mainly based on characteristic distributions of charges, potentials, and dipoles. Scales assigning geometric properties (volume, surface area, bulkiness) appear to be less important around a membrane/nonmembrane border according to the filter system. Surprisingly, "refractivity" is highly connected to many input units, whereas "hydrophobicity" is not. These 2 scales are not significantly correlated \((r = 0.03)\) but, on the other hand, "refractivity" is correlated to both "volume" \((r = 0.81)\) and "surface area" \((r = 0.81)\) (Table 1). Therefore, structurally relevant features can be recognized by the network even if the properties "volume" or "surface area" are not highly connected (Fig. 1B).

Figures 2 and 3A give representative examples of predictions in 2 human test set sequences, the precursor sequence of a single-spanning integral membrane protein, biliary glycoprotein \( I \) (Kuroki et al., 1991), and the sequence of a multispanning membrane protein, rhodopsin (Nathans & Hogness, 1984). The mean hydrophobicity plot using the scale of Engelman et al. (1986) and a 19-residue window is also shown for comparison. The neural filter system produces 2 predictions scanning the sequence from the N-terminus to its C-terminus \((N \rightarrow C)\) and vice versa \((C \rightarrow N)\). Putative transmembrane regions are indicated by minima in the resulting plots. The single transmembrane re-

### Table 1. Correlation coefficients of the 7 physicochemical property scales used for sequence encoding

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1.0</td>
<td>-0.83</td>
<td>-0.85</td>
<td>0.00</td>
<td>-0.25</td>
<td>-0.31</td>
<td>0.03</td>
</tr>
<tr>
<td>II</td>
<td>1.0</td>
<td>0.72</td>
<td>-0.43</td>
<td>-0.23</td>
<td>-0.56</td>
<td>-0.43</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>1.0</td>
<td>0.11</td>
<td>0.33</td>
<td>-0.20</td>
<td>0.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>1.0</td>
<td>0.95</td>
<td>0.78</td>
<td>0.81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>1.0</td>
<td>0.64</td>
<td>0.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>1.0</td>
<td>0.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VII</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*See text for calculation scheme. I, hydrophobicity (Engelman et al., 1986); II, hydrophilicity (Hopp & Woods, 1981); III, polarity (Jones, 1975); IV, volume (Zamyatin, 1972); V, surface area (Chothia, 1975); VI, bulkiness (Jones, 1975); VII, refractivity (Jones, 1975).*
Prediction of membrane-spanning amino acid sequences

Fig. 1. A: Scheme of the sequence windows investigated. The numbers indicate the relative sequence positions. The arrow defines the membrane/nonmembrane border. Amino acids buried in the membrane are shown by gray circles. B: Optimized neural network for prediction of transmembrane segments in amino acid sequences of human integral membrane proteins. The flow of information is top-down, from the input (physicochemical properties $P_k$) to the single output unit $O$. The arrows indicate the putative sequence position of membrane/nonmembrane transition within the investigated sequence windows. The connectivity between the amino acid properties and the first layer units ($I_i$), the connectivity between the units of the first and the second network layer ($H_j$), and the number of units in the first layer ($I_i$) and in the second layer ($j$) were subjected to systematic optimization by the structure evolution algorithm. A first layer unit $I_i$ was allowed to focus on only 1 of the 7 physicochemical amino acid properties. The transformation function of the network that determines the actual output value for a given sequence window (input) is shown below. $S(x)$ is the common sigmoidal unit transfer function (Fermi function).

region of biliary glycoprotein I can be identified by a broad minimum in each of the 2 network predictions (Fig. 2).

The filter system identifies longer transmembrane sequences than found in biochemical experiments (Figs. 2, 3). It is very likely that this is an effect based on noisy training data and network errors and, therefore, is of no biological significance.

The 7 transmembrane sequences of human rhodopsin cannot unambiguously be found by the neural network (Fig. 3A). Five of them are clearly indicated. Two of them, the fifth and the seventh one, are not recognized correctly. The hydrophobicity plot as specified above leads to a similar result. Here, the second and the third, and the sixth and the seventh transmembrane segment cannot be resolved. Looking at the plots produced by the 2 methods, however, allows the correct assignment of predictions to actual membrane-spanning sequences. This holds for all of the 11 test set sequences. Prediction of transmembrane regions
in the precursor sequence of bacteriorhodopsin from *Halobacterium salinarium* (PIR1 RAHSB) (Oesterhelt & Stoeckenius, 1971; Khorana et al., 1979) leads to slightly more pronounced signals (Fig. 3B).

Several very powerful methods exist for predicting transmembrane helical structures (Popot & Engelman, 1990; von Heijne & Manoil, 1990; White & Jacobs, 1990; Cronet et al., 1993; Taylor et al., 1994), whereas β-strand prediction is still not successful. It can be assumed that the transmembrane sequences investigated in our analysis adopt helical structures within the membrane. The filter system, therefore, probably focuses on the recognition of transmembrane helical structures only, which is
supported by the failure to predict strands in the sequence of Escherichia coli porin (data not shown). This is a disadvantage of the system. As a consequence, future experiments must aim at developing a neural filter for β-strand regions of integral membrane proteins.

We conclude that a description of amino acid sequences in terms of physicochemical properties is a good choice for developing feature detectors in a reasonable way, and structure evolution provides for a systematic development of both architecture and weights of the neural filter system. A diskette containing the prediction program for DOS computers is available from the authors on request.

Acknowledgments

We thank Georg Büldt, Gerhard Müller, Ingo Rechenberg, Mark Acht mann, Michael Herdy, Thomas Görne, and Johannes Schuchhardt for encouragement and helpful discussions. Markus Hauswald helped prepare the sequence data. The project has been supported by grants from the FCI, the BMFT, and the DFG.

References


