SCLPred: Membrane vs Non-Membrane Prediction Literature Review

Liam Ellinger
Completed 3rd July, 2019

Contents

1 Introduction

This project seeks to create an artificial neural network capable of making predictions on whether or not a given protein is located on the cell membrane or elsewhere via deep learning. The number of sequenced proteins has increased, but the number of proteins annotated with useful information has not increased as much. Tools like this project aim to bridge that gap with predictions until the protein's subcellular localisation can be experimentally determined. The application predicts if a given protein will be localised to any of the cell's membranes or not.

1.1 Protein Overview

1.1.1 Proteins and Protein Targeting

Proteins are chains of 20 different amino acids synthesised according to instructions stored in DNA, and they are responsible for structure, signalling, and breakdown of other molecules [1]. They are manufactured in the ribosomes, and then transported to where the cell needs them based on certain signals present in their sequence of amino acids [2].

1.1.2 Membrane Proteins

Proteins sent to the membrane of the cell have special tasks that set them apart from others. For example, membrane proteins are used to help the immune system identify the body's own cells as well as any cells that have marked themselves for destruction, as well as passive and active transport and other things. [3]. They also can act as receptors for information the body is trying to send to its cells [3]. Because of these things, membrane proteins are common targets for drugs, and their dysfunction can cause many different problems, and even disease [3].

1.1.3 Target Peptides

Each protein has stored information to tell the cell where to put it once it has been manufactured. One important one is the signal peptide, which sends the protein to the secretory pathway [4]. Other signals in the protein's sequence can influence targeting as well. The different places these signals send proteins can be seen in figure 2. Each branch of this graph is caused by a different signal. For example, one signal may cause the protein to move from the cytosol to the ER, and then a different one will direct it from there to the

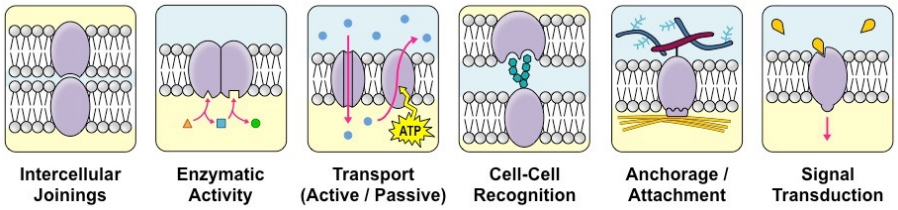

Figure 1: The functions of a membrane protein.
Retrieved from: https://ib.bioninja.com.au/

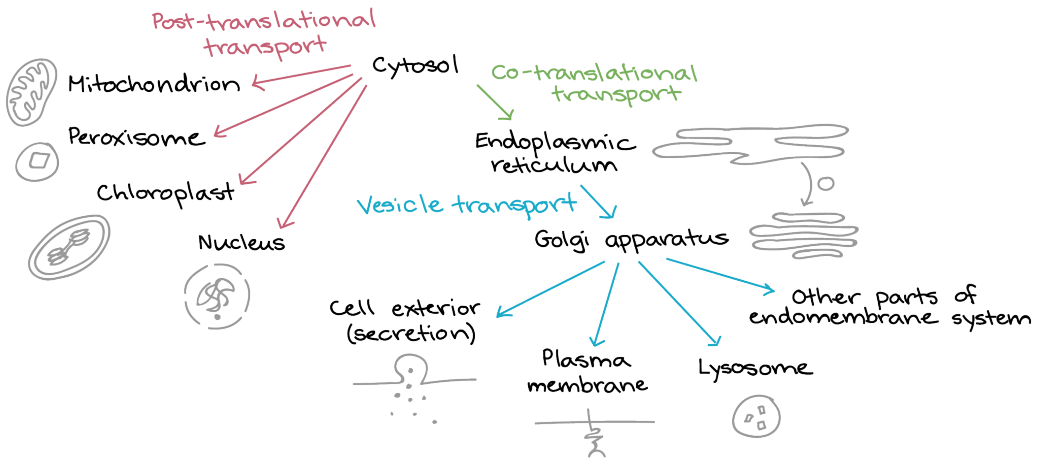

Figure 2: The different places a protein may be based on signal peptide or patch. Retrieved from: https://www.khanacademy.org/

Golgi apparatus to the plasma membrane [5]. There are many different localisation signals, and many have specific locations in the protein's amino acid sequence they appear. A list of signals and their targets, as well as the location the target can be found in the amino acid sequence can be found in the table below. Some signals are removed by cell machinery after the proteins have reached their destinations and the signal is no longer needed.

List of signals and their targets.

Signal	Target	Location
Signal Peptide	Endomembrane System via ER (removed once targeting completed)	N-Terminus
ER-Retention Signal	Endoplasmic Reticulum (ER)	C-Terminus
Nuclear Localisation Signal	Nucleus	Anywhere
Nuclear Export Signal	Cytosol from Nucleus (some have NLS and NES and switch back and forth)	Anywhere
Nucleolar Localisation Signal	Nucleolus	N-Terminus [6]
Mitochondrial Targeting Signal	Mitochondria (removed once targeting completed)	N-Terminus
PTS1	Peroxisome	C-Terminus
PTS2	Peroxisome	N-Terminus

All info from http://proline.bic.nus.edu.sg/spdb/ except otherwise cited.

1.2 Computational Techniques Overview

This project uses a convolutional neural network (CNN), which is a type of deep learning [7]. A CNN handles its data in a way that leads to less resources used, faster computation and the possibility of complicated input that would be impossible to process using a traditional feed forward neural network [7]. The different components of a CNN will be detailed below.

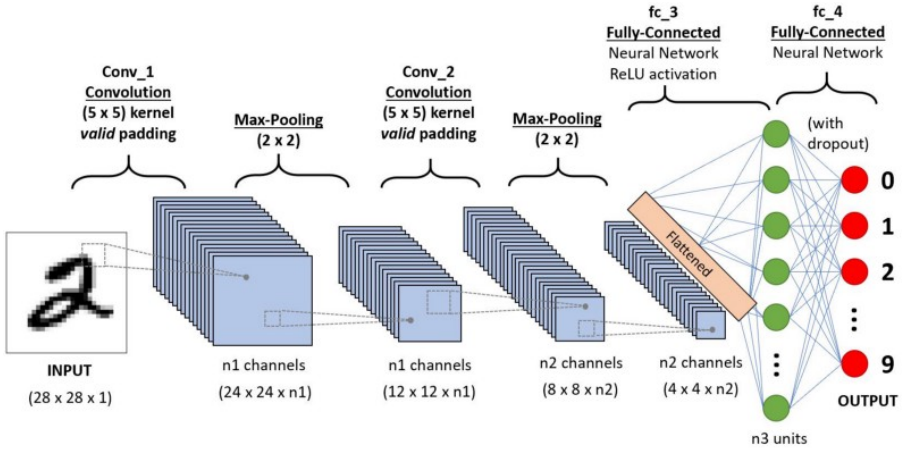

Figure 3: An example convolutional neural network for classifying handwritten digits
Retrieved from: https://towardsdatascience.com/

1.2.1 The Kernel Layer(s)

A convolutional neural network takes in data in sections. These sections are then simplified down into one value or set of values. The values of all the sections are arranged into a matrix that the next part of the network can use [8]. In the example of an image, a chunk could be a 3x3 pixel area of the image. The purpose of this is to reduce the number of parameters the network has to keep track of. In a 1k by 1k pixel image, there will be 3 million individual data points (3 colour values per pixel, 1000^2 pixels). This would require 3 million input neurons and would require an extremely large amount of computing power [7]. Convolutional neural networks allow us to process large datasets like these that would not otherwise be possible to operate on. Figure 3 shows an example network with its kernel layers. It should be noted that this project uses a one-dimensional dataset instead of a traditional two-dimensional set like the one shown in the image. This project uses two kernel layers, but any number can be used generally.

1.2.2 The Pooling Layer

The pooling layer simply serves to compress the data further, and relies only on a hard-coded mathematical formula dependent on the network's purpose. They are present after each kernel layer. [8].

1.2.3 The Fully Connected Layer

The last part of the network is a traditional deep neural network called the fully connected layer. As the name implies, each neuron is connected to every neuron in the next layer. This can be many layers or just one plus the output layer. This is the layer that takes the simplified data from the kernel and pooling layers and generates the final prediction [7].

1.3 Background Research

1.3.1 Bioinformatics and Project Background

Bioinformatics concerns the use of computers to aid learning in biology. It relies on the ability to represent biological objects as code, starting in the 1950s with protein sequencing and eventually moving to DNA sequencing as well [9]. Today bioinformatics covers a multitude of different subtopics, but the one relevant to this project is protein subcellular localisation prediction. Machine learning can be used to predict certain properties of biological objects. This project uses deep learning to predict if a protein will be targeted to one of the cell's membranes. It does this by examining a set of pre-labelled data, learning what patterns the different groups in the data have, and then using those patterns to make a prediction of the properties of new data. This requires a vast quantity of data to examine, which is contained in UniProt, with 146 million proteins sequenced and annotated [10]: perfect for use as training data for this project's neural network. There are other databases available that other predictors take advantage of, such as LocDB [11]. These datasets are usually modified to get only data of a certain type, or to remove redundancy to make the predictor's training data more diverse. Redundancy reduction can be accomplished using a tool like BLAST and it derivatives, which will search a set of proteins for those a given percentage similar to the inputted comparison protein [12].

1.3.2 Overview of Similar Projects

Many projects in this area are iterative. They usually improve on a past algorithm, be it their own previous version or a totally different predictor. Another newer strategy is combining the results of many predictors to make a new, more accurate predictor. Several examples of these things can be found in the list of current predictors. Most recent predictors have an above 80% accuracy, with some having higher accuracy on certain parts of the data or other classifications. The number of classifications each network is capable of is highly variable, with some having only 2 and others having upwards of 20. The high performing predictors all use machine learning, and within this there are two categories, those dealing with sequence data and those dealing with annotation data [13]. Sequence based predictors use information contained directly in the protein's sequence of amino acids, whereas annotation based predictors use metadata about the protein, such as gene ontology [13]. Gene Ontology data contains descriptions of cellular components and molecular functions of gene products, among other things [13]. The thing that has pushed newer predictors up and over the capabilities of the older ones has been deep learning. Previously these predictors relied on data that was already processed and designated into certain groups by humans. With deep learning, the predictor can automatically classify the proteins into these groups, simplifying the process significantly [14]. The most common deep learning techniques for this are deep neural networks, convolutional neural networks (which this project is using), and recurrent neural networks [14].

1.3.3 List of Current Predictors

These predictors were chosen by examining the predictors released in 2017 and onward that classified proteins into membrane and non-membrane categories. Several other older predictors were included if they were very commonly cited and benchmarked against the newer predictors as well.

WoLF PSORT (2007) This predictor extends the PSORT method, detailed next. It converts the amino acid sequence into numerical vectors and classifies them using a weighted k-nearest neighbour classifier. It uses a wrapper to select only the features relevant to the current task. It uses UniProt version 45, and is divided into animal, plant and fungi. WoLF PSORT was published in 2007, and uses the UniProt data from that year [15].

Available at: https://wolfpsort.hgc.jp/

LocTree2(2012) & LocTree3(2014) This predictor classifies eukaryotes into 18 total classes. It uses PSI-BLAST and multiple support vector machines in a hierarchical fashion as detailed in figure 3. Each SVM has a binary output, but they are all arranged to form the final output. The accuracy of LocTree2 is 65% for its 18 classes, but as for its first prediction, which is membrane vs non-membrane, just like this project, is 94% [16]. LocTree3 is simply an enhanced version of LocTree2 that has a better accuracy of 80 $\pm$ 4 % for eukaryotic proteins in their 18 classes [17].

Available at: https://rostlab.org/services/loctree3/

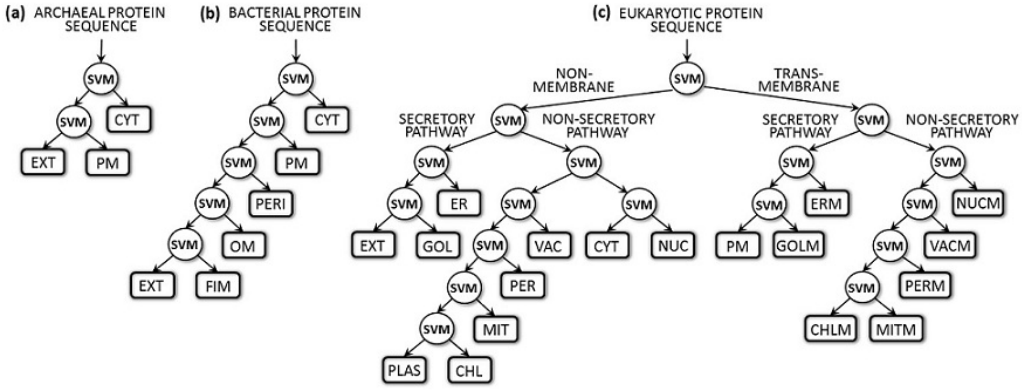

Figure 4: LocTree2 Classification Hierarchy [16]

MDLoc (2015) This predictor uses an approach they call a dependency based generative model. It can classify proteins into more than one output class. It uses a dataset called DBMLoc. As this predictor can place things into more than one class, comparing its accuracy to other predictors is somewhat not applicable [18].

Available at: http://128.4.31.235/

DeepLoc(2017) This predictor uses a recurrent neural network and the UniProt dataset, and analyses the protein sequence directly, as opposed to relying on annotations. It predicts 10 locations, and has a another mode that just classifies membrane bound proteins versus soluble ones. In its 10-class mode it has an accuracy of 78%, and 92% for membrane versus non-membrane [19].

Available: http://www.cbs.dtu.dk/services/DeepLoc/

HPSLPred(2017) This predictor has 10 classes including the cell membrane. It uses data from UniProtKB, and reports a 75% success rate. It uses an ensembled multi label classifier and only works on human proteins [20].

Available at: http://server.malab.cn/HPSLPred/

SubCons (2017) This predictor uses an ensemble of 4 predictors(CELLO2.5, LocTree2, MultiLoc2 and SherLoc2) for its predictions, therefore taking advantage of sequence and annotation based predictions. It reports an F1-Score of 0.79. It places proteins into 9 different classes including membrane, and uses the Mass-Spec dataset, the SLHPA dataset and UniProt [21].

Available at: http://subcons.bioinfo.se/pred/

FUEL-mLoc (2017) This predictor can also predict more than one location for a protein, and handles all eukaryotes and some prokaryotes. It uses gene ontology for its predictions, and is therefore an annotation based predictor. It uses two new databases for its datasets: ProSeq and ProSeq-GO. It has a 79.3% accuracy for eukaryotes, and a 74.3% accuracy for humans. It uses an elastic net multi label classifier for its predictions [22].

Available at: http://bioinfo.eie.polyu.edu.hk/FUEL-mLoc/index.html

BUSCA (2018) This predictor combines 8 other predictors(DeepSig, SChloro, TPpred3, BetAware, MemLoci, PredGPI, BaChelLo, ENSEMBLE3.0) to produce its results. It uses different pipelines for each type of input, for example eukaryotes. The predictors are chosen so that the prediction gets more refined as the system runs. Its eukaryotic workflow can be seen in the figure below. It uses methods for identifying signal and transit peptides, GPI anchors, transmembrane domains and tools for discriminating location of globular and membrane proteins. For animals and fungi it has 9 different locations it can place a protein in, and 16 for plants. It works for plants, animals and gram positive and negative bacteria. They report an F1 accuracy score of 0.49 [23].

Available at: http://busca.biocomp.unibo.it/

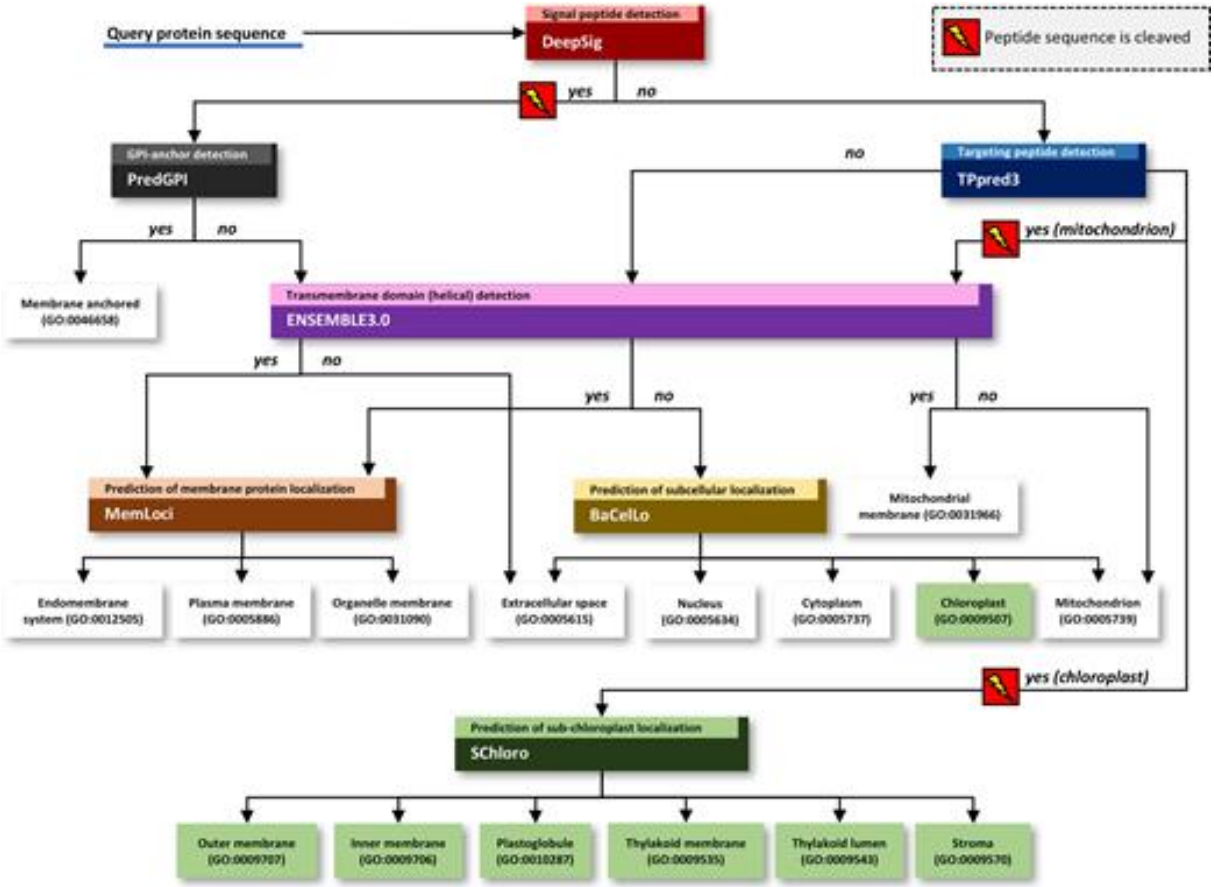

Figure 5: BUSCA's workflow for eukaryotic prediction [23]

References

[1] Jeremy M. Berg, John L. Tymoczko, and Lubert Stryer. Protein structure and function. *Biochemistry. 5th edition*, 2002.

[2] Elliott M. Kanner, Irene K. Klein, Martin Friedlander, and Sanford M. Simon. The amino terminus of opsin translocates "posttranslationally" as efficiently as cotranslationally. *Biochemistry*, 41(24):7707–7715, Jun 2002.

[3] Markus Sällman Almén, Karl JV Nordström, Robert Fredriksson, and Helgi B Schiöth. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biology*, 7:50, Aug 2009.

[4] Günter Blobel and Bernhard Dobberstein. Transfer of proteins across membranes. *The Journal of Cell Biology*, 67(3):835–851, Dec 1975.

[5] Tom A. Rapoport. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*, 450(7170):663–669, Nov 2007.

[6] Andreas Birbach, Shannon T. Bailey, Sankar Ghosh, and Johannes A. Schmid. Cytosolic, nuclear and nucleolar localization signals determine subcellular distribution and activity of the nf-κb inducing kinase nik. *Journal of Cell Science*, 117(16):3615–3624, Jul 2004.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*, page 1097–1105. Curran Associates, Inc., 2012.

[8] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv:1312.4400 [cs]*, Dec 2013. arXiv: 1312.4400.

[9] Jeff Gauthier, Antony T. Vincent, Steve J. Charette, and Nicolas Derome. A brief history of bioinformatics. *Briefings in Bioinformatics*, 2018.

[10] Alex Bateman. Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, Jan 2019.

[11] Shruti Rastogi and Burkhard Rost. Locdb: experimental annotations of localization for homo sapiens and arabidopsis thaliana. *Nucleic Acids Research*, 39(Database issue):D230–D234, Jan 2011.

[12] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, Oct 1990.

[13] Leyi Wei, Yijie Ding, Ran Su, Jijun Tang, and Quan Zou. Prediction of human protein subcellular localization using deep learning. *Journal of Parallel and Distributed Computing*, 117:212–217, Jul 2018.

[14] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in Bioinformatics*, 18(5):851–869, Sep 2017.

[15] Paul Horton, Keun-Joon Park, Takeshi Obayashi, Naoya Fujita, Hajime Harada, C.J. Adams-Collier, and Kenta Nakai. Wolf psort: protein localization predictor. *Nucleic Acids Research*, 35(Web Server issue):W585–W587, Jul 2007.

[16] Tatyana Goldberg, Tobias Hamp, and Burkhard Rost. Loctree2 predicts localization for all domains of life. *Bioinformatics*, 28(18):i458–i465, Sep 2012.

[17] Tatyana Goldberg, Maximilian Hecht, Tobias Hamp, Timothy Karl, Guy Yachdav, Nadeem Ahmed, Uwe Altermann, Philipp Angerer, Sonja Ansorge, Kinga Balasz, and et al. Loctree3 prediction of localization. *Nucleic Acids Research*, 42(W1):W350–W355, Jul 2014.

[18] Ramanuja Simha, Sebastian Briesemeister, Oliver Kohlbacher, and Hagit Shatkay. Protein (multi-)location prediction: utilizing interdependencies via a generative model. *Bioinformatics*, 31(12):i365–i374, Jun 2015.

[19] José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, Nov 2017.

[20] Shixiang Wan, Yucong Duan, and Quan Zou. Hpslpred: An ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. *PROTEOMICS*, 17(17–18):1700262, 2017.

[21] M. Salvatore, P. Warholm, N. Shu, W. Basile, and A. Elofsson. Subcons: a new ensemble method for improved human subcellular localization predictions. *Bioinformatics*, 33(16):2464–2470, Aug 2017.

[22] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. Fuel-mloc: feature-unified prediction and explanation of multi-localization of cellular proteins in multiple organisms. *Bioinformatics*, 33(5):749–750, Mar 2017.

[23] Castrense Savojardo, Pier Luigi Martelli, Piero Fariselli, Giuseppe Profiti, and Rita Casadio. Busca: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Research*, 46(W1):W459–W466, Jul 2018.