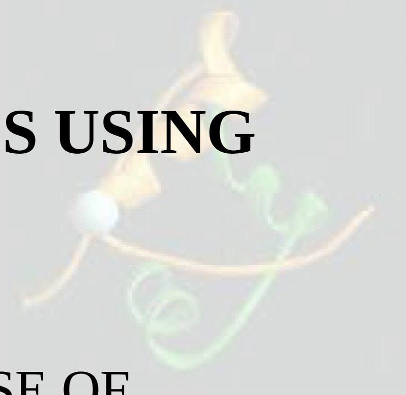


IDENTIFICATION OF SIGNAL PEPTIDES USING A HIDDEN MARKOV MODEL

IST APPLICATION TO A LYTIC HYDROLASE OF BIOTECHNOLOGICAL IMPORTANCE

Lisete Sousa *Dep. of Statistics and Op. Research and CEA - Lisbon Univ.*

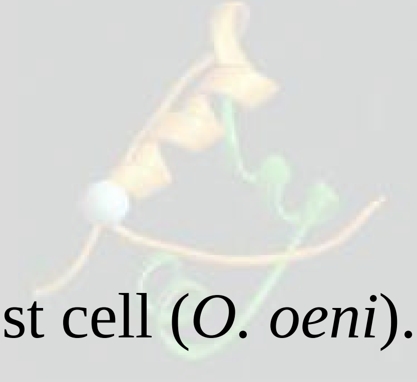
M. A. Amaral Turkman *Dep. of Statistics and Op. Research and CEA - Lisbon Univ.*
Wolfgang Urfer *Dep. of Statistics - Dortmund Univ.*
Mário Santos *Dep. of Plant Biology - Lisbon Univ.*

Partially sponsored by FCT - MCT

TOPICS

- Biological background
  - fOg44 lysin
  - Proteins
  - Signal peptides and signal anchors
- Hidden Markov models (HMMs)
  - Basic concepts
  - Model structure
  - Searching for the hidden sequence
- Application - SignalP

Biological Background



fOg44 lysin

- fOg44 bacteriophage accomplishes lysis of the host cell (*O. oeni*).
- Lysis happens by the concerted action of a lytic hidrolase known as lysin (Lys44).
- During an attempt to overproduce Lys44, São-José, *et al.* (2000) detected the production of two proteins, rather than a single polypeptide, in *E. coli* extracts.

Maybe the hidrophobic N-terminal region of the fOg44 lysin functions as a cleavable signal peptide

Protein structure

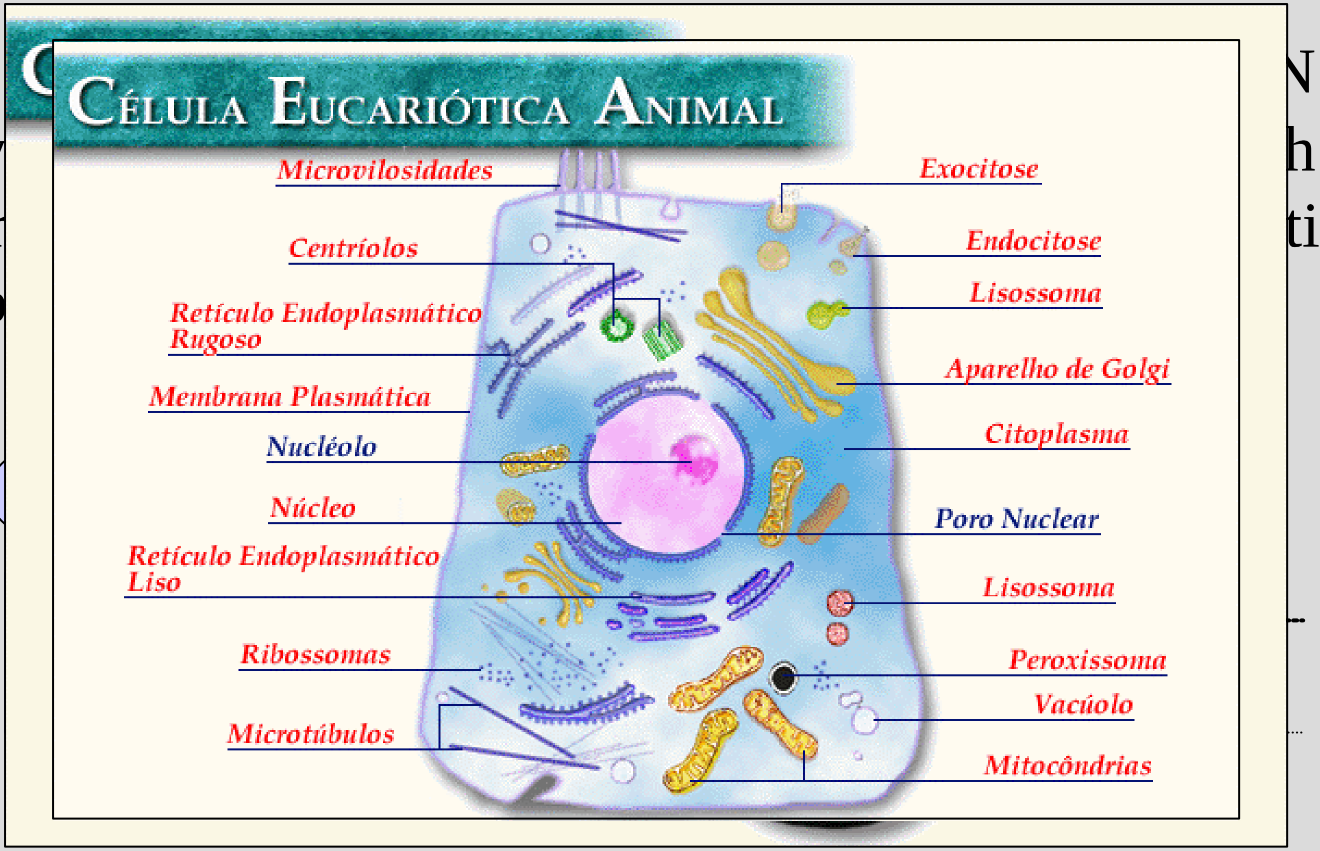
- Primary structure (amino acids sequence)
- Secondary structure (α-helix, β-sheet,...)
- Tertiary structure (three-dimensional structure)
- Quaternary structure (group of polypeptides)

Signal peptide

The general secretory pathway (GSP) is a mechanism for protein secretion in both prokaryotic (Gram-positive and Gram-negative) and eukaryotic cells.

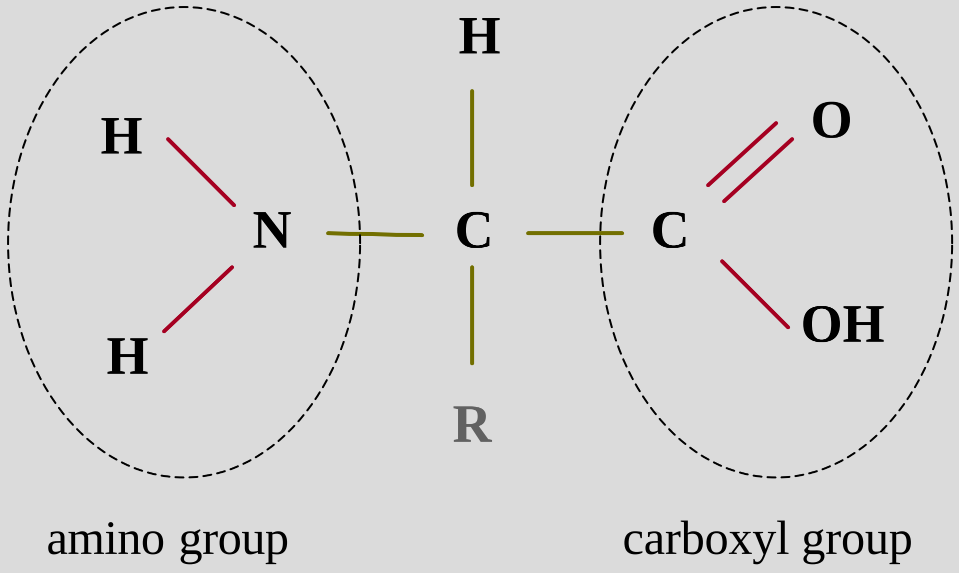
The entry ... N-terminal peptide ty... h is cleaved fr... tion across the memb...

N-terminal



C-terminal

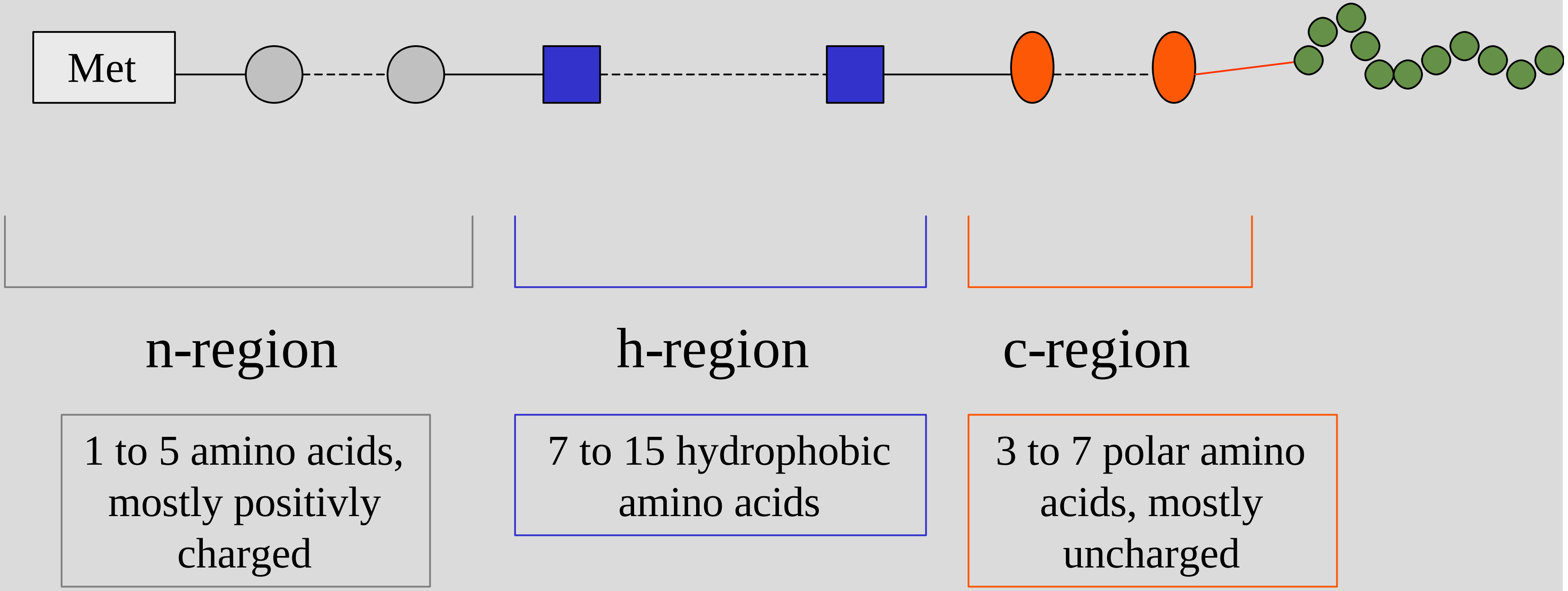
Amino acid general structure



- ○ There are 20 different amino acids in proteins.
- ○ R - side chain specifying the amino acid.
- ○ The amino acids can be hydrophobic or hydrophilic. They can also be charged and each one has a specific size.

Signal peptide structure

amino acid sequence



n-region

1 to 5 amino acids, mostly positivly charged

h-region

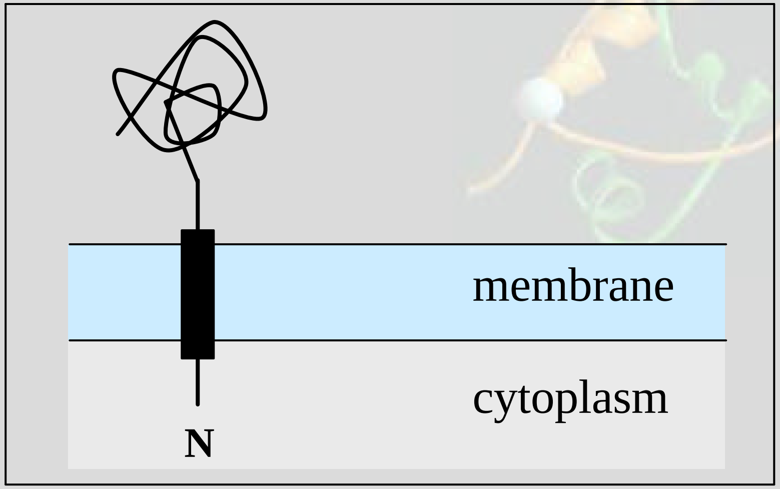
7 to 15 hydrophobic amino acids

c-region

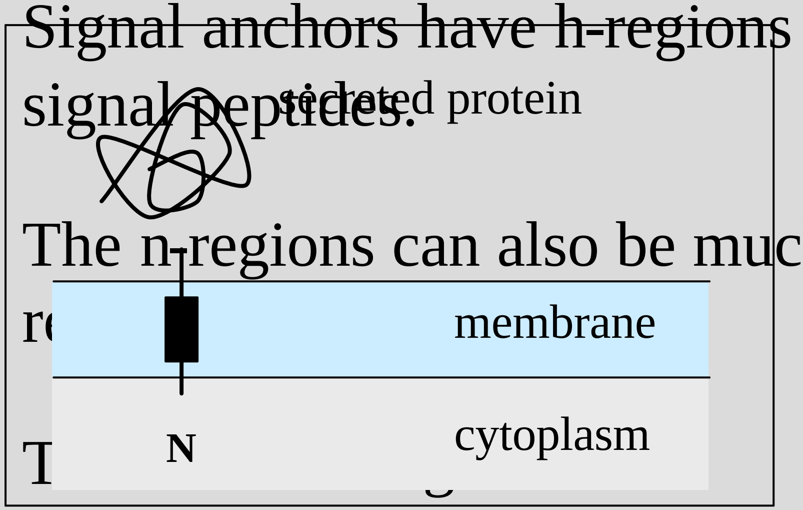
3 to 7 polar amino acids, mostly uncharged

Signal anchors

The signal anchor is not cleaved off and the protein is anchored to the membrane



- Signal anchors have h-regions longer than those of cleaved signal peptides.
- The n-regions can also be much longer, up to more than 100 re
- T

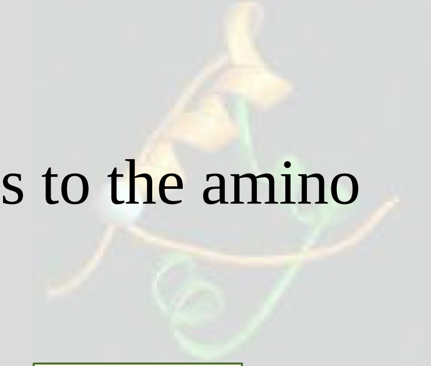


The signal peptide is cleaved off and the mature protein released

Hidden Markov Models

- Signal peptide prediction involves two tasks:
  - Given that the sequence is a signal peptide, locate the cleavage site
  - Discriminate between secretory proteins and non-secretory proteins
- Other methods:
  - Weighting matrices
  - Neural networks

Basic concepts



$\{X_k,\ k=1,...,N\}$ first-order Markov chain where *k* refers to the amino acid position in the sequence

$X_1, X_2, \ldots, X_N$ - sequence of visited states (hidden) regions

$A_1, A_2, \ldots, A_N$ - sequence of emitted symbols (known) amino acids

⇨ A set of 43 states:

$$\boldsymbol{S} = \{n_1, n_2, \ldots, n_8, h_1, h_2, \ldots, h_{20}, c'_1, \ldots, c'_4, c_1, \ldots, c_6, m_1, m_2, m_3, m_4, m_5\} =$$

$$= \mathcal{N} \cup \mathcal{H} \cup \mathcal{C} \cup \mathcal{M}$$

⇨ A set of 20 observation symbols:

$\mathcal{A}$ = {the 20 distinct amino acids}

⇨ Transition probability matrix, $\Phi = [\Phi(i,j)]$:

$$\Phi(i,j) = P(X_{k+1} = j \mid X_k = i) \qquad i\,,\,j \in \mathcal{S}$$

⇨ Emission probability matrix, $H = [H(i,a)]$:

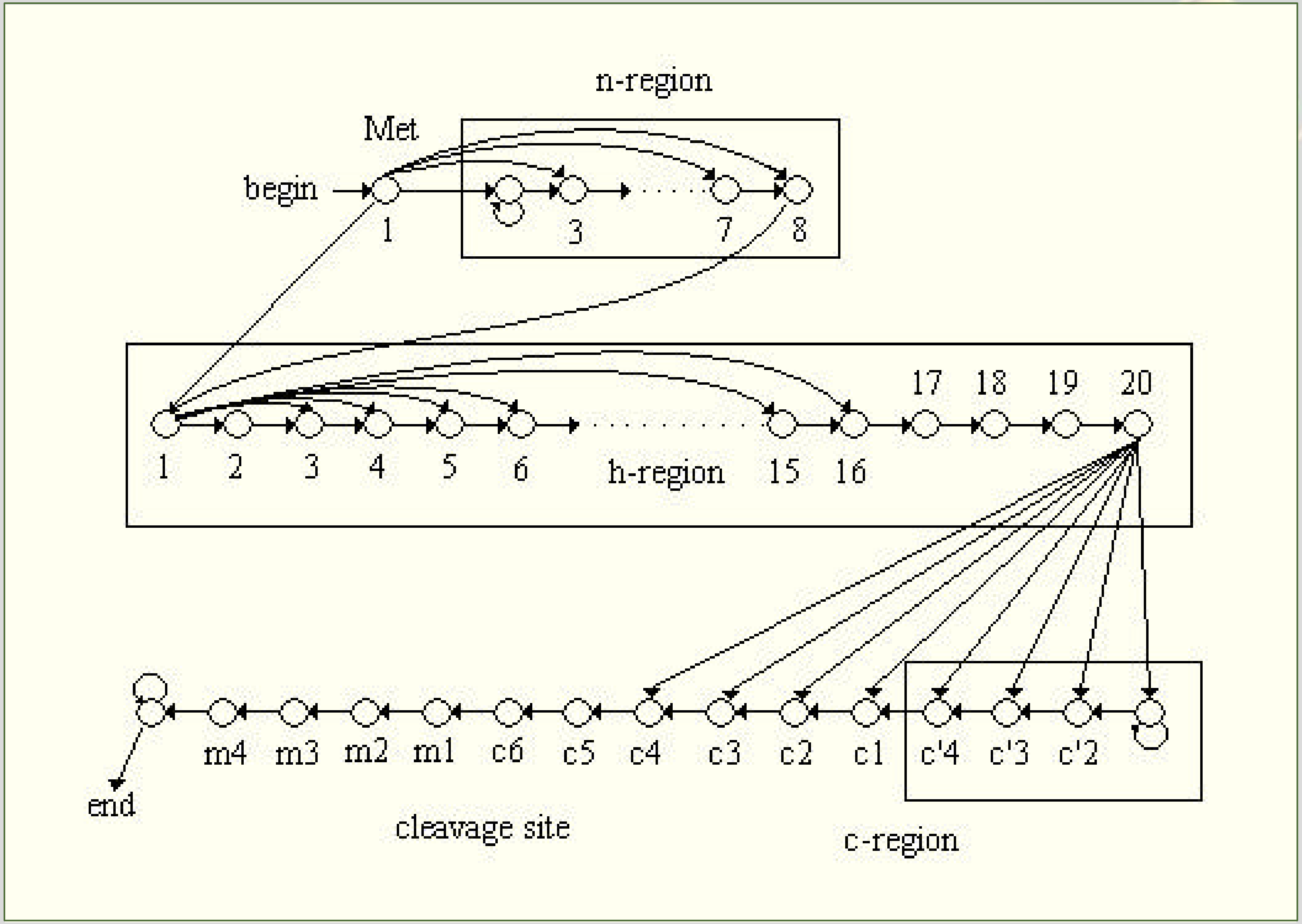
$$H(i,a) = P(A_k = a \mid X_k = i) \qquad i \in \mathcal{S}e \quad a \in \mathcal{A}$$

⇨ Initial distribution vector $\pi = (\pi_i)$:

$$\pi_i = P(X_1 = i) \qquad i \in \mathcal{A}$$

Note: Usually $K \in \{1,...,70\}$ because almost all signal peptides are shorter than 70.

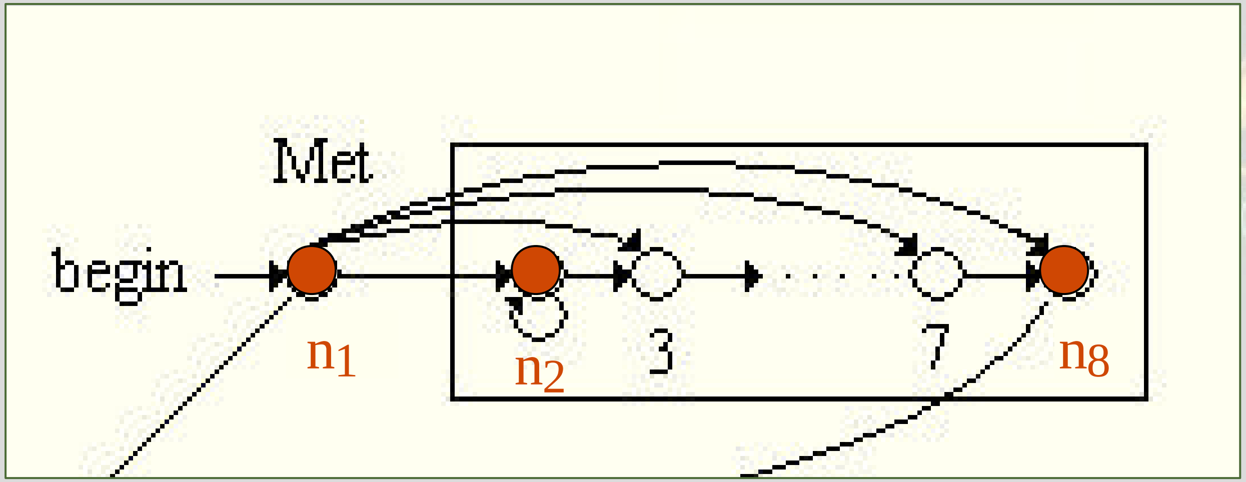
Model structure



Training set: 1665 signal peptides

Nielsen and Krogh, 1998

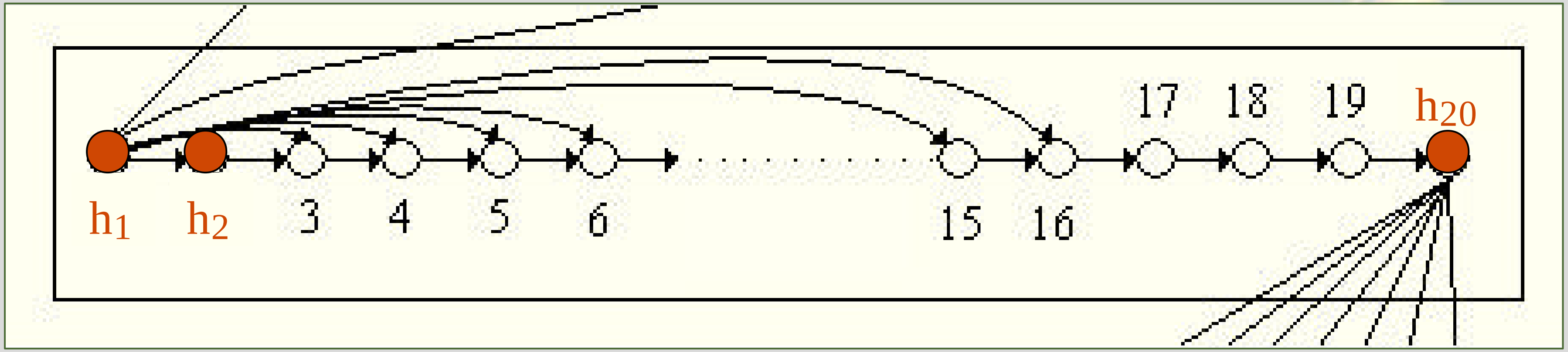
n-region



- The n-region is typically between 2 and 7 amino acids long, but can be significantly longer.
- It is modelled by an array of 8 states, of which the last 7 are tied to each other:

$$\mathcal{N} = \{n_1, n_2, \ldots, n_8\}$$

h-region

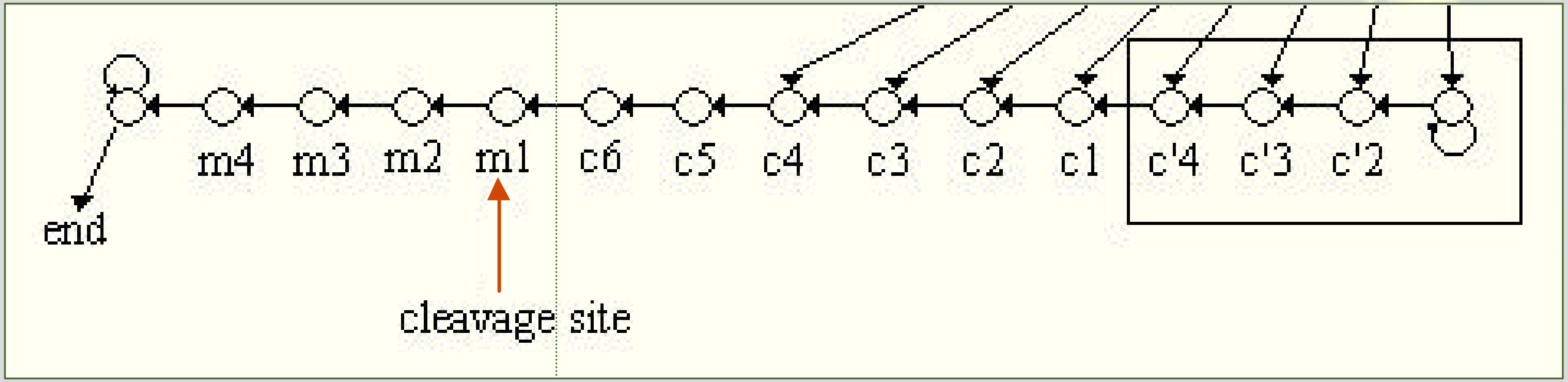


- The minimum length of the h-region is 6 amino acids and the maximum 20, with very few exceptions.

- It is modelled by an array of 20 states, all tied to each other:

$$\mathscr{H} = \{h_1, h_2, \ldots, h_{20}\}$$

c-region



- The c-region is by definition at least 3 amino acids long.
- It is modelled by an array of 10 states, of which the first 4 are tied to each other :

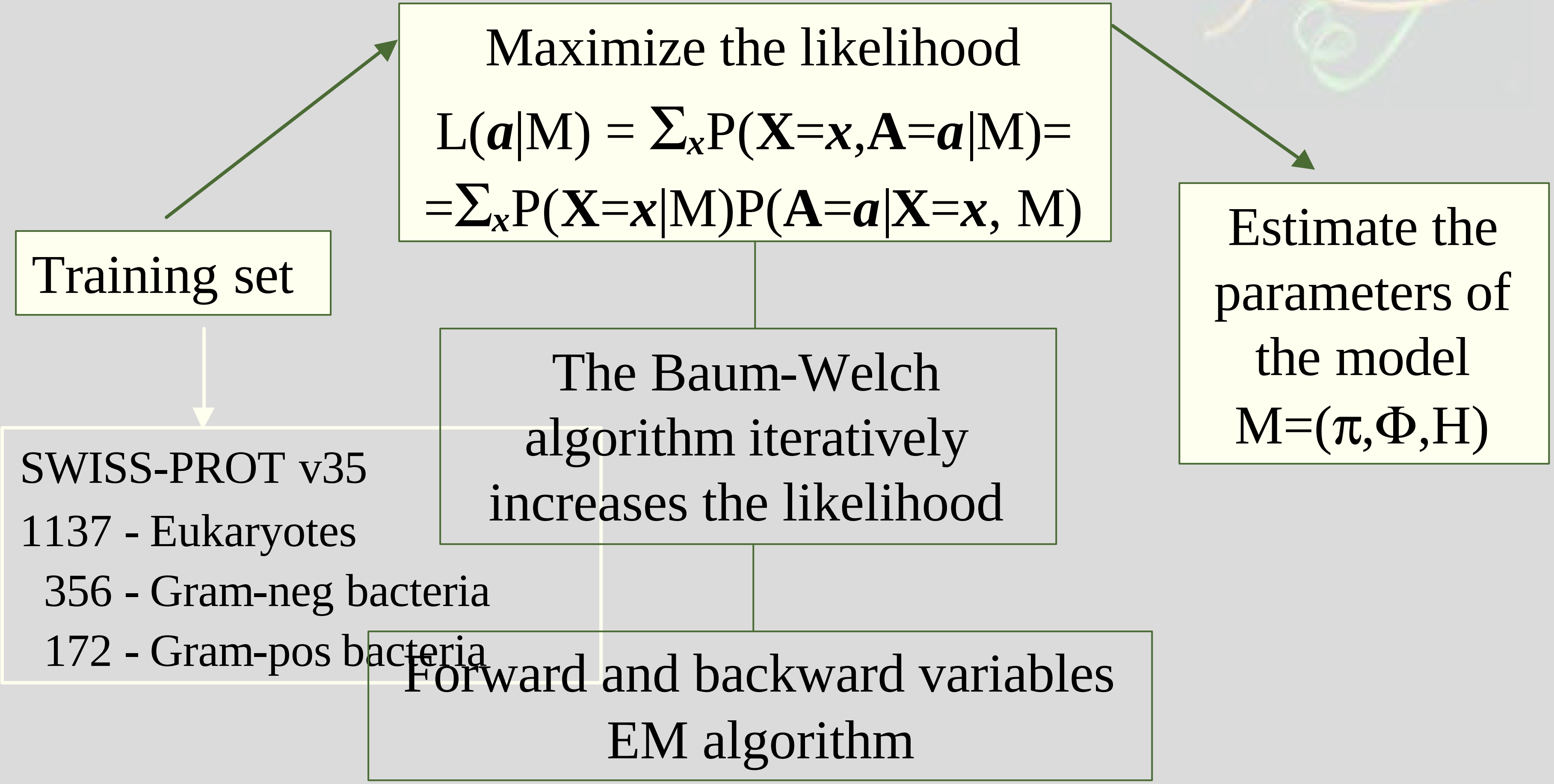
$$\mathcal{C} = \{c'_1, \ldots, c'_4, c_1, \ldots, c_6\}$$

$$\mathcal{M} = \{m_1, m_2, m_3, m_4, m_5\}$$

Searching for the hidden sequence

$\boldsymbol{a} = (a_1, a_2,\ldots,a_N)$

$\boldsymbol{x} = (x_1, x_2,\ldots,x_N)$



Training set

SWISS-PROT v35

1137 - Eukaryotes

356 - Gram-neg bacteria

172 - Gram-pos bacteria

Maximize the likelihood

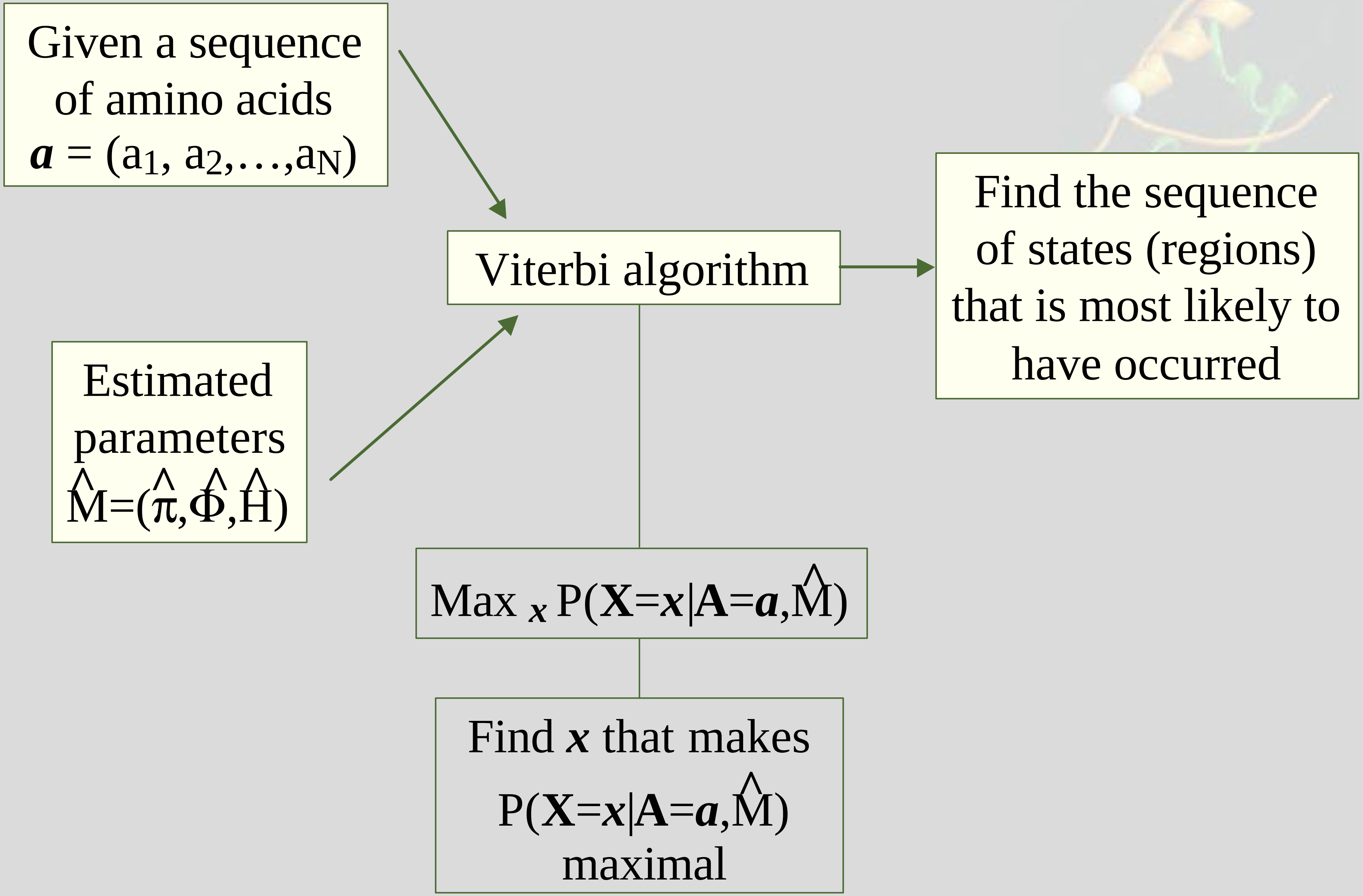
$$L(\boldsymbol{a}|M) = \Sigma_{\boldsymbol{x}} P(\mathbf{X}=\boldsymbol{x},\mathbf{A}=\boldsymbol{a}|M)=$$
$$=\Sigma_{\boldsymbol{x}} P(\mathbf{X}=\boldsymbol{x}|M)P(\mathbf{A}=\boldsymbol{a}|\mathbf{X}=\boldsymbol{x}, M)$$

The Baum-Welch algorithm iteratively increases the likelihood

Forward and backward variables
EM algorithm

Estimate the parameters of the model $M=(\pi,\Phi,H)$

Sousa *et al*. (2001)



Given a sequence of amino acids $\boldsymbol{a} = (a_1, a_2,\ldots,a_N)$

Estimated parameters $\hat{M}=(\hat{\pi},\hat{\Phi},\hat{H})$

Viterbi algorithm

Find the sequence of states (regions) that is most likely to have occurred

$\text{Max}_{\boldsymbol{x}}\, P(\mathbf{X}=\boldsymbol{x}|\mathbf{A}=\boldsymbol{a},\hat{M})$

Find ***x*** that makes $P(\mathbf{X}=\boldsymbol{x}|\mathbf{A}=\boldsymbol{a},\hat{M})$ maximal

The most probable path is used for assigning a region to each amino acid in the sequence

P (position *k* corresponds to region $\mathcal{R}$ | position *k-1* corresponds to region $\mathcal{R}^*$ and residue in position *k* corresponds to amino acid a_k)=

$$= \Sigma_{j\in \mathcal{R}} P(X_k=j \mid X_{k-1}=i,\ A_k= a_k\ ,\ \hat{M})$$

and to predict the cleavage site

P(position *k* corresponds to the cleavage site | position *k-1* corresponds to region $\mathcal{R}^*$ and residue in position *k* corresponds to amino acid a_k) = $P(X_t = m_1 \mid X_{t-1}=i,\ A_t=a_t\ ,\ \hat{M})$

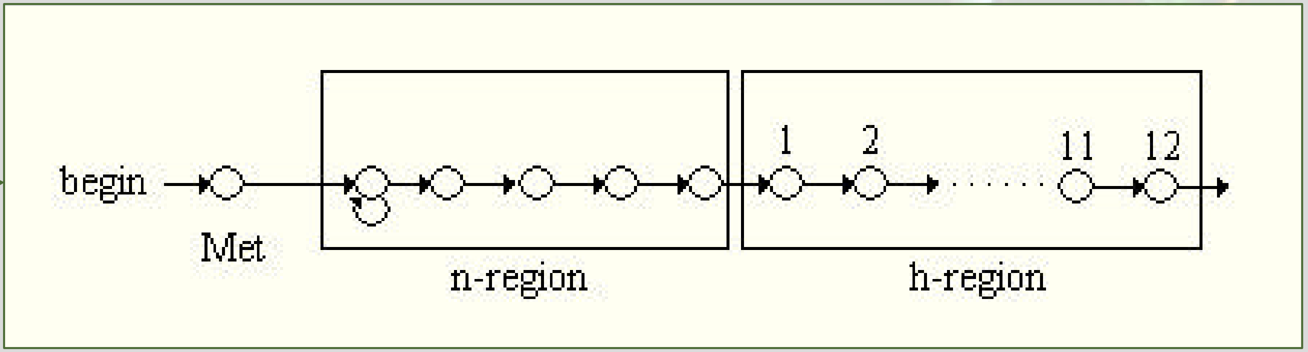
$i \in \mathcal{R}^*$

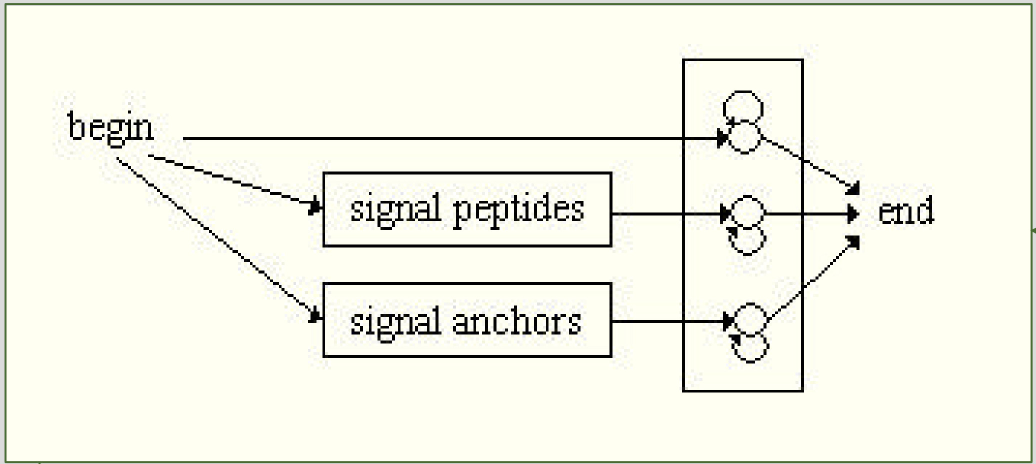
$\mathcal{R} \in \{\mathcal{N}, \mathcal{H}, \mathcal{C}\}$

$\mathcal{R}^* \in \{\mathcal{N}, \mathcal{H}, \mathcal{C}, \mathcal{M}\}$

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

signal anchors





combined model

The whole model is trained from all types of sequences in the training set (1665 SP, 67 SA, 1937 N-S)

Application

Sequence of fOg44 lysin?

http://www.ncbi.nlm.nih.gov

FASTA FORMAT

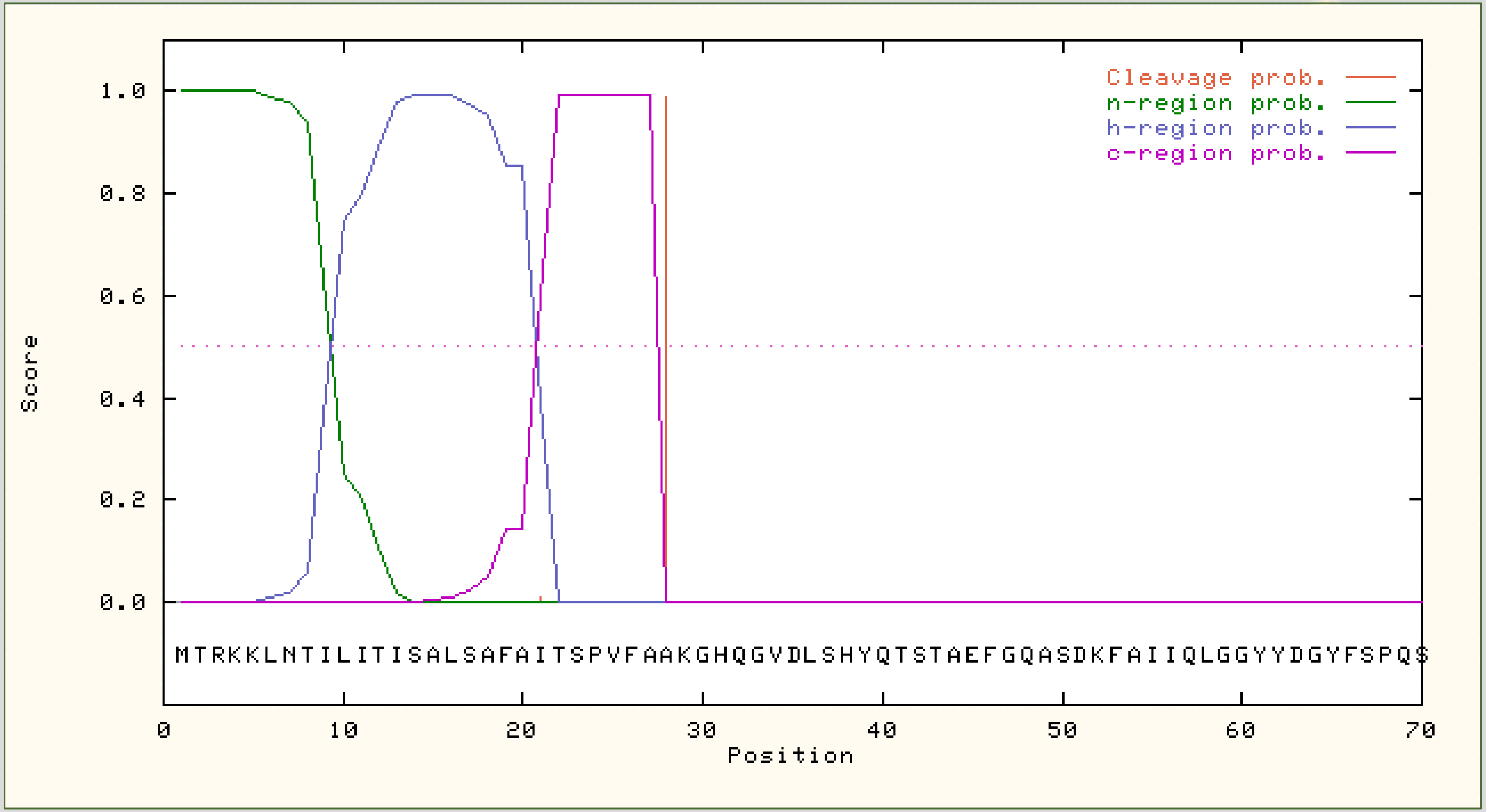
432 amino acids

```
>gi|4204413|gb|AAD10705.1|Lys44[Oenococcusoenitemperate bacteriophage fOg44]
MTRKKLNTILITISALSAFAITSPVFAAKGHQGVDLSHYQTSTAEFGQASDKFAIIQLGG
YYDGYFSPQSTYATQVASTIAQGKRAHTYIYSQFSSNAQADQILNYYFPKVQIPKFSIVA
LDVESGNPNTASVEYALAKIKFAGYTPVLYGYKSFLTAHLDLASIAKTYPLWLAEYPN
YNVTTSPNYNYFPSYDNIGIFQFTSTYKAGGLDGDIDLTGITDNGYKGTTTASTGGTAV
KTTTSTPAVKAGQQANNTPKSSITVGDTVKVNFSASKWSTGESIPSWVKGKSYKVLQV
SGNNVLLAGLSSWISKSNVEILLTTSTAAKISAPSSTGYYTVRSGDTLGAIAAKYGTTYQ
KLASLNGIGSPYIIIPGEKLKVSGSVSSSSASYYKVASGDTLSAIASKYGTSVSKLVSLNG
LKNANYIYVGQTLRIK
```

http://www.cbs.dtu.dk

Nielsen and Krogh, 1998

Submit the amino acid sequence to SignalP v2.0 in order to predict if the sequence is a signal peptide and, if so, where it will be cleaved.



P(position *k* corresponds to the cleavage site | position *k-1* corresponds to region λ^* and residue in position *k* corresponds to amino acid a_k) = $P(X_t = m_1 \mid X_{t-1} = 1, A_t = a_t, \hat{M})$

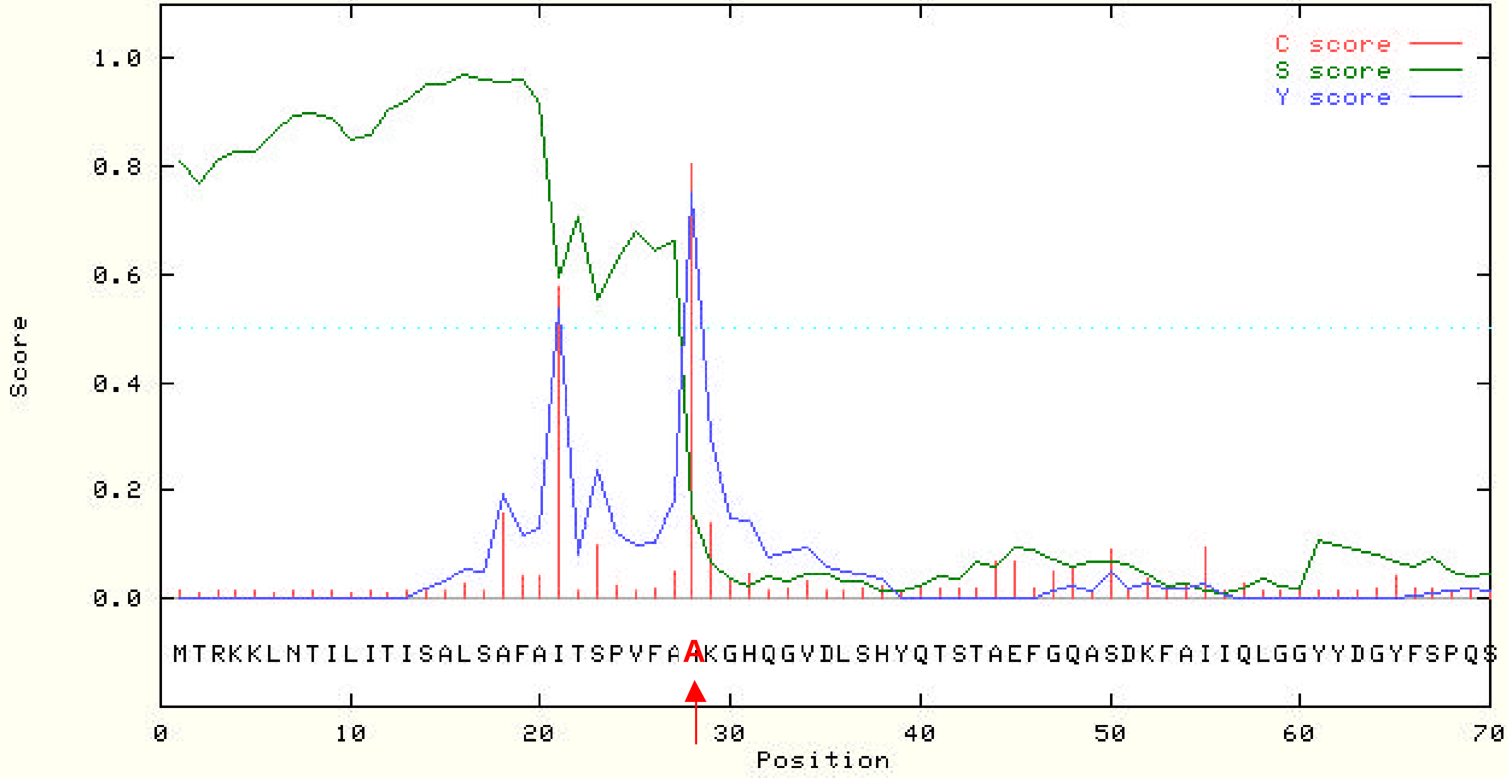
>gi_4204413_gb_AAD10705.1_

Prediction: Signal peptide

Signal peptide probability: 1.000

Max cleavage site probability: 0.990 at 28

Neural networks output



```
>gi_4204413_   length
# Measure  Position   Value   Cutoff   signal petide?
  max. C      28      0.804   0.50        Yes
  max. Y      28      0.752   0.32        Yes
  max. S      16      0.969   0.90        Yes
  mean S     1-27     0.825   0.44        Yes
# Most likely cleavage site between pos. 27 and 28: VFA-AK
```

Conclusions

- Lys44 is, in fact, a signal peptide.
- Cleavage site between residues 27 and 28.
- 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.

Future Work ...

- We intent to apply different approaches:
  - Hidden neural networks
  - Bayesian networks
  - Combine hidden Markov models and neural networks

- These approaches shall be applied to define transmembrane protein topology.

References

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. *Journal of Bacteriology* 182, 5823-5831.

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.

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