

Structure d'un signal peptide :

- région N-ter chargée positivement (n-region) suivie
- région hydrophobe (h-region) suivie
- région neutre mais polaire (c-region)

Règles autour du site de clivage :

Positions -3 et -1 doivent être de petits acides aminés neutres.

Signal P version 1 : neural network (dans version 2, ajout d'un HMM):

2 réseaux construits : un pour prédire le site de clivage et un pour prédire si l'acide aminé appartient ou pas à un signal peptide.

Réseaux avec zéro ou une couche de neurones cachés dont le nombre varie de 2 à 10.

Réseaux entraînés avec différentes séquences :

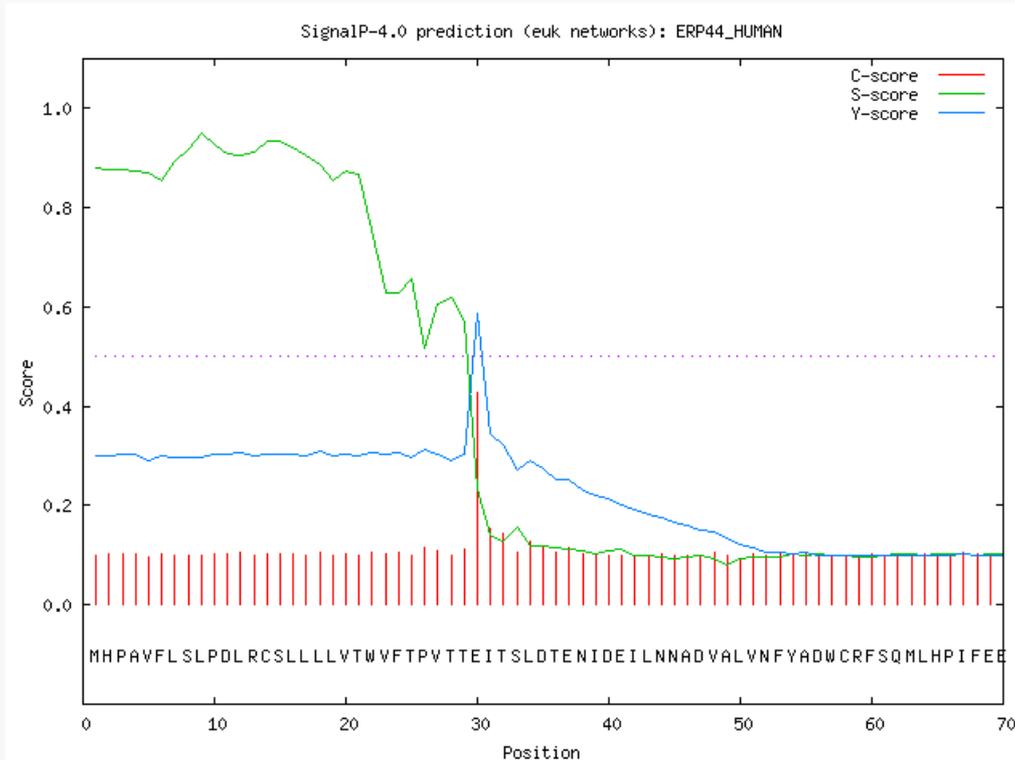
- eucaryotes
- procaryotes Gram+
- procaryotes Gram-

Donc différents réseaux pour chacun des groupes.

Résultat de SignalP v4

ERP44_HUMAN

SignalP result:



Measure Position Value

max. C 30 0.427

max. Y 30 0.586

max. S 9 0.950

mean S 1-29 0.821

D 1-29 0.713

ERP44_HUMAN SP= 'Yes' Cleavage site between pos. 29 and 30: VTT-EI D= 0.713 D-cutoff= 0.450 Networks= SignalP-noTM

C-score : obtenu par le neural network dédié à l'identification du site de clivage

S-score : obtenu par le neural network signal peptide/no signal peptide

Y-score : combine les deux scores permettant une meilleure prédiction du site de clivage :

$$Y_i = \sqrt{C_i \Delta_d S_i} \quad (1)$$

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

$$\Delta_d S_i = \frac{1}{d} \left(\sum_{j=1}^d S_{i-j} - \sum_{j=0}^{d-1} S_{i+j} \right) \quad (2)$$

mean-S : moyenne des scores de S de la position N-ter à la position ayant le Y max; mean-S > 0.5 → signal peptide (version 1 et 2)

D score : $D = wY_{\max} + (1-w)S_{\text{mean}}$
D > D_{cut} → signal peptide (w et D_{cut} optimisés sur les jeux tests)
version 3 et 4 de SignalP

HMM de SignalP version 2

Les états dans les boîtes ombrées sont liés, c'est-à-dire qu'ils ont la même distribution d'acides aminés.

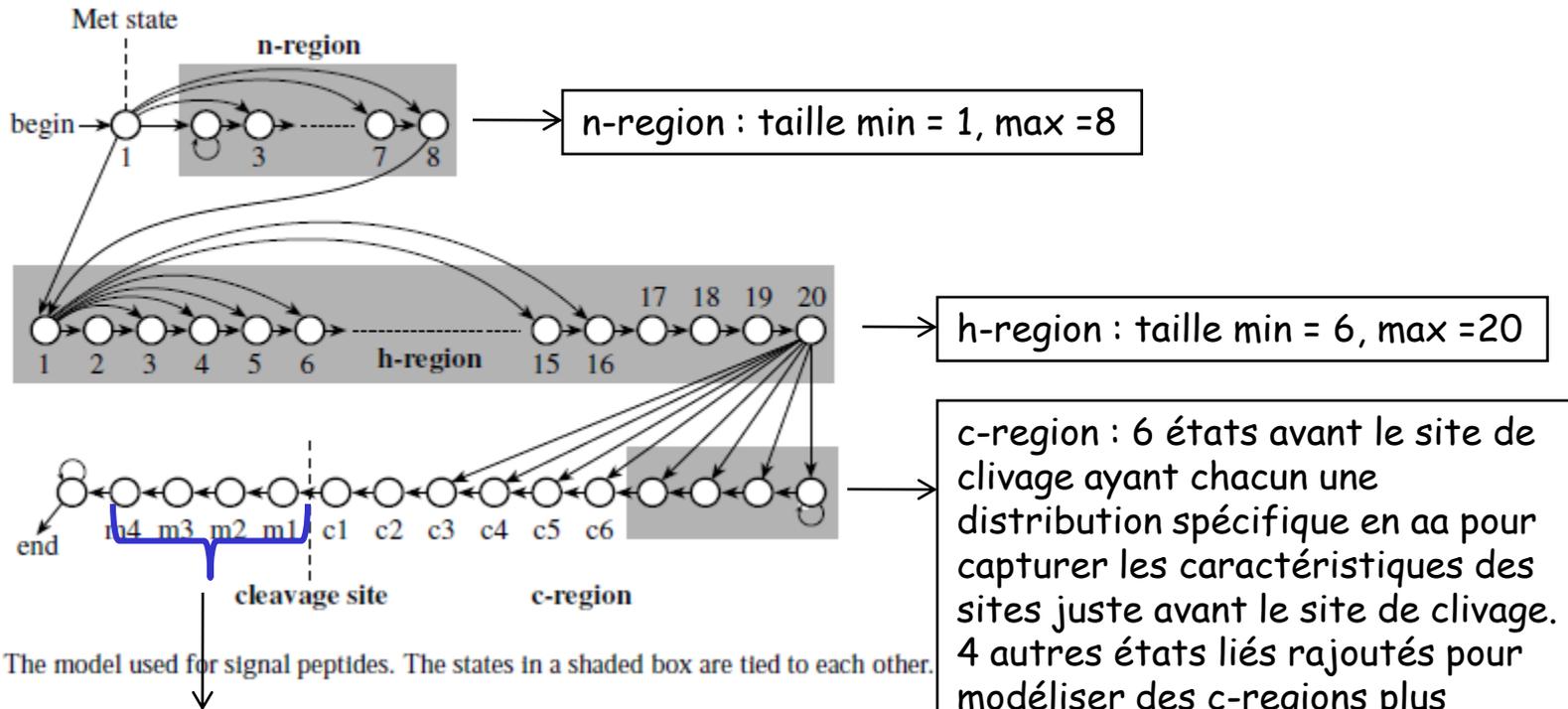


Figure 2: The model used for signal peptides. The states in a shaded box are tied to each other.

4 états modélisant les 4 premières positions de la protéine mature

(extrait de Nielsen and Krogh : **Prediction of signal peptides and signal anchors by a hidden Markov model** In J. Glasgow et al., eds., Proc. Sixth Int. Conf. on Intelligent Systems for Molecular Biology, 122-130. AAAI Press, 1998)

Problème : distinguer un peptide signal d'une hélice transmembranaire localisé en N-ter et qui peut soit être un signal d'ancrage ou pas.

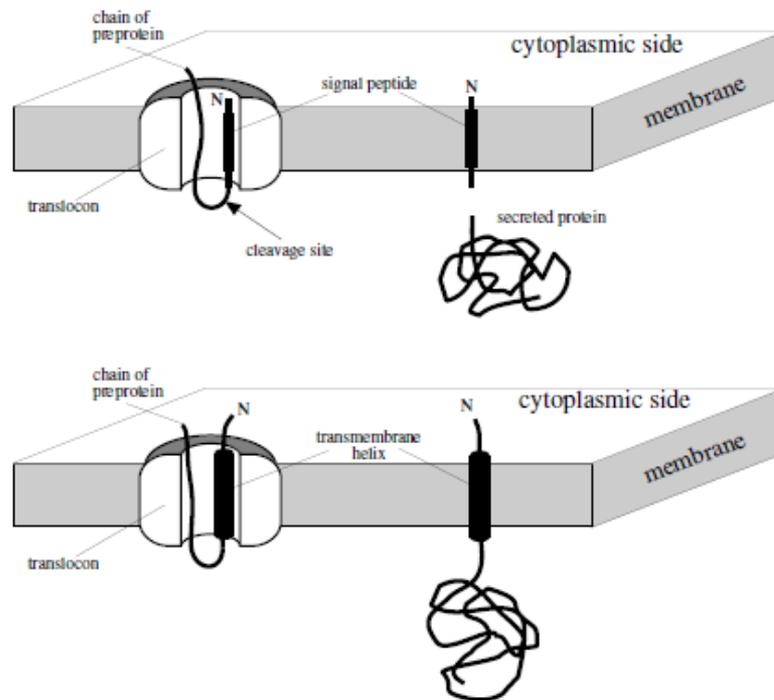


Figure 1: Cartoons of a signal peptide (above) and a signal anchor (below), and how they are translocated by the translocon. After translocation the signal peptide is cleaved off and the mature protein released, whereas the signal anchor is not cleaved off and the protein is anchored to the membrane.

(extrait de Nielsen and Krogh : **Prediction of signal peptides and signal anchors by a hidden Markov model** In J. Glasgow et al., eds., Proc. Sixth Int. Conf. on Intelligent Systems for Molecular Biology, 122-130. AAAI Press, 1998)

Plusieurs tentatives pour résoudre le problème :

- version 3 : pour séquences eucaryotes : 3 sous-modèles (signal peptide, signal anchor et autres protéines)
- version 4 : 2 sortes de données négatives :
 - ✓ premier groupe le même qu'avant protéines cytoplasmiques et nucléaires
 - ✓ deuxième groupe : séquences n'étant pas des peptides signal mais contenant un fragment transmembranaire (TM) dans les 70 premiers aa.



2 réseaux de neurones :

- ✓ SignalP-TM : entraîné avec comme données négatives le deuxième groupe (TM).
- ✓ SignalP-noTM : entraîné sur le premier groupe.

Si SignalP-TM prédit 4 positions ou plus comme étant membranaires, SignalP-TM est utilisé pour la prédiction final, sinon SignalP-noTM est utilisé.