FragAnchor: A Large-Scale Predictor of Glycosylphosphatidylinositol Anchors in Eukaryote Protein Sequences by Qualitative Scoring

Guylaine Poisson[1]*, Cedric Chauve[2,3,4], Xin Chen[1], and Anne Bergeron[2]

[1] *Department of Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, HI 96822, USA;* [2] *CGL & Département d'informatique, Université du Québec à Montréal, Montréal QC, Canada;* [3] *LACIM, Université du Québec à Montréal, Montréal QC, Canada;* [4] *Department of Mathematics, Simon Fraser University, Burnaby BC, Canada.*

A glycosylphosphatidylinositol (GPI) anchor is a common but complex C-terminal post-translational modification of extracellular proteins in eukaryotes. Here we investigate the problem of correctly annotating GPI-anchored proteins for the growing number of sequences in public databases. We developed a computational system, called FragAnchor, based on the tandem use of a neural network (NN) and a hidden Markov model (HMM). Firstly, NN selects potential GPI-anchored proteins in a dataset, then HMM parses these potential GPI signals and refines the prediction by qualitative scoring. FragAnchor correctly predicted 91% of all the GPI-anchored proteins annotated in the Swiss-Prot database. In a large-scale analysis of 29 eukaryote proteomes, FragAnchor predicted that the percentage of highly probable GPI-anchored proteins is between 0.21% and 2.01%. The distinctive feature of FragAnchor, compared with other systems, is that it targets only the C-terminus of a protein, making it less sensitive to the background noise found in databases and possible incomplete protein sequences. Moreover, FragAnchor can be used to predict GPI-anchored proteins in all eukaryotes. Finally, by using qualitative scoring, the predictions combine both sensitivity and information content. The predictor is publicly available at http://navet.ics.hawaii.edu/~fraganchor/NNHMM/NNHMM.html.

Key words: bioinformatics, post-translational modification, proteome analysis

Introduction

The complexity of information contained in protein sequences is a major challenge for large-scale analysis of proteomics data: some characteristics are easily identifiable, such as general hydrophobicity, whereas others are well hidden, such as the presence of short segments endowed with a specific function but having been affected by numerous mutations during their evolution. Identifying and classifying correctly these characteristics often requires novel approaches that assist standard tools of analysis. Furthermore, the structure of protein sequences translated from genes is not sufficient to assess the overall complexity of their functions. Indeed, post-translational modifications (PTMs) can spawn, for example, changes of activity, cellular localization, or protein interaction (*1*). PTMs such as glycosylphosphatidylinositol (GPI) anchors are fundamental for understanding biological functions of proteins; however, studies are suffering from a shortage of efficient methods that could allow to identify them in large-scale analyses (*2*). The prediction of PTM in protein sequences is then an integral part of an in-depth study fostering the understanding of biological functions, which turns out to be an important step in the annotation of proteomes.

Glycosylation is one of the most common and complex forms of PTMs (*3*, *4*). It is classified into three categories: N-glycosylation, O-glycosylation, and the attachment of a glycolipide (GPI) to the C-terminus of a protein. A GPI anchor is a type of membrane attachment discovered fairly recently. Its occurrences in eukaryotic cells were identified in the 1980s through the works of several researchers (*5*–*9*). Among proteins with GPI anchors one finds enzymes, adhesive proteins, receptors, activation antigens, and so on (*10*, *11*). Currently the exact function of this at-

***Corresponding author.**
E-mail: guylaine@hawaii.edu

tachment is not well characterized (*12*). Nonetheless, its conservation across great taxonomic diversity (yeasts, protozoa, plants, vertebrates, and even archeo-bacteria) suggests some important functionality (*13*). Proteins with GPI anchors reveal a very interesting feature that they are exclusively extracellular. This PTM thus provides interesting information for the annotation of new sequences by specifying their cellular localizations. Moreover, this property of GPI-anchored proteins opens the way to several potential applications. For example, in the genome of *Plasmodium falciparum* (human malaria parasite), several proteins that are attached to the membrane by a GPI anchor can be used as vaccine candidates (*14*).

Proteins linked to the membrane by a GPI anchor are not easy to identify with traditional sequence alignment and pattern recognition approaches that are used in computational biology. Indeed, there is no clear constant, approximate, or repetitive patterns, and similarity analysis yields poor results (results not shown). Nonetheless, some general rules have been identified. For example, GPI-anchored proteins have an N-terminal signal for translocation across the endoplasmic reticulum. However, we discovered that this signal is absent or not clearly predicted by computational tools in nearly 7% of the annotated GPI-anchored proteins in the Swiss-Prot database (http://www.expasy.org/sprot/). 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 into 4 regions (*15*): (1) an unstructured linker region of about 10 residues; (2) a region of small residues, including the GPI attachment and cleavage site; (3) a spacer region, following the cleavage site, of about 7 amino acids (a.a.); (4) a hydrophobic tail next to the spacer region, completing the C-terminus (Figure 1).

Such sequence features suggest that rule-based approaches should be efficient in predicting GPI-

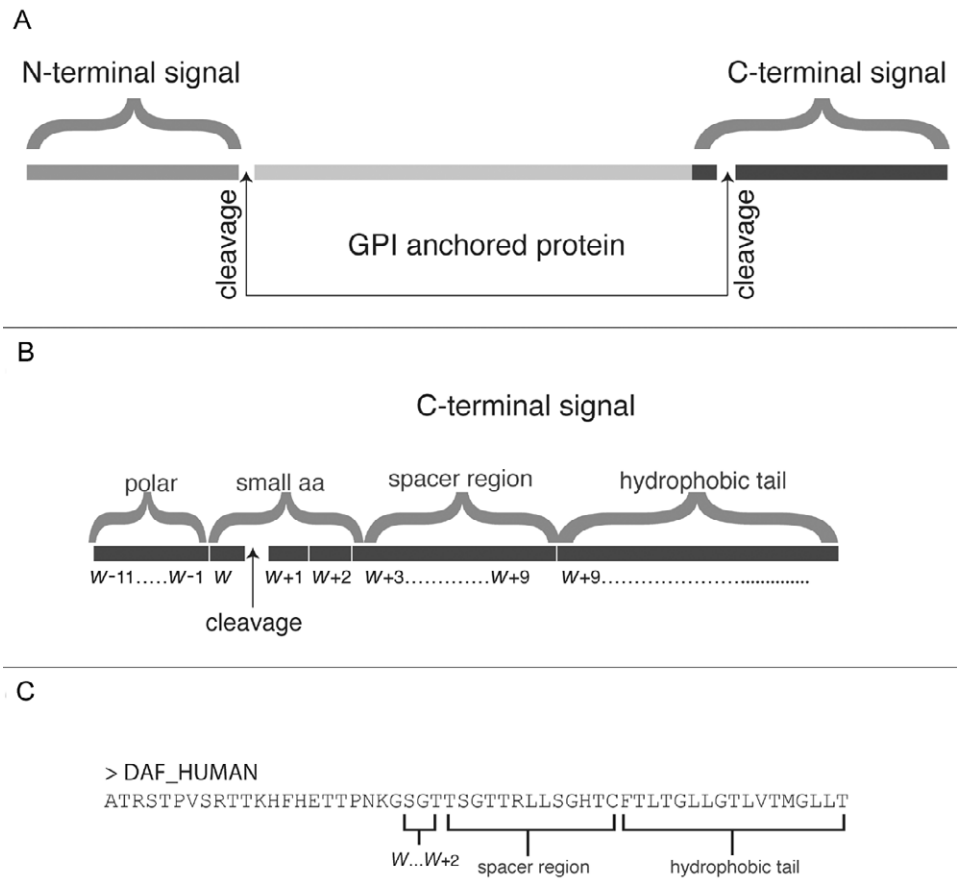

Fig. 1 Structure of GPI-anchored proteins. **A**. Signals of GPI-anchored proteins at the N-terminus and the C-terminus, respectively. **B**. The C-terminal signal can be further decomposed, from left to right, as a polar region, a sequence of three small amino acids (the anchor site is at position W), a spacer region, followed by a hydrophobic tail. **C**. Illustration of the structure with a fragment of the DAF_HUMAN protein.

anchored proteins, but there are many exceptions that newly identified GPI-anchored proteins depart from these rules (for example, unpredictable N-terminal signals, C-terminal signals with overlapping spacer region and hydrophobic tail, or length of the spacer region outside the parameters), lowering automatically the sensitivity of the predictions based on methods using such rules in a strict way.

Currently, there exist three different predictors publicly available: big-π (*15*) from the University of Vienna (available in two versions, one for metazoa and one for protozoa), DGPI from the University of Geneva (retrieved from http://129.194.185.165/dgpi/), and the large-scale annotator GPI-SOM (*16*) from the University of Bern. The first two also predict potential cleavage sites while the third does not. These tools target the C-terminal and N-terminal signals of protein sequences, or require verifying the presence of the N-terminal signal before submitting sequences to the predictor. This affects automatically their performance in the presence of fragmented or partial sequences, which happen to be more and more common in databases. In particular, metagenome projects generate many fragmented sequences representing only partial proteins, complicating the annotation process. Moreover, with metagenome projects, the inability to relate many of the new sequences to a specific taxonomic group does not favor the use of a group-specific tool like big-π. Therefore, large-scale annotation tools, which are able to predict the presence of a particular motif at different levels of precision (that is, to assess each prediction in a qualitative way) with minimal information, are then needed to address this new reality, since tools that are too restrictive or specialized often lack the flexibility needed to make correct predictions. For example, a recent analysis of the *P. falciparum* proteome shows that big-π, DGPI, and GPI-SOM generated poor results in predicting proteins selected as biologically validated or highly probable candidates (*14*). Accordingly, our system was designed to offer the precision of cleavage site prediction tools and the flexibility needed for a large-scale annotator in a noisy environment.

Results and Discussion

We developed a system, called FragAnchor, based on the tandem use of a neural network (NN) and a hidden Markov model (HMM). NN is used to select potential GPI-anchored sequences and HMM classifies the selected sequences by a qualitative scoring scheme. The sequences selected by NN are annotated as highly probable (Class 1), probable (Class 2), weakly probable (Class 3), or potential false positive (Class 4). HMM is also used to predict the position of the cleavage site in the sequence.

Firstly, NN and HMM were trained separately with different training sets. This choice was made to ensure an optimal training set for each method. For NN, a validation test was performed with 134 GPI sequences and 134 non-GPI sequences from the Swiss-Prot database release 49.0. This test showed that, for a threshold of selection set to 0.90 (the possible values for this threshold range from 0.0 to 1.0), NN had a precision of 93% and a positive correlation coefficient between prediction and observation (*17*) of 0.86, implying that the predictions are quite accurate. With this validation test set, NN was able to correctly predict 89.47% of the GPI-anchored sequences and 96.27% of the sequences that were not annotated as GPI-anchored. These results, along with the area under the receiver operating characteristic (ROC) curve, are good indicators suggesting that NN has an acceptable degree of precision and an interesting generalization power.

We also performed additional tests of NN with a dataset containing 593 sequences annotated as GPI-anchored in the Swiss-Prot database, 265 membrane transport protein sequences, as well as 111 cytoplasm and nuclear protein sequences (Table 1). The result showed that NN was able to predict 91.06% of the annotated GPI sequences in Swiss-Prot. Note that the training sequences, as well as the sequences that are very similar to some ones present in the training set, are present in this dataset and these results cannot be used to assess precisely the quality of the predictor. However, these tests indicated that NN had a specificity of 95%, which means that 5% of the predictions were false positive. For HMM alone, we performed similar tests and the results we obtained showed a higher specificity but a lower sensitivity (results not shown).

For the tandem system (see Materials and Methods), we ran a validation test by using four different sets of sequences (Table 2). The first two are positive sets (GPI-anchored proteins) while the next two are negative sets (proteins that are not GPI-anchored). The first positive set contains 121 sequences among the 134 sequences used to test NN (13 sequences used in the training set of HMM were discarded). The second positive set contains 30 sequences presented in a

Table 1 Prediction results of the neural network on different sequence sets

Sequence set	No. of sequences	Predicted GPI (%)
Cytoplasm and nuclear	111	98.20
Membrane transport	265	92.45
GPI sequences in Swiss-Prot	593	91.06

Table 2 Prediction results of the tandem system on different sequence sets

Sequence set	No. of sequences	Predicted GPI (%)	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)	Invalid (%)	Rejected by NN (%)
First positive set	121	88.43	66.94	11.57	4.13	5.79	0.00	11.57
P. falciparum	30	96.67	20.00	26.67	10.00	40.00	0.00	3.33
Cytoplasm and nuclear	1,873	3.20	0.21	0.16	0.27	2.56	0.08	96.80
Membrane transport	4,587	6.21	0.76	0.68	0.44	4.29	0.59	93.79

Table 3 Comparative prediction results between big-π, GPI-SOM, and FragAnchor

Sequence set	No. of sequences	Predicted GPI (%)				
		FragAnchor (all classes)	FragAnchor +SignalP	FragAnchor (Classes 1–3)	GPI-SOM	big-π
First positive set	121	88.43	82.64	82.64	82.64	60.33
P. falciparum	30	96.67	80.00	56.67	73.33	46.67
Cytoplasm and nuclear	1,873	3.20	0.85	0.64	1.49	0.54
Membrane transport	4,587	6.21	3.24	1.88	2.83	0.50

recent analysis of the *P. falciparum* proteome (*14*). The first negative set contains 1,873 sequences with a subcellular localization annotated as "cytoplasmic" from the human protein localization database (http://locate-human.imb.uq.edu.au/). The second negative set contains 4,587 sequences retrieved from the Swiss-Prot database by using the keywords "transmembrane", "transport", or "secreted" and are not annotated as GPI-anchored.

These tests revealed an interesting prediction power of the tandem system and more importantly a remarkable flexibility offered by the qualitative scoring. The test on the first positive set (121 sequences) revealed a sensitivity of 88.43% to the GPI anchor signal if we accept all predictions, and 82.64% if we want to be stricter by eliminating the potential false positive class. The specificity varied from 99.89% to 93.89% depending on the type of sequences and classes. HMM also gives an annotated prediction with a potential cleavage site. A test on 330 sequences with an annotated cleavage site showed that the tandem system correctly predicted 75% of them. In the 25% incorrectly predicted cleavage sites, 58% were predicted at a distance of at most three amino acids from the annotated anchor site. An important point is that many of the GPI anchor sites in the Swiss-Prot curated database are based on automated prediction.

In the test performed with 30 sequences of those proposed by Gilson *et al* (*14*) from the *P. falciparum* proteome, FragAnchor was able to predict 29 of them as GPI-anchored protein sequences (96.67%). However, it is interesting to note that 40% of these predictions were classified as potential false positive (Class 4).

The tests with the two negative sets gave very good results as 96.80% and 93.79% of the sequences were rejected by NN, respectively, while HMM classified most of the incorrectly selected sequences as potential false positives. These examples illustrate the importance and usefulness of the qualitative scoring of the predictions of FragAnchor.

The sequences of the four test sets described above were also submitted to two publicly available predictors of GPI-anchored proteins: big-π (*15*) and GPI-SOM (*16*) (Table 3). We did not compare FragAnchor with DGPI because the design of this tool makes large dataset analysis difficult. Since big-π and GPI-SOM require the use of SignalP (*18*) for the N-terminal signal peptide prediction, we included an evaluation of FragAnchor combined with SignalP to help the comparison process. However, the use of SignalP is not required before using FragAnchor. Compared with big-

π, FragAnchor is more sensitive to the GPI-anchoring pattern even if we consider only the first three classes (highly probable, probable, and weakly probable). For the first set of 121 GPI-anchored sequences, FragAnchor correctly predicted 88.43% of them (82.64% in Classes 1–3) while big-π accepted 60.33%. For the newly analyzed *P. falciparum* sequences, as pointed out in Gilson *et al* (*14*), most predictors failed to predict a significant number of sequences known to bear GPI anchor. In Table 3, we can see that FragAnchor has a very high capacity of prediction with only one sequence wrongly predicted. Even when discarding the potential false positive sequences, FragAnchor recognizes more GPI-anchored sequences than big-π does. One of the reasons of the better performance of FragAnchor is the fact that it does not need the presence of the N-terminal signal, which is not clearly detectable in 6 of the 30 *P. falciparum* sequences.

On the other hand, FragAnchor shows a lower capacity of discrimination for the negative sets when compared with big-π, which is understandable due to the high stringency of big-π. However, this high stringency is accompanied by a poor capacity of prediction for the GPI anchor signal. Nonetheless, we achieved a specificity of FragAnchor comparable to that of GPI-SOM and big-π. The specificity was computed by adding the proportions of the first three classes (Table 2), for example, for the cytoplasm and nuclear proteins, it is $0.21 + 0.16 + 0.27 = 0.64$. For the membrane transport proteins, FragAnchor gave more false positive predictions. However, for this set of proteins, it is possible that some of them will eventually be validated as real GPI-anchored proteins in laboratory experiments.

When compared with GPI-SOM, FragAnchor is more accurate. GPI-SOM predicted fewer GPI-anchored sequences from the first positive set of 121 sequences, with a prediction rate of 82.64% compared with 88.43% for FragAnchor. It also failed to predict a significant number of the *P. falciparum* sequences with only 73.33% of correctly predicted GPI-anchored sequences. For the negative sets, the use of FragAnchor combined with SignalP showed that FragAnchor achieved results close to GPI-SOM (better for the first negative set, slightly lower for the second negative set).

More interestingly, the analysis of *P. falciparum* GPI-anchored sequences showed that a large portion of the GPI-anchored sequences (40%) fell in the potential false positive prediction, proving the usefulness of keeping those potentially wrongly predicted sequences in a specific class. An analysis of the sequences rejected by GPI-SOM showed that 50% of them were found in the potential false positive class from FragAnchor. In addition, the analysis result of *P. falciparum* sequences demonstrated that 4 of the 11 biochemically validated GPI-anchored sequences did not have an N-terminal signal peptide recognized by SignalP. This fact confirms the effectiveness of an annotation not based on the prediction of this signal. In short, the comparison with other tools is not easy due to the uniqueness of the tandem method. However, it reveals an interesting power of prediction where, in a single run, FragAnchor can have a very low number of potential false positive predictions with a much higher number of GPI-anchored sequences predicted.

In Figure 2, we show an example of output from the tandem system (FragAnchor predictor). NN first selects a protein sequence and then HMM parses the sequence in order to predict a potential cleavage site. The first predicted anchor site for that sequence (5NTD_HUMAN) is at position 549. This is the position annotated in the Swiss-Prot database. In the output, the predictor proposes a maximum of three possible anchor sites, ranked by decreasing scores.

We selected 29 eukaryote proteomes available from the EBI Intergr8 portal (*19*) for large-scale analysis. The results showed that, in general, the percentage of GPI-anchored proteins in a proteome is between 0.21% and 2.01% when only the highly probable class is considered (Table 4). These results are comparable with a study performed by Eisenhaber *et al* (*20*) that estimated the proportion of the *Arabidopsis thaliana* proteome consisting of GPI-anchored proteins at 0.75%.

Conclusion

The comparison of FragAnchor with big-π and GPI-SOM shows that FragAnchor can be used at different levels of precision, allowing a more selective prediction. In addition, it offers a high level of flexibility with respect to the organism considered. Studies like the *P. falciparum* proteome analysis by Gilson *et al* (*14*), where the available predictors failed to predict the biochemically proposed new GPI-anchored proteins, show that there is a need for a more flexible and universal predictor. The flexibility and accuracy achieved by FragAnchor offer an excellent annotator for the GPI anchor signal designed for all metagenome projects.

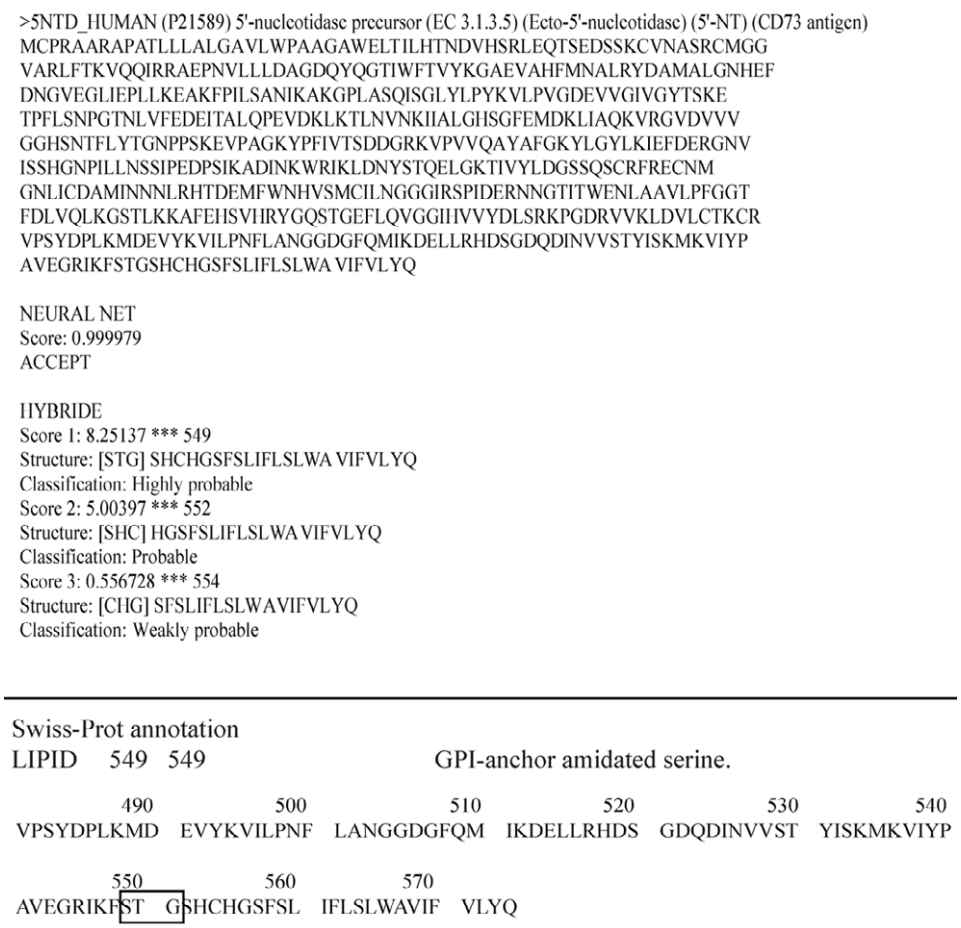

Fig. 2 Example of the output from FragAnchor and comparison to the Swiss-Prot annotation.

One of the main goals of this large-scale prediction tool is to offer greater flexibility to the users. We propose a tool that offers different levels of sensitivity to the GPI anchor signal. The tests reveal that the combination of the two machine learning approaches yields very good results. The tandem system is a very general (all eukaryotes) tool for the annotation of GPI-anchored proteins on a large scale by using minimal information. It produces prediction with a qualitative annotation allowing the user to choose the strength of precision that he/she wants. The less sensitive classes may contain sequences with unusual GPI anchor signals, which can yield to new discoveries in the PTM research area.

Materials and Methods

Neural network predictor

Neural networks are designed to classify patterns through a learning process that allows defining class boundaries in a non-parametric way. The 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; (3) output a value to other neurons or back to the external environment.

The Stuttgart Neural Network Simulator (javaNNS 1.1) developed at the University of Stuttgart (http://www-ra.informatik.uni-tuebingen.de/downloads/JavaNNS/) was used to design and train our NN predictor. The learning set for this experiment is 79 sequences from the Swiss-Prot database release 49.0 that are annotated as GPI-anchored proteins, with a length of the C-terminal section to be 50 a.a., and another 79 sequences of proteins that are known not to be GPI-anchored proteins. The choice of the length of the C-terminal section was based on an analysis of Swiss-Prot GPI-anchored proteins, which showed a maximum length of 45 a.a. for the C-terminal signal. To give some flexibility to the system, we selected an input vector of 50 a.a. Since neural networks accept numerical input, and given the importance of molecular weight and hydrophobicity

Table 4 Proportion of highly probable GPI-anchored proteins predicted by FragAnchor for 29 eukaryote proteomes

Organism	Proteome size (sequences)	Predicted GPI (Class 1) (%)
Anopheles gambiae	15,135	0.61
Arabidopsis thaliana	33,862	0.83
Ashbya gossypii	4,720	0.91
Aspergillus fumigatus	9,906	0.89
Brachydanio rerio	11,863	0.58
Caenorhabditis briggsae	13,192	0.70
Caenorhabditis elegans	22,362	0.66
Candida glabrata	5,180	1.31
Cryptococcus neoformans	6,442	0.90
C. neoformans var. *neoformans*	6,442	0.90
Debaryomyces hansenii	6,311	0.82
Dictyostelium discoideum	13,049	1.29
Drosophila melanogaster	16,302	0.79
Encephalitozoon cuniculi	1,909	0.21
Gallus gallus	11,820	0.93
Gibberella zeae	11,638	1.13
Guillardia theta	598	0.33
Homo sapiens	38,039	0.74
Kluyveromyces lactis	5,312	1.02
Leishmania major	8,010	1.02
Mus musculus	32,901	0.82
Paramecium tetraurelia	6,311	0.82
Plasmodium falciparum	5,253	0.19
Plasmodium yoelii yoelii	7,755	0.22
Rattus norvegicus	11,820	0.93
Saccharomyces cerevisiae	5,801	0.95
Schizosaccharomyces pombe	4,964	0.68
Tetraodon nigroviridis	27,810	0.55
Yarrowia lipolytica	6,525	2.01

in the anchoring process, we encoded each amino acid with its hydropathy on the Kyte and Doolitle scale (*21*), as well as its molecular weight. The last step is the normalization of the input vectors. We have applied a simple min-max normalization:

$$v' = \frac{(v - min)}{(max - min)} * (max' - min') + min'$$

where v' represents the normalized value of data v, min' and max' represent the minimal and maximal value of the target interval, while min and max are the minimal and maximal value of our actual data intervals: $[-4.5, 4.5]$ for hydropathy, and $[75.07, 204.23]$ for molecular weight.

An alignment of sequences consisting of the last 50 residues of GPI sequences was done with ClustalW (*22*). It was visually analyzed to target non-redundant sequences in order to eliminate the risk of bias toward a type of sequence too abundantly represented. We retained 79 sequences annotated as GPI-anchored from the Swiss-Prot database. In order to discriminate accurately between GPI and non-GPI sequences for the model, we retained mostly experimentally verified sequences. We also selected a set of 79 sequences that have a very low probability of being GPI-anchored (such as cytoplasm and nucleic proteins) or have similar features but without the GPI anchor (membrane transport proteins). The selection of the sequences was manually made to ensure a low level of similarity and to verify the quality of the annotation as *not* GPI-anchored sequences. The combination of these two datasets defines our training set.

To validate the model, we used a test set composed of 134 GPI-anchored sequences. This set features few redundant sequences and is disjoint from the training set. We also selected 134 sequences with a very low probability of being GPI-anchored.

Our validation set of 268 sequences does not represent the complexity of actual sequence databases. Therefore, we selected in the Swiss-Prot database the following types of sequences:

(1) Sequences that have structures or physicochemical properties relatively similar to GPI-anchored and potential false positive proteins. We selected 265 membrane transport proteins with similar hydrophobicity characteristics, making them good candidates for discrimination purpose.

(2) Sequences with a very low probability of being GPI-anchored. GPI-anchored proteins are exclusively extracellular, so we selected 111 cytoplasmic protein sequences based on their subcellular localization.

The architecture of the neural network used in FragAnchor is a multilayered perceptron using the RPROP (Resilient Back Propagation) learning algorithm (*23*). The input layer is composed of 100 neurons, corresponding to 2 input values for each of 50 amino acids. A hidden layer of 150 neurons encodes the classification process, and the output layer contains 1 neuron, giving a score to each sequence. The number of neurons in the hidden layer is the optimal architecture selected over different test architectures (results not shown).

The learning process consisted 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 potential GPI-anchored protein.

Hidden Markov model predictor

HMM is a probabilistic model equivalent to a regular grammar that is built by using machine-learning algorithms, from a set of sequences (amino acid sequences in this study) called the training set. Given an amino acid sequence, the automaton reads it and computes the following data: (1) a score expressing the likelihood that a new sequence is similar to the ones in the training set; (2) a most likely path in the automaton that highlights possible features of the structure of this sequence (*24*).

Based on the known structure of the GPI signal (Figure 1), we designed our HMM predictor following the approach used in SignalP (*18*). The underlying graph of HMM is composed of three parts, corresponding to the GPI signal features (Figure 3). The score associated to a sequence is the log-score of the likelihood, normalized by the length of the sequence.

The parameters for HMM were estimated by 100 iterations of the Baum-Welch algorithm (*25*). Pseudocounts were used for the two parts of HMM corresponding to the spacer region and the hydrophobic tail. For each of these regions, the distribution of the amino acids after the initial training in the given region was used to correct for the rare states that had a null emission probability for some amino acids. The training set for HMM was a set of 87 sequences from Swiss-Prot that are annotated as GPI-anchored proteins. The model was validated using 500 bootstrap replications on our training set. For each replication, a dataset composed of 87 sequences (repetitions allowed) was generated. These models were trained and

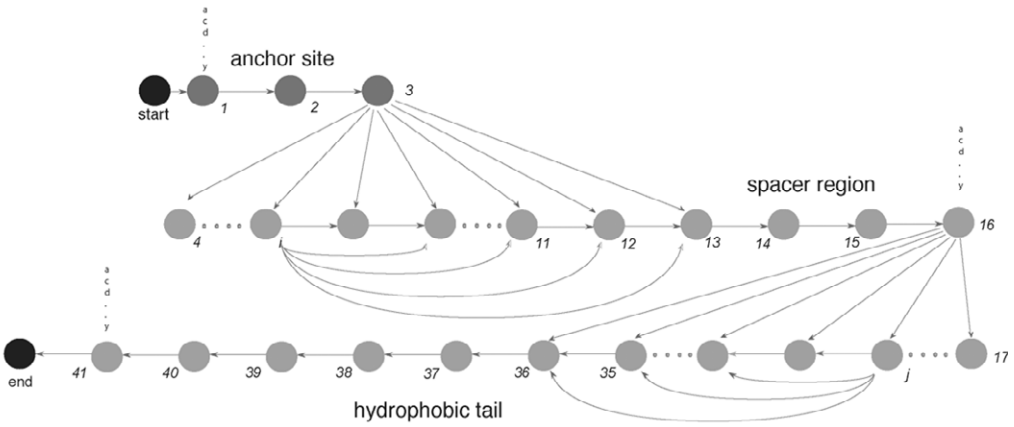

Fig. 3 The hidden Markov model for the GPI anchor signal representation. The first three states model the three amino acids around the anchor site. From the third state, transitions are possible to states 4 to 13, modeling the variable length spacer region. Finally, the hydrophobic tail, also a variable length region, is modeled by states 17 to 41.

tested using a validation set of 66 sequences.

In the experimental setup, each analyzed sequence was run through HMM with each putative cleavage site. These putative cleavage sites were identified with a sliding window that detects groups of three amino acids of small molecular weight. This is how HMM can also predict cleavage sites. The Forward algorithm computed the likelihood of a sequence, and the Viterbi algorithm (*26*) computed the most probable path for each sequence. The three best, in terms of scores, possible cleavage sites were recorded.

Tandem system

In the tandem system (Figure 4), the sequences selected by NN (the ones with a score greater than 0.90) are presented to HMM. The HMM score obtained on each sequence is then used to classify the sequence. This classification ranges from "highly probable" to "potential false positive" (Table 5). Each class of annotation was defined using a graphical representation of false-positive and true-positive predictions at all discrimination threshold, the ROC curve. This qualitative annotation allows us to keep the high sensitivity (proportion of all true positives correctly predicted) obtained with NN, together with the specificity (proportion of all true negatives correctly predicted) and the capacity of HMM to identify some features of the GPI signal, such as the cleavage site. The major advantage of using the qualitative scoring in a large-scale annotator is its fast and easy identification of the best predictions and the possible false positives. NN can detect the GPI anchor motif efficiently but its architecture does not give us any information about the primary structure of the motif, while the use of HMM provides this information. Consequently, we can refine the prediction depending on the quality of the signal present in the protein sequence.

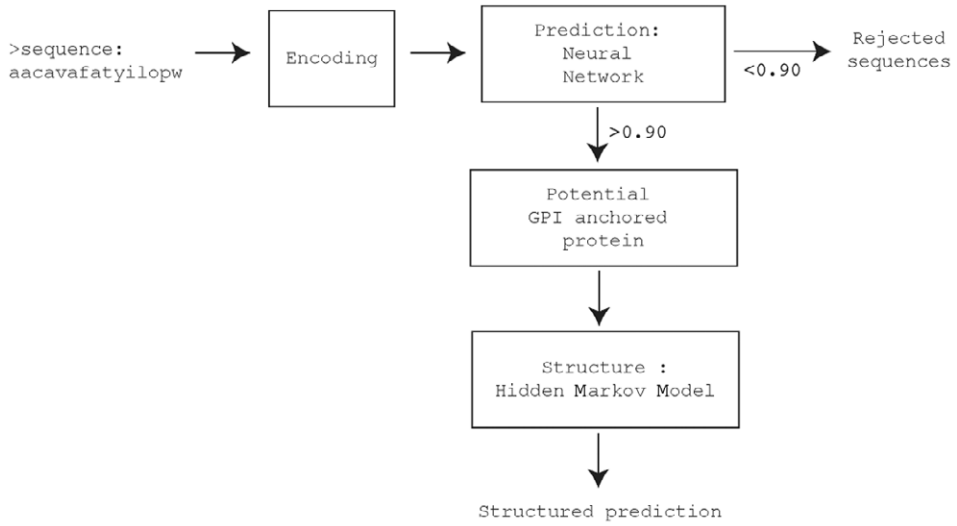

Fig. 4 The structure of the tandem system.

Table 5 Qualitative annotation of the tandem system

Category	Class	HMM score
Probable	highly probable (Class 1)	score $\geq$ 5.40
	probable (Class 2)	2.20 $\leq$ score $<$ 5.40
	weakly probable (Class 3)	0.20 $\leq$ score $<$ 2.20
Potential	potential false positive (Class 4)	score $<$ 0.20

Acknowledgements

We thank Fathey Sarhan, Mounir Boukadoum, Mathieu Blanchette, Pierre Poirier, and Marc LePape for reviewing and providing useful comments and suggestions. This work was partly supported by the National Institutes of Health of USA (Grant No. 5 P20 RR16467) and by the GP/PhD scholarship from the FCAR (now NATEQ) fund of Canada.

Authors' contributions

GP conceived of the study, participated in its design, performed the analysis for the neural network and the tandem method, and drafted the manuscript. CC participated in the design and coordination of the study, performed the analysis for the hidden Markov model, and also drafted the manuscript. XC performed the automation of the method and designed the web site. AB participated in the design and coordination of the study and also drafted the manuscript. All authors contributed to the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Seo, J. and Lee, K.J. 2004. Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. *J. Biochem. Mol. Biol.* 37: 35-44.
2. Mann, M. and Jensen, O.N. 2003. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* 21: 255-261.
3. Nalivaeva, N.N. and Turner, A.J. 2001. Post-translational modifications of proteins: acetylcholinesterase as a model system. *Proteomics* 1: 735-747.
4. Spiro, R.G. 2002. Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bounds. *Glycobiology* 12: 43R-56R.
5. Futerman, A.H., *et al.* 1985. Identification of covalently bound inositol in the hydrophobic membrane-anchoring domain of Torpedo acelylcholinesterase. *Biochem. Biophys. Res. Commun.* 129: 312-317.
6. Roberts, W.L. and Rosenberry, T.L. 1985. Identification of covalently attached fatty acids in the hydrophobic membrane-binding domain of human erythrocyte acetylcholinesterase. *Biochem. Biophys. Res. Commun.* 133: 621-627.
7. Tse, A.G., *et al.* 1985. A glycophospholipid tail at the carboxyl terminus of the Thy-1 glycoprotein of neurons and thymocytes. *Science* 230: 1003-1008.
8. Ferguson, M.A., *et al.* 1985. *Trypanosoma brucei* variant surface glycoprotein has a sn-1,2-dimyristyl glycerol membrane anchor at its COOH terminus. *J. Biol. Chem.* 260: 4963-4968.
9. Ferguson, M.A., *et al.* 1988. Glycosylphosphatidylinositol moiety that anchors *Trypanosoma brucei* variant surface glycoprotein to the membrane. *Science* 239: 753-759.
10. Chatterjee, S. and Mayor, S. 2001. The GPI-anchor and protein sorting. *Cell. Mol. Life Sci.* 58: 1969-1987.
11. Hooper, N.M. 2001. Determination of glycosylphosphatidylinositol membrane protein anchorage. *Proteomics* 1: 748-755.
12. Ikezawa, H. 2002. Glycosylphophatidylinositol (GPI)-anchored proteins. *Biol. Pharm. Bull.* 25: 409-417.
13. Low, M.G., *et al.* 1999. GPI-anchored biomolecules—an overview. In *GPI-Anchored Membrane Proteins and Carbohydrates* (eds. Hoessli, D.C. and Ilangumaran, S.), pp. 1-14. Landes, Austin, USA.
14. Gilson, P.R., *et al.* 2006. Identification and stoichiometry of glycosylphosphatidylinositol-anchored membrane proteins of the human malaria parasite *Plasmodium falciparum*. *Mol. Cell. Proteomics* 5: 1286-1299.
15. Eisenhaber, B., *et al.* 1999. Prediction of potential GPI-modification sites in proprotein sequences. *J. Mol. Biol.* 292: 741-758.
16. Fankhauser, N. and Mäser, P. 2005. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics* 21: 1846-1852.
17. Mattews, B.W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405: 442-451.
18. Nielsen, H. and Krogh, A. 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6: 122-130.
19. Kersey, P., *et al.* 2005. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.* 33: D297-302.
20. Eisenhaber, B., *et al.* 2003. Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for *Arabidopsis* and rice. *Plant Physiol.* 133: 1691-1701.
21. Kyte, J. and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157: 105-132.
22. Thompson, J.D., *et al.* 1994. CLUSTALW: improving the sensibility of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.
23. Riedmiller, M. and Braun, H. 1992. RPROP—a fast adaptive learning algorithm. In *Proceedings of the Seventh International Symposium on Computer and Information Sciences*, pp. 279-285. Antalya, Turkey.
24. Baldi, P. and Brunak, S. 2001. *Bioinformatics: The Machine Learning Approach* (second edition). MIT Press, Cambridge, USA.
25. Baum, L.E., *et al.* 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* 41: 164-171.
26. Viterbi, A.J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory* 13: 260-269.