Hidden Markov Models and Neural Networks: Identifying Signal Peptides and Transmembrane Protein Topology

by Lisete Sousa[1], Mário Santos[2], Maria Antónia Amaral Turkman[1] and Wolfgang Urfer[3]

[1]Department of Statistics and Operation Research and CEA, University of Lisbon,
[2]Department of Plant Biology, University of Lisbon,
[3]Department of Statistics, University of Dortmund

Abstract: These report presents two methods for the identification of signal peptides and their cleavage sites. The first method is based on based neural networks and the second on hidden Markov models. The transmembrane protein topology can also be identified by a method based on hidden Markov models, which is described here in detail. The methodologies are then applied to identify a signal peptide in fOg44 lysin and to determine the topology of the transmembrane protein holin, also in fOg44 virus. Finally an outlook for possible improvements in these methodologies combining, somehow, hidden Markov models and neural networks.

Keywords: Hidden Markov models, combined neural networks, signal peptide, cleavage site, transmembrane protein topology

1 Introduction

The number of papers concerning methods for predicting protein structure has grown extraordinarily in the last few years. This probably reflects the efforts of computational biologists to fill the gap between the explosion in available biological data and the relatively slow speed at which experiments can reveal protein structures and untangle structure/function relationships. Hidden Markov models and neural networks have become increasingly popular for signal peptide and transmembrane protein topology prediction, among others. In protein data context the hidden Markov model consists of a sequence of unobservable states following a first-order Markov chain. Each state emits a symbol (an amino acid) and only the symbol sequence is directly observed. The neural network presented here is, in fact, the combination of two different networks allowing the recognition of signal peptides and their cleavage sites.

To show the applicability of these models and their importance two proteins of the same bacteriophage are analyzed. It is known that all tailed bacteriophages with double-stranded DNA genomes appear to accomplish lysis of the host cell by the concerted action of a peptidoglycan hydrolase (referred to as endolysin or lysin) and a small hydrophobic protein (holin) presumed to form specific lesions upon oligomerization in the membrane. The later function seems essential to allow access of the lytic enzyme to the cell wall compartment since in the phage lysins examined so far, the presence of a signal peptide that would target them to the translocase of the general secretory pathway (GSP) has never been demonstrated. Unlike most lysins, the N-terminal region of the Oenococcus oeni phage fOg44 lysin (Lys44) seems to function as an export signal (signal peptide).

To decide about the existence of a signal peptide in Lys44, the hidden Markov model and the neural network algorithm are applied. The models indicate, with high probability, the presence of a signal peptide with cleavage site between residues 27 and 28. Experimental data confirm this prediction and thus the suitability of the models. The hidden Markov model is once again applied but this time to predict the topology of fOg44 holin, detecting two transmembrane helices. Since the topology of fOg44 holin is not experimentally determined yet, neural networks are also applied for comparison showing slightly different results.

2 Protein Targeting

Proteins are always produced in the cell cytosol. However, many of them are meant to operate in different compartments and must be recognized by cellular components which mediate their targeting to their final destination. The correct recognition and localization of such proteins requires the presence of particular features in their primary, secondary or tertiary structure that can function as signals. Typical examples are membrane and secreted proteins.

Proteins which perform their function within the lipid environment of biological membranes have at least one, but usually several continuous segments (about 18 - 22 amino- -acyl residues) of high hydrophobicity, separated by stretches of hydrophilic and charged amino-acids. While the latter are exposed to either side of the membrane, the former remain inserted within the lipid bilayer, forming *transmembrane domains*.

2.1 Signal Peptides

The general secretory pathway (GSP) is a mechanism for protein secretion found in both eukaryotic and prokaryotic cells. The entry to the GSP is controlled by the signal peptide, an N-terminal peptide typically between 15 and 40 amino acids long, which is cleaved from the mature part of the protein during translocation across the membrane. Translocation takes place via multiprotein complex known as the translocon or translocation apparatus.

The most characteristic common feature of signal peptides is a stretch of hydrophobic amino acids called the *h-region*. The region between the initiator Met (methionine) and the h-region, the *n-region*, is typically one to five amino-acids in length, and normally carries positive charge. Between the h-region and the cleavage site is the *c-region*, which consists of three to seven polar, but mostly uncharged, amino acids. Close to the cleavage site a more specific pattern of amino acids is found: the residues at positions -3 and -1 (relative to the cleavage site) must be small and neutral for cleavage to occur correctly.

Signal peptides from different proteins do not share a strict consensus sequence; in fact, the sequence similarity between them is rather low. Bacterial signal peptides are longer than their eukaryotic counterparts, and those of Gram-positive bacteria are longer than those of Gram-negative bacteria (which have an outer membrane in addition to the cytoplasmic membrane).

2.2 Signal Anchors

Some proteins have sequences that initiate translocation in the same way as signal peptides do, but are not cleaved by signal peptidase. As the rest of the polypeptide chain is translocated trough the membrane, the resulting protein remains anchored to the membrane by the hydrophobic region, with a short N-terminal cytoplasmic domain. The uncleaved signal peptide is known as a signal anchor, and the resulting protein is known as a type II membrane protein.

The distinction between a true signal peptide and an N-terminal membrane anchor is often elusive by simple inspection of the protein primary sequence. Signal anchors have h-regions longer than those of cleaved signal peptides and the n-regions can also be much longer than 100 residues. Interestingly, experiments have shown that it is possible to convert a cleaved signal peptide to a signal anchor merely by lengthening the h-region.

2.3 Transmembrane Proteins

Transmembrane protein amino acids have three main locations: in the transmembrane helix core (in the hydrophobic tail region of the membrane), in the membrane helix caps (in the head region of the membrane), and in loops. Nevertheless, residues have different distributions on the different sides of the membrane. On the non-cytoplasmic side there are short and long loops.

Most transmembrane alpha-helices are encoded by an unusual long stretch of hydrophobic residues. This compositional bias is imposed by the constraint that residues buried in lipid membranes must be suitable for hydrophobic interactions with the lipids.

The orientations of the transmembrane helices, i.e. whether they run inwards or outwards, give the overall "topology" of the protein. It is known that the positively charged residues arginine and lysine play a major role in determining the orientation as they are mainly found in non-transmembrane parts of the protein (loops) on the cytoplasmic side, often referred to as the *positive-inside rule*.

Note that the first transmembrane segment on a few membrane proteins may also function as a signal peptide and its cleavage is required for proper function in these cases.

3 Statistical Methods

As new protein sequences are permanently being deposited in Databanks, the number of experimentally determined signal peptides, membrane anchors or transmembrane proteins topology is increasingly available for comparisons and for the development of reliable and stronger predictive schemes.

Two methods have been successfully applied to the recognition of signal peptides and their cleavage sites: combined neural networks and hidden Markov models. In contrast to neural networks, one of the advantages of the hidden Markov models is that it is usually very easy to build biological knowledge into the model in an intuitive way.

Hidden Markov models have also been used to predict membrane protein topology. One of the main advantages of these models is that it is possible to model helix length, which as only been done fairly crudely in most other methods, by setting upper and lower limits for the length of a membrane helix. Note that the hydrophobic region contained in the signal peptide (that target a protein for export) can easily be mistaken for a transmembrane region by a prediction program.

Models are estimated from the training data. The topology of the proteins in the training data is experimentally determined by biochemical and genetic methods that are not always entirely reliable. The accuracy of the model is tested by cross-validation. The resulting models can then be applied to the analysis of whole genomes and other large data sets.

3.1 Combined Neural Networks

The combined neural networks approach to the recognition of signal peptides and their cleavage sites was developed by Nielsen *et al.* (1997). It uses one network to recognize the cleavage site and another network to distinguish between signal peptides and non-signal peptides.

Neural networks

Advances in neurophysiology and new experimental techniques have greatly enhanced our understanding of the anatomy of the human brain and the physical and chemical processes occurring within it. Furthermore, mathematical models and algorithms have been designed to mimic the information processing and knowledge acquisition methods of the human brain. These models are called *neural networks*. As the name implies, neural networks consist of neurons connected into networks. Because each neuron has a large number of dendrites, many signals can be received by the neuron simultaneously. To each individual signal corresponds a synaptic strength, *weight*. A group of neurons producing a set of outputs simultaneously is called a *layer*.

As each neuron j produces its own net input Net_j (function of all signals, s_{ji}, and of all weights, w_{ji}) and output signal Out_j, these individual signals of one layer can be combined to vectors, the net input signal vector, $\boldsymbol{Net}$, and the output vector, $\boldsymbol{Out}$. The output vector, $\boldsymbol{Out}$, can be used as an input vector, $\boldsymbol{Inp}$ or $\boldsymbol{X}$, to another layer of neurons. The layers above the passive input layer are usually referred to as the *hidden* layers, because they are not directly connected to the outside world as the input units and the output neurons are, as shown in figure 3.1. If every unit sends its output to higher layers than its own and receives its input from lower layers than its one, that is the so called *feed-forward* network.

The basic operation of a neuron is always the same: it collects a net input and transforms it into the output signal via a *transfer* function (Zupan and Gasteiger, 1993); the only thing one has to choose in advance is the number of layers, and the number of neurons in each layer. The input or output variables can be: real numbers (preferable in the range from 0 to 1, or -1 to +1); binary numbers, i.e. 0 and 1; or bipolar numbers, i.e. -1 and +1. The aim is to find the appropriate vector of weights, $\boldsymbol{W}$ from the training data. Nevertheless, a learning procedure is needed to recursively improve the weights and thus the output vector. One of the most applied strategies for the correction of weights is the learning method *back-propagation of errors*. The attractiveness of the back-propagation method comes from the well defined and explicit set of equations for weight corrections. These equations are applied throughout the layers, beginning with the correction of the weights in the last (output) layer, and then continuing backwards towards the input layer.

Network design

Nielsen *et al.* (1997), used a network design in which the input symbols are the 20 amino acids and a special spacer symbol for regions between proteins; the output symbol is a score between $\mathbf{0}$ and $\mathbf{1}$ for each amino acid in the sequence. The output from the signal peptide/non-signal peptide network, the $\boldsymbol{S}$-score, can be interpreted as an estimate of the probability of the position belonging to the signal peptide, while the output from the cleavage site/non-cleavage site network, the $\boldsymbol{C}$-score, can be interpreted as an estimate of the probability of the position being the first in the mature protein (position +1 relative to the cleavage site).

A diagram of the basic network is shown in Figure 3.1. The processing units are arranged in layers, with the input units shown on the bottom and output unit shown at the top. The network is given a contiguous sequence of $\boldsymbol{m}$ amino acids (typically $\boldsymbol{m} = \mathbf{13}$). The goal of the network is to correctly predict if the middle amino acid belongs to the signal peptide (for the $\boldsymbol{S}$-score network) or if it is the first in the mature protein. The network can be considered a "window" with $\boldsymbol{m}$ positions that moves through the protein, 1 amino acid at a time. The input layer is arranged in $\boldsymbol{m}$ groups. Each group has 21 units, each unit representing 1 of the amino acids (or spacer). For a local encoding of the input sequence, 1 and only 1 input unit in each group, corresponding to the appropriate amino acid at each position, is given a value $\mathbf{1}$, and the rest are set to $\mathbf{0}$. This is called a *local coding scheme* because each unit encodes a single item.

Network training procedure

Initially, the weights in the network are assigned randomly. The performance is gradually improved by changing the weights using the back-propagation learning algorithm. During the training, the output values are compared with the desired values, and the weights in the network are altered by gradient descent to minimize the error.

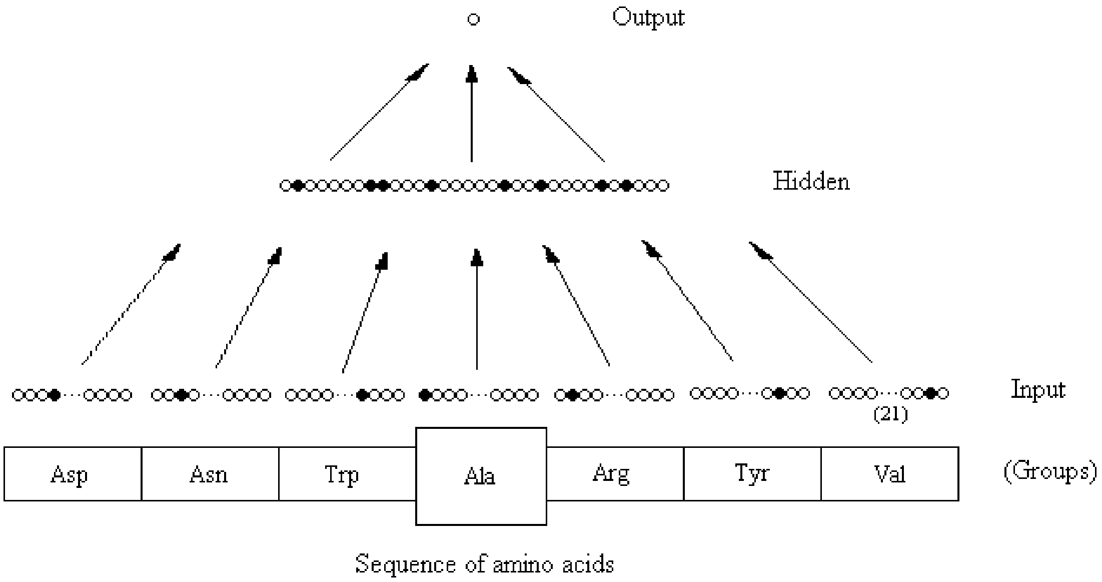

Figure 3.1 - A diagram of the network architecture illustrating the case of $\boldsymbol{m = 7}$ input groups, with 21 units per group. Information from the input layer is transformed by an intermediate layer of hidden units to produce the output unit.

Nielsen *et al.* (1997), used a data set of 1418 proteins for the training. The data set was divided into prokaryotic and eukaryotic entries and the prokaryotic date set was further divided into Gram-positive and Gram-negative bacteria. The sequence of the signal peptide and the first 30 amino acids of the mature protein from the secretory proteins were included in the data set. The first 70 amino acids of each sequence were used from the non-secretory proteins. The neural networks were feed-forward networks with zero or one layer of 2 to 10 hidden units, trained using back-propagation with a slightly modified error function.

To access the performance of the method to distinguish between secretory and non-secretory proteins, the correlation coefficient was calculated (Mathews, 1975):

$$\boldsymbol{C} = \frac{(\boldsymbol{p} \times \boldsymbol{n}) - (\boldsymbol{u} \times \boldsymbol{o})}{\sqrt{(\boldsymbol{n}+\boldsymbol{u})(\boldsymbol{n}+\boldsymbol{o})(\boldsymbol{p}+\boldsymbol{u})(\boldsymbol{p}+\boldsymbol{o})}},$$

where $\boldsymbol{p}$ and $\boldsymbol{n}$ are the correctly predicted positives and negative examples and $\boldsymbol{u}$ and $\boldsymbol{o}$ are similarly the incorrectly predicted positives and negatives. The test performances were calculated by cross validation; each data set was divided into five approximately equal sized parts and then every network run was carried out with one part as test data and and the other four parts as training data.

The $\boldsymbol{C}$-problem was best solved by networks with asymmetric windows (windows including more positions upstream than downstream of the cleavage site); the $\boldsymbol{S}$-problem, on the other hand, is best solved by symmetric or approximately symmetric windows.

Combined neural networks

If there are several $\boldsymbol{C}$-score peaks of comparable strength the true cleavage site may often be found by inspecting the $\boldsymbol{S}$- score curve in order to see which of the $\boldsymbol{C}$-score peaks coincides best with the transition from the signal peptide to the non-signal peptide region. In order to formalize this and improve prediction, Nielsen *et al.* (1997), present a geometric average of the $\boldsymbol{C}$-score and a smoothed derivative of the $\boldsymbol{S}$-score, termed the $\boldsymbol{Y}$*-score*:

$$\boldsymbol{Y_i = \sqrt{C_i \Delta_d S_i}},$$

where $\boldsymbol{\Delta_d S_i}$ is the difference between the average $\boldsymbol{S}$-score of $\boldsymbol{d}$ positions before and $\boldsymbol{d}$ positions after position $\boldsymbol{i}$:

$$\boldsymbol{\Delta_d S_i = \frac{1}{d}\Big(\sum_{j=1}^{d} S_{i-j} \sum_{j=0}^{d-1} S_{i+j}\Big)}$$

3.2 Hidden Markov Models

Consider a stochastic system consisting of $\boldsymbol{N}$ distinct states, each of wich produces output according to different probabilistic rules. A set of probabilities can be associated with each state to govern transitions from one state to another. A comprehensive description of the system requires the specification of the current state and all the previous states. If the system is modelled as an $\boldsymbol{m}$-order Markov chain, there is the assumption that the next state of the system depends only on the current state and the previous $\boldsymbol{m-1}$ states. In most natural systems, the Markov states cannot be detected directly, but only indirectly through the observation of the output, *emissions*. Thus, the term "hidden" is used.

When dialling with protein sequence data, it is often useful to think of HMMs as generative models that can "emit" protein sequences by randomly going from state to state, and in each state emit an amino acid according to the distribution for that state. For a given sequence one can calculate for instance the most probable way this sequence was generated by the model, or the total probability that it was generated by the model at all. Because it is a probabilistic model, one can use standard methods like maximum likelihood to determine the model parameters (Sousa *et al.*, 2001).

3.2.1 Predicting Signal Peptides and Signal Anchors

As it was said in the previous section, signal peptide prediction involves two tasks: (1) given that the sequence is a signal peptide, locate the cleavage site; and (2) discriminate between secretory proteins with signal peptides and non-secretory proteins. A hidden Markov model (HMM) can be applied for both prediction tasks. For signal peptides the model´s design is so that it has parts corresponding to each of the three regions of a signal peptide and such that reasonable length constraints are hard-wired in the model.

Another advantage of the HMM approach is that the HMM can easily be extended by adding other modules to the model. The signal peptide module is combined with a model of signal anchors, in order to make a model that is good at discriminating between signal peptides and anchors.

To predict signal peptides and signal anchors by an HMM, Nielsen and Krogh (1998) make the states of the model correspond to the unobserved regions in the signals ($\boldsymbol{n}$, $\boldsymbol{h}$ and $\boldsymbol{c}$). Associated with each state is a distribution over the 20 amino acids.

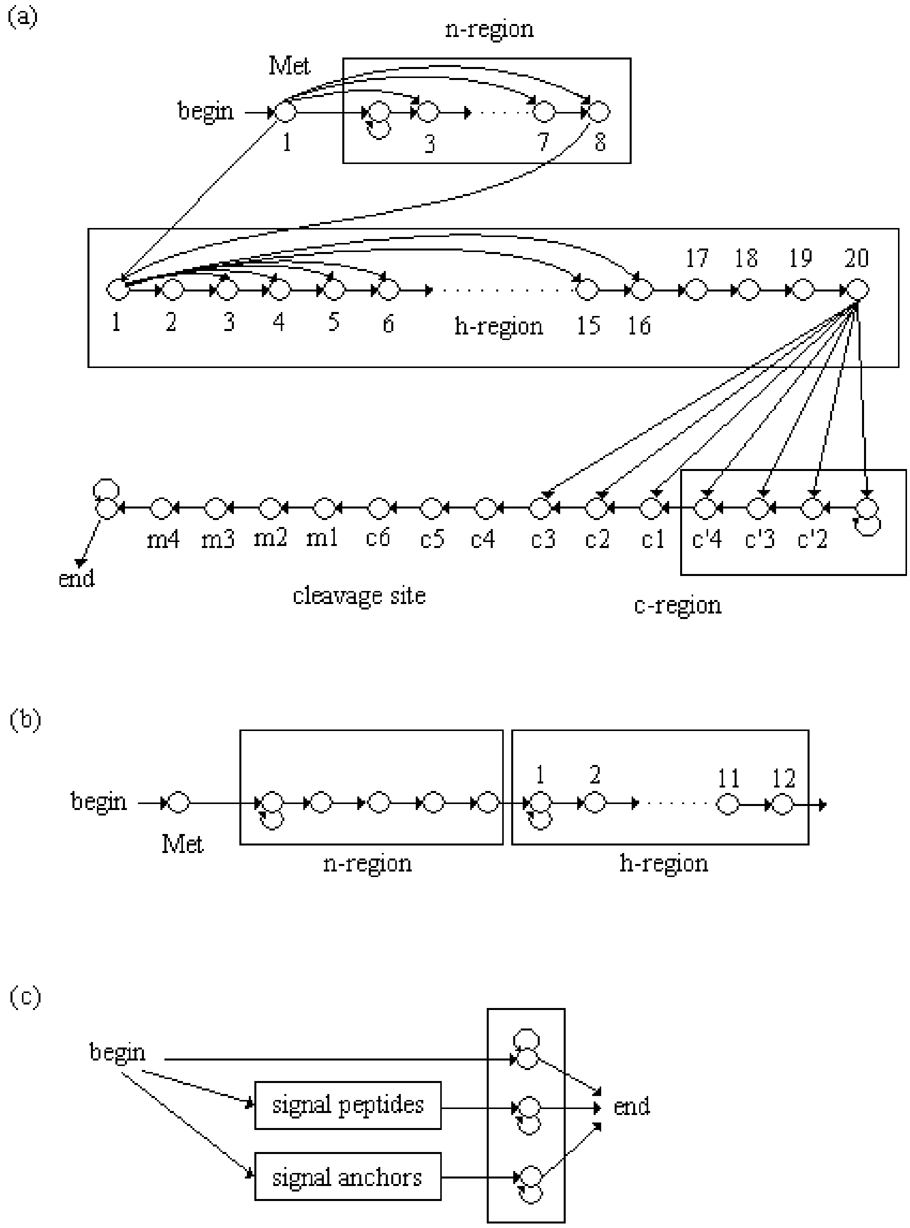

Figure 3.2.1 - The models used for signal peptides (***a***), signal anchors (***b***), and the combined model (***c***). The states in a box are tied to each other.

Model structure

To construct the topology of the hidden Markov model it is necessary to have an idea of the length and amino acids distribution of the three different regions regions in a signal peptide. This can be done using the training data. In this data set all the signal peptides were assigned an h-region ranging from 6 to 20 in length. The c-region is by definition at least three residues long, whereas the n-region is typically between 2 and 7 long, but can be significantly longer.

The regions defined in this way were used while designing the model shown in Figure 3.2.1 $(\boldsymbol{a})$. It implements an explicit modelling of the length distribution of the h-region with an array of 20 states, where there is a transition from the first state directly to each of the following 15 states, which means that the minimum length of the h-region is 6 and the maximum 20. All these states are *tied*, which means that they have the same amino acid distribution. The n-region is modelled by an array of 8 states, of which the last 7 are also tied to each other (but use another distribution than the h states). The first state has probability one for Met, because all the proteins in the data sets begin with Met. The c-region is modelled by an array of 6 states prior to the cleavage site, in which each state has a specific distribution to capture the pattern of amino acids just before the cleavage site (states c_1 to c_6 in Figure 3.2.1 $(\boldsymbol{a})$).

Components of the hidden Markov model

Basically, in HMMs there is, first, a sequence of states visited, denoted by $\boldsymbol{x_1, x_2, x_3, ...}$, and second, a sequence of emitted symbols, denoted by $\boldsymbol{y_1, y_2, y_3, ...}$. Often the sequence $\boldsymbol{y_1, y_2, y_3, ...}$ is known but the sequence $\boldsymbol{x_1, x_2, x_3, ...}$ is unknown. In such a case the sequence $\boldsymbol{x_1, x_2, x_3, ...}$ is called *hidden*.

Here, the HMM for signal peptides consists of the following five components:

(1) A set of 43 states $\mathscr{S} = \boldsymbol{\{n_1, ..., n_8, h_1, ..., h_{20}, c'_1, ..., c'_4, c_1, ..., c_6, m_1, ...m_5\}}$, where $\boldsymbol{n}$ refers to n-region residues, $\boldsymbol{h}$ to h-region, $\boldsymbol{c}$ to c-region and $\boldsymbol{m}$ refers to the residues located after the cleavage site, in the mature protein.

(2) An alphabet of 20 distinct observation outputs $\mathscr{A} = \{$all the 20 amino acids$\}$.

(3) The transition probability matrix $\boldsymbol{\Phi} = [\boldsymbol{\Phi(i, j)}]$, where,
$\boldsymbol{\Phi(i, j) = P(X_{k+1} = j | X_k = i), i, j} \in \mathscr{S}$.

(4) The emission probabilities: for each state $\boldsymbol{i} \in \mathscr{S}$ and $\boldsymbol{a} \in \mathscr{A}$,

$\boldsymbol{H(i, a) = P(}$state i emits an output $\boldsymbol{a) = P(A_k = a | \ X_k = i)}$.

For tied states, the emission probabilities are the same for each amino acid, e.g., $\boldsymbol{P(A_k = a | X_k = n_b) = P(A_k = a | X_k = n_c)}$, for any $\boldsymbol{a} \in \mathscr{A}$ and $\boldsymbol{b, c} \in \{2, ..., 8\}$.

(5) An initial distribution vector $\boldsymbol{\pi} = [\boldsymbol{\Phi(i)}]$, where $\boldsymbol{\Phi(i) = P(X_1 = i)}$.

Since the amino acid in the first state $(\boldsymbol{n_1})$ is *Met*, $\boldsymbol{\pi}$ is a column of zeros, except for the first element referring to $\boldsymbol{\Phi(n_1)}$ which is one.

Training the hidden Markov model

For training the model, data sets were made for four types of proteins: signal peptides, signal anchors, cytoplasmic and nuclear proteins (the last two as non-secretory proteins). All sets were grouped in subsets for eukaryotes, Gram-positive bacteria and Gram-negative bacteria. Proteins in all sets were truncated after 70 residues, which is the region chosen to model because almost all signal peptides are shorter then 70. All the data sets were then homology reduced so that no two sequences were homologous within a set.

The parameters of the model were then estimated from the training data by the Baum-Welch algorithm, which is a maximum likelihood procedure that iteratively increases the total likelihood of the training data (for detail see Ewens and Grant, 2001). Nielsen and Krogh (1998) did the training with the labelled data, such that the cleavage site was always correctly positioned during training, but the model was left to find out for itself where to put the boundaries between n-, h- and c- regions. To predict the cleavage site for a new sequence, the most probable path through the trained model was found by the standard Viterbi algorithm. The most probable path was also used for assigning a region to each amino acid in the sequence.

The accuracy of the HMM was tested by five-fold cross validation. The estimated model can be then used to predict the location of the cleavage site, which it finds correctly in nearly 70% of signal peptides in the training set.

Discrimination between signal peptides, signal anchors and non-secretory proteins

To discriminate between signal peptides, signal anchors and non-secretory proteins, the model was augmented by a model of anchors as shown is Figure 3.2.1 ***(b)***, ***(c)***. The structure of this model is the like the model for signal peptides, but the n- and h-regions are simpler and the c-region is of course omitted.

The whole model was trained from all types of sequences (signal peptides, signal anchors, cytoplasmic and nuclear). The most likely path through the combined model yields a prediction of which of the three classes the protein belongs to. The HMM correlation coefficient for discrimination between signal peptides and signal anchors was 0,74. For discrimination between signal peptides and non-secretory proteins the correlation coefficient was 0,94.

3.2.2 Predicting Transmembrane Protein Topology

By defining states for transmembrane helix residues and other states for residues in loops, residues on either side of the membrane, and connecting them in a cycle, it is possible to produce a model that in architecture closely resembles the biological system that is being modelled. If the model parameters are tuned to capture the biological reality, the path of a protein sequence through the states with the highest probability should be able to predict the true topology.

The hidden Markov model (HMM) is very well suited for prediction of transmembrane helices because it can incorporate hydrophobicity, charge bias, helix lengths and "grammatical constraints" into one model for which algorithms for parameter estimation and prediction already exist. "Grammatical constraints" refer to a "grammar" followed by helical membrane proteins in which cytoplasmic and non cytoplasmic loops have to alternate.

Sonnhammer *et al.* (1998), to predict transmembrane protein topology consider the states of the model corresponding to the unobserved topology, with each state representing residues belonging to one of 7 structural categories. Although there are three main locations of a residue (transmembrane helix core and helix caps, and loops), due to the different residue distributions on the different sides however, seven different states are used: one for the helix core, two for caps on either side, one for loops on the cytoplasmic side, one each for short and long loops on the non-cytoplasmic side, and one for "globular domains" in the middle of each loop. Associated with each state is a distribution over the 20 amino acids.

Model structure

The layout of the model is is shown if Figure 3.2.2. The amino acid emission probabilities of all states of the same type are tied to each other, i.e., they are estimated collectively.

The transmembrane helix is modelled by two cap regions of 5 residues each, surrounding a core region of variable length 5-25 residues. The loops between the helices are modelled by modules that contain 2× 10 states in a ladder configuration, and one self-looping state. The idea is that the 10 first states should contain most of the topogenic signals (bias, in amino acid usage), while larger globular domains are modelled in a simple way by the single self-looping state, which has a neutral amino acid distribution.

Components of the hidden Markov model

When modelling transmembrane proteins topology, the HMM consists of the following five components:

(1) A set of 96 states $\mathscr{S} = \{\boldsymbol{h_1, ..., h_{25}, cc_1, ..., cc_5, cn_1, ..., cn_5, l_1, ...l_{20}, sl_1, ..., sl_{20}, ll_1, ..., ll_{20}, g}\}$, where $\boldsymbol{h}$ refers to residues in the helix core, $\boldsymbol{cc}$ in the helix caps on the cytoplasmic side, $\boldsymbol{cn}$ in the helix caps on the non-cytoplasmic side, $\boldsymbol{l}$ in loops, $\boldsymbol{sl}$ in short loops, $\boldsymbol{ll}$ in long loops) and $\boldsymbol{g}$ in globular domains.

(2) An alphabet of 20 distinct observation outputs $\mathscr{A} = \{$all the 20 amino acids$\}$.

(3) The transition probability matrix $\boldsymbol{\Phi} = [\boldsymbol{\Phi(i,j)}]$, where $\boldsymbol{\Phi(i,j) = P(X_{k+1} = j|X_k = i), i, j} \in \mathscr{S}$.

(4) The emission probabilities: for each state $\boldsymbol{i} \in \mathscr{S}$ and $\boldsymbol{a} \in \mathscr{A}$,

$\boldsymbol{H(i,a) = P(}$state i emits an output $\boldsymbol{a) = P(A_k = a|\ X_k = i)}$.

(5) An initial distribution vector $\boldsymbol{\pi} = [\boldsymbol{\Phi(i)}]$, where $\boldsymbol{\Phi(i) = P(X_1 = i)}$.

Training the hidden Markov model

As for signal peptide prediction, the parameters of the model were estimated from the training data by the Baum-Welch algorithm. Sonnhammer *et al.* (1998) also did the training with the labelled data. The accuracy of the HMM was tested by ten-fold cross

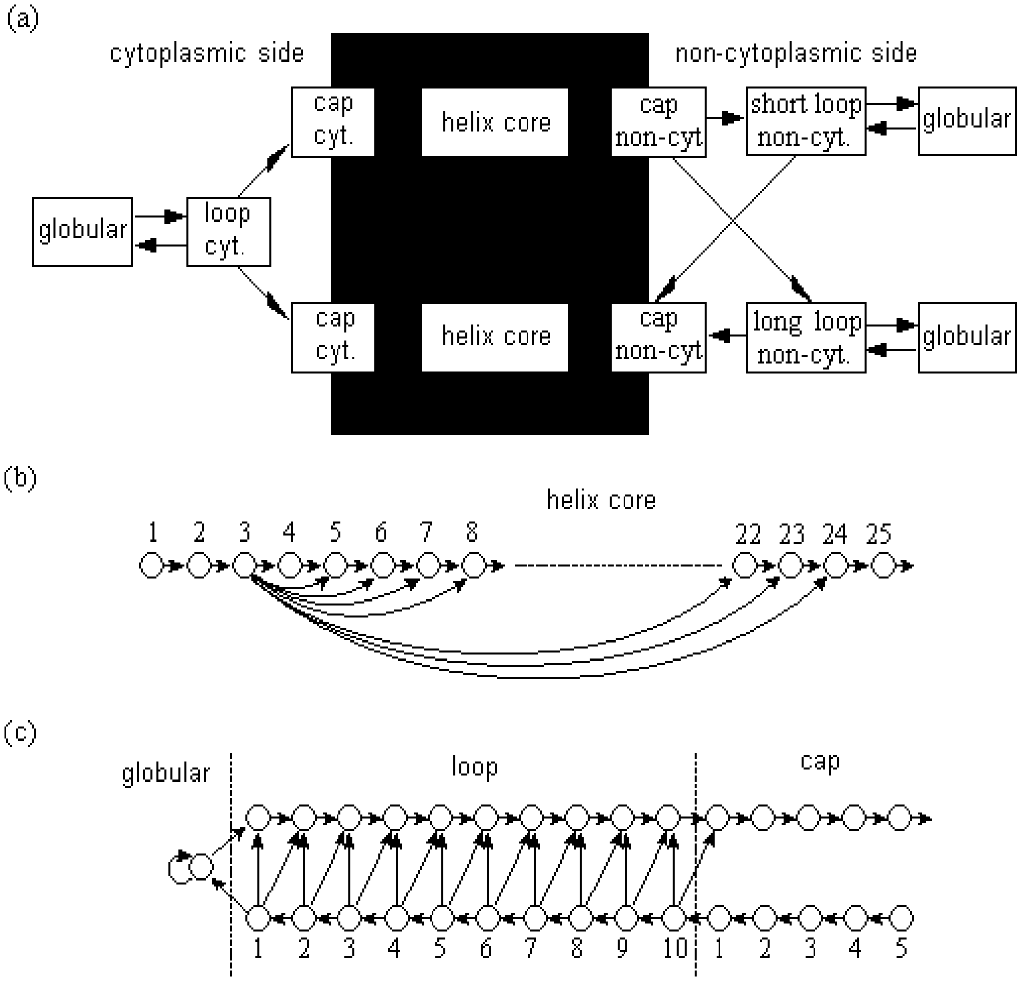

Figure 3.2.2 - (***a***) The overall layout of the HMM for transmembrane protein topology. (***b***) The structure of the inside loop, outside loop and helix cap models. (***c***) The structure of the model for the helix core.

validation and predicted correctly all the transmembrane segments for 77% of the proteins, regardless of their orientation.

The model was trained from a set of 160 membrane proteins, most of which have experimentally determined topology.

Discrimination between membrane and non-membrane proteins

The discrimination between membrane and non-membrane proteins investigated by Krogh *et al.* (2001), is based on the expected number of residues in the transmembrane helices. If this number is high, the probability that it is a helical membrane protein is also high. A threshold value can be determined from the data and used for discrimination.

4 Applications

The good quality of wine is ensured by a proper fermentation. The bacterium *Oenococcus oeni* is an important player in this process due to its ability to convert malic acid into lactic acid, in the presence of high ethanol concentrations, thus leading to a reduction in wine acidity. Bacteriophages such as fOg44 can destroy *O. oeni* cells, impairing malolactic fermentation. Successful bacteriophage attack depends on a phage product, a lytic hydrolase known as lysin. The fOg44 lysin (Lys44) belongs to a family of lysosymes with the capacity to cleave 6-O-acetylated peptidoglycans such as the present in the cell walls of bacterial pathogens such as Staphylococcus aureus, which are not hydrolysed by other enzymes. Studies on the fOg44 lysin and holin may therefore give insights on how to prevent oenococcal lysis and, at the same time, may lead to its use as a food preservative due to its action on bacterial pathogenic contaminants.

São-José *et al.* (2000), described the sequence of the lysin and holin genes from the *Oenococcus oeni* bacteriophage fOg44 and noted that the N-terminal region of its putative lysin (Lys44) was highly hydrophobic. However, during an attempt to overproduce Lys44 in an easily purifying form (as histidine-tagged fusion product, His-Lys44), they detected the production of two proteins rather than a single polypeptide, in *E. coli* extracts. They then observed that only the larger product reacted with commercial ant-His_6 antibodies, suggesting that a processing event had removed part of the N-terminal region in a fraction of synthesized proteins. From these preliminary observations came the idea that the hydrophobic N-terminal region of the fOg44 lysin, could indeed be functioning as a cleavable signal peptide.

4.1 Prediction of a Signal Peptide and Cleavage site in fOg44 Lysin

To predict the signal peptide structure and the cleavage site, São-José *et al.* (2000) used neural networks and hidden Markov models through the public domain SignalP v2.0 (http://www.cbs.dtu.dk/services/SignalP) confirming the experimental results.

Both algorithms indicate, with high probability, the presence of a peptidase cleavage site between residues 27 and 28 of Lys44, as it is shown in Figure 4.1. In the neural network method the C- and Y-scores are high at position 28, while the S-score is high before the cleavage site and low thereafter. The hidden Markov model output provides not only a prediction of the presence of a signal peptide and the position of the cleavage site, but also an approximate assignment of n-, h- and c-regions within the signal peptide. These are shown in the graphical output as probabilities for each position being in one of these three regions.

The term ”endolysins” has been traditionally used to designate bacteriophage-encoded peptidoglycan hydrolases, owing to their cytoplasmic localization as long as membrane integrity is maintained. However, the results strongly suggest that the *O. oeni* bacteriophage fOg44 encodes a secretory lytic enzyme, or exolysin, which is structurally competent for export through the GSP. Primary structure analysis predicted that the first 27 residues of fOg44 should function as a signal peptide in both Gram-positive and Gram-negative hosts.

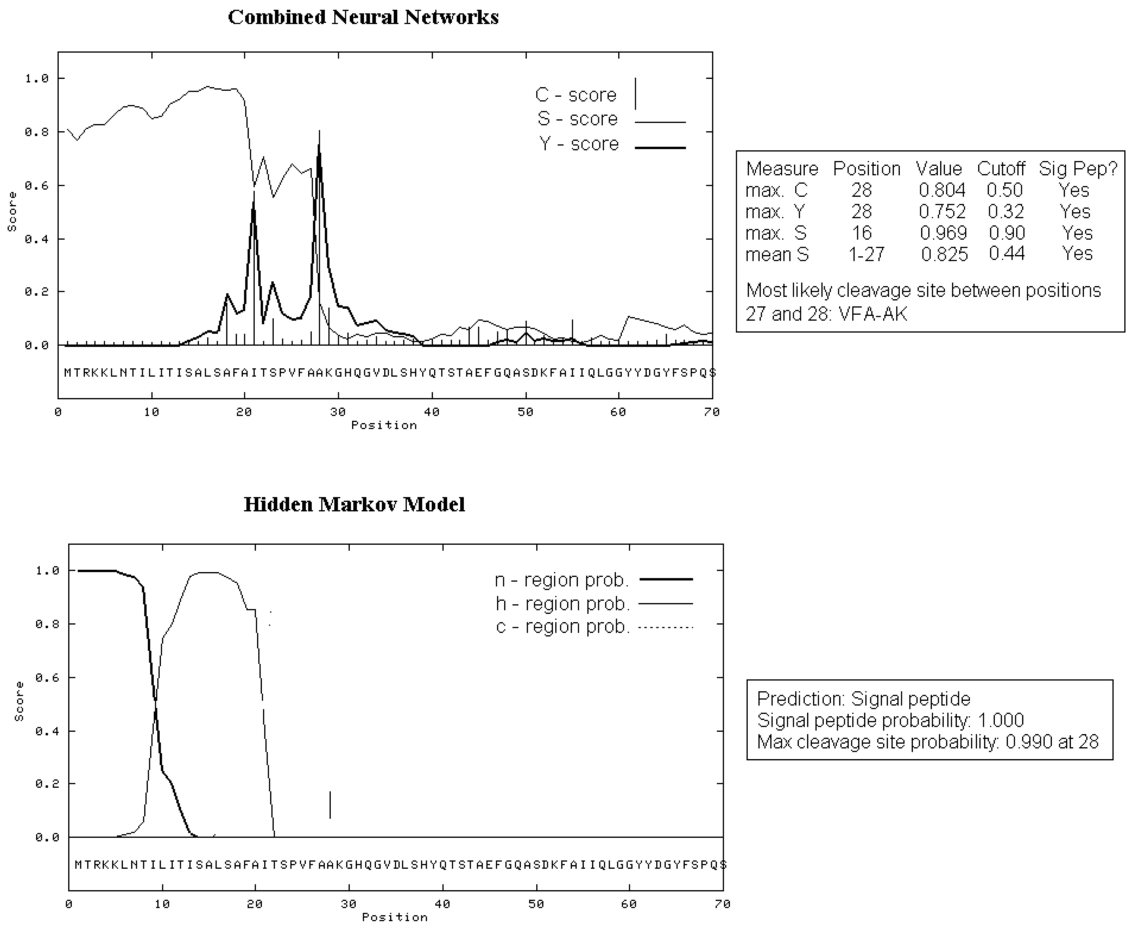

Figure 4.1 - SignalP v2.0 output using neural networks and hidden Markov models for Lys44

4.2 Prediction of Transmembrane Topology in fOg44 Holin

It is known that fOg44 holin is a transmembrane protein, nevertheless its topology is not experimentally determined, yet. The sequence of amino acids constituting holin, was analyzed by a the hidden Markov model for trough the public domain TMHMM2.0 (http://www.cbs.dtu.dk/services/TMHMM). The results are shown in Figure 4.2. The plot shows the posterior probabilities of inside/outside/transmembrane helix. Here one can see possible weak transmembrane helices that were not predicted, and one can get an idea of the certainty of each segment in the prediction. The model predicts two transmembrane helices: the first corresponding to the amino acids in positions 5-27 and the second from the 37th to the 54th amino acid. The N-terminal of the protein stays in the cytoplasmic side of the membrane running outwards and then inwards again.

For comparison neural networks (Rost *et al.*, 1995) were also applied trough the prediction server PHDhtm (http://cubic.bioc.columbia.edu/predictprotein/submit_adv.html). This method predicts two transmembrane helices but not exactly composed by the same residues as for TMHMM prediction. Here, the first transmembrane helix corresponds to the amino acids in positions 10-27 and the second from the 35th to the 52nd amino acid, but with the same in/out orientation as for TMHMM.

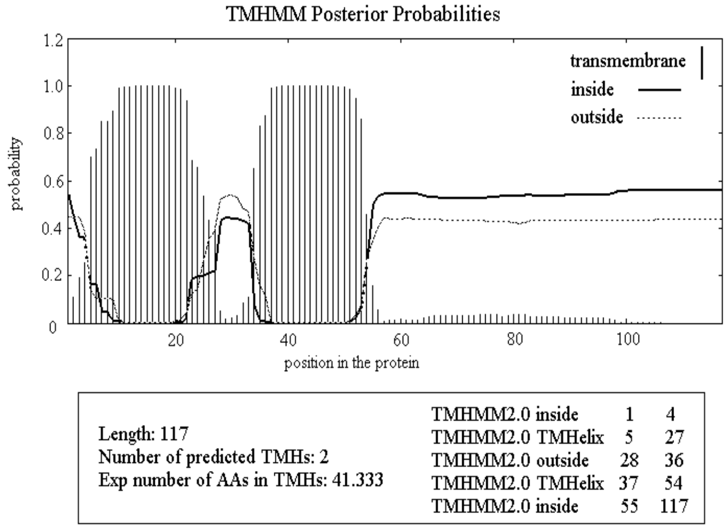

Figure 4.2 - TMHMM2.0 output for fOg44 holin

5 Outlook

A variety of tools are available to predict the topology of transmembrane proteins. Möller *et al.* (2001) evaluated the performance of the currently best known and most widely used methods for the prediction of transmembrane regions in proteins. Their results show that TMHMM is currently the best performing transmembrane prediction program. Nevertheless, they did not include in their study the TM126 model (Liò and Goldman, 1999), in which a particular secondary structure different from the structure considered by Sonnhammer *et al.* (1998), in TMHMM, is considered.

In future we develop different approaches to identify transmembrane protein topology, namely, some methodologies that have been applied to speech recognition, such as: hidden neural networks, bayesian networks and some combinations of hidden Markov models and neural networks.

6 Acknowledgements

This research has been supported by the German Research Council through the collaborative Research Center "Reduction of Complexity in Multivariate Data Structures" (DFG: SFB 475) and by the Portuguese Foundation for Science and Technology through the grant SFRH/BD/4845/2001 to L.S.

7 References

Ewens, W.J. and Grant, G.R. (2001): *Statistical Methods in Bioinformatics: An Introduction.* Springer-Verlag New York, Inc.

Krogh, A., Larsson, B., von Heijne G. and Sonnhammer, E.L.L. (2001): Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Bio.* 305, 567-580.

Liò, P. and Goldman, N. (1999): Using protein structural information in evolutionary inference: transmembrane proteins. *Mol. Biol. Evol.* 16, 1696-1710.

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

Möller, S., Croning, M.D.R. and Apweiler, R. (2001): Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17, 646-653.

Nielsen, H., Englebrecht, J., Brunak, S. and von Heijne, G. (1997): Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1-6.

Nielsen, H. and Krogh, A. (1998): Prediction of signal peptides and signal anchors by a hidden Markov model. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, J. Glasgow *et al.*, eds., AAAI Press, Calif., 122-130.

Rost, B., Casadio, R., Fariselli, P. and Sander, C. (1995): Prediction of helical transmembrane segments at 95% accuracy. *Prot. Science* 4, 521-533.

Sonnhammer, E.L., von Heijne, G. and Krogh, A. (1998): A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, J. Glasgow *et al.*, eds., AAAI Press, Calif., 175-182.

São-José, C., Parreira, R., Vieira, G. and Santos, M.A. (2000): The N-terminal region of the *Oenococcus oeni* bacteriophage fOg44 lysin behaves as a bona fide signal peptide in *Escherichia coli* and as a *cis*-inhibitory element, preventing lytic activity on Oenococcal cells. *J. Bacteriology* 182, 5823-5831.

Sousa, L., Santos, M.A., Turkman, M.A.A. and Urfer, W. (2001): *Bayesian Analysis of Protein Sequence Data.* Technical Report 48/01, University of Dortmund.

Zupanm, J. and Gasteiger, J. (1993):*Neural Networks for Chemists: An Introduction.* VCH.