Hybrid System NN/HMM for Large-scale GPI-anchored Protein Prediction

Guylaine Poisson, Anne Bergeron and Cedric Chauve

Laboratoire de Combinatoire et d'Informatique Mathématique
Université du Québec à Montréal
Montréal, Québec, Canada
anne|chauve|poisson@lacim.uqam.ca

Introduction

A glycosyl phosphatidyl-inositol (GPI) anchor is a C-terminal post-translational modification of proteins. Here, we investigate the problem of correctly annotating GPI-anchored protein for thegrowing number of sequences in public databases. Like many other protein sequence signatures, the biological rules that direct the primary structure of GPI-anchored protein are not exact [1], offering many challenges to automated systems of annotation. We developed a hybrid system based on the tandem use of Neural Network and Hidden Markov Model methods. The Neural Network selects thenpotential GPI-anchored protein in the dataset, and the Hidden Markov Model parses the signal and refines the prediction. The combination of the two methods reveals an interesting predicting power. The system reveals to be 93% accurate for all the GPI-anchored proteins annotated in the Swiss-Prot database. The distinctive feature of the system is that it targets only the C-terminal end of proteins, making it less sensible to the background noise found in databases. Moreover, the system is not focused on a particular taxonomic group: It can be used to predict GPI-anchored proteins in all eukaryotes (plants, animals, Fungi, protozoa etc.). Finally, by using qualitative scoring, the predictions combine both sensitivity, and information content.

Structure of GPI

Proteins linked to the membrane with a GPI-anchor are not easily identified with traditional pattern recognition approaches used in computational biology. There is an absence of constant, approximated or repetitive patterns, and similarity analysis yields poor results. However, some general rules have been identified. Besides the N-terminal signal, the C-terminal GPI signal, cleaved off at the time of the addition of the GPI-lipid anchor, can be further broken down in 4 regions [2] : 1) A unstructured linker region of about 10 residues; 2) A region of small residues, including the GPI-attachment and cleaving site; 3) A spacer region, following the cleaving site, of about 7 amino acids; 4) A hydrophobic tail next to the spacer region, completing the C-terminal end (Figure 1).

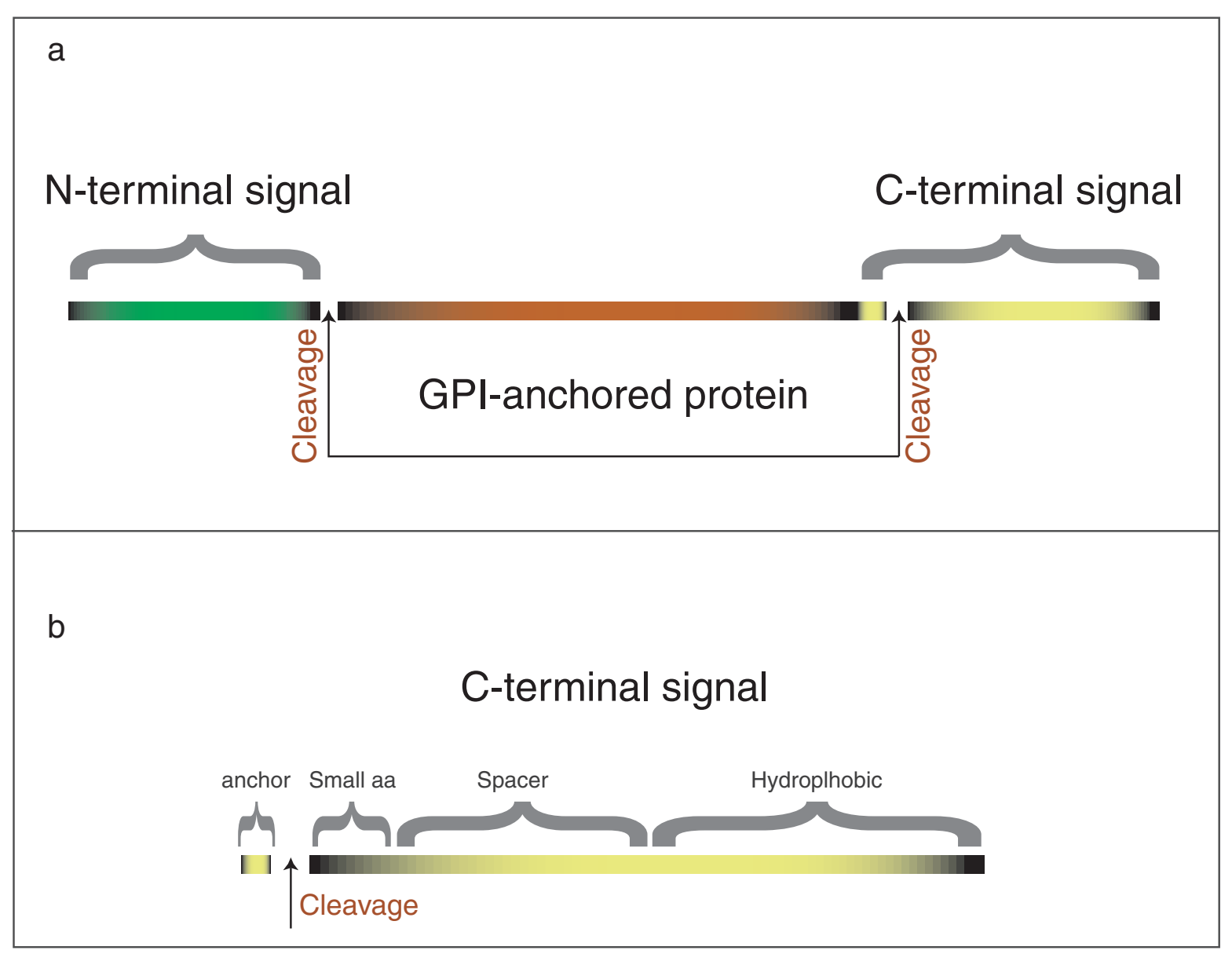

Figure 1: Structure of GPI-anchored protein. a: The protein with the 2 sequence signals before cleavage. b: The GPI signal in the C-terminal part of the protein. The anchor remains in the protein after cleavage of the signal.

Such exact rules suggest rule based approaches, but newly identified GPI-anchor proteins depart from these rules, lowering the specificity of the predictions established by the use of these methods: the spacer region and the hydrophobic tail can overlap, and the length of the spacer can vary outside the parameters.

Neural Network Predictor

Neural Networks are constructed to classify patterns through a learning process that allows to define classes boundaries in a non-parametric way. Their basic elements are artificial neurons that 1) accept numerical input from other neurons, or from the external environment, 2) process their input with a transfer function, and 3) output a value to other neurons, or backto the external environment.

The learning set for this experiment is the 50 amino acids C-terminal sections of a set of 163 sequences of SWISSPROT that are annotated as GPI-anchored proteins and 163 sequences of protein known NOT to be a GPI-anchored protein. Since neural networks accept numerical input, and given the importance of molecular weight and hydrophobicity in the anchoring process, we encode each amino acid with its hydrophobicity on the Kyte and Doolite scale, and its molecular weight.

The architecture of artificial neurons used in this study is a multilayered perceptron using the RPROP (Resilient back propagation) learning algorithm. The input layer is composed of 100 neurons, corresponding to two values for each of 50 amino acids. A hidden layer of 150 neurons encodes the classification process, and the output layer contains only one neuron, giving a score to each sequence (Figure 2).

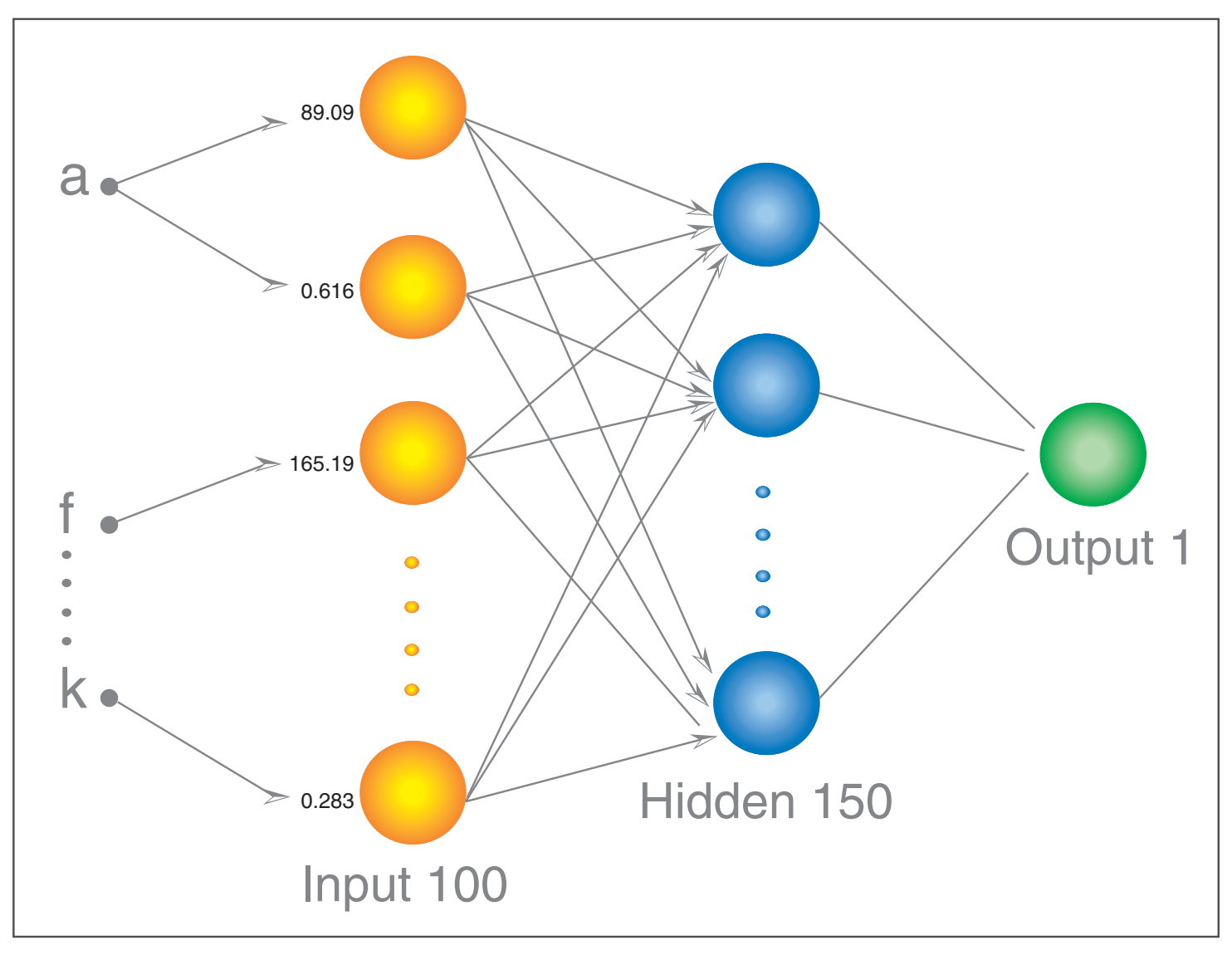

Figure 2: The multilayered perceptron neural network model. The input is a protein sequence where each amino acid is replaced by its hydropathy value and its molecular weight.

The learning process consists in gradually adjusting the weights of the processing functions in order to obtain good scores on known GPI-anchored proteins. A score greater than 0.90 indicates that the network has identified a GPI-anchored protein.

The Hidden Markov Model Predictor

A Hidden Markov Model is a stochastic automaton that is built, using machine-learning algorithms, from a set of amino acids sequences, called the training set. Given an amino acids sequence, the automaton processes it and computes a score that expresses the probability that the new sequence is similar to the ones in the training set. Moreover, the HMM is able to predict putative cleavage sites for the given sequence and to rank them according to their relevance with respect to the training set.

Basically, an HMM processes an amino acids sequence as follows: starting with an initial state, each state of the automaton processes one letter of the sequence and sends the rest of the sequence stochastically to another state. A sequence thus generates one or several path through the automaton. Each state assigns a probability to an amino acid, and the probability associated to a path is the product of these probabilities and thoses of the transitions. The log-odd of the sum of the probabilities of all possible paths for a sequence, normalized with respect to the length of the sequence, gives the score of the sequence. Based on preliminary experiments, we classify as GPI sequences having a score greater or equal to 5.

The layout of the automaton follows the rules governing the structure of GPI. Three states process the cleavage site and the next two amino acids, up to 12 states process the spacer region, and the remaining states analyze the hydrophobic tail (Figure 3).

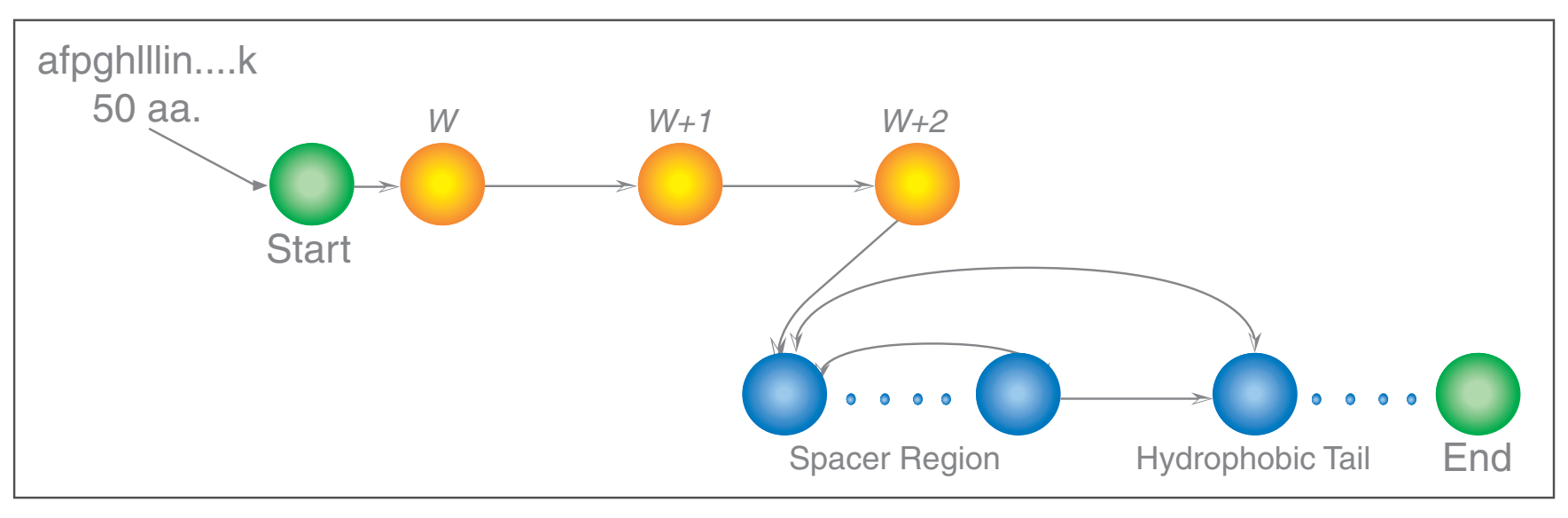

Figure 3: The Hidden Markov Model. The input is a protein sequence of at most 50 amino acids. *W* represent the anchor site position.

The main parameters of a HMM are the distributions of probabilities associated with the states and transitions. These are adjusted through a learning process in which sequences that correspond to GPI-anchored proteins are run repeatedly through the automaton in order to increase their score. The training set for the HMM was a set of 210 sequences of SWISSPROT that are annotated as GPI-anchored proteins.

In the experimental setup, each sequence was run through the automaton starting with each putative cleavage site. These putative cleavage sites are identified with a sliding window that detects groups of three amino acids of small molecular weight. This is how the HMM can also predict cleavage sites.

Hybrid System

In the hybrid system, the sequences selected by the neural network are presented to the HMM. The HMM score obtained by each sequence will next be used to annotate the prediction, based on a predefined scale. This scale ranges from "Highly probable" to "Potential false positive" (table 1).

Classes	HMM score
Highly probable	score > 5.40
Probable	5.39 > score > 2.20
Weakly probable	2.19 > score > 0.20
Potential false positive	score < 0.19

Table 1: Annotation scale based on HMM score.

This type of annotation allows us to keep the high sensibility, obtained with the neural network, together with the specificity and the capacity to structure of the HMM.

Results and Discussion

We ran eight series of comparative tests. The first four are not GPI proteins. All others are GPI-anchored proteins (Table 2).

Test File	Prediction	Highly probable	Probable	Weakly probable	Potential false positive
Cytoplasmic_Nuclear	0,01	0,0	0.0	0.0	1.0
Transmembrane	0.06	0.18	0.09	0.18	0.55
Transport_Protein	0.07	0.17	0.0	0.0	0.83
Random	0.04	0.01	0.02	0.08	0.89
Metazoa	0.92	0.84	0.09	0.03	0.04
Protozoa	0,95	0,60	0,25	0,05	0,10
Plant	0.95	0.77	0.13	0.05	0.05
Fungus	0.96	0.84	0.09	0.03	0.04

Table 2: Results of the hybrid system annotation on the Neural Network pr/dictions

Those tests reveal an interesting predicting power for the hybrid system. The mean sensibility of the system is 0.938 if we accept all prediction, and 0.876 if we want to be stricter by eliminating the sequences in the potential false positive class. In this last case, the specificity goes from 0.957 to 0.99. The HMM also gives an annotated prediction with a potential cleavage site. A test on 330 sequences with an annotated cleavage site shows that the hybrid system can correctly predict 75% of them. In the 25% falsely predicted cleavage sites, 58% had a predicted site less then 3 amino acids apart.

The test sequences were also submitted to a publicly available predictor of GPI-anchored proteins, called big-π [3] (table 3). Compared to big-π, the hybrid system is usually more sensible to the GPI-anchoring pattern. On the other hand, big-p is highly specific and can almost surely eliminate non-GPI proteins, but it misses many real GPIs.

Not GPI	big-π Prediction
Cytoplasmic_Nuclear	0,0
Transmembrane	0.0
Transport_Protein	0.0
Random	——

GPI	big-π Prediction
Metazoa	0.72
Protozoa	0,64
Plant	0.95
Fungus	0.86

Table 3: big-π test results

Some annotated GPI-anchored proteins were rejected by both predictors, and we intend to examine them more closely to establish whether they are indeed GPIs.

Finally, the tests revealed that the combination of the two machine learning approaches yields good results. The weaknesses of one are readily compensated by the strengths of the other. The hybrid system is a very general (all eukarya) tool for annotation of GPI-anchored protein on a large scale. It produces prediction with a qualitativeannotation letting the user decide the strength of annotation he wants. The less sensitive classes can contain sequences with unusual GPI-anchor signal, which can yield to new discovery in the post-translational modification research area.

Reference

[1] Eisenhaber, B., Brok, P. and Eisenhaber, F. 1999. Prediction of potentiel GPI-modification Sites in Proprotein Sequences. J. Mol. Biol 292:741-758.