Nucleic Acids Research, 2003, Vol. 31, No. 13 **3631–3634**
DOI: 10.1093/nar/gkg537

Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-Π, NMT and PTS1

Frank Eisenhaber*, Birgit Eisenhaber, Werner Kubina, Sebastian Maurer-Stroh, Georg Neuberger, Georg Schneider and Michael Wildpaner

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

Received February 11, 2003; Revised and Accepted March 27, 2003

ABSTRACT

Many posttranslational modifications (N-myristoylation or glycosylphosphatidylinositol (GPI) lipid anchoring) and localization signals (the peroxisomal targeting signal PTS1) are encoded in short, partly compositionally biased regions at the N- or C-terminus of the protein sequence. These sequence signals are not well defined in terms of amino acid type preferences but they have significant interpositional correlations. Although the number of verified protein examples is small, the quantification of several physical conditions necessary for productive protein binding with the enzyme complexes executing the respective transformations can lead to predictors that recognize the signals from the amino acid sequence of queries alone. Taxon-specific prediction functions are required due to the divergent evolution of the active complexes. The big-Π tool for the prediction of the C-terminal signal for GPI lipid anchor attachment is available for metazoan, protozoan and plant sequences. The myristoyl transferase (NMT) predictor recognizes glycine N-myristoylation sites (at the N-terminus and for fragments after processing) of higher eukaryotes (including their viruses) and fungi. The PTS1 signal predictor finds proteins with a C-terminus appropriate for peroxisomal import (for metazoa and fungi). Guidelines for application of the three WWW-based predictors (http://mendel.imp.univie.ac.at/) and for the interpretation of their output are described.

INTRODUCTION

For researchers who want to analyze the occurrence of a potential PTS1 signal, of a putative GPI lipid anchor attachment or myristoylation sites in their target protein sequences, this text provides application and output interpretation guidelines for the WWW-servers big-Π, NMT and PTS1. The methodology behind those tools and their validation is described in great detail elsewhere (Table 1) except for the new big-Π plant predictor (B. Eisenhaber, M. Wildpaner, C.J. Schultz, G.H.H. Borner, P. Dupree and F. Eisenhaber, manuscript submitted). In the following, we summarize aspects that are important from the user's point of view.

A number of sequence motifs at the termini of proteins encode signals for targeting to cellular compartments and for posttranslational modifications. The N-terminal signal peptide responsible for export into the ER is the most well known, the mitochondrial and the chloroplast signals are also N-terminally located. In contrast, the peroxisomal targeting signal PTS1 is C-terminal. Many posttranslational modifications are attached N-terminally (N-myristoylation) or C-terminally (GPI lipid anchors, farnesylation, geranylgeranylation), to name just a few (1).

Despite the functional importance of these sequence signals, the theoretical methods for their prediction from the sequence of query proteins has received less general attention than those for studying globular domains. With the concept of homology, the assumption of a common ancestor originating a family of sequentially similar sequences in an evolutionary process involving gene duplications and mutations, function can be assigned to globular domains (having a typical length of 100–150 amino acids) by annotation transfer from experimentally studied sequence family members (2). Unfortunately, the signals for subcellular targeting and posttranslational modification are located in relatively short (<40 amino acids), non-globular regions with typical amino acid compositional bias and interpositional correlations. Therefore, the measures for quantifying remote sequence similarity cannot be directly applied for family classification of these signals.

Even in the absence of knowledge of the active complex responsible for translocation or modification of the substrate protein, the sequence requirements for productive binding with the active protein complex can be derived from the variability of sequences of experimentally verified substrate protein sequences. If the learning set is large, procedures of unsupervised, automated learning successfully extract complex

*To whom correspondence should be addressed. Tel: +43 179730557; Fax: +43 17987153; Email: frank.eisenhaber@imp.univie.ac.at

 © *Oxford University Press 2003; all rights reserved*

Table 1. Big-Π, NMT, PTS1: web URL, taxonomic range and prediction accuracy

WWW-server, URL and taxonomic range	Sensitivity (%)	False-positive prediction rate (%)
Big-Π (8,10,16)		
http://mendel.imp.univie.ac.at/gpi/gpi_server.html (metazoa/protozoa)	80	0.2
http://mendel.imp.univie.ac.at/gpi/plant_server.html (plants)	94	0.1
NMT (7,9)		
http://mendel.imp.univie.ac.at/myristate/ (non-fungal eukarya + viruses/fungi)	>95	0.5
PTS1 (19,20)		
http://mendel.imp.univie.ac.at/myristate/ (fungi, metazoa, other eukarya)	>90	0.2–0.8

All servers request a single fasta-formatted sequence and a taxon selection as input. Only for taxonomic ranges where the amount of learning data (especially the number of non-redundant substrate sequences) is sufficient, predictors have been developed as indicated in this table. Details of the prediction functions and their validation procedures are described in detail in the references indicated. Scores calculated from different predictors are not directly comparable. In a practical application, it is also not clear from the score alone how large is the risk of wrongly accepting the prediction as correct. Therefore, the score is translated into probabilities of false-positive prediction calculated for the queries analyzed (9,16,17) following the BLAST E-value style (18). Two values characterize the accuracy of the corresponding predictors. Sensitivity or coverage is the probability to predict the biological property for a true target (with a threshold score = 0). The false-positive prediction rate (which complements the specificity to 100%) assesses the probability that the biological property is assigned to a random, non-related protein sequence (with a threshold score = 0).

sequence patterns [for example, in the case of SIGNALP (3), the current standard for signal peptide prediction]. The same methodology is considerably less powerful if the learning set is an order of magnitude smaller and less reliable as for the mitochondrial or chloroplast targeting signal (4,5), especially for rejecting false-positive predictions.

If the sequence motif in the substrate protein is considered from the view point of productive binding with the active complex, simple physical conditions for the rejection of non-permissive query sequences can be formulated (6,7). Typically, a core of the sequence motif with several positions of amino acid type conservation is necessary for binding in the active site of the modifying enzyme or the recognition site of the translocator. Conformational flexibility in the motif region is required to adapt to the catalytic cleft. The sequence environment of the core has to provide accessibility of the sequence signal, mechanical linkage to the remainder of the substrate protein and appropriate interaction with the aqueous or membrane surrounding of the active complex. A combined score function with profile terms (for evaluating amino acid type preferences) and physical property terms (with only non-positive scores for rejecting unsuitable queries) can successfully discriminate queries even in cases of single-residue mutations that affect modification efficiency (1,8,9).

BIG-Π: PREDICTION OF THE C-TERMINAL GPI LIPID ANCHOR MOTIF IN METAZOAN, PROTOZOAN AND PLANT SEQUENCES

Posttranslational modification with a GPI lipid anchor consists of two reactions executed by the transamidase complex in the endoplasmic reticulum, the attachment of the GPI moiety to the carboxyl terminus (ω-site) of the polypeptide after proteolytic cleavage of a C-terminal propeptide. Typically, a GPI lipid anchored protein is finally moved to the extracellular side of the cytomembrane via vesicular transport. The classical sequence pattern consists of four regions defined by the preferred pattern of physical properties of amino acid side chains (6,10). (i) The region $\omega-11\ldots\omega-1$ is a flexible, polar linker. This stretch has been hypothesized to occupy a channel in the transamidase complex. In the structural model of the transamidase (11), access to the active site cleft of the cysteine protease PIG-K/gpi8 is regulated by the endoplasmic lumenal domain of PIG-T, a β-propeller structure with a central hole. (ii) The region $\omega-1\ldots\omega+2$ has volume constraints and is occupied preferentially by small residues. (iii) The spacer region $\omega+3\ldots\omega+9$ is composed of moderately polar residues. (iv) The typical hydrophobic tail begins with $\omega+9$ or $\omega+10$ and extends up to the C-terminal end.

The big-Π tool (Table 1) evaluates the concordance of a query with this sequence motif. In the output, the primary and, if available, the secondary ω-sites are reported. Together with their sequence position, the prediction quality [strong prediction or twilight zone (8)], the score and the probability of false positive prediction are presented. In the case of sequences without GPI lipid anchor motif, the nevertheless best site is listed. In either case, a detailed description of score components is shown that allows the evaluation of the agreement with amino acid type profile and with physical pattern properties and, especially, to analyze reasons for negative predictions. Therefore, the big-Π predictor is well suited for designing mutations aimed at abolishing GPI lipid anchoring capacity. For example, modified query sequences where the putative site is substituted by a residue with large side chain or with more immobile backbone can be tested prior to the experiment.

A positive prediction by big-Π does not necessarily mean capacity for GPI lipid anchoring *in vivo*. Big-Π assesses only the concordance of the C-terminus with the GPI lipid anchor modification motif. In the evaluation of the prediction outcome, the issue of ER export signal should receive special independent attention. One can routinely check for signal leaders (3) but alternative export signals [see, for example (12)] should also be taken into account.

Further, the function parameterization relies on the small set of known GPI lipid anchor modified proteins. Thus, a largely negative physical property term ('profile independent score') can be considered a sure sign for the absence of the GPI anchor motif because only a handful of very stably derived parameters enter this term (8). In contrast, a small profile score can also be a result of the still limited learning set with biased amino acid

type preferences and, consequently, an insufficiently general profile matrix.

NMT: PREDICTION OF N-MYRISTOYLATION OF N-TERMINAL GLYCINES FOR HIGHER EUKARYOTE, VIRAL AND FUNGAL QUERY SEQUENCES

N-terminal N-myristoylation is the attachment of a myristoyl anchor to an N-terminal glycine by a myristoyltransferase (NMT) for modulation of interaction of the modified protein with intracellular membranes or with other proteins. At least the N-terminal 17 residues of the substrate protein experience amino acid type variability restrictions for N-myristoylation (7). Positions 1–6 with glycine in the leading position fit the binding pocket of the NMT, positions 7–10 interact non-specifically with the NMT's surface at the mouth of the catalytic cavity, and positions 11–17 form a hydrophilic linker. Thus, in addition to the segment physically interacting with the NMT, 10–11 more residues in a linker region experience weaker sequence variability restrictions and contribute to the recognition motif.

The NMT predictor (Table 1) scores the agreement of a query N-terminus with the N-myristoylation pattern and returns the corresponding probability of false-positive prediction (Fig. 1). We distinguish reliably predicted targets (score ≥ 0), twilight zone predictions ($0 >$ score ≥ -2), and proteins that are predicted not to be NMT targets. It should also be noted that, for example in the case of viral polyproteins, internal glycines become N-terminal after protein processing and are myristoylated. Optionally, possible myristoylation at internal glycines (in typical processing patterns) may be analyzed, too (9).

The N-myristoylation signal is commonly applied to target proteins to membranes. The NMT predictor can be used for testing protein constructs with engineered N-terminal N-myristoylation motif prior to the experiment. With the complete output of score components, the agreement with the physical property pattern can be checked in detail (for example, the suitability of the linker region) and it becomes easy to examine the effects of changes in the construct.

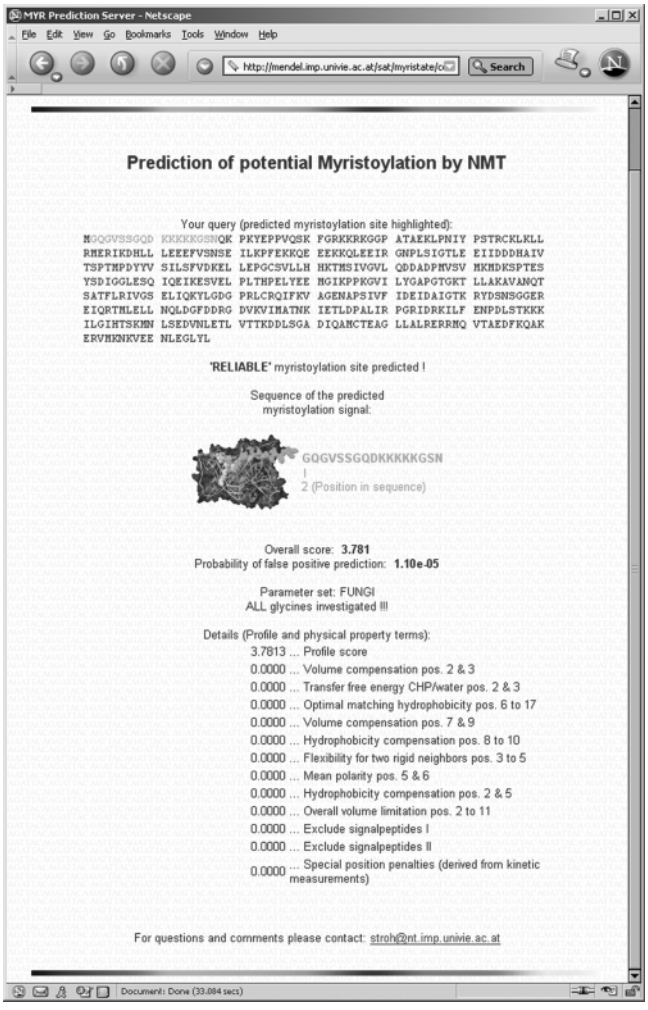

Figure 1. Example output of the NMT predictor. The information generated upon sequence submission is similarly structured for all three servers. As example, the server output is presented for the yeast 26 S protease regulatory subunit 4 homologue (RPT2, SWISS-PROT accession P40327). For control purposes, the complete sequence is returned first with the examined motif highlighted. After the general classification of the prediction (reliable, twilight zone, not predicted) and the overall score and probability of false positive prediction, the components of the score function are listed. In this case, no deviation from the physical property pattern was measured although the protein is not part of the learning set. N-myristoylation of the RPT2 protein has been predicted (9) and the experimental verification reported (15).

PTS1: PREDICTION OF THE PTS1 PEROXISOMAL IMPORT SIGNAL FOR HIGHER EUKARYOTES AND FUNGI

To date, two different signals that can trigger peroxisomal import have been characterized, termed PTS1 and PTS2. PTS1, the major targeting signal, consists of the three C-terminal amino acids (mainly the canonical tripeptide S/A/C-K/R/H-L, but not only) that bind to the inner cavity of the receptor molecule Pex5 in addition to several residues further upstream (13) that either interact with the surface of Pex5 or serve as a short conformationally unrestricted linker to the remainder of the protein.

The concordance with this motif is searched for using an algorithm implemented in the PTS1 signal server (Table 1). Reliably predicted targets should have a non-negative total score; queries with a score larger than -10 are considered as twilight zone hits. In all other cases, the protein is predicted not to have a PTS1 signal. We must emphasize that the server analyzes exclusively the concordance of the query's C-terminus with the generalized PTS1 motif as described above. The PTS1 signal competes with other signals if contained in the sequence. Proteins with dual localizations, including for example a peroxisomal and a mitochondrial fraction (14), are known; proteins with a strong signal peptide are most probably co-translationally exported to the ER.

ACKNOWLEDGEMENTS

The authors are grateful for generous support from Boehringer Ingelheim. This project has been partly funded by the Austrian National Bank (Österreichische Nationalbank), by the Fonds

zur Förderung der wissenschaftlichen Forschung Österreichs (FWF P15037) and by the Austrian Gen-AU bioinformatics integration network sponsored by BM-BWK.

REFERENCES

1. Eisenhaber,F., Eisenhaber,B. and Maurer-Stroh,S. (2003) Prediction of post-translational modifications from amino acid sequence: problems, pitfalls, methodological hints. In Andrade,M.M. (ed.), *Bioinformatics and Genomes: Current Perspectives.* Horizon Scientific Press, Wymondham, pp. 81–105.
2. Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.*, **283**, 707–725.
3. Nielsen,H., Brunak,S. and von Heijne,G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.*, **12**, 3–9.
4. Emanuelsson,O., Nielsen,H. and von Heijne,G. (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.*, **8**, 978–984.
5. Emanuelsson,O., von Heijne,G. and Schneider,G. (2001) Analysis and prediction of mitochondrial targeting peptides. *Methods Cell Biol.*, **65**, 175–187.
6. Eisenhaber,B., Bork,P. and Eisenhaber,F. (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.
7. Maurer-Stroh,S., Eisenhaber,B. and Eisenhaber,F. (2002) N-terminal N-myristoylation of proteins: refinement of the sequence motif and its taxon-specific differences. *J. Mol. Biol.*, **317**, 523–540.
8. Eisenhaber,B., Bork,P. and Eisenhaber,F. (1999) Prediction of potential GPI-modification sites in proprotein sequences. *J. Mol. Biol.*, **292**, 741–758.
9. Maurer-Stroh,S., Eisenhaber,B. and Eisenhaber,F. (2002) N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. *J. Mol. Biol.*, **317**, 541–557.
10. Eisenhaber,B., Bork,P. and Eisenhaber,F. (2001) Post-translational GPI lipid anchor modification of proteins in kingdoms of life: analysis of protein sequence data from complete genomes. *Protein Eng.*, **14**, 17–25.
11. Eisenhaber,B., Maurer-Stroh,S., Novatchkova,M., Schneider,G. and Eisenhaber,F. (2003) Enzymes and auxiliary factors for GPI lipid anchor biosynthesis and post-translational transfer to proteins. *Bioessays*, **25**, 367–385.
12. Denny,P.W., Gokool,S., Russell,D.G., Field,M.C. and Smith,D.F. (2000) Acylation-dependent protein export in Leishmania. *J. Biol. Chem.*, **275**, 11017–11025.
13. Lametschwandtner,G., Brocard,C., Fransen,M., Van Veldhoven,P., Berger,J. and Hartig,A. (1998) The difference in recognition of terminal tripeptides as peroxisomal targeting signal 1 between yeast and human is due to different affinities of their receptor Pex5p to the cognate signal and to residues adjacent to it. *J. Biol. Chem.*, **273**, 33635–33643.
14. Holbrook,J.D., Birdsey,G.M., Yang,Z., Bruford,M.W. and Danpure,C.J. (2000) Molecular adaptation of alanine : glyoxylate aminotransferase targeting in primates. *Mol. Biol. Evol.*, **17**, 387–400.
15. Kimura,Y., Saeki,Y., Yokosawa,H., Polevoda,B., Sherman,F. and Hirano,H. (2003) N-terminal modifications of the 19S regulatory particle subunits of the yeast proteasome. *Arch. Biochem. Biophys.*, **409**, 341–348.
16. Eisenhaber,B., Bork,P., Yuan,Y., Loffler,G. and Eisenhaber,F. (2000) Automated annotation of GPI anchor sites: case study *C. elegans*. *Trends Biochem. Sci.*, **25**, 340–341.
17. Cserzo,M., Eisenhaber,F., Eisenhaber,B. and Simon,I. (2002) On filtering false positive transmembrane protein predictions. *Protein Eng.*, **15**, 745–752.
18. Altschul,S.F., Boguski,M.S., Gish,W. and Wootton,J.C. (1994) Issues in searching molecular sequence databases. *Nature Genet.*, **6**, 119–129.
19. Neuberger,G., Maurer-Stroh,S., Eisenhaber,B., Hartig,A. and Eisenhaber,F. (2003) Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. *J. Mol. Biol.*, **328**, 567–579.
20. Neuberger,G., Maurer-Stroh,S., Eisenhaber,B., Hartig,A. and Eisenhaber,F. (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol.*, **328**, 581–592.