Computational approaches for peroxisomal proteins localisation

Marco Anteghini, Vitor Martins dos Santos

June 2022

Marco Anteghini[1,2,3,*], Vitor AP Martins dos Santos[1,4],

1 LifeGlimmer GmbH, Berlin, Germany
2 Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen WE, The Netherlands.
3 Zuse Institute Berlin, Berlin, Germany.
4 Bioprocess Engineering, Wageningen University & Research, Wageningen WE, The Netherlands.

***Corresponding authors**
marco.anteghini@wur.nl

Abstract

Computational approaches are practical when investigating putative peroxisomal proteins and for sub-peroxisomal protein localisation in unknown protein sequences. Nowadays, advancements in computational methods and Machine Learning (ML) can be used to fasten the discovery of novel peroxisomal proteins and can be combined with more established computational methodologies. In this chapter, we explain and list some of the most used tools and methodologies for novel peroxisomal protein detection and localisation.

Introduction

Advancements in organelle-specific research are possible also thanks to use case-specific tools such as for sub-peroxisomal and sub-mitochondrial protein localisation [1–5]. These tools, nowadays easily accessible and user-friendly, allow researchers to perform fast and accurate screening while looking for new peroxisomal and mitochondrial proteins [1–5]. Alternatively, general methods for protein sequence localisation can be handy if re-adapted for specific use cases [6–11].

For example, a general peroxisomal protein search from a given set of FASTA sequences can start by detecting the predicted subcellular localisation using DeepLoc-2.0 [6]. After filtering for predicted peroxisomal protein, a researcher can either look for known Peroxisomal Targeting Signals (PTSs) to further filter the dataset [3], or retrieve a list of candidates for future analysis or experimental validations [12].

PTSs are consensus motifs found in many peroxisomal proteins. Specific receptors recognise a PTS and bind to a region of the peroxisomal protein [13]. The known PTSs are: 1) PTS1. The PTS1 receptor is encoded by the PEX5 gene [14] is defined as the final dodecamer with a focus on the terminal tripeptide [15]; 2) PTS2. It is an N-terminal targeting signal and its receptor is encoded by the PEX7 gene (a co-receptor is also involved in the peroxisomal import) [16]; 3) mPTS. It is a cis-acting targeting signal specific for peroxisomal membrane proteins. Its mechanism is still poorly understood [17]. The algorithms defined in Schülter et al. (2009) [3] can detect these different PTSs, and the PTS1 can now be easily and accurately detected also on https://organelx.hpc.rug.nl/fasta/compute_in_pts, as described in recent works [1, 5].

In this chapter, we list a number of practical tools accompanied by specific use cases and a workflow on how to perform a complete peroxisomal protein localisation search. The workflow presented here is supported by a service bundle and a practical study example [18].

Materials

Use case-specific tools

- PeroxisomeDB. The PEROXISOME DATABASE (PeroxisomeDB) organise and integrates curated information about peroxisomes. That includes genes, proteins, molecular functions, metabolic pathways and their related disorders [3]. Related prediction tools are also available at http://www.peroxisomedb.org/. In the scope of this chapter, we report three main tools for different PTSs detection: 1) PTS1 binding sites; 2) PTS2 binding sites; 3) Pex19BS binding sites. All the three programs rely on multiple sequence alignments where the input sequence or the input BLOCK is aligned with a predefined BLOCK that contains a specific category of proteins (e.g. proteins containing PTS1).

- In-Pero. A computational pipeline that discriminates between matrix and membrane proteins [1]. In-Pero relies on a Support Vector Machine classifier trained on the statistical representation of protein sequences obtained by combining two deep-learning embeddings (UniRep + SecVec) [19, 20]. In-Pero can be executed locally following the instruction available at https://github.com/MarcoAnteghini/In-Pero or on the dedicated web server available at https://organelx.hpc.rug.nl/fasta/compute_in_pero.

- In-Mito. A computational pipeline that allows classifying the four sub-mitochondrial compartments: matrix, internal-membrane, inter-membrane space and external membrane [1]. In-Mito relies on a Support Vector Machine classifier trained on the statistical representation of protein sequences obtained by combining two deep-learning embeddings (UniRep + SecVec) [19, 20]. In-Mito can run locally following the instruction available at https://github.com/MarcoAnteghini/In-Mito or on the dedicated web server available at https://organelx.hpc.rug.nl/fasta/compute_in_mito.

- DeepMito. A computational method for predicting sub-mitochondrial localisation based on a convolutional neural network architecture [2]. Given an input protein, DeepMito can discriminate the four sub-mitochondrial compartments: matrix, internal-membrane, inter-membrane space and external membrane. DeepMito is available at http://busca.biocomp.unibo.it/deepmito/.

General tools for subcellular localisation and transmembrane detection

- TMHMM2.0 and DeepTMHMM. TMHMM2.0 is a membrane protein topology prediction method based on a hidden Markov model (HMM) [8]. It predicts transmembrane helices and discriminates between soluble and membrane proteins. The tool is available at https://services.healthtech.dtu.dk/service.php?TMHMM-2.0. DeepTMHMM is a novel version of the TMHMM predictor. It is the most complete and currently the best-performing method for the membrane protein topology prediction [21]. The model encodes the primary amino acid sequence by a pre-trained language model and decodes the topology by a state-space model to produce topology and type predictions. DeepTMHMM is available at https://dtu.biolib.com/DeepTMHMM.

- Phobius. Combined transmembrane topology and signal peptide predictor [11]. The predictor relies on a HMM that models the different sequence regions of a signal peptide and the different regions of a transmembrane protein in a series of interconnected states. Phobius is available at https://phobius.sbc.su.se/

- DeepLoc-2.0 [6]. Multi-localization prediction tool based on a pre-trained protein language model that uses a three-stage deep learning approach for sequence classification. 1) The feature representation for each amino acid in the sequence is generated. 2) Attention-based pooling stage produces a single representation for the whole sequence. 3) the prediction stage uses a classifier to output the subcellular labels. DeepLoc-2.0 is available at https://services.healthtech.dtu.dk/service.php?DeepLoc-2.0

- PSORT. A computer program that predic protein localisation sites in cells and its last version is WoLF PSORT[7]. WoLF PSORT converts protein amino acid sequences into numerical localisation features; based on sorting signals, amino acid composition and functional motifs. A k-nearest neighbour classifier is used for the final prediction. The webserver is available at https://psort.hgc.jp/

- TargetP-2.0. Deep Learning method to identify N-terminal sorting signals, which direct proteins to the secretory pathway, mitochondria, and chloroplasts or other plastids [10]. The method relies on Bi-directional Recurrent Neural Networks (BiRNN) with Long short-term memory (LSTM) cells and a multi-attention mechanism [22]. TargetP-2.0 is available at https://services.healthtech.dtu.dk/service.php?TargetP-2.0.

Methods

Peroxisomal protein candidates detection

The workflow for a typical analysis represented as a service bundle is visible in Figure 1 and also available (with functional links for each tool) at https://tess.elixir-europe.org/workflows/peroxisomal-candidates-detection. For an accurate analysis, it is recommended to first look for known PTSs when available and then proceed with further filtering steps. After the PTS detection, we can investigate the presence of transmembrane regions in the aminoacid sequence and filter the results according to the detected PTS. In particular, we can exclude membrane proteins while checking for PTS1 or PTS2. Afterwards, if stringent filtering is required, it is recommended to analyse the candidates with other subcellular localisation tools (see the 'General tools for subcellular localisation and transmembrane detection' section) and remove the proteins with unexpected predicted localisation.

Alternatively, we can start our analysis directly from the subcellular localisation prediction and then run the predicted peroxisomal proteins with a sub-peroxisomal classification tool that does not consider PTS motifs [1]. As reported in Figure 1, after the subcellular localisation prediction, if we obtain mitochondrial proteins, it is possible to either run DeepMito or In-Mito, while we can execute In-Pero for the peroxisomal proteins [1, 2]. These tools discriminate the sub-organelle compartments, which are 4 in case of mitochondria (matrix, internal-membrane, inter-membrane space, external membrane) and 2 in case of peroxisomes (matrix, membrane) [1, 2].

As a final step for further validation, selected sequences can be screened for conservation of the potential PTS1/PTS2 using BLAST (the last version while writing this chapter is BLAST+ 2.13.0) [23, 24] .

An example of a complete pipeline was performed in the recent work of Kamoshita et al. (2022) [18]. For simplicity we report here a summarised computational pipeline: 1) The Danio rerio proteome was downloaded from UniProt (https://www.uniprot.org/) [25] and screened for proteins carrying a PTS1 at the very C-terminus matching the consensus motif [ASCNPHTG]-[RKHQNSL]-[LMIVF]; 2) Among 46,848 proteins, 2,638 proteins matching the

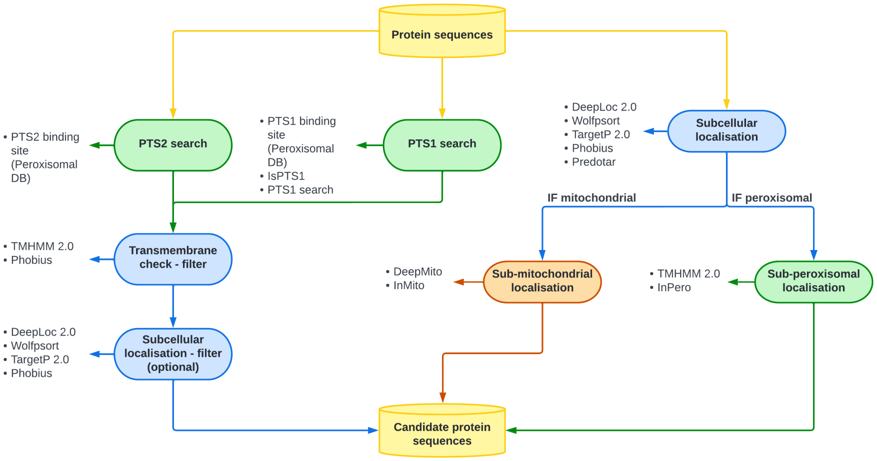

Figure 1: Workflow and Service Bundle of a standard peroxisomal and mitochondrial protein localisation analysis. The workflow starts with the initial dataset containing protein sequences in FASTA format. The starting point is the root of the graph 'Protein sequences'. Each node represents a step of the analysis. Its associated tools are visible on the left of the node. The workflow converges in the final node 'Candidate protein sequences', where candidate protein sequences are selected for future analyses or experimental validation.

pattern were identified and filtered for non membrane proteins with TMHMM Server v. 2.02 [8] (1966 protein left); 3) The 1,966 protein sequences were further analysed with WoLF PSORT (Package Command Line Version 0.2) [7] and entries with Endoplasmic Reticulum as possible subcellular localisation were removed (1,171 sequences left); 4) The identified proteins were further analysed by PTS1 predictor algorithms [3] and sequences which produced no hit with the "metazoa" or "general" modus of the software were removed (371 proteins left); 5) Finally, the obtained entries were manually curated, integrating information from Zebrafish specific datasets and considered for experimental validation.

Protein sequence embeddings for subcellular and sub-peroxisomal protein localisation

Notes:

1. Most of the tools presented in this chapter are designed for Eukaryotes. Some of them can be used for prokaryotic organisms as well (e.g. DeepTMHMM [21]). Note that peroxisomes are only present in Eukaryotes. We advise the user to check the specifications of each tool in the original web server or paper.

2. In this chapter, we list some of the available tools for mitochondrial protein detection. Important components of the organelle division machinery present a dual localisation (peroxisomal and mitochondrial). Moreover, both organelles have proven to be in continuous interplay [26]. For an accurate peroxisomal protein localisation search, it is advised to look into mitochondrial localisation too.

References

[1] Anteghini M, dos Santos VM and Saccenti E. In-Pero: Exploiting Deep Learning Embeddings of Protein Sequences to Predict the Localisation of Peroxisomal Proteins. International Journal of Molecular Sciences 2021; 22:6409.

[2] Savojardo C, Bruciaferri N, Tartari G et al. DeepMito: accurate prediction of protein sub-mitochondrial localization using convolutional neural networks. Bioinformatics 2019;36:56–64.

[3] Schlüter A, Real-Chicharro A, Gabaldón T et al. PeroxisomeDB 2.0: an integrative view of the global peroxisomal metabolome. Nucleic Acids Research 2009;38:D800–D805.

[4] Claros MG and Vincens P. Computational Method to Predict Mitochondrially Imported Proteins and their Targeting Sequences. European Journal of Biochemistry 1996;241:779–786.

[5] Anteghini M, Haja A, Martins dos Santos VA et al. OrganelX Web Server for Sub-Peroxisomal and Sub-Mitochondrial protein localisation. bioRxiv 2022;.

[6] Thumuluri V, Almagro Armenteros JJ, Johansen A et al. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Nucleic Acids Research 2022;ISSN 0305-1048.

[7] Horton P, Park KJ, Obayashi T et al. WoLF PSORT: protein localization predictor. Nucleic Acids Research 2007;35:W585–W587.

[8] Krogh A, Larsson B, von Heijne G and Sonnhammer EL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. Journal of Molecular Biology 2001;305:567–580. ISSN 0022-2836.

[9] Small I, Peeters N, Legeai F and Lurin C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. PROTEOMICS 2004;4:1581–1590.

[10] Almagro Armenteros JJ, Salvatore M, Emanuelsson O et al. Detecting sequence signals in targeting peptides using deep learning. Life Science Alliance 2019;2.

[11] Käll L, Krogh A and Sonnhammer EL. A Combined Transmembrane Topology and Signal Peptide Prediction Method. Journal of Molecular Biology 2004;338:1027–1036.

[12] Schrader TA, Islinger M and Schrader M. Detection and Immunolabeling of Peroxisomal Proteins. In Methods in Molecular Biology. Springer New York, 2017;113–130.

[13] Gould SG, Keller GA and Subramani S. Identification of a peroxisomal targeting signal at the carboxy terminus of firefly luciferase. Journal of Cell Biology 1987;105:2923–2931.

[14] Kiel JA, Emmrich K, Meyer HE and Kunau WH. Ubiquitination of the Peroxisomal Targeting Signal Type 1 Receptor, Pex5p, Suggests the Presence of a Quality Control Mechanism during Peroxisomal Matrix Protein Import. Journal of Biological Chemistry 2005;280:1921–1930.

[15] Brocard C and Hartig A. Peroxisome targeting signal 1: Is it really a simple tripeptide? Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 2006;1763:1565–1573.

[16] Kunze M. The type-2 peroxisomal targeting signal. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 2020;1867:118609.

[17] Van Ael E and Fransen M. Targeting signals in peroxisomal membrane proteins. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 2006;1763:1629–1638. ISSN 0167-4889. Peroxisomes: Morphology, Function, Biogenesis and Disorders.

[18] Kamoshita M, Kumar R, Anteghini M et al. Insights Into the Peroxisomal Protein Inventory of Zebrafish. Frontiers in Physiology 2022;13. ISSN 1664-042X.

[19] Alley E, Khimulya G, Biswas S et al. Unified rational protein engineering with sequence-based deep representation learning. Nature Methods 2019; 16.

[20] Heinzinger M, Elnaggar A, Wang Y et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics 2019;20.

[21] Hallgren J, Tsirigos KD, Pedersen MD et al. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. bioRxiv 2022;.

[22] Lin Z, Feng M, Santos CNd et al. A structured self-attentive sentence embedding. arXiv preprint arXiv:170303130 2017;.

[23] Altschul SF, Gish W, Miller W et al. Basic local alignment search tool. Journal of Molecular Biology 1990;215:403–410.

[24] Camacho C, Coulouris G, Avagyan V et al. BLAST+: architecture and applications. BMC Bioinformatics 2009;10.

[25] Consortium TU. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 2020;49:D480–D489.

[26] Schrader M, Costello J, Godinho LF and Islinger M. Peroxisome-mitochondria interplay and disease. Journal of Inherited Metabolic Disease 2015;38:681–702.