Protein Engineering vol.14 no.1 pp.17–25, 2001

Post-translational GPI lipid anchor modification of proteins in kingdoms of life: analysis of protein sequence data from complete genomes

Birgit Eisenhaber[1,2,3], Peer Bork[1,4] and Frank Eisenhaber[2]

[1]Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle-Straβe 10, D-13122 Berlin-Buch, Germany, [2]Research Institute of Molecular Pathology, Dr. Bohr-Gasse 7, A-1030 Vienna, Austria and [4]European Molecular Biology Laboratory, Meyerhofstrasse1, Postfach 10.2209, D-69012 Heidelberg, Germany

[3]To whom correspondence should be addressed at: Research Institute of Molecular Pathology, Dr. Bohr-Gasse 7, A-1030 Vienna, Austria
E-mail: birgit.eisenhaber@nt.imp.univie.ac.at

To investigate the occurrence of glycosylphosphatidylinositol (GPI) lipid anchor modification in various taxonomic ranges, potential substrate proteins have been searched for in completely sequenced genomes. We applied the big-π predictor for the recognition of propeptide cleavage and anchor attachment sites with a new, generalized analytical form of the extreme-value distribution for evaluating false-positive prediction rates. (i) We find that GPI modification is present among lower and higher Eukaryota (~0.5% of all proteins) but it seems absent in all eubacterial and three archaeobacterial species studied. Four other archaean genomes appear to encode such a fraction of substrate proteins (in the range of eukaryots) that they cannot be explained as false-positive predictions. This result supports the possible existence of GPI anchor modification in an archaean subgroup. (ii) The frequency of GPI-modified proteins on various chromosomes of a given eukaryotic species is different. (iii) Lists of potentially GPI-modified proteins in complete genomes with their predicted cleavage sites are available at http://mendel.imp.univie.ac.at/gpi/gpi_genomes.html. (iv) Orthologues of known transamidase subunits have been found only for Eukarya. Inconsistencies in domain structure among homologues some of which may indicate sequencing errors are described. We present a refined model of the transamidase complex.

Keywords: genome annotation/GPI lipid anchor attachment/GPI modification prediction/post-translational modification/transamidase complex

Abbreviations: GPI, glycosylphosphatidylinositol; TM, transmembrane; ER, endoplasmic reticulum

Introduction

The evolution of the cellular glycosylphosphatidylinositol (GPI) modification machinery and its occurrence in various taxonomic ranges remains an open question despite the intensive research efforts of approximately two decades. Almost all experimentally verified GPI-modified proteins are from animals (Ferguson and Williams, 1988; Udenfriend and Kodukula, 1995) or their viruses (Zhou *et al.*, 1997). Only singular examples from plants (Vai *et al.*, 1993; Morita *et al.,* 1996; Takos *et al.,* 1997; Youl *et al.,* 1998; Oxley and Bacic, 1999; Sherrier *et al.,* 1999), fungi (Vai *et al.,* 1993; Guadiz *et al.,* 1998; Popolo and Vai, 1999) and archaeobacteria (Kobayashi *et al.,* 1997) have also been reported.

In this work, we extend our previous analysis of the *Caenorhabditis elegans* genome (Eisenhaber *et al.,* 2000) and investigate the occurrence of GPI-modified proteins throughout the taxonomic spectrum from two points of view. (i) We analyze publicly available complete genomes/chromosomes for proprotein sequences and report lists of potentially GPI-anchored proteins. The existence of a significant number of such proteins can be considered as an indirect hint that the given organism possesses an enzyme complex for GPI post-translational modification. (ii) We search for orthologues of subunits of the transamidase complex executing the GPI modification. To summarize, our results indicate that GPI modification is common among Eukaryota and possibly also among a subset of archaean species but probably absent among all other Archaea and all Eubacteria.

Theory: estimation of false positive rates for GPI lipid anchor modification prediction

To be recognized as substrate of the GPI modification enzyme complex, a specific C-terminal sequence motif in the proprotein sequence appears necessary and sufficient, given the protein is exported from the cytoplasma to the endoplasmic reticulum (ER). Chemical linking of the GPI moiety to the C-terminal residue (ω-site) of the polypeptide chain occurs by a trans-amidation reaction after proteolytic cleavage of a C-terminal propeptide from the proprotein (Udenfriend and Kodukula, 1995; Sharma *et al.*, 1999; Meyer *et al.*, 2000). The sequence motif obtained from a comparison of known substrate proteins (Eisenhaber *et al.*, 1998) includes four sequence elements:

(i) an unstructured linker region of about 11 residues (ω–11 ... ω–1) for connection of the substrate protein with the catalytic cavity of the transamidase complex;
(ii) a region of four preferably small residues (ω–1 ... ω+2) fitting the catalytic cleft and including the ω-site for propeptide cleavage and GPI-attachment;
(iii) an, on average, moderately polar spacer region (ω+3 ... ω+9); and
(iv) a hydrophobic tail beginning, as a rule, with ω+9 or ω+10 up to the C-terminal end.

It should be noted that all four signal elements are necessary for recognition by the transamidase complex; even a single-residue substitution may change an otherwise 100% GPI-anchored protein to a completely non-anchored version (Eisenhaber *et al.*, 1999). The general theme is varied by subtle taxon-specific differences among the sequence motifs for GPI-modification (Moran and Caras, 1994; Udenfriend and Kodukula, 1995; Eisenhaber *et al.*, 1998; Takos *et al.*, 2000). For example, highest efficiency of cleavage has been observed for asparagine at the ω-site for yeast (Nuoffer *et al.*, 1993) but for serine in the case of Mammalia (Micanovic *et al.*, 1990) although both amino acid types are permitted in both taxons.

We have recently developed a reliable GPI-site annotation tool, the big-π predictor (Eisenhaber *et al.*, 1999). In its composite prediction function ***S*** parametrized for animal sequences (separately for Metazoa and Protozoa), available knowledge about the GPI modification sites has been incorporated. The total score ***S*** consists of two parts:

$$\boldsymbol{S} = \boldsymbol{S_{profile}} + \boldsymbol{S_{ppt}} \tag{1}$$

The profile-dependent section $\boldsymbol{S_{profile}}$ evaluates the concordance with the weak amino acid type preferences in the learning set at single alignment positions (Eisenhaber *et al.*, 1998; Sunyaev *et al.*, 1999). The physical property term score $\boldsymbol{S_{ppt}}$ describes the conservation of physical properties in the GPI modification signal arising from the interaction of few or many sequence positions.

The efficiency of a prediction method can be evaluated with the indicators ‘sensitivity’ (or the rate of false negative predictions) and ‘selectivity’ (or the rate of false positive predictions). The big-Π predictor was shown to recognize 100% of all examples with full experimental verification of the ω-site in the learning set and above 80% of all proteins, if examples with indirect experimental evidence of GPI modification and with sequence similarity considerations are also included (Eisenhaber *et al.*, 1999); thus, the rate of false negative predictions is thought to be below 20% for animal sequences.

In our previous paper (Eisenhaber *et al.*, 1999), the resulting total score ***S*** of the prediction function has been translated into the probability (***P*** value) of a false positive GPI-site prediction with the help of an extreme value distribution (Altschul *et al.*, 1994). The probability of having incidentally a score ***S*** for a given sequence larger than a threshold $\boldsymbol{S_{th}}$ is

$$\boldsymbol{P}(\boldsymbol{S}>\boldsymbol{S_{th}}) = 1 - \exp\{-\exp[-\boldsymbol{f}(\boldsymbol{S_{th}})]\} \tag{2}$$

with

$$\boldsymbol{f}(\boldsymbol{S_{th}}) = \lambda(\boldsymbol{S_{th}}-\boldsymbol{u}) \tag{3}$$

The parameters λ and ***u*** have been determined from the empirical distribution function of scores calculated from sets of SWISS-PROT sequences without the keyword ‘GPI-anchor’. Statistical tests have shown that the regression between

$$-\ln[-\ln(1-\boldsymbol{P_{observed}}(\boldsymbol{S}>\boldsymbol{S_{th}}))] \text{ and } \boldsymbol{f}(\boldsymbol{s_{th}})$$

is valid if calculated over the whole argument space [(see legend to Fig. 1 in Eisenhaber *et al.* (1999)]. At the same time, it was obvious that the fit does dramatically overestimate the probability of false positive prediction in the interesting range of scores $\boldsymbol{S_{th}} > -10$ (see Fig. 1 in that paper). Nevertheless, a ***P*** value estimated close to 0.01 appeared to us a reasonably low risk for a researcher investigating the possibility of GPI anchoring for a single sequence (Eisenhaber *et al.*, 1999).

This extreme-value distribution approximation is too crude for the analysis of genomes having in the order of 2000 genes since the expected rate of false positives would reach about 20 proteins. The key to a more accurate statistical description can be found if correlation between terms constituting the total score ***S*** is taken into account. We recall that the maximum of a normally distributed random variable is extreme-value distributed and this is true also for sums of normally distributed and non-correlated random variables (Kendall and Stuart, 1977). The latter condition is not well observed for extreme scores in our case. For example, the volume terms $\boldsymbol{T}_0$, $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ [described in Methods of Eisenhaber *et al.* (1999)] have a strong tendency to be small or large together but it is physically unlikely that one of the terms is extremely large whereas the remaining two are small. Therefore, the number of realizations of large scores is smaller than expected from the theoretically assumed distribution. Such positive correlation among score components results in a dependency of ***P*** not only on $\boldsymbol{T}_0$, $\boldsymbol{T}_1$, ... but also on $\boldsymbol{T}_0\cdot\boldsymbol{T}_1$,$\boldsymbol{T}_0\cdot\boldsymbol{T}_1\cdot\boldsymbol{T}_2$, etc. This effect can be included in the analytical form of the distribution (Equation 2) with a polynomial exponent ***f*** instead of just a linear fit; i.e., ***P*** may depend on $\boldsymbol{S_{th}}$, $\boldsymbol{S^2_{th}}$ etc.

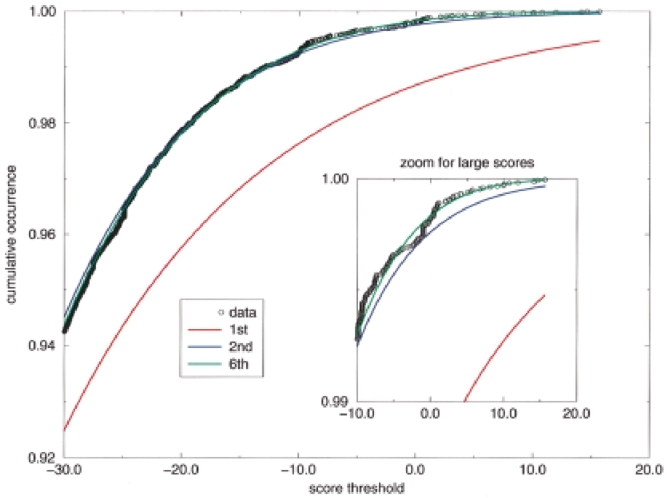

Fig. 1. Approximation of the empirical score distribution function calculated for non-GPI-annotated metazoan sequences with statistical functions. Any sequence motif may occur incidentally in a non-related protein; therefore, the statistical significance of a match between query sequence and pattern must be evaluated. This figure illustrates the empirical distribution function $\boldsymbol{P_{observed}}(\boldsymbol{S} \leqslant \boldsymbol{S_{th}})$ of best scores *S* for the 23989 non-annotated metazoan sequences (with a length ⩾ 55 residues and without the keyword GPI-anchor) in the rel. 37 of SWISS-PROT (black circles) for thresholds $\boldsymbol{S_{th}} > -30$ (for $\boldsymbol{S_{th}} > -10$ in the insert). This set has been selected as test set of ‘unrelated’ sequences. Additionally, we show the theoretical distribution functions calculated with equations (2) and (4) for polynomials with ***n*** = 1 (red, linear fit, classical extreme-value distribution), ***n*** = 2 (blue, quadratic fit) and ***n*** = 6 (green). It should be emphasized that even the new, more accurate approximations probably overestimate the rate of false predictions since some of the examples with positive scores among the 23989 sequences tested may be truly GPI anchored proteins.

$$\boldsymbol{f}(\boldsymbol{S_{th}}) = \sum_{i=1}^{n} \lambda_i(\boldsymbol{S_{th}}-\boldsymbol{u})^i \tag{4}$$

here, ***n*** denotes the degree of the polynomial function.

Indeed, it is possible to have a good fit both over the whole argument range and for score thresholds above $\boldsymbol{S_{th}} > -10$ at the same time without changing the general form of the extreme-value distribution (2) but using (4) instead of (3) for the exponent. In the case of the metazoan score function, the residual ***R*** of the least-square fit

$$\boldsymbol{R} = \sum_{j=1}^{k} [-\ln(-\ln(1-\boldsymbol{P_{observed}}(\boldsymbol{S}<\boldsymbol{S_{th,j}})))-\boldsymbol{f}(\boldsymbol{S_{th,j}})]^2\sigma_j^{-2} \tag{5}$$

(***k*** is the number of sequences without keyword ‘GPI-anchor’ in SWISS-PROT release 37, ***k*** = 23989 for Metazoa and 1062 for Protozoa, $\forall \boldsymbol{j}:\sigma_j = 1$ for calculating the residual) over the whole argument space is ***R*** = 2383.3 for ***n*** = 1, 61.1 for ***n*** = 2, 17.9 for ***n*** = 3, 17.8 for ***n*** = 4, 5.8 for ***n*** = 5, 4.5 for ***n*** = 6, and 4.5–4.2 for ***n*** between 7 and 10. If the unchanged

Table I. Estimates of false positive prediction rates as a function of score thresholds

Prediction class	Score threshold	$n = 1$	$n = 2$	$n = 6$
(a) For the metazoan parametrization				
A	28.15	0.0025	5.7×10^{-5}	5.8×10^{-7}
B	16.41	0.0050	3.0×10^{-4}	4.5×10^{-5}
C	9.54	0.0075	7.3×10^{-4}	2.7×10^{-4}
D	4.66	0.0100	0.0014	7.3×10^{-4}
Positive score	0.00	0.0132	0.0024	0.0017
S	–4.86	0.0175	0.0042	0.0035
(b) For the protozoan parametrization				
A	29.69	0.0025	1.3×10^{-4}	2.5×10^{-11}
B	16.21	0.0050	5.7×10^{-4}	5.4×10^{-6}
C	8.31	0.0075	0.0013	2.2×10^{-4}
D	2.70	0.0100	0.0022	0.0012
Positive score	0.00	0.0115	0.0028	0.0022
S	–8.24	0.0175	0.0080	0.0077

The probabilities of $\boldsymbol{P}(\boldsymbol{S} <$ score threshold) are listed for the $\boldsymbol{n} = 1$ linear fit (Eisenhaber *et al.*, 1999), the quadratic fit ($\boldsymbol{n} = 2$) and the fit with a 6th-degree polynomial function (equation 4) are listed.

polynomial coefficients are used to calculate the residual $\boldsymbol{R}$ just over the range $\boldsymbol{S}_{th} > -10$, then we find $\boldsymbol{R} = 628.8$ for $\boldsymbol{n} = 1$, 30.3 for $\boldsymbol{n} = 2$, 7.5 for $\boldsymbol{n} = 3$, 7.0 for $\boldsymbol{n} = 4$, 2.6 for $\boldsymbol{n} = 5$, 2.5 for $\boldsymbol{n}$ between 6 and 10. Obviously, already the quadratic function corrects the fit dramatically. Further increase of $\boldsymbol{n}$ results only in gradual improvement but no essential change in the residual can be found for $\boldsymbol{n} > 6$. The χ^2-homogeneity test (Kendall and Stuart, 1977) cannot distinguish between the goodness of fit for different polynomial degrees over the whole argument space [for χ^2 computation, the variance σ_j in (5) of the observed data was estimated using the three sequence datasets described in the legend of Fig. 1 in Eisenhaber *et al.* (1999)]. But, for $\boldsymbol{S}_{th} > -10$, the regressions for $\boldsymbol{n} = 1$ and $\boldsymbol{n} = 2$ (significance = 1.00) as well as for $\boldsymbol{n} = 3$ (significance > 0.10) and $\boldsymbol{n} = 4$ (significance > 0.05) are clearly worse than those for $\boldsymbol{n} \geqslant 5$ (significance $< 10^{-10}$). We have found the same behaviour for the protozoan score function (data not shown).

In Table I, the estimated rates of false positives for the prediction classes A, B, C, D, and S defined in Eisenhaber *et al.* (Eisenhaber, *et al.*, 1999) are presented (see also Figure 1). We give values both for $\boldsymbol{n} = 2$ and $\boldsymbol{n} = 6$, the latter being used in the following sections of this paper. The $\boldsymbol{P}$ values for a zero total score are about an order of magnitude lower than those calculated in our previous work for both $\boldsymbol{n}$. In a genome with about 2000 genes, we will expect an incidental occurrence of ~2 proteins with $\boldsymbol{S} \geqslant 0$ and ~1 protein with prediction class D or better. It should be emphasized that the set of non-annotated proteins from SWISS-PROT consists probably not only of unrelated proteins but comprises a few real GPI-anchored proteins; hence, even this assessment of false positives may be an overestimation.

The new method for $\boldsymbol{P}$ value computation has been implemented in an updated version of the big-Π predictor and is available at http://mendel.imp.univie.ac.at/gpi/gpi_server.html.

Calculational methods

GPI motifs in proprotein sequences have been searched for with the big-π predictor (Eisenhaber *et al.*, 1999). A sequence was considered a hit if at least one of the two parametrizations (for protozoan or for metazoan proteins) was sufficient to recognize concordance with the motif's properties.

From the total score $\boldsymbol{S}$, the probability ($\boldsymbol{P}$ value) of a false positive GPI-site prediction is calculated with the help of an extreme value distribution (equation 2). For qualitative comparison of prediction results, predictions are labelled with ratings A, B, C, D and S (Table I). All sequences with a $\boldsymbol{P}$ value above the S-threshold are not predicted as potential GPI-anchored proteins (label N). Additionally, all sequences having (i) a negative total score $\boldsymbol{S}$ and a profile-independent score $\boldsymbol{S}_{ppt}$ below –8 or (ii) the S-label and a profile-independent score $\boldsymbol{S}_{ppt}$ below –12 are also excluded as possible GPI-modification candidates and are labelled I (Eisenhaber *et al.*, 1999).

For predictions of TM regions and the orientation of the protein with respect to the membrane sides (cytoplasmic/outside cytoplasmic), the TOPPRED2 (Claros and von Heijne, 1994) suite has been applied. Signal peptides have been recognized with SIGNALP (Nielsen *et al.*, 1997; Nielsen *et al.*, 1999). For coiled coil and secondary structure predictions, the tools COILS2 (Lupas, 1997) and PREDATOR (Frishman and Argos, 1997) have been used. Sequence homology considerations have been applied to unify structural predictions within a given family of proteins; sequentially similar regions are supposed to adopt the same structure.

Sequence similarities to globular domains were searched for with the Blast/PSI-Blast (Altschul *et al.*, 1997) and PFAM/PFAM-FRAGMENT (Bateman *et al.*, 2000) tools. Unfortunately, the C13 alignment in PFAM included also the signal peptide region when this research was carried out. For our purpose, the initial part of this alignment has been excluded. Sequence motif searches in sequence databases have been carried out with BIOMOTIF from G. Mennessier (http://www.infobiogen.fr/doc/bioMotifdoc/bioMotif_ref.htm).

Results

I. Occurrence of potentially GPI-modified proteins in kingdoms of life

We have applied the big-π predictor on the protein sequences derived from 17 eubacterial, seven archaebacterial, and two eukaryotic complete genomes as well as from several complete chromosomes of parasitic protozoa (chromosome 1 of *Leishmania major*, chromosomes 2 and 3 of *Plasmodium falciparum*). The genome data except for five genomes have been taken from http://ncbi.nlm.nih.edu/genbank/genomes (Tatusova *et al.*, 1999) as from October 1999. The results including complete lists of entries predicted to describe GPI-modified proteins for all studied genomes/chromosomes as well as the genome references are available at http://mendel.imp.univie.ac.at./gpi/gpi_genomes.html.

In all eubacterial genomes, we have never observed a good prediction for a GPI-modified protein with a $\boldsymbol{P}$ value < 0.00027 (labels A, B, or C; see Table I) or more than one hit with label D (P value < 0.0012 for protozoan function and < 0.0008 for the metazoan function). Only for the sets of proteins encoded by the genomes of *Mycobacterium tuberculosis* and *Campylobacter jejuni*, we found just two hits with label D. The same observation has been found for *Methanococcus jannaschii*, *Pyrococcus abyssi* (both with zero hits with label D or better) and *Pyrococcus furiosus* (one hit with label D and none better than D). It should be noted that the expected rate of false positives is ~1 hit with label D or better (Table I).

This is in sharp contrast to data for the remaining archaebacterial and all eukaryotic genomes. A summary of these

Table II. Number of potential proproteins for GPI modification in eukaryotic and archaebacterial genomes

Genome	M	N	G	H_1	$G+H_1$	H_2	$G+H_1+H_2$
Caenorhabditis elegans	19126	18952	41	45	86 (0.45%)	37	123 (0.64%)
Leishmania major (ch. 1)	79	79	0	0	0 (0.00%)	0	0 (0.00%)
Plasmodium falciparum (ch. 2, 3)	430	425	1	1	2 (0.47%)	0	2 (0.47%)
Saccharomyces cerevisiae	6530	6483	27	14	41 (0.63%)	14	55 (0.84%)
Aeropyrum pernix	2694	2687	4	12	16 (0.59%)	13	29 (1.08%)
Archaeoglobus fulgidus	2407	2357	5	3	8 (0.33 %)	2	10 (0.42%)
Methanobacterium thermoautotrophicum	1869	1846	5	6	11 (0.59%)	7	18 (0.42%)
Methanococcus jannaschii	1715	1703	0	0	0 (0.00%)	3	3 (0.17%)
Pyrococcus abyssi	1765	1760	0	1	1 (0.06%)	4	5 (0.28%)
Pyrococcus furiosus	2208	2073	1	0	1 (0.05%)	3	4 (0.18%)
Pyrococcus horikoshii	2064	2056	6	5	11 (0.53%)	7	18 (0.87%)

The column headings have the following meaning: M is the total number of proteins described for the given genome/set of chromosomes. N denotes the number of proteins with sufficient quality for prediction (length >55 residues, without non-amino acid letters at the C-terminus). G, H_1, and H_2 are the number of predictions with different levels of certainty. In the case of G, label D or better is required. For the twilight zone predictions H_1 and H_2 (label S), the total prediction score is non-negative for H_1 and negative for H_2. The number $G+H_1$ is considered the total number of predicted GPI-anchored proteins (Eisenhaber *et al.*, 1999). An upper limit for the number of proteins for GPI modification encoded by the given genome is given by $G+H_1+H_2$.

results is given in Table II. In the case of *Caenorhabditis elegans* and *Saccharomyces cerevisiae*, dozens of highly probable hits (with label D or better, column G) were found. The observed frequency of hits with label D or better is 0.0021 for *C.elegans*, 0.0041 for *S.cerevisiae*, 0.0015 for *Aeropyrum pernix*, 0.0025 for *Archaeoglobus fulgidus*, 0.0027 for *Methanobacterium thermoautotrophicum*, and 0.0029 for *Pyrococcus horikoshii*. The expected false positive rate for archaean species to compare with is below 0.0008 since the protozoan function did not give hits with label D or better at all (Table I); thus, the observed values are 2–4 times larger.

Predictions with label S have scores near zero and belong to the twilight zone (Eisenhaber *et al.*, 1999). Whereas the terms in our prediction function used for scoring physical sequence properties result always in non-positive contributions to the total score, the profile component may add positively or negatively to the total score. Since the latter contribution could not be parametrized stably for a complete jackknife test, it has a lower reliability for excluding the possibility of GPI modification (Eisenhaber *et al.*, 1999). Therefore, the criterion of a non-negative total score appears reasonable for subselection of likely hits in the twilight zone. This method has been successfully applied for predicting the outcomes of mutation experiments enlisted in the big-π mutation database (Eisenhaber *et al.*, 1999). In Table II, the two types of prediction have been listed as H_1 (S-labelled predictions with score ≥ 0) and H_2 (S-labelled predictions with score <0). We consider the value $G+H_1$ as number of true predictions, the numbers G and $G+H_1+H_2$ as lower and upper bounds for the amount of GPI-modified proteins (Table II).

The absolute number of proproteins to be GPI-modified may depend on the absolute number of genes in the genome. The relationship between the fraction of proteins predicted to be GPI-modified $(G+H_1)/M$ and the total number of proteins M (about the total number of genes) in the genome is shown in Figure 2. Clearly, the genomes cluster in two groups.

The frequency of hits with positive score is below 0.18% for all eubacterial genomes except for *Mycoplasma genitalium* (1 hit out of 480 proteins, 0.21%) and *Escherichia coli* (10 hits out of 4289 proteins, 0.23%); thus, it is below or very near to the expected rate of false positive predictions (0.22%) for all eubacterial genomes. The same is true for the genomes of *Methanococcus jannaschii* (0.00%), *P.abyssi* (0.06%) and *P.furiosus* (0.05%). The remaining four archaean and the two complete eukaryotic genomes have total frequencies of hits above 0.45% (except for *A.fulgidus* with 0.33%).

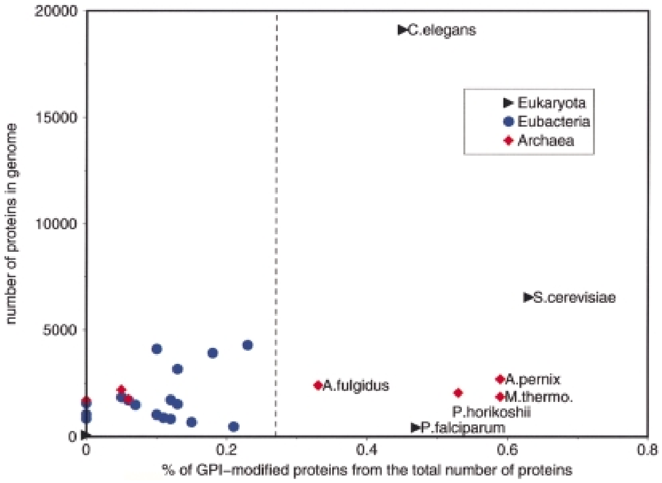

Fig. 2. Fraction of GPI-anchored proteins versus genome size. The relationship of $(G+H_1)/M$ *versus* the number of proteins M (see Table I) is illustrated. Each blue circle, red diamond, and black triangle represents the data for a given complete genome/set of chromosomes of an eubacterial, archaean, or eukaryotic organism respectively. The eubacterial genomes studied are those of *Aquifex aeolicus*, *Bacillus subtilis*, *Borrelia burgdorferi*, *Campylobacter jejuni*, *Chlamydia pneumoniae*, *Chlamydia trachomatis*, *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Helicobacter pylori J99*, *Mycobacterium tuberculosis*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Rickettsia prowazekii*, *Synechocystes* sp., *Thermotoga maritima*, and *Treponema pallidum*. The remaining markers without identifier are those for the archaea *Methanococcus jannaschii*, *Pyrococcus abyssi*, *Pyrococcus furiosus*, and for chromosome 1 of *Leishmania major.* The dashed line visualizes the obvious clustering of genomes with respect to the fraction of GPI-anchored proteins in the total genome.

Whereas the chromosomes 2 and 3 of *Plasmodium falciparum* have a typical content of proproteins predicted to be GPI-modified (0.47%), there are none found in chromosome 1 of *L.major*, a parasite known to have many GPI-anchored surface antigens (Smith *et al.*, 1997; Descoteaux and Turco, 1999). This finding might indicate that, even if GPI-modified proteins exist in any eukaryotic species, these proteins may occur with non-equal frequency on various chromosomes of that species. Indeed, we found that the chromosomal distribution of GPI-anchored proteins in the two known complete eukaryotic genomes is also not even. For example, the fraction of proproteins to be GPI-modified among the total number of proteins is only 0.19% for chromosome 3 but 0.72% for chromosome 10 of the metazoon *C.elegans* worm. In the case of the fungus *S.cerevisiae*, the fraction is between 0.00% (chromosome 6) and 2.72% (chromosome 1) for all 16 chromosomes of yeast (the predicted fractions in percentages are 2.72, 0.22, 0.58, 0.59, 1.07, 0.00, 0.16, 0.72, 0.85, 0.49, 0.86, 1.40, 0.80, 0.44, 0.49, and 0.20 for chromosomes 1–16 respectively).

For completeness, we mention the previous attempt of Caro *et al.* (Caro *et al.*, 1997) to find GPI-modified proteins in yeast with a very simple GPI motif description. They searched for sequences with Asn or Ser close to the C-terminal followed by a hydrophobic region. Our predictions agree in 39 out of 58 cases thought by Caro *et al.* to be GPI-modified if only the condition of a near zero scoring physical term is required (see http://mendel.imp.univie.ac.at/gpi/gpi.g/comp_klis.html). In 19 sequences, our predictor notes the absence of at least one of the necessary four sequence signals.

Finally, it should be noted that the accuracy of our prediction of numbers of GPI-modified proteins depends heavily on the correctness and the completeness of the protein library derived from the genome sequence. Especially for eukaryotic genomes, this condition is likely to be fulfilled only partially.

Although a cleavable signal peptide is not obligatory for GPI modification (Howell *et al.*, 1994) and alternative pathways of export from the cytoplasma exist, we checked the occurrence of N-terminal signal peptides among our hits with the SIGNALP algorithm (Nielsen *et al.*, 1999) using the combined *Y*score and *S*mean criterion. We found that 100% of the hits in *P.falciparum*, 61.0% of all predicted sequences (63.0% of those with $S > 0$) in *C.elegans* and 72.7% (80.5% of those with $S > 0$) among those from *S.cerevisiae* are predicted to have typical signal peptides. These numbers agree well with the sensitivity of SIGNALP over large test sets which was estimated near 70% (Nielsen *et al.*, 1997; Nielsen *et al.*, 1999). There is no good predictor for signal peptides of archaean species (Nielsen *et al.*, 1999); nevertheless, we predicted signal peptides for 33 (21) out of the 75 (46 with $S > 0$) hits for the four archaean species using SIGNALP which group together with eukaryotic genomes. In any case, the set of predicted archaean substrate proteins seems dramatically enriched with potential extracellular proteins compared with fractions of exported proteins encoded in eubacterial genomes estimated ~15% and in the genome of *M.jananaschii* predicted ~2% (34 out of 1715 proteins) by Nielsen *et al.* (Nielsen *et al.*, 1999) with SIGNALP.

II. Searches for subunits of the transamidase complex

Whereas the occurrence of many substrate proproteins is only an indirect hint, the detection of genes encoding subunits of the putative transamidase complex is a more direct proof for the existence of GPI post-translational modification in the given organism. Unfortunately, only two genes have been experimentally characterized as necessary for attachment of pre-synthesized complete GPI moieties to substrate proteins: (i) Gaa1/Gpaa1 in yeast (Hamburger *et al.*, 1995), mouse and human (Hiroi *et al.*, 1998; Inoue *et al.*, 1999; Ohishi *et al.*, 2000); and (ii) Gpi8 in yeast (Benghezal *et al.*, 1996; Meyer *et al.*, 2000), *Leishmania mexicana* (Hilley *et al.*, 2000), mouse and human (Yu *et al.*, 1997; Ohishi *et al.*, 2000). We have searched in sequence databases for orthologues with Blast/Psi-Blast (Altschul *et al.*, 1997) runs as well as with hidden Markov models (Durbin *et al.*, 1998).

Gpi8p-like proteins

Seven Gpi8p-like proteins from orthologous genes have been identified in one protozoan, two fungal, one plant, and three animal species: *L.mexicana* (accession no. CAB55340.1), *S.cerevisiae* (P49018), *Schizosaccharomyces pombe* (CAB57844.1), *Arabidopsis thaliana* (AC003981), *C.elegans* (P49048), *Drosophila melanogaster* (O46047), and human (Q92643, O14822). This group is delimited by a gap in the Blast *P* value ($< 10^{-42}$) from the following legumains, vacuole-processing enzymes, haemoglobinases and other cysteine proteases (*P* value $> 10^{-26}$). Additionally, ESTs and genomic fragments translating into protein segments with high levels of identity (>61%) to one of the Gpi8 proteins were found for many plants, for example *Lycopersicon esculentum* (AW032278, AW222470), *Gossypium arboreum* (AW68353) and *Medicago truncatula* (AW775802), the fungi *Neurospora crassa* (AI328293) and *Pisolithus tinctorius* (L38785), the nematode *Trichuris muris* (AW288376) and the alveolate *Cryptosporidium parvum* (AQ450372, identity 51%) as well as for a number of animal species.

All seven proteins appear to have a common structure (Table IIIa, Figure 3a, see Methods) as sequence analysis methods reveal. A signal peptide (sequence positions ca. 1–30, used for export in the endoplasmic reticulum) (Nielsen *et al.*, 1999) is followed by a C13-type Cys-endopeptidase domain (ca. 30–310) having a catalytic dyad His-Cys, another segment with unknown function consisting of about 30 residues, a single transmembrane region (Claros and von Heijne, 1994), and a cytoplasmic C-terminal tail. The borders of the protease domain and the catalytic residues have been refined using sequence similarities to legumains, vacuole-processing enzymes, gingipain R and haemoglobinases (Chen *et al.*, 1998; Eichinger *et al.*, 1999). Gpi8 has experimentally been shown to cleave the C-terminal propeptide of subtrate proteins (Ramalingam *et al.*, 1996; Sharma *et al.*, 1999; Hilley *et al.*, 2000; Meyer *et al.*, 2000; Ohishi *et al.*, 2000) even without a GPI anchor moiety being present.

Given the otherwise excellent sequence similarity, our sequence comparison results indicate that eukaryotic proteins obtained just as a result of DNA sequencing must be viewed with caution. The two sequences from *Caenorhabditis elegans* and *D.melanogaster* appear too short (compare Table IIIa). In both cases, the C13 peptidase domain is predicted being complete with the PFAM and the PFAM-FRAGMENT domain database tools. The C-terminal tail including the transmembrane domain might be missing (possibly, the last exon).

The leishmanian protein is also reported being smaller (~50 residues) than the remaining homologues from yeast to human (Hilley *et al.*, 2000), in the opinion of the authors, due to the missing C-terminal transmembrane region. Sensitive sequence analysis methods suggest a possibly alternative interpretation. The sequence may be without the C-terminal part of the endopeptidase domain in agreement with (i) the discrepancy of the C13 domain borders predicted by searching the PFAM and PFAM-FRAGMENTS domain databases (Bateman *et al.*, 2000); and (ii) the prediction of a C-terminal TM region with TOPPRED2 (Claros and von Heijne, 1994). In the light of the strict structural conservation of Gpi8p from fungi to human, the exception of the *Leishmania mexicana* sequence appears requiring further experimental studies.

Gaa1p/Gpaa1p-like proteins

Seven Gaa1/Gpaa1 orthologues have been found in fungi and

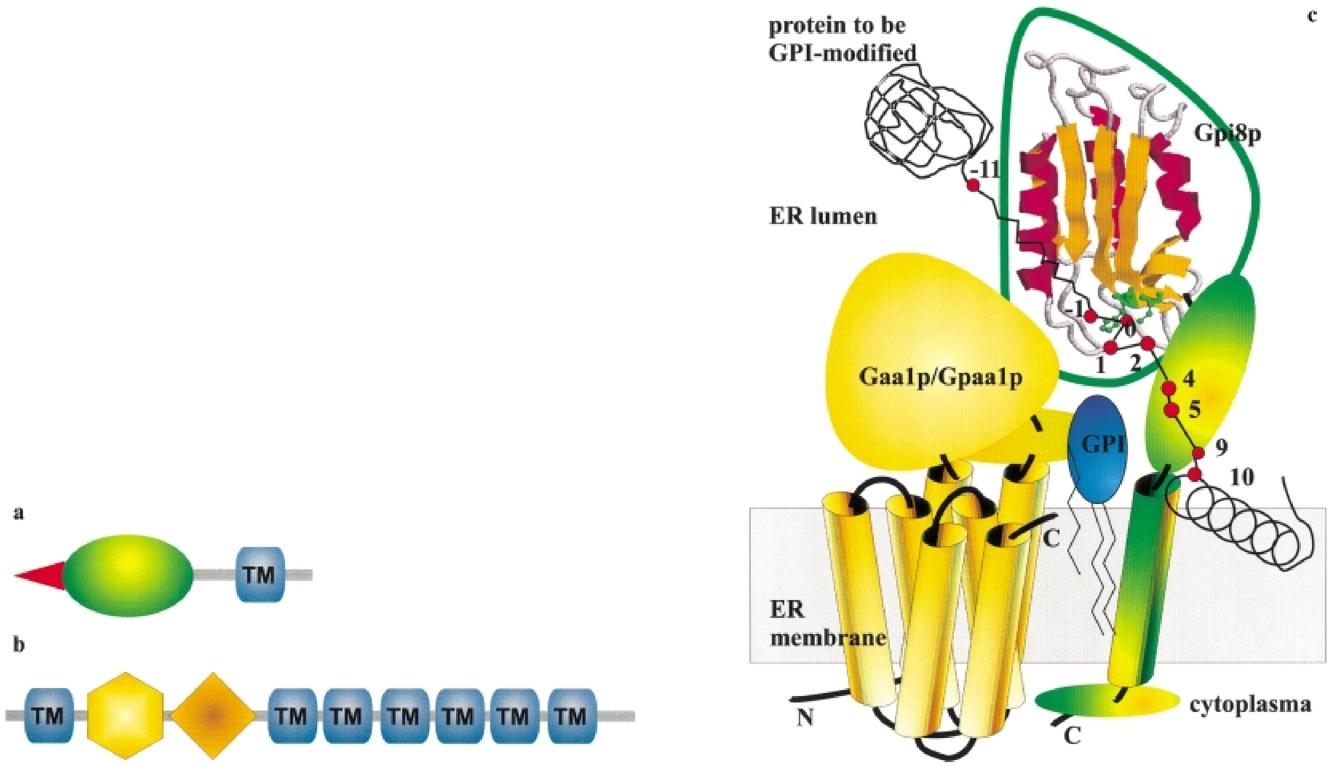

Fig. 3. The transamidase complex in eukaryotes. (**a**) Domain structure of Gpi8p. (**b**) Domain structure of Gaa1p/Gpaa1p. (**c**) The transamidase complex with a substrate protein. Schematically, the structures of the Gpi8p and the Gaa1p/Gpaa1p proteins and their interaction with a substrate protein are illustrated. After cleavage of the signal peptide, the transamidase Gpi8p (shown mainly in green, see Table IIIa) is located in the ER lumen but attached to its membrane by a single transmembrane region. The N-terminal endopeptidase domain is distantly related to caspases and gingipain R (Chen *et al*., 1998; Eichinger *et al*., 1999) in accordance to secondary structure predictions and sequence similarity. In this figure, elements of 1ibc (Rano *et al*., 1997) have been used to model the endopeptidase domain. Among all residues with functional side chains, only a histidine and a cysteine (ca. 40 residues apart) and a few small residues nearby are strictly conserved in an alignment with other C13 proteases. These two catalytic residues (shown as ball-and-stick in green) form the catalytic dyad (Chen *et al*., 1998; Sharma *et al*., 1999; Meyer *et al*., 2000). Two sequence segments on both sides of the TM region remain structurally and functionally uncharacterized. The C-terminus is located in the cytoplasma. The Gaa1p/Gpaa1p protein (shown in yellow) is a seven-transmembrane region protein. It seems exported to the ER without signal peptide cleavage. The N-terminus of the polypeptide chain is in the cytoplasma, the C-terminus is located in the ER lumen. Two or three globular domains located in the ER activate the (acylated) GPI anchor (shown in blue) and supply it to the Gpi8 protein and to the cleaved polypeptide. The substrate protein (shown as continuous black line with a few residues emphasized by red circles) is recognized by the transamidase alone. An unstructured region (ω–11 ... ω–1) connects the globule of the substrate protein with the catalytic cavity comprising residues ω–1 ... ω+2. Amino acid type preferences in the spacer region (ω+3 ... ω+9) at ω+4 and ω+5 indicate that these two residues may fit to a special binding site and play a role in substrate recognition. The mainly hydrophobic residues from ω+9 or ω+10 to the C-terminal appear to be bound by a hydrophobic pocket of Gpi8p and/or to dive into the membrane. Since many mammalian hydrophobic tails are preferentially composed of leucines, an α-helical transmembrane structure of the hydrophobic tail was originally anticipated. This idea has been criticised (Wang *et al*., 1999). Many recent examples of short hydrophobic tails (<20 residues) in lower animals, plants and fungi with large fractions of β-branched amino acid residues question this view, too (Eisenhaber *et al*., 1999).

animals: in *Saccharomyces cerevisiae* (P3012), *Schizosaccharomyces pombe* (CAB65611.1), *C.elegans (*AF039720), *D.melanogaster* (AAF46094.1), *Leishmania major* (CAB86709.1), mouse *(*BAA82589.1), and human (4504079). A fungal protein fragment (from *Mycosphaerella graminicola*, EST translation of AW180777) has high similarity to yeast Gaa1 (*P* value 10^{-28}). All seven proteins appear to have the same 7-transmembrane segment structure (Table IIIb, Figure 3b) as revealed by structural predictions and sequence homology considerations. In accordance with our SIGNALP (Nielsen *et al*., 1999) and TOPPRED2 (Claros and von Heijne, 1994) predictions, none of the five proteins has a signal peptide; thus, the N-terminal hydrophobic region (residues ca. 20–40) appears to form a single transmembrane region with the polypeptide N-terminus in the cytoplasma. TOPPRED2 and secondary structure (Frishman and Argos, 1997) predictions and homology considerations suggest that the C-terminal of Gaa1/Gpaa1 beginning with residue ca. 360 is just composed of six transmembrane regions and interconnecting loops. The sequence segment involving residues ca. 45–360 is without a low complexity region (Wootton, 1994) and probably comprises 2–3 globular domains located in the endoplasmic reticulum. In accordance with experimental evidence (Hamburger *et al*., 1995; Hiroi *et al*., 1998; Ohishi *et al*., 2000), this part of Gaa1/Gpaa1 appears to activate the GPI moiety for attachment to the cleaved polypeptide.

Again, the worm sequence being the predicted protein from a high-throughput genome sequencing effort appears inaccurate and the result of an in-silico fusion of two independent proteins. The sequence segment including the residues ca. 285–886 is a normal Gaa1/Gpaa1 protein. The N-terminal part (ca. 1–285) represents another protein having a DHHC zinc finger structure (sequence positions 75–139 identified with PFAM) and a coiled coil region [sequence positions 258–283 using COILS2 (Lupas, 1997)]. The cellular localization of such proteins is cytoplasmic or even nuclear (Mesilaty-Gross *et al*., 1999; Putilina *et al*., 1999). Also the *L.major* sequence appears incomplete at the N-terminal part of the globular segment, the

Table III. Orthologues of subunits of the putative transamidase complex

(a) Structure of Gpi8 proteins

Species	Accession no.	Sequence length	Signal peptide	C13-type Cys-protease	Catalytic His	Catalytic Cys	TM region
Homo sapiens	Q92643	396	1–29	30–329	165	207	367–387
*Caenorhabditis elegans**	P49048	322	1–28	29–320	159	201	Not found
*Drosophila melanogaster**	O46047	326	1–25	26–326	165	207	Not found
*Leishmania mexicana**	CAB55340.1	349	1–32	33–248	174	216	251–271
Schizosaccharomyces pombe	CAB57844.1	380	1–20	21–310	145	187	356–386
Saccharomyces cerevisiae	P49018	411	1–23	24–337	157	199	380–400
Arabidopsis thaliana	AC003981	428	1–20	21–329	150	192	345–365

(b) Structure of Gaa1/Gpaa1 proteins

Species	Accession no.	Sequence length	N-terminal TM region	Globular segment	Onset of 6 TM regions
Homo sapiens	gi4504079	621	22–42	45–365	368
Mus. musculus	BAA82589.1	621	22–42	45–365	368
*Caenorhabditis elegans**	AF039720	885	303–323	325–650	652
Drosophila melanogaster	AAF46094.1	674	26–46	47–390	391
*Leishmania major**	CAB86709.1	464	15–35	36–290	292
Schizo-Saccharomyces pombe	CAB65611.1	581	18–38	40–365	369
Saccharomyces cerevisiae	P39012	614	24–44	45–350	353

The domain structure of the proteins is described using the residue position numbers in the respective sequences. The prediction techniques used are listed in the Methods section. Entries labelled with '*' are possibly contaminated with sequence artefacts or differ from the general pattern due to species-specific particularities (see Results II).

latter region is only ~250 residues long (the remaining Gaa1/Gpaa1 proteins have ~320 residues) and it is homologous to the C-terminal of the Gaa1/Gpaa1 globular segment. This may not necessarily point to a sequencing error but can represent a species-specific particularity as in the case of Gpi8p of *L.mexicana* and possibly other protozoan parasites.

The available data on the putative transamidase complex in eukaryotes are summarized in Figure 3a–c which may be viewed as refinement of our model presented in Figure 2 of Eisenhaber *et al*. (Eisenhaber *et al*., 1998).

Searches for Gpi8p-like proteins in Archaea

Unfortunately, dedicated hidden Markov model searches in the non-redundant protein database using models extracted from Gaa1p/Gpaa1p, Gpi8p or just C13 endopeptidase multiple sequence alignments of various length did not result in significant archaean hits. Therefore, we decided to utilize the knowledge of the transamidase structure (suggesting the conservation of the catalytic mechanism) and the condition of completeness of the seven archaean genomes to narrow down the circle of candidate proteins that might constitute subunits for the transamidase complex (Figure 3c).

In the case of Gpi8p, we know from the multiple alignment of GPI modification transamidases, legumains, vacuole-processing enzymes and haemoglobinases that the residues of the catalytic dyad are surrounded by a few small and/or polar residues. This catalytic motif also known from the tertiary structures 1ibc (Rano *et al*., 1997) and 1cvr (Eichinger *et al*., 1999) (Figure 3c). Two four-residue motifs a_1HGa_4 (with a_1 and a_4 out of ADGSN) and $a_1Ca_3a_4$ (with a_3 among ADEGNQS) being 30–50 residues apart are observed in all sequences. A BIOMOTIF search (see Methods) with this apparently necessary (but surely not sufficient) cysteine protease description in a database consisting of all proteins from the seven complete archaean genomes failed to find any hit for the three genomes of *Methanococcus jannaschii*, *Pyrococcus abyssi*, *P.furiosus* (clustering with Eubacteria in Figure 2) but also for *P.horikoshii*. The motif is repeated in the hypothetical protein AF1758 (gi2648792) with 193 residues from *Archaeoglobus fulgidus*, in three proteins of *Aeropyrum pernix* (the hypothetical dipeptide transporter AP000059 (BAA79269.1) with 286 residues, the hypothetical proteins AP000060 (BAA79545.1) with 222 residues and AP000063 (BAA81133) with 156 residues and an aminotransferase (AE000897, AAB85815.1) of *Methanobacterium thermoautotrophicum*. Some of these proteins (notably AF1758 and AP000063) share other sequence features with transamidases such as predicted hydrophobic β-strands in front of both the histidine and the cysteine residues allowing the formation of a parallel β-sheet. In the case of AP000060, a single TM region (160–180) close to the C-terminus is predicted by TOPPRED (Claros and von Heijne, 1994). Besides these indicative suggestions, no strong evidence for Gpi8p-like proteins in archaean genomes could be detected. It can also not be excluded that non-orthologous proteins execute a catalytic activity similar to the one considered.

The existence of GPI modification in taxonomic groups and implications for the evolution of kingdoms

If the appearance of the putative transamidase complex for GPI-modification was a unique event in the evolution of species, at least among eukaryotes, and if the enzymes' specificity remained sufficiently conserved in biological time scales, then it should be possible to recognize a considerable number (possibly, not all) of proprotein sequences with potential GPI-anchoring in non-animal complete genomes even with prediction functions not parametrized for the given taxon. In this context, it is of interest to note that our prediction of GPI anchoring of the plant arabinogalactan proteins Q41071 and Q40380 (Eisenhaber *et al*., 1999) has been supported by recent experimental confirmation of GPI modification for a number of AGPs very recently (Sherrier *et al*., 1999; Oxley and Bacic, 1999).

As is visible in Figure 2, the genomes studied for the occurrence of GPI modification substrate proteins cluster in two groups with respect to the fraction of potentially GPI-modified proteins in the total genome: (group 1) eukaryotic genomes and the archaean genomes of *Archaeoglobus fulgidus*, *Aeropyrum pernix*, *M.thermoautotrophicum*, and *P.horikoshii*, (group 2) the eubacterial genomes and those of the archaea *Methanococcus jannaschii*, *P.abyssi*, and *P.furiosus*. We have also demonstrated with statistical criteria that this clustering is not incidental: the fractions of predicted substrate proteins can be explained with the rate of false positive predictions for all genomes of the second group whereas the frequency of hits is 2–4 higher for the genomes in group 1.

The GPI modification device seems missing in the case of the second group of genomes; i.e. the few predictions for substrate proteins appear (i) false positives or (ii) to belong to proteins having incidentally the GPI modification motif which is not exploited in the given cellular context. Additionally, (iii) cases of gene transfer from organisms with GPI modification can also not be excluded. This general result was not obvious since phosphatidylinositol lipid anchors are common in many Eubacteria. There are even lipoarabinomannans (LAMs) on the plasma membrane of mycobacteria (such as *Mycobacterium tuberculosis*) resembling GPI-anchored proteins since they represent complex, multiply branched carbohydrate polymers terminated by a phosphatidylinositol lipid anchor (Brennan and Nikaido, 1995; Ilangumaran *et al*., 1995). We agree that it appears reasonable to search for other biological functions for (glycosyl-) phosphatidylinositol moieties and to consider GPI anchoring as just one evolutionary possibility of their use (Sevlever *et al*., 1999).

Having a large number of substrate proteins in their genome, organisms of the first group are supposed to have a protein GPI-modification machinery with substrate specificity similar to those of animal enzymes and, possibly, with common evolutionary origin. We can conclude from the sequence analysis of Gaa1/Gpaa1 and Gpi8 proteins that the molecular apparatus for GPI modification is sequentially and structurally highly conserved among lower and higher eukaryotes (Figure 3c). GPI anchoring appears an ubiquitously used mechanism for tethering proteins to the non-cytoplasmic side of membranes in eukaryotes even in species where this pathway has not yet been experimentally studied.

Surprisingly, our data suggest that the Archaea are heterogeneous with respect to the ability of GPI modification of proteins. Whereas three species appear not to have a typical GPI lipid anchoring mechanism due to the vanishingly small number of substrate proproteins, we predict that four out of seven archaean genomes studied contain 2–4 times more potential substrate proteins for GPI modification than can be explained by the expected rate of false positive predictions. Although the C-terminal signal found by us in a number of archaean protein sequences might be used in a yet unknown cellular context, the experimental report (Kobayashi *et al*., 1997) of a GPI-modified protein in the archaeobacterium *Sulfolobus acidocaldarius* (although relying only on the PI-PLC test) suggests that the existence of a GPI anchor protein modification machinery in the four species is more likely. It should also be noted that several high-scoring proprotein hits for other, not yet completely sequenced archaean species such as the halobacterial surface protein CSG_HALHA (P08198, label D) have been found searching SWISS-PROT/SP-TrEMBL (Eisenhaber *et al*., 1999). Although we could not assign transamidase-like enzymatic activity to gene products in the four archaean genomes with high statistical significance, it should be noted that, even in much better studied cases such as *E.coli*, the assignment of 10 catalytic activities among the ~150 necessary for amino acid synthesis known to exist in that bacterium was not possible despite knowledge of the full genome (Selkov *et al*., 2000).

The available data might be not considered completely sufficient to finally state the existence of GPI anchoring within an archaean subgroup but this hypothesis is consistent with the available experimental and sequence analysis data. Our findings also indicate that the evolutionary history of kingdoms of life may not be reduced to the question whether Eukaryota or Eubacteria are closer to Archaea. The results support the view that only a subgroup of the Archaea has a common ancestor with the eukaryotic branch.

Acknowledgements

We are grateful to Göran Bucht, Mike A. Ferguson, Frans M. Klis, and Jens G. Reich for valuable advice and encouragement. Birgit Eisenhaber and Frank Eisenhaber thank Boehringer–Ingelheim International for support.

References

Altschul,S., Boguski,M., Gish,W. and Wootton,J.C. (1994) *Nature Genet*., **6**, 119–129.

Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res*., **25**, 3389–3402.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) *Nucleic Acids Res.*, **28**, 263–266.

Benghezal,M., Benachour,A., Rusconi,S., Aebi,M. and Conzelmann,A. (1996) *EMBO J*., **15**, 6575–6583.

Brennan,P.J. and Nikaido,H. (1995) *Annu. Rev. Biochem*., **64**, 29–63.

Caro,L.H.P., Tettelin,H., Vossen,J.H., Ram,A.F.J., van den Ende,H. and Klis,F.M. (1997) *Yeast*, **13**, 1477–1489.

Chen,J.-M., Rawlings,N.D., Stevens,R.A.E. and Barrett,A.J. (1998) *FEBS Lett*., **441**, 361–365.

Claros,M.G. and von Heijne,G. (1994) *Comput. Appl. Biosci*., **10**, 685–686.

Descoteaux,A. and Turco,S.J. (1999) *Biochim. Biophys. Acta*, **1455**, 341–352.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchinson,G. (1998) *Biological Sequence Analysis*, 1st edn. Cambridge, Cambridge University Press.

Eichinger,A., Beisel,H.-G., Jacob,U., Huber,R., Medrano,F.-J., Banbula,A., Potempa,J., Travis,J. and Bode,W. (1999) *EMBO J*., **18**, 5453–5462.

Eisenhaber,B., Bork,P. and Eisenhaber,F. (1998) *Protein Eng*., **11**, 1155–1161.

Eisenhaber,B., Bork,P. and Eisenhaber,F. (1999) *J. Mol. Biol*., **292**, 741–758.

Eisenhaber,B., Bork,P., Yuanping,Y., Löffler,G. and Eisenhaber,F. (2000) *Trends Biochem. Sci*., **25**, 340–341.

Ferguson,M.A. and Williams,A.F. (1988) *Annu. Rev. Biochem*., **57**, 285–320.

Frishman,D. and Argos,P. (1997) *Proteins*, **27**, 329–335.

Guadiz,G., Haidaris,C.G., Maine,G.N. and Simpson-Haidaris,P.J. (1998) *J. Biol. Chem*., **273**, 26202–26209.

Hamburger,D., Egerton,M. and Riezman,H. (1995) *J. Cell. Biol*., **129**, 629–639.

Hilley,J.D., Zawadzki,J.L., McConville,M.J., Coombs,G.H. and Mottram,J.C. (2000) *Mol. Biol. Cell,* **11**, 1183–1195.

Hiroi,Y., Komuro,I., Chen,R., Hosoda,T., Mizuno,T., Kudoh,S., Georgescu,S.P., Medof,M.E. and Yazaki,Y. (1998) *FEBS Lett*., **421**, 252–258.

Howell,S., Lanctôt,C., Boileau,G. and Crine,P. (1994) *J. Biol. Chem*., **269**, 16993–16996.

Ilangumaran,S., Arni,S., Poincelet,M., Theler,J.-M., Brennan,P.J., ud-Din,N. and Hoessli,D.C. (1995) *J. Immunol*., **155**, 1334–1342.

Inoue,N., Ohishi,K., Endo,Y., Fujita,T., Takeda,J. and Kinoshita,T. (1999) *Cytogenet. Cell Genet*., **84**, 199–205.

Kendall,M. and Stuart,A. (1977) *The Advanced Theory of Statistics*, 1st edn. Griffen, London.

Kobayashi,T., Nishizaki,R. and Ikezawa,H. (1997) *Biochim. Biophys. Acta*, **1334**, 1–4.

Lupas,A. (1997) *Curr. Opin. Struct. Biol*., **7**, 388–393.

Mesilaty-Gross,S., Reich,A., Motro,B. and Wides,R. (1999) *Gene,* **231**, 173–186.

Meyer,U., Benghezal,M., Imhof,I. and Conzelmann,A. (2000) *Biochemistry*, **39**, 3461–3471.

Micanovic,R., Gerber,L.D., Berger,J., Kodukula,K. and Udenfriend,S. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 157–161.

Moran,P. and Caras,I.W. (1994) *J. Cell. Biol.*, **125**, 333–343.

Morita,N., Nakazato,H., Okuyama,H., Kim,Y. and Thompson,G.A.,Jr (1996) *Biochim. Biophys. Acta*, **1290**, 53–62.

Nielsen,H., Brunak,S. and von Heijne,G. (1999) *Protein Eng.*, **12**, 3–9.

Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) *Protein Eng.*, **10**, 1–6.

Nuoffer,C., Horvath,A. and Riezmann,H. (1993) *J. Biol. Chem.*, **268**, 10558–10563.

Ohishi,K., Inoue,N., Maeda,Y., Takeda,J., Riezman,H. and Kinoshita,T. (2000) *Mol. Biol. Cell*, **11**, 1523–1533.

Oxley,D. and Bacic,A. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 14246–14251.

Popolo,L. and Vai,M. (1999) *Biochim. Biophys. Acta*, **1426**, 385–400.

Putilina,T., Wong,P. and Gentleman,S. (1999) *Mol. Cell. Biochem.*, **195**, 219–226.

Ramalingam,S., Maxwell,S., Medof,M.E., Chen,R., Gerber,L.D. and Udenfriend,S. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 7528–7533.

Rano,T.A., Timkey,T., Peterson,E.P., Rotonda,J., Nicholson,D.W., Becker,J.W., Chapman,K.T. and Thornberry,N.A. (1997) *Chem. Biol.*, **4**, 149–155.

Selkov,E., Overbeek,R., Kogan,Y., Chu,L., Vonstein,V., Holmes,D., Silver,S., Haselkorn,R. and Fonstein,M. (2000) *Proc. Natl Acad. Sci. USA*, **97**, 3509–3514.

Sevlever,D., Pickett,S., Mann,K.J., Sambamurti,K., Medof,M.E. and Rosenberry,T.L. (1999) *Biochem. J.*, **343**, 627–635.

Sharma,D.K., Vidugiriene,J., Bangs,J.D. and Menon,A.K. (1999) *J. Biol. Chem.*, **274**, 16479–16486.

Sherrier,D.J., Prime,T.A. and Dupree,P. (1999) *Electrophoresis*, **20**, 2027–2035.

Smith,T.K., Sharma,D.K., Crossman,A., Dix,A., Brimacombe,J.S. and Ferguson,M.A.J. (1997) *EMBO J.*, **16**, 6667–6675.

Sunyaev,S.R., Eisenhaber,F., Rodchenkov,I.V., Eisenhaber,B., Tumanyan,V.G. and Kuznetsov,E.N. (1999) *Protein Eng.*, **12**, 387–394.

Takos,A.M., Dry,I.B. and Soole,K.L. (1997) *FEBS Lett.*, **405**, 1–4.

Takos,A.M., Dry,I.B. and Soole,K.L. (2000) *Plant J.*, **21**, 43–52.

Tatusova,T.A., Karsch-Mizrachi,I. and Ostell,J. (1999) *Bioinformatics*, **15**, 536–543.

Udenfriend,S. and Kodukula,K. (1995) *Annu. Rev. Biochem.*, **64**, 563–591.

Vai,M., Lacanà,E., Gatti,E., Breviario,D., Popolo,L. and Alberghina,L. (1993) *Curr. Genet.*, **23**, 19–21.

Wang,J., Maziarz,K. and Ratnam,M. (1999) *J. Mol. Biol.*, **286**, 1303–1310.

Wootton,J.C. (1994) *Curr. Opin. Struct. Biol.*, **4**, 413–421.

Youl,J.J., Bacic,A. and Oxley,D. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 7921–7926.

Yu,J., Nagarajan,S., Knez,J.J., Udenfriend,S., Chen,R. and Medof,M.E. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 12580–12585.

Zhou,J., Dutch,R.E. and Lamb,R.A. (1997) *Virology*, **239**, 327–339.

Received June 16, 2000; accepted October 13, 2000