Institut Suisse de Bioinformatique (ISB)
Swiss Institute of Bioinformatics (SIB)

Evaluation of programs for the prediction of GPI-anchor in proteins.

Report of Practical Training
performed at Swiss-Prot group,
Swiss Institute of Bioinformatics
CMU, 1 Michel-Servet
CH-1211 Genève 4 - Suisse
submitted for:
The Master Degree in Proteomics and Bioinformatics
by Natalia V. Sernova.

Jury:
Dr. Anne-Lise Veuthey (responsable de stage)
Dr. David Perret
Dr. Patricia Palagi
Prof. Amos Bairoch (président de jury)
To be defended at the University of Geneva, June 16, 2006 at 4:00 PM

Table of contents:

SUMMARY.

Purposes:

- UniProtKB/Swiss-Prot is a curated database of proteins, which integrates biological information retrieved from the biomedical literature. In addition, UniProtKB/Swiss-Prot annotation relies on sequence analysis tools and protein feature prediction methods, since genome sequencing projects are providing large amounts of potential protein coding regions without any experimental characterization. Thus, promoting the development of reliable predictive methods, as well as evaluating their performances is an essential task for improving the database quality.
- Protein modifications play crucial structural and functional roles (Han *et al*, 1992), so their identification is very important. Experimental data are available for limited number of proteins, so prediction is the only way to deal with huge amount of protein sequence data. Any PTM prediction program for UniProtKB/Swiss-Prot annotation has to be evaluated, and since new data information may help to improve program performance, evaluation has to be done regularly.

Methods:

- Evaluation implies the estimation of prediction method performance on experimentally proved negative and positive protein datasets, making proper dataset construction a priority. Being constructed, besides evaluation the datasets can be also used to develop new prediction programs, as well as to improve existing methods.
- In the present work we evaluated the performance of three currently existing programs, predicting GPI-anchor protein modification (namely BigPI, DGPI and GPI-SOM). Protein data for evaluation datasets were extracted from the UniProtKB/Swiss-Prot, the most accurate and up-to-date protein knowledgebase currently available.
- This work was started with data from UniProtKB/Swiss-Prot release 48.1 (September 2005), and near finishing results were updated according to the data from release 49.7 (May 16, 2006). The script for later fully automated updates is provided by the author.

Results:

- The most reliable program in terms of its specificity is BigPI, which was investigated on experimentally verified metazoan dataset. Lack of verified data prevented us from proper investigation of other taxons, as well as some program features like cleavage site prediction.

Conclusions:

- Better datasets are needed to create better prediction programs, although currently existing programs, like BigPI, may already give some hints on what proteins experimentalists should concentrate to have verified GPI-anchored protein sequences.

Work objectives:

These studies were initiated to pursue the following goals:

- to evaluate currently existing GPI-anchor predictive methods
- to update experimentally verified datasets for the purpose to develop new prediction programs, as well as to improve existing tools
- to develop automatic update tools

To achieve these goals, the subject was sub-divided into the following tasks:

- to create positive and negative dataset of experimentally verified proteins, being/not being GPI-anchored, respectively.
- to write Perl-scripts, which would access web-sites of three GPI-anchor prediction programs and parse the output, what will allow to calculate specificity/sensitivity of the three programs.
- to analyse and compare the results of all three programs.

1. Introduction.

Epigraph:

"Increased confidence that a protein is putatively GPI-anchored should encourage more researchers to experimentally verify this modification, which in turn will allow the construction of a predictor with even higher accuracy in the future." Eisenhaber *et al*, 1998

Almost all proteins analyzed to date carry some post-translational modifications (PTMs). The modified protein function is often strongly affected, or even determined by these modifications (Blom *et al*, 2004). Increased knowledge about the potential PTMs of a target protein may deepen our understanding of the molecular processes it is involved in, and ultimately of its function. High-throughput methods for the identification of PTMs are being developed, in particular within the fields of proteomics and mass spectrometry.

However, most of these methods are still in their infancy, and to cut down on the number of experimental steps by integrating computational approaches into the validation procedures is indeed advantageous. Many advanced methods for the prediction of PTMs exist and many are made publicly available.

The current work is dedicated to evaluation of performance for three existing GPI-anchor prediction programs.

In this introduction we will talk about:

- the importance of post-translational modifications for protein function with a description of GPI-anchor PTM as an example;
- UniProtKB/Swiss-Prot database as a reliable source of PTM-information, with its daily data curation which is necessary to maintain data high quality;
- existing experimental GPI-detection techniques as well as computational GPI-prediction methods

1.1 The PTMs and their importance for protein function

Many years ago, in the early days of molecular biology, the function of a protein was typically known before the sequence of amino acids encoded by the gene was determined. Nowadays the situation is reversed, and as long as sequences have been accumulating in the databases, corresponding protein functional analysis/prediction will remain the most important issue. Currently the number of experimentally validated examples of post-translational modifications (PTMs) grows tremendously, since PTMs make a particular protein molecule unique in terms of its structure and function.

Three main classes of post-translational modifications (the name is misleading because the modifications may also occur before and during protein synthesis) are the following:

- cleavage
- linkage
- cross-linking

and can be combined – for instance, glycosylphosphotidylinositol (GPI) – anchor attachment implies both cleavage and linkage. (Farriol-Mathis *et al*, 2004)

Proteins appear to be modified several times along their life-time (Eisenhaber at al, chapter 5, 2003). Most proteins cannot perform their molecular function as unmodified folded polypeptides. In most cases, proteins need to acquire permanent or transient molecular features in order to become functional.

There is enormous amount of known DNA-sequences (as well as complete genomes) nowadays, which raise a question of PTM-prediction from amino acid sequence for genome annotation (with further goal of protein function prediction). PTMs influence protein size, hydrophobicity and other physico-chemical properties; can change, enhance or block a specific activity; can also target the protein to the specific subcellular location. Because of that, PTM-prediction from protein sequence data is very important.

The paradigm has been that protein sequence determines its structure, and knowing the structure yields the functional information (Bork *et al*, 1998; Attwood, 2000). A somewhat complementary approach is currently being developed with bioinformatics, namely the idea of making use of protein features and then assign function using the features in an integrated fashion (Jensen *et al*, 2002; Jensen *et al*, 2003).

Such features include global properties such as molecular weight, isoelectric point, localization signals and information about potential PTMs. This novel approach of including PTMs in function prediction is based on the assumption that overall structure and function of a mature protein in a cell is influenced significantly by PTMs. The aim is then to use a number of correlated features to predict the functional category of a protein.

1.2 Challenges in PTM-prediction.

PTM-prediction nowadays is still a difficult task, although many prediction programs are created to deal with a huge amount of sequence data. We will try to describe the main problems, which can be encountered on this way.

- **Datasets** – often do not contain sufficient number of sequences with verified feature, but often redundant in terms of containing multiple related sequences, and there are approaches helping to deal with it and reduce this redundancy (Hobohm *et al*, 1992). But it may lead to the loss of information. Sometimes a few substitutions in the sequence completely abolish the PTM-process – folate receptor families serve as a good example (Eisenhaber *et al*, 1999). Special means, like PSIC, “position specific independent counts” (Sunyaev *et al*, 1999), has to be used to overcome this problem by taking into account both sequence- and position-specific weighting in profile extraction from alignment (Sunyaev *et al*, 1999).
- **Structural aspects**. Since fold recognition/structure prediction are not reliable so far, prediction of PTM which somehow depends on structural features is not reliable either. If the sequence motif is inside the protein globule, it is unlikely be modified; so some people consider for prediction only PTM occurred in N- or C-terminal unstructured regions, which are easier accessible for the modifying enzyme(s). GPI-anchor is among them. (Eisenhaber *et al*, 2003)

- **Modification process itself**. The complexity of the process and the number of enzymes involved determine the chance of successful predictions. Often PTM implies a sequence of events, which make a successful prediction even less probable. GPI-anchoring include transport to the ER by leader peptide mechanism, cleavage of C-terminal propeptide and attachment of the anchor itself. (Udenfriend and Kodukula, 1995), so it is complex enough to be a real challenge.

The first and the most important question for any PTM-prediction program is datasets, which it was based on and with which it was evaluated. Now we will discuss the databases, which can be a proper source for those datasets.

1.3 Post-translational modification data resources. UniProtKB/Swiss-Prot

Virtually any of the 20 natural amino acids may be modified by some type of PTM as evidenced by the many examples shown in the RESID database (Garavelli, 2003). The polypeptide chain is subject to many different types of post-translational processing in different cellular compartments, including the nucleus, cytosol, endoplasmic reticulum and Golgi apparatus. These modifications may confer various structural and functional properties to the affected proteins.

A database of protein post-translational modifications with descriptive, chemical, structural and bibliographic information is available: RESID (http://www.ebi.ac.uk/RESID/) (Garavelli, 2003). But the key repository of protein sequences modified by PTMs has been the UniProtKB/Swiss-Prot (Wu *et al*, 2006, Apweiler *et al*, 2004, Farriol-Mathis *et al*, 2004, Boeckmann, 2005), and other PTM specific databases have also emerged recently.

There are three features, which distinguish UniProtKB/Swiss-Prot database from other protein sequence databases: annotation, minimal redundancy and abundant references to other databases. We will discuss the annotation feature in detail, because it is particularly important for the whole work presented here.

For methods which are based on experimental data, the prediction accuracy is strongly limited by the amount and redundancy of the underlying data. High quality annotated data is the key for the development of PTM classifiers, but unfortunately most of the data in the databases is annotated based on similarity, and rarely on experimental evidence.

The amount of data required for obtaining a high quality prediction obviously depends on the diversity of modified protein sequences, and it is difficult to make a reliable estimation of the "sufficient" amount of data, although some statistical suggestions were presented in the literature (Eisenhaber *et al*, 2004). Another related issue is how well a particular algorithm is able to construct a model of the modified sequence from a limited set of known sites.

1.3.1 Swiss Prot annotation.

The annotation is mainly found in the comment lines (CC), in the feature table (FT) and in the keyword lines (KW). (Farriol-Mathis *et al*, 2004)

- The comment block (CC lines) which is used by curators to indicate useful information in a free text format. Each CC line belongs to a topic describing the feature. For instance, "SUBCELLULAR LOCATION" is a topic related to the subcellular location of the mature protein.
- Keywords (KW lines). Allowed keywords are part of a controlled vocabulary and the presence of a keyword in a KW line is usually linked to the presence of other topic-related lines in the protein entry.
- The feature table (FT lines) describes regions or sites of interest in a sequence, such as post-translational modifications, enzyme active sites, and other characteristics, either reported in the cited references or predicted by sequence analysis tools.

1.3.2 Non-experimental qualifiers.

Three major types of non-experimental qualifiers in comment (CC) lines and feature table (FT lines) indicate that the information given is not based on experimental data:

- Potential
- Probable
- By similarity

The information provided in a UniProtKB/Swiss-Prot entry can be either experimentally verified or based on prediction. In the first case, the corresponding line describing the feature doesn’t have any qualifiers. In contrast, the presence of the “Probable”, “By similarity” or “Potential” qualifiers indicates that the feature was not proven experimentally. “Probable” means, that although the presence of a specific sequence feature has never been directly proven, there are some indirect experimental clues which suggest that the protein has a high probability of having it. The “Potential” qualifier indicates that the feature has been discovered by running a prediction program on the protein sequence. “By similarity” is used when an ortholog sequence has been experimentally shown to have this specific feature.

It is of particular importance to make test, validation and evaluation datasets on entries without non-experimental qualifiers, and just on those which contain experimentally verified sequence feature(s). Now we will look at one informative example of post-translational modifications, GPI-anchoring, and will discuss what is known about it and proteins, which carry it.

1.4 GPI-anchor PTM

A **GPI anchor,** which is stands for glycosylphosphatidylinositol anchor, is a common modification that is covalently linked to the C-terminus of proteins.. (Takeda *et al*, 1995, Ferguson, 1999, Hooper *et al* 1999, Horejsi *et al*, 1999) It is composed of a hydrophobic *phosphatidyl inositol* group linked through a carbohydrate containing linker (glucosamine and mannose linked to phosphoryl ethanolamine residue) to the C-terminal amino acid of a mature protein**.** The structure of the GPI-anchor (Fig.1) is similar among different organisms except for the addition of carbohydrate side groups (Masterson et al, 1989; see also appendix for the details of GPI-anchor chemical structure). The two fatty acids within the hydrophobic phosphatidylinositol group anchor the protein to the membrane.

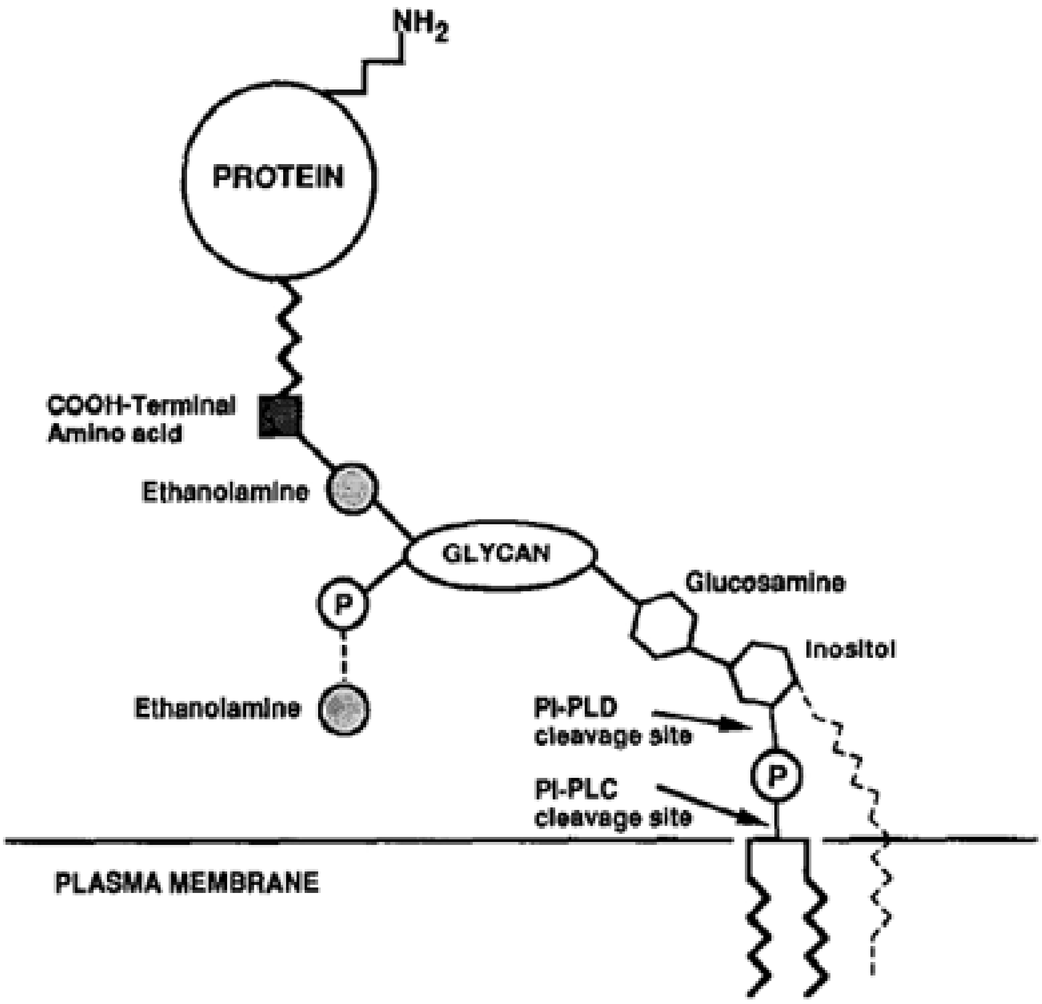


Figure 1 Structure of a typical GPI-anchored protein.

(From Udenfriend & Kodukula, 1995b)

1.4.1 The mechanism of GPI-anchor attachment to the protein.

GPI-anchoring is a mechanism for tethering eukaryotic proteins to cellular membrane. (Beghdadi-Rais *et al*, 1993) Those proteins are extracellular and otherwise would be secreted.

A protein, which is supposed to become GPI-anchored, before the processing contains hydrophobic sequences both at its N- and C-termini (Udenfriend *et al*, 1995a, Yan *et al*, 1998) : N-terminal peptide, which is targeting the newly synthesized protein to the ER, and later removed by signal peptidase, and C-terminal peptide, which is removed by putative GPI-transamidase, at the same time attaching GPI-anchor moiety to the last residue of remaining part, so called ω-site. Thus, features required for the protein to become GPI-anchored are the following (Eisenhaber *et al*, 1998). (See also figure 2A, describing this list.)

- Export signal (N-terminal signal peptide)
- Recognition signal for transamidase complex (four main regions):
  - (I) a polar and flexible linker region (~ 11 residues – ω-11…ω-1)
  - (II) a region of small residues (ω-1…ω+2) with the ω-site
  - (III) a spacer region (ω+4…ω+9) with moderately polar flexible residues
  - (IV) a hydrophobic tail from ω+9 – ω+10 up to the C-terminal end

(Roman digits are region identifiers.)

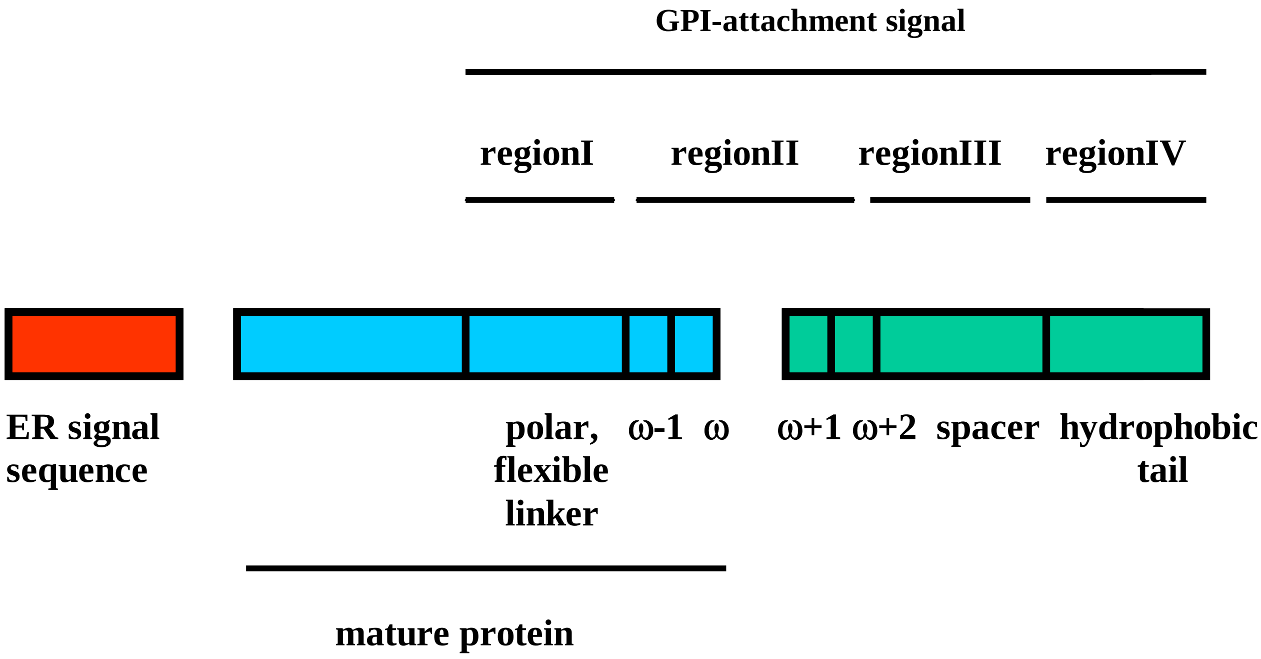


Figure 2A. The C-terminal GPI lipid anchor signal on the sequence level (adapted from Eisenhaber *et al*, 2003)

The entry to the GPI-modification reaction is directed entirely by a C-terminal sequence signal. It has been proven experimentally that C-terminal sequence fragment is sufficient to make the protein GPI-anchored, (Eisenhaber, 1998). Mature protein is then translocated with secretory vesicles to be immobilized on the extracellular side of the plasma membrane.(Wang *et al*, 1999) The fact that a protein is GPI-anchored alone determines its cellular localization and limits its range of possible molecular functions. (Eisenhaber *et al*, 1998).

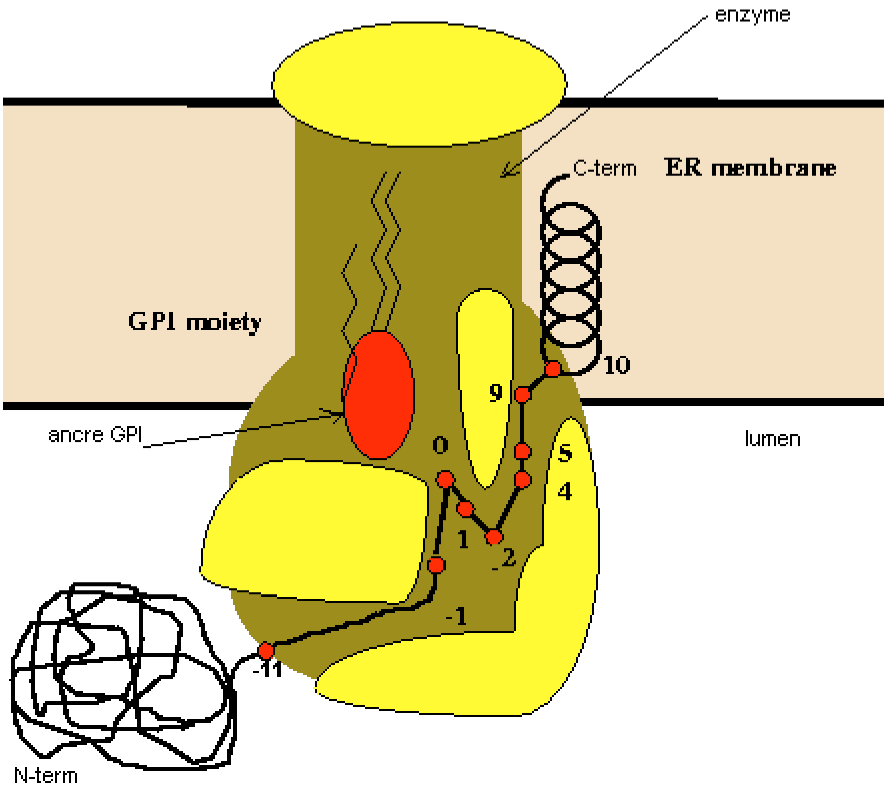


Figure 2B. 3D-model for GPI-anchor attachment (Eisenhaber *et al*, 1998).

1.4.2 Experimental way to prove GPI-anchoring for the protein

Experimental scientists face significant difficulties, trying to prove, if a protein is GPI-anchored, which ideally include the necessity to determine its attachment site as well. (Eisenhaber, 1999) Unambiguous answer can be given only by mass spectrometry. The process will involve protease digestion of the suspected protein, subsequent separation of GPI-anchored peptide and its sequencing by tandem mass-spectrometry (Taguchi *et al*, 1999 a and b, Haas et al, 1996, Omaetxebarria, 2006). But this technique is expensive plus requires purified protein, which often is not needed for any other purpose; that is why scientists usually prefer simpler, but unfortunately ambiguous approaches. At the protein level, GPI anchoring was originally demonstrated by protein release from the cell surface by PI-PLC (phosphatidylinositol specific phospholipase C). This is still the simplest method, but it is not always reliable because, for example, in some forms of GPI, such as human erythrocyte acetylcholinesterase (Roberts WL, Myher J J, *et al*, 1988, Roberts WL, Santikarn S, *et al*, 1988), an additional palmitoyl group on the inositol results in

resistance to PI-PLC. Similarly, the procyclic acidic repetitive protein (PARP) of *Trypanosoma brucei* also contains an additional fatty acid (Field MC, *et al*, 1991).
Protein metabolic labeling by radioactive elements of GPI such as ethanolamine, inositol, or fatty acids is alternative, frequently used method of establishing that a protein is GPI-anchored. Incorporation of at least two of the above components should be used because there are other mechanisms by which either ethanolamine or fatty acids can be incorporated into proteins (Howard AD *et al* ,1987, Micanovic R *et al*, 1988, Ogata S *et al*, 1988, Ogata S *et al*, 1990). Furthermore, one must demonstrate that the radioactivity incorporated is not a metabolite of the radioactive precursor.

Only if all of the “ambiguous” procedures were successfully carried out, the researcher can be fairly certain that he is dealing with a GPI protein. However, the residue in the nascent protein to which the GPI is attached must still be determined.

Relatively few of the many known GPI proteins have had their ω sites determined experimentally because this process involves isolation and purification of the protein and enzymatic fragmentation, isolation, and sequencing of the peptide(s) containing a GPI moiety. As with NH2-terminal processing, the cleavage sites of COOH-terminally processed nascent proteins are now generally deduced.

In those rare cases, where the ω site had been determined experimentally, GPI-containing peptides released after protease treatment were detected in the lysates by one of two methods. Where the protein could possibly be labeled metabolically, either with radioactive ethanolamine, inositol, or fatty acid, the peptide was detected and purified by monitoring radioactivity in the protease digest (Micanovic R *et al*, 1988, Caras IW, 1991, Ogata S *et al*, 1988, Ogata S *et al*, 1990). Production of a site-directed antibody upstream and close to the expected ω site was used to detect and isolate the GPI peptide in digests of placental alkaline phosphatase (PLAP) (Bailey *et al*, 1988, Micanovic R *et al*, 1988). It is also possible to use site-directed mutagenesis of potential ω-residues. (Furukawa *et al*, 1997)

There can be some ambiguities for these investigations as well. For example, site determination may be further complicated by the existence of minor, alternate ω-sites in addition to the main one. (Yan, 1995; Bucht, 1996). It is also known, that cellular determination for GPI-anchoring may affect only a fraction of protein molecules in the cell, and a fraction of molecules could be, for instance, cleaved and released to the medium, which was investigated for folate receptor (FR) beta (Wang *et al*, 1997). Carboxy-terminal peptide in FR-beta is efficiently proteolyzed intracellularly by a pathway that is independent of GPI signal recognition, and protein molecules without C-terminal part are secreted. Definitely, plenty of this kind of mechanisms remain to be discovered.

This explains why such a small amount of proteins is experimentally proven to be GPI-anchored.

1.4.3 GPI-anchoring prediction program algorithms

Another way to figure out if the suspected protein is GPI-anchored is prediction. So far there are 3 programsfor the prediction of GPI-anchoring and the site for attachment of GPI-moiety. These programs are Big-PI (Eisenhaber*et al*., 1999, which now exists in kingdom-specific flavors:
(http://mendel.imp.univie.ac.at/gpi/gpi_server.html for metazoa 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),
DGPI (Kronegg and Buloz, 1999, http://129.194.185.165/dgpi/index_en.html) and
GPI-SOM (Fankhauser *et al*, 2005, http://gpi.unibe.ch/). They are based on either machine learning techniques such as neural networks (GPI-SOM) or rules depicted from the analysis of the biological sequences of experimentally known targeted proteins (BigPI and DGPI).

A concise description of all 3 programs was made in a recent paper (Fankhauser at al, 2005).

BigPI-algorithm is based on sequence properties extracted from a positive set, as well as DGPI-algorithm. BIG-PI and DGPI both predict GPI-anchoring, investigating amino acid composition around ω-site. According to Fankhauser *et al*, 2005, both programs are useful to predict attachment site in proteins, which are known to be GPI-anchored. GPI-SOM is supposed to predict GPI-anchoring for unknown protein, using Kohonen's SOM (self organising map) approach. In the rest of this chapter we will consider those algorithms in detail.

DGPI algorithm is based on amino acid hydrophobicity estimation. The scale of hydrophobicity used is that of Kyte and Doolittle (Kyte at all, 1982). Two filters are imposed – low pass and median.

The idea of low pass or mean filtering is simply to replace each amino acid hydrophobicity value with the mean (`average') value of its neighbors within a sequence window, including itself. This has the effect of eliminating values, which are unrepresentative of their surroundings.

The median filter is normally used to reduce noise, somewhat like the mean filter. However, it often does a better job than the mean filter of preserving useful detail in the sequence. Instead of simply replacing the value with the *mean* of neighbouring amino acid values, it replaces it with the *median* of those values. The median is calculated by first sorting all the amino acid values from the surrounding neighbourhood within the window into numerical order and then replacing the amino acid value being considered with the middle value.

Five rules have to be fulfilled for the protein sequence to be predicted as GPI-anchored:

The first rule is that the protein must contain a signal at the N-terminus (because the modification takes place at the ER).

The second rule states the presence of hydrophobic part of a certain length (minimum 13 amino acids). The authors used a combination of the low-pass filter and median filter for analysis. Coefficient of weighting of these filters was given heuristically and the best report/ratio remains the fifty-fifty. The window which gave the best results is 15.

The third rule relates to the hydrophylic part. Its length is given according to the same method as for the hydrophobic part (two filters, windows of 15).

The fourth rule relates to the number of amino acids (approximately 8 amino acids) between the end of the hydrophobic part and the site of anchoring.

The fifth rule allows a position improvement of the cleavage site found by rule 4. For that, the distribution of amino acids around the cleavage site was used. Authors analyzed 172 annotated UniProtKB/Swiss-Prot GPI-anchored proteins. and recomputed previous distribution of amino acid frequencies at ω, ω+1, ω+2 sites (Udenfriend & Kodukula, 1995), because numbers in Udenfriend & Kodukula table didn’t give correct DGPI predictions for verified sequences. The new table is provided on the web-site http://129.194.185.165/dgpi/algorithme_en.html.

BigPI-algorithm is also based on sequence analysis, but depends on many different parameters, not only hydrophobicity. There are 4 slightly different versions of the program, dedicated to 4 eukaryotic taxonomic groups, metazoa, protozoa, fungi and plants. The general GPI-terminal signal scheme is the same for all four, but differs in taxon-specific details.

The GPI modification sequence motif for BigPI appeared to be described in terms of physical properties such as length requirements and average hydrophobicity (e.g. for the C-terminal segment), sometimes involving interactions of several sequence positions (Caras & Weddel, 1987; Moran *et al.*,1991; Udenfriend & Kodukula, 1995a,b; Furukawa *et al.*, 1997; Eisenhaber *et al.*, 1998).

Since profile constructions, based on alignments of sequence segments around ω-site are not successful in terms of their failure to find potential proproteins in the database, the facts that GPI-sequence signal is not well characterized by amino acid type preferences, and single residue substitution may completely reverse the modification, a sophisticated scoring function should be applied.

Therefore, a final scoring function S consists of two parts: $S=S_{profile} + S_{ppt}$.

A profile-dependent section $S_{profile}$ evaluates the concordance with the weak amino acid type preferences in the learning set at single alignment positions. The relative occurrences of amino acids of particular type at a given motif position were determined. A new profile extraction technique (PSIC: position-specific independent counts) which assigns both sequence and alignment position-specific weights (Eisenhaber *et al.*, 1998; Sunyaev *et al.*, 1999) was applied. The profile score $S_{profile}$ is composed of weighted subscores for specific sequence regions (see Figure 2A) and two penalties.

Physical property terms, which compose the score S_{ppt}, describe the conservation of physical properties in the GPI-modification signal arising from the interaction of few or many sequence positions. Those terms include:

- side-chain volume limitations and mutual volume compensation effects for residues ω-1 . . . ω+2 ;
- backbone flexibility requirements within the segment ω-1 . . . ω+2;
- propeptide length ranges (from ω+1 to the C end);
- spacer region (ω+3 ... ω+8) hydrophilicity and sequence volume per residue;
- hydrophobicity limits averaged over the C-terminal hydrophobic region and conditions for even distribution of hydrophobic residues;

- the presence of aliphatic hydrophobic residues (LVI-contents in the tail) and the absence of long stretches of residues with a flexible backbone in the C-terminal hydrophobic tail.

S_{ppt} is organized in such a manner that clear deviations from value ranges in the learning set of proproteins are penalized. The form of physical terms in S_{ppt} reflects BigPI authors model for requirements of the protein binding site in the transamidase complex executing the GPI modification.

A scoring function, developed for the animal predictor (Eisenhaber *et al*, 1999) was slightly modified for plant and fungi predictor; some terms were introduced and some were removed. For all predictors, four signal region-specific profile terms are in $S_{profile}$ but they are parameterized using the alignment of C termini of corresponding taxonomy learning set sequences. For the plant predictor, when the analysis of the plant-specific C-terminal pattern in learning set sequences was carried out, it was found out that, in addition to the previously described terms in the S_{ppt} component (physical property terms; see Eisenhaber *et al*., 1999), seven new terms for plant-specific features can be introduced. Seven terms in S_{ppt} have been changed in their functional form for fungi predictor as well, replaced or newly introduced as compared with the animal predictor.

When a query sequence was submitted, the predictor scans the C-terminal 55 residues (the same length for all taxonomies), calculates a score for each position (assuming it being the ω-site) and selects the best scoring sites. A score $S \geq 2$ indicates a reliably predicted site (prediction label P). Typically, such a score requires an almost zero S_{ppt} and a sufficiently positive $S_{profile}$. Scores $S \leq -2$ are interpreted as resulting from sequences without capacity for GPI lipid anchoring (label N). The remaining scores $-2 < S < 2$ belong to a twilight zone where clear prediction is difficult (label S). Nevertheless, low S_{ppt} subscores ($S_{ppt} \leq -12$ for positive S and $S_{ppt} \leq -8$ for negative S) are used to discriminate unlikely transamidase substrates within this group (prediction quality I). The scores are translated into probabilities of false motif detection.

GPI-SOM program is based on different principles then DGPI and BigPI ones. The web-site of the program is http://www.unibe/gpi.ch. Kohonen neural networks are known to be powerful tools for classification of hidden information in large datasets. Learning happens classically by adjusting the weights of the connections (synapses) between units (neurons). But Kohonen self-organizing maps (SOMs), or Kohonen neural networks (Kohonen, 1988), were employed because they learn by unsupervised training, in contrast to the classical feed-forward networks. This means that the network self-organizes during the training, and distributes the sequences in a map without knowing to which classes they belong. Only at the end of the training the sequences are labelled with their known classes. The neurons are then assigned to classes according to the sequences that excited them, and by inspection of the resulting map it is possible to verify if clustering of the different classes occurred.

It was necessary to evaluate different numerical representation formats for amino acid sequence. Finally sequence information had to be read by the input layer of a neural network, and there was a dilemma – either substantial loss of information or huge amount of data and long computation time.

The authors of GPI-SOM tried several different input formats for representation of positional transformation of protein sequences. They started with slightly modified Virtual potential (VP) concept from Aires-de-Sousa *et al* (2003), who proposed VP formula for DNA sequences; in GPI-SOM the formula was adapted to Virtual Potentials for amino acids. But finally GPI-SOM authors used principally different formula for calculation of each amino acid occurrence and introduced so called “zentriole”. 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

$Z(A) = 2^{-n}\sum_{i=1}^{n} 2^{i-1} P_{Ai}$, where n is the number of occurrences for a particular amino acid A within 32-residue fragment considered.

For amino acids not occurring in the input sequence, *Z* equals zero.

Several techniques were used by GPI-SOM authors to identify and minimize the number of significant C-terminal residues for GPI-anchoring, and as a result just 22 important amino acid positions in C-terminal protein sequence were selected out of last 32 positions. For a given protein zentriols were calculated, taking into consideration only amino acids in these 22 positions. This approach provided 20 zentriols (some value for each amino acid present and zero for the absent ones). Additional input format for GPI-protein recognition, relative hydrophobicity for those

22 important amino acid position at the protein C-terminus was also proposed and investigated. Finally, the combination of a zentriol for each of the 20 amino acids (occurring in 22 important amino acid positions at the protein C-terminus) with a collinear representation of relative hydrophobicity for each of these 22 positions was accepted as a suitable input format. Already by itself, 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 ($Z + H$).

Final input vector contained 44 components : calculated values from the distribution of all 20 amino acids (zentriols), hydrophobicity values for the 22 important C-terminal amino acid positions and 2 extra units. The two extra-units were added in order to better distinguish GPI-anchoring signals from transmembrane domains, one unit for the quality of a putative ω site and one for its position.

Thus, the final GPI-prediction program, GPI-SOM, was implemented as a Kohonen SOM with an input layer with 44 neurons.

2. Materials and methods

2.1 Annotation of GPI-anchor in the UniProtKB/Swiss-Prot

Architecture of the UniProtKB/Swiss-Prot and of a UniProtKB entry is explained in detail in (Farriol-Mathis *et al.*, 2004) and in the UniProtKB/Swiss-Prot user manual (http://www.expasy.org/sprot/userman.html). To build the datasets required for the evaluation of the GPI-anchor prediction programs, we used annotations provided within entries of the UniProtKB/Swiss-Prot. This work was started with UniProtKB/Swiss-Prot Release 48.1 (September, 2005) and finally updated with new data from Release 49.7 (May, 2006). We required only experimentally proven sequences/entries to be included in datasets.

Entries for GPI-anchored proteins contain information on both the fact of GPI-anchoring and the attachment site. GPI-annotation is represented in several lines – key words (KW), comments (CC) and feature table (FT).

Statement in KW-line on GPI-anchoring is independent of the fact of experimental verification, and no non-experimental qualifiers are added, even if the protein is not experimentally proven to be GPI-anchored. See the example below.

```
KW   GPI-anchor;
```

We considered CC-line as the most informative one for protein GPI-status. Concerning the fact of GPI-anchor modification, CC-line looks as the following:

```
CC    -!-SUBCELLULAR LOCATION: Attached to the membrane by a GPI-
anchor.
```

This kind of CC-line was considered as experimental proof of GPI-anchoring. The entries with non-experimental qualifiers in the end of this line we did not consider as proven, and didn't include them into the positive dataset.

Cleavage site information is indicated in FT   LIPID line. Example is shown below.

```
FT   LIPID         174    174         GPI-anchor amidated asparagine
```

There are only 43 entries with experimentally proven cleavage sites, others contain non-experimental qualifiers in FT LIPID-line.

The protein which is supposed to become GPI-anchored, has to be transported to the endoplasmic reticulum of the cell to be modified with the enzyme. That's why the presence of the signal-peptide, which is indicated in FT   SIGNAL-line is important to consider (see example below).

```
FT   SIGNAL         1     36
```

This kind of entry without any non-experimental qualifier in the FT SIGNAL-line was considered as experimentally verified secreted protein.

2.2 Building datasets

Taking into account considerations above, we built the negative and positive datasets used later to evaluate the program performance. We used grep-function with indicated regular expressions for the purpose of database search.

2.2.1 Positive dataset creation:

First we selected entries from the whole SP database with regular expression

`KW.*GPI-anchor`; ("GPI-anchor biosynthesis." entries have the expression in their key-word line, so are not included). To have only experimentally verified proteins with GPI-anchor, out of those entries we selected ones without:

```
CC.*GPI-anchor\(By.*\s+.*similarity\)
CC.*GPI-anchor.\*\s+.*\(Potential\)
CC.*GPI-anchor.*\s+.*\(Probable\)
```

which can be generalized with `CC.*GPI-anchor.*\s+.*\(` .

In order to avoid introduction of a bias in the evaluation procedure, we subtracted the entries used for program training from the entries extracted from the UniProtKB/Swiss-Prot to build the positive datasets. We did not exclude proteins whose sequence has similarity greater than 90% with proteins of a training dataset, as it is usually done with datasets, because even a point mutation may prohibit GPI-anchoring (Eisenhaber *et al*, 1999). The remaining proteins are those which carry experimentally verified GPI-anchor and which the considered prediction programs have not been trained with.

2.2.2 Dataset with experimentally verified cleavage site.

Entries with proven cleavage site were selected in the following way. Dataset with GPI-anchor key-word `KW.*GPI-anchor;` was checked for the expression below.

```
FT.*LIPID.*\(.*\)\.
```

GPI-entries without the expression were selected (we wanted entries without non-experimantal qualifiers, which are always in parentheses).

2.2.3 Negative dataset creation.

We required protein N-terminal sequencing on amino acid level to make sure that leader peptide existed for these proteins and was properly cleaved. We wanted to make sure, that the protein can be transported to the ER and encounter the proper enzyme.

Again, we started with the whole Swiss-Prot database, and took all entries from SP with the following expression: `FT    SIGNAL\s+\d+\s+\d+\s*\n`

We wanted to exclude entries if any which were annotated to be GPI-anchored. So we selected entries without `KW.*GPI-anchor;`

We wrote a Perl script to find all entries where amino acid protein sequence starts at (+1) amino acid position relative to the end of signal peptide. The result was improved with the addition of entries with `RP.*Sequence of N-terminus`.

In addition, we required proteins in the negative dataset to be secreted. A condition `CC.*SUBCELLULAR LOCATION.*Secreted protein` was imposed.

2.2.4 Special case: transmembrane proteins

We investigated the performance of all three programs on the dataset of transmembrane proteins, which are a known source of false-positive predictions (Dalley and Bulleid, 2003, Fankhauser *et al*, 2005). The dataset was made of proteins, spanning the membrane one time with their N-terminus being extracellular and intracellular (type I and type II transmembrane proteins, respectively) and of proteins, having their C-terminal region imbedded to the membrane (type IV membrane proteins). The following expressions were used for the search:

```
CC   SUBCELLULAR LOCATION: Type I membrane protein\.\n
CC   SUBCELLULAR LOCATION: Type II membrane protein\.\n
```

For type I and II proteins we took only metazoan entries with signal peptide, indicated in FT-line, although allowing non-experimental qualifiers there.

Type IV metazoan proteins were also selected, using the following expression and later filtering for metazoan OC line.

`CC    SUBCELLULAR LOCATION: Type IV membrane protein\.\n` Those entries were not filtered for signal peptide presence, since there is no signal peptide in them, but importance of these entries for program evaluation is significant.

2.2.5 Automated dataset development and update

Since the grep-search is semi-automated and requires human intervention, in the end we decided to develop alternative fully automated tool(s) for the search of GPI-anchored proteins in UniProtKB/Swiss-Prot and for the check of their experimental verification.

Two scripts were developed for this purpose.

The first script is performing the full-text search in UniProtKB/Swiss-Prot, retrieves the entries and saves them in a file in Swiss-Prot format. The output is ready to use by the second script, which filters the entries to retain ones with key-word “GPI-anchor” and to remove afterwards ones with non-experimental qualifiers for GPI-anchor in CC-line, SUBCELLULAR LOCATION. Its output can be AC-list of verified entries or verified entries themselves.

Script 1 (for Perl-code, see appendix)**.**

The search is done with LWP::User Agent perl module and implies three steps.

The first step calls one of the expasy-web-sites for the full-text search in UniProtKB/Swiss-Prot (http://www.expasy.ch/cgi-bin/sprot-search-ful). We propose to search just for “GPI-anchor”.

The second step saves the entries found in a file in the directory `outgoing` of the ExPASy anonymous ftp server.

The third step downloads the file from the ftp-server and saves it on a user-computer.

The second step is possible only if the first one was successful, since the second step uses the URL, retrieved from the first step output. The third step is possible only if the second one was successful and the file with entries found was saved on ftp-server.

Script 2 (for Perl-code, see appendix)**.**

The second script was written to replace multiple calls for grep-function from CRISP-program, which was used initially to build all the datasets. As an input the script uses the file with entries in Swiss-Prot format, so it can be merged with the first script without any modifications.

The major engine of the script is Swiss Knife perl-module. Key element of the script is analysis of KW-line (an array of key-words) and CC-lines (a hash-table), exactly the same idea which was used with grep-function initially, but now with different tools, analyzing several conditions from different entry lines on the same run. We select entries with KW “GPI-anchor” and filter them for non-experimental qualifiers in CC-line, SUBCELLULAR LOCATION, mentioning GPI-anchor.

2.3 Datasets used for development of the GPI-prediction programs.

We analyzed all the datasets, which were used for development of the 3 programs. Those datasets have to be subtracted from the final datasets to give a clean update. We will discuss their contents in Results section.

2.4 Access to GPI-prediction-programs.

Three currently available programs – BigPI, DGPI and GPI-SOM were analyzed. Perl web-scripts, calling corresponding web-sites and analyzing the output from them were written. See Figure 3 for the description of the algorithm of the scripts. Perl-script to call DGPI-program is provided in the appendix.

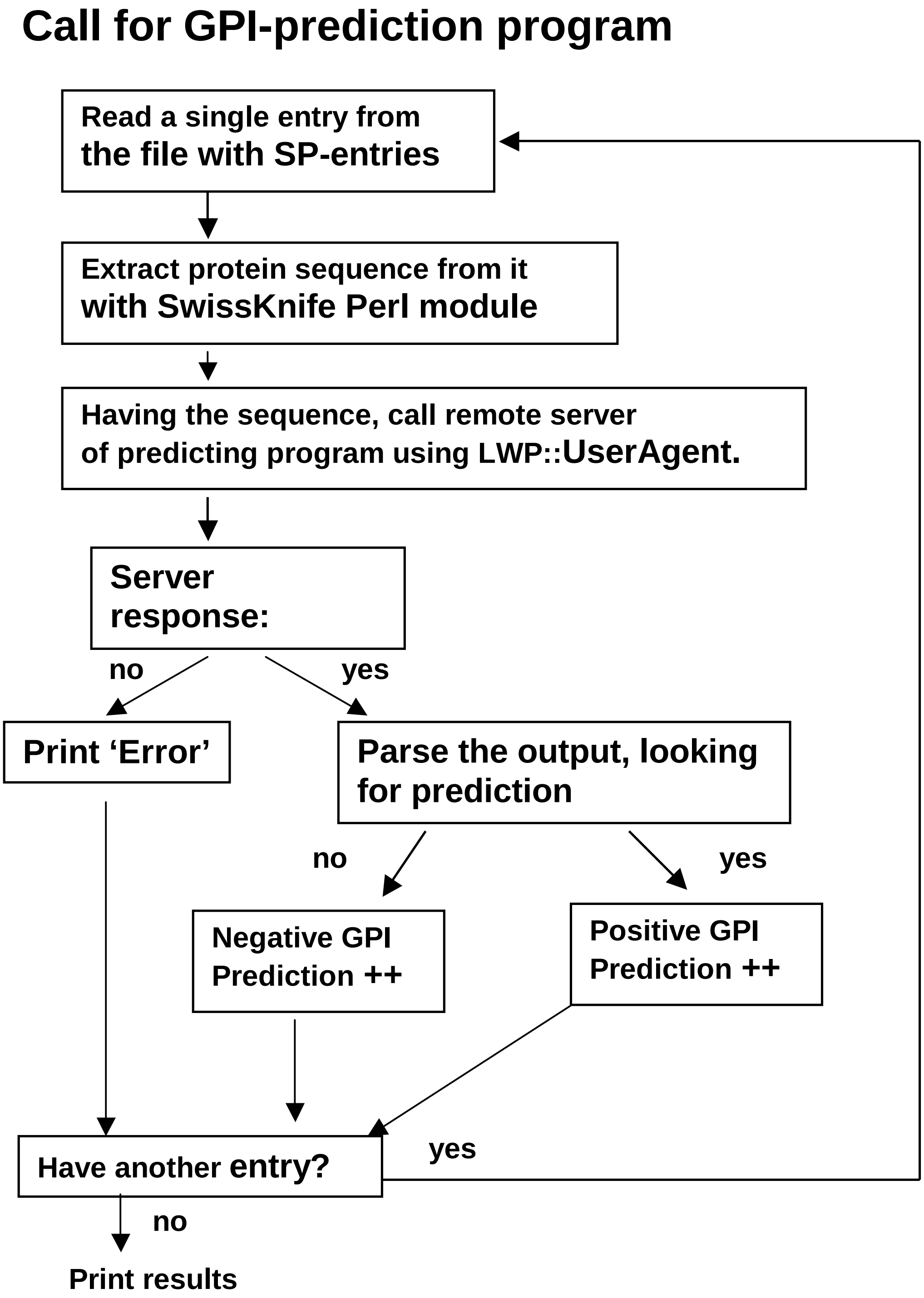

Call for GPI-prediction program
Read a single entry from the file with SP-entries
Extract protein sequence from it with SwissKnife Perl module
Having the sequence, call remote server of predicting program using LWP::UserAgent.
Server response:
no
yes
Print 'Error'
Parse the output, looking for prediction
no
yes
Negative GPI Prediction ++
Positive GPI Prediction ++
Have another entry?
yes
no
Print results

Figure 3. General algorithm used for accessing web-sites of 3 GPI-predicting programs.

For program evaluation, entries from positive and negative datasets were used as an input. Sequence and accession number information were retrieved from the entry with Swiss Knife Perl-module, and then submitted to the web-sites in fasta-format. Slightly different scripts were written to call each prediction program. The key element of all the three scripts was LWP::UserAgent module. When we create a User-Agent object, it connects to a server, sends requests, receives responses and masters the received data.

2.5 Program performance evaluation

In order to assess each prediction program performance, several statistical values were calculated. In the following formulas, abbreviations stand for: TP, true positive; FP, false positive; TN, true negative; FN, false negative.

The sensitivity (*Sn*) is defined as the proportion of true positive items out of all items in a positive set and can be calculated with:

$$Sn = \frac{TP}{TP + FN}$$

The specificity (*Sp*) is the proportion of true negatives out of all negatives and can be calculated with:

$$Sp = \frac{TN}{TN + FP}$$

The accuracy (*Acc*) is the proportion of all true predictions (positive and negative) on positive and negative examples and is given by:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Finally, Mathews' correlation coefficient (*MCC,* Matthews, 1975) is given by:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \quad \text{or}$$

$$MCC = \frac{(TP + FN + TN + FP) \times TP - (TP + FN)(TP + FP)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

It varies between –1 and +1; a value of +1 indicating there is a perfect agreement between observations and predictions; a value of 0 indicating that predictions are not better as they were randomly generated; a value of –1 indicating that predictions are in total disagreement with what is observed.

3. Results

3.1 Datasets.

3.1.1 Analysis of the datasets, used for the program training.

Contents of all positive training datasets, filtered for GPI-verified sequences and UniProtKB/Swiss-Prot accession-numbers, are displayed in Table 1.

BigPI (metazoa/protozoa) dataset consists of 167 sequences, with 136 out of them being experimentally verified (row 1, Table 1); BigPI (fungal) and BigPI (plant) datasets are based on mostly non-verified entries:
plant dataset: 254 total/ 42 experimentally verified (Table 1, row 2);
fungal dataset: 219 total/ 19 experimentally verified (Table 1, row 3).

DGPI authors didn't mention the dataset they used for their program creation, but it was done in Swiss-Prot with Swiss-Prot entries, and approximately at the same time when BigPI was written, so we considered DGPI dataset, used for program development to be the same, as verified BigPI original dataset for metazoa/protozoa, so only subtraction of BigPI training set from final positive dataset was done.

GPI-SOM required both positive (training and validating ones) and negative sets. GPI-SOM positive training set and GPI-SOM positive validation set together consisted of 358 proteins, taken from Genbank, out of which just 87 have 100% sequence identity with some UniProtKB/Swiss-Prot entries, and just 75 out of these 87 are experimentally verified. (Table 1, row 5). Used sequences were subtracted from the final positive dataset. We used different principles to create negative dataset, than those that were used by GPI-SOM authors (for instance, they used cytosolic protein sequences without signal-peptide), so GPI-SOM negative dataset was not taken into account at all.

Winding up, we will give a few comments about training datasets quality. Positive sets used for program creation contain a lot of unverified sequences (see Fig.4). It immediately created some doubts about programs quality, since this is the case for all three programs. Besides this, there are other examples of irregularities. GPI-SOM, non-taxon-specific program, used mostly protein sequences from *Arabidopsis thaliana* as a positive set, which is an obvious bias, which should have been avoided. One of GPI-SOM negative training set consists of proteins without signal peptide, which is not appropriated. If a protein doesn't contain a signal peptide, it cannot be transported to the ER and will never encounter the modifying enzyme. Whatever C-terminal sequence it has, in nature this protein will never become GPI-anchored, although it may contain a proper C-terminal part for anchoring. This approach may lead to false predictions.

Table 1. Taxonomic contents of the training datasets:

Row	Datasets	Metazoa	Protozoa	Fungi	Plants	Total
Positive datasets, used for the program development						
1	BigPI/DGPI meta/prot	96	37	3	0	136
2	BigPI plants	0	0	0	42	42
3	BigPI fungi	0	0	19	0	19
4	Verified ω-site, training	19	15	1	0	35
5	SOM positive	16	14	3	42	75

Table 1 is based only on experimentally verified GPI UniprotKB/Swiss-Prot entries. The training datasets of BigPI/DGPI and GPI-SOM overlap significantly; in Table 1 we showed just experimentally verified Swiss-Prot entries, but a significant percentage of entries in the training datasets used are not verified.

3.1.2 Updated datasets.

Since all the datasets had to be updated, we created our own positive and negative datasets. The goal was to make datasets of protein sequences which programs have never encountered before.

Positive dataset. First we started with “grep”-search procedure, described in “Materials and methods”, but finally we wrote two Perl-scripts to make it automatic (for Perl-code, see Appendix), and updated our previoius results. UniProtKB/Swiss-Prot release 49.7 contain 389 proven GPI-anchored proteins. Its taxonomic contents, before and after subtraction of “already seen” entries, is presented in updated dataset section, Table 2, rows 1 and 3, respectively.

Negative dataset. Proteins with determined N-terminal amino-acid sequence, proven cleaved signal-peptide, with the comment “Secreted” in CC SUBCELLULAR LOCATION line and without any mentioning of GPI-anchor anywhere in corresponding entry were selected. (See row 4, Table 2). The requirement of secretion will give some additional confidence, that the protein is not GPI-anchored, since only the presence of signal peptide does not guarantee it. To verify that, we looked at our positive set of verified metazoan entries – 262. There are 23 entries with comment “Secreted protein”, but either “After cleavage: Secreted protein” or

“Secreted protein” is a comment for a different splice-isoform. So “Secreted protein” is a proper additional selection criteria, which would imply the absence of GPI-anchoring.

As we said in the previous section, for negative dataset we used different principles, than authors of GPI-SOM, that is why our dataset was totally different, and we didn’t need any subtraction.

Transmembrane proteins dataset. See Table 2, rows 6 and 7.

We made two datasets of eukaryotic transmembrane proteins, as described in “Materials and methods”. The first one – for non-plant type I and II transmembrane proteins with signal peptide (1156 entries), the second one – for non-plant type IV transmembrane proteins without signal peptide (29 entries).

Dataset of proteins with verified GPI-cleavage site. Unfortunately, only 8 entries out of 43 eukaryotic proved cleavage site entries were not used in a training set, so this studies don’t really show if the cleavage site prediction is reliable or not. For taxonomic contents see Table 1, row 4 (training set) and Table 2, row 2 (total set before subtraction) and row 5 (evaluation set).

Table 2. Taxonomic contents of newly created datasets.

Row	Datasets	Metazoa	Protozoa	Fungi	Plants	Total
Updated datasets before subtraction						
1	Positive	262	66	19	42	389
2	Verified ω-site	26	16	1	0	43
Updated datasets after subtraction						
3	Positive	158	27	0	0	185
4	Negative	227	0	17	13	257
5	Verified ω-site, evaluat	7	1	0	0	8
6	Transmem I & II, SIG	1143	10	3	0	1156
7	Transmem IV, NO SIG	23	3	3	0	29

Table 2 is based only on experimentally verified GPI UniprotKB/Swiss-Prot entries.

Highlighted column in Table 2 shows updated metazoan datasets, later used for program evaluation. Table 2 clearly demonstrates the lack of verified entries for plants and fungi, and limited amount of protozoan entries, that is why we decided to proceed just with metazoan datasets.

On the way of creation updated datasets, we encountered a number of difficulties. There were two main problems for the positive set:

- alternative splicing

Analyzing positive entries with “GPI-anchor” key word, we encountered a problem with alternative splicing. Sometimes the sequence displayed in the entry was not the sequence of the spliced isoform, which would become GPI-anchored, there was just a reference to the anchored isoform inside the entry. We removed these entries from the positive set. Also it was not stated clearly for a few times which isoform is anchored. As far as we know, this problem has been already fixed by curators (Nathalie Farriol-Mathis, personal communications).

- lack of verified entries

The general difficulty, which we encountered – the small number of experimentally proved entries with GPI-anchor, especially for fungi and protozoa.

The problem with negative set is that no scientist ever states, that the protein is not GPI-anchored. To conclusively prove that a site is negative under all conditions is impossible, but to know that it is negative at least in some contexts would be useful. Since there is no experimental verification of the fact, that the protein is not GPI-anchored, we based our negative set construction on the entries, where the presence of signal-peptide was experimentally verified by N-terminal amino acid sequencing and without any mentioning of GPI-anchoring, and selected only secreted proteins from these entries. None of those entries contain mass spectrometry measurements, which would exclude the possibility of GPI-anchoring, which is unfortunate.

3.2 Evaluation of the GPI-anchor prediction programs

GPI-anchor prediction programs are involved in the annotation procedure of the UniProtKB/Swiss-Prot, so their evaluation is of high importance for providing a high-quality annotated database. The results obtained are presented in the following paragraphs. There is different taxonomic specificity for existing GPI-prediction programs in the way they were developed. BigPI implies different versions for metazoa, protozoa, plants and fungi, DGPI and GPI-SOM make no difference between different taxonomies. For protozoa we observed very few entries, and all known verified entries for fungi and plants were used for the development of at least one of the three programs, that is another reason why the evaluation was performed only for the most numerous dataset – metazoan. It allowed us to compare conclusions of metazoan-specialized (BigPI) and non-specialized (DGPI and GPI-SOM) programs as well, and also to compare results of BigPI programs, written for different taxonomies than the datasets used.

First we compared three programs performance for the fact of GPI-anchoring on positive and negative metazoan datasets and calculated sensitivity, specificity, accuracy and Matthew's correlation coefficient for the programs (see "Materials and methods", section 2.5 for the formulas). Results are presented in Table 3.

Table 3. Prediction of the fact of GPI-anchoring by BigPI, DGPI and GPI-SOM on metazoan datasets.

	POSITIVE SET		NEGATIVE SET		Sensitivity	Specificity	Matt cor cf	Accuracy
	TP	**FN**	**FP**	**TN**	**Sn**	**Sp**	**MCC**	**ACC**
BigPI	93	65	0	227	0.589	1.000	0.676	0.831
DGPI	119	39	14	213	0.753	0.938	0.715	0.862
GPI-SOM	122	36	12	215	0.772	0.947	0.743	0.875

Calculations of sensitivity/specificity show, that only BigPI satisfies the requirements of very low FP-value. Let's not forget, that BigPI is a specialized program, which distinguish between taxonomic groups, and it seems to be a clear advantage. BigPI is currently used in Anabelle (a tool in UniProtKB/Swiss-Prot annotation platform which gives curators ways to choose the right decision concerning a particular protein feature) for GPI-anchor prediction. BigPI is the most specific program, and partially because of that the least sensitive. DGPI and GPI-SOM are less specific (0.95), but subsequently more sensitive. More than 5% of false positive predictions is a lot, that is why the other two programs are not used in UniProtKB/Swiss-Prot curation procedure. Accuracy for all three programs is good and have about the same value for all three, about 0.85; Matthews correlation coefficients (MCC) are also similar, about 0.7. MCC is symmetric with respect to false positive and false negative values, Fp and Fn, and since it is almost equal for all the programs, we expected, that programs, which gain in specificity, would lose in sensitivity and visa versa (basic statistical property).

That's interesting, that DGPI and GPI-SOM have close sensitivity and specificity values, although are based on different principles.

3.3 Cleavage site prediction comparison

All three programs do prediction of the ω-site, and we wanted to proceed with the evaluation of this feature as well. There are just 26 verified metazoan entries. Program performance for them is presented in Table 4.

Table 4. Summary of the cleavage site prediction results (metazoa only):

	DGPI	BigPI	GPI-SOM
Evaluation set : 7 sequences			
Correct site prediction	4	6	4
Incorrect site prediction	2	1	3
NO site	1	0	0
Total	7	7	7
Training set : 19 sequences			
Correct site prediction	13	17	10
Incorrect site prediction	5	2	9
NO site	1	0	0
Total	19	19	19
Sum of both sets : 26 sequences			
Correct site prediction	17	23	14
Incorrect site prediction	7	3	12
NO site	2	0	0
Total	26	26	26

We looked at the site prediction regardless of the result of GPI-anchoring prediction, since they are independent for programs. All three programs may predict the site correctly or incorrectly or give no site prediction for the protein sequence (three possibilities). Comparison of the programs results for the training and evaluation cleavage site sets did not show a dramatic difference. When we look just at the results for 7 entries, which were not used for the program development (see Table 4, results for evaluation set), we see that all the programs make mistakes in the site prediction. But there is roughly the same proportion of wrong site predictions for the training set of BigPI and DGPI as well, and a little bit less than 50% of false predictions for GPI-SOM (see predictions for training set). That is why we decided to make conclusions about this feature, looking at the whole dataset of 26 metazoan proteins with verified ω-site (see predictions for the sum of both sets).

We didn't calculate sensitivity and specificity for cleavage site predictions, since there are very few verified sites and it wouldn't have any statistical meaning, so we just looked at the percentage of correct answers.

BigPI predicts ω-site correctly in about 90% of the cases. DGPI is also a good program to predict ω-site (about 65% of correct predictions), but in more than one third of the cases it is wrong, and a couple of times it cannot recognize the sites at all. GPI-SOM is the worst program out of these three in site predictions, it is wrong in about 45% of the cases.

Unfortunately, there are not enough entries to state something about statistical significance of these results, but overall conclusion is clear – more data are needed.

There are some confusing conclusions which all three programs make, since all three predict the fact of GPI-anchoring and the corresponding cleavage site(s) independently. So the conclusion of the program can be "No GPI-anchoring", but it still predicts the site of cleavage. It is a really useful thing that the programs do, because it may give some clues why the prediction was negative (BigPI, for example, gives numerical scores for each physical term), and allow even "in silico" mutations of the protein. Certainly, the programs have to be upgraded for this kind of studies, but this is just a very useful initial step.

3.4 Transmembrane proteins

We also checked if three investigated programs make mistakes for transmembrane proteins, which are not GPI-anchored. The dataset was built, according to the rules from "Materials and methods" section. Fankhauser *et al* argued, that since both C-terminal transmembrane proteins and proteins, carrying GPI-anchor, have highly hydrophobic C-terminus, GPI-prediction programs sometimes give false positive predictions for such proteins. To evaluate this assumption, we looked at program perfomance on typeI/typeII transmembrane protein dataset (see Table 5). Since transmembrane proteins form another negative set, here we can talk about "transmembrane specificity". We cannot add transmembrane proteins to our main negative dataset (Table 2, line 4), because they were built according to different principles. The only Fp prediction for BigPI is P10379. In the corresponding Swiss-Prot entry there is no indication of the fact of GPI-anchoring for this protein.

Table 5. "Specialized" specificity, calculated on transmembrane proteins, type I and II.

	TRANSMEMBRANE NEGATIVE SET		SPECIFICITY
	FP	TN	SP
BigPI	1	1142	0.9991
DGPI	59	1084	0.9484
GPI-SOM	71	1072	0.9379

We also made a dataset of metazoan type IV transmembrane proteins. There are just 23 entries selected, and these proteins don't carry signal peptide (that is why they are rejected by DGPI and GPI-SOM). To demonstrate significance for the programs the presence of signal peptide in the sequence, we attached artificial signal to all 23 sequences and ran prediction programs again. Results are presented on the diagram 1.

BigPI doesn't look at the presence of signal peptide, just at the C-terminal region, so predictions are the same with and without the signal. It is interesting, that three entries, Q9UNK0, O88983 and Q9Z2Q7, are predicted to be GPI-anchored according to BigPI with good scores. This is the case when closer look is necessary, since those proteins can potentially carry GPI-anchor despite being transmembrane, and these are interesting cases when the very good performance program – BigPI – gives some false positive predictions.

If we compare predictions by DGPI and GPI-SOM on the same sequences, but with artificial signal and without signal, we see, that these two programs make their conclusions about GPI-anchoring, based mostly on the presence of N-terminal signal, and GPI-signal itself is almost ignored.

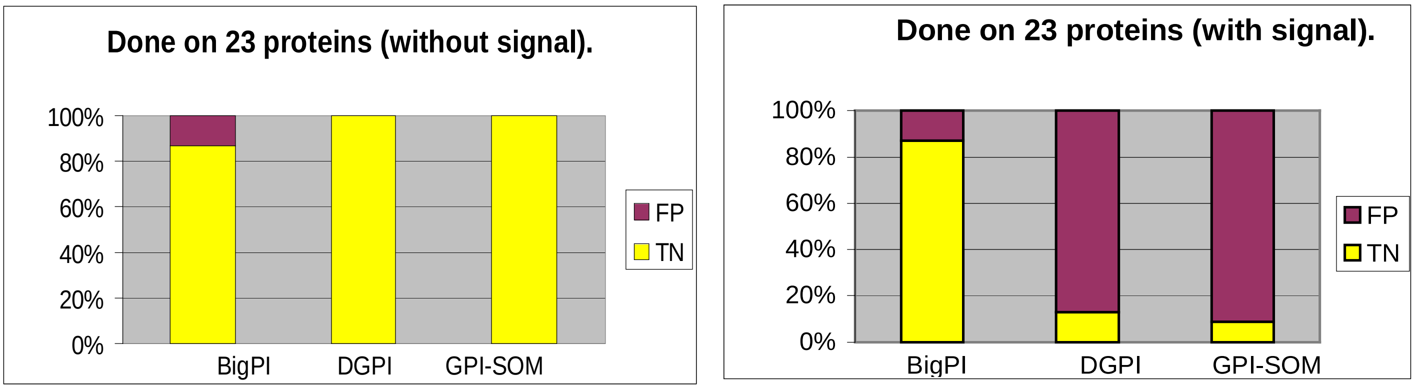
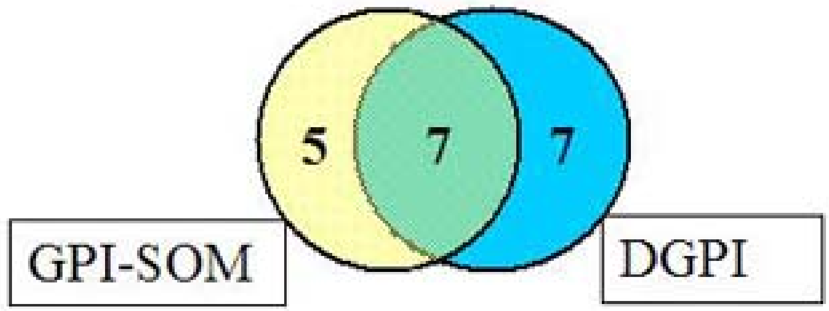


Diagram 1. Comparison of GPI-predictions on Type IV membrane proteins with and without artificial signal sequence.

3.5 Comparative analysis of program performance

We thought, it could be useful to see, if there are some tendencies in mutual mistakes, like –"if one program prediction is wrong for some sequence, the second program prediction for this sequence is also always wrong". First we looked at comparison of Fp-predictions between 3 programs on metazoan negative set. See Figure 5 and Table A1 in Appendix for the details.

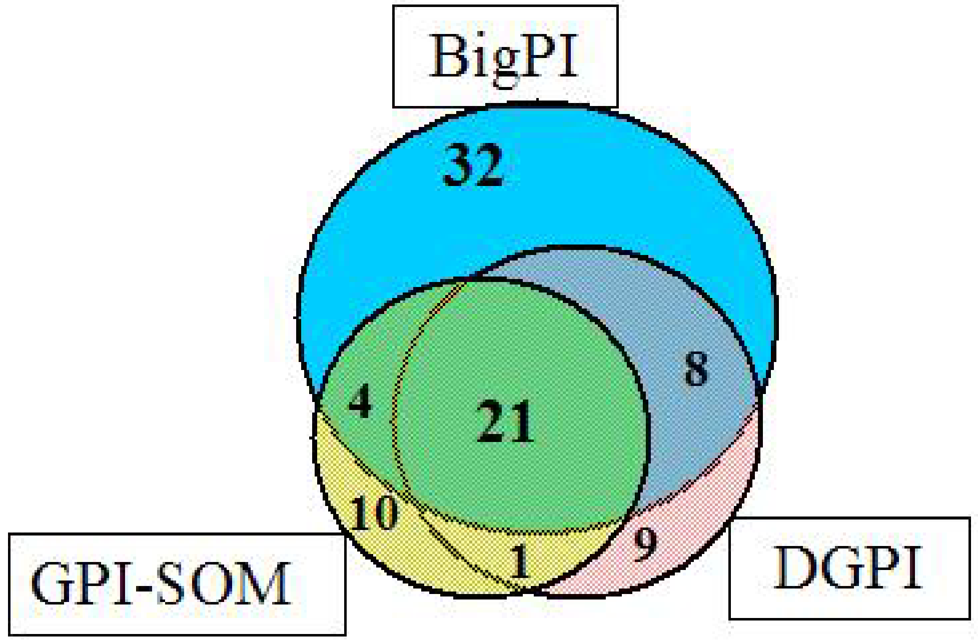
Figure 5. Relations between programs for Fp-predictions.



BigPI is the best program, no false positive prediction for 227 negative entries. Although two other programs, GPI-SOM and DGPI, are based on different principles, total number of Fp-predictions they give is about the same (12 and 14, respectively, for metazoan negative set). Some entries (7) are the same for both programs, the other incorrectly predicted entries are different. The reason of both phenomena can be the target of further investigation and improvement of the programs.

We also looked at comparison of Fn-predictions between three programs on metazoan positive set. See Figure 6 and Table A2 in Appendix. The goal is to compare Fn prediction and to figure out, which predictions are shared and try to understand what it can be useful for.

Figure 6. Relations between programs for Fn-predictions.



When we look at the Figure 6 and compare cases of negative answers for the positive set (see also table A2 in Appendix), we see, that DGPI and GPI-SOM share 21 Fn predictions with BigPI, and have groups of 9 and 10 Fn-predictions for DGPI and GPI-SOM, respectively, unshared with any other program. Comparing overlap between DGPI/BigPI (8) and GPI-SOM/BigPI (4) Fn-predictions, the impression is that DGPI approach is closer to BigPI one, than GPI-SOM to BigPI. Of course, there are very few data to make this conclusion, but what we know about the program algorithms, may suggest that it is right to some extent. What is an interesting fact - DGPI and GPI-SOM share only 1 Fn-prediction, while sharing many false positives (Figure 5). It may suggest, that programs accept sequences according to similar principles, but reject them because of different criteria. Another striking thing, which was expected, is a lot of Fn predictions by BigPI alone (32), what is a directly linked to its very high specificity value.

The most interesting thing, which could be done here is an investigation of correlation between the program conclusions (between scores). Unfortunately, only one program, BigPI, gives numerical score, so it was not possible.

Another useful thing, which can be derived from this kind of analysis, is the following. Let's look at the BigPI case and analyze whether the program for one taxonomic group can be used for another one.

3.6 Evaluation of improvement provided by taxon-specific predictor

Before we started our work, there was just metazoan/protozoan BigPI-predictor, implemented in Anabelle, and metazoan predictor was used for protein sequences from plants and fungi. Now all four BigPI programs are implemented in Anabelle. We decided to measure the improvement which was acquired with the introduction of taxon-specific programs.

Table 6. BigPI predictions for the verified plant set.

	NEW predictor	OLD predictors	
	BIG PI plants	BIG PI metazoa	BIG PI protozoa
Tp-prediction	30	20	13
Fn- prediction	12	22	29
Total	42	42	42

We analyzed the relations between taxon-specific BigPI-predictors on 42 verified plant sequences (Eisenhaber et al, 2003, Borner et al). Metazoan/protozoan (old) predictor results were compared with plant (new) predictor result for each sequence. All the plants sequences predicted by metazoan program are predicted by plant program, but not visa versa – some of the sequences, predicted by plant program, are not recognized by metazoan one. For this dataset, metazoan predictor never says that the protein is not GPI-anchored, when plant predictors says that it is anchored, it verifies former use of metazoan predictor for plants.

Metazoan predictor recognizes 20 of 42 proven sequences, 9 of which are shared with protozoan predictor. Although protozoan predictor recognizes less sequences as GPI-anchored (13 total), it does recognize 4 additional sequences which metazoan predictor doesn't. What is important – all 24 sequences, predicted by meta/pro BigPI for plants are recognized by plant BigPI, and 6 additional sequences are identified. Clear advantage of taxon-specific program use is observed. However it has to be noticed, that those 42 sequences were in plant-BigPI training set. BigPI is not a program, based on mashine-learning, so training implies for it just choice of a threshold to minimize Fp-rate, so we think we can make some conclusions based on these "training" data.

Nowadays plant BigPI is already implemented in Anabelle, so this investigation is just confirming that is was useful to do it, because it is a huge improvement

We performed the same evaluation for verified fungi set (Eisenhaber et al, 2003). The problem was that just 19 fungi sequences in UniProtKB/Swiss-Prot are proven to be GPI-anchored, but let's see what we had for them.

There are now at least 43 verified fungi-entries known, according to Eisenhaber et al, 2004, simply in the UniProtKB/Swiss-Prot most of them are still marked with non-experimental qualifiers in CC-line. To be consistent, we decided to proceed only with those proven 19 sequences from UniProtKB/Swiss-Prot and look, as before for plant sequences, what is the difference between old and new BigPI-predictors.

Table 7. BigPI predictions for the verified fungi set.

	NEW predictor	OLD predictors	
	BIG PI fungi	BIG PI metazoa	BIG PI protozoa
Tp prediction	18	11	4
Fn prediction	1	8	15
Total	19	19	19

We analyzed the relations between taxon-specific BigPI-predictors on 19 verified fungal entries. Metazoan/protozoan predictor results were compared with fungal predictor result for each sequence.

Protozan predictor gives too many false negatives to be applicable for fungi, just 4 fungal sequences are recognized. These 4 sequences are shared with metazoan predictor, so probably there was no point to use protozoan predictor for fungi. In its turn, metazoan predictor recognizes 11 sequences from this fungal set. But fungal predictor performance is very good for the fungal set, it recognizes 18 out of 19 sequences. Those 18 include 11 sequences, recognized by metazoan predictor. Again, there is a clear advantage in using fungal predictor for fungi – much more true positive predictions, than would be for metazoan predictor on fungal set. It should be noticed, that 19 sequences are from the fungal BigPI training set. Fungal predictor is now also implemented in Anabelle.

4. Discussion and concluding remarks.

Let's make some overview of all GPI-prediction programs from the user position. None of the programs seem to be really user friendly.

- **BigPI**
  - Doesn't allow to download a file with sequences, just cut and paste fasta-formatted sequences into the on-line form
  - Doesn't check for the presence of a signal peptide in proprotein sequence
  - There is a critical limit of 500 amino acids per line; if more, the program just ignores the rest of the sequence without further warnings
  - No fasta-header in plant- and fungi-output predictions, just the number of amino acids in the sequence submitted
- **DGPI**

- Allows to download just a single protein sequence, although in variety of formats
- Doesn’t give the resulting score, according to which the prediction is made

- **GPI-SOM**
    - Doesn’t give any numerical scores, just a graphical map
    - Shows the cleavage site position from the C-terminus of the protein sequence

The properties listed make the programs inconvenient for users and have to be fixed sooner or later.

Particular results, especially numerical results, were already discussed in the Result-section in the end of each sub-section.

The two scripts for the automatic GPI-dataset update are to be integrated into the generic procedure of updating Biosapiense datasets.

Evaluation of all three programs was performed before just once by Fankhauser et al, 2005, the authors of the program GPI-SOM, The datasets they used for their evaluation were not composed of verified GPI-anchored proteins.

To conclude, we can list the following items:

Evaluation sets suffer from insufficient number of verified data in the database. We considered possible to perform the evaluation only for metazoan datasets.

For metazoa the best prediction program is BigPI – no Fp. It has a special web-site, dedicated to metazoan sequences, which is a clear advantage over non-specialized programs. Exact cleavage site prediction performance is difficult to evaluate – lack of independent dataset.

Comparison of metazoan and plant BigPI-predictor performance for verified plant dataset, and metazoan and fungal BigPI-predictor performance for verified fungal dataset showed, that taxon-specialized BigPI predictors give a huge improvement in GPI-anchoring prediction. Metazoan predictor never contradicts with negative predictions of specialized predictors on corresponding datasets, but gives more false negative predictions.

All the programs make mistakes with transmembrane proteins, BigPI - mostly for type IV transmembrane proteins, and DGPI and GPI-SOM make a lot of false-positive predictions for type I and II transmembrane proteins.

The kind of program evaluation performed shows clearly a “chicken and egg” problem. Good prediction program can be written only if good datasets are provided, but datasets themselves, at least indirectly, depend on prediction programs. Ideally the programs should give experimental scientists some hints, which proteins to investigate to enlarge those datasets.

The Swiss-Prot annotation as we see it is a process of extraction of information about the protein from scientific literature together with its gathering/collection from other sources, which lead to the information digestion, rearrangement and reassessment and finally give rise to a comprehensive description of the protein. Databases like Swiss-Prot, which are committed to contain annotated entries, are of great importance to experimental biologist community, since their existence make scientist life easier. Scientific database contents should be experimentally proven, or at least can be trusted. The extent of this trust has to be clearly indicated (an indispensable role which non-experimental qualifiers play in Swiss-Prot entries). The amount of experimentally verified information available is growing, but still limited. Prediction programs are supposed to partially fill the existing gaps, at least for the time being. Evaluation of this kind of programs should lead to the choice of a proper technique if any, which will allow minimal amount of false positive predictions, mistakes, which may cause at least a significant loss of efforts. False negative prediction can be considered just as a temporary loss of information until new data arrival, and not as a mistake.

Acknowledgements.

Many-many thanks to Dr.Anne-Lise Veuthey (my supervisor), Dr.David Perret, Dr.Patricia Palagi and Prof. Amos Bairoch, as well as to all the other colleagues from the Swissprot PTM-group and the whole SwissProt. It was a great pleasure for me to work inside the community for some period of time (unfortunately, a very short period!).

REFERENCES

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.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* Jan 1;32 (Database issue):D115-9.

Apweiler R., Bairoch A., Wu C.H. (2004) Protein sequence databases. *Curr. Opin.Chem. Biol.* 8:76-80.

Attwood, T., *Int. J. Biochem. Cell Biol.* 2000, *32*, 139–155.

Bailey CA, Howard A, Micanovic R, Berger J, Heimer E, Felix A, Gerber L, Brink L, Udenfriend S. Site-directed antibodies for probing the structure and biogenesis of phosphatidylinositol glycan-linked membrane proteins: application to placental alkaline phosphatase. Anal Biochem. 1988 May 1;170(2):532-41.

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt).Nucleic Acids Res. 2005 Jan 1;33 (Database issue):D154-9.

Bairoch A. (2000) Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times! *Bioinformatics* 16:48-64.

Bairoch A., Boeckmann B., Ferro S., Gasteiger E. (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinform.* 5:39-55.

Beghdadi-Rais C, Schreyer M, Rousseaux M, Borel P, Eisenberg RJ, Cohen GH, Bron C, Fasel N. (1993) Carboxyl terminus structural requirements for glycosylphosphatidylinositol anchor addition to cell surface proteins. J of Cell Science 105 831-840.

Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics. 2004 Jun;4(6):1633-49. Review.

Boeckmann B, Blatter MC, Famiglietti L, Hinz U, Lane L, Roechert B, Bairoch A. Protein variety and functional diversity: Swiss-Prot annotation in its biological context.C R Biol. 2005 Oct-Nov;328(10-11):882-99. Epub 2005 Jul 28. Review.

Boeckmann B., Bairoch A., Apweiler R., Blatter M.C., Estreicher A., Gasteiger E.,Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M. (2003) The UniProtKB/Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003.*Nucleic Acids Res.* 31:365-370.

Bork, P., Dandekar T., Diaz-Lazcoz, Y., Eisenhaber, F. *et al.*, *J. Mol. Biol.* 1998, *283*, 707–725.

Borner GH, Lilley KS, Stevens TJ, Dupree P. Identification of glycosylphosphatidylinositol-anchored proteins in Arabidopsis. A proteomic and genomic analysis.Plant Physiol. 2003 Jun;132(2):568-77. Epub 2003 May 1.

Bucht G, Hjalmarsson K. Residues in Torpedo californica acetylcholinesterase necessary for processing to a glycosyl phosphatidylinositol-anchored form. Biochim Biophys Acta. 1996 Feb 8;1292(2):223-32.

Butikofer P, Vassella E, Ruepp S, Boschung M, Civenni G, Seebeck T, Hemphill A, Mookherjee N, Pearson TW, Roditi I. Phosphorylation of a major GPI-anchored surface protein of Trypanosoma brucei during transport to the plasma membrane.J Cell Sci. 1999 Jun;112 (Pt 11):1785-95.

Caras IW, Weddell GN, Davitz MA, Nussenzweig V, Martin DW Jr. Signal for attachment of a phospholipid membrane anchor in decay accelerating factor. Science. 1987 Nov 27;238(4831):1280-3.

Cross G.(1990) Glycolipid anchoring of plasma membrane proteins. Annu Rev Cell Biol 6,1-39

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.

Eisenhaber B, Bork P, Eisenhaber F. Post-translational GPI lipid anchor modification of proteins in kingdoms of life: analysis of protein sequence data from complete genomes .Protein Eng. 2001 Jan;14(1):17-25.

Eisenhaber B, Bork P, Eisenhaber F. Prediction of potential GPI-modification sites in proprotein sequences.J Mol Biol. 1999 Sep 24;292(3):741-58.

Eisenhaber B, Eisenhaber F, Maurer-Stroh S, Neuberger G. Prediction of sequence signals for lipid post-translational modifications: insights from case studies. Proteomics. (2004a)Jun;4(6):1614-25.

Eisenhaber B, Schneider G, Wildpaner M, Eisenhaber F. A sensitive predictor for potential GPI lipid modification sites in fungal protein sequences and its application to genome-wide studies for Aspergillus nidulans, Candida albicans, Neurospora crassa, Saccharomyces cerevisiae and Schizosaccharomyces pombe. J Mol Biol. (2004b) Mar 19;337(2):243-53.

Eisenhaber B, Wildpaner M, Schultz CJ, Borner GH, Dupree P, Eisenhaber F. Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for Arabidopsis and rice. Plant Physiol. 2003 Dec;133(4):1691-701.

Eisenhaber F, Eisenhaber B, Kubina W, Maurer-Stroh S, Neuberger G, Schneider G, Wildpaner M. Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-Pi, NMT and PTS1. Nucleic Acids Res. 2003 Jul 1;31(13):3631-4.

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.

F. Eisenhaber, B. Eisenhaber, and S. Maurer-Stroh. (2003) Prediction of Post-Translational modifications from Amino Acid Sequence: Problems, Pitfalls and Methodological Hits. In Bioinformatics and Genomes: Current Perspectives.

Fankhauser N, Maser P. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. Bioinformatics. 2005 May 1;21(9):1846-52. Epub 2005 Feb 2.

Farriol-Mathis N, Garavelli JS, Boeckmann B, Duvaud S, Gasteiger E, Gateau A, Veuthey AL, Bairoch A. Annotation of post-translational modifications in the Swiss-Prot knowledge base.Proteomics. 2004 Jun;4(6):1537-50.

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.

Field MC, Menon AK, Cross GA. A glycosylphosphatidylinositol protein anchor from procyclic stage Trypanosoma brucei: lipid structure and biosynthesis. EMBO J. 1991 Oct;10(10):2731-9.

Field MC, Menon AK, Cross GA. Developmental variation of glycosylphosphatidylinositol membrane anchors in Trypanosoma brucei. In vitro biosynthesis of intermediates in the construction of the GPI anchor of the major procyclic surface glycoprotein.J Biol Chem. 1992 Mar 15;267(8):5324-9.

Furukawa Y, Tsukamoto K, Ikezawa H. Mutational analysis of the C-terminal signal peptide of bovine liver 5'-nucleotidase for GPI anchoring: a study on the significance of the hydrophilic spacer region. Biochim Biophys Acta. 1997 Sep 4;1328(2):185-96.

Garavelli, J., *Nucleic Acids Res.* 2003, *31*, 499–501.

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.

Haas R, Jackson BC, Reinhold B, Foster JD, Rosenberry TL. Glycoinositol phospholipid anchor and protein C-terminus of bovine erythrocyte acetylcholinesterase: analysis by mass spectrometry and by protein and DNA sequencing. Biochem J. 1996 Mar 15;314 (Pt 3):817-25.

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

Han, K. K., Martinage, A., *Int. J. Biochem.* 1992, *24*, 1928.

Hobohm U, Scharf M, Schneider R and Sander C (1992) Protein science, 1, 409-417.

Hooper NM. Detergent-insoluble glycosphingolipid/cholesterol-rich membrane domains, lipid rafts and caveolae (review). Mol Membr Biol. 1999 Apr-Jun;16(2):145-56. Review.

Horejsi V, Drbal K, Cebecauer M, Cerny J, Brdicka T, Angelisova P, Stockinger H. GPI-microdomains: a role in signalling via immunoreceptors. Immunol Today. 1999 Aug;20(8):356-61. Review.

Howard AD, Berger J, Gerber L, Familletti P, Udenfriend S. Characterization of the phosphatidylinositol-glycan membrane anchor of human placental alkaline phosphatase.Proc Natl Acad Sci U S A. 1987 Sep;84(17):6055-9.

Jensen, L., Gupta, R., Blom, N., Devos, D. *et al.*, *J. Mol. Biol.*2002, *319*, 1257–1265.

Jensen, L., Ussery, D., Brunak, S., *Genome Res.* 2003, *13*, 2444–2449.

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) http://129.194.185.165/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.

Masterson, W. J., Doering, T, L, Hart, G. Wo and Englund, P. (1989) *Cell* 56, 793-800

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.

Micanovic R, Bailey CA, Brink L, Gerber L, Pan YC, Hulmes JD, Udenfriend S. Aspartic acid-484 of nascent placental alkaline phosphatase condenses with a phosphatidylinositol glycan to become the carboxyl terminus of the mature enzyme. Proc Natl Acad Sci U S A. 1988 Mar;85(5):1398-402.

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

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

Ogata S, Hayashi Y, Misumi Y, Ikehara Y. Membrane-anchoring domain of rat liver 5'-nucleotidase: identification of the COOH-terminal serine-523 covalently attached with a glycolipid. Biochemistry. 1990 Aug 28;29(34):7923-7.

Ogata S, Hayashi Y, Takami N, Ikehara Y. Chemical characterization of the membrane-anchoring domain of human placental alkaline phosphatase. J Biol Chem. 1988 Jul 25;263(21):10489-94.

Omaetxebarria MJ, Hagglund P, Elortza F, Hooper NM, Arizmendi JM, Jensen ON. Isolation and Characterization of Glycosylphosphatidylinositol-Anchored Peptides by Hydrophilic Interaction Chromatography and MALDI Tandem Mass Spectrometry. Anal Chem. 2006 May 15;78(10):3335-3341.

Roberts WL, Myher JJ, Kuksis A, Low MG, Rosenberry TL. Lipid analysis of the glycoinositol phospholipid membrane anchor of human erythrocyte acetylcholinesterase. Palmitoylation of inositol results in resistance to phosphatidylinositol-specific phospholipase C. J Biol Chem. 1988 Dec 15;263(35):18766-75.

Roberts WL, Santikarn S, Reinhold VN, Rosenberry TL. Structural characterization of the glycoinositol phospholipid membrane anchor of human erythrocyte acetylcholinesterase by fast atom bombardment mass spectrometry. J Biol Chem. 1988 Dec 15;263(35):18776-84.

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

Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations.Protein Eng. 1999 May;12(5):387-94.

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.

Taguchi R, Hamakawa N, Maekawa N, Ikezawa H.J. Application of electrospray ionization MS/MS and matrix-assisted laser desorption/ionization-time of flight mass spectrometry to structural analysis of the glycosyl-phosphatidylinositol-anchored protein. Biochem (Tokyo). (1999a) Aug;126(2):421-9.

Taguchi R, Yamazaki J, Takahashi M, Hirano A, Ikezawa H. Identification of a new glycosylphosphatidylinositol-anchored 42-kDa protein and its C-terminal peptides from bovine erythrocytes by gas chromatography-, time-of-flight-, and electrospray-ionization-mass spectrometry. Arch Biochem Biophys. (1999b) Mar 1;363(1):60-7.

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

Udenfriend and Kodukula (1995b) How glycosylphosphatidylinosital-anchored membrane proteins are made. Annu. Rev. Biochem. 1995, 64:563-91

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

Wang J, Maziarz K, Ratnam M. Recognition of the carboxyl-terminal signal for GPI modification requires translocation of its hydrophobic domain across the ER membrane.J Mol Biol. 1999 Mar 12;286(5):1303-10.

Wang J, Shen F, Yan W, Wu M, Ratnam M. Proteolysis of the carboxyl-terminal GPI signal independent of GPI modification as a mechanism for selective protein secretion. Biochemistry. 1997 Nov 25;36(47):14583-92.

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. The Universal Protein Resource (UniProt): an expanding universe of protein information.Nucleic Acids Res. 2006 Jan 1;34(Database issue):D187-91.

Yan W, *et al* (1998) The hydrophobic domains in the Carboxyl-terminal Signal for GPI modification and in the Amino-terminal leader peptide have similar structural requirements. JMB, 275. 25-33.

7.**Appendix.**

7.1 Perl-scripts

```
### Script 1.
# The program is supposed to provide update to the GPI-anchor
# positive datasets. It was written for detection of GPI-anchored
sequences,
# but can easily be adapted to other protein modifications


#!/usr/bin/perl

use strict;

# Modules used
use LWP::UserAgent;
use HTTP::Cookies;
my $ua;
my $line;
my $file_number;
my $file="new_GPI.txt";
my $file_format="sp";
my $key="GPI-anchor";
my $selection_yes="on";
my $wait;

$ua = new LWP::UserAgent;

# Cookies
    $ua->cookie_jar(
        HTTP::Cookies->new(
            file => 'mycookies.txt',
            autosave => 1
        )
    );

# First round - find references to all the entries in the current
release,
# which have the expression “GPI-anchor” anywhere. We will filter
them
# again with another script for experimentally verified entries.

      my $WEB = 'http://www.expasy.ch/cgi-bin/sprot-search-ful';
      my $req = new HTTP::Request POST => $WEB;
      $req->content_type('application/x-www-form-urlencoded');
      $req->content("SEARCH=$key&S=$selection_yes");
      my $res = $ua->request($req);
      if ($res->is_success) {
            my $output1 = $res->content;
#           $output1 =~ s/<.*?>//g;
            print $output1;
# parse the output to find the current file number
            if($output1=~/\/cgi-bin\/sprot_to_ftp\.pl\?(\d+)/) {
                  $file_number=$1;
                  sleep(1);

            # Second round - saving entries found in SP-format in the
file
            # with provided name on ongoing expasy-server
            my $srv_ftp='http://www.expasy.ch/cgi-
bin/sprot_to_ftp.pl?'.$file_number;

#sleep for some time to let the job finish
            sleep(1);
            $req = new HTTP::Request POST => $srv_ftp;
            $req->content_type('application/x-www-form-urlencoded');
            $req->content("filename=$file&format=$file_format");
            $res=$ua->request($req);
#           print $res->content;
            if ($res->is_success) {
                  my $output2 = $res->content;
#                 $output2 =~ s/<.*?>//g;
                  print $output2;
                  sleep(1);

                  # Third round - download the file with entries.
                  # URL for GET should be absolute!
                        $req = HTTP::Request->new(GET =>
                  'ftp://www.expasy.org/outgoing/new_GPI.txt');
                        $req->header('Accept' => 'text/html');
                        $req->content;
                        $res=$ua->request($req);
                        print "CONTENT OF THE FILE:"."\n";
#                       print $res->content;
                        if ($res->is_success) {
                              my $output3 = $res->content;
#                             $output3 =~ s/<.*?>//g;
                              print $output3;
# print the output to another file
                              open(FILE, ">result_GPI.txt");
                              print FILE $output3;
                        }
                        else {
                              print "Error3: " . $res->status_line .
"\n";
                              }
                        }

                  }
            else {
                  print "Error2";
            }
      }

      else {
            print "Error1";
      }
```

```
### Script 2
### This program has been written for selecting experimentally
# verified entries from the file with entries in Swiss-Prot format

# MAIN

# Modules used:
use SWISS::Entry;
use strict;
use warnings;

print "\n";
# Command line usage
print " Usage : perl entry_file [output_file]\n";

my $entry_file;
my $output_file;

my $metazoa=0;
my $protozoa=0;
my $ fungi=0;
my $plants=0;
my $alt_splis=0;
my $proven=0;
my $non_proven=0;
my $taxon;

my @metazoa=();
my @protozoa=();
my @fungi=();
my @plants=();
my @alt_splis=();

# manage input/output file
$entry_file=shift @ARGV or die "you must provide input file!";
print"\n";
# if output-file name is not provided, default name will be used
if (scalar @ARGV>0) {
      $output_file=shift @ARGV;
}
else  {
      $output_file="CC_output.txt"
}

my $count=0;
my $positive;

# record separator \n is replaced by //
$/ = "\n//\n";

open (FILIN, "$entry_file") or die "cannot open file $!\n";
open (FILOUT, ">$output_file") or die "cannot open file $!\n";

# for each Swiss-Prot entry in the input file
while (<FILIN>) {

# read an entry and store it in the $entry variable
      my $entry = SWISS::Entry->fromText($_);

# store the AC number in the entry_AC variable
      my $entry_AC = $entry->AC;
```

```
# store the sequence in the entry_SQ variable
      my $entry_SQ = $entry->SQ;
my $conclusion;

# to select just metazoan entries
      if ($entry->OCs->size) {
            if ($entry->OCs->head eq "Eukaryota") {
                   $taxon=${$entry->OCs->list}[1];
                  if ($taxon eq "Metazoa") {
                        $metazoa++;
                        push(@metazoa, $entry_AC);
                  }
                  if ($taxon ne "Metazoa") {next;}
            }

      }

# Find out if GPI-anchor is a key-word
      foreach my $kw ($entry->KWs->elements) {
            if ($kw->text eq 'GPI-anchor') {
                  $conclusion=1;
#                 print $kw->text."\n";
            }
# select alternative splicing entries, which may need manual
verification
            if($kw->text eq 'Alternative splicing') {
                  print $kw->text."\n";
                  $alt_splis++;
                  push(@alt_splis, $entry_AC);
            }
      }

if ($conclusion==0) {
      next;
}

# Print all CCs
my @CC = $entry->CCs->elements;

      foreach my $CC (@CC) {
            print "TOPIC ".$CC->{topic}."\t";
         if ($CC->{topic} eq 'SUBCELLULAR LOCATION') {
            if($CC->{comment}=~/GPI-anchor\s+(\((?:By
similarity|Potential|Probable)\))/){
                  print "-----------------------------------------
      "."\n";
                  $non_proven++."\n";
                  print "NON"."\t".$non_proven."\t".$CC-
>{comment}."\n".$taxon."\t";
                  print $entry_AC."\t"." NON-EXPERIMENTAL!!!"."\n";
            }
            else {
                  print "-----------------------------------------
"."\n";
                  $proven ++;
                  print $proven."\t".$CC->{comment}."\n".$taxon."\t";
                  print $entry_AC."\t"." Verified
experimentally!"."\n";
            }
         }
      }

}
print "Metazoa"."\t".$metazoa.”\n”;
for (my $i=0; $i<scalar @alt_splis; $i++){
      print $alt_splis[$i]."\n";
}



# This script was written to call DGPI web-site
# and parse the output.
use strict;
# Modules used
use SWISS::Entry;
use SWISS::FTs;
use LWP::UserAgent;
# record separator \n is replaced by //
local $/="//\n";
my $count=0;
my $count_posit=0;
my $count_negat=0;
my $CGI;
my $ua;
my $line;

while(<>) {
$CGI = 'http://129.194.185.165/DGPI/DGPI';

# call UserAgent
$ua = new LWP::UserAgent;

      my $entry = SWISS::Entry->fromText($_);
      my $entry_AC = $entry->AC;
      my $entry_SQ = $entry->SQ;
      my $entry_ID = $entry->ID;
#     print $entry_AC;
#     print $entry_SQ."\n";
      my $fasta_SQ = ">".$entry_AC."\n".$entry_SQ;
# send new request
      my $req = new HTTP::Request POST => $CGI;
      $req->content_type('application/x-www-form-urlencoded');
      $req->content("protein=$entry_SQ");
      my $res = $ua->request($req);
      if ($res->is_success) {
            my $output = $res->content;
            $output =~ s/<.*?>//g;
            print "##################################"."\n";
            print $fasta_SQ."\n";

#                 print $output;
#           print "##################################"."\n";

#split output into lines
            foreach $line (split /\n/, $output) {

                  if($line=~/This protein is GPI-anchored/) {
                        print "AC  ". $entry_AC."        ". "ID   ".
                        $entry_ID."\n";
#                       print $entry_AC."        ".$entry_ID."\n";
                        $count_posit++;
                        $count++;
                        print "FOUND!"."\n";
                  }
                  elsif ($line=~/This protein is not GPI-anchored/) {
                        print "AC  ".$entry_AC."\n";
                        $count_negat++;
                        $count++;
                        print "NO!"."\n";
                  }

            }


      } else {
            print "Error";
      }
}
print "\n";
print "There were ".$count." entries. Out of them:"."\n";
print "POSITIVE : ".$count_posit."\n";
print "NEGATIVE : ".$count_negat."\n";.
```

7.2 Additional tables and images

Table A1. False-positive resuls comparison between 3 programs. Blank field means correct answer. Only entries with comment “Secreted protein” are shown.

AC	DGPI	GPI SOM
P15086	YES	YES
P82592	YES	
Q9Y0X9	YES	YES
P82290	YES	YES
Q8T9R8	YES	YES
P37889	YES	YES
Q90WJ8	YES	
O02380	YES	
P33575	YES	
Q9PTT2	YES	
P22922	YES	YES
Q00193	YES	YES
Q8SPQ0	YES	
Q965E2	YES	
P33577		YES
P81264		YES
Q9N2D3		YES
Q08380		YES
Q8IOL5		YES
Total Fp	14	12

Table A2. False negative results comparison for the 3 programs on verified positive set.

	DGPI	BIGPI	GPI-SOM
O08523		NO	
O08524		NO	
O08602			NO
O08603			NO
O08604			NO
O43895		NO	
O46647	NO	NO	NO
O75015	NO	NO	NO
O75326		NO	NO
O75443	NO	NO	
O88310	NO	NO	
O95980	NO		
P05186	NO		
P06858	NO	NO	NO
P07140			NO
P08571	NO		
P08582			NO
P09487			NO
P10675		NO	
P11151	NO	NO	NO
P11152	NO	NO	NO
P11153	NO	NO	NO
P11602	NO	NO	NO
P13265	NO	NO	NO
P14384		NO	
P22063		NO	
P25291		NO	NO
P28685		NO	NO
P35459		NO	
P38566	NO	NO	
P48733	NO		
P48794	NO	NO	NO
P49060	NO	NO	NO

P49291		**NO**	
P49923	**NO**	**NO**	**NO**
P51780	**NO**		
P52797		**NO**	
P52961			**NO**
P54097		**NO**	
P55031	**NO**	**NO**	**NO**
P56161	**NO**	**NO**	**NO**
P58019		**NO**	
P58020	**NO**	**NO**	**NO**
P83408	**NO**	**NO**	**NO**
P83456	**NO**	**NO**	**NO**
P91885		**NO**	
P91887		**NO**	
P97527		**NO**	
P97528	**NO**	**NO**	
Q06000	**NO**	**NO**	**NO**
Q12860		**NO**	
Q14210		**NO**	
Q28106		**NO**	
Q28680		**NO**	
Q28896		**NO**	
Q29524	**NO**	**NO**	**NO**
Q5R5C1		**NO**	
Q60935	**NO**	**NO**	**NO**
Q62803	**NO**	**NO**	**NO**
Q62997	**NO**	**NO**	
Q64389		**NO**	
Q80ZA0	**NO**	**NO**	
Q8JG54			**NO**
Q8N6Q3			**NO**
Q8R4F1		**NO**	
Q8R4G0	**NO**		
Q8WWA0	**NO**	**NO**	
Q93070		**NO**	**NO**
Q95323		**NO**	
Q95333		**NO**	

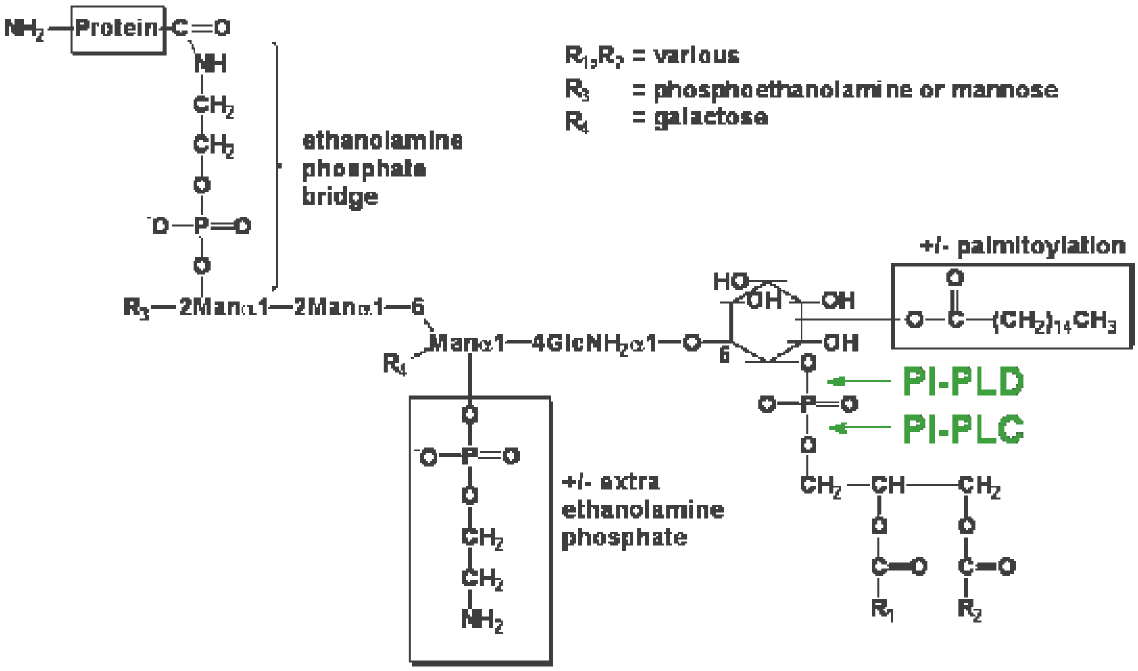
Q96PL2		NO	
Q98892	NO		
Q9BZM4		NO	
Q9BZM6		NO	
Q9D733		NO	
Q9DA79		NO	
Q9JI58			NO
Q9JJT2		NO	
Q9NHB0	NO	NO	NO
Q9QUG3	NO	NO	
Q9QUR8		NO	
Q9R0S3	NO		
Q9VN14		NO	
Q9YH85	NO		NO
Q9Z0J1	NO		
TOTAL:	39	65	36

Table A3. Program performance for cleavage-site prediction (43 metazoan sequences)

ENTRIES	Proved	best site	BIG PI	best site	DGPI	best site	GPI-SOM
	Site	BIG PI	prediction	DGPI	prediction	SOM	Prediction
P21589	549		549		549		549
P21588	551		551		551		551
P22748	284		284		284		281
P08174	353		353		353	350	NO GPI
P04058	564		564		563		564
P07692	567		567		566		567
P15328	234		234		234		230
P14207	230		230		230		230
P31997	320		326		325		319
P09326	220		220		206		218
P18181	217		217		217		217
P10252	217		217	217	NO GPI		217
P04216	130		130		NO SITE		139
O62643	130		130		NO SITE		139
P01831	131		131		131		140
P01830	130		130		130		139
P16444	385		385		385		384
P31358	36		36		36		36
Q09553	205		205		205		206
P05187	506		NO GPI		506		506
Q60506	231		231		231		231
Q60468	231		231		231		231
P04273	231		231		231		231
Q95333	649	656	NO GPI		658		649
P13987	102		102		102		121
O62680	98		98		103		98

If a program predicted ω-site, it is indicated, and highlighted in dark-grey if it is the same as experimentally verified ω-site. In case a program didn’t predict the fact of GPI-anchoring, it is indicated “NO GPI”. In some cases programs predict cleavage site differently from a proved one (it is not highlighted in this case) or are not able to predict it at all (as indicated “NO SITE”). Accsession numbers of 7 entries, which were not in BigPI-learning set, are highlighted.

GPI-anchor detailed chemical structure.



Chemical structure of a GPI anchor typically found attached to proteins at the external surface of mammalian plasma membranes. The glycan core, Man(α1→2)-Man(α1→6)-Man(α1→4)-GlcN, is glycosidically linked to the 6-hydroxyl group of phosphatidylinositol. Possible modifications can occur at R_1, R_2, R_3 and R_4, as well as palmitoylation of the inositol ring. The sites of anchor cleavage by the phospholipases PI-PLC and PI-PLD are indicated.

(Taken from http://www.uoguelph.ca/~fsharom/research/gpi.html)