doi:10.1016/S0022-2836(03)00553-9 J. Mol. Biol. (2003) 330, 443–456

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

In Silico Prediction of the Peroxisomal Proteome in Fungi, Plants and Animals

Olof Emanuelsson[1], Arne Elofsson[1], Gunnar von Heijne[1]* and Susana Cristóbal[2]

[1]*Stockholm Bioinformatics Center, AlbaNova University Center, Department of Biochemistry and Biophysics Stockholm University, S-106 91 Stockholm, Sweden*

[2]*Department of Cell and Molecular Biology, Biomedical Center, Uppsala University Box 596, SE-751 24 Uppsala Sweden*

**Corresponding author*

In an attempt to improve our abilities to predict peroxisomal proteins, we have combined machine-learning techniques for analyzing peroxisomal targeting signals (PTS1) with domain-based cross-species comparisons between eight eukaryotic genomes. Our results indicate that this combined approach has a significantly higher specificity than earlier attempts to predict peroxisomal localization, without a loss in sensitivity. This allowed us to predict 430 peroxisomal proteins that almost completely lack a localization annotation. These proteins can be grouped into 29 families covering most of the known steps in all known peroxisomal pathways. In general, plants have the highest number of predicted peroxisomal proteins, and fungi the smallest number.

© 2003 Elsevier Science Ltd. All rights reserved

Keywords: peroxisome; proteome; prediction; protein sorting; subcellular location

Introduction

Peroxisomes, along with glyoxysomes of plants and glycosomes of trypanosomes, belong to the microbody family of organelles. These three types of microbodies exist in different cellular environments and possess distinct specialized functions. They house an important set of enzymes within their single membrane, and at the very least they all contain one hydrogen-peroxide-producing oxidase and a catalase to decompose the hydrogen peroxide.[1] Peroxisomes contain enzymes involved in lipid metabolism, such as β-oxidation of fatty acids, synthesis of cholesterol, bile acids and plasmalogens in mammals, in the glyoxylate cycle in plants, in methanol oxidation in yeasts,[2,3] and part of the glycolytic pathway in kinetoplastid parasites.[4]

The importance of the peroxisome is underscored by the existence of numerous human genetic disorders associated with peroxisomal defects. Lack of single peroxisomal enzymes is the cause for several human diseases.[5,6] However, the most severe peroxisomal disorders originate from defects in peroxisome biogenesis, with the simultaneous loss of several metabolic functions. These disorders, known as the peroxisomal biogenesis disorders (PBDs), such as Zellweger syndrome, are genetically heterogeneous with 12 known complementation groups.[7]

The biogenesis of peroxisomal matrix proteins is fairly well understood.[8–10] Both the protein targeting and import mechanism into microbodies and the components required for peroxisomal biogenesis are evolutionarily conserved.[11,12] Peroxisomal proteins are nuclear-encoded and synthesized in the cytosol on free polyribosomes.[1] Peroxisomes acquire their matrix proteins by post-translational import from the cytosol *via* two pathways that rely on two kinds of conserved peroxisomal targeting signals (PTS). The majority of peroxisomal matrix proteins have a PTS1 at their extreme carboxyl terminus, consisting of just three amino acids—SKL—or a conservative variant thereof.[8,11] A few peroxisomal enzymes (malate dehydrogenase, citrate synthase, acyl-CoA oxidase and 3-ketoacyl-CoA-thiolase) are known to use a different targeting signal, the amino-terminally located PTS2, which is a bipartite signal with the consensus sequence [RK]-[LVI]-x5-[HQ]-[LA].[13]

Although most peroxisomal matrix proteins use PTSs for their targeting, there are a few proteins that lack a canonical targeting signal and that might enter the peroxisomal matrix by "piggybacking" on other proteins bearing PTSs.[14,15]

Abbreviations used: PBD, peroxisomal biogenesis disorders; PTS, peroxisomal targeting signal; TPR, tetratricopeptide repeats; MCC, Matthews correlation coefficient.

E-mail address of the corresponding author: gunnar@dbb.su.se

0022-2836/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved

The peroxisomal protein import machinery requires around 20 PEX genes and their products, the peroxins.[8] PTS1 interacts with the tetratricopeptide repeats (TPRs) of the receptor Pex5p.[16] Proteins bearing PTS2 bind to the WD40 motifs of Pex7p.[17] After binding of their cargo proteins, these receptors are thought to interact with a docking complex consisting of Pex13p and Pex14p, where both pathways seem to merge[18,19] and then shuttle back into the cytosol for the next round of targeting. At present, several peroxins involved in the protein import machinery have been characterized, however, little is known about the principles of the translocation process.[10]

Considering the biological and medical importance of the peroxisome, methods for identifying peroxisomal proteins from the amino acid sequence is an important challenge in bioinformatics. Previous attempts to predict peroxisomal localization based on amino acid sequence include PSORT,[20,21] a knowledge-based predictor using a decision tree to sort proteins among several different compartments. In PSORT, the PTS1 motif [AS]-[HKR]-L is used as a marker for peroxisomal location along with amino acid composition over the entire protein. The performance on peroxisomal proteins is modest, in the sense that many peroxisomal proteins are missed. Cai *et al*.[22] applied a support vector machine (SVM) to predict protein localization based on both amino acid composition and sequence. Though the overall performance was fair, the results for the peroxisomal subset was poor. Geraghty *et al*.[23] used a pattern-based method to scan the *Saccharomyces cerevisiae* ORFs for potential peroxisomal proteins. Including both PTS1 and PTS2 motifs in their search, they found 18 new potential peroxisomal proteins. GFP fusions allowed them to confirm that about half of these proteins were truly located in the peroxisome. Another way to predict PTS1-containing protein is to use the PROSITE[24,25] pattern, [ACGNST]-[HKR]-[AFILMVY], for microbody C-terminal targeting signals, but this pattern also finds many non-peroxisomal proteins.

For lack of data on PTS2 proteins, we have chosen to focus on proteins carrying the C-terminal PTS1. In this work, we have (i) constructed an amino acid sequence based predictor for PTS1-mediated peroxisomal targeting, including both a motif identification step and a machine learning module, (ii) coupled our predictor with the subcellular localization predictors TargetP and TMHMM to improve performance, (iii) applied our prediction scheme on the predicted proteins (ORFs) of eight eukaryotic genomes, (iv) searched for and clustered homologs between the predicted sets from these eight genomes in order to reinforce localization prediction in a manner inspired by phylogenetic profile analysis,[26–28] and finally (v) expanded the clusters by searching for proteins with domain composition identical with the proteins in the clustered sets.

By combining these different approaches, we identify a set of strongly predicted peroxisomal proteins from eight eukaryotic genomes. This set enables us to make cross-genomic comparisons and to make an initial guess at the contents of the different peroxisomal proteomes.

Construction of the PeroxiP Predictor

The PTS1 motif and the data sets

As described in Methods, we initially extracted 152 peroxisomal proteins with a true PTS1 as well as 308 non-peroxisomal proteins with a PTS1-like C-terminal tripeptide from Swiss-Prot.[29] It cannot be ruled out that PTS1 from animal, plants and fungi could show organelle specificity within the microbody family, however, the data sets would be considerably reduced by additional subdivision into glyoxysomal and peroxisomal proteins. Also, it has been shown that glyoxysomal proteins can be imported into both animal[30] and plant

Table 1. The 35 PTS1 motifs from SWISS-PROT set of 152 peroxisomal proteins. Three out of the 35 PTS1 motifs were excluded from the accepted set of motifs, since these motifs contained at one of their positions an amino acid that was found only once at that position in the entire SWISS-PROT peroxisomal set (see the text)

PTS1 motif	Excluded	Occurrences in SWISS-PROT sets	
		Peroxisomal	Non-peroxisomal
AHL		3	5
AKA		1	19
AKF		1	3
AHI		1	9
AKL		19	21
AKM		3	1
AKV		1	19
ANL		2	5
ARF		1	2
ARL		4	9
ARM		5	–
ARY	Yes	1	1
ASL	Yes	1	23
CKL		2	4
HRL		1	1
HRM		2	1
KKL		2	6
NKL		4	10
PHL		1	8
PKL		3	4
PRL		1	3
SHL		10	18
SKF		2	7
SKI		6	2
SKL		42	20
SKM		4	1
SKV		1	24
SNL		2	9
SQL		4	4
SRL		14	11
SRM		4	–
THL		1	18
TKL		1	23
TKV		1	4
YRM	Yes	1	13

peroxisomes.[31] Thus, we decided to pool all microbody PTS1 proteins.

We found 35 different PTS1 motifs in our peroxisomal set, Table 1. The "original" PTS1-motif, SKL, and one-residue variations thereof, are the most abundant among the motifs, and in the sequence logos, (Figure 1), the degree of conservation at the PTS1 site is shown to be quite high, as expected. To minimize the effect of potential sequencing and annotation errors, three out of the 35 PTS1 motifs were excluded from the accepted set of motifs. At one of their positions, these motifs contained an amino acid that was found only once at that position in the entire set. In order to further optimize the performance on the test set, more PTS1 motifs could have been excluded (e.g. SKV which is only present once in the true peroxisomal set but in 24 non-peroxisomal proteins). However, the main aim here was not to optimize the performance, but to use the PTS1-filtering step to remove obvious false positives while losing only a minimal set of peroxisomal proteins.

We also constructed a relaxed consensus motif, by allowing any amino acid seen in each of the last three positions, i.e. [ACHKNPST]-[HKNQRS]-[AFILMV]. Comparing this with the PROSITE[24,25] motif for microbody targeting, [ACGNST]-[HKR]-[AFILMVY], some differences are found, predominantly due to our motif being more permissive. The only two exceptions to this are that our motif lacks glycine in the first position and tyrosine in the last position

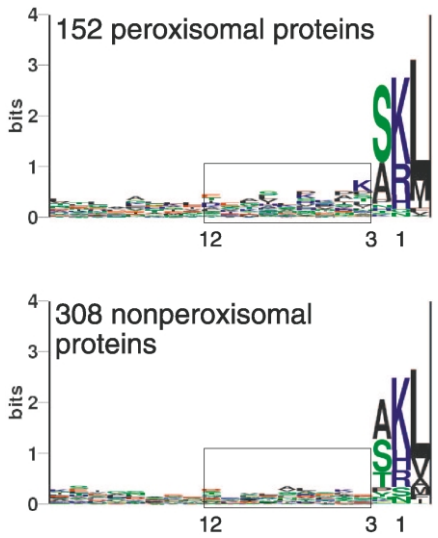


Figure 1. Sequence logos of C-terminal part of 152 peroxisomal proteins with a PTS1 (upper panel), and 308 non-peroxisomal proteins with a PTS1-like motif (lower panel). The 9-mers used in the pattern recognition module of the predictor are framed. The plots show the degree of conservation, I, which is calculated based upon the Shannon entropy,[64] H:

$$I = H_{\max} - H = \log_2 20 + \sum_{s \in A,C,D,\ldots,Y} f_s \log_2 f_s$$

where the f are the frequencies of each of the $n = 20$ amino acids at that particular position. The height of each letter within each bar corresponds to the frequency of that particular amino acid at that position. A totally conserved position would correspond to $I = \log(20) = 4.3$ bits.

The region next to the C-terminal tripeptide is much less conserved than the PTS1, but it has been shown that this region does play a role in the binding of a PTS1-containing protein to the Pex5 receptor.[16,32] Presumably, the C-terminal PTS1 must be exposed on the surface of the protein, and the adjacent region may also influence the strength of binding to the receptor. Comparing the logos of the peroxisomal and non-peroxisomal proteins (Figure 1) reveals that the degree of conservation in the adjacent region is slightly but detectably higher in the peroxisomal set. Obviously, the fact that the number of non-peroxisomal proteins in SWISS-PROT with a PTS1-like motif is approximately twice the number of true peroxisomal proteins with a PTS1, is by itself an indication that the signaling system for peroxisomal localization cannot solely be a question of having the right three residues at the C terminus.

Development of the predictor

The PeroxiP predictor consists of a preprocessing module that employs other localization predictors to remove potential false positives, a motif identification module that compares the C terminus of the query protein to a list of approved PTS1 motifs, and a pattern recognition module that analyses the sequence region adjacent to the PTS1 and that makes use of machine learning techniques. The modules are applied in this order (Figure 2) but will be presented in the reverse order for clarity.

Pattern recognition module

After some unsuccessful attempts at constructing a purely neural network-based predictor (see Methods), we decided to separate the PTS1 motif identification and the machine-learning based pattern analysis of the adjacent region. In the predictor, we can then choose to include or exclude the pattern analysis of the adjacent region.

Thus, the machine-learning pattern recognition techniques were applied only to the 9-mer adjacent to the C-terminal tripeptide and not to the tripeptide itself (Figure 1). We used both standard feed-forward neural networks (NN) with sigmoidal neurons,[33,34] and SVMs with polynomial kernels.[35] Input data were encoded using sparse encoding,[36] and the overall amino acid composition of the protein was used as input along with the sequence of the 9-mer (Methods). Fivefold cross-validation was used in the training, thus avoiding testing the predictor on the same data as it was trained on.

The final neural network architecture was based on a heuristic search in parameter space, trying various learning rates, number of training cycles, and number of hidden nodes. The resulting network was trained for 250 cycles with a learning rate of 0.005, with two hidden nodes, and one

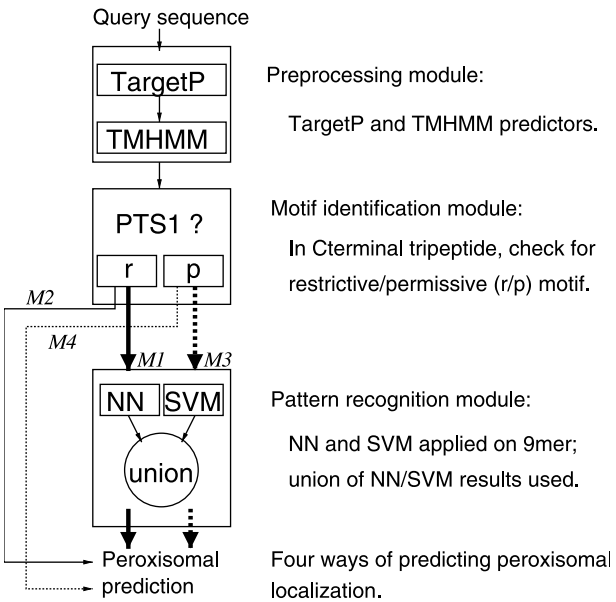


Figure 2. The PeroxiP prediction scheme (not including the clustering step). A query protein is first processed through TargetP and TMHMM to remove secreted and trans-membrane proteins. The next step is the motif identification module where the presence of a PTS1 motif is checked. There are two versions of the PTS1 motif; r, restrictive, corresponding to the 32 true PTS1 motifs found in the peroxisomal SWISS-PROT dataset, and p, permissive, which accepts the relaxed consensus motif [ACHKNPSTY]-[HKNQRS]-[AFILMV]. A query protein that passes the motif check is processed through the pattern recognition module, as indicated by the thick arrows labeled M1 (method 1, continuous arrow, uses restrictive PTS1 motif) and M3 (method 3, broken arrow, uses permissive PTS1 motif). The pattern recognition module examines the sequence region adjacent (framed in Figure 1) to the PTS1, using neural network (NN) and support vector machine (SVM) modules. It suffices that the protein is predicted as peroxisomal by either of these two modules (i.e. we use the union of NN and SVM predictions). For methods 2 and 4, the pattern recognition module is by-passed, as indicated by the thin arrows labeled M2 (method 2, continuous arrow, uses restrictive PTS1 motif) and M4 (method 4, broken arrow, uses permissive PTS1 motif).

output node taking values in the [0,1] interval. A cutoff of 0.40 yielded reasonable performance both in terms of specificity and sensitivity.

The SVM module employs a polynomial kernel. The score of a certain protein is produced by processing its representation through the SVM and is in the interval $[-1,1]$. It was found that -0.15 is a suitable cutoff. A protein was predicted to be peroxisomal if at least one of its NN or SVM scores was above the corresponding cutoff (i.e. we used the union of NN and SVM predictions), as this improved the performance over either of the individual methods.

Motif identification module

The motif identification module, which precedes the pattern recognition module (Figure 2) checks for the presence of a PTS1-motif at the C terminus at each query sequence. Two versions of the module were developed: one that only accepts the 32 motifs from the SWISS-PROT set (Table 1) and one that accepts the relaxed consensus motif that was constructed from these 32, [ACHKNPST]-[HKNQRS]-[AFILMV], in total 288 different motifs. Proteins without an accepted motif are directly predicted as non-peroxisomal, and only proteins with an accepted motif are further processed through the pattern recognition module.

Preprocessing module

Even though dual location of peroxisomal proteins occurs, it is highly unlikely that proteins located in the peroxisome also are secreted out of the cell. Therefore, we added a preprocessing step to our prediction schema: proteins predicted to be secreted with high reliability (TargetP[37] prediction with reliability coefficient of 1–3, corresponding to a specificity of 0.97) were excluded from analysis of the motif identification and pattern recognition modules, as were also proteins predicted to contain at least one trans-membrane region (using TMHMM v2.0[38] to predict trans-membrane proteins).

Combining motif identification and pattern recognition modules

The two different motif requirements—the restrictive, only allowing the 32 PTS1s from the SWISS-PROT set, and the permissive, allowing any protein matching the relaxed consensus motif—result in two versions of the predictor with different expected sensitivity and specificity. This was further expanded into in total four versions, method 1–method 4 (Table 4), by allowing the prediction to by-pass the entire pattern recognition module (Figure 2) and thus predicting peroxisomal localization based only on the presence of a PTS1 motif. Many of the proteins predicted this way are certainly wrong (false positives), but on the other hand, the number of missed proteins (false negatives) should be marginal. All four versions are used in our genome-wide predictions, but unless otherwise stated, only the most restrictive method (method 1) is discussed in the next section.

Performance of the predictor

Using the optimized cutoffs, and including the preprocessing module (TargetP and TMHMM predictions), a Matthews correlation coefficient (MCC) of 0.50 was recorded for the test set. With these cutoffs, the sensitivity was 0.78 and the specificity 0.64. The predictor performs better than PSORT,[20,21] which has an MCC of 0.44 on the same set (including the TargetP pre-processing to improve performance). The results are summarized in Table 2.

Table 2. Performance of PeroxiP (method 1) and PSORT[20] predictors, including preprocessing but without clustering

	Test set			Human SWISS-PROT set	
Predictor	MCC	Sens.	Spec.	Sens.	Spec.
PeroxiP	0.50	0.78	0.64	0.50	0.64
PSORT	0.44	0.47	0.76	0.25	0.54

Sens., sensitivity, which is the fraction of proteins known to belong to a specific compartment that actually are predicted to that compartment, $tp/(tp + fn)$; spec., specificity, which is the fraction of proteins predicted to a specific compartment that actually belongs to that compartment, $tp/(tp + fp)$; MCC, Matthews correlation coefficient, $MCC = (tp^{*}tn - fp^{*}fn)/\sqrt{(tn + fn)(tn + fp)(tp + fn)(tp + fp)}$ which is one for a perfect prediction and zero for a random assignment. Tp, true positive; fn, false negative; fp, false positive; tn, true negative. The test and human SWISS-PROT sets are described in the text.

To get an estimate of the "life-like" performance on a larger and more realistic set, we tried PeroxiP on the set of all human SWISS-PROT proteins with subcellular location annotated (SWISS-PROT release 40.17). This set contained 28 matrix peroxisomal proteins (after exclusion of 13 membrane spanning or associated proteins) and 5174 other proteins. PeroxiP reached a sensitivity of 0.50 and a specificity of 0.64 on this set (method 1; for methods 2–4, sensitivity was higher and specificity was lower, data not shown). PSORT was also tested and obtained a sensitivity of 0.25 and a specificity of 0.54.

Taking a closer look at the 28 peroxisomal proteins in the human SWISS-PROT set, we found that three (O00116, O14832, P09110) were annotated to be targeted to the peroxisome by a PTS2 signal. These proteins were not predicted to be peroxisomal by PeroxiP, since it does not handle PTS2-containing proteins. Of the remaining 25 proteins, between 14 and 21 were predicted to be peroxisomal by the various versions of PeroxiP. One protein (P34913) was erroneously predicted as non-peroxisomal, since in the pre-processing it was predicted by TargetP to be secreted and thus excluded from further analysis. Two proteins (Q9UHK6, Q13907) were missed by the methods 1 and 2 (that use the strict motif requirement), since they both had one of the three PTS1s that were removed in the construction of the training data. Still, using even the most relaxed motif (methods 3 and 4), three proteins were excluded because of the motif (Q03426, -DGL; O60683, -HYR; P57989, -VRV). For at least one of these proteins (O60683), there is some evidence for membrane association. It should be noted that none of these are actually annotated to contain a PTS1, only to be peroxisomal. Some of the proteins in the human SWISS-PROT set were also present in the PeroxiP training and test set; for these proteins, the test set scores were used in the calculation of performance (Table 2).

Predicting the Peroxisomal Proteome *in Silico*

Recently, attempts to systematically identify peroxisomal proteins have been carried out in different model organisms. In greening cotyledons of *Arabidopsis thaliana*, 53 proteins were analyzed by MALDI-TOF mass spectrometry and 29 were identified.[39] In *S. cerevisiae*, 19 soluble proteins were identified from 1D-gel electrophoresis.[40] Using 2D-gels from liver and kidney tissues from *Mus musculus* we have separated approximately 70 proteins (S.C., unpublished data). Taking all this into account, a reasonable estimate would be that the peroxisomal proteome should in most organisms contain between a few tens and one hundred different proteins.

For the *in silico* predictions, eight sequenced eukaryotic genomes were collected (see Table 3) and processed through all four versions of PeroxiP, Figure 2. Using the most restrictive method (method 1), we found around 60 proteins in most of the genomes (Table 4) with the notable exception of *Schizosaccharomyces pombe*, where only ten proteins were found. As pointed out by one of the referees it has not been shown that *S. pombe* actually contains a peroxisome. Its low number of potential peroxisomal proteins could therefore be seen as an indication that *S. pombe* might not contain a peroxisome. Using the most permissive method (method 4) where the only requirement is a C-terminal tripeptide that coincides with the relaxed motif, the number of proteins predicted to be peroxisomal ranged from 210 (*S. pombe*) to 1592 (*Oryza sativa*).

Cross-species clustering

A cross-species clustering procedure, based on

Table 3. Web addresses for the eight eukaryotic genomes that were scanned for potential peroxisomal proteins

Genome	Internet resource	No. ORFs
S. cerevisiae	http://mips.gsf.de/proj/yeast/CYGD/db/index.html/all_contig.pep.fa	6449
S. pombe	ftp://ftp.sanger.ac.uk/pub/yeast/pombe/Protein_data/pompep	4962
A. thaliana	ftp://ftpmips.gsf.de/cress/arabiprot/arabi_all_proteins_v120302.gz	25,826
O. sativa	ftp://ftp.tigr.org/pub/data/o_sativa/irgsp/PUBLICATION_RELEASE/	41,915
C. elegans	ftp://ftp.sanger.ac.uk/pub/C.elegans_sequences/WORMPEP/	20,414
D. melanogaster	ftp://ftp.ebi.ac.uk/pub/databases/edgp/sequence_sets/aa_gadfly.dros.Z	13,729
M. musculus	ftp://ftp.ensembl.org/pub/current_mouse/data/fasta/pep	28,097
H. sapiens	ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/protein/	38,051

Table 4. The number of predicted PTS1-targeted peroxisomal proteins in eight eukaryotic genomes using the four versions of our predictor before clustering

	Motif r/p	NN/SVM module	SC	SP	AT	OS	CE	DM	MM	HS	Total
Method 1	r	Yes	27	10	61	102	61	58	59	44	422
Method 2	r	No	64	36	198	249	164	116	198	240	1265
Method 3	p	Yes	77	53	337	574	251	156	217	243	1908
Method 4	p	No	277	210	1146	1592	755	482	947	1427	6836

Motif r/p: r, restrictive motif (accepts only the exact 32 PTS1 motifs from the SWISS-PROT set); p, permissive motif (accepts the relaxed consensus motif that was constructed from these 32, in total 288 different motifs). NN/SVM module, is the pattern recognition module working on the PTS1-adjacent region employed. SC, *Saccharomyces cerevisiae*; SP, *Schizosaccharomyces pombe*; AT, *Arabidopsis thaliana*; OS, *Oryza sativa*; CE, *Caenorhabditis elegans*; DM, *Drosophila melanogaster*; MM, *Mus musculus*; HS, *Homo sapiens*.

the Pfam domain contents in the predicted proteins, was then performed in order to find orthologous proteins that were predicted to be peroxisomal in more than one organism. The assumption is that conservation of function and localization is correlated and thus that a prediction of a protein being peroxisomal is strengthened if also its orthologs in other organisms are predicted to be peroxisomal. A somewhat similar phylogenetic profile method has previously been used to predict subcellular location of a subset of mitochondrial proteins, by using the phylogenetic distribution of proteins among several prokaryotes.[41]

The idea of making use of protein domain contents for predicting subcellular localization was introduced by Mott *et al*.[42] Their method, which does not include any cross-species comparisons, is based on the observation that proteins containing particular domains often are co-localized. Here, we make use of this observation, and use it to reinforce a subset of the prediction results by searching for homologs among the predicted peroxisomal sets from all eight genomes.

We examined different clustering methods to construct the phylogenetic profiles of predicted peroxisomal proteins. First, ordinary BLAST[43,44] searches were carried out between all proteins in the sets predicted by methods 1–4 for the eight organisms. The proteins were then clustered to produce clusters of orthologous peroxisomal proteins. This clustering was transitive, i.e. if proteins A and B, and B and C are considered homologs, then also proteins A and C are considered homologs even if they do not have a pairwise BLAST alignment score satisfying the cutoff. Secondly, we used the Pfam-A database release 7.2,[45] and assigned where possible Pfam domain(s) to every predicted protein. Then, proteins containing the same Pfam domain were clustered. A protein that contains more than one Pfam domain will end up in more than one cluster, which was encountered for 113 out of the 291 proteins with a Pfam assignment (method 1). For both approaches, we used an *e*-value cutoff of 0.001.

It turned out that to a large extent the clusters constructed using BLAST and Pfam were similar. Therefore, we decided to use the Pfam clustering first, and only use the BLAST clustering for all proteins that completely lacked Pfam domain assignments. Functional assignments to the BLAST clusters were done using the Pcons method.[46] It was found that the single BLAST cluster (of proteins lacking a Pfam domain assignment) detected in at least three genomes from method 1 was a remote homolog to the hydrolase superfamily (Pcons score of 3.9 to Prolyl Oligopeptidase (PDB code: 1E5T). Very few clusters contained proteins from all eight genomes—only ten clusters even for the sets predicted with the most permissive method (method 4) and none for the sets predicted with the most restrictive method. This could partly be explained by the low number of PTS1 peroxisomal proteins found in *S. pombe*. Incomplete annotation of the C terminus in the genomes might also have affected the performance of our predictor. Further, even in ubiquitous biochemical pathways in microbodies like β-oxidation, some proteins follow the PTS2 pathway,[47] and in the biosynthesis of cholesterol it has been postulated that proteins completely lacking PTS1 and PTS2 signals may enter peroxisomes by alternative mechanisms not yet defined.[48] All these proteins were obviously missed in our predictions.

Analysis of clusters

With the clusters in hand, we set out to find a cluster cutoff that would enable us to infer a set of more reliably predicted peroxisomal proteins. Comparing the Pfam domains of the clusters with known peroxisomal functions, we decided that a reasonable cutoff for method 1 was to demand the domain/orthologs to be present in the predicted peroxisomal sets from at least three genomes, Table 6.

An obvious improvement of the clustering procedure would be to consider more specific phylogenetic patterns that take into account known differences between peroxisomal functions in different organisms. However, in this work, we decided to focus on the more ubiquitous peroxisomal pathways and leaving the lower, sparse part of the list of clusters for later analysis.

For the data sets predicted using method 1, we found 28 clusters above the cluster cutoff including both Pfam and BLAST generated clusters (Table 6). To analyse the performance we manually classified all Pfam-domains detected as "potentially

peroxisomal" or "most likely not peroxisomal" (see Table 5 and Supplementary Material). Obviously this classification is quite rough, as the same Pfam domains might exist in both peroxisomal and non-peroxisomal proteins, and because our manual classification might contain errors. Although limited in its accuracy this analysis should give an indication of the performance of the different methods, and aid in the production of a set of more reliably predicted peroxisomal proteins.

Among the 28 accepted clusters built on the results from method 1, 27 (96%) were deemed compatible with known peroxisomal functions (Table 5). One cluster (Pfam domain LIM) represents a function most likely not peroxisomal (it is a metal-binding region that most frequently is involved in transcription regulation). The fraction of "potentially peroxisomal" domains decreased significantly for the clusters constructed on the results from methods 2 to 4, Table 6.

Translating the results to the protein level, for method 1 clusters, 149 proteins were found whereof 145 (97.3%) contain a Pfam domain that is compatible with known peroxisomal functions, Table 6. For method 4 clusters, there is a total of 1446 proteins above the cluster cutoff, whereof 529 (36.6%) contain Pfam domains compatible with peroxisomal functions.

Next, we investigated the functional annotations of these proteins or ORFs (as found in the downloaded genomic sets; Table 3) to find out how many of them had a known or suspected subcellular location. Of the 149 proteins from method 1, 15 were annotated as being peroxisomal, and only one had a non-peroxisomal subcellular location annotation. Thus, using method 1 we are able to suggest 133 new potential peroxisomal proteins from the eight genomes.

In an attempt to assess the sensitivity of the clustering, we found that 63% of the 152 PTS1-containing proteins in SWISS-PROT had a homolog in our final set of 149 proteins (method 1 clusters), while for the method 4 clusters a homolog for 73% was found. These figures suggest that we were able to pick up a relatively large proportion of the known peroxisomal (and PTS1-targeted) proteins already using the most restrictive version of our prediction scheme, while the increase in sensitivity was relatively small when using the other methods.

Conversely, 24 of the proteins in the predicted set of 149 proteins (method 1 clusters) did not match any of the SWISS-PROT PTS1 proteins (one being annotated as containing a PTS1, though; this somewhat strange situation is probably because this sequence has not been entered in SWISS-PROT yet). With PeroxiP and the following clustering procedure we were able to infer their subcellular localization as being peroxisomal with a high reliability, and for the proteins with a homolog, we were able to further support a suggested peroxisomal localization. The predicted sets are available as Supplementary Material.

Table 5. Domain classification and the number of organisms that the most ubiquitous Pfam[45] domains were present in

Pfam ACC	Pfam ID	P/N	M1	M2	M3	M4
PF00106	adh_short	P	7	7	8	8
PF00378	ECH	P	6	6	6	6
PF00441	Acyl-CoA_dh	P	6	6	6	6
PF00501	AMP-binding	P	6	6	8	8
PF00561	abhydrolase	P	5	6	5	7
PF01070	FMN_dh	P	5	5	5	5
PF00069	pkinase	P	4	8	6	8
PF00070	pyr_redox	P	4	6	5	7
PF00412	LIM	N	4	4	6	7
PF00702	Hydrolase	P	4	3	5	6
PF00725	3HCDH	P	4	6	4	6
PF00755	Carn_acyltransf	P	4	5	5	5
PF01014	Uricase	P	4	4	4	4
PF01266	DAO	P	4	5	4	5
PF01756	ACOX	P	4	6	4	6
PF02036	SCP2	P	4	5	4	5
PF02551	Acyl_CoA_thio	P	4	5	4	5
PF02737	3HCDH_N	P	4	6	4	6
PF00056	ldh	P	3	3	4	6
PF00108	thiolase	P	3	3	3	4
PF00293	NUDIX	P	3	5	4	8
PF00578	AhpC-TSA	P	3	3	4	5
PF01274	Malate_synthase	P	3	3	3	3
PF01565	FAD_binding_4	P	3	3	3	4
PF01575	MaoC_dehydratas	P	3	4	3	4
PF02770	Acyl-CoA_dh_M	P	3	3	4	4
PF02803	thiolase_C	P	3	3	3	3
(BLAST)	"Hydrolase"	P	3	3	3	3

The Table is limited to the 27 Pfam domains that were above the cluster cutoff for method 1 (i.e. present in at least three organisms). As a 28th domain, the cluster found in the BLAST clustering of proteins completely lacking any Pfam domain has been added. Its functional annotation (Pfam ID column) was done using Pcons[46] (see the text). P/N, peroxisomal/non-peroxisomal; M1, method 1; M2, method 2; M3, method 3; M4, method 4.

Table 6. Clustering results

		No. clusters		No. proteins		
PeroxiP version	Cluster cutoff	Tot.	Perox.	Total unclustered	w. cluster cutoff	Perox.[a]
Method 1	3	28	27	422	149	145 (97.3%)
Method 2	4	32	22	1265	277	221 (79.8%)
Method 3	4	45	26	1908	418	239 (57.2%)
Method 4	6	56	20	6836	1446	529 (36.6%)

The number of predicted PTS1-targeted peroxisomal proteins in eight eukaryotic genomes using clustering and cluster cutoff, based on the four versions of PeroxiP. Both Pfam and BLAST clusters included. Perox., peroxisomal.

[a] According to classification in Table 5.

Expanding the predicted sets

As is clear from the above, the PTS1 signal used in method 1 (and 2) is too restrictive to identify all peroxisomal proteins, while the relaxed motif

used in methods 3 and 4 is too promiscuous. Although peroxisomes are ubiquitous organelles in eukaryotic cells, many peroxisomal matrix proteins are species-specific. Also within an organism, there are specialized members of the microbody family of organelles. For instance, higher plants possess several classes of peroxisomes and glyoxysomes that are present at distinct development stages and serve different metabolic roles.[49]

Among the seven metabolic pathways analyzed in next section (β-oxidation, α-oxidation, synthesis of glycerolipids, isoprenoid biosynthesis, degradation of amino acids, degradation of purines, and the glyoxylate cycle) only three—β-oxidation, isoprenoid biosynthesis and purine metabolism—are present in all eukaryotes. However, including solely proteins extracted from more than three genomes using method 1, we believe that we have identified a significant fraction of the peroxisomal proteome at a quite high specificity.

We can expand the prediction to all genomes by identifying homologs between the proteins in the method 1 clusters above the cutoff (28 clusters, 149 proteins, Tables 5 and 6) and all proteins from method 4 (before clustering), i.e using the relaxed PTS1 motif and by-pass the pattern recognition module. Homologs were identified by assuming that proteins sharing a set of Pfam domains were homologous (see Methods). Homology detection led to a final set of 430 predicted peroxisomal proteins. These 430 proteins are divided into 23 groups of homologous proteins classified by their function and species in Table 7. In addition to these 23 groups of peroxisomal proteins, searching the method 4 sets, we have identified six additional groups of proteins that have a known peroxisomal function.

According to these results, the peroxisomal proteome from plants is larger than in mammals which in turn is larger than in the fungal genomes. The latter is in agreement with the current knowledge on peroxisomal function, as some pathways are not present in yeast but only in higher organisms. It is also not surprising that plants have more peroxisomal proteins than animals as they have several types of microbodies using the same targeting system, including leaf peroxisomes, glyoxysomes and unspecialized peroxisomes.[31] Strangely, for some reason twice as many peroxisomal proteins were predicted for *M. musculus* than for *Homo sapiens*, although all mammals are believed to have identical peroxisomal function. This is most likely due to the incomplete annotation state of both these genomes. It is possible that the *M. musculus* from TIGR used here, contains more exact C-terminal annotation than the *H. sapiens* genome from ENSEMBL.

Biological Implications

Our prediction method has succeeded in extracting the majority of peroxisomal proteins described in literature, moreover, several novel proteins with possible roles in peroxisomal biochemistry have additionally been found. In this final section, we analyse the predictions for several peroxisomal pathways in some more detail.

β-Oxidation

The β-oxidation pathways show differences in subcellular distribution in different organisms. In fungi, it is strictly a peroxisomal process,[50] while in mammals and plants, peroxisomes contain two fatty acid β-oxidation pathways—the predominant route for straight chain fatty acid breakdown and an alternative route for branched chain fatty acids—and in addition a fatty acids β-oxidation pathway which acts on signaling lipids that are only poorly oxidized by plant mitochondrial pathways.[3,51,52] Nematodes resemble mammals in that they possess both peroxisomal and mitochondrial β-oxidation.[53]

Proteins A1–A11 in Table 7 correspond to all enzymes involved in β-oxidation and several auxiliary proteins from this pathway bearing PTS1 motif. Acyl-CoA oxidase (A2) was not found in yeast; however, it is known that this protein follows alternative targeting routes in various organisms. At least in rat, mouse and human, it contains a canonical PTS1, whereas in the yeasts *S. cerevisiae*, *C. tropicalis*, *C. maltosa* and *Y. lipolytica* neither a PTS1, nor a PTS2 motif is present.[47] Two copies of A2 are found in the human genome corresponding to the straight chain acyl-CoA oxidase and the branched chain oxidase or pritanoyl-CoA oxidase.[3] It is noteworthy that several A2 proteins found in plants are related to short chain acyl-CoA oxidases, a unique feature in plant β-oxidation. Our finding is in agreement with the suggestion that short chain acyl-CoA oxidases might have arisen from acyl-CoA dehydrogenase, gaining a peroxisomal targeting signal during evolution.[54]

Interestingly, our method has retrieved candidates for bi-functional protein 2 enzyme (DBP) (A3b) both in *Caenorhabditis elegans* and *Drosophila melanogaster*; so far, solely described in mammalian peroxisomes.[55] An example of a missing protein is thiolase 2 (A4) from *H. sapiens*. Although human A4 contains a PTS1 motif, it was not found by PeroxiP, since the sequence was not annotated in the original genome data. This example highlights that the annotation of all eukaryotic genomes is still far from being completely accurate. However, it also indicates that cross-genomics studies is a possible approach to circumvent this problem.

Peroxisomes also house enzymes to convert polyunsaturated or unsaturated fatty acids into appropriate β-oxidation substrates (A5–A8). The restrictive method 1 found candidates for protein A5–A7. Proteins displaying a very weak Pfam domain could have been missed, as is the case of enoyl-CoA isomerase (A7) from *S. cerevisiae*. On the other hand, isocitrate dehydrogenase (A8) was

Table 7. Expanding the clustered sets

Code	Protein	Domains	Pathway	SC	SP	AT	OS	CE	DM	MM	HS
A1	Acyl-CoA synthetase	AMP-binding	β-Oxidation	1	1	14	15	6	4	9	4
A2	Acyl-CoA oxidase	[Acyl-CoA_dh_N] [Acyl-CoA_dh_M] [Acyl-CoA_dh] [ACOX]	β-Oxidation	0	0	6	6	15	5	5	2
A3a	LBP (Bifunctional protein)	[ECH] 3HCDH_N [TrkA-N] 3HCDH	β-Oxidation	0	0	3	2	1	1	1	1
A3b	DBP (Bifunctional protein)	adh_short [MaoC_dehydratas] [SCP2]	β-Oxidation	1	0	0	0	2	2	1	0
A4a	pTH2 and SCPx	[thiolase] [thiolase_C] [ketoacyl-synt_C] SCP2	β-Oxidation	0	0	1	0	1	2	3	1
A5 + A7	2,4-Dienoyl-CoA isomerase and Δ3-enoyl-CoA isomerase	[ACBP] ECH	β-Oxidation	0	0	6	2	3	2	4	1
A6	2,4-Dienoyl-CoA reductase	adh_short [adh_zinc]	β-Oxidation	1	2	17	5	9	2	6	2
A8	Isocitrate dehydrogenase	ISODH	β-Oxidation	Only found by method 4							
A9	Catalase	Catalase	β-Oxidation	Only found by method 4							
A10	Carnitine acyltransferase (CAT)	Carn_acyltransferase	Aux β-oxidation	1	0	0	0	2	2	3	1
A11	Acyl-CoA thioesterase	[cNMP_binding] Acyl_CoA_thio	Aux β-oxidation	1	0	1	1	3	0	1	0
B1	2-Hydroxyphytanoyl-CoA lyase	TPP-N TPPenzyme TPP-C	α-Oxidation	Only found by method 4							
B2	α-Methylacyl-CoA racemase	CAIB-BAIF	α-Oxidation	Only found by method 4							
C1	Acetoacetyl-CoA thiolase	thiolase thiolase_C	Isoprenoid biosynthesis	0	0	0	0	2	1	0	0
C2	Phosphomevalonate kinase	[FHA] pkinase	Isoprenoid biosynthesis	7	5	34	20	10	5	10	7
C3 + C4	Isopentenyl-pyrophosphate isomerase (IPP) and farnesyl-pyrophosphate transferase (FPP)	NUDIX	Isoprenoid biosynthesis	1	1	6	5	1	1	6	2
D1	Acyltransferase	Acyltransferase	Glycerolipid synthesis	Only found by method 4							
D2	Alkyl DHAP synthase	FAD_binding_4 [FAD-oxidase_C]	Glycerolipid synthesis	0	0	1	3	1	1	0	0
E1	D-Amino acid oxidase	DAO [pyr_redox] [Amino_oxidase]	Amino acid metabolism	0	0	0	1	3	2	1	1
F1	Urate oxidase	Uricase	Purine metabolism	0	0	1	0	0	1	1	1
G1	Isocitrate lyase	ICL	Glyoxylate cycle	Only found by method 4							
G2	Malate synthase	Malate_synthase	Glyoxylate cycle	2	0	1	2	0	0	0	0
G3	Glyoxylate oxidase	FMN_dh [Glu_synthase] [IMPDH_C]	Glyoxylate cycle	0	0	2	1	1	1	2	0
G4	Malate dehydrogenase	ldh [Semialdhyde_dh] ldh_C	Glyoxylate cycle	1	0	0	2	1	2	1	1
P1	Hydrolase	Hydrolase	–	1	2	0	1	0	0	0	0
P2	Epoxide hydrolase	[Hydrolase] Abhydrolase	–	1	0	6	11	1	4	3	1
P3	Alkyl hydroperoxide reductase	AhpC-TSA	Antioxidant function	1	1	0	2	0	0	1	1
P4	Acyl-CoA dehydrogenase or short chain dehydrogenase	pyr_redox	–	2	2	9	5	0	3	1	0

(*continued*)

Table 7 *Continued*

Code	Protein	Domains	Pathway	SC	SP	AT	OS	CE	DM	MM	HS
X1	LIM	[NB-ARC] [AAA] LIM [homeobox]	–	0	0	2	0	2	1	6	4
Sum	In total: 430 proteins			21	14	110	84	64	42	65	30

The resulting 29 groups of proteins are classified according to their function and divided into seven pathways: β-oxidation, α-oxidation, synthesis of glycerolipids, isoprenoid biosynthesis, degradation of amino acid, degradation of purines and glyoxylate cycle, (A-G). For each protein a set of compatible domains are described, where domains within square brackets ([ ]) are not present in all the individual proteins. The complete list of all domains compatible with a specific group of proteins is available from the Supplementary Material. Of these groups 23 were found in the clusters based on method 1 and the rest were present in the set from method 4. Four groups of proteins (P1–P4) are not classified into peroxisomal pathways as their exact functions in peroxisomes are unknown to us. A fifth group (X1) contains the LIM domain which we believe is most likely not peroxisomal.

only predicted in *S. cerevisiae* using method 1, while it was found in six genomes with method 2.

α-Oxidation

Despite efforts to elucidate the α-oxidation in peroxisomes, the individual steps of this pathway have remained unknown until recently. No enzymes related to this route were found using method 1. However, applying method 2, 2-hydroxyphytanoyl-CoA lyase (B1) and α-methyl-CoA racemase (B2) were predicted in several organisms. The other known components of this pathway are either membrane proteins or follows the PTS2 targeting pathway, and were therefore not found.

Isoprenoid biosynthesis

The isoprenoid biosynthesis pathway is present in most organisms, but it is not an entirely peroxisomal pathway. Proteins C2–C4 were retrieved with method 1, while acetoacetyl-CoA thiolase (C1) was only found in two genomes. Phosphomevalonate kinase (C2), bearing a pkinase domain, was found in all genomes. However, the numbers in Table 7 for C2 are certainly overestimations as we would expect to find at most a couple of proteins of this type per genome. Here, we find up to 34 (for *A. thaliana*). This is most likely because several types of not well-studied kinase-containing proteins also share this domain. Those proteins have recently been identified from plant peroxisomes and phosphorylation of peroxisomal proteins was discussed as a possible novel peroxisomal function.[39] In human, isoprenyl-pyrophosphate isomerase (IPP) (C3) and farnesyl-pyrophosphate transferase (FPP) (C4) has been demonstrated to be targeted to peroxisomes *via* the PTS1 receptor.[56] Candidates for these proteins have been found in all genomes. They all bear a NUDIX domain, which is present in a variety of different proteins. Other enzymes in this pathway do not contain any identifiable PTS1 or PTS2 motifs, and it is likely that these proteins gain access to the matrix by piggybacking with another PTS containing protein.[57]

Other pathways

The glycerolipid synthesis pathway contains two enzymes with PTS1 motifs. Dihydroxiacetone phosphate (DHAP) acyltransferase (D1) was found in the proteins from method 4 in four organisms, whereas alkyl DHAP synthase (D2) was retrieved with method 1. The absence of these proteins from the human genome is in agreement with changes in targeting pathway upon evolution. In human, protein D2 bears PTS2 motif, whereas it is targeted *via* PTS1 in the other genomes.[58] Another ubiquitous reaction in peroxisomes is carried out by D-amino acid oxidase. Method 1 found candidates with PTS1 motifs in all genomes except *A. thaliana*, in which the enzyme has a PTS2, and in yeast.

The end product of purine metabolism varies from species to species. The degradation of purines to urate is common to all animal species, but the degradation of urate is much less complete in higher animals. Urate oxidase (uricase) (F1) is located in the peroxisome in all animals.[59] We have predicted F1 in four genomes. In yeast, this enzyme has a non-canonical PTS1 motif, and was therefore not found by PeroxiP.

Finally, plant glyoxysomes and yeast peroxisomes contain enzymes from the glyoxylate cycle. The glyoxylate shunt is a bypass of the tricarboxylic acid cycle that permits gluconeogenesis starting from acetyl-CoA, which is generated following fatty acid catabolism.[60] Malate synthase (G2) was correctly found in all expected genomes except *S. pombe*. Predictions for glyoxylate oxidase (G3) and malate dehydrogenase (G4) were retrieved from several genomes. Curiously, G4 from *A. thaliana* bears a PTS2 motif, while in *O. sativa* it seems to have a PTS1-like motif. Both, G3 and G4 contain Pfam domains common to several protein families, hence method 1 also found candidates in several animal genomes. Further analysis of these proteins and those classified under code P1–P4, will help us to identify novel peroxisomal functions.

Several families absent in plants

The largest fraction of proteins belongs to plants,

which is not surprising as these have several types of microbodies that use a similar import pathway.[31] However, ten of the families detected by method 1 are not found in plants. Several explanations for this is possible, including the use of a PTS1 signal not detected by our motif recognition module, the use of the PTS2 pathway, the discrepancies in domain sets, the incomplete and still erroneous status of the genome sequences, or, obviously, because some of these functions might not be performed in plants. In general, it seems as if the peroxisomal proteome that we detect is less clearly defined in plants.

Missed proteins

As seen in Table 7, six families of peroxisomal proteins that are known to use the PTS1 pathway were missed by the clustering method, and found only in the method 4 sets. Most of these proteins were missed, since in several genomes the proteins did not match the restrictive PTS1 motif set (Table 1), but some proteins were missed due to the "pattern recognition module" operating on the region adjacent to the C-terminal tripeptide (Figures 1 and 2).

Conclusions

We have developed a scheme for predicting peroxisomal localization of proteins, using information from the sequence in the form of the PTS1 motif and a machine learning-based module that analyses the PTS1-adjacent region. The PeroxiP predictor alone seems to work better than previous attempts at constructing peroxisomal localization predictors. The main difficulty of predicting PTS1-targeted peroxisomal proteins is that PTS1-like C-terminal tripeptides are found in many non-peroxisomal proteins, but presumably does not function as targeting signals because they are not properly exposed on the surface of the folded protein. We found that the specificity of PeroxiP can be substantially improved by a screening step where the TargetP and TMHMM predictors are used to remove secretory and integral membrane proteins, respectively. By combining TargetP, TMHMM and PeroxiP with a homology-based clustering procedure that makes use of the presence of identical Pfam domains as an indication of homology, we have been able to predict 430 peroxisomal proteins in eight different genomes, many of which lack localization annotation.

Methods

Data sets for predictor construction

Sequence data were collected from SWISS-PROT release 39.27.[29] PTS1-containing sequences were searched for among the entries containing the annotation "SUBCELLULAR LOCATION: *{PEROXISOMAL|GLYOXYSOMAL|GLYCOSOMAL}" using the label "MICROBODY TARGETING SIGNAL" where found, or otherwise including sequences with clear annotation about peroxisomal location and with a C-terminal tripeptide similar to any confirmed PTS1. One hundred and fifty-six sequences were found this way. In a manual control, four proteins were removed from the set (SWISS-PROT AC number O14313 and P14293, membrane proteins; P32573 and P38139, little evidence for the annotation of peroxisomal location). The final set of peroxisomal proteins with a PTS1 signal thus consisted of 152 sequences, whereof 21 with annotations about potential dual location (Q00614, P32796, Q13011, O35459, Q62651, P41689, P31029, O35423, P09139, Q9UHK6, O09174, P70473, mitochondrial/peroxisomal; O95822, P12617 cytoplasmic/mitochondrial/peroxisomal; P22307, P32020, P11915, Q07598, cytoplasmic/mitochondrial (internal targeting signal)/peroxisomal; P52826, P43155, P47934 endoplasmic reticulum/mitochondrial/peroxisomal). Two sequences (Q00925 and Q01962) were reported to contain both a C-terminal PTS1 and an N-terminal PTS2. The negative set was collected by searching among all eukaryotic proteins in SWISS-PROT release 39.27 for non-peroxisomal proteins that at their C terminus exhibited any of the 35 tripeptides found among the PTS1 tripeptides in the peroxisomal set. Three hundred and eight proteins were found in this way.

Proteins that had an C-terminal tripeptide in which one of the three positions contained an amino acid found only once at that position in the entire set of 152 peroxisomal proteins were removed. This was done as an attempt to remove the potentially most uncertain motifs (due to, e.g. sequencing or annotation errors). In total, this resulted in the exclusion of three motifs, -YRM, -ASL, and -ARY and the reduction of the set of known proteins with accepted PTS1 from 152 to 149 proteins, and the set of known non-peroxisomal proteins with PTS1-like C-terminal tripeptide from 308 to 271 proteins.

We performed redundancy reduction of the data sets in two steps. First, the 9-mers adjacent to (but not including; see next section) the C-terminal PTS1 signal were demanded to differ at two or more positions, and subsequently the Hobohm algorithm 2[61] was applied for protein removal. Second, the pairwise amino acid identity over the entire sequences was demanded to be at most 25%. BLASTClust[43,44] was used for protein clustering and removal, leaving 90 peroxisomal and 151 non-peroxisomal proteins in the redundancy reduced set used for the training and testing of neural networks and SVMs.

Neural network and support vector machine training

We decided on separating the PTS1 motif recognition and the machine-learning based pattern analysis of the adjacent region. This was done, since a pilot study had shown that the prediction was very dominated by the actual PTS1 motif, and that the partitioning into the five sets in the cross-validation was very sensitive to which particular motifs that ended up in which set. Accordingly, a laborious balancing of the motifs among the negative and positive sets would be required, thus in effect more or less forcing the prediction to be based on other parts of the sequence. If this balancing would have failed just slightly, then when applying the prediction on a whole-genome scale, most proteins presenting

any of the PTS1 motifs would probably be predicted as peroxisomal thus producing many false positives.

Thus, the machine-learning pattern recognition techniques were applied only on the 9-mer adjacent to the C-terminal tripeptide and not on the tripeptide itself, Figure 1. We used a standard feed-forward neural network (NN) with sigmoidal neurons[33,34] (Billnet implementation†,[62]), and a SVM with polynomial kernel[35] (SVM-LIGHT implementation[63]). Input data were encoded using sparse encoding,[36] yielding 20 input nodes (values) for each amino acid position and setting to one the node corresponding to the amino acid present at that position, letting the others take on the value zero. Overall amino acid composition information was also used as input and for this 20 values (nodes) were added to the input, each representing the frequency of one amino acid in the query protein (multiplied by 10 to yield numbers of the same order of magnitude as the sparse sequence encoding input units).

Fivefold cross-validation was used in the training. NN training was stopped when the error on the training set had reached a plateau, as decided by visual inspection. The number of epochs was chosen to be 250; the exact choice of stopping point is not critical, since the error is stable over a long range, also for the test set (data not shown). SVM training is a convex optimizations problem and thus guaranteed to find the global minimum given the problem and the constraints.

Classification of peroxisomal Pfam domains

In total 77 Pfam domains were found to be above the "cluster cutoff" in a cluster constructed from any of the four sets initially predicted by PeroxiP (i.e. using methods 1–4). For each of these domains, a manual classification into two groups "possibly peroxisomal" and "most likely not peroxisomal" was done by examining the annotated function of proteins containing this domain and comparing these to literature of described peroxisomal functions. The assignments for the 27 domains that were above the cluster cutoff for method 1 can be found in Table 5 (which also includes the one cluster of proteins without a Pfam domain assignment, that were clustered using BLAST, see Table legend). The complete list is posted on the Supplementary Material.

Expanding the predicted set

To expand the predictions to all genomes we wanted to identify all homologs with a possible PTS1 signal in all the eight genomes. First, all proteins that were found with method 1, using the subsequent clustering and applying the cluster cutoff, were extracted, and all Pfam domains in these proteins were collected. Secondly, the proteins found using method 4 (before the clustering) were tested for the presence of these Pfam domains. Proteins were considered as homologs if they contained at least one of the 27 Pfam domains (and the relaxed PTS1 signal). This yielded a final set of 430 proteins.

From these proteins we classified 29 different biochemical reactions (in seven pathways) believed to be peroxisomal (described in Table 7). It was found that some Pfam-domains, for instance thiolase, might belong to proteins performing different functions, i.e. taking part in more than one step in the pathway, and that proteins in different genomes might not contain the exact set of domains, due to mis-classification or domain rearrangements. Therefore, to finally classify all proteins into one of the 29 reactions of the different pathways, a list of compatible domains was defined for each group of proteins. For instance, if the thiolase domain is found with an SCP2 domain it was assumed that the protein was pTH2 (A4) but if the protein did not contain the SCP2 domain the protein is most likely acetoacetyl-CoA thiolase (C1). The exact classification of domains for each protein family can be found in Supplementary Material.

† http://www.iit.demokritos.gr/vasvir/billnet/

Acknowledgements

This project were supported by grants from The Swedish Research Council (G.vH., A.E. and S.C.), The Swedish Foundation for Strategic Research (A.E. and G.vH.) and the Carl Trygger foundation (S.C. and A.E.).

References

1. Lazarow, P. B. & Fujiki, Y. (1985). Biogenesis of peroxisomes. *Annu. Rev. Cell Biol.* **1**, 489–530.
2. van den Bosch, H., Schutgens, R. B., Wanders, R. J. & Tager, J. M. (1992). Biochemistry of peroxisomes. *Annu. Rev. Biochem.* **61**, 157–197.
3. Wanders, R. J., Vreken, P., Ferdinandusse, S., Jansen, G. A., Waterham, H. R., van Roermund, C. W. & van Grunsven, E. G. (2001). Peroxisomal fatty acid α and β-oxidation in humans: enzymology, peroxisomal metabolite transporters and peroxisomal diseases. *Biochem. Soc. Trans.* **29**, 250–267.
4. Opperdoes, F. R. & Borst, P. (1977). Localization of nine glycolytic enzymes in a microbody-like organelle in *Trypanosoma brucei*: the glycosome. *FEBS Letters*, **80**, 360–364.
5. Wanders, R. J. A., Barth, P. G., and Heymans, H. S. A. (2001). The Peroxisome. *The Metabolic and Molecular Bases of Inherited Disease*. (Scriver, C. R. & Sly, W. S., eds), vol. 2, chapt. 15, pp 3219–3256, McGraw-Hill, New York.
6. Wanders, R. J. A., Jakobs, C. & Skejldal, O. H. (2001). The Peroxisome. *The Metabolic and Molecular Bases of Inherited Disease*. (Scriver, C. R. & Sly, W. S., eds), vol. 2, chapt. 15, pp 3303–3322, McGraw-Hill, New York.
7. Gould, S. G., Valle, D. & Raymond, G. V. (2001). The Peroxisome. *The Metabolic and Molecular Bases of Inherited Disease*. (Scriver, C. R. & Sly, W. S., eds), vol. 2, chapt. 15, pp 3181–3217, McGraw-Hill, New York.
8. Subramani, S., Koller, A. & Snyder, W. B. (2000). Import of peroxisomal matrix and membrane proteins. *Annu. Rev. Biochem.* **69**, 399–418.
9. Titorenko, V. I. & Rachubinski, R. A. (2001). The life cycle of the peroxisome. *Nature Rev. Cell Biol.* **2**, 357–386.
10. Gould, S. J. & Collins, C. (2002). Peroxisomal-protein import: is it really that complex? *Nature Rev. Cell Biol.* **3**, 382–389.
11. Gould, S. J., Keller, G. A., Hosken, N., Wilkinson, J. &

Subramani, S. (1989). A conserved tripeptide sorts proteins to peroxisomes. *J. Cell Biol.* **108**, 1657–1664.
12. Subramani, S. (1993). Protein import into peroxisomes and biogenesis of the organelle. *Annu. Rev. Cell Biol.* **9**, 445–478.
13. Swinkels, B. W., Gould, S. J., Bodnar, A. G., Rachubinski, R. A. & Subramani, S. (1991). A novel, cleavable peroxisomal targeting signal at the amino-terminus of the rat 3-ketoacyl-CoA thiolase. *EMBO J.* **10**, 3255–3262.
14. Purdue, P. E. & Lazarow, P. B. (2001). Peroxisome biogenesis. *Annu. Rev. Cell Dev. Biol.* **17**, 701–752.
15. Titorenko, V. I., Nicaud, J.-M., Wang, H. H. C. & Rachubinski, R. A. (2002). Acyl-CoA oxidase is imported as a heteropentameric, cofactor-containing complex into peroxisomes of *Yarrowia lipolytica*. *J. Cell Biol.* **156**, 481–494.
16. Gatto, G. J., Jr, Geisbrecht, B. V., Gould, S. J. & Berg, J. M. (2000). Peroxisomal targeting signal-1 recognition by the TPR domains of human PEX5. *Nature Struct. Biol.* **7**, 1091–1095.
17. Neer, E. J., Schmidt, C. J., Nambudripad, R. & Smith, T. F. (1994). The ancient regulatory-protein family of WD-repeat proteins. *Nature*, **371**, 297–300.
18. Girzalsky, W., Rehling, P., Stein, K., Kipper, J., Blank, L., Kunau, W. H. & Erdmann, R. (1999). Involvement of Pex13p in Pex14p localization and peroxisomal targeting signal 2-dependent protein import into peroxisomes. *J. Cell Biol.* **144**, 1151–1162.
19. Albertini, M., Rehling, P., Erdmann, R., Girzalsky, W., Kiel, J., Veenhuis, M. & Kunau, W. H. (1997). Pex14p, a peroxisomal membrane protein binding both receptors of the two PTS-dependent import pathways. *Cell*, **89**, 83–92.
20. Nakai, K. & Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897–911.
21. Horton, P. & Nakai, K. (1997). Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *ISMB*, **5**, 147–152.
22. Cai, Y. D., Liu, X. J., Xu, X. B. & Chou, K. C. (2002). Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. *J. Cell. Biochem.* **84**, 343–348.
23. Geraghty, M. T., Bassett, D., Morrell, J. C., Gatto, G. J., Jr, Bai, J., Geisbrecht, B. V. *et al.* (1999). Detecting patterns of protein distribution and gene expression *in silico*. *Proc. Natl Acad. Sci. USA*, **16**, 2937–2942.
24. Bucher, P. & Bairoch, A. (1994). A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation. *ISMB*, **2**, 53–61.
25. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K. & Bairoch, A. (2002). The PROSITE database, its status in 2002. *Nucl. Acids Res.* **30**, 235–238.
26. Gaasterland, T. & Ragan, M. (1998). Microbial genescapes: phyletic and functional patterns of orf distribution among prokaryotes. *Microb. Comp. Genomics*, **3**, 199–217.
27. Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D. & Yeates, T. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
28. Liberles, D. A., Thoren, A., von Heijne, G. & Elofsson, A. (2002). The use of phylogenetic profiles for gene predictions. *Curr. Genomics*, **3**, 131–138.
29. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
30. Trelease, R. N., Choe, S. M. & Jacobs, B. L. (1994). Conservative amino acid substitutions of the C-terminal tripeptide (Ala-Arg-Met) on cottonseed isocitrate lyase preserve import *in vivo* into mammalian cell peroxisomes. *Eur. J. Cell Biol.* **65**, 269–279.
31. Olsen, L. J., Ettinger, W. F., Damsz, B., Matsudaira, K., Webb, M. A. & Harada, J. J. (1993). Targeting of glyoxysomal proteins to peroxisomes in leaves and roots of a higher plant. *Plant Cell*, **5**, 941–952.
32. Lametschwandtner, G., Brocard, C., Fransen, M., Veldhoven, P. V., Berger, J. & 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.
33. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error backpropagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition in Foundations* (Rumelhart, D., McClelland, J. & Group, P. R., eds), vol. 1, pp. 318–362, MIT Press, Cambridge, MA.
34. Minsky, M. & Papert, S. (1968). *Perceptrons: An Introduction to Computational Geometry*, MIT Press, Cambridge, MA.
35. Vapnik, V. N. (1998). *The Nature of Statistical Learning Theory*, Wiley, New York.
36. Qian, N. & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**, 865–884.
37. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016.
38. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.
39. Fukao, Y., Hayashi, M. & Nishimura, M. (2002). Proteomic analysis of leaf peroxisomal proteins in greening cotyledons of *Arabidopsis thaliana*. *Plant Cell Physiol.* **43**, 689–696.
40. Schafer, H., Nau, K., Sickmann, A., Erdmann, R. & Meyer, H. E. (2001). Identification of peroxisomal membrane proteins of *Saccharomyces cerevisiae* by mass spectrometry. *Electrophoresis*, **22**, 2955–2968.
41. Marcotte, E. M., Xenarios, I., van Der Bliek, A. M. & Eisenberg, D. (2000). Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **97**, 12115–12120.
42. Mott, R., Schultz, J., Bork, P. & Ponting, C. P. (2002). Predicting protein cellular localization using a domain projection method. *Genome Res.* **12**, 1168–1174.
43. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
44. Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Yi, Y. I. W. *et al.* (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucl. Acids Res.* **29**, 2994–3005.
45. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller,

L., Eddy, S. R. *et al.* (2002). The pfam protein families database. *Nucl. Acids Res.* **30**, 276–280.
46. Lundström, J., Rychlewski, L., Bujnicki, J. & Elofsson, A. (2001). Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.* **10**, 2354–2362.
47. Koller, A., Spong, A. P., Luers, G. H. & Subramani, S. (1999). Analysis of the peroxisomal acyl-CoA oxidase gene product from *Pichia pastoris* and determination of its targeting signal. *Yeast*, **15**, 1035–1044.
48. Olivier, L. M. & Krisans, S. K. (2000). Peroxisomal protein targeting and identification of peroxisomal targeting signals in cholesterol biosynthetic enzymes. *Biochim. Biophys. Acta*, **1529**, 89–102.
49. Hayashi, M., Toriyama, K., Kondo, M., Kato, A., Mano, S., De Bellis, L. *et al.* (2000). Functional transformation of plant peroxisomes. *Cell Biochem. Biophys.* **32**, 295–304.
50. Kunau, W. H., Dommes, V. & Schulz, H. (1995). Beta-oxidation of fatty acids in mitochondria, peroxisomes, and bacteria: a century of continued progress. *Prog. Lipid Res.* **34**, 267–342.
51. Graham, I. A. & Eastmond, P. J. (2002). Pathways of straight and branched chain fatty acid catabolism in higher plants. *Prog. Lipid Res.* **41**, 156–181.
52. Farmer, E. E., Weber, H. & Vollenweider, S. (1998). Fatty acid signaling in *Arabidopsis*. *Planta*, **206**, 167–174.
53. Gurvitz, A., Langer, S., Piskacek, M., Hamilton, B., Ruis, H. & Hartig, A. (2000). Predicting the function and subcellular location of *Caenorhabditis elegans* proteins similar to *Saccharomyces cerevisiae* beta-oxidation enzyme. *Yeast*, **17**, 188–200.
54. Hayashi, H., De Bellis, L., Ciurli, A., Kondo, M., Hayashi, M. & Nishimura, M. (1999). A novel acyl-CoA oxidase that can oxidize short-chain acyl-CoA in plant peroxisomes. *J. Biol. Chem.* **274**, 12715–12721.
55. Lensink, M. F., Haapalainen, A. M., Hiltunen, J. K., Glumoff, T. & Juffer, A. H. (2002). Response of SCP-2L domain of human MFE-2 to ligand removal: binding site closure and burial of peroxisomal targeting signal. *J. Mol. Biol.* **323**, 99.
56. Aboushadi, N., Engfelt, W. H., Paton, V. G. & Krisans, S. K. (1999). Role of peroxisomes in isoprenoid biosynthesis. *J. Histochem. Cytochem.* **47**, 1127–1132.
57. Kovacs, W. L., Olivier, L. M. & Krisans, S. K. (2002). Central role of peroxisomes in isoprenoid biosynthesis. *Prog. Lipid Res.* **41**(5), 369–391.
58. Wanders, R., Dekker, C., Hovarth, V. A., Schutgens, R. B., Tager, J. M., Van Laer, P. & Lecoutere, D. (1994). Human alkyldihydroxyacetonephosphate synthase deficiency: a new peroxisomal disorder. *J. Inherit. Metab. Dis.* **17**, 315–318.
59. Hayashi, S., Fujiwara, S. & Noguchi, T. (2000). Evolution of urate-degrading enzymes in animal peroxisomes. *Cell Biochem. Biophys.* **32**, 123–129.
60. Kornberg, H. (1996). The role and control of the glyoxylate cycle in *E. coli*. *Biochem. J.* **99**, 1–11.
61. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.
62. Perantonis, S. J. & Virvilis, V. (2000). Efficient perceptron learning using constrained steepest descent. *Neural Netw.* **13**, 351–364.
63. Joachims, T. (1999). Making large-scale SVM learning practical. In *Advances in Kernel Methods—Support Vector Learning* (Schölkopf, B., Burges, C. & Smola, A., eds), MIT Press, Cambridge, USA.
64. Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (see also pp. 623–656).

Edited by M. Levitt

(Received 30 January 2003; received in revised form 9 April 2003; accepted 12 April 2003)



Supplementary Material for this paper comprising two Tables and three lists of protein sequences is available on Science Direct