

# Identifying Novel Peroxisomal Proteins

John Hawkins<sup>1,2</sup>      Donna Mahony<sup>3</sup>      Stefan Maetschke<sup>1,2</sup>  
Mark Wakabayashi<sup>1,2</sup>      Rohan D Teasdale<sup>3</sup>      Mikael Bodén<sup>2</sup>

December 18, 2006

<sup>1</sup> ARC Centre for Complex Systems.

<sup>2</sup> School of Information Technology and Electrical Engineering

<sup>3</sup> Institute for Molecular Bioscience and ARC Centre in Bioinformatics.

*The University of Queensland, St Lucia, Queensland 4072, Australia.*

Contact: John Hawkins

Email: [jhawkins@itee.uq.edu.au](mailto:jhawkins@itee.uq.edu.au)

PH: +61 7 3365 1636

FAX: +61 7 3365 4999

## Abstract

Peroxisomes are small subcellular compartments responsible for a range of essential metabolic processes. Efforts in predicting peroxisomal protein import are challenged by species variation and sparse sequence data sets with experimentally confirmed localization.

We present a predictor of peroxisomal import based on the presence of the dominant peroxisomal targeting signal one (PTS1), a seemingly well-conserved but highly unspecific motif. The signal appears to rely on subtle dependencies with the preceding residues. We evaluate prediction accuracies against two alternative predictor services, PEROXIP and the PTS1 PREDICTOR. We test the integrity of prediction on a range of prokaryotic and eukaryotic proteomes lacking peroxisomes. Similarly we test the accuracy on peroxisomal proteins known to not overlap with training data. The model identified a number of proteins within the RIKEN IPS7 mouse protein dataset, as potentially novel peroxisomal proteins. Three were confirmed in vitro using immunofluorescent detection of myc-epitope-tagged proteins in transiently transfected BHK-21 cells (Dhrs2, Serhl and Ehhadh).

The final model has a superior specificity to both alternatives, and an accuracy better than PEROXIP and on par with PTS1 PREDICTOR. Thus, the model we present should prove invaluable for labeling PTS1 targeted proteins with high confidence. We use the predictor to screen several additional eukaryotic genomes to revise previously estimated numbers of peroxisomal proteins.

Available at <http://pprowler.itee.uq.edu.au>.

## 1 Introduction

Peroxisomes are relatively small compartments that are particularly abundant in liver cells and neurons. In plants they occur in large numbers as metabolically specialized microbodies. The peroxisome plays an important role in lipid, ethanol and glyoxylate metabolism and detoxification of reactive oxygen species. They are believed to be essential for coping with oxidative stress. Peroxisomal disorders often involve abnormal accumulation of long fatty acids which subsequently impacts membrane structure [34]. The mechanisms of protein localization are one potential avenue to be explored for disease-control [21].

There are numerous mechanisms of subcellular localization, depending on the organelle and the function of the protein. In general they all rely on the existence of some form of targeting signal,

which is recognized by chaperone proteins that then effect the localization [22, 11, 31]. However, the process is complicated by the dynamical aspects of cellular life. Targeting to the endoplasmic reticulum is achieved by an N-terminal signal peptide that is recognized by a signal recognition particle as it emerges from the ribosome. The localization process thus occurs in tandem with translation, and will dominate other functional localization pathways, including those for the peroxisome.

Targeting to the peroxisome only happens after the completion of translation on free ribosomes. The transport of peroxisomal matrix proteins is believed to rely on a small number of sequence motifs. The dominating targeting signal is the PTS1 which appears at the C-terminus. The PTS1 consists of a strongly conserved tri-peptide but there are constraints that range over a larger region exposed to the Pex5 that play a central role in import [24, 17, 3, 20].

In recent efforts, computational methods have been used in genome-wide screens for peroxisome-like sequence features to single-out previously unknown proteins and to subsequently identify novel peroxisomal enzymes and putative regulatory proteins [30]. More broadly, subcellular localization has been approached as a multi-class prediction problem. Models that predict localization of proteins on basis of their amino acid sequences can be used to screen and annotate unknown proteins and to guide experimental efforts to where they are best needed. Several computational approaches have shown promise including manually programmed expert systems and data-driven machine learning models [22, 11, 31]. Machine learning models of protein localization are often based on probabilistic techniques [9, 18, 32], neural networks [13, 5, 10, 4] and support vector machines [16, 7, 26].

Besides general subcellular localization predictors [23, 26, 8], there are few predictors that specialize in predicting peroxisomal targeting: PEROXIP by Emanuelsson, Elofsson, Heijne and Cristóbal [12], and PTS1 PREDICTOR by Neuberger, Maurer-Stroh, Eisenhaber, Hartig and Eisenhaber [24], both address only the PTS1 signal that is located at the C-terminus. Both exclude the nona-peptide motif (PTS2) and membrane-bound peroxisomal proteins (mPTS) due to the lack of experimental sequence data.

PEROXIP has three stages in its processing of candidate sequences. An initial step filters out sequences with predicted signal peptides (to eliminate those proteins co-translationally inserted into the ER) and those sequences predicted to have a membrane spanning region. Secondly, a motif identification module examines the C-terminus of the sequence and filters out those proteins that do not fit a general pattern of approved tri-peptide motifs. In the final step a machine learning model examines the nine amino acids preceding the PTS1 motif and provides a prediction of whether the sequence is peroxisomal or not. In the PEROXIP model the machine learning module consists of a neural network and a support vector machine (SVM) that are trained independently and operate as a miniature ensemble. The final output is the union of their individual outputs, i.e. the sequence is predicted as peroxisomal if either the neural network or the SVM indicates so [12].

Neuberger *et al.* manually formulated a scoring function using a set of sequences experimentally confirmed to interact with the Pex5 complex responsible for PTS1 import. They also analyzed a larger sequence set extracted from public databases. The scoring function that is used by PTS1 PREDICTOR, considers a “composite profile term that evaluates concordance with amino acid preferences” and a set of physical property patterns that were identified from the data sets [24]. Due to some species-specific variation in targeting (via PTS1), three versions were developed: Metazoa, Fungi and General. Both sensitivity and specificity were reportedly very high.

In previously reported simulations we demonstrated that the prediction accuracy of a PEROXIP-like architecture could be improved by including the highly conserved tri-peptide in the window of residues supplied to the machine learning module. The model was making use of inter-dependencies between the tri-peptide and the nine preceding amino acids [33]. The SVM model was refined by investigating the use of an amino acid composition window, and exploring the space of kernel parameters more thoroughly. The final predictor, PTS1PROWLER, utilizes a fifth order polynomial kernel with a separately trained logistic output function [15].

In the present paper we review the structure of our final model and perform a number of analyses to illustrate the accuracy and ability of PTS1PROWLER. We compare PTS1PROWLER against

PEROXIP and PTS1 PREDICTOR using the most recent updates of SWISS-PROT (guaranteed to not overlap with the training data of any model), and by screening proteomes of organisms known not to contain peroxisomes, both eukaryotes and prokaryotes. Furthermore, a selection of eukaryotic proteomes are screened. We report on updated estimates of the number of peroxisomal matrix proteins in each organism. Due to the model’s high specificity, our screens should contain few false positives. Finally we screen the RIKEN IPS7 mouse protein dataset for potentially novel peroxisomal proteins and confirm three of the predictions in vitro using immunofluorescent detection of myc-epitope-tagged proteins in transiently transfected BHK-21 cells.

The PTS1PROWLER prediction service is integrated into the PROTEIN PROWLER suite available at <http://pprowler.itee.uq.edu.au>. Related sequence data is made available at the web site.

## 2 Materials and Methods

### 2.1 Replication of the PeroxiP Data Set

The data set of PTS1-containing peroxisomal proteins and non-peroxisomal proteins, as identified by Emanuelsson *et al.* was extracted from SWISS-PROT R39.27. The initial data set included 152 proteins identified as peroxisomal, and 308 identified as non-peroxisomal. Redundancy reduction had not been performed on this data set. The training and test data was finalized as detailed by Emanuelsson *et al.*

Highly similar proteins were removed such that each pair of proteins differed in at least two positions in the nine residues preceding the C-terminal tripeptide.

The final stage of redundancy reduction was performed using BLASTClust. In order to reproduce a data set of the same size we found that a similarity threshold of 1.675 was required (using BLASTClust’s -s option). **By our estimations this corresponds to between 80-85% similarity over the aligned region, which by default is set to 90% of the length of the sequences.**

The final number of these was 90 peroxisomal proteins and 160 non-peroxisomal proteins. These numbers differ only slightly to the reported data set of 90 and 151 used as training and test data in PEROXIP.

### 2.2 New Data Set

The new data set was collected from SWISS-PROT R45. We followed a similar process to that of PEROXIP to extract an initial set of peroxisomal sequences. All SWISS-PROT entries were searched for those containing a "SUBCELLULAR LOCATION" annotation in the comments field that included any of "PEROXISOM", "GLYOXYSOM", or "GLYCOSOM" using a case-insensitive search. The proteins were also required to identify a "Microbody targeting signal" in the feature table, indicating a PTS1 targeted protein. This gave an initial set of 202 proteins.

The proteins found were then filtered manually for proteins not likely to be targeted by a PTS1 and for membrane proteins. Full details of the curation process are available in the online documentation of the predictor.

The negative set was generated by extracting from SWISS-PROT R45 all eukaryotic proteins with a C-terminal tripeptide identical to that of one of the initially identified peroxisomal proteins, and a subcellular localization not specified as peroxisomal, glyoxysomal, or glycosomal. These proteins were then manually curated for potentially fallacious annotations, full details of which are given in the online documentation.

The final data set contained 206 peroxisomal proteins and 348 non-peroxisomal proteins. Redundancy reduction was performed by similar processes to those performed on the PEROXIP data set (above). Because the predictor was trained and tested on the complete C-terminal twelve residues, these residues were tested across both the positive and negative sets to ensure no pairs of sequences had less than two differing residues. This reduced the data set to 157 and 239 proteins for the peroxisomal and non-peroxisomal sets, respectively. Clustering of the remaining sets using BLASTClust with a similarity threshold of 1.675 **(80-85% similarity over the aligned region**

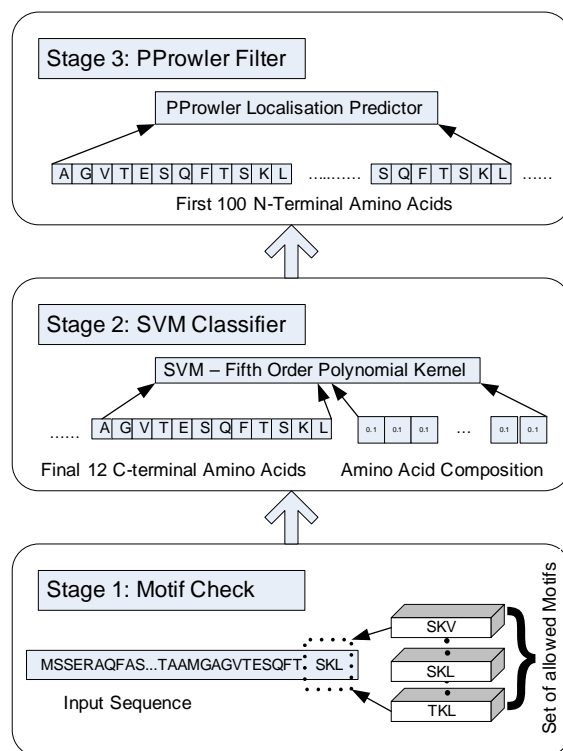


Figure 1: Overview of the three stages of the final PTS1PROWLER model. The first stage filters out sequences without a PTS1 motif. The second stage runs the SVM classifier over the final 12 residues with an orthonormal encoding and an amino acid composition window. If this stage produces a positive prediction, then the third stage is executed, by which the PROTEIN PROWLER application examines the sequence for a potential Signal Peptide. If the PROTEIN PROWLER application gives a score greater than 0.85 then the prediction is changed to non-peroxisomal, otherwise the output from stage two is given.

of, minimum 90% of the sequence) and extraction of representatives from each cluster (the same procedure used to replicate Emanuelsson’s data set) resulted in a final testing and training set of 124 peroxisomal proteins and 214 non-peroxisomal proteins. The ratio of positives and negatives was therefore similar to those of PEROXIP’s set, with the overall size of the data set increased by approximately 40%.

## 2.3 PTS1Prowler Development

As shown in Figure 1, the overall structure of PTS1PROWLER is similar to that of PEROXIP. Processing occurs in the three distinct stages. An initial motif filter rejects sequences with a C-terminal tripeptide not occurring amongst peroxisomal proteins in SWISS-PROT R45. A machine learning module analyses the final 12 residues of the sequence and gives probabilistic prediction as to whether the protein is PTS1 targeted. Finally a signal peptide prediction application is used to filter proteins that are likely to be secreted.

### 2.3.1 Performance Metrics

Performance is evaluated using several metrics. The Matthews correlation coefficient [19] (MCC) is a performance statistic that takes into account the numbers of true positives ( $tp$ ), true negatives

( $tn$ ), false positives ( $fp$ ) and false negatives ( $fn$ ). It is calculated as follows:

$$\frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fn)(tp + fp)(tn + fp)(tn + fn)}} \quad (1)$$

The *sensitivity* is a measure of the classifiers predilection for always identifying positive examples, calculated thus:

$$\frac{tp}{tp + fn} \quad (2)$$

The *specificity* of the classifier is a measure of the classifier’s tendency to avoid making false positive predictions, calculated as follows:

$$\frac{tn}{tn + fp} \quad (3)$$

### 2.3.2 The Machine Learning Module

Perhaps the most crucial stage of the model is the machine learning module which decides whether a potential sequence with a valid PTS1 motif, is genuinely a peroxisomal protein. This component has been the focus of our development efforts.

It has been shown that essential sequence dependencies occur within the last twelve residues of PTS1 targeted proteins, in particular between the PTS1 motif and the preceding 9 residues [24, 1]. In our initial studies of this problem we established that the machine learning module could be improved considerably by allowing it to learn the dependencies between the PTS1 motif and the preceding 9 residues [33].

The PEROXIP predictor was reported to have employed an amino acid composition window, as well as the 9-mer preceding the PTS1 motif, as input to the machine learning module [12]. The amino acid composition encoding is formed by summing the number of occurrences of each amino acid in the entire protein, and then producing a probability vector with each element giving the likelihood that a random position in the sequence would contain the corresponding amino acid. The amino acid composition has been shown to be at least correlated crudely with localization [25], and had been used as an encoding for machine learning prediction of localization to various subcellular locations (including the peroxisome), with moderate success [29]. In a previous study we established that the contribution of the amino acid composition to SVM classification varied with kernel choice, however for **all but one** kernel (Gaussian  $\gamma = 0.01$ ) it improved performance, with an average improvement of 0.07 points of MCC across a range of kernels [15].

By default a support vector machine will produce a binary classification on a two class problem. For many reasons it is desirable that a classifier service produces a posterior probability score for a given input. This can be achieved with SVMs by fitting a logistic model to the output of the trained classifier [28]. For this model the output probability of the target class is given by:

$$P(y = 1|f(x)) = \frac{1}{1 + e^{Af(x)+B}} \quad (4)$$

Where  $f(x)$  is the output of the SVM for sample  $x$ , and the parameters  $A$  and  $B$  must be estimated.

In the previous study we also established that the use of the logistic output function not only failed to impede the performance of the model; it improved performance for almost all kernel choices. The capacity for logistic outputs to improve performance was noted by Platt, but is deemed to be difficult to predict in advance [28]. Following Platt’s recommendation and examples, we used an internal cross validation within the training set (three fold) to estimate the parameters for the logistic function [28]. This means that two thirds of the training set are used for training the SVM and the final third is used for estimating the parameters of the logistic function via the maximum likelihood method.

The final implementation of PTS1PROWLER is the product of all these insights into peroxisomal protein prediction. We use a fifth order polynomial kernel without lower order terms and a logistic

output function. The input is given via two windows one containing the final twelve residues and another containing the amino acid composition. Following previous results we use the orthonormal encoding for the primary sequence window [12, 33]. The optimization procedures for adapting the SVMs and the logistic output models are drawn from Weka library of machine learning tools [35].

In our exploration of the various free parameters available to fine tune the SVM we found that the the complexity constant  $C$  and the round-off error  $\epsilon$  had very little effect on the results. Only at extreme values did the performance of the model change, and then only in a deleterious fashion. Minor improvements were achieved by tuning the tolerance parameters so that the average MCC increased from 0.741 to 0.749, with a final value of  $T = 0.038$ . The insensitivity to the setting of regularization parameters indicate that the separation in the feature space defined by the polynomial kernel is simple. However it is worth noting that due to computation time constraints it was assumed that optimal values for these parameters could be established independently of one another.

### 2.3.3 Filtering Signal Peptides

The final stage of the model filters out proteins that possess a signal peptide, because they will not have an opportunity to interact with Pex5 even if they have a functional PTS1.

In PTS1PROWLER the filtering of secreted proteins is performed by the PROTEIN PROWLER sub-cellular localization model [4]. However, in our model this step is performed last rather than first. Its use is deferred as it is more computationally expensive than either the motif filter or the SVM driven classification. The reordering has no effect on the classification, only the efficiency of the model.

Running PROTEIN PROWLER over the training sets revealed only one peroxisomal protein P24552 (D-amino-acid oxidase) with a significant prediction of containing a signal peptide (0.832). A threshold value of 0.85 was chosen as the cutoff for the PROTEIN PROWLER filter stage, yielding no predictions from the positive set and 49 predictions from the negative training set.

Because no peroxisomal proteins are predicted as secreted by the PROTEIN PROWLER filter, the sensitivity of the model does not decline by including the filter. The PROTEIN PROWLER filter increased specificity by an average of 0.020 (an increase of 2%) and the MCC by an average of 0.017 (an increase of 2%). Decreasing the threshold could further increase the gain in accuracy, but may result in reduced sensitivity when exposed to a larger data set.

### 2.3.4 Training and Testing

The entire model was trained and tested using the jack knife procedure, in which the data set is separated into as many sets as there are samples. In each a set a single example is set aside for testing. **The procedure involves creating a model for each set, training each model separately and then testing on the single test case. All of the test results are then used to produce an overall estimate of the performance of the technique on the problem.** Usually this testing procedure precludes the use of multiple runs and the reporting of standard deviations, simply because the data can be split only one way and an SVM (unlike a neural net) is not sensitive to initialization conditions and will thus perform identically on repeated exposures to the same data. However, the use of the logistic output function, requires that the training data be split between that used to train the SVM and that used to fit the logistic model. Hence, our jack knife simulations are performed five times with different seeds and we report mean value of all metrics over the five runs.

## 2.4 In Vitro Peroxisome Subcellular Localization Assay

N-terminally tagged myc-gene of interest expression constructs were generated using a modified overlapping PCR methodology [2]. Generated expression constructs comprised of a promoter fragment, the cDNA of interest and a terminator fragment. The PCR primers, reactions and amplification conditions used have been reported previously [2]. *BHK* – 21 (Syrian golden hamster. kidney

**fibroblasts, ATCC CCL 10**) cells were cultured in Dulbecco's Modified Essential Media (DMEM, Life Technologies Inc., Grand Island, NY, USA) supplemented with 10% fetal bovine serum (Cambrex, NJ, USA) and 2 mM L-glutamine (Life Technologies Inc) and maintained in a humid, 5% CO<sub>2</sub> environment at 37°C. Sub-confluent BHK-21 cells cultured on glass coverslips coated with 0.01% poly-L-lysine (Sigma Aldrich St Louis, MO, USA) were transiently transfected with PCR expression constructs diluted with OptiMEM (Life Technologies Inc.) using Lipofectamine2000 (Life Technologies Inc.) as per the manufacturer's instructions. Twenty hours post-transfection BHK-21 cells were fixed in 4% paraformaldehyde (Sigma Aldrich) in phosphate-buffered saline (PBS) for 20 minutes prior to permeabilization in 0.1% Triton-X 100 (Sigma Aldrich) in PBS for 5 minutes followed by 3 washes in PBS. Blocking was performed in 2% bovine serum albumin (BSA, Sigma Aldrich in PBS) prior to incubation with the monoclonal anti-myc antibody (Cell Signaling Technology, Beverly, MA, USA, 1:3000 in 2% BSA) and rabbit anti-catalase (Abcam, Cambridge, UK, 1:2000 in 2% BSA) for 60 minutes at room temperature. The cells were subsequently washed in PBS prior to incubation with the secondary antibodies goat anti-mouse IgG-cy3 conjugated antibody (Zymed, San Francisco, USA, diluted 1:600 in 2% BSA) and goat anti-rabbit IgG-Alexa-488 antibody (Molecular Probes Inc. Eugene, OR, USA) for 45 minutes at room temperature and the coverslips were mounted using MO-WIOL (Calbiochem, Nottingham, UK). Representative images of the observed localization patterns were captured for each construct using an Olympus AX-70 upright fluorescence microscope. Images were prepared using Adobe Photoshop CS2 (Adobe Systems, USA).

## 2.5 Latest Peroxisomal Proteins

To perform an independent test of the generalization of the model compared to its competitors we extracted a set of peroxisomal proteins not present in the training set.

The protein sequences were extracted from SWISS-PROT R48 (excluding TREMBL) filtered such that the SUBCELLULAR LOCATION line of the CC field ended with "Peroxisomal." These proteins were then filtered for the version addition identifiers "46, Created", "47, Created" "48, Created". Proteins meeting these criteria were further filtered for a MOTIF entry in the feature table (FT) containing the string "Microbody targeting signal". This resulted in a final set of 17 protein sequences.

## 2.6 Eukaryotic Genome Screens

The data used for the eukaryotic genome screens was obtained from a number of different sources. In each case we used the data as it was obtained without any filtering. The URLs for each proteome file are listed below.

- *S. pompe*  
ftp.sanger.ac.uk/pub2/yeast/pombe/Protein\_data/pompep
- *S. cerevisiae*  
ftpmips.gsf.de/yeast/sequences/Scerevisiae\_prot\_2006200
- *D. melanogaster*  
ftp.ebi.ac.uk/pub/databases/edgp/sequence\_sets/aa\_gadfly.dros.Z
- *C. elegans*  
ftp.sanger.ac.uk/pub2/wormbase/sequences/WORMPEP/wormpep
- *A. thaliana*  
ftpmips.gsf.de/cress/arabiprot/arabi\_all\_proteins\_v020204.gz
- *M. musculus*  
ftp.ensembl.org/pub/current\_mouse/data/fasta/pep/Mus\_musculus.NCBIM34.jul.pep.fa.gz

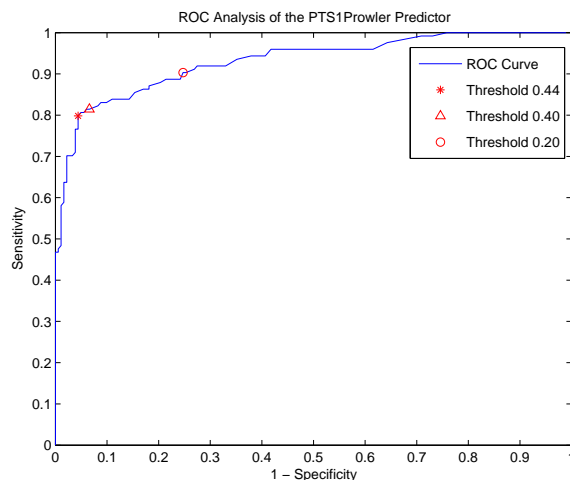


Figure 2: A Receiver Operating Characteristic curve of the performance of PTS1PROWLER. We have marked the location of three different output thresholds. At the threshold 0.44 the predictor is operating with maximal MCC. Using 0.4 gives a slight increase in Sensitivity at the cost of Specificity. Finally 0.2 raises the Sensitivity of the predictor to 0.9.

- *H. sapiens*  
[ftp.ncbi.nih.gov/genomes/H\\_sapiens/protein/Gnomon\\_prot.fsa.gz](ftp.ncbi.nih.gov/genomes/H_sapiens/protein/Gnomon_prot.fsa.gz)
- *O. sativa*  
[http://www.tigr.org/tdb/e2k1/osa1/data\\_download.shtml](http://www.tigr.org/tdb/e2k1/osa1/data_download.shtml)

## 3 Results

### 3.1 PTS1Prowler

Emanuelsson *et al.* published values of 0.50, 0.78, and 0.64 for MCC, sensitivity and specificity respectively (values were obtained from a single run of five-fold cross-validation. When trained on the same data set, PTS1PROWLER yielded an average MCC of 0.76. With the updated data set the average MCC is 0.77, which is a 54% improvement on PEROXIP. It is interesting to note that the sensitivity of PTS1PROWLER is identical to that of the original PEROXIP (0.78), all of the improvement in performance has come from an increase of 45% in the specificity of the model (0.93).

An ROC analysis was performed using the average of the outputs from the five simulations (see Figure 2). The ROC analysis revealed that the optimal threshold for a positive prediction is around 0.44 yielding an MCC of 0.78. In the name of making a conservative estimate we do not use this as the final estimate of MCC, but we recommend the threshold value for users of PTS1PROWLER.

### 3.2 The Latest Peroxisomal Protein Test

In order to perform an independent test of the sensitivity of PTS1PROWLER we procured the set of peroxisomal proteins added to the SWISS-PROT database since the development of the training sets. The 17 sequences that were identified, were presented to PTS1PROWLER, PEROXIP and PTS1 PREDICTOR. Results are shown in Table I.

For PEROXIP we report only their Method 1, as all other methods involve relaxation of constraints seriously compromising specificity, and as such do not offer a reasonable measure of the PEROXIP's accuracy.

Novel-Peroxisomal Protein Prediction

| Protein     | PEROXIP<br>Method 1 | PTS1 PREDICTOR |              | PTS1PROWLER |                |                 |
|-------------|---------------------|----------------|--------------|-------------|----------------|-----------------|
|             |                     | Raw            | $\theta = 0$ | Raw         | $\theta = 0.5$ | $\theta = 0.22$ |
| NDX8_CAEEL  | —                   | 9.17           | ✓            | 0.71        | ✓              | ✓               |
| SUOX_ARATH  | —                   | -3.06          | —            | 0.41        | —              | ✓               |
| PECR_CAVPO  | —                   | 6.63           | ✓            | 0.95        | ✓              | ✓               |
| DECR2_MOUSE | —                   | 4.39           | ✓            | 0.34        | —              | ✓               |
| PECR_HUMAN  | —                   | 6.90           | ✓            | 0.88        | ✓              | ✓               |
| OPR3_LYCES  | —                   | 11.88          | ✓            | 0.40        | —              | ✓               |
| OPR3_ARATH  | ✓                   | 9.46           | ✓            | 0.69        | ✓              | ✓               |
| URIC_MACFA  | ✓                   | 14.13          | ✓            | 1.00        | ✓              | ✓               |
| NUD12_HUMAN | —                   | -17.77         | —            | 0.00        | —              | —               |
| URIC_AOTTR  | ✓                   | 13.72          | ✓            | 1.00        | ✓              | ✓               |
| URIC_MACMU  | ✓                   | 13.72          | ✓            | 1.00        | ✓              | ✓               |
| DECR2_RAT   | —                   | 3.53           | ✓            | 0.61        | ✓              | ✓               |
| DCMC_RAT    | ✓                   | 4.56           | ✓            | 1.00        | ✓              | ✓               |
| ACOX_PICPA  | —                   | 3.33           | ✓            | 0.00        | —              | —               |
| NUDT7_MOUSE | ✓                   | -12.22         | —            | 0.69        | ✓              | ✓               |
| DCMC_MOUSE  | ✓                   | 2.47           | ✓            | 1.00        | ✓              | ✓               |
| ACOX4_ARATH | ✓                   | 11.03          | ✓            | 0.63        | ✓              | ✓               |
| Total       | 8                   |                | 14           |             | 12             | 15              |
| Accuracy    | 47%                 |                | 82%          |             | 71%            | 88%             |

Table 1: The prediction results for the three PTS1 predictors on experimentally verified PTS1 targeted proteins released since R45 until R48 of SWISS-PROT. When applicable, classifications are made on basis of the predictors' raw output according to the specified thresholds  $\theta$  (PTS1 PREDICTOR default  $\theta = 0$  and PTS1PROWLER default  $\theta = 0.5$ , compensated  $\theta = 0.22$ ).

| Genus       | Proteins | PEROXIP | PTS1PROWLER |
|-------------|----------|---------|-------------|
| Trichomonas | 15       | 0       | 0           |
| Giardia     | 35       | 2       | 0           |
| Entamoeba   | 55       | 0       | 0           |

Table 2: PEROXIP and PTS1PROWLER applied to all available proteins in Swiss-Prot Release 47.5 from three genera of eukaryotes for which there is no evidence of peroxisomes. The number of available proteins is displayed in the first column, followed by the number of proteins predicted to be peroxisomal by PEROXIP and PTS1PROWLER.

On data used for model development, PTS1 PREDICTOR has been reported to have a much higher sensitivity than PEROXIP or PTS1PROWLER. However, the specificity of PTS1 PREDICTOR was not established. The ROC analysis on PTS1PROWLER enables the determination of an adjusted threshold that on the data set used for development gives the same sensitivity, 0.896, as reported for PTS1 PREDICTOR (general version, henceforth PTS1 PREDICTOR-General). With the adjustment the accuracy of PTS1PROWLER on the novel PTS1 sequences improves from 12/17 (with 0.5 as threshold) to 15/17 (with 0.22 as threshold), one better than PTS1 PREDICTOR-General.

Interestingly, PTS1 PREDICTOR and PTS1PROWLER differ in their predictions. By using the logical OR-function on the output of the two, **all but one** of the proteins are predicted as PTS1 targeted, the elusive protein is NUD12\_HUMAN protein which was not predicted as peroxisomal by any of the three. The terminal tripeptide of this protein is PNL, which is not included in the acceptance set for either PEROXIP or PTS1PROWLER, hence it is automatically rejected by both classifiers.

### 3.3 Organisms Without Peroxisomes

Although peroxisomes are present in most eukaryotic cells, there are numerous species of parasitic eukaryotes which are believed not to possess a peroxisome. For Protozoa of the genus Giardia, Trichomonas (agent of Vaginitis), or Amoebas of the genus Entamoeba (agents of dysentery and ulceration of the colon and liver) there is no evidence for the presence of a peroxisome [27]. Thus, their proteomes should not contain proteins with a valid PTS1 except by mere chance.

The proteomes of peroxisome-less organisms offer an opportunity to test the integrity of classifiers. By feeding the set of known proteins from each genus through both PEROXIP and PTS1PROWLER we report how many proteins are fallaciously labeled peroxisomal, thus giving an independent estimate of the specificity of the models. **The data sets for this study were retrieved from SWISS-PROT R47.5 (within each taxonomy, excluding protein fragments)**. For PEROXIP we conservatively use only Method 1 predictions, in which the most restrictive criteria are used. For PTS1PROWLER we use the neutral cutoff of 0.5.

The results displayed in Table II reinforce the findings of the preceding simulations. **PEROXIP identified 2 proteins as peroxisomal, a false positive rate of 1.9%, whereas PTS1PROWLER identified none.** Due to restrictions in the number of sequences that may be presented to the PTS1 PREDICTOR service, we were unable to easily test PTS1 PREDICTOR this way.

To provide an estimate of the accuracy of PTS1PROWLER in relation to PTS1 PREDICTOR, PTS1PROWLER was tested on the entire proteomes of a range of prokaryotic organisms, all of which were screened previously by PTS1 PREDICTOR [24]. This test provides an additional assessment of the predictors' relative specificity due to the fact that prokaryotes completely lack subcellular compartments and, consequently, have no selective conservation of peroxisomal targeting signals.

The results of the PTS1PROWLER tests are shown in Table III. The collection of protein sequences were downloaded from the NCBI whole genome retrieval site.

Neuberger *et al.* report the total number of false positives for the same genomes (but older ver-

| Species  | No of Proteins | Predictions | FP Rate |
|--|----------------|-------------|---------|
| <i>Escherichia coli</i> K12                            | 5,379          | 7           | 0.13    |
| <i>Pseudomonas aeruginosa</i> PA01                     | 5,567          | 2           | 0.04    |
| <i>Helicobacter pylori</i> -strain J99                 | 1,491          | 1           | 0.07    |
| <i>Staphylococcus epidermidis</i> ATCC 12228           | 2,419          | 3           | 0.12    |
| <i>Staphylococcus aureus</i> subsp. <i>aureus</i> N315 | 2,588          | 2           | 0.08    |
| <i>Mycobacterium tuberculosis</i> CDC1551              | 4,189          | 2           | 0.05    |
| Totals   | 21,633         | 17          | 0.08    |

Table 3: Number of Prokaryotic Proteins within each of the selected organisms predicted to be peroxisomal by PTS1PROWLER.

| Positive Predictions |            |            | Negative Predictions |            |            | Total Proteins |
|----------------------|------------|------------|----------------------|------------|------------|----------------|
| Peroxi               | Non-peroxi | Unknown SL | Peroxi               | Non-peroxi | Unknown SL |                |
| 33                   | 13         | 48         | 38                   | 12452      | 20938      | 33,451         |

Table 4: The results of screening the RIKEN IPS7 data set with PTS1PROWLER.

sions with a slightly lower total) for their three predictors as follows: Metazoa 70/20,544 (0.34%), Fungi 39/20,544 (0.19%), and General 152/20,544 (0.74%). Thus, with a false positive rate on prokaryotic data of 0.08% our predictor has superior specificity to all three of the PTS1 PREDICTOR models [24]. However, the sensitivity of PTS1 PREDICTOR is given as 0.896, significantly higher than our model using the default threshold. In order to obtain a more objective comparison of the two models, we again adjusted the threshold (0.22) to achieve a comparable sensitivity. With the adjusted threshold PTS1PROWLER yielded an additional 35 false positives, taking the total to 52. This gives a false positive rate of 0.24%, which is still superior to PTS1 PREDICTOR-Metazoa and PTS1 PREDICTOR-General.

### 3.4 Experimental Validation

The model was used to screen the entire RIKEN IPS7 mouse protein dataset, a high quality proteome derived exclusively from full-length transcripts [6, 14]. This data set was chosen for the selection of candidates for experimental validation. The results of the screen are shown in Table IV. The results are broken into three categories based on the entries in the LOCATE database (as of 05 May 2006) [14]. The label *Peroxi* is given for proteins that have some evidence of being peroxisomal, either by direct experiment or through similarity. *Non-peroxi* applies to proteins that have some evidence that they are located elsewhere. Finally *Unknown* is given to proteins for which there is no available information.

Of the 48 proteins of no known subcellular localization, five were chosen for experimental validation. Mammalian expression constructs were engineered to encode the full-length ORF of the five proteins. The subcellular localization of the myc-epitope-tagged proteins was determined using immunofluorescent detection of the myc epitope in transiently transfected BHK-21 cells.

Three of the proteins (PA113040.1 Dhrr2, PA226887.1 Serhl, PA74692.8 Ehhadh) localized to distinct cytoplasmic puncta (Figure 3 A, D and G). Co-staining with a rabbit anti-catalase antibody (Figure 3 B, E and H) that detects endogenous peroxisome protein confirmed these three proteins are localized to the peroxisome (Figure 3 C, F and I).

Two of the constructs computationally predicted to be peroxisomal proteins were found to have subcellular localizations outside of the peroxisome (data not shown). One protein (PA103003.1 Zfp438) localized to the nucleus and protein (PA111509.2 Hmgcl) showed cytoplasmic localization. The former was only weakly predicted as peroxisomal (0.52) has since been identified as containing

Eukaryotic Genome Screens for Peroxisomal Proteins

| Species                | PEROXIP  |           |            | PTS1PROWLER |           |            |
|------------------------|----------|-----------|------------|-------------|-----------|------------|
|                        | Proteins | Predicted | Percentage | Proteins    | Predicted | Percentage |
| <i>S. pompe</i>        | 4,962    | 10        | 0.20%      | 4,990       | 6         | 0.12%      |
| <i>S. cerevisiae</i>   | 6,449    | 27        | 0.42%      | 6,720       | 21        | 0.31%      |
| <i>D. melanogaster</i> | 13,729   | 58        | 0.42%      | 14,080      | 48        | 0.34%      |
| <i>C. elegans</i>      | 20,414   | 61        | 0.28%      | 22,730      | 60        | 0.26%      |
| <i>A. thaliana</i>     | 25,826   | 61        | 0.24%      | 26,639      | 62        | 0.23%      |
| <i>M. musculus</i>     | 28,097   | 59        | 0.21%      | 36,471      | 65        | 0.18%      |
| <i>H. sapiens</i>      | 38,051   | 44        | 0.12%      | 37,605      | 59        | 0.16%      |
| <i>O. sativa</i>       | 41,915   | 102       | 0.24%      | 61,250      | 95        | 0.16%      |

Table 5: The number of proteins within each of the eight selected eukaryotic proteomes that were predicted to be peroxisomal by PEROXIP and PTS1PROWLER.

a zinc finger motif. The later protein has been generally predicted to be mitochondrial, hence poses something of a general problem for predictors. Overall the PTS1PROWLER application enabled the productive identification of novel peroxisomal proteins within the mouse proteome.

### 3.5 Eukaryotic Proteome Tests

The proteomes of the eight eukaryotic species studied by Emanuelsson *et al.* were presented to PTS1PROWLER to provide an updated estimate of the number and proportion of peroxisomal proteins in these organisms (see Table V and Figure 4).

As an aside, Emanuelsson *et al.* report on an unexpected discrepancy between numbers found in human and mouse believed to stem from C-terminal annotation inaccuracies. We do not observe such a pronounced dip in the plot of predictions made by PTS1PROWLER (see Figure 4).

## 4 Conclusion

We have taken the overall three stage classifier design of Emanuelsson *et al.* and focused on improving the machine learning stage of the process. The final model is a finely tuned support vector machine with a polynomial kernel of order five. Unlike Emanuelsson we include the PTS1 tri-peptide in the input window and additionally we train the model to produce a probability by fitting a logistic function to the output of the SVM. The combination of all of these modifications resulted in a predictor with an average MCC of 0.77, an improvement of 54% on the results reported for the original PEROXIP predictor. ROC curve analysis reveals that the optimum cutoff value for accepting a protein as peroxisomal occurs at about 0.44. To achieve a comparable sensitivity to that reported for PTS1 PREDICTOR, a threshold around 0.22 should be used.

We have performed numerous analyses of the performance of our model. **Notably, by testing on eukaryotic data for organisms without a peroxisome, we found that our model made no false positives, whereas PEROXIP operated with a false positive rate of 1.9%. We used prokaryotic proteomes to replicate the tests in Neuberger and colleagues' published benchmarks. Using the optimal output threshold our model operated with significantly fewer false positives than any of the three models in the PTS1 PREDICTOR suite.** Both Emanuelsson *et al.* and Neuberger *et al.* compare their models against general predictors like PSORT and report that such are inferior to the specialized predictors.

We test the sensitivity of our model using the very latest peroxisomal proteins from SWISS-PROT and found that using the default threshold PTS1PROWLER is significantly better than PEROXIP, but slightly worse than PTS1 PREDICTOR. However, with the adjusted threshold (matching PTS1 PREDICTOR's reported sensitivity of 0.896), it found one more positive than PTS1 PREDICTOR. Importantly, this adjustment did not compromise specificity to a great extent. The

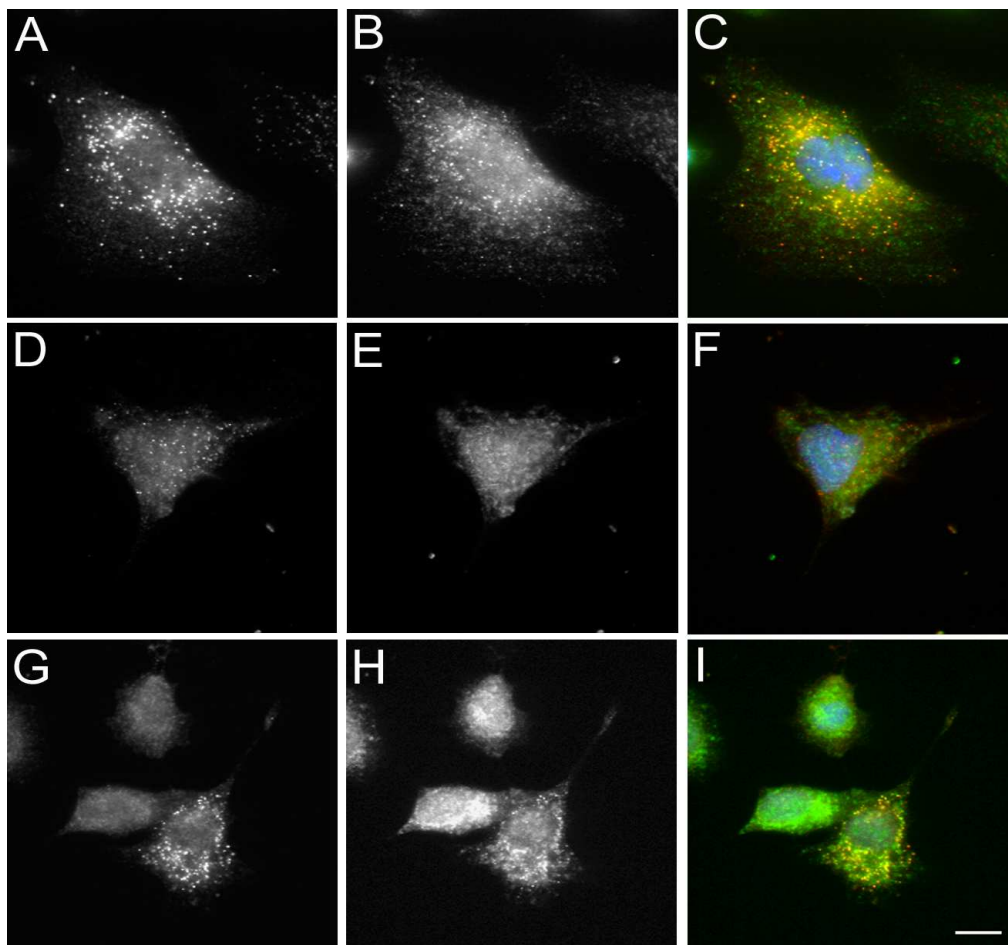


Figure 3: Experimental validation of predicted peroxisomal proteins. Sub-confluent BHK-21 cells were transiently transfected with N-terminal, myc-epitope-tagged expression constructs generated by overlapping PCR. These constructs were expressed for 20 h prior to fixation, immunodection and subsequent visualization as described in the Materials and Methods. Three over-expressed proteins with previously unknown subcellular localization showed localization to distinct cytoplasmic puncta (A, D and G). Co-labeling with a catalase antibody (B, E and H) revealed co-localization to peroxisomal structures (merged images C, F and I). The scale bar represents  $10\mu\text{m}$ .

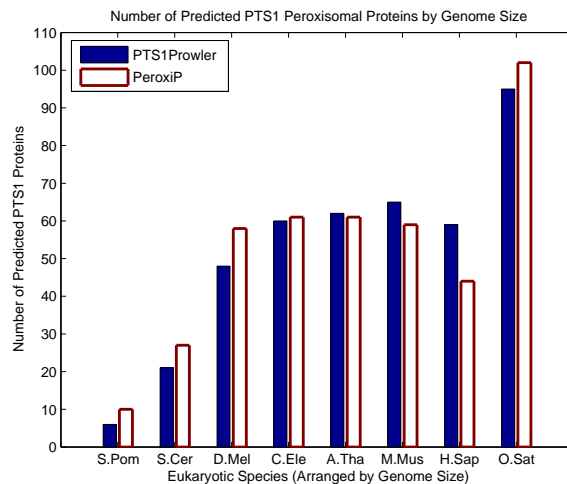


Figure 4: Eukaryotic Proteome Predictions. Graph of the number of PTS1PROWLER and PEROXIP predicted peroxisomal matrix proteins for a range of eukaryotic organisms sorted by proteome size.

comparison on false positive rates on prokaryotic sequence data showed that the lowered threshold only brought PTS1PROWLER’s false positive rate to 0.24%, **still below PTS1 PREDICTOR-General’s 0.74% and PTS1 PREDICTOR-Metazoa’s 0.34%, however, slightly worse than PTS1 PREDICTOR-Fungi’s 0.19%**. It should be noted that the data set of new peroxisomal proteins is very small compared to those used for the specificity tests, hence this test has weaker significance.

We ran our model on the RIKEN IPS7 mouse protein dataset to test for novel peroxisomal proteins. Five proteins of unknown subcellular localization were identified as potentially peroxisomal and were selected for experimental validation. Their subcellular localization was determined using immunofluorescent detection of the myc-epitope-tagged proteins and three were confirmed to be novel peroxisomal proteins.

Finally we presented our classifier with the data sets of a number of eukaryotic proteomes in order to provide an update on the numbers of peroxisomal proteins found therein. The absolute numbers were much on par with that produced by the original PEROXIP study, even though the number of sequences in these data sets has increased. In general we found that our model has downgraded the number of predictions as a function of proteome size and predicts a smoother relationship between proteome size and number of peroxisomal proteins.

## Acknowledgements

This work was supported by the funds from the Australian Research Council, through the ARC Centre for Complex Systems and the ARC Centre for Bioinformatics, and the Australian National Health and Medical Research Council; R.D.T. is supported by an NHMRC R. Douglas Wright Career Development Award.

## References

- [1] L. Amery, M. Fransen, K. De Nys, G. P. Mannaerts, and P. P. Van Veldhoven. Mitochondrial and peroxisomal targeting of 2-methylacyl-coa racemase in humans. *J. Lipid Res.*, 41(11):1752–1759, 2000.
- [2] R. N. Aturaliya, J. L. Fink, M. J. Davis, M. S. Teasdale, K. A. Hanson, K. C. Miranda, A. R. R. Forrest, S. M. Grimmond, H. Suzuki, M. Kanamori, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, and R. D. Teasdale. Subcellular localization of mammalian type II membrane proteins. *Traffic*, 7(5):613–625, 2006.
- [3] A. Baker and I. A. Sparkes. Peroxisome protein import: some answers, more questions. *Current Opinion in Plant Biology*, 8(6):640–647, 2005.

- [4] M. Bodén and J. Hawkins. Prediction of subcellular localisation using sequence-biased recurrent networks. *Bioinformatics*, 21:2279–2286, 2005.
- [5] Y.-D. Cai, X.-J. Liu, and K.-C. Chou. Artificial neural network model for predicting protein subcellular location. *Computers & Chemistry*, 26(2):179–182, 2002.
- [6] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, and S. e. a. Batalov. The Transcriptional Landscape of the Mammalian Genome. *Science*, 309(5740):1559–1563, 2005.
- [7] K. C. Chou and Y. D. Cai. Using functional-domain composition and support vector machines for prediction of protein subcellular location. *Journal of Biological Chemistry*, 29:45765–45769, 2002.
- [8] K. C. Chou and Y. D. Cai. Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *Journal of Cellular Biochemistry*, 90(6):1250–1260, 2003.
- [9] A. Drawid and M. Gerstein. A bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *Journal of Molecular Biology*, 301(4):1059–1075, 2000.
- [10] J. Dyrlov Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak. Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*, 340(4):783–795, 2004.
- [11] O. Emanuelsson. Predicting protein subcellular localisation from amino acid sequence information. *Briefings in Bioinformatics*, 3(4):361–376, 2002.
- [12] O. Emanuelsson, A. Elofsson, G. von Heijne, and S. Cristobal. In silico prediction of the peroxisomal proteome in fungi, plants and animals. *Journal of Molecular Biology*, 330(2):443–456, 2003.
- [13] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence,. *Journal of Molecular Biology*, 300(4):1005–1016, 2000.
- [14] J. L. Fink, R. N. Aturaliya, M. J. Davis, F. Zhang, K. Hanson, M. S. Teasdale, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, and R. D. Teasdale. LOCATE: a mouse protein subcellular localization database. *Nucl. Acids Res.*, 34(1):D213–217, 2006.
- [15] J. Hawkins and M. Bodén. Predicting peroxisomal proteins. In *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 469–474, Piscataway, November 2005. IEEE.
- [16] S. J. Hua and Z. R. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, 2001.
- [17] G. Lametschwandtner, C. Brocard, M. Fransen, P. Van Veldhoven, J. Berger, and A. Hartig. 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. *Journal of Biological Chemistry*, 273(50):33635–33643, 1998.
- [18] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, 20(4):547–556, 2004.
- [19] B. W. Matthews. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, 405:442–451, 1975.
- [20] P. Michels, J. Moyersoen, H. Krazy, N. Galland, M. Herman, and V. Hannaert. Peroxisomes, glyoxysomes and glycosomes. *Molecular Membrane Biology*, 22(1 - 2):133–145, 2005.
- [21] J. Moyersoen, J. Choe, E. Fan, W. G. Hol, and P. A. Michels. Biogenesis of peroxisomes and glycosomes: trypanosomatid glycosome assembly is a promising new drug target. *FEMS Microbiology Reviews*, 28(5):603–643, 2004.
- [22] K. Nakai. Protein sorting signals and prediction of subcellular localization. *Advances in Protein Chemistry*, 54:277–344, 2000.
- [23] K. Nakai and P. Horton. PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Sciences*, 24(1):34–35, 1999.
- [24] G. Neuberger, S. Maurer-Stroh, B. Eisenhaber, A. Hartig, and F. Eisenhaber. Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. *Journal of Molecular Biology*, 328(3):567–579, 2003.
- [25] K. Nishikawa, Y. Kubota, and T. Ooi. Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. *Journal Of Biochemistry*, 94(3):981–995, 1983.
- [26] K.-J. Park and M. Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, 19(13):1656–1663, 2003.
- [27] M. Parsons, T. Furuya, S. Pal, and P. Kessler. Biogenesis and function of peroxisomes and glycosomes. *Molecular and Biochemical Parasitology*, 115(1):19–28, 2001.
- [28] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 2000.

- [29] A. Reinhardt and T. Hubbard. Using neural networks for prediction of the subcellular location of proteins. *Nucl. Acids Res.*, 26(9):2230–2236, 1998.
- [30] S. Reumann, C. Ma, S. Lemke, and L. Babujee. Araperox. a database of putative arabidopsis proteins from plant peroxisomes. *Plant Physiol.*, 136(1):2587–2608, 2004.
- [31] G. Schneider and U. Fechner. Advances in the prediction of protein targeting signals. *Proteomics*, 4(6):1571–1580, 2004.
- [32] M. S. Scott, D. Y. Thomas, and M. T. Hallett. Predicting subcellular localization via protein motif co-occurrence. *Genome Res.*, 14(10a):1957–1966, 2004.
- [33] M. Wakabayashi, J. Hawkins, S. Maetschke, and M. Bodén. Exploiting targetting signal dependencies in the prediction of pts1 peroxisomal proteins. In M. Gallagher, J. Hogan, and F. Maire, editors, *Intelligent Data Engineering and Automated Learning - IDEAL 2005: 6th International Conference*, volume 3578 of *Lecture Notes in Computer Science*, pages 454–461. Springer, 2005.
- [34] S. Weller, S. J. Gould, and D. Valle. Peroxisome biogenesis disorders. *Annual Review of Genomics and Human Genetics*, 4(1):165–211, 2003.
- [35] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2000.