Automated annotation of GPI anchor sites: case study *C. elegans*

Functional characterization of a hypothetical protein as a potential substrate for glycosylphosphatidylinositol (GPI) lipid anchor post-translational modification is valuable, even if no other sequence-homology-based assignment of biological activity can be made. GPI anchor attachment implies that the protein (more correctly, an essential fraction of the protein's population) is located at the surface of biological membranes in the extracellular space, inside the (secretory) vesicular system of the Golgi apparatus or the endoplasmic reticulum. Therefore, its cellular function must be restricted to pathways and cascades of those compartments.

Recognition of the GPI modification signal in potential proprotein sequences: the public 'big-Π predictor' WWW server

Chemical linkage of the GPI moiety to the C-terminal residue (ω-site) of the polypeptide chain occurs by a transamidation reaction involving the simultaneous proteolytic cleavage of a C-terminal propeptide (typically 17–31 residues) from the proprotein[1,2]. We have found in a meta-analysis of the available sequence and literature data that the GPI modification signal is carried by the sequence segment from ω−11 to the C terminus of the proprotein[2]. The typical GPI modification signal motif contains four essential elements (see Fig. 1): (1) an unstructured linker region of ~11 residues (ω−11 ... ω−1); (2) a region of small residues (ω−1 ... ω+2) including the ω-site for propeptide cleavage and GPI attachment; (3) a spacer region (ω+3 ... ω+9) of moderately polar residues; and (4) a hydrophobic tail beginning with ω+9 or ω+10 up to the C-terminal end.

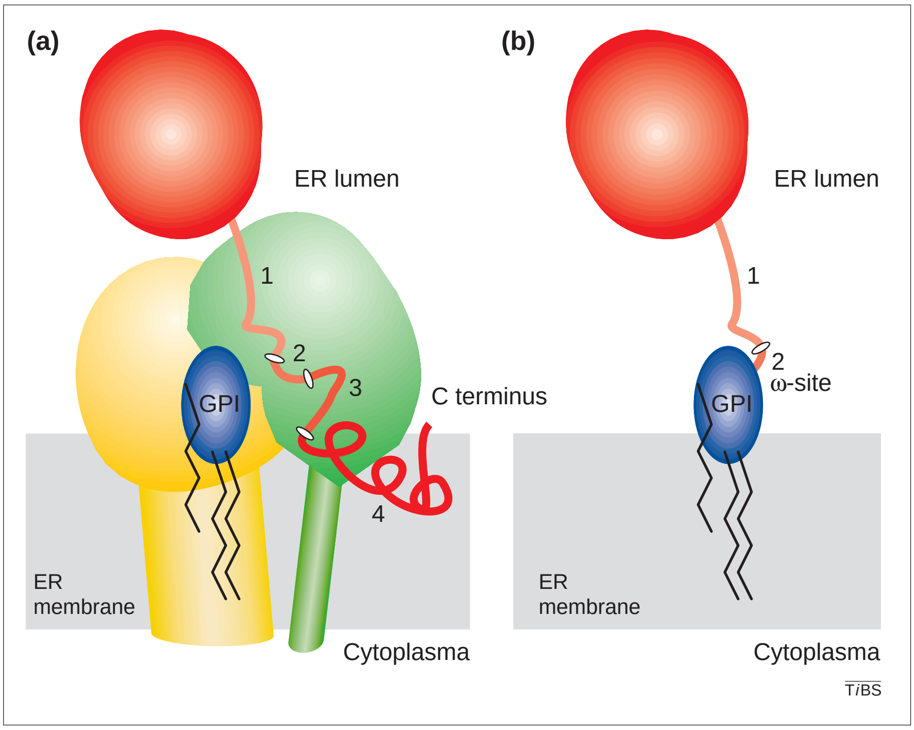

Figure 1
Post-translational modification of a substrate protein with a GPI lipid anchor. **(a)** The C-terminal tail of a substrate protein [red, located in the endoplasmic reticulum (ER) lumen] needs to be easily unfolded to fit into the catalytic cleft of the catalytic subunit (green). Probably, another subunit (yellow) supplies the GPI lipid anchor moiety (blue) to the free ω-site residue after propeptide cleavage. The four regions of the C-terminal signal motif are numbered (1–4) and have different colors red (from light- to dark-red). The inter-regional boundaries are indicated. **(b)** After release from the putative transamidase complex, only region 1 and partially region 2 remain with the post-translationally modified protein. A GPI lipid anchor being attached to the now C-terminal ω-site residue fixes the protein to the non-cytoplasmic side of the membrane.

The GPI modification signal is not well characterized by specific amino acid type preferences, apart from the ω and ω+2 positions where small residues such as Ala, Asn, Asp, Cys, Gly and Ser are statistically favoured[1,2]. A motif description in terms of physical properties of amino acid side chains and multi-residue correlations is much more adequate[2,3]. Additionally, the absence of one of the single necessary signal elements is not compensated by the rest of the sequence; a single-residue substitution might change an otherwise 100% GPI-anchored protein to a completely non-anchored version. Many such examples have been noted among the almost 400 mutations registered in the big-Π mutation database, which is available at http://mendel.imp.univie.ac.at/gpi/gpi.m/gpi.mut.html (Ref. 3).

As the first sufficiently reliable GPI-site annotation tool, the big-Π predictor[3] both analyses a query sequence for the occurrence of a GPI anchor post-translational modification and predicts the best potential ω-site(s). The input consists of the proprotein sequence alone. The algorithm employs a composite prediction function: $S = S_{\text{profile}} + S_{\text{physical\_pattern}}$, incorporating both special profile-based score terms S_{profile} (Ref. 4), evaluating the residual amino acid type preferences, as well as terms $S_{\text{physical\_pattern}}$, assessing the concordance of sequence segments with conservation patterns of physical properties of amino acid side chains. The latter contribution contains functions for evaluating (1) ω-site region properties (side-chain volume effects[2] and backbone flexibility near the cleavage site), (2) hydrophilicity and side-chain volume in the spacer region, (3) degree of hydrophobicity and backbone flexibility of the hydrophobic tail, as well as the even distribution of hydrophobic residues in that region[3].

There are subtle taxon-specific differences among the sequence motifs for GPI-modification[1,5,6], but, unfortunately, the available sequence data were only sufficient to parametrize score functions for metazoan and protozoan proteins reliably[3]. The accuracy of the big-Π predictor was tested with a jack-knife approach for the learning sets of known

0968–0004/00/$ – See front matter © 2000, Elsevier Science Ltd. All rights reserved. PII: S0968-0004(00)01601-7

GPI-modified proteins (i.e. leaving the sequence for prediction out of the procedure for prediction function parameter computation) and, independently, for the set of (permissive and non-permissive) GPI-modification motif mutations collected from the literature. In all cases, the prediction accuracy was $>80\%$. Searches in protein databases (SWISS-PROT, SP-TrEMBL) among unrelated sequences revealed only a single, clearly false-positive, hit (with a P value <0.01)[3]. This prediction success is especially remarkable because, first, the quality of the experimental data described in the literature and in databases are very uneven[3]; and, second, the variety of available learning sequences is currently insufficient to derive a profile parameter set that is 100% stable for jack-knife testing. It should be emphasized that all parameters in $S_{\text{physical\_pattern}}$ were determined by *ad hoc* physical considerations, without numerical optimization for higher prediction accuracy.

Our prediction tool is now available as a public WWW server at http://mendel.imp.univie.ac.at/gpi/gpi_server.html. Another implementation of the server can be reached at http://www.embl-heidelberg.de/~beisenha/gpi/gpi_server.html. In its latest version, the server returns both the final conclusion (with or without GPI modification, sequence position of ω-sites, total scores and P values), as well as the details for each term of the score function. In this way, the user obtains information about the sequence properties that caused the automatic predictor to exclude or to justify GPI modification for the given query.

Annotation of potentially GPI-modified proteins encoded in the complete genome of *C. elegans*

We have applied the big-Π predictor on the protein sequences derived from the genome of *Caenorhabditis elegans* (ftp://ftp.sanger.ac.uk/pub/databases/wormpep/wormpep17). The complete prediction results are available at http://mendel.imp.univie.ac.at/gpi/gpi.g/gpi.C_elegans.html. Among the total of 19216 protein sequences, our computer tool identified 41 proteins for which GPI modification is almost sure ($P < 0.01$). Another 82 proteins were found in the twilight zone defined with $P < 0.0175$ (Ref. 3). All these molecules are potential candidates for experimental testing of their GPI modification.

The terms $S_{\text{physical\_pattern}}$ of the prediction function used for scoring physical properties are constructed in a manner so that they are always equal to or lower than zero. All deviations from the ideal physical property pattern are penalized with negative score contributions. Only the profile component S_{profile} can add positive, negative or zero (in the special case of a sequence indifferent to the profile) contributions to the total score. Therefore, taking into account that the learning dataset size does not allow a fully reliable parametrization of the profile term S_{profile}, the criterion of a non-negative total score S appears reasonable for subselection of likely hits in the twilight zone. This method has been applied successfully for predicting the outcomes of mutation experiments enlisted in the big-Π mutation database[3]. In the case of *C. elegans*, 45 among the 82 twilight zone hits have a non-negative score. Thus, we predict that the genome of *C. elegans* encodes 86 (= 41 + 45) GPI-modified proteins (0.45% of the total number of proteins).

It is especially interesting that we detected 63 hypothetical proteins (in accordance with their database annotation) in the genome of *C. elegans*, which appear post-translationally modified by a GPI anchor. We found also that the chromosomal distribution of GPI-anchored proteins in *C. elegans* is not even. For example, the fraction of proproteins that appear GPI-modified among the total number of proteins is 0.19% for chromosome 3 and 0.72% for chromosome 10 of the worm. Thus, the big-Π predictor can be a helpful tool in genome annotation.

Although the parameter sets of the prediction function are specific for animal sequences, it is likely that the prediction tool is also helpful in finding at least some candidate GPI-anchored proteins in other species. There is both a biological and a methodical justification. First, among eukarya, a common evolutionary origin of the cellular GPI-modification machinery might be hypothesized. The general outline of the sequence motif for GPI modification is similar among species repeating the four-element scheme described above[1,2]. Second, the subtle species-specific differences affect mainly the profile-dependent part of the score function. At the same time, there is little difference in the parameters for physical terms between Metazoa and Protozoa; thus, these parameters are probably transferable to other species.

The research community is invited to use the server. Feedback about discrepancies between predictions and experimental data is welcome (email Birgit.Eisenhaber@nt.imp.univie.ac.at). This will help us to continue our work on improving the prediction function and to provide a better service.

Acknowledgements

We are grateful to numerous researchers, especially Göran Bucht, Mike A. Ferguson, Frans M. Klis and Jens G. Reich, for valuable advice, encouragement, for providing manuscripts prior to publication and unpublished experimental data used for testing the predictor.

References

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

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

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

4 Sunyaev, S.R. *et al.* (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.* 12, 387–394

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

6 Caro, L.H.P. *et al.* (1997) *In silicio* identification of glycosyl-phosphatidylinositol-anchored plasma-membrane and cell wall proteins of *Saccharomyces cerevisiae. Yeast* 13, 1477–1489

BIRGIT EISENHABER

Max-Delbrück-Centrum für Molekulare
Medizin, Robert-Rössle-Straβe 10, D-13122
Berlin-Buch, Germany; Research Institute of
Molecular Pathology, Dr. Bohr-Gasse 7,
A-1030 Vienna, Austria; and European
Molecular Biology Laboratory,
Meyerhofstrasse1, Postfach 10.2209,
D-69012 Heidelberg, Germany.
Email: Birgit.Eisenhaber@nt.imp.univie.ac.at

PEER BORK

Max-Delbrück-Centrum für Molekulare
Medizin, Robert-Rössle-Straβe 10, D-13122
Berlin-Buch, Germany; and European
Molecular Biology Laboratory,
Meyerhofstrasse1, Postfach 10.2209,
D-69012 Heidelberg, Germany.

YUANPING YUAN

European Molecular Biology Laboratory,
Meyerhofstrasse1, Postfach 10.2209,
D-69012 Heidelberg, Germany.

GERALD LÖFFLER AND
FRANK EISENHABER

Research Institute of Molecular Pathology,
Dr. Bohr-Gasse 7, A-1030 Vienna, Austria.