BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 9 2005, pages 1846–1852
doi:10.1093/bioinformatics/bti299

Sequence analysis

Identification of GPI anchor attachment signals by a Kohonen self-organizing map

Niklaus Fankhauser and Pascal Mäser*

Institute of Cell Biology, University of Bern, CH-3012 Bern, Switzerland

Received on September 29, 2004; revised on January 25, 2005; accepted on January 27, 2005
Advance Access publication February 2, 2005

ABSTRACT

Motivation: Anchoring of proteins to the extracytosolic leaflet of membranes via C-terminal attachment of glycosylphosphatidylinositol (GPI) is ubiquitous and essential in eukaryotes. The signal for GPI-anchoring is confined to the C-terminus of the target protein. In order to identify anchoring signals *in silico*, we have trained neural networks on known GPI-anchored proteins, systematically optimizing input parameters.

Results: A Kohonen self-organizing map, GPI-SOM, was developed that predicts GPI-anchored proteins with high accuracy. In combination with SignalP, GPI-SOM was used in genome-wide surveys for GPI-anchored proteins in diverse eukaryotes. Apart from specialized parasites, a general trend towards higher percentages of GPI-anchored proteins in larger proteomes was observed.

Availability: GPI-SOM is accessible on-line at http://gpi.unibe.ch. The source code (written in C) is available on the same website.

Contact: pascal.maeser@izb.unibe.ch

Supplementary information: Positive training set, performance test sets and lists of predicted GPI-anchored proteins from different eukaryotes in fasta format.

INTRODUCTION

Anchoring of proteins to the extracellular surface of the plasma membrane via glycosylphosphatidylinositol (GPI) is widespread among eukaryotes. GPI-anchored proteins range from small peptides to large antigens and fulfill a variety of cellular functions. Some are receptors for external signals, e.g. Nogo receptor (Fournier *et al*., 2001) or Trail decoy receptors (Sheridan *et al*., 1997), others for nutrients such as the folate receptor (Lacey *et al*., 1989). Extracellular proteases and other enzymes may be GPI-anchored (Netzel-Arnett *et al*., 2003). Structural surface proteins with a GPI anchor are of particular importance as antigens of eukaryotic parasites (Ferguson, 1999). There are also GPI-anchored proteins of unknown function, such as the prion protein involved in bovine spongiform encephalopathy (Stahl *et al*., 1987). GPI-anchoring is essential for cell function and development, indicated by the fact that null mutations in GPI synthesis are lethal to the yeast *Saccharomyces cerevisiae* (Hamburger *et al*., 1995; Sutterlin *et al*., 1998). Mice lacking GPI synthesis fail in their development at early embryonic stages (Nozaki *et al*., 1999).

Proteins destined to receive a GPI-anchor carry a C-terminal signal sequence. This sequence is sufficient for GPI-anchor attachment, as has been demonstrated by gene fusion experiments (Caras *et al*., 1987). Furthermore, heterologous expression systems revealed that the GPI-anchor attachment signal is generally recognized across eukaryotic kingdoms, though not necessarily in all instances (Moran and Caras, 1994; Meyer *et al*., 2002). Signal sequences were functional from *Pneumocystis carinii* in COS cells (Guadiz *et al*., 1998), from *Homo sapiens* in *Trypanosoma brucei* (Butikofer *et al*., 1999) and from rat in *Pichia pastoris* (Morel and Massoulie, 1997). However, the C-termini from known GPI-anchored proteins cannot be aligned to a consensus sequence. The GPI anchor attachment signal is cleaved during protein processing and the preassembled GPI core structure is covalently attached to the new C-terminus of the target protein, termed omega (ω) site (Takeda and Kinoshita, 1995). Since these reactions take place in the lumen of the endoplasmic reticulum (ER), a C-terminal GPI anchor-attachment signal only makes sense in the context of an N-terminal export sequence. The canonical tool for prediction of the latter type of signal is SignalP, a program that uses hidden Markov models and a neural network (Nielsen *et al*., 1997). Two programs are available for computational prediction of C-terminal GPI-anchoring signals, Big-PI (Eisenhaber *et al*., 1999, http://mendel.imp.univie.ac.at/sat/gpi/gpi_server.html) and DGPI (Kronegg and Buloz, 1999, http://129.194.185.165/dgpi/). Both are based on the amino acid composition around the ω site (Udenfriend and Kodukula, 1995; Eisenhaber *et al*., 1998). Such programs are most useful when predicting the ω site of proteins known to be GPI-anchored. For screening of unknown proteins, however, it is difficult to balance between false positive and false negative errors. Big-PI now exists in kingdom-specific flavors (http://mendel.imp.univie.ac.at/gpi/gpi_server.html for animals or protozoa, http://mendel.imp.univie.ac.at/gpi/fungi_server.html for fungi, http://mendel.imp.univie.ac.at/gpi/plant_server.html for plants).

Neural networks of the Kohonen type, also termed self-organizing maps (SOMs), are powerful tools for classification of hidden information in large datasets (Kohonen, 2001). As with classical feed-forward networks, learning in SOMs happens by adjusting the weights of the connections (synapses) between units (neurons). But in contrast to feed-forward nets, SOMs learn in an unsupervised manner, guaranteeing minimal bias from the investigator. Thus SOMs will distinguish patterns without knowing if and how many different patterns the input contains. Furthermore, SOM output can easily be visualized as a two-dimensional map. Biological applications range from clustering of microarray data (Toronen *et al*., 1999) to analysis of whale songs (Murray *et al*., 1998). SOMs have successfully been

*To whom correspondence should be addressed.

 © The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

source: https://doi.org/10.7892/boris.115848 | downloaded: 12.10.2024

applied for classification of DNA sequences based on codon usage (Kanaya *et al*., 2001) (Supek and Vlahovicek, 2004), nucleotide frequencies (Abe *et al*., 2003), or virtual potentials (Aires-de-Sousa and Aires-de-Sousa, 2003). SOM analysis of protein sequences was carried out using bipeptide composition as input (Ferran and Ferrara, 1992; Ferran *et al*., 1994).

Encouraged by the facts that the GPI anchor attachment signal (1) carries universal features and (2) is confined to the C-terminus of the target protein, we implemented neural network approaches for identification of GPI-anchoring signals. Here, we present a case study for development and systematic optimization of a SOM that recognizes GPI-anchored proteins from diverse eukaryotes.

SYSTEMS AND METHODS

Hardware

The University of Bern Linux cluster Ubelix (http://ubelix.unibe.ch) was used for running multiple experiments in parallel in order to optimize network architecture and input parameters. The final program GPI-SOM and its web interface (http://gpi.unibe.ch) are running on an AMD64 gentoo Linux server.

Neural networks

All neural networks were implemented with the artificial neural network library (ANNLIB) (A.Hoekstra, M.A.Kraaijveld, D.de Ridder, W.F.Schmidt, Pattern Recognition Group, Delft University of Technology) and written in C. PNG image files of two-dimensional maps were generated using the GD graphics library (http://www.Boutell.com). The web interface was written in Perl-cgi.

Training and evaluation sets

The positive training and evaluation sets consisted of proteins that had been experimentally shown to be GPI-anchored. These included 110 proteins of all four eukaryote kingdoms selected via Entrez from GenBank, supplemented with a set of 248 GPI-proteins from *Arabidopsis thaliana*, kindly provided by P.Dupree, University of Cambridge (Borner *et al*., 2003). The positive test sets for Table 2 were (e) a list of GPI-anchored proteins downloaded from the website of B.Eisenhaber, University of Vienna (http://mendel.imp.univie.ac.at/gpi/gpi.p/gpi.swp), excluding those already present in our positive training and validation sets, and (f) recently published, experimentally verified GPI-anchored proteins that none of the tested programs had encountered before.

The negative training and evaluation sets consisted of 256 known cytosolic and 128 transmembrane proteins of all eukaryote kingdoms, 25 of which had a transmembrane domain near their C-terminus. The negative test sets for Table 2 were selected from GenBank by text-based searches. For the set N–TM–C, only transmembrane proteins with an N-terminal export signal predicted by SignalP as well as a hydrophobic C-terminus were selected. All protein sets were homology-reduced with a Perl script that uses the Smith/Waterman algorithm (Smith and Waterman, 1981) to find any two sequences that have an alignment score above a certain percentage of the shorter sequence's selfmatch score. The shorter sequence will be removed in order to create a set of non-homologous proteins. The threshold for sequence removal was set to 50% for the negative set and 80% for the C-terminal 32 amino acids of GPI anchored proteins.

Random sequences between 80 and 400 amino acids in length (random distribution) were generated based on the amino acid frequencies of the predicted *S.cerevisiae* proteome (A, 0.055; C, 0.013; D, 0.058; E, 0.065; F, 0.045; G, 0.050; H, 0.022; I, 0.066; K, 0.073; L, 0.096; M, 0.021; N, 0.061; P, 0.043; Q, 0.039; R, 0.045; S, 0.090; T, 0.059; V, 0.056; W, 0.010; Y, 0.034), with a Perl script utilizing random numbers from http://random.org.

Table 1. Selected formats of sequence representation, their corresponding numbers of input residues (AAs), numbers of cells in the input layer and their performance as indicated by validation error (FP, false positives; FN, false negatives) of feed forward networks trained by back-propagation

Interface	AAs	Input cells	FN (%)	FP (%)
2D	32	640	3.1	3.2
H	32	32	4.7	12
VP	32	20	13	15
VP + H	32	52	1.6	7.2
Z	32	20	3.1	6.4
$Z + H$	32	52	3.9	3.2
$Z + H + \omega$	32	54	3.1	2.4
$Z + H + \omega$	22	44	3.1	1.6

Input elements: 2D, two-dimensional interface; H, hydrophobicity; VP, virtual potential; Z, zentriole; ω, omega site.

Proteome files

Predicted proteins from completely sequenced eukaryotic genomes were obtained from ftp.ebi.ac.uk (*A.thaliana*, *Drosophila melanogaster*, *S.cerevisiae*, *Schizosaccharomyces pombe*), ftp.ensembl.org (*Caenorhabditis elegans*, *H.sapiens*, *Anopheles gambiae*), ftp.ncbi.nlm.nih.gov (*Encephalitozoon cuniculi*, *Mus musculus*), ftp.sanger.ac.uk (*T.brucei* chromosome 1), ftp.tigr.org (*T.brucei* chromosome 2), and www.plasmodb.org (*Plasmodium falciparum*).

ALGORITHMS

Network architecture and training

Pilot experiments for optimizing input parameters were run as feed-forward networks for sake of speed. These networks contained variable numbers of input units (depending on input format; Table 1), one hidden layer of 10 units, a second one of 5, and 2 units in the output layer. These networks were trained by back-propagation with a constant learning rate of 0.001 (a gradually decreasing learning rate was tried out but did not perform better). The weights of all connections were initially set at random. After each round of training, all weights were updated by back-propagation and saved to a separate file. After 5000 rounds, weight values yielding minimal validation error were restored to avoid over-training of the network (i.e. minimizing training error at the cost of validation error; Kohonen, 2001). Protein sets had been split 2:1 training to validation.

Kohonen SOMs were also trained for 5000 rounds starting from random weights, but updating of weights was restricted to the winning unit and its neighbors (radius scaled by the Gaussian function of distance). After each cycle, the winning units were determined for the validation sets and the number of units responding to sequences from both positive and negative sets was taken as a negative measure of quality. The map was saved only when the number of such undecided units was lower than in any previous step. Thus, upon completion of training, the network had been stored optimized with respect to validation. For visual evaluation, every unit was represented as a colored square according to class and intensity representing how often a particular unit had won.

Sequence representation

A number of different input formats were investigated (see Implementation section). Virtual potentials (VP) for amino acids were

calculated in analogy to the formula proposed for DNA sequences (Aires-de-Sousa and Aires-de-Sousa, 2003). The VP at the C-terminal position of a preceding window sized 32 was used as input. For three occurrences of amino acid A at positions p_{A1}, p_{A2}, p_{A3}, the VP equals $((p_{A1}{}^{-1} + p_{A2})^{-1} + p_{A3})^{-1}$, where p counts upwards from 1 starting at distance 32 from the C-terminus. The zentriole Z of a given amino acid A represents its average position weighed by its proximity to the C-terminus. For three occurrences of A at positions p_{A1}, p_{A2}, p_{A3} counted upwards from 1 starting at distance 32 from the C-terminus, Z was defined as $((p_{A1}/2 + p_{A2})/2 + p_{A3})/2$, which generalizes to

$$Z(\mathrm{A}) = 2^{-n} \sum_{i=1}^{n} 2^{i-1} p_{\mathrm{A}i}. \quad (1)$$

For amino acids not occurring in the input sequence, Z equals zero. The quality of a putative omega site was assessed by a scoring matrix for the triplet $\omega, \omega + 1, \omega + 2$, based on known ω sites (Gerber *et al*., 1992; Kodukula *et al*., 1993; Udenfriend and Kodukula, 1995; Eisenhaber *et al*., 1998). Top scores were attributed to serine followed by alanine and glycine. Hydrophobicity scores of amino acids were derived from Kyte and Doolittle (1982).

Automated filling up of the map

Empty units in a SOM that had not been hit during training were classified according to their surroundings. Scores for GPI and non-GPI of all units within a radius of three around the empty one were multiplied with a distance factor (3, 1.5, or 1 beginning with the innermost layer) and summed up. If the difference between the two sums was >1, the unit was assigned to the higher-scoring class; otherwise it was left undecided.

IMPLEMENTATION

Optimizing sequence representation

Transformation of biological sequence data into a form that can be read by the input layer of a neural network inevitably causes a substantial loss of information, since it is not practicable to express molecular structure in numbers. We have evaluated different numerical representation formats of amino acid sequences for identification of GPI proteins from their 32 C-terminal residues. Beginning with collinear versions, where input neurons directly represent individual amino acid positions, a two-dimensional interface of 20 binary input units for each of the 32 positions was tried. The resulting network performed with an accuracy of ~97%, but it was impractical because of the large amounts of data and long computation times (Table 1). Computation was accelerated by representing each position with a single unit instead of twenty; in that case, amino acids were substituted by their relative hydrophobicity (Kyte and Doolittle, 1982). However, this increased the number of wrong predictions, particularly false positive ones (Table 1). Addition of an input unit for the local alignment score to a reference GPI signal sequence (the last 31 amino acids of pig renal dipeptidase, GenBank accession P22412) did not reduce validation errors (not shown).

Virtual potentials have been used for positional transformation of DNA sequences (Aires-de-Sousa and Aires-de-Sousa, 2003). We have adapted this concept to amino acids. This transformation obviously reduced input size and computation time compared to collinear representations, but resulted in only ~85% correct predictions.

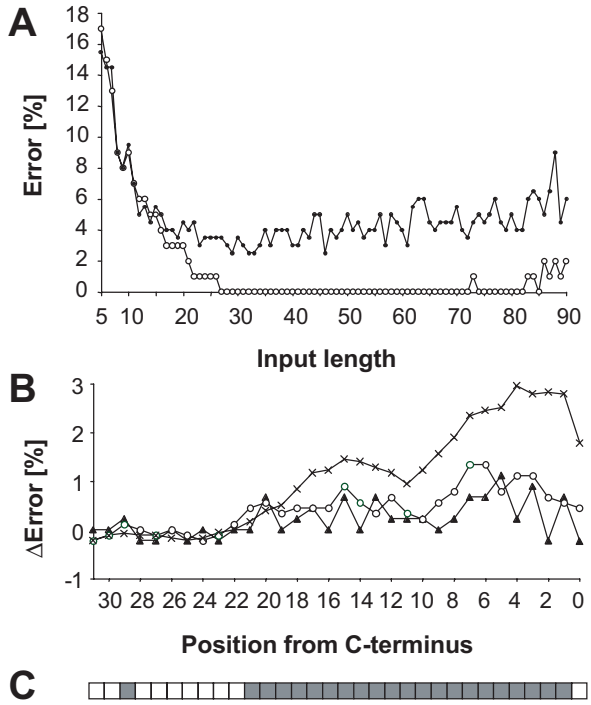

Fig. 1. Selection of input residues from the C-terminus with feed-forward networks. (**A**) Prediction accuracy in function of input length. The average percentage of false positives and false negative predictions on the training sets (white circles) and validation sets (black circles) is plotted against the number of amino acids counted from the C-terminus. Validation error was minimal at an input size of 29. (**B**) Simulated mutagenesis of the presumed signal. Single positions (black triangles), pairs (white circles), or groups of four amino acids (crosses) were masked sequentially and the performance of the network was evaluated as average of positive and negative validation errors between masked and original input sequence. (**C**) 22 important positions (filled squares) were used as input for the Kohonen map GPI-SOM.

These high error rates were, however, substantially reduced by the addition of input units for relative hydrophobicity (H) at each position (Table 1). Thus the combination of a positionally transformed parameter (VP) for each of the 20 amino acids with a collinear representation (H) for each position of the C-terminus appeared to be a suitable input format for recognizing GPI-anchored proteins, while neither VP nor H alone performed well. Related to the virtual potential is the concept of the zentriole (Z), a C-proximally weighed average position (described under Algorithms). Already by itself, the zentriole input format performed promisingly well and combined with hydrophobicity values of each position, it achieved minimal error rates. Further studies and optimization were, therefore, carried out with this type of input vector ($Z + H$).

Narrowing down on the signal sequence

In order to streamline input data in respect to signal recognition, a fast feed-forward network was repeatedly trained and evaluated with C-terminal fragments from the GPI positive sets, each time increasing the length of input sequences by one (Fig. 1A). Initially, both training error and validation error decreased with increasing length of input sequence, reaching a minimum at 29 amino acids.

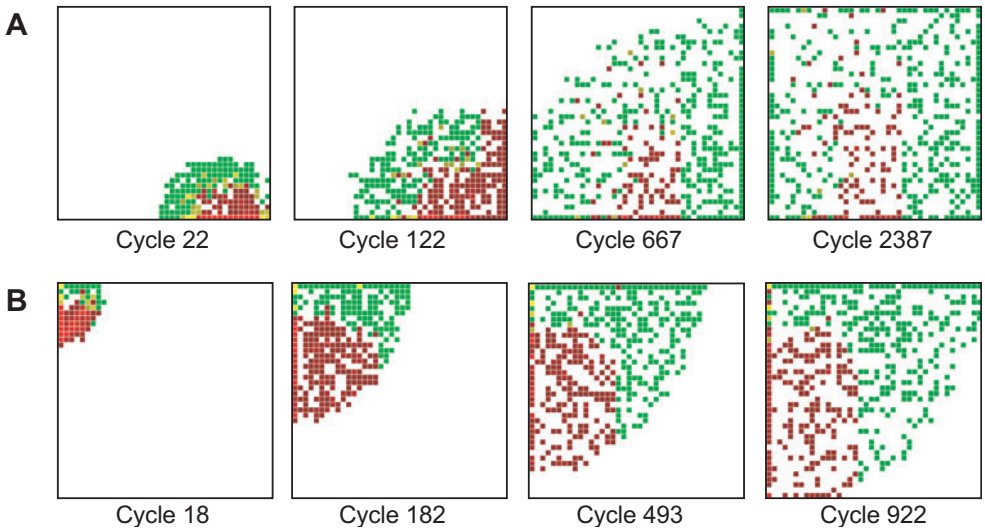

Fig. 2. Unsupervised learning by self-organizing Kohonen maps. Using the input residues outlined in Figure 1C, Kohonen SOMs of size 40 × 40 were trained with different amino acid representation formats. (**A**) Representing each residue with its Kyte-Doolittle hydrophobicity resulted in poor separation and slow convergence of self-organization. (**B**) Hydrophobicity of each residue combined with the zentriole Z for each amino acid performed much better. Clear and fast separation of GPI-proteins and non-GPI proteins was observed. Note that coloring took place after training; during self-organization, the SOM is not told which sequences are GPI and which are not. Color intensity indicates how often a particular unit was activated (green, GPI; red, non-GPI; yellow, activated by members of either set).

From 32 residues upwards, however, the validation error rose again, indicative of excessive information. Therefore only the 32 C-terminal amino acids were selected as input for further analyses.

By performing an *in silico* mutagenesis experiment, we investigated which of the 32 C-terminal residues best distinguished a GPI-anchored protein as such. A sliding window that represented any amino acid as 'X' was used to mask each position in turn (X was assigned the hydrophobicity of alanine). As expected, prediction accuracy decreased with increasing window size (Fig. 1B). The amino acids far from the C-terminus were, with a few exceptions, less significant than the ones near it (Fig. 1B). Based on these data, positions to be presented to the network were selected and the most efficient combination was identified empirically. It was an input vector of 22 residues (Fig. 1C) which, when fed into the network, performed even better than the vector of all 32 C-terminal amino acids (Table 1).

The most frequent source of false positives were integral membrane proteins with a transmembrane domain within the last 30 amino acids. In order to better distinguish GPI-anchoring signals from transmembrane domains, two extra units were added to the input layer: one for the quality of a putative ω site and one for its position. This further decreased error rates (Table 1). Thus, the final input vector contained 44 components ($Z + H + \omega$; Table 1).

GPI-SOM

The final GPI anchoring signal prediction program GPI-SOM was implemented as a Kohonen SOM with an input layer of 44 neurons as described above. Square output maps of side length 10 did not provide enough room for both classes to segregate (not shown). With increasing side length there was a clearer separation of GPI and non-GPI proteins, until at length 40 the number of units in the map that were excited by proteins from both positive and negative sets was minimal. Figure 2 shows the process of self-organization during training. After a few cycles it became evident that the classes were separating using the zentriole plus hydrophobicity input vector ($Z + H$; Fig. 2B), illustrating that prediction of GPI-anchoring is solvable with a SOM. The collinear hydrophobicity vector alone did not distinguish clearly between GPI-positive and -negative proteins and the SOM took longer to reach minimal ambiguity (H; Fig. 2A).

After training, blank units in the map were classified based on their surroundings (see Algorithms). Since there were more than twice as many units in the SOM than sequences in the training sets, the majority of units was assigned only after training. While the units inside the GPI (blue) and non-GPI (green) regions were straightforward to assign, 11 of the units in between the two areas had to be left 'undecided' (red in Fig. 3). If such a blank unit is hit by a test sequence, there will be no prediction made (classified 'uncertain'). Furthermore, there was an inactive region of 185 blank units at the edge of the map that no input sequence has activated so far (Fig. 3). GPI-SOM is accessible via http://gpi.unibe.ch and accepts batch input in fasta format.

Evaluation of different GPI-prediction programs

A series of positive and negative test sets consisting of proteins from all eukaryote kingdoms were used to assess sensitivity and selectivity of the GPI-anchoring prediction programs BigPI, DGPI, GPI-SOM, and its corresponding feed-forward network ($Z + H + \omega$). Since a target protein must have an N-terminal ER export signal to receive its GPI anchor all programs were combined with SignalP (HMM version; Nielsen and Krogh, 1998), except for DGPI which already considers the N-terminus of the target protein. Prediction of GPI-anchored proteins based on their C-termini alone is not sensible since GPI-anchoring signals are only meaningful inside the ER (a presumed C-terminal GPI anchor attachment sequence, even a perfect one, is meaningless in the absence of an N-terminal export sequence).

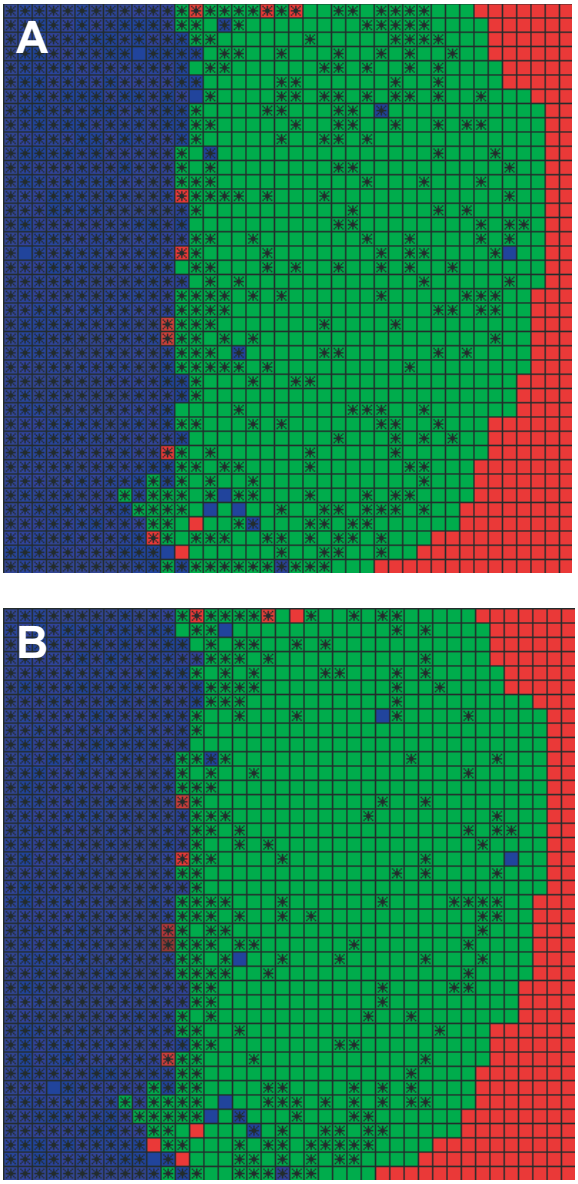

Fig. 3. The final map GPI-SOM. The map of 40 × 40 units was filled completely as described in Algorithms, and subdivided into three types of fields: GPI (green), non-GPI (blue) and undecided (red). This allowed fast scanning of large datasets. Black dots represent hits for (**A**) the predicted proteome of *S.cerevisiae* (5864 proteins) and (**B**) the same number of random sequences of the same amino acid frequencies as *S.cerevisiae* proteins. Intensity indicates how often a unit was hit. In the online version (http://gpi.unibe.ch) each unit is clickable, producing a list of the proteins that activated it.

Thus only proteins predicted to have both N- and C-terminal signals were classified as GPI-anchored.

As shown in Table 2, Big-PI was extremely specific, with hardly any false positive predictions throughout the negative test sets. The other programs also performed well, except against transmembrane proteins with an N-terminal export sequence plus a hydrophobic C-terminus (row d). These are the proteins most closely resembling GPI-anchored ones (Dalley and Bulleid, 2003) and, accordingly, the false positive error rates were around 30%. Regarding sensitivity, the feed forward network and GPI-SOM performed best. BigPI exhibited the highest rate of false negative predictions, presumably the price for its excellent specificity.

Table 2. Performance of GPI-anchoring prediction programs

	BigPI	DGPI	FF	GPI-SOM
Negative sets				
(a) Cytosolic	0	1.5	2.0	0.5
(b) Secreted	0	1.5	2.9	1.5
(c) TM	0	0	2.5	0.6
(d) N–TM–C	1.9	27	32	34
Positive sets				
(e) GPI	17	17	4	4
(f) new GPI	48	14	2.4	4.8

GPI-SOM and its corresponding feed-forward network (FF) are compared to Big-PI (Eisenhaber *et al.*, 2003) and DGPI (Kronegg and Buloz, 1999) using different test sets: (a) cytosolic proteins (196 sequences); (b) secreted proteins (68 sequences); (c) transmembrane proteins (159 sequences); (d) transmembrane proteins with N-terminal export signal and hydrophobic C-terminus (107 sequences); (e) GPI-anchored proteins not present in our positive training and validation sets (75 sequences); (f) recently published GPI-proteins which none of the programs had seen before (42 sequences). All test sets are available as supplementary material. BigPI, FF and GPI-SOM were combined with the HMM output of SignalP; DGPI already considers the N-terminus on its own. Numbers are the percentage of false predictions.

Genome-wide surveys for GPI-anchored proteins

GPI-SOM combined with SignalP was used in genome-wide surveys for GPI-anchored proteins in a number of eukaryotes. The *S.cerevisiae* proteome is shown as an example in Figure 3A. Of the total 5864 sequences, GPI-SOM predicted 438 positives, 121 of which were assigned N-terminal signals by SignalP resulting in 2.1% predicted GPI-anchored proteins. As stated above, the 307 proteins with predicted C-terminal signal but lacking an N-terminal one cannot be classified false positives; such predictions are meaningless (in order to test C-terminal predictions experimentally, the respective proteins would need to be fused to an ER export signal). For comparison, 5864 random sequences of the same amino acid frequencies as yeast proteins are shown in Figure 3B. GPI-SOM predicted 437 positives, of which only 8 (0.14%) were also predicted to possess an N-terminal export sequence by SignalP.

Most organisms appeared to have between 2 and 3% GPI-anchored proteins. Notable exceptions were *E.cuniculi* with only 0.5% and *T.brucei* with 5.6% predicted GPI-proteins (Fig. 4). Both are highly specialized parasites. Among the remaining organisms, a trend was observed toward a higher percentage of GPI-anchored proteins in organisms with larger proteomes (Fig. 4). Lists of predicted GPI-anchored proteins for different organisms are available as Supplementary information or from the GPI-SOM website.

DISCUSSION

SOMs are powerful tools for the detection and the classification of hidden patterns, but applications to proteins are hampered by the size of input data and by the inherent problem that conversion of

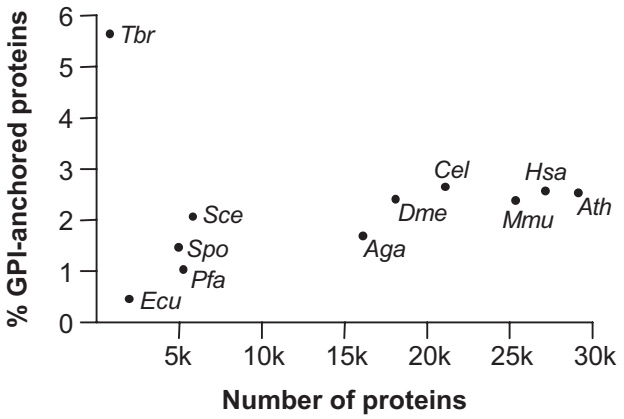

Fig. 4. Genome-wide predictions of GPI proteins. Predicted proteomes of completely sequenced eukaryotic genomes were screened for presumably GPI-anchored proteins with GPI-SOM. Since anchoring signals are only meaningful in the ER, only hits for which SignalP predicted an N-terminal export signal were counted as GPI-proteins. Organisms with more genes tended to have a higher percentage of GPI-anchored proteins (Aga, *A.gambiae*; Ath, *A.thaliana*; Cel, *C.elegans*; Dme, *D.melanogaster*; Ecu, *E.cuniculi*; Hsa, *H.sapiens*; Mmu, *M.musculus*; Pfa, *P.falciparum*; Sce, *S.cerevisiae*; Spo, *S.pombe*; Tbr; *T.brucei* chromosomes 1 and 2).

amino acid sequences to numerical format causes substantial loss of chemical information. Signal sequences localized within proteins, however, may be suitable targets for neural networks (Nielsen *et al*., 1997). This has been demonstrated by the good performance of the feed-forward network SignalP in predicting N-terminal export sequences (Bendtsen *et al*., 2004). Here we present GPI-SOM, a self-organizing map that recognizes C-terminal GPI-anchor attachment signals with good accuracy. It was developed by systematic, target-oriented optimization of input parameters and network architecture. Input consists of 44 numbers: the zentrioles for each amino acid (20 units), hydrophobicity of selected C-terminal positions (22 units), and quality and position of the best match for a putative ω site (2 units). The zentriole represents the average position of a given amino acid weighed by C-terminal proximity, a transformation somewhat related to the concept of virtual potentials (Aires-de-Sousa and Aires-de-Sousa, 2003). The output layer is a square map of 1600 units, where anchored and non-anchored proteins clearly separate (Fig. 2). The map was finalized by an algorithm that categorized empty or ambiguous units based on their surroundings. The good performance of GPI-SOM indicates that, in principal, the problem of GPI-anchoring signal prediction is solvable with a SOM.

GPI-SOM had a sensitivity of ~0.96 (Table 2). Selectivity is less straightforward to assess and depends greatly on the nature of the negative test proteins (Table 2). The main source of false positive predictions were integral membrane proteins with a transmembrane domain at their C-terminus (Table 2, row d). This is an inherent problem with GPI-anchoring signals; indeed, it has been shown experimentally that one point mutation may suffice to convert an anchor attachment signal to a transmembrane domain (Dalley and Bulleid, 2003). Misinterpretation of integral membrane proteins for anchored ones might be minimized by excluding sequences with multiple predicted transmembrane domains. However, we refrained from doing so since it cannot be excluded on the assumption that a protein has a C-terminal GPI anchor in addition to internal transmembrane domains.

A drawback of neural networks is that the machine is learning but not the investigator. In most cases, it is impossible to track the connections of a trained network and determine which input features are the most important. We have circumvented this problem by systematically altering input data. Thus, varying the length of input sequence (Fig. 1A) followed by a simulated mutagenesis experiment (Fig. 1B) identified crucial positions in the C-terminus distinguishing GPI-anchored from non-anchored proteins. This allowed maximal prediction accuracy with minimal input data (Table 1).

Surveys for GPI-anchored proteins were carried out in eukaryotes for which unbiased protein sets from completely sequenced chromosomes were available. Most species had between 2 and 3% predicted GPI-proteins (Fig. 4). Genome-wide prediction of GPI-proteins is critical because the error rates of GPI-SOM are in the same order of magnitude as the percentages of GPI-anchored proteins in a given proteome. Thus the predicted numbers of GPI-proteins have to be taken with caution. Also prediction of N-terminal signal sequences with SignalP, which is a prerequisite for prediction of GPI-anchor attachment sites, involves a certain error. Nevertheless, genome-wide comparisons between different species may yield insights into their use of GPI anchors. There appeared to be a trend towards higher percentages of GPI-anchored proteins in larger proteomes. No such trend was observed in transmembrane proteins (Ward, 2001). Top and bottom positions in Figure 4 were taken by the parasitic protozoa *T.brucei* (5.6% GPI-proteins) and *E.cuniculi* (0.5% GPI-proteins), respectively. *T.brucei* proliferate extracellularly in the mammalian bloodstream. Evading the host's immune system by variation of their surface coat, *T.brucei* spp. have a repertoire of several hundred genes for GPI-anchored surface glycoproteins (Donelson, 2003). The microsporidian *E.cuniculi*, in contrast, is an obligate intracellular parasite and might, therefore, not be expected to possess GPI-anchored proteins at all. However, GPI-SOM in combination with SignalP identified 9 candidate proteins with N- and C-terminal signals, among which a proteinase and proteins similar to oligosaccharide deacetylase, glucosyltransferase, and glucan glucosidase (see Supplementary data). *E.cuniculi* lacks several of the enzymes involved in GPI synthesis and attachment; but surprisingly, it has a predicted protein with high similarity to phosphatidylinositol N-acetylglucosaminyltransferase (GPI2), catalyzing the first step in GPI synthesis (GenBank accession NP_597633 has a p-value of 2e-119 against PFAM entry PF06432). Whether the nine *E.cuniculi* proteins predicted to receive an anchor are false positives or whether some of these proteins actually get anchored to the host cell membrane remains to be investigated.

In summary, GPI-SOM is a new approach towards computational prediction of GPI-anchoring signals and provides a welcome addition to the existing programs Big-PI and DGPI which predict GPI anchor attachment sites based on statistical expectation.

ACKNOWLEDGEMENTS

We wish to thank Isabel Roditi, Peter Bütikofer, Walter Senn and Daniel Stalder for the helpful advice, and Paul Dupree for the provision of sequences of GPI-anchored *Arabidopsis* proteins. This work was supported by a Swiss National Science Foundation research professorship grant to P.M.

REFERENCES

Abe,T. *et al*. (2003) Informatics for unveiling hidden genome signatures. *Genome Res.*, **13**, 693–702.

Aires-de-Sousa,J. and Aires-de-Sousa,L. (2003) Representation of DNA sequences with virtual potentials and their processing by (seqrep) Kohonen self-organizing maps. *Bioinformatics*, **19**, 30–36.

Bendtsen,J.D. *et al*. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.

Borner,G.H. *et al*. (2003) Identification of glycosylphosphatidylinositol-anchored proteins in *Arabidopsis*. A proteomic and genomic analysis. *Plant Physiol.*, **132**, 568–577.

Butikofer,P. *et al*. (1999) Phosphorylation of a major GPI-anchored surface protein of *Trypanosoma brucei* during transport to the plasma membrane. *J. Cell Sci.*, **112** (Pt 11), 1785–1795.

Caras,I.W. *et al*. (1987) Signal for attachment of a phospholipid membrane anchor in decay accelerating factor. *Science*, **238**, 1280–1283.

Dalley,J.A. and Bulleid,N.J. (2003) The endoplasmic reticulum (ER) translocon can differentiate between hydrophobic sequences allowing signals for glycosylphosphatidylinositol anchor addition to be fully translocated into the ER lumen. *J. Biol. Chem.*, **278**, 51749–51757.

Donelson,J.E. (2003) Antigenic variation and the African trypanosome genome. *Acta Trop.*, **85**, 391–404.

Eisenhaber,B. *et al*. (1998) Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. *Protein Eng.*, **11**, 1155–1161.

Eisenhaber,B. *et al*. (1999) Prediction of potential GPI-modification sites in proprotein sequences. *J. Mol. Biol.*, **292**, 741–758.

Eisenhaber,F. *et al*. (2003) Prediction of lipid posttranslational modifications and localization signals from protein sequences: Big-pi, nmt and pts1. *Nucleic Acids Res.*, **31**, 3631–3634.

Ferguson,M.A. (1999) The structure, biosynthesis and functions of glycosylphosphatidylinositol anchors, and the contributions of trypanosome research. *J. Cell. Sci.*, **112** (Pt 17), 2799–2809.

Ferran,E.A. and Ferrara,P. (1992) Clustering proteins into families using artificial neural networks. *Comput. Appl. Biosci.*, **8**, 39–44.

Ferran,E.A. *et al*. (1994) Self-organized neural maps of human protein sequences. *Protein Sci.*, **3**, 507–521.

Fournier,A.E. *et al*. (2001) Identification of a receptor mediating nogo-66 inhibition of axonal regeneration. *Nature*, **409**, 341–346.

Gerber,L.D. *et al*. (1992) Phosphatidylinositol glycan (pi-g) anchored membrane proteins. Amino acid requirements adjacent to the site of cleavage and pi-g attachment in the COOH-terminal signal peptide. *J. Biol. Chem.*, **267**, 12168–12173.

Guadiz,G. *et al*. (1998) The carboxyl terminus of *Pneumocystis carinii* glycoprotein a encodes a functional glycosylphosphatidylinositol signal sequence. *J. Biol. Chem.*, **273**, 26202–26209.

Hamburger,D. *et al*. (1995) Yeast gaa1p is required for attachment of a completed GPI anchor onto proteins. *J. Cell Biol.*, **129**, 629–639.

Kanaya,S. *et al*. (2001) Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E.coli* O157 genome. *Gene*, **276**, 89–99.

Kodukula,K. *et al*. (1993) Biosynthesis of glycosylphosphatidylinositol (GPI)-anchored membrane proteins in intact cells: specific amino acid requirements adjacent to the site of cleavage and GPI attachment. *J. Cell Biol.*, **120**, 657–664.

Kohonen,T. (2001) *Self-Organizing Maps*, 3rd edn, Springer Series in Information Sciences, Vol. 30, Springer, Berlin.

Kronegg,J. and Buloz,D. (1999) Detection/prediction of GPI cleavage site (GPI-anchor) in a protein (DGPI).

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Lacey,S.W. *et al*. (1989) Complementary DNA for the folate binding protein correctly predicts anchoring to the membrane by glycosyl-phosphatidylinositol. *J. Clin. Invest.*, **84**, 715–720.

Meyer,U. *et al*. (2002) The glycosylphosphatidylinositol (GPI) signal sequence of human placental alkaline phosphatase is not recognized by human gpi8p in the context of the yeast GPI anchoring machinery. *Mol. Microbiol.*, **46**, 745–748.

Moran,P. and Caras,I.W. (1994) Requirements for glycosylphosphatidylinositol attachment are similar but not identical in mammalian cells and parasitic protozoa. *J. Cell Biol.*, **125**, 333–343.

Morel,N. and Massoulie,J. (1997) Expression and processing of vertebrate acetylcholinesterase in the yeast *Pichia pastoris*. *Biochem. J.*, **328** (Pt 1), 121–129.

Murray,S.O. *et al*. (1998) The neural network classification of false killer whale (*Pseudorca crassidens*) vocalizations. *J. Acoust. Soc. Am.*, **104**, 3626–3633.

Netzel-Arnett,S. *et al*. (2003) Membrane anchored serine proteases: a rapidly expanding group of cell surface proteolytic enzymes with potential roles in cancer. *Cancer Metastasis Rev.*, **22**, 237–258.

Nielsen,H. and Krogh,A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park CA, pp. 122–130.

Nielsen,H. *et al*. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, **8**, 581–599.

Nozaki,M. *et al*. (1999) Developmental abnormalities of glycosylphosphatidylinositol-anchor-deficient embryos revealed by Cre/LoxP system. *Lab. Invest.*, **79**, 293–299.

Sheridan,J.P. *et al*. (1997) Control of trail-induced apoptosis by a family of signaling and decoy receptors. *Science*, **277**, 818–821.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Stahl,N., Borchelt,D.R., Hsiao,K. and Prusiner,S.B. (1987) Scrapie prion protein contains a phosphatidylinositol glycolipid. *Cell*, **51**, 229–240.

Supek,F. and Vlahovicek,K. (2004) Inca: synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics*, **20**, 2329–2330.

Sutterlin,C. *et al*. (1998) *Saccharomyces cerevisiae* gpi10, the functional homologue of human pig-b, is required for glycosylphosphatidylinositol-anchor synthesis. *Biochem. J.*, **332** (Pt 1), 153–159.

Takeda,J. and Kinoshita,T. (1995) GPI-anchor biosynthesis. *Trends Biochem. Sci.*, **20**, 367–371.

Toronen,P. *et al*. (1999) Analysis of gene expression data using self-organizing maps. *FEBS Lett.*, **451**, 142–146.

Udenfriend,S. and Kodukula,K. (1995) Prediction of ω site in nascent precursor of glycosylphosphatidylinositol protein. *Methods Enzymol.*, **250**, 571–582.

Ward,J. (2001) Identification of novel families of membrane proteins from the model plant *Arabidopsis thaliana*. *Bioinformatics*, **17**, 560–563.