

# PTS1Prowler Peroxisomal Protein Datasets

Mark Wakabayashi & John Hawkins & Stefan Maetschke & Mikael Bodén

## Abstract

In this paper we outline the development of a updated and highly curated data set for peroxisomal localisation via the PTS1 motif. We perform this task by following the general curation and redundancy reduction guidelines outlined for the development of the original PeroxiP datasets. Further more, using the published non-redundancy reduced PeroxiP data, we replicated the original PeroxiP dataset.

## 1 Introduction

The localisation of proteins to the peroxisomal matrix is known to be generally dependant upon the presence of one of two dominant motifs. The vast majority of peroxisomal matrix proteins rely on a motif on the C-terminal end of the sequence called the PTS1 signal, it is often described as SKL with variation, but it has also been described using a number of different PROSITE patterns, each allowing more or less flexibility in the allowed substitutions. A much smaller number of proteins rely on an N-terminal bipartite signal with the following consensus sequence [RK]-[LVI]-x5-[HQ]-[LA] called the PTS2 motif. Even though these motifs are highly conserved in peroxisomal proteins, they are also present in a great many non peroxisomal proteins, thus it is known that there are other elements of the sequence that are involved. We publish two datasets for the specific purpose of improving the quality of machine learning experiments in the area of peroxisomal protein subcellular localisation prediction.

## 2 Datasets

### 2.1 Replication of Emanuelsson’s dataset

The dataset of PTS1-containing peroxisomal proteins and non-peroxisomal proteins, as identified by Emanuelsson *et al.* was available on the internet. These were extracted from SWISS-PROT release 39.27. The dataset included 152 proteins identified as peroxisomal, and 308 identified as non-peroxisomal. This dataset could not be directly used to replicate the PeroxiP model, however, as the manual motif reduction and redundancy reduction had not been performed on this dataset.

The motif reduction was accomplished by removing the proteins in the positive set with residues in the C-terminal tripeptide that only occurred once in that position. This included proteins with SWISS-PROT accession numbers Q13907, Q9UHK6, and Q00922, with motifs -YRM, -ASL, and -ARY respectively. This filter should also exclude Q9P8Q7, whose motif -AKA contains the only A in the final position, but this protein and its corresponding motif were not removed by Emanuelsson. In order to replicate the dataset used to train and test PeroxiP, this was not removed. This filtering left 149 positives and 271 negatives.

Both steps of redundancy reduction were also performed. 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. This caused a reduction to 116 positives and 182 negatives. The final stage of redundancy reduction was performed using BLASTClust. We initially used the reported threshold of 25% given by Emanuelsson *et al.* however this value reduced the positive set to a mere 43 proteins. In order to reproduce a dataset of the same size we found that a similarity threshold of 1.675 was required. The first protein of each cluster was taken as a representative, and the final

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

## 2.2 New dataset construction

The new dataset was collected from SWISS-PROT release 45. We followed a similar process to that described by Emanuelsson et al 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. It was found that the reported search for the complete words, such as "PEROXISOMAL", missed some proteins that were included in the published PeroxiP dataset, and these truncated search strings were therefore used. 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. SWISS-PROT protein P97562, added after release 39.27, has experimental evidence suggesting it is not PTS1 targeted [2]. As a known non-PTS1 targeted protein, this protein was removed from the peroxisomal set and added to the non-peroxisomal set. Proteins O19094 and P11466 were also removed, as they have been judged as unlikely to be PTS1 targeted due to the hydrophobic composition of the adjacent residues potentially concealing the C-terminal tripeptide [3]. Neither protein had any literature supporting their being labelled as a "potential" peroxisomal protein. Furthermore, O19094 bore a C-terminal tripeptide, -PHL, occurring only once in the peroxisomal proteins in SWISS-PROT release 45. Three proteins, O14313, P14293, and P14292 were identified as membrane proteins and moved to the non-peroxisomal set (of which only P14292 occurred in our reduced version of the Emanuelsson dataset). Two further proteins, P56577 and P56578 (both in the Emanuelsson dataset), were not labelled as membrane associated, but were identified as peroxisomal on the basis of similarity to known peroxisomal membrane proteins [5]. These were therefore also removed from the set of peroxisomal proteins, however they were not added to the non-peroxisomal set as there was no experimental evidence available to conclusively show that these are peroxisomal membrane proteins.

A further set of proteins in the dataset were identified as having little or weak experimental evidence in support of peroxisomal localisation, or containing C-terminal tripeptides that occur infrequently in the peroxisomal set without supporting literature for that PTS1. This included Q00922, P58044, Q9BXS1, Q13907, and P38139 - a set that could be expanded by increasing the requirements on supporting literature. Although removal of these proteins from the set yielded no significant change in performance, the set of motifs used in the model's filter was reduced as many of these were identified as potentially non-peroxisomal due to their unusual PTS1 motif. The reduction of the set of accepted PTS1 motifs would reduce the number of positively identified peroxisomal proteins when applied to a larger set if any removed motifs were true PTS1 motifs. Emanuelsson *et al.* removed Q9UHK6 due to its unusual motif, -ASL, however it has been experimentally determined [1] that this is a legitimate PTS1. To avoid such mistakes and to maintain sensitivity in the model, it was therefore decided to retain these proteins in the positive set.

The negative set was generated by extracting from SWISS-PROT release 45 all eukaryotic proteins with a C-terminal tripeptide identical to that of one of the initially identified peroxisomal proteins, and a subcellular localisation not specified as peroxisomal, glyoxysomal, or glycosomal. 573 proteins were found in this way. A pairwise residue comparison of these to the peroxisomal proteins showed that two proteins with no annotation indicating peroxisomal localisation were highly similar to peroxisomal proteins. P42861, annotated as being localised to the cytoplasm, differed by only three residues in its C-terminal twelve residues to those of P13377 - a glycosomal protein. The associated literature [4] confirms that P42861 is partially distributed to the glycosome, and it was consequently moved to the positive set of peroxisomal proteins. The protein Q86TX2 was found to differ in only one residue in the C-terminal 12-mers to the peroxisomal protein P49753. The differing residue is in the position eight residues from the C-terminal end, where arganine is present in Q86TX2 and histidine in P49753. This position has been shown, [3] p.575, to have a

Motif	Motif Distribution		New dataset	
	Peroxisomal	Non-peroxisomal	Peroxisomal	Non-peroxisomal
AHL	5	7	3	0
AKA	1	19	3	11
AKF	1	3	1	1
AKI	1	9	1	9
AKL	19	21	28	9
AKM	3	1	3	2
AKV	1	19	1	1
ANL	2	5	2	7
ARF	1	2	1	2
ARL	4	9	4	9
ARM	5	0	5	0
ARY	1	1	1	1
ASL	1	23	1	22
CKL	2	4	6	1
CRL	0	0	2	0
HRL	1	1	1	1
HRM	2	1	2	1
HRV	2	6	1	1
NHL	0	0	1	2
NKF	0	0	1	6
NKL	4	10	4	12
PKL	3	4	5	6
PRL	1	3	4	10
SHL	10	18	11	17
SKF	2	7	2	7
SKI	6	2	6	6
SKL	42	20	53	22
SKM	4	1	4	2
SKV	1	24	1	24
SNL	2	9	4	24
SNM	0	0	2	2
SQL	4	4	9	4
SRL	14	11	22	7
SRM	4	0	4	0
THL	1	18	1	18
TKL	1	23	1	28
YRM	1	13	1	13

Table 1: The distribution of PTS1 motifs within the PeroxiP dataset and our new dataset.

preference for positively charged residues in PTS1 targeted proteins. It is therefore highly likely that Q86TX2 is PTS1 targeted. This protein was removed from the negative set, but because of the lack of experimental support for this prediction and the redundancy with P49753 that would result, it was not added to the positive set.

In addition to these removals, two sets of proteins were removed due to evidence that they are localised to the peroxisome. Hydroxymethylglutaryl-CoA lyases have been demonstrated (Ashminara *et al.* 1994) to be present in the peroxisome but the relevant SWISS-PROT entries, Q29448, P35915, P35914, P38060, and P97519, do not reflect this. These proteins were moved to the set of peroxisomal proteins. Similarly, The isocitrate dehydrogenases Q9Z2K9, Q9Z2K8, O88844, P41562, P50217, and P50218 were moved to peroxisomal set because of experimental evidence (Geisbrecht and Gould, 1999, and Yoshihara *et al.* 2001).

The resulting dataset contained 206 peroxisomal proteins and 564 non-peroxisomal proteins. Given this overrepresentation of non-peroxisomal proteins and to increase the quality of the dataset, the non-peroxisomal set was cleaned of all proteins whose subcellular location was qualified as "potential", "probable", or "by similarity". This resulted in a final negative set of 348 proteins, a sufficient size given the size of the peroxisomal set.

Redundancy reduction was performed by similar processes to those performed on Emanuelsson's dataset (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 dataset 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 and extraction of representatives from each cluster (the same procedure used to replicate Emanuelsson's dataset) 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 Emanuelsson's set, with the overall size of the dataset increased by approximately 40%.

## 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] E. Baumgart, J. Vanhooren, M. Fransen, F. VanLeuven, H. Fahimi, P. VanVeldhoven, and G. Mannaerts. Molecular cloning and further characterization of rat peroxisomal trihydroxycoprostanoyl-coa oxidase. *Biochemical Journal*, 320(1):115–121, 1996.
- [3] 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.
- [4] K. Nyame, C. D. Do-Thi, F. R. Opperdoes, and P. A. M. Michels. Subcellular distribution and characterization of glucosephosphate isomerase in leishmania mexicana mexicana. *Molecular and Biochemical Parasitology*, 67(2):269–279, 1994. TY - JOUR.
- [5] H. Yasueda, T. Hashida-Okado, A. Saito, K. Uchida, M. Kuroda, Y. Onishi, K. Takahashi, H. Yamaguchi, K. Takesako, and K. Akiyama. Identification and cloning of two novel allergens from the lipophilic yeast, malassezia furfur,. *Biochemical and Biophysical Research Communications*, 248(2):240–244, 1998.