Article No. jmbi.1999.3069 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 292, 741–758

JMB

Prediction of Potential GPI-modification Sites in Proprotein Sequences

Birgit Eisenhaber[1,2,3]*, Peer Bork[1,2] and Frank Eisenhaber[1,2,3]

[1]*European Molecular Biology Laboratory, Meyerhofstrasse1 Postfach 10.2209 D-69012, Heidelberg, Federal Republic of Germany*

[2]*Max-Delbrück-Centrum für Molekulare Medizin Robert-Rössle-Straße 10 D-13122, Berlin-Buch, Federal Republic of Germany*

[3]*Research Institute of Molecular Pathology, Dr. Bohr-Gasse 7 A-1030, Vienna, Republic of Austria*

*Corresponding author

Glycosylphosphatidylinositol (GPI) lipid anchoring is a common posttranslational modification known mainly from extracellular eukaryotic proteins. Attachment of the GPI moiety to the carboxyl terminus (ω-site) of the polypeptide follows after proteolytic cleavage of a C-terminal propeptide. For the first time, a new prediction technique locating potential GPI-modification sites in precursor sequences has been applied for large-scale protein sequence database searches. The composite prediction function (with separate parametrisation for metazoan and protozoan proteins) consists of terms evaluating both amino acid type preferences at sequence positions near a supposed ω-site as well as the concordance with general physical properties encoded in multi-residue correlation within the motif sequence. The latter terms are especially successful in rejecting non-appropriate sequences from consideration. The algorithm has been validated with a self-consistency and two jack-knife tests for the learning set of fully annotated sequences from the SWISS-PROT database as well as with a newly created database "big-Π" (more than 300 GPI-motif mutations extracted from original literature sources). The accuracy of predicting the effect of mutations in the GPI sequence motif was above 83%. Lists of potential precursor proteins which are non-annotated in SWISS-PROT and SPTrEMBL are presented on the WWW-page http://www.embl-heidelberg.de/beisenha/gpi/gpi_prediction.html The algorithm has been implemented in the prototype software "big-Π predictor" which may find application as a genome annotation and target selection tool.

Keywords: GPI lipid anchor attachment; posttranslational modification; GPI modification prediction; amino acid index; genome annotation

Introduction

Molecular biology and taxonomy of GPI-anchored proteins

Posttranslational modification with a glycosylphosphatidylinositol (GPI) lipid anchor is an important mechanism for tethering proteins of eukaryotic organisms (Ferguson & Williams, 1988) and their viruses (Zhou *et al.*, 1997) to cellular membranes. Most examples described in the literature (Gerber *et al.*, 1992) and in sequence databases (Bairoch & Apweiler, 1999) are of metazoan or parasitic protozoan origin but few other proteins from plants (Vai *et al.*, 1993; Morita *et al.*, 1996; Takos *et al.*, 1997) or fungi (Vai *et al.*, 1993; Guadiz *et al.*, 1998) have been reported. The GPI-modification pathway appears common also to some non-eukaryotic organisms such as archaeobacteria (Kobayashi *et al.*, 1997) and it is not finally excluded for some lines of eubacteria (Brennan & Nikaido, 1995; Ilangumaran *et al.*, 1995).

Typically, the posttranslational processing for GPI anchoring includes two steps (Gerber *et al.*, 1992). First, the preproproteins are targeted to the endoplasmic reticulum (ER) from the cytoplasm after their ribosomal synthesis *via* the signal peptide pathway, although alternative translocation mechanisms appear also possible (Howell *et al.*, 1994). Second, attachment of the GPI moiety to the

Abbreviations used: GPI, glycosylphosphatidylinositol; ER, endoplasmic reticulum; PSIC, position-specific independent counts; CM, comments for mutations; LAMS, lipoarabinomannans.

E-mail address of the corresponding author: b_eisen@nt.imp.univie.at

Please correspond to Dr Eisenhaber at the Research Insitute of Molecular Pathology, Austria.

0022-2836/99/380741-18 $30.00/0

carboxyl terminus (ω-site) of the polypeptide occurs by a transamidation reaction within the lumen of the endoplasmatic reticulum following proteolytic cleavage of a C-terminal propeptide from the proprotein (Udenfriend & Kodukula, 1995a,b). The entry to the GPI-modification reaction is directed by a C-terminal sequence signal (Moran *et al.*, 1991; Coyne *et al.*, 1993; Udenfriend & Kodukula, 1995a,b; Bucht & Hjalmarsson, 1996; Furukawa *et al.*, 1997; Yan *et al.*, 1998; Eisenhaber *et al.*, 1998). Subsequently, the mature proteins are described to be translocated with secretory vesicles and to be immobilised on the extracellular side of the plasma membrane (Nosjean *et al.*, 1997). It is not clear whether some types of GPI-anchored proteins may stay inside the vesicular system of the endoplasmic reticulum/Golgi during their whole life-time to execute their cellular function in this compartment.

Difficulties in experimental verifications of GPI lipid anchoring

Knowledge of a protein's GPI modification is very valuable, since it defines the subcellular localisation and limits the range of possible cellular functions. The GPI modification has also great medical importance (Ilangumaran & Robinson, 1996; Nosjean *et al.*, 1997). At the same time, the number of experimentally verified GPI-modified proteins is increasing more slowly than the total eukaryotic sequence data by several orders of magnitude (Eisenhaber *et al.*, 1998; Bairoch & Apweiler, 1999). Although the fraction of GPI-anchored proteins encoded in the genomes is still unknown, the trend appears to be the result of the discrepancy between the dramatic technological improvements for DNA and protein sequencing and the experimental difficulties in verifying GPI-posttranslational modifications. The latter task requires the demonstration of an existing GPI anchor for the given protein (P1), as well as the specification of the amino acid residue carrying the GPI-moiety (exact ω-site) (P2). Thus, there is little hope that the reports of experimentally verified ω-sites will multiply in a near future, and a sequence-based prediction algorithm as presented here would be the method of choice for the selection of targets for further experimental studies.

The first experimental problem (P1) is usually solved to a certain degree of confidence with solubilisation tests involving phospholipase (type C or D) cleavage of the GPI anchor. It should be noted that the sensitivity of this test depends on many factors including the anchor microheterogeneity (Taguchi *et al.*, 1994, 1999), the acylation state of the anchor which may change during the protein's lifetime (Chen *et al.*, 1998), the eukaryotic cell line studied, and the bacterial source of the phospholipase C (Low *et al.*, 1988). To add another level of complication, the cellular determination for GPI anchoring is not just an all-or-nothing decision, but may affect only a fraction of the population of protein molecules of a given sort due to competition with other independent pathways such as secretion with or without cleavage of the C-terminal propeptide (Wang *et al.*, 1997).

The determination of the exact ω-site (P2) is a much more laborious, non-standard experimental effort involving diverse techniques adapted to the specificity of the protein studied. The most direct, unambiguous approach includes proteinase digestion of the protein into smaller peptides under appropriate conditions, the separation of the GPI-labelled peptide, and the physico-chemical characterisation of this peptide and the GPI-modified amino acid residue in it with radioactive labelling, chemical peptide sequencing, composition analysis, NMR, mass spectrometry, and the like (Killeen *et al.*, 1988; Clayton & Mowatt, 1989; Misumi *et al.*, 1990; Stahl *et al.*, 1990; Moran *et al.*, 1991; Nuoffer *et al.*, 1991; Sugita *et al.*, 1993). Sometimes, experimental reports emphasise the existence of minor, alternate ω-sites in addition to the preferred GPI lipid anchor location which further complicate the site determination (Yan & Ratnam, 1995; Bucht & Hjalmarsson, 1996).

Solution: GPI-modification motif recognition from proprotein sequences

To conclude, tool development for the prediction of potential GPI-modification sites in proprotein sequences is not just a logical academic consequence of scientific developments, but represents an urgent practical need. It is required for the functional annotation of genomes (Bork *et al.*, 1998; Eisenhaber & Bork, 1998b) and for the selection of specific extracellular proteins among the wealth of sequence data as pharmaceutically important targets or for biotechnological applications.

Here, we present a knowledge-based algorithm evaluating the degree of presence of the C-terminal sequence signal in a query proprotein sequence based on sequence properties extracted from a learning set. If the proprotein sequence is a potential candidate for GPI modification, the algorithm determines also the best possible ω-sites. With our new technique, large-scale database searches for potentially GPI-anchored protein are feasible for the first time.

The paper is organised as follows: in Theory, the motivation for a new prediction function, its general structure, and the *P*-value-based classification of prediction results are described. All detail which should suffice to program the method is described in Methodological Details. Results contains two types of data. First, the method is validated with (1) a self-consistency test, (2) a jack-knife test, and (3) with predictions for about 300 mutations and natural sequence polymorphisms collected from the original literature (the most serious test). The second group of results includes predictions for non-annotated proteins. Finally, we discuss the possibilities for applications and improvements of the new method. Complete lists of learning set sequences, of prediction results, of numerical values for prediction function parameters as well as the big-Π mutation database are available on the WWW-page http://www.embl-heidelberg.de/~beisenha/gpi/gpi_prediction.html associated with this work.

Theory: Outline of the Prediction Function

The GPI-modification sequence motif

The nature of the GPI-modification signal carried by the C-terminal sequence segment has been investigated in detail by site-directed mutations in several exemplary model proteins (see, for example, Moran *et al.*, 1991; Coyne *et al.*, 1993; Bucht & Hjalmarsson, 1996; Furukawa *et al.*, 1997; Yan *et al.*, 1998). A meta-analysis of this data as well as a study of proprotein sequences in protein sequence databases (Eisenhaber *et al.*, 1998) revealed the following four sequence signal elements: (1) an unstructured linker region of about 11 residues ($\omega - 11 \ldots \omega - 1$); (2) a region of small residues ($\omega - 1 \ldots \omega + 2$), including the ω-site for propeptide cleavage and GPI-attachment; (3) a spacer region ($\omega + 3 \ldots \omega + 9$) of moderately polar residues; and (4) a hydrophobic tail beginning with $\omega + 9$ or $\omega + 10$ up to the C-terminal end.

Each of the sequence signal elements appears to represent a necessary requirement for recognition as a database study confirmed (Eisenhaber *et al.*, 1998) although, in a minority of publications (mostly before 1990; for a detailed analysis see Discussion), a contradictory opinion is reported (Waneck *et al.*, 1988; Orchansky *et al.*, 1988; Kurosaki & Ravetch, 1989; Santillán *et al.*, 1992; Engle *et al.*, 1995).

Why do traditional sequence analysis methods fail?

Simple profile searches (Bork & Gibson, 1996), for example with WiseTools (Birney *et al.*, 1996), based on alignments of sequence segments containing regions around the ω-site are not successful in selecting potential proproteins from the SWISS-PROT database. More than two-thirds of the first 100 hits in our tests are clear false positive predictions. This is the result of a reasonable total score from other sequence regions compensating, for example, for the absence of a suitable propeptide cleavage site. Experimental data on some single-residue mutations changing the efficiency of GPI anchoring by several orders of magnitude (Caras & Weddel, 1989; Moran *et al.*, 1991; Coyne *et al.*, 1993; Bucht & Hjalmarsson, 1996; Furukawa *et al.*, 1997; Yan *et al.*, 1998) also indicate that simple sum scores from profile-sequence alignments alone are not a good prediction tool, since they are not very sensitive to single-residue substitutions. Additionally, our database study (Eisenhaber *et al.*, 1998) has shown that the GPI-modification sequence signal is not well characterised by amino acid type preferences (except for a few positions close to the ω-site).

The composite prediction function

The GPI modification sequence motif may be better 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, 1989; Moran *et al.*, 1991; Udenfriend & Kodukula, 1995a,b; Furukawa *et al.*, 1997; Eisenhaber *et al.*, 1998).

Therefore, we have formulated a more sophisticated score function S which consists of two parts:

$$S = S_{\text{profile}} + S_{\text{ppt}} \qquad (1)$$

A profile-dependent section S_{profile} evaluates the concordance with the weak amino acid type preferences in the learning set at single alignment positions. We applied a recently validated, powerful 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). This method can extract more information from alignments containing subsets of similar sequences than traditional algorithms.

Another class of terms composing the score S_{ppt} (physical property term) describes the conservation of physical properties in the GPI-modification signal arising from the interaction of few or many sequence positions: (1) side-chain volume limitations and mutual volume compensation effects for residues $\omega - 1 \ldots \omega + 2$ expected to be located within the catalytic cleft of the putative GPI modification transamidase; (2) backbone flexibility requirements within the segment $\omega - 1 \ldots \omega + 2$; (3) propeptide length ranges (from $\omega + 1$ to the C end); (4) spacer region ($\omega + 3 \ldots \omega + 8$) hydrophilicity and sequence volume per residue; (5) hydrophobicity limits averaged over the C-terminal hydrophobic region and conditions for even distribution of hydrophobic residues; (6) the presence of aliphatic hydrophobic residues (LVI-contents in the tail) and the absence of long stretches of residues with a flexible backbone (GS-content in a window) in the C-terminal hydrophobic tail.

There are taxon-specific differences in the GPI-modification motif (Moran & Caras, 1994). Given the limitations of the learning set, we have derived function parametrisations for Metazoa and Protozoa. The procedure for compiling the learning set is not trivial and it is described in detail in Methodological Details.

The probability of false positive predictions

For a given query sequence, first any (C-terminal) residue is assumed to be a potential ω-site and the associated score S (equation (1)) is computed. The best site (with the highest score) among all sequence positions is selected as the prediction. The total score S is translated into the probability of a false-positive GPI-site prediction (Figure 1) with the help of an extreme value distribution (Altschul *et al.*, 1994). The probabilistic interpretations of scores has the advantage that prediction thresholds can be reasonably introduced and scores with differently parametrised functions may be compared.

For the purpose of qualitative comparison of prediction results, we label predictions with ratings A, B, C, and D corresponding to thresholds of P-values below 0.0025, 0.0050, 0.0075, and 0.010, respectively. Generally, ω-sites predicted with labels A-D can be considered as quite reliable, since the error of false positive predictions is <1 %, and the corresponding scores are always clearly positive (Table 1); i.e. sequences predicted as GPI-modified should have some compliance with the profile and no serious motif defect that might have resulted in a large negative S_{ppt}.

The label S(pecial) is assigned to predictions with a P-value above 0.010 but below 0.0175. All sequences with a P-value above the S-threshold ($P > 0.0175$) are not predicted as potential GPI-anchored proteins (label N). Additionally, all sequences having (1) a negative total score S and a profile-independent score S_{ppt} below -2ρ or (2) the S-label and a profile-independent score S_{ppt} below -3ρ are also excluded as possible GPI-modification candidates (label I). The value ρ has been set *ad hoc* to 4, since this is about the absolute value of the profile penalty in S_{profile} for an amino acid having a rare occurrence at a given position (Sunyaev *et al.*, 1999). With this condition, we acknowledge that non-compliance with general physical requirements to the sequence is more important than a reasonable total S score, since the latter may

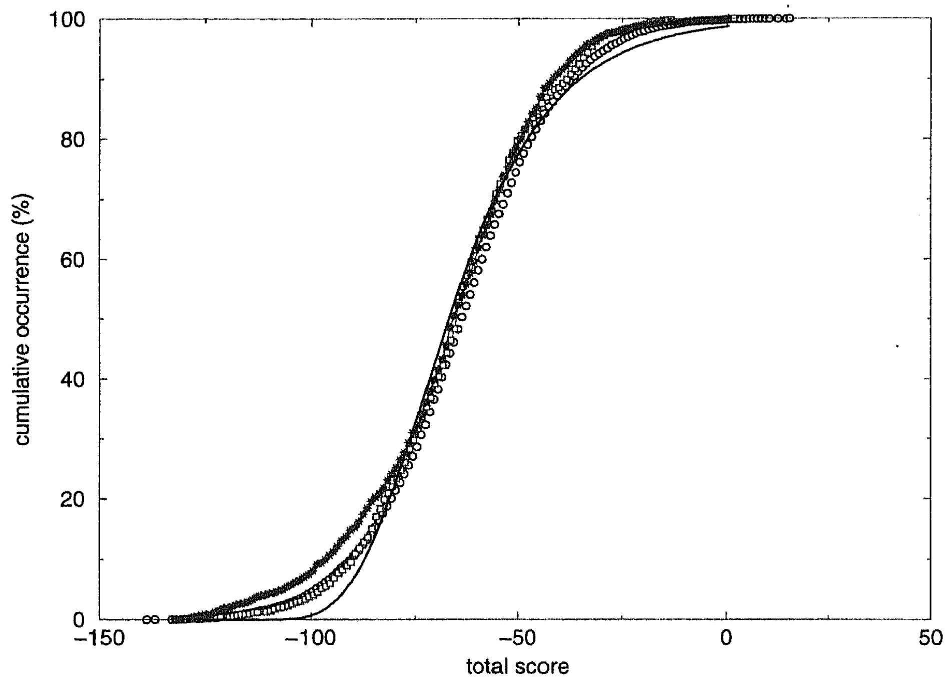

Figure 1. The extreme-value distribution of scores calculated from non-GPI-modified proteins. Any type of sequence motif may occur incidentally in non-related proteins; thus, the statistical significance of a match between query sequence and motif needs to be evaluated. We assume that the values of each term of the score function are normally distributed for large sequence arrays; therefore, the total score also follows a normal distribution. The prediction algorithm selects the site corresponding to the maximal score. Just as the sum of many independent random variables results naturally in a normal distribution, the maximum of many independent random variables yields an extreme value distribution. This Figure shows the distribution functions of best scores for metazoan sequences with a length ⩾55 residues. Black circles correspond to the set of all sequences without the keyword GPI-anchor in the rel. 37 of SWISS-PROT (23989 sequences) (curve $P_{\text{observed}}(S \leqslant T)$, the observed frequency of scores S below a threshold T). The red squares represent the set of sequences having the pattern "CYTOPLASM" in a comment line (2106 sequences). Green stars illustrate the distribution function for a set of nuclear proteins (with the keyword "NUCLEAR PROTEIN", 3191 sequences). All three curves are very close to each other and to the theoretical distribution function of maximal scores S below a threshold T:

$$P(S \leqslant T) = \exp\{-\exp[-\lambda(T - u)]\}$$

The correlation coefficient between $\ln[-\ln(P_{\text{observed}}(S \leqslant T))]$ and $-\lambda(T - u)$ is 0.95 and the regression is validated by the t-tests for the regression coefficients and Fisher's test for comparison with the function average for all significances 0.001-0.05. In the diagram, we show all distribution functions multiplied with 100%. It should be noted that the extreme value distribution is below the experimental curves in the case of very high scores. Therefore, the theoretical significances do even overestimate the probabilities of random motif occurrence.

Table 1. Self-consistency test of the learning set

	Metazoa		Protozoa	
	Score threshold	Number	Score threshold	Number
Total number of sequences		120		38
Predicted as A ($P < 0.0025$)	$\geqslant 28.15$	18	$\geqslant 29.69$	4
Predicted as B ($P < 0.0050$)	$\geqslant 16.41$	14	$\geqslant 16.21$	24
Predicted as C ($P < 0.0075$)	$\geqslant 9.54$	42	$\geqslant 8.31$	3
Predicted as D ($P < 0.0100$)	$\geqslant 4.66$	22	$\geqslant 2.70$	3
Predicted as S ($P < 0.0175$)	$\geqslant -4.86$	12	$\geqslant -8.24$	0
Not predicted		12		4
Total prediction (A-D) (%)		96 out of 120 (80.0)		34 out of 38 (89.5)
Total prediction (A-S) (%)		108 out of 120 (90.0)		34 out of 38 (89.5)

The prediction function parameters have been calculated using the whole learning set. With this function, predictions were made for all members of the learning set. The score thresholds listed are determined from the theoretical extreme value distribution fitted to the empirical distribution functions for non-related metazoan and protozoan sequences (see Figure 1) given the predefined significance thresholds.

be achieved due to a good S_{profile} (a "sufficient" condition) at not so important sequence positions. The major purpose of S_{ppt} consists of the efficient exclusion of unsuitable sequences from consideration (a "necessary" condition).

The remaining S-labelled predictions are in a twilight zone (total score is near zero, Table 1) and need experimental verification. As a rule, the sequences also do not have dramatic problems with the physical description of the motif. Their low total score may just be the result of an incomplete profile due to the limited variety of sequences in the learning set but also a false positive prediction.

Results

Prediction function validation: self-consistency test for the learning set

The results of predicting the learning set with the prediction function extracted from the same data are described in Table 1 and in detail on our WWW page. The prediction of a potential GPI modification with less than 1% probability of a false positive decision could be made for 80% of metazoan and 89.5% of protozoan entries. Including the S-labelled predictions, the rate increases to 90% for Metazoa.

In total 16 proteins (four protozoan, 12 metazoan) are predicted as being non-compatible with the GPI-modification motif. The sequence of PAG1_TRYBB (Q01889) misses a suitable ω-site. Three other protozoan entries as well as ten metazoan entries lack an appropriately constructed hydrophobic tail. The metazoan protein 5NTD_DISOM (P29240) has an extraordinarily hydrophobic spacer. THY1_CHICK (Q07212) misses an acceptable ω-site. In both latter cases, the sequences are unusually dissimilar to the C-terminal region of family members 5NTD_HUMAN (P21589) (Misumi *et al.*, 1990), 5NTD_BOVIN (Q05927) (Suzuki *et al.*, 1993), 5NTD_RAT (P21588) (Ogata *et al.*, 1990) and THY1_RAT (P01830) (Tse *et al.*, 1985) with partly or completely experimentally verified ω-site. It should be noted that the annotations in all 16 entries are not based on experimental data but, in fact, represent "informed guesses" (Nielsen *et al.*, 1999). Hence, our prediction result may be interpreted as source for doubts with respect to these annotations or indicate a possible sequencing error.

For proteins predicted as GPI modified, the coincidence of predicted and annotated ω-sites was tested. Only in seven cases (one protozoan and six metazoan entries), the predicted site deviates from the annotated one; thus, the rate of site prediction is 97.1% for Protozoa (out of 34) and 94.4% for Metazoa (out of 108). It should be noted that, for all 20 fully experimentally verified GPI-anchored proteins (see Methodological Details), the algorithm predicted both the fact of GPI modification and the site correctly.

For two proteins, experimental evidence for minor, alternate ω-sites is available. We predict position 235 (label D) as secondary site for FOL1_HUMAN (P15328, major site is predicted at 234, label C) and position 565 (label D) as secondary site for ACES_TORCA (P04058, major site is predicted at 564, label C) in full agreement with the literature sources (Yan & Ratnam, 1995; Bucht & Hjalmarsson, 1996).

Prediction function validation: jack-knife test for the learning set

The jack-knife test (leaving the predicted sequence out of the learning procedure) is a more serious test for the power of our prediction function (Table 2). In our first jack-knife procedure (test 1 in Table 2), the prediction function was completely re-computed after leaving out a single entry in the learning set. Among protozoan proteins, 30 out of 38 (79.0% with labels A-D) remain predicted as candidates for GPI anchoring. The same is true for 101 out of 120 metazoan entries (69.2% with labels A-D, 84.2% with labels A-S). It should be emphasised that all negative predictions which were

Table 2. Jack-knife test of the learning set

	Metazoa		Protozoa	
	Test 1	Test 2	Test 1	Test 2
Total number of sequences	120	55	38	19
Predicted as A ($P < 0.0025$)	17	2	3	1
Predicted as B ($P < 0.0050$)	10	7	19	12
Predicted as C ($P < 0.0075$)	31	19	5	3
Predicted as D ($P < 0.0100$)	25	14	3	1
Predicted as S ($P < 0.0175$)	18	6	0	0
Not predicted	19	7	8	2
Total prediction (A-D) (%)	83 out of 120 (69.2)	42 out of 55 (76.4)	30 out of 38 (79.0)	17 out of 19 (89.5)
Total prediction (A-S) (%)	101 out of 120 (84.6)	48 out of 55 (87.3)	30 out of 38 (79.0)	17 out of 19 (89.5)

The jack-knife test 1 was performed over the whole learning set and the prediction function was re-computed each time after leaving a single protein out. In the case of jack-knife test 2, we cycled over all sequences in the largest subset of non-homologous sequences only. The profile terms (in S_{profile}) were computed from the whole learning set but the parametrisation of the physical terms (in S_{ppt}) was redetermined after leaving one protein out of the original largest subset of non-related proteins.

above the thresholds in the self-consistency test (four protozoan and seven metazoan entries) suffered from the profile parametrisation but not from the slight changes of parameters computed for the physical property terms from the reduced learning set. It is clear that an exhaustive representation of the profile components is only possible with a dramatically enlarged learning set.

Despite the fact that the learning set contains subsets of similar sequences, the PSIC profile method applied (see Methodological Details) is able to compute a profile matrix with both sequence and position-specific weightings; thus, even small sequence deviations serve as additional sources of information and, consequently, the profile matrices differ for each jack-knife-tested sequence. At the same time, this jack-knife test cannot be considered as stringent as in the case of learning sets composed only of sequentially non-related sequences, since the S_{ppt} terms were not really subjected to jack-knifing.

To verify that the parameters of the physical property terms are not an overfitting of the small datasets, we performed a second jack-knife test (test 2 in Table 2) just over the largest subset of non-homologous sequences. The parameters for S_{ppt} were re-calculated with a leaving-one-out algorithm, but the profile remained unchanged as calculated from the whole learning set. The prediction accuracy is very near to (for Metazoa: 87.4% compared with 90%) and identical with (for Protozoa, 89.5% in both cases) the value from the self-consistency test; thus, even our small learning set is sufficient to determine the parameters of S_{ppt} reliably.

Prediction function validation: prediction of natural polymorphisms

Isoforms of precursor proteins are a natural source of sequential polymorphism. Ratnam and collegues (Shen *et al.*, 1995; Wang *et al.*, 1997) have studied the GPI-modification efficiency for the α, β, and γ-isoforms of the human folate receptor in great detail. The first isoform is obligatorily GPI-anchored, the third one is constitutively secreted whereas the β-isoform is distributed between a soluble, extracellular and a GPI-anchored states in roughly the same proportion. Ratnam and collegues (Shen *et al.*, 1995; Wang *et al.*, 1997) suppose that the GPI-modification pathway and another, yet to be identified intracellular pathway for C-terminal proprotein processing and secretion, compete for the β-isoform of the folate receptor precursor. Our prediction results are in satisfactory agreement with the experimental data (Table 3): The scores and *P*-values follow the experimentally determined GPI-modification efficiencies. The α and β-isoforms are predicted as candidates for GPI anchoring (with a worse score and *P*-value for the β-isoform). The computed ω-sites coincide with the annotated ones. The γ-isoform is not predicted as complying with the GPI motif consensus. It is noteworthy that this protein carries the keyword "GPI-ANCHOR" in SWISS-PROT but is without ω-site annotation in the feature table.

Table 3. Natural polymorphism: isoforms of the human folate receptor

Isoform			Prediction			
SWISS-PROT entry						
ID	Accession	Type	Score	*P*-value	Site label	Experiment
FOL1_HUMAN	P15328	α	15.68	5.22×10^{-3}	C	100% GPI
FOL2_HUMAN	P14207	β	11.82	6.56×10^{-3}	C	51% GPI
FOL3_HUMAN	P41439	γ	−45.58	1.78×10^{-1}	N	Secreted

The experimental data on the efficiency of GPI-modification of the human folate receptor has been taken from Wang *et al.* (1997) for the α and β-isoforms, and from Shen *et al.* (1995) for the γ-isoform.

Prediction function validation: prediction of mutation data

The big-Π mutation database

We scanned the scientific literature for reports with quantitative data on mutations in the GPI motif region in exemplary proteins and constructed a database named big-Π (big-PI, Birgit's GPI motif mutation database) in SWISS-PROT format with our own system of IDs, accession numbers, and experimental data lines associated with a new "CM" (comments for mutations) token. The database is accessible over the WWW. Our effort received unexpected support from Bucht *et al.* (1999) who supplied their new mutation data for the human acetylcholine esterase precursor prior to publication. At its current status, the GPI motif mutation library contains 293 mutations (Berger *et al.*, 1988; Caras *et al.*, 1989; Su & Bothwell, 1989; Micanovic *et al.*, 1990; Moran *et al.*, 1991; Lowe, 1992; Beghdadi-Rais *et al.*, 1993; Coyne *et al.*, 1993; Kodukula *et al.*, 1993; Furukawa *et al.*, 1994, 1997; Moran & Caras, 1994; Okuyama *et al.*, 1995; Yan & Ratnam, 1995; Bucht & Hjalmarsson, 1996; Wang *et al.*, 1997; Yan *et al.*, 1998; Wilbourn *et al.*, 1998; Aceto *et al.*, 1999; Bucht *et al.*, 1999; Tomassetti *et al.*, 1999) tested in metazoan, and 65 mutations (Nuoffer *et al.*, 1993) tested in fungal cell systems. The precursor proteins the mutation of which are described in our database belong to 13 different families of evolutionarily related proteins. In the evaluation of the effect of mutations on the efficiency of GPI modification, we rely on the qualitative assessments, the authors by since it is difficult for us to evaluate the experimental error range. For two publications (Kodukula *et al.*, 1993; Bucht *et al.*, 1999), we considered only mutations resulting in more than 10% of wild-type activity as GPI-anchored, since: (1) some mutations are more efficient than the wild-type by a factor of 2-4; (2) this is about the range of the experimental accuracy; and (3) we determined the values from a graphical representation in the case of one publication (Bucht *et al.*, 1999).

Prediction of the effect of GPI motif mutations: statistics

The set of metazoan mutation data is a serious test for our prediction function since, in many cases, just a single-residue mutation reverses the GPI-anchoring efficiency from 100% to zero. The prediction results are described in Table 4 and on the WWW with respect to the mutation series and to GPI-modification pathway permissive and breaking mutations. We considered all predictions with labels A-D as efficient precursors for GPI anchoring. In the case of the known twilight zone (label S), we rejected predictions with negative scores as potentially GPI-modified proteins.

To our surprise, the algorithm predicted the impact of mutations on the GPI-anchoring ability of the proteins considered correctly in 244 out of 293 cases (83.3%). The rate of correct prediction does not depend on whether pathway-supporting (84.5%) or pathway-breaking (81.1%) mutations are studied.

As a tendency, the prediction rate is associated with the series of mutations (literature source) and the type of precursor protein being mutated. There are two outliers with 0% (series hGH_VSG; Moran & Caras, 1994) and 61.9% (series Ly6A; Su & Bothwell, 1989); the rest are between 70% and 100%. We observe a tendency that predictions appear better for mutation sets from more recent papers and from groups having a longer history of experimental work in the GPI anchor field compared with those having only a single publication and, maybe, a small set of experimentally studied mutations; thus, singular and systematic experimental inaccuracies may also introduce noise. The experimental techniques and standards appear much improved during the last decade and more recent data seems more reliable than that from before 1990.

Prediction of the effect of GPI motif mutations: analysis of false predictions

It should be emphasised that species or cell line-specific requirements to proprotein sequences which are not incorporated in our general metazoan-specific function cannot be excluded and will certainly play an important role. For example, contradictory mutation data have been published concerning the occurrence of valine, glutamate, threonine, and proline residues in the region $\omega-1\ldots\omega+2$ and especially at the ω-site (mutation series hGH_DAF29, PLAP, CD46, FA10_PLAP, ACES_HUMAN). Our prediction function agrees with the tenor of the experimental papers which tends to disfavour their occurrence; thus, singular exceptions are not predicted. A special problem is a cysteine residue at the cleavage site. From the viewpoint of residue volume and backbone flexibility, this residue should perfectly suit the requirements. Moran *et al.* (1991) also express their surprise that this mutation is not permissive in their experiment, whereas cysteine residues seem to be efficient ω-sites in the series PLAP (Berger *et al.*, 1988) and in THY1 proteins (Tse *et al.*, 1985). In rare cases (mutation LAA in $\omega\ldots\omega+2$ in series PLAP and mutations CD, CL, FA, GW, YT at $\omega+1$ and $\omega+2$ in series ACES_HUMAN), the profile score S_{profile} in the prediction function S causes a prediction in contrast to the experimental criterion; thus, the limited learning set of sequences plays a role.

Further, the prediction function is adjusted to a learning set of naturally occurring sequences known to be GPI anchored. Therefore, fusion proteins with C-terminal hydrophobic tails from parasitic protozoa (Moran & Caras, 1994) are rejected as possible candidates due to the low profile score in the tail region (series hGH_VSG). Also, artificial

Table 4. Prediction results for mutated precursors of GPI-modified proteins

Mutation				Prediction results			
Series	Site	Tail	Rest	With GPI-anchor	No GPI-anchor	Accuracy (%)	References
DAF	x	x	x	4/5	1/1	83.3	Caras *et al.* (1989)
hGH-DAFx	x		x	5/5	1/1	92.9	Moran *et al.* (1991)
hGH-DAF28x	x	x			3/3		Wilbourn *et al.* (1998)
hGH-xDAFx	x		x	0/1	4/4		Caras *et al.* (1989)
hGH-DAF29x group 1	x			5/5		84.2	Moran *et al.* (1991)
hGH-DAF29x group 2	x			0/2	1/2		Moran *et al.* (1991)
hGH-DAF29x group 3	x				10/10		Moran *et al.* (1991)
hGH_VSG	x			0/4		0.0	Moran & Caras (1994)
PLAP I	x			14/16	14/17	84.8	Moran & Caras (1994) Kodukula *et al.* (1993) Micanovic *et al.* (1990)
PLAP II		x		2/3	4/5	75.0	Berger *et al.* (1988)
Folate receptor I	x	x	x	3/3	2/3	91.2	Wang *et al.* (1997)
Folate receptor II		x		8/8			Yan *et al.* (1998)
Folate receptor III		x	x	4/4			Yan *et al.* (1998)
Folate receptor IV		x		5/6			Yan *et al.* (1998)
Folate receptor V	x	x	x	6/7			Yan & Ratnam (1995)
Folate receptor VI	x			2/2	1/1		Tomasetti *et al.* (1999)
ACES_TORCA	x	x	x	12/12	4/4	100.0	Bucht & Hjalmarsson (1996)
ACES_HUMAN I			x	6/6	4/4	100.0	Bucht *et al.* (1999)
ACES_HUMAN II	x			16/19	14/16	85.7	Bucht *et al.* (1999)
CD46 (fusion)	x	x	x	26/35	2/5	70.0	Coyne *et al.* (1993)
THY1			x	4/4	2/3	85.7	Behdadi-Rais *et al.* (1993)
CAH4	x			2/2	2/2	100.0	Okuyama *et al.* (1995)
5′-NTD		x	x	10/10	5/7	88.2	Furutawa *et al.* (1994, 1997)
FA10_PLAP	x	x	x	6/8	3/4	75.0	Lowe (1992)
miniPLAP	x			3/3		100.0	Aceto *et al.* (1999)
MP/uPAR	x			11/12	5/6	88.9	Aceto *et al.* (1999)
LY6A		x	x	4/5	4/8	61.6	Su & Bothwell (1989)
				158/187	86/106		
				84.5	81.1		
Total				244/293		83.3	

The names of the mutation series are identical with those in the WWW-page associated with this work. It is marked with an x if the series contains mutations in the ω-site region, in the hydrophobic tail or in some other part of the GPI motif (Rest). The prediction results are summarised separately for GPI-modification permissive and GPI-modification inhibitive mutations (as *a*/*b*; i.e. *a* correct predictions from all *b* examples of the given series). The accuracy is the fraction of correct predictions among all expressed as a percentage.

proproteins with extreme propeptide lengths and ω-sites shifted deeply to the N terminus such as in mutation database entries PLAP_35 (Berger *et al.*, 1988) and hGH_DDAF17 (Caras *et al.*, 1989) are not predicted due to the functional form of propeptide length term T_5 (see Methodological Details). It is not clear whether naturally existing precursors with long propeptides do exist (Caras, 1991; Wang *et al.*, 1999).

The penalties for non-permissive spacer sequences ($\omega + 3 \cdots \omega + 8$) appear to be too small (wrong prediction for some spacer mutations in series CD46, FA10_PLAP, and THY1), but the experimental data are not sufficiently conclusive to make a final suggestion. For example, lower thresholds for spacer hydrophobicity are not well supported by the learning data. Up to three charged residues fit nicely into the spacer of some examples in the learning set. The dipeptidase MDP1_HUMAN (P16444) has a fully verified experimental ω-site but three residues (HRH) at positions $\omega + 6 \ldots \omega + 8$ in the spacer. Similarly, upper thresholds for spacer hydrophobicity cannot easily be derived.

Singular mutations in the N-terminal linker region, partly far away from the supposed cleavage site, in the series LY6A (Su & Bothwell, 1989) remain completely unnoticed by our prediction function. Likewise, a few polar or charged residues in the C-terminal hydrophobic region are normally tolerated in the available learning set and, therefore, also by our prediction function. But surprisingly, they stop GPI modification in some examples (series folate receptor I and FA10_PLAP).

In the majority of examples, the predicted ω-site of the mutated protein coincides with the annotated one in the wild-type sequence, although we observe several cases of small shifts. Unfortunately, none of the mutation series is associated with a direct structural study for locating the ω-site. Our meta-analysis of experimental mutation studies indicates that regular direct determinations of propeptide cleavage and GPI anchor attachment sites would greatly increase the value of the data, reduce possibly biased interpretations of experimental output, and ease the correct design of new mutation experiments.

Prediction function application: searches for potential precursors of GPI-anchored proteins in SWISS-PROT, SWISS-NEW, and SPTrEMBL

We applied the computer software developed for the search of candidate precursors in general protein databases. We analysed (1) SWISS-PROT (rel. 37) complemented, on our WWW-page, with SWISS-NEW from 12th of April, 1999, and (2) SPTrEMBL (rel. 9). The prediction function with metazoan and with protozoan parametrisation was applied separately. As the taxonomic classifier "PROTOZOA" disappeared since SWISS-PROT release 37, we searched for entries outside the groups of Metazoa, fungi, viruses, archaeobacteria, bacteria, and Planta. The remaining selected entries were checked manually.

Prediction statistics in SWISS-PROT

Within SWISS-PROT, we considered separately: (1) entries carrying the keyword "GPI-ANCHOR"; (2) entries having the keyword but no ω-site annotation in the feature table; and (3) entries without the keyword. The numbers of entries describing predicted potential precursors of GPI-modified proteins are listed in Table 5 and on the WWW-page. Surprisingly, our prediction function did not detect many of the sequence entries labelled with the keyword GPI-ANCHOR (Table 5 and Figure 2) as potential precursors (only 72.4 % for Metazoa and 65.4 % for Protozoa), especially if the site has not been annotated with "FT LIPID" (60.0 % for Metazoa and 45.9 % for Protozoa). This is in sharp contrast to the results for the learning set and the mutation data (clearly above 80 %). Possibly, our prediction function S (equation (1)) is not sufficiently general due to the limited learning set. But several other reasons may also contribute to this result. First of all, the sequence part of the entry may contain not the precursor sequence but that of the mature protein, of another splicing version or sequence isoform. Sequencing or annotation errors may also play a role (e.g. the γ-isoform of the human folate receptor being annotated as GPI-anchored in SWISS-PROT in contradiction to literature reports; Shen *et al.*, 1995). Both issues are difficult to check in an automated manner. But we should also admit that most rejected examples failed due to a very low S_{ppt} score; thus, there is little compliance with the property pattern from the learning set. A considerable fraction of the GPI-anchor annotations appears not reliable.

Prediction of non-annotated protozoan proteins in SWISS-PROT and SPTrEMBL

Non-annotated precursor proteins are probably of great biological interest (see the WWW-page for lists). Although the protozoan learning set consists almost exclusively of sequences from *Trypanosoma*, we found potential precursor proteins for GPI

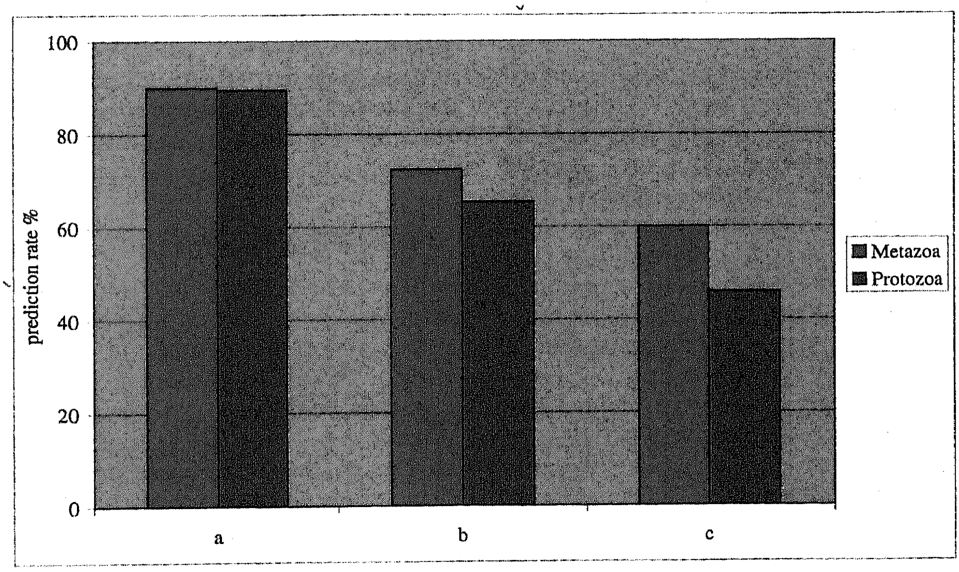

Figure 2. Prediction rates of proteins annotated as GPI-anchored in SWISS-PROT. The prediction rates of proteins described in database entries with different annotation status are visualised: (a) with the keyword GPI-ANCHOR and a feature table note for the ω-site (in fact the learning set); (b) with the keyword; and (c) with the keyword but without a respective feature table note. With decreasing quality (detail) of the annotation status, the prediction rate of the proteins as being GPI lipid-anchored goes down dramatically. Probably, a considerable number of entries carries non-reliable annotations with respect to the GPI modification.

Table 5. Prediction of potential GPI-modification sites

	Metazoa (%)	Protozoa (%)
SWISS-PROT (rel. 37) with "KW GPI-ANCHOR"		
A-D	140/243 (57.6)	51/78 (65.4)
A-S	176/243 (72.4)	51/78 (65.4)
SWISS-PROT (rel. 37) with "KW GPI-ANCHOR" but without "FT LIPID"		
A-D	46/115 (40.0)	17/37 (45.9)
A-S	69/115 (60.0 %)	17/37 (45.9 %)
SWISS-PROT (rel. 37) no "KW GPI-ANCHOR"		
A-D	15	0
A-S	70	7
SPTrEMBL (rel.)		
A-D	112	56
A-S	259	88

Numbers of database entries with sequences predicted as having the GPI-modification motif are listed. If possible, this number is compared with all proteins having the same annotation status.

modification of many other Protozoa including representatives of *Entamoeba, Toxoplasma, Plasmodium, Eimeria,* and *Leishmania.* Very often, both the S_{profile} and S_{ppt} scores have small absolute values, i.e. the physical property motif is conserved but the amino acid type occurring in the sequence differ from expectations calculated in the small learning set. The circumsporozoite surface proteins of *Plasmodium* are suspected to be GPI-anchored (Moran & Caras, 1991), our prediction for six entries is in agreement here. To mention just for curiosity, we found the entry Q27673 of an ectometalloproteinase (*Leishmania amazonensis*) with an annotated C-terminal signal leader peptide beginning with position 574! We predict position 573 as potential ω-site; thus, the 'signal peptide" appears in fact a propeptide cleaved by a GPI modification transamidase.

Prediction of non-annotated metazoan proteins in SWISS-PROT and SPTrEMBL

The metazoan prediction function seems much more noisy than the protozoan one. We found a large number of hits but only a small portion with labels A-D. Some observations deserve special attention. Dimethylaniline monoxygenases (e.g. P97872, P17636, P49326), heme oxygenases (e.g. P14901, P06762), and an aminopeptidase N (O46156) are located in the endoplasmic reticulum. The hyaluronidase LUCA2 functions in lysosomes. The sequences of these isoforms appear permissive for GPI anchoring. These may be examples of GPI-modified proteins without translocation to the extracellular side of the plasmalemma. Further, we predict a number of cancer (Q13421, O46156) and cell differentiation (prostate stem cell antigen O43653, megacaryocyte potentiating factor Q14859) markers, a serotonin receptor (Q29005), plant nodulins (e.g. P25226), a *Caenorhabditis elegans* carboxypeptidase (P52716), etc.

Among the many S-labelled predictions, we consider a large number as false positives, since the biological context of the protein function appears little compatible with GPI anchoring. As in the strange example of an extracellular cytochrome (Hettmann *et al.*, 1998), the generally accepted function need not always fit easily with the observed cellular localisation. For example, mitochondrial cellular localisation was found with automatic annotation analysis (Eisenhaber & Bork, 1998a; Eisenhaber & Bork, 1999) for 11 entries with similar sequences describing lipid-binding subunits of ATPases and 13 homologous examples of subunits of NADH-ubiquinone oxidoreductases (one of those is the only observed false prediction with label D). It should be noted that our prediction function tests only the existence of a possible GPI-modification motif in the C-terminal sequence. Maybe, some of those proteins would really undergo GPI modification if their targeting signal would be changed to a signal leader peptide forcing translocation to the endoplasmic reticulum but, just due to the cellular context, the motif has never been checked by appropriate enzyme systems. As another possibility, the prediction function is confused as a result of a C-terminal hydrophobic region with a preceding loop matching the physical requirements of a propeptide cleavage and attachment site.

Application of the prediction function to non-Zoa

The application of the prediction functions developed for metazoan sequences to those from other, non-metazoan taxonomic subdivisions cannot be really considered a prediction (more a creative suggestion), since the parametrisation is hardly adequate (the same is true for the application of the protozoan prediction function in a non-protozoan taxonomic lineage). The results presented on our WWW-page may be examined by experimentalists searching for GPI-modification pathways in organisms currently considered to be lacking it. Several viral, eubacterial, archaeobacterial, and fungal proteins are promising candidates. For example, the halobacterial surface glycoprotein CSG_HALHA (P08198, label D) appears a very good hit.

Discussion

History of attempts for GPI-modification prediction

We found in a search for appropriate literature references that the prediction of potential GPI modifications as possible posttranslational modifications relying only on sequence data has received little attention as a specific bioinformatics research task.

Antony & Miller (1994) proposed a formalism to locate the ω-site if both the sequence of the proprotein (after signal sequence cleavage) and the amino acid sequence composition of the mature protein are known. Thus, the costly procedure of experimental identification of GPI-modified proteins is still necessary. Udenfriend & Kodukula (1995a,b) developed an algorithm using only amino acid type preferences in the $\omega \ldots \omega + 2$ region. Of course, this signal alone is not sufficient for large-scale database searches. But as one element, a similar term is contained also in our prediction function. The authors of PSORT2 (Nakai & Kanehisa, 1997; Nakai & Horton, 1999) predict potential precursors as having just a transmembrane region with very short tail (type 1a membrane protein); thus, other requirements such as ω-site restrictions are not considered. The idea of penalising the absence of C-terminal hydrophobic tails has found a continuation in our work.

Alternatively, Chou & Elrod (1999) propose to use the amino acid composition of a protein for predicting its membrane association including possible GPI anchoring and report 66-81 % accuracy for some test sets. The practical application of their algorithm appears impossible. The amino acid composition of a protein changes at different levels of protein processing and maturation. Additionally, single mutations may target the protein to another location (there are many examples of re-directions between a GPI-anchored and a secreted states in our mutation database) but a single mutation changes the amino acid composition very little.

Scope of possible applications of our prediction technique

Thus, our prediction technique can well be considered as the first method of integrating all sequence requirements known today for the GPI-modification motif. We included terms evaluating amino acid type preferences at given motif positions but also terms judging the conservation of physical properties in the query sequence which represent correlation between few or many motif positions. The score S_{ppt} proved especially helpful in removing many false positives that could not be distinguished as such in our attempts with standard sequence analysis methods. Our predictions labelled with A-D have a considerable reliability. The accuracy of prediction of the mutation data appears a good estimate of the sensitivity of our method (>83 %). Compared with the size of the databases, the number of proteins in the twilight zone (with label S) is very small and appears suitable for detailed consideration (high selectivity). Thus, our prediction algorithm is an appropriate tool for target selection and for genome annotation purposes (Bork *et al.*, 1998; Eisenhaber & Bork, 1998b). A detailed summary of GPI-anchored proteins in available complete genomes and chromosomes is in preparation.

It should be noted that even general requirements of the GPI-modification motif are questioned in a few but not very recent reports: (1) the wild-type mouse protein Qa-2 was reported as being GPI anchored (Waneck *et al.*, 1988) although a polar spacer region (except for one residue D295) is completely absent. Additionally, our metazoan profile proved to be not compatible with the C terminus of Qa-2. In accordance with our model, the substitution D295V inhibits GPI-modification.

(2) Three proteins, two artificial constructs with an identical C terminus (Orchansky *et al.*, 1988; Santillán *et al.*, 1992) and the intestinal alkaline phosphatase isoform IAP-2 (Engle *et al.*, 1995), are reported to be GPI-modified although the suspected precursor sequence completely lacks a hydrophobic C-terminal tail. At least in one case (Engle *et al.*, 1995), the experimental data appear doubtful, since the standard deviation is almost 50 % of the mean (see their Table 4).

(3) In the wild-type FcγRIII-1 isoform of the IgG Fc receptor, the sequence FSPP is supposed to be the $\omega - 1 \ldots \omega + 2$ region (Kurosaki & Ravetch, 1989) whereas the remaining literature (see the review by Eisenhaber *et al.*, 1998) discusses the possibility of a single proline residue surrounded by tiny residues near the ω-site. Our predictor selects S201 (putative ω-site region SSFS) as best although the sequence is discarded as untypical for GPI-modification due to low hydrophobicity and the incompatibility with the profile of the hydrophobic tail.

Unfortunately, these reports have not been confirmed by later publications relying on more sophisticated techniques. Just the phospholipase C test has been used in the original studies, a direct ω-site determination as described in Introduction has not been carried out. If these results are not artefacts, we suppose the existence of an alternative GPI-modification pathway not requiring the classical motif. It might also be that the proteins are anchored with non-GPI phospholipid anchors which are also sensitive to phospholipase C.

Further improvement of GPI-modification prediction appears hampered mainly by the current status of experimental data. A substantial fraction of corresponding annotations in protein databases seems doubtful. In its current formulation, the metazoan-specific function looks more noisy than the protozoan-specific one, since the fraction of predictions in the twilight zone is higher, i.e. we expect relatively more falsely annotated data in the corresponding learning set. It would be necessary to accumulate more verified data on ω-sites (in direct structural experiments) and associated efficiencies of GPI anchor attachment (possibly, together with fractions of the same protein which moved along alternative pathways). Also, structural data with respect to the GPI-modification transamidase complex would be helpful in recognising directions for function improvement.

Given the status of the learning data, we believe that continued "optimisation" of prediction function parameters, for example, with methods of machine learning, is not appropriate. The appearance of possibly higher prediction rates as a result of smoother fitting to the doubtful part of the data is not accompanied with a physical interpretation, does not improve the understanding of the sequence motif studied and does not help in the design of new mutation experiments.

The big-Π predictor software for prediction of GPI modification sites in precursor sequences will be made publicly available as a WWW server during 1999.

The biological role of GPI-anchoring

The biological role of GPI lipid anchors appears not fully uncovered yet, and is probably not restricted to mechanical tethering of eukaryotic proteins at the plasmalemma only. Especially if additional transmembrane segments do exist, our predictions suggest that some proteins which remain inside the ER/golgi/lysosomes for their whole life-cycle may also be GPI anchored. It can be imagined that the fixation of the C terminus may function as protection against peptidases in a hostile environment.

Our twilight zone predictions for some eubacterial proteins indicate that lineages of Eubacteria having a pathway similar to the GPI modification for proteins might also exist. Lipoarabinomannans (LAMs) on the plasma membrane of mycobacteria (such as *Mycobacteria tuberculosum* and *Mycobacteria leprae*) resemble GPI-anchored proteins, since they represent complex, multiply branched carbohydrate polymers of mannose and arabinose terminated by a phosphatidylinositol lipid anchor (Brennan & Nikaido, 1995; Ilangumaran *et al.*, 1995). But the similarity is not complete, since the typical glycosyl-*N*-α-1-inositol linkage in GPI moieties is not present in LAMS. Generally, phosphatidylinositol lipid anchors are fairly common in Eubacteria. Thus, it cannot be excluded that some bacterial strains have experimented with GPI anchors during evolution.

Methodological Details

Creation of the learning set

Statistics of GPI-anchor annotation in SWISS-PROT

We searched for entries of annotated precursors of proteins with GPI-modification (with keyword GPI-ANCHOR and feature table entry FT LIPID) in SWISS-PROT (rel.36) and in SWISS-NEW (as of 27th of January, 1999). A total of 188 entries were detected. After exclusion of duplications, cases with incomplete or erroneous annotation, etc., 176 examples remained. The non-annotated carbonic anhydrase IV precursor entry (CAH4_HUMAN, P22748) was also included, since we found the full experimental verification of its GPI-modification site incidentally in a reference (Okuyama *et al.*, 1995). The annotated ω-sites for the acetylcholine esterases from *Torpedo californica* (ACES_TORCA, P04058) and *Torpedo marmorata* (ACES_TORMA, P07692) have been corrected in accordance with recent literature data (Bucht & Hjalmarsson, 1996). Thus, the primary learning set consists of 177 sequence entries. The list of all entries with detailed commentaries to the selection procedure and annotation corrections is available *via* the WWW.

Quality of GPI-anchor annotations

The quality of the learning data is limiting for the success of the prediction technique to be developed. Nielsen *et al.* (1999) have thoroughly analysed various types of errors that may result in wrong database annotations. In our case, the completeness of experimental checks for the ω-site annotation is the major point and remains to be improved. The current status of verification of the fact of GPI-modification and the exact determination of the ω-site has been discussed (as from August 1998) in detail in our previous publication (Eisenhaber *et al.*, 1998) where a list of 18 completely verified entries has been supplied. Additionally, we found carbonic anhydrase IV and the THY1 glycoprotein of rat (Tse *et al.*, 1985). The situation has not been significantly improved since then. This does not necessarily mean that the ω-site assignment in all other entries is not reliable, since the experimental results may be contained in papers not referred to in the annotation or may not be published at all; thus, the sources are extremely difficult for us to trace. For a large number of sequences, only indirect data such as mutation experiments indicating a possible ω-site are available. In many cases, strong sequence similarities to verified entries indicate that the annotated assignments are reliable.

Also, we found that all sequences in the learning set have an N-terminal signal leader except for five cases. Three of those (VSA8_TRYBB, P06017; VSE2_TRYBR, P26335; VSM0_TRYBB, P07209) are known fragments. The precursor sequences VSG7_TRYBR (P02898) and DAF_PONPY (P49457) are possibly also incomplete.

Therefore, we can expect that the consensus signal from all sequences of the learning set is reasonably characteristic for precursors of GPI-modified proteins in general; if not for all, then, at least, for a subset of them. At the same time, there might be a few entries in the learning set with wrongly annotated ω-sites or even sequences falsely labelled as GPI-anchored.

Taxonomic classification of the learning set

Another problem is the taxonomic classification within the 177 entries of the primary learning set. Few examples are available for viruses (one) and fungal proteins (ten). The remaining ones are from Metazoa (126) and Protozoa (40). Due to the limited dataset and the dependence of the GPI-modification signal on taxonomic branches (Eisenhaber *et al.*, 1998), we elaborated prediction functions only for the two largest taxonomic subdivisions.

Expert editing of the learning set: removal of crude annotation errors

We have compared the sequences in the learning set with the accumulated expert knowledge of the GPI sequence motif (Eisenhaber *et al.*, 1998). Two protozoan and six metazoan entries annotated as "POTENTIAL" precursors (the result of an "expert decision") had to be excluded from learning due to their extreme propeptide length (outside the range of 17-31 residues) and the absence of an alternative reasonable ω-site (see the associated WWW page); therefore, the final learning set consists of 38 protozoan and 120 metazoan entries. For another two protozoan and ten metazoan entries with extreme propeptide length, such an alternative site could be identified. In a further 11 cases of metazoan sequence entries, the ω-site had to be shifted due to reasons of sequence similarity with other entries or because the site was occupied by a large and/or positively charged amino acid type. It should be noted that all 23 entries of ω-site editing were annotated as potential or by similarity but not on the basis of an experimental ω-site determination.

The profile score term S_{profile}

General structure of the profile score term

The profile score S_{profile} is composed of subscores S_{region} for specific sequence regions ($S_{-11\ldots-2}$, $S_{-1\ldots+2}$, $S_{-+3\ldots+9}$, and $S_{+10\ldots\text{C-end}}$) and two penalties ($C_{-1\ldots2}$ and $C_{+10\ldots\text{C-end}}$):

$$S_{\text{profile}} = \begin{cases} \alpha_{\text{profile}}\alpha_{-11\ldots-2}S_{-11\ldots-2}+ \\ \alpha_{\text{profile}}\alpha_{-1\ldots+2}S_{-1\ldots+2}+ \\ \alpha_{\text{profile}}\alpha_{+3\ldots+9}S_{+3\ldots+9}+ \\ \alpha_{\text{profile}}\alpha_{+10\ldots\text{C-end}}S_{+10\ldots\text{C-end}}+ \\ C_{-1\ldots+2} + C_{+10\ldots\text{C-end}} \end{cases} \tag{2}$$

The factor α_{profile} is determined with a normalisation condition taking the different lengths of the sequence regions into account:

$$\alpha_{\text{profile}}\ \text{profile\_length} = \sum_{\text{region}} \alpha_{\text{region}}\ \text{region\_length} \tag{3}$$

The weightings α_{region} serve mainly for equilibrating the influence of the ω-site region and the C-terminal hydrophobic tail relative to each other due to their different sequence length (see below). The absence of either of the two signals is penalised with a large negative term:

$$C_{\text{region}} = \begin{cases} 0 & \text{if } S_{\text{region}} \geqslant \text{threshold} \\ A_{\text{region}} & \text{otherwise} \end{cases} \tag{4}$$

which effectively reduces the total score beyond positive prediction levels (A_{region} is a sequence region-specific constant parameter). The profile subscores:

$$S_{\text{region}} = \sum_{i\in\text{region}} S_i \tag{5}$$

are independently summed over all sequence positions i within one of the sequence regions $\omega-11\ldots\omega-2$, $\omega-1\ldots\omega+2$, $\omega+3\ldots\omega+9$, and $\omega+10\ldots$ end relative to a supposed ω-site in the query sequence. The profile values S_i are extracted from a gapless multiple alignment relative to the ω-site (from $\omega-15$ to the C-terminal end) of proprotein sequences described as precursors of GPI-anchored proteins in SWISS-PROT separately for each taxonomic subdivision (Eisenhaber *et al.*, 1998; Sunyaev *et al.*, 1999).

Sensitive profile extraction

From the alignments of the proprotein sequences, the relative occurrences $p(a,i)$ of amino acid types a at given motif positions i (Eisenhaber *et al.*, 1998) are determined. With these values, a profile matrix for the sequence segment $\omega-11\ldots\omega+25$ has been computed (see below). For positions more C-terminal than $\omega+25$, this procedure becomes difficult or impossible due to the lack of data, since most sequences have shorter propeptides.

It should be emphasised that these alignments contain many highly similar alignments with little sequence variations affecting often only a few positions. To extract a maximum of information, we used PSIC, a recently validated, powerful new profile extraction technique which assigns both sequence and alignment position-specific weights (Eisenhaber *et al.*, 1998; Sunyaev *et al.*, 1999). In brief, we compute an effective number $n(a,i)_{\text{eff}}$ of observations of amino acid type a at alignment position i and determine $p(a,i)$ as:

$$p(a,i) = \frac{n(a,i)}{\sum_b n(b,i)_{\text{eff}}} \tag{6}$$

The summation is carried out over all amino acid types b. The value $n(a,i)_{\text{eff}}$ is thought to depend on the overall similarity of sequences having the common amino acid type a in the alignment column considered. The frequency of alignment positions $f(a,i)$ in the subset of sequences having the same amino acid type a at alignment position i is used as similarity measure and is set equal to the probability of identical alignment positions for $n(a,j)_{\text{eff}}$ in random sequences. The solution of the equation:

$$f(a,i) = \sum_b q_b^{n(a,i)_{\text{eff}}} \tag{7}$$

for $n(a,j)_{\text{eff}}$ estimates the number of independent observations of amino acid a at position j in the alignment. The value q_b is the default frequency of amino acid type b in a sequence database. The final profile matrix S_i (a) which enters equation (5) is calculated as (Sunyaev *et al.*, 1999):

$$S_i(a) = \ln\frac{p(a,i)}{q_a} \tag{8}$$

The $S_{profile}$ parametrisation

The weightings in equations (2) and the constants in equation (4) have been determined on an *ad hoc* basis and after analysing the sequence region subscores S_{region} for all sequences in the learning dataset. We have set $\alpha_{-1\ldots2} = 3\alpha_{-10\ldots\text{C-end}} = 3$, since the ω-site region has on average only a third of the length of the hydrophobic tail. The other two weights $\alpha_{-11\ldots-1}$ and $\alpha_{+3\ldots+8}$ were set equal to 0.5, since these two sequence signals appear less prominent in their importance. As threshold in equation (4), the value 0.1 was accepted. The penalties are $A_{-1\ldots+2} = -\rho$ and $A_{+10\ldots\text{C-end}} = -\rho$. The constant ρ is the standard penalty scale and it is used also in many physical property terms. It is set equal to 4, since this is about the same value compared with the absolute sequence profile score for an unfavourably occupied position in S_{profile}.

The physical property score S_{ppt}

General structure of the term S_{ppt}

The functional form of multiple residue correlation terms with respect to physical properties composing S_{ppt} is selected in such a manner that clear deviations from value ranges in the learning set of proproteins are penalised. At the same time, compliance with the consensus signal extracted from the learning set results in a zero score (but not in positive scores). Herewith, we recognise that the form of physical terms in S_{ppt} reflects our rough understanding of requirements of the polypeptide binding site in the transamidase complex executing the GPI modification but a possible specific role of different amino acid types at certain sequence positions might be not well discerned or differ among species (Eisenhaber *et al.*, 1998).

In its current formulation, S_{ppt}:

$$S_{\text{ppt}} = \sum_{\substack{j=0;\\ j\neq 3,11}}^{14} \alpha_j T_j \tag{9}$$

includes the following terms (the numbering coincides with the prediction results available on WWW pages; $T_3 = C_{-1\ldots+2}$ and $T_{11} = C_{+10\ldots\text{C-end}}$) describing: (1) the side-chain volume limitations and mutual volume compensations for residues $\omega-1\ldots\omega+2$ within the catalytic cleft (terms T_0, T_1, and T_2); (2) backbone flexibility requirements within the segment $\omega-1\ldots\omega+2$ (term T_4); (3) propeptide length ranges (from $\omega+1$ to the C-end, term T_5); (4) spacer region $(\omega+3\ldots\omega+8)$ hydrophilicity (term T_6) and sequence volume per residue (term T_7); (5) hydrophobicity limits averaged over the C-terminal hydrophobic region (terms T_8 and T_9) and conditions for even distribution of hydrophobic residues (terms T_{10} and T_{12}); and (6) the presence of aliphatic hydrophobic residues (LVI-contents in the tail, term T_{13}) and the absence of long stretches of residues with a flexible backbone (GS-content in a window, term T_{14}) in the C-terminal hydrophobic tail.

Each of the 13 physical conditions enters the sum (equation (9)) in the form of the natural logarithm of a probability distribution function to be comparable with scores from profile computations. A Gauss or a Boltzmann-like distribution was assumed for values outside the allowed value ranges.

Weights α_j in equation (9) are allowed in the algorithm but are equal to unity except for the term T_6 (see below). All function parameters have been set on the basis of *ad hoc* physical considerations or as averages over the learning set. Any form of further parameter optimisation will reflect more the particularities of the available limited sequence dataset than, at the present time, an understanding of functioning of the transamidase complex which is still on a very approximate level. The purpose of introducing S_{ppt} consists of excluding sequences as unlikely candidates for GPI anchoring due to untypical integral sequence properties compared with the learning set. It should be noted that, in contrast to the profile score S_{profile}, which has 20 parameters per alignment position (from $\omega-11$ to $\omega+25$, a total of $20[(\omega+25)-(\omega-11)+1]=740$ profile matrix elements), all physical terms together have less than 30 parameters (except for sequence region parameters and term weightings). Therefore, they are expected to describe the GPI modification motif in a more general form which is not so dependent on the nature of the specific sequence examples in the learning set.

Parametrisation of physical property terms

Volume terms T_0, T_1, T_2, and T_7

The database analysis of precursors of GPI-modified proteins (Eisenhaber *et al.*, 1998) has revealed that the allowed side-chain volume within the ω-site region and, to a lesser extent, also in the spacer is limited. This result, which justifies physical terms T_0, T_1, T_2, and T_7, is even more convincing for the current, slightly larger learning set. High correlation with amino acid size characteristics (corr. coeff. > 0.70) has been observed at positions $\omega-1$, ω, $\omega+1$, $\omega+2$, $\omega+3$, $\omega+6$, and $\omega+7$ (more weakly on positions $\omega+4$ and $\omega+5$) for Metazoa and, except for $\omega+7$, also for Protozoa. Generally, we observe an avoidance of large aromatic residues in both sequence regions. This result justifies the terms T_0 and T_7 with the functional form:

$$T_j(V) = \begin{cases} 0 & \text{if} \quad V \leqslant V_j \\ -(V-V_j)^2/(2\sigma_j^2) & \text{if} \quad V > V_j \end{cases} \tag{10}$$

with $j = 0,7$. Obviously, $T_j(V)$ is the natural logarithm of a Gauss function for sufficiently large arguments. The volume V is calculated as sum of the residue volumes (in accordance with the scale by Harpaz *et al.* (1994)) at the selected sequence positions (here, the regions $\omega-1\ldots\omega+2$ and $\omega+3\ldots\omega+8$ correspondingly). V_j and σ_j are parameters calculated as averages over the largest subset of non-related proteins (see below); i.e. the sum volume of the residues in the learning set and its r.m.s.d., respectively. In the case of T_7, both V and the parameters are calculated as mean values per sequence position.

The volume compensation effect within the cleavage site is statistically significant in accordance with Fisher's F-criterion (see equation (3) by Eisenhaber *et al.* (1998) and the discussion therein) for the positions $\omega-1$, $\omega+1$, and $\omega+2$ (set 1, $F = 1.62$) as well as for $\omega-1$ and $\omega+2$ (set 2, $F = 1.86$) in the case of Metazoa ($F_{\text{critical}} = 1.60$ for 5 % significance). The results for Protozoa ($F_{\text{critical}} = 2.23$ for 5 % significance) are not significant due to the smaller number of examples but they are still remarkable: $F = 1.86$ for set 1 comprising $\omega-1$, ω, and $\omega+1$; $F = 2.01$ for set 2 comprising $\omega-1$ and $\omega+1$. Hence, we feel supported to introduce terms T_1 and T_2 (for sequence position sets 1 and 2, respectively) in the form of equation (10).

Backbone flexibility term T_4

Our sequence analysis revealed that none of the sequences in the learning set had more than one residue out of the set including proline, threonine or valine (PTV) within the sequence region $\omega-1\ldots\omega+2$. This fact may be explained by the necessity for some backbone flexibility to accommodate the chain in the catalytic cleft. Ring closures of the side-chain with the backbone or β-branched side-chains reduce this flexibility. The functional form of T_4 is:

$$T_4 = \begin{cases} 0 & \text{if } n^{(PTV)}_{-1\ldots2} \leqslant 1 \\ -\rho\,(n^{(PTV)}_{-1\ldots2} - 1) & \text{otherwise} \end{cases} \quad (11)$$

The value $n^{(PVT)}_{-1\ldots2}$ is the total number of residues with reduced backbone flexibility.

Propeptide length term T_5

All experimentally verified sites have propeptide lengths between 17 and 31 residues (Eisenhaber *et al.*, 1998). There are some experimental indications that GPI modification of internal residues resulting in overly long propeptides might be possible for artificial fusion proteins (Caras, 1991; Wang *et al.*, 1999) but, as Caras (1991) has pointed out, the close C-terminal localisation of the ω-site might represent an evolutionary adaptation rather than a strict functional constraint. Until further experimental clarification of the issue, we penalise extreme propeptide lengths with the term:

$$T_5 = -\rho[g(l_{\min} - l) + g(l - l_{\max})] \quad (12)$$

where the function g is defined as:

$$g(x) = \begin{cases} 0 & \text{if } z \leqslant 0 \\ x & \text{if } x > 0 \end{cases} \quad (13)$$

the value l is the propeptide length ($l_{\min} = 17$; $l_{\max} = 31$). The term T_5 as well as T_4 can be interpreted as natural logarithms of a Boltzmann distribution function.

Hydrophilicity term T_6 *for the spacer region*

We summed the hydrophobicity of the residues in the spacer region $\omega + 3 \ldots \omega + 8$ (in accordance with the hydrophobicity scale by Nakashima & Nakashima (1992)), computed the average hydrophobicity H per residue for the sequence segment, and repeated this procedure for all sequences in the learning set. The distribution of H is clearly shifted towards low hydrophobicity values. To penalise large deviations towards very hydrophobic spacers, we introduced term T_6 in a form similar to equation (10):

$$T_6(H) = \begin{cases} 0 & \text{if } H \leqslant H_6 \\ -(H - H_6)^2/(2\sigma_6^2) & \text{if } H > H_6 \end{cases} \quad (14)$$

Here, H is the average per residue of the spacer in the query sequence, and H_6 and σ_6 are the mean value and the r.m.s.d. of H over the largest subset of non-homologous proteins in the learning set (see below). The weight α_6 is equal to 3.

Terms T_8, T_9 *and* T_{13} *for average hydrophobicity of hydrophobic tail*

In principally the same way, we generate negative scores for overly hydrophilic C-terminal tails. One function ($j = 8$) controls the whole region (from $\omega + 10$ to the C terminus); another one ($j = 9$) is additionally designed for propeptides extending the limits of the profile (from $\omega + 26$ to the C terminus). The functional form is:

$$T_j(H) = \begin{cases} -(H - H_j)^2/(2\sigma_j^2) & \text{if } H < H_j \\ 0 & \text{if } H \geqslant H_j \end{cases} \quad (15)$$

Again, H is the average hydrophobicity per residue of the sequence region considered in the query sequence and H_j and σ_j are the mean values and the r.m.s.d. of H over the largest subset of non-homologous proteins in the learning set (see below). If a sequence region considered is smaller than eight residues, it is extended to this level on the N-terminal side, since eight residues is the smallest possible hydrophobic tail (Eisenhaber *et al.*, 1998).

The sequence examples in the learning set achieve a high hydrophobicity of the C-terminal tail due to a high content of aliphatic amino acids but not with aromatic ones. Therefore, we require a minimal content of leucine, valine, and isoleucine ($\geqslant 30\,\%$) in the sequence region ($\omega + 10 \ldots$ C terminus). If this condition is not fulfilled, a large penalty is assigned with $T_{13} = -3\rho$.

We feel that, as a tendency, the hydrophobic tails of many proteins in the learning set are too short (especially among Protozoa) to form a transmembrane (-helix (standard length of 21 amino acid residues). Therefore, the suggestion by Wang *et al.* (1999) that the tail does not interact with the lipid bilayer might be true. Further clarification of the molecular interactions will help to decide whether it is necessary to design terms that distinguish C-terminal hydrophobic tails from transmembrane regions.

Terms T_{10} *and* T_{12} *for even distribution of hydrophobicity along the hydrophobic tail*

A high average hydrophobicity over the C-terminal hydrophobic tail may also be achieved by a single cluster of extremely hydrophobic residues in addition to a long stretch of polar residues. Analysis of the learning set shows that this is not the case, polar or even charged residues are always surrounded by groups of non-polar residues. The even distribution can be quantified by sliding a small window over the sequence and calculating the average hydrophobicity per residue in the window. We found that almost all sequences have at least five consecutive windows of length 4 with an average hydrophobicity above 9.00 using the scale by Nakashima & Nishikawa (1992). The absence of this property is penalised in S_{ppt} with $T_{10} = -3\rho$.

Very polar windows are also extremely rare. The corresponding function T_{12} was adjusted to the learning set. We penalize windows of length 3 with an average hydrophobicity below 1.50 with $T_{12} = -\rho$ and extremely polar windows (average hydrophobicity below 0.50) with $T_{12} = -3\rho$.

Backbone flexibility term T_{14} *in the C-terminal tail*

In contrast to the ω-site region, the C-terminal tail of the learning set examples is characterised by the avoidance of residues with tiny side-chains such as glycine and serine, at least in the form of clusters, i.e. C-terminal segments with increased backbone flexibility may not favour the adjustment of the substrate polypeptide in the transamidase binding site. Based on this observation, we penalise the occurrence of sequence windows of length 4 with at least three tiny residues (glycine and/or serine residues) with $T_{14} = -3\rho$.

Largest subset of sequentially non-related proteins

Averages of physical properties and their r.m.s.d. have been calculated for the largest subset of sequences with maximal pairwise sequence identity below 30 % (only for the sequence segments $\omega - 30 \ldots \omega - 1$ and $\omega + 3 \ldots \omega + 8$, since the regions $\omega \ldots \omega + 2$ and $> \omega + 8$ are known to be compositionally biased). This is necessary to balance for redundancy due to subgroups of highly similar sequences. We followed published algorithms (Heringa *et al.*, 1992; Hobohm *et al.*, 1992). The resulting set consists of 55 sequences for Metazoa but only 18 proteins for Protozoa. Detailed data on parameters computed for all 13 physical property terms are listed on the WWW pages for this work.

Algorithmic details

In its current implementation of the algorithm in the form of computer software, the C-terminal 55 amino acid residues of a query sequence are taken and are analysed whether or not they present a syntactically correct protein sequence. Then, all sequence positions having a distance between 10 and 40 residues to the C-terminal end are analysed as potential ω-sites, the respective total score is calculated and the two best alternatives are determined. The scores are translated into probabilities of false motif detection with an extreme value distribution (Altschul *et al.*, 1994):

$$P(\text{score} \geqslant S) = 1 - \exp\{-\exp[-\lambda(S-u)\} \qquad (16)$$

The parameters λ and u have been determined using the best score of all SWISS-PROT entries (separately for Protozoa and Metazoa) without the annotation of a GPI modification (Figure 1). We assume that formula (16) can be applied although not all score components entering S can be considered as fully independent from each other. The quantitative assessment of the influence of their weak mutual correlation on the probability P is difficult from the theoretical point of view. Therefore, we acknowledge that the use of formula (16) is an approximation.

Acknowledgement

The authors are grateful to J.G. Reich for constant support during this work. We acknowledge especially the contribution of G. Bucht, P. Wikström, and K. Hjalmarsson who supplied us their quantitative data on 45 mutations in the GPI-modification signal sequence region of the human acetylcholine esterase proprotein prior to publication.

References

Aceto, J., Kieber-Emmons, T. & Cines, D. B. (1999). Carboxy-terminal processing of the urokinase receptor: implications for substrate recognition and glycosylphosphatidylinositol anchor addition. *Biochemistry*, **38**, 992-1001.

Altschul, S., Boguski, M., Gish, W. & Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119-129.

Antony, A. & Miller, M. E. (1994). Statistical prediction of the locus of endoproteolytic cleavage of the nascent polypeptide in glycosylphosphatidylinositol-anchored proteins. *Biochem. J.* **298**, 9-16.

Bairoch, A. & Apweiler, R. (1999). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucl. Acids Res.* **27**, 49-54.

Beghdadi-Rais, C., Schreyer, M., Rousseaux, M., Borel, P., Eisenberg, R. J., Cohen, G. H., Bron, C. & Fasel, N. (1993). Carboxyl terminus structural requirements for glycosyl-phosphatidylinositol anchor addition to cell surface proteins. *J. Cell Sci.* **105**, 831-840.

Berger, J., Howard, A. D., Brink, L., Gerber, L. D., Hauber, J., Cullen, B. R. & Udenfriend, S. (1988). COOH-terminal requirements for the correct processing of a phosphatidylinositol-glycan anchored membrane protein. *J. Biol. Chem.* **263**, 10016-10021.

Birney, E., Thompson, J. D. & Gibson, T. J. (1996). PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucl. Acid Res.* **24**, 2730-2739.

Bork, P. & Gibson, T. J. (1996). Applying motif and profile searches. *Methods Enzymol.* **266**, 162-184.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. & Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707-725.

Brennan, P. J. & Nikaido, H. (1995). The envelope of mycobacteria. *Annu. Rev. Biochem.* **64**, 29-63.

Bucht, G. & Hjalmarsson, K. (1996). Residues in *Torpedo californica* acetylcholinesterase necessary for processing to a glycosyl phophatidylinositol-anchored form. *Biochim. Biophys. Acta*, **1292**, 223-232.

Bucht, G., Wikström, P. & Hjalmarsson, K. (1999). Optimising the signal peptide for GPI modification of human acetylcholinesterase using mutational analysis and peptide-QSAR. *Biochim. Biophys. Acta*, **1431**, 471-482.

Caras, I. W. (1991). An internally positioned signal can direct attachment of a glycophospholipid membrane anchor. *J. Cell Biol.* **113**, 77-85.

Caras, I. W. & Weddel, G. N. (1989). Signal peptide for protein secretion directing glycophospholipid membrane anchor attachment. *Science*, **243**, 1196-1198.

Caras, I. W., Weddel, G. N. & Williams, S. R. (1989). Analysis of the signal for attachment of a glycophospholipid membrane anchor. *J. Cell Biol.* **108**, 1387-1396.

Chen, R., Walter, E. I., Parker, G., Lapurga, J. P., Millan, J. L., Ikehara, Y., Udenfriend, S. & Medof, M. E. (1998). Mammalian glycophosphatidylinositol anchor transfer to proteins and posttransfer deacylation. *Proc. Natl Acad. Sci. USA*, **95**, 9512-9517.

Chou, K.-C. & Elrod, D. W. (1999). Prediction of membrane protein types and subcellular locations. *Proteins: Struct. Funct. Genet.* **34**, 137-153.

Clayton, C. E. & Mowatt, M. R. (1989). The procyclic acididc repetitive proteins of *Trypanosoma brucei*. *J. Biol. Chem.* **264**, 15088-15093.

Coyne, K. E., Crisci, A. & Lublin, D. M. (1993). Construction of synthetic signals for glycosyl-phosphatidylinositol anchor attachment. *J. Biol. Chem.* **268**, 6689-6693.

Eisenhaber, F. & Bork, P. (1998a). Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol.* **8**, 169-170.

Eisenhaber, F. & Bork, P. (1998b). Sequence and structure of proteins. In *Biotechnology Recombinant Proteins, Monoclonal Antibodies and Therapeutic Genes* (Mountain, A., Ney, U. & Schomburg, D., eds), vol. 5a, pp. 47-86, VCH-Wiley, Weinheim.

Eisenhaber, F. & Bork, P. (1999). Computer-evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics*, **15**, in the press.

Eisenhaber, B., Bork, P. & Eisenhaber, F. (1998). Sequence properties of GPI-anchored proteins near the ω-site: constraints for the polypeptide binding site of the putative transamidase. *Protein Eng.* **11**, 1155-1161.

Engle, M. J., Mahmood, A. & Alpers, D. H. (1995). Two rat intestinal alkaline phosphatase isoforms with different carboxyl-terminal peptides are both membrane-bound by a glycan phosphatidylinositol linkage. *J. Biol. Chem.* **270**, 11935-11940.

Ferguson, M. A. & Williams, A. F. (1988). Cell-surface anchoring of proteins *via* glycosyl-phosphatidylinositol structures. *Annu. Rev. Biochem.* **57**, 285-320.

Furukawa, Y., Tamura, H. & Ikezawa, H. (1994). Mutational analysis of the COOH-terminal hydrophobic domain of bovine liver 5′-nucleotidase as a signal for glycosylphosphatidylinositol (GPI) anchor attachment. *Biochim. Biophys. Acta*, **1190**, 273-278.

Furukawa, Y., Tsukamoto, K. & Ikezawa, H. (1997). 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*, **1328**, 185-196.

Gerber, L. D., Kodukula, K. & Udenfriend, S. (1992). Phosphatidylinositol glycan (PI-G) anchored membrane proteins. *J. Biol. Chem.* **267**, 12168-12173.

Guadiz, G., Haidaris, C. G., Maine, G. N. & Simpson-Haidaris, P. J. (1998). The carboxyl terminus of *Pneumocystis carinii* glycoprotein A encodes a functional glycosylphosphatidylinositol signal sequence. *J. Biol. Chem.* **273**, 26202-26209.

Harpaz, Y., Gerstein, M. & Chothia, C. (1994). Volume changes on protein folding. *Structure*, **2**, 641-649.

Heringa, J., Sommerfeldt, H., Higgins, D. & Argos, P. (1992). OBSTRUCT: a program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. *Comput. Appl. Biosci.* **8**, 599-600.

Hettmann, T., Schmidt, C. L., Anemüller, S., Zähringer, U., Moll, H., Petersen, A. & Schäfer, G. (1998). Cytochrome b558/566 from the archaeon *Sulfolobus acidocaldarius*. *J. Biol. Chem.* **273**, 12032-12040.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409-417.

Howell, S., Lanctôt, C., Boileau, G. & Crine, P. (1994). A cleavable N-terminal signal peptide is not a prerequisite for the biosynthesis of glycosylphosphatidylinositol-anchored proteins. *J. Biol. Chem.* **269**, 16993-16996.

Ilangumaran, S. & Robinson, P. J. (1996). Transfer of exogeneous glycosylphosphatidylinositol (GPI)-linked molekules to plasma membranes. *Trends Cell Biol.* **6**, 163-167.

Ilangumaran, S., Arni, S., Poincelet, M., Theler, J.-M., Brennan, P. J., ud-Din, N. & Hoessli, D. C. (1995). Integration of mycobacterial lipoarabinomannans into glycosylphosphatidylinositol-rich domains of lymphomonocytic cell plasma membranes. *J. Immunol.* **155**, 1334-1342.

Killeen, N., Moessner, R., Arvieux, J., Willis, A. & Williams, A. F. (1988). The MRC OX-45 antigen of rat leukocytes and endothelium is in a subset of the immunoglobulin superfamily with CD2, LF-3 and carcinoembryonic antigens. *EMBO J.* **7**, 3087-3091.

Kobayashi, T., Nishizaki, R. & Ikezawa, H. (1997). The presence of GPI-linked protein(s) in an archaeobacterium, Sulfolobus acidocaldarius, closely related to eukaryotes. *Biochim. Biophys. Acta*, **1334**, 1-4.

Kodukula, K., Gerber, L. D., Amthauer, R., Brink, L. & Udenfriend, S. (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.

Kurosaki, T. & Ravetch, J. V. (1989). A single amino acid in the glycosyl phosphatidylinositol attachment domain determines the membrane topology of FcγRIII. *Nature*, **342**, 805-807.

Low, M. G., Stiernberg, J., Waneck, G. L., Flavell, R. A. & Kincade, P. W. (1988). Cell-specific heterogeneity in sensitivity of phosphatidylinositol-anchored membrane antigens to release by phospholipase C. *J. Immunol. Methods*, **113**, 101-111.

Lowe, M. E. (1992). Site-specific mutations in the COOH-terminus of placental alkaline phosphatase: a single amino acid change converts a glycosylphosphatidylinositol-glycan-anchored protein to a secreted protein. *J. Cell Biol.* **116**, 799-807.

Micanovic, R., Gerber, L. D., Berger, J., Kodukula, K. & Udenfriend, S. (1990). Selectivity of the cleavage/attachment site of phosphatidylinositol-glycan-anchored membrane proteins determined by site-specific mutagenesis at Asp-484 of placental alkaline phosphatase. *Proc. Natl Acad. Sci. USA*, **87**, 157-161.

Misumi, Y., Ogata, S., Ohkubo, K., Hirose, S. & Ikehara, Y. (1990). Primary structure of human placental 5′-nucleotidase and identification of the glycolipid anchor in the mature form. *FEBS Letters*, **191**, 563-569.

Moran, P. & Caras, I. W. (1991). Fusion of sequence elements from non-anchored proteins to generate a fully functional signal for glycosylphosphatidylinositol membrane anchor attachment. *J. Cell. Biol.* **115**, 1595-1600.

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

Moran, P., Raab, H., Kohr, W. J. & Caras, I. W. (1991). Glycophospholipid membrane anchor attachment. *J. Biol. Chem.* **266**, 1250-1257.

Morita, N., Nakazato, H., Okuyama, H., Kim, Y. & Thompson, G. A., Jr (1996). Evidence for a glycosylinositolphospholipid-anchored alkaline phophatase in the aquatic plant Spirodela oligorrhiza. *Biochim. Biophys. Acta*, **1290**, 53-62.

Nakai, K. & Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* **24**, 34-35.

Nakai, K. & Kanehisa, M. (1997). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897-911.

Nakashima, H. & Nishikawa, K. (1992). The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. *FEBS Letters*, **303**, 141-146.

Nielsen, H., Brunak, S. & von Heijne, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **12**, 3-9.

Nosjean, O., Briolay, A. & Roux, B. (1997). Mammalian GPI proteins: sorting, membrane residence and functions. *Biochim. Biophys. Acta*, **1331**, 153-186.

Nuoffer, C., Jenö, P., Conzelmann, A. & Riezmann, H. (1991). Determinants for glycophospholipid anchoring of the *Saccharomyces cerevisae* GAS1 protein to the plasma membrane. *Mol. Cell. Biol.* **11**, 27-37.

Nuoffer, C., Horvath, A. & Riezmann, H. (1993). Analysis of the sequence requirements for glycosylphosphatidylinositol anchoring of *Saccharomyces cerevisiae* Gas1 protein. *J. Biol. Chem.* **268**, 10558-10563.

Ogata, S., Hayashi, Y., Misumi, Y. & Ikehara, Y. (1990). Membrane-anchoring domain of rat liver 5′-nucleotidase: identification of the COOH-terminal serine-523 covalently attached with a glycolipid. *Biochemistry*, **29**, 7923-7927.

Okuyama, T., Waheed, A., Kusumoto, W., Zhu, X. L. & Sly, W. S. (1995). Carbonic anhydrase IV: role of removal of C-terminal domain in glycosylphosphatidylinositol anchoring and realization of enzyme activity. *Arch. Biochem. Biophys.* **320**, 315-322.

Orchansky, P. L., Escobedo, J. A. & Williams, L. T. (1988). Phosphatidylinositol linkage of a truncated form of the platelet-derived growth factor receptor. *J. Biol. Chem.* **263**, 15159-15165.

Santillán, G. E., Sandoval, M. J., Chernajovsky, Y. & Orchansky, P. L. (1992). Conversion of human interferon-β from a secreted to a phosphatidylinositol anchored protein by fusion of a 17 amino acid sequence to its carboxyl terminus. *Mol. Cell. Biochem.* **110**, 181-191.

Shen, F., Wu, M., Ross, J. F., Miller, D. & Ratnam, M. (1995). Folate receptor type γ is primarily a secretory protein due to lack of an efficient signal for glycosylphosphatidylinositol modification: protein characterization and cell type specifity. *Biochemistry*, **34**, 5660-5665.

Stahl, N., Baldwin, M. A., Burlingame, A. L. & Prusiner, S. B. (1990). Identification of glycoinositol phospholipid linked and truncated forms of the scrapie prion protein. *Biochemistry*, **29**, 8879-8884.

Su, B. & Bothwell, A. L. M. (1989). Biosynthesis of a glycosylphosphatidylinositol-glycan-linked membrane protein:signals for posttranslational processing of the Ly-6E antigen. *Mol. Cell. Biol.* **9**, 3369-3376.

Sugita, Y., Nakano, Y., Oda, E., Noda, K., Tobe, T., Miura, N. H. & Tomita, M. (1993). Determination of carboxy-terminal residue and disulphide bonds of MACIF (CD59), a glycosyl-phosphatidylinositol-anchored membrane protein. *J. Biochem.* **114**, 473-477.

Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G. & Kuznetsov, E. N. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.* **12**, 387-394.

Suzuki, K., Furukawa, Y., Tamura, H., Ejiri, N., Suematsu, H., Taguchi, R., Nakamura, S., Suzuki, Y. & Ikezawa, H. (1993). Purification and cDNA cloning of bovine liver 5′-Nucleotidase, a GPI-anchored protein, and its expression in COS cells. *J. Biochem.* **113**, 607-613.

Taguchi, R., Hamakawa, N., Harada-Nishida, M., Fukui, T., Nojima, K. & Ikezawa, H. (1994). Microheterogeneity in glycosylphosphatidylinositol anchor structures of bovine liver 5′-nucleotidase. *Biochemistry*, **33**, 1017-1022.

Taguchi, R., Yamazaki, J., Takahashi, K., Hirano, A. & Ikezawa, H. (1999). 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.* **363**, 60-67.

Takos, A. M., Dry, I. B. & Soole, K. L. (1997). Detection of glycosyl-phosphatidylinositol-anchored proteins on the surface of Nicotiana tabacum protoplasts. *FEBS Letters*, **405**, 1-4.

Tomassetti, A., Bottero, F., Mazzi, M., Miotti, S., Colnaghi, M. I. & Canevari, S. (1999). Molecular requirements for attachment of the glycosylphosphatidylinositol anchor to the human alpha folate receptor. *J. Cell Biol.* **72**, 111-118.

Tse, A. G. D., Barclay, A. N., Watts, A. & Williams, A. F. (1985). A glycophospholipid tail at the carboxyl terminus of the Thy-1 glycoprotein of neurons and thymocytes. *Science*, **230**, 1003-1008.

Udenfriend, S. & Kodukula, K. (1995a). How glycosylphosphatidylinositol-anchored membrane proteins are made. *Annu. Rev. Biochem.* **64**, 563-591.

Udenfriend, S. & Kodukula, K. (1995b). Prediction of omega site in nascent precursor of glycosylphosphatidylinositol protein. *Methods Enzymol.* **250**, 571-582.

Vai, M., Lacanà, E., Gatti, E., Breviario, D., Popolo, L. & Alberghina, L. (1993). Evolutionary conservation of genomic sequences related to the GGP1 gene encoding a yeast GPI-anchored glycoprotein. *Curr. Genet.* **23**, 19-21.

Waneck, G. L., Stein, M. E. & Flavell, R. A. (1988). Conversion of a PI-anchored protein to an integral membrane protein by a single amino acid mutation. *Science*, **241**, 697-699.

Wang, J., Shen, F., Yan, W., Wu, M. & Ratnam, M. (1997). Proteolysis of the carboxyl-terminal GPI signal independent of GPI modification as a mechanism for selective protein secretion. *Biochemistry*, **36**, 14583-14592.

Wang, J., Maziarz, K. & Ratnam, M. (1999). Recognition of the carboxyl-terminal signal for GPI modification requires translocation of its hydrophobic domain across the ER membrane. *J. Mol. Biol.* **286**, 1303-1310.

Wilbourn, B., Nesbeth, D., Wainwright, L. J. & Field, M. C. (1998). Proteasome and thiol involvement in quality control of glycosylphosphatidylinositol anchor addition. *Biochem. J.* **332**, 111-118.

Yan, W. & Ratnam, M. (1995). Preferred sites of glycosylphosphatidylinositol modification in folate receptors and constraints in the primary structure of the hydrophobic portion of the signal. *Biochemistry*, **34**, 14594-14600.

Yan, W., Shen, F., Dillon, B. & Ratnam, M. (1998). The hydrophobic domains in the carboxyl-terminal signal for GPI modification and in the amino-terminal leader peptide have similar structural requirements. *J. Mol. Biol.* **275**, 25-33.

Zhou, J., Dutch, R. E. & Lamb, R. A. (1997). Proper Spacing between Heptad repeat B and the transmembrane domain boundary of the paramyxovirus SV5 F protein is critical for biological activity. *Virology*, **239**, 327-339.

Edited by G. von Heijne

(Received 10 May 1999; received in revised form 21 July 1999; accepted 27 July 1999)