Predicting the route: from protein sequence to sorting in eukaryotic cell

Marco Salvatore
Predicting the route: from protein sequence to sorting in eukaryotic cell

Marco Salvatore

Abstract
Proteins need to be localised in the correct compartment of a eukaryotic cell to function correctly. Therefore, a protein needs to be transported to the right location. Specific signals present in the protein sequence direct proteins to different subcellular localisations. The correct transport is essential for the life of the cell, while, possible errors during the transport can cause irreversible damage and interfere with the activities of surrounding proteins. For more than 30 years, the development of methods to identify the localisation of proteins using both experimental and computational approaches has been an important research area. The objective of this thesis is to develop better computational methods for the classification of the subcellular localisation of eukaryotic proteins. I first describe the development of a consensus method, SubCons, which improves the subcellular prediction of human proteins. Next, I present the SubCons web-server as well as an additional benchmark using protein annotation from novel mass-spectrometry studies in two eukaryotic organisms Mus musculus and Drosophila melanogaster. Then, I present the new version of TargetP and how deep learning can improve the identification of N-terminal sorting signals by focusing on relevant biological signatures. Finally, I describe the development of a novel method for sub-nuclear localisation prediction. Here, I show that the performance of a deep convolutional neural network is improved when using an augmented dataset of homologous proteins.

Keywords: eukaryotic cell, sorting signals, subcellular localisation, machine learning, biological sequence analysis, bioinformatics.

Stockholm 2019
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-171434

Department of Biochemistry and Biophysics

Stockholm University, 106 91 Stockholm
PREDICTING THE ROUTE: FROM PROTEIN SEQUENCE TO SORTING IN EUKARYOTIC CELL

Marco Salvatore
Predicting the route: from protein sequence to sorting in eukaryotic cell

Marco Salvatore
To Luce, Enrichetta and my family
Contents

Abstract i
List of papers iii
Abbreviations v
1 Introduction 1
2 Cell organisation and protein sorting 5
  2.1 Cytoplasm .......................... 5
  2.2 Protein sorting mechanism ........ 6
    2.2.1 Co-translational sorting ...... 7
    2.2.2 Post-translational sorting ... 11
3 Approaches to study subcellular localisation of proteins 17
  3.1 Experimental techniques and databases for determining localisation sites ........ 17
    3.1.1 Immunofluorescence Microscopy and the “Cell” Human Protein Atlas .......... 18
    3.1.2 Localisation of Organelle Proteins by Isotope Tagging ......................... 19
    3.1.3 UniProt .................................. 20
  3.2 Computational techniques for determining localisation sites ......................... 20
3.2.1 Supervised machine learning ...................................... 21
3.2.2 Decision Trees ...................................................... 22
3.2.3 Random Forests ...................................................... 22
3.2.4 Artificial Neural Networks ........................................ 23
3.2.5 Deep Neural Networks ............................................. 24
3.3 Practical machine learning techniques ............................... 29
  3.3.1 Data Encoding in Bioinformatics ................................. 29
  3.3.2 Cross-Validation .................................................... 30
  3.3.3 Evaluation of the performance ................................... 31

4 History and concepts of subcellular localisation prediction .......... 33
  4.1 Signal-based .......................................................... 33
  4.2 Global property-based ............................................... 34
  4.3 Homology-based ....................................................... 35
  4.4 Deep learning-based ................................................ 35

5 Results and Discussion .................................................. 37
  5.1 SubCons: a new ensemble method for improved human subcellular localisation predictions (Paper I). .............. 38
  5.2 The SubCons web-server: A user-friendly web interface for state-of-the-art subcellular localisation prediction (Paper II). ................................................ 39
  5.4 Improved sub-nuclear prediction by Deep Learning using an augmented dataset (Paper IV). .......................... 42

Sammanfattning ............................................................... 45

Acknowledgements ........................................................... 47

References ................................................................. 55
Abstract

Proteins need to be localised in the correct compartment of a eukaryotic cell to function correctly. Therefore, a protein needs to be transported to the right location. Specific signals present in the protein sequence direct proteins to different subcellular localisations. The correct transport is essential for the life of the cell, while, possible errors during the transport can cause irreversible damage and interfere with the activities of surrounding proteins. For more than 30 years, the development of methods to identify the localisation of proteins using both experimental and computational approaches has been an important research area. The objective of this thesis is to develop better computational methods for the classification of the subcellular localisation of eukaryotic proteins. I first describe the development of a consensus method, SubCons, which improves the subcellular prediction of human proteins. Next, I present the SubCons web-server as well as an additional benchmark using protein annotation from mass-spectrometry studies in two eukaryotic organisms (*Mus musculus* and *Drosophila melanogaster*). Then, I present the new version of TargetP and how deep learning can improve the identification of N-terminal sorting signals by focusing on relevant biological signatures. Finally, I describe the development of a novel method for sub-nuclear localisation prediction. Here, I show that the performance of a deep convolutional neural network is improved when using an augmented dataset of homologous proteins.
List of papers

Papers included in the thesis


2. The SubCons webserver: A user friendly web interface for state-of-the-art subcellular localisation prediction.


4. Improved sub-nuclear prediction by Deep Learning using an augmented dataset.
Additional papers

1. The classification of orphans is improved by combining searches in both proteomes and genomes.

2. Why do eukaryotic proteins contain more intrinsically disordered regions?

* = denotes joint first authorship

Reprints were made with permissions from publishers.
Abbreviations

biNLS = classical bipartite nuclear localisation signal
CNN = convolutional neural network
cTP = chloroplast transit peptide
CV = cross Validation
DL = deep learning
DNA = deoxyribonucleic acid
FFN = feed forward network
GPU = graphics processing unit
IF = immunofluorescence microscopy
IMP = inner membrane peptidase
lTP = thylakoid transfer domain
LSTM = long-short term memory cell
SP = signal peptide
MIP = mitochondrial intermediate peptidase
ML = machine learning
mTP = mitochondrial targeting peptide
NLS = classical monopartite nuclear localisation signal
NLP = natural language processing
PST1-PST2 = peroxisomal targeting signal 1 or 2
ReLU = rectified linear units
RF = random forest
RNN = recurrent neural network
RNA = ribonucleic acid
Sec/SPI-Sec/SPII = Sec substrates cleaved by SPase I or II
Tat/SPI = Tat substrates cleaved by SPase I
Chapter 1

Introduction

“A cell is the most basic unit of life” [1] and all living things are made of cells.

The cell was discovered for the first time in 1653 and presented in the book “Micrographia” by Robert Hooke in 1665 [2]. In 1674, Antony van Leeuwenhoek observed the first single-cell organism [3]. However, only later in 1838-1839, two scientists, Matthias Jakob Schleiden and Theodor Schwann introduced the “cell theory”. Finally, in 1858, Rudolf Ludwig Carl Virchow published the book “Die Cellularpathologie”. Here, he proposed that cells arise from the division of pre-existing cells and that all organisms are made of cells [4]. Since then, cell biology has progressed in steps to understand and characterise cells.

Cells can be subdivided into two types, namely prokaryotic and eukaryotic. The main difference between a prokaryotic and a eukaryotic cell is the presence in the latter of a well-defined nucleus surrounded by the nuclear envelope which separates the interior of the nucleus from the cytoplasm.

Despite this difference, the cell organisation is similar, and they share common features, including that they all self-replicate and are separated from the extracellular space by the cell membrane, which allows substances to go in and out of the cell. The cell membrane is present in every type of cell.
The eukaryotic cell is highly organised (Figure 1.1) and contains several membrane-enclosed organelles. There is an internal division where the cytoplasm is organised into several compartments, such as endoplasmic reticulum, Golgi apparatus, lysosome, mitochondrion, nucleus, peroxisome and chloroplast (only plant cells) (Figure 1.1b). There are also non-membrane bound compartments both within the nucleus (nucleoli, Cajal bodies, nuclear speckles, PML body) and the cytoplasm (processing bodies, granules and stress granules).

![Figure 1.1: The figure shows the organisation and compartmentalisation of the eukaryotic cell. Figure (a) illustrates the animal cell while figure (b) the plant cell.](image)

Every organelle carries out a specialised function. Nucleic acids (DNA and RNA) are the molecules responsible for storing and expressing the genetic material. DNA carries all the information required to build and maintain the cell, while RNA express the information stored in DNA.

However, without proteins, processes for expressing and maintaining the genetic information are not possible. Proteins are responsible for several processes, ranging from assisting in genome replication to cell division. They carry out specific functions according to their localisation in the cell. Therefore, a protein needs to be in the right location.
Specific signals present in the sequence are responsible for directing the protein to the correct destination (Figure 2.1). The correct transport of the protein is essential for cellular life, while erroneous transport can cause irreversible damage and interfere with the activities of surrounding proteins [5]. Consequently, it is crucial to correctly identify and annotate the localisation of a protein using experimental or computational approaches.

The objective of the thesis is to develop better computational methods for the classification of the subcellular localisation of eukaryotic proteins. I begin my thesis by presenting the two essential ways (co- and post-translational targeting) used by the proteins to reach their final destination (chapter 2). Next, I describe experimental and computational approaches to identify subcellular localisation, with an emphasis on machine learning algorithms utilised during my doctoral studies (chapter 3). I also introduce a few important concepts in the history of the field (chapter 4), and, finally, I present my contribution to the area of subcellular localisation and sorting signals prediction (chapter 5).
Chapter 2

Cell organisation and protein sorting

2.1 Cytoplasm

The cytoplasm is the part of the cell ending at the cell membrane, where all organelles are placed into a gel-like substance called the cytosol (made of about 80% of water) [6]. It is rich in sodium, potassium, calcium, salt, enzymes, proteins and RNA. Many activities take place in the cytoplasm such as cellular movement, glycolysis, phagocytosis, apoptosis, and protein synthesis [1, 6].

The cytoplasm is also the place wherein eukaryotic translation occurs and where the ribosomes are located. Ribosomes can be free in the cytoplasm (post-translation targeting) or bound to the endoplasmic reticulum membrane (co-translational targeting). From the cytoplasm, proteins destined to mitochondria, peroxisomes, chloroplasts and the nucleus use the post-translational targeting. The co-translational targeting is used by proteins directed to the endoplasmic reticulum where they enter the secretory pathway [1].
2.2 Protein sorting mechanism

Protein sorting is the mechanism that allows proteins to reach the correct place inside or outside of the cell.

Specific signals, called signal or targeting peptides, are responsible for directing the protein to the correct destination (Figure 2.1). The best-studied signals are the pre-sequences found at the N-terminal region and are composed of basic and hydrophobic amino acids. Additionally, proteins can have internal or C-terminal targeting peptides, like peroxisomes do.

Figure 2.1: The figure shows the signals responsible for directing the protein to the correct destination. 1 = signal peptide, 2 = mitochondrial targeting peptide, 3 = mitochondrial targeting peptide plus mitochondrial intermediate peptidase, 4 = mitochondrial targeting peptide plus inner membrane peptidase, 5 = chloroplast transit peptide, 6 = chloroplast transit peptide plus thylakoid transfer domain, 7 = classical monopartite nuclear localisation signal of SV40 large T antigen, 8 = classical bipartite of nucleoplasmin, 9 = peroxisomal targeting signal, 10 = peroxisomal targeting signal.
A particular type of signals, the signal patches, are composed of parts separated in the sequence and become functional when the protein folds. Also, protein modifications can induce targeting (i.e. glycosylation for proteins destined to the Golgi apparatus).

Further, proteins might have multiple subcellular localisations or might end up in the final localisation through the binding of other proteins that contain the information [7]. However, all of them fall into two main mechanisms, known as co- and post-translation sorting.

2.2.1 Co-translational sorting

Proteins undergoing co-translation sorting contain a short positively-charged N-terminal segment, a central hydrophobic region and a polar C-terminal segment known as the signal peptide [8].

By using this signal, proteins reach the endoplasmic reticulum, where they enter the secretory pathway, which is the way the cell secrets proteins into the extracellular environment (Figure 2.2).

Endoplasmic reticulum

The endoplasmic reticulum is a membrane-bound system of tubules and flattened sacks (cisternae) only present in eukaryotic cells that extend from the nuclear membrane throughout the cytoplasm [10].

It consists of two contiguous membrane domains; the rough (covered by ribosomes) and the smooth (without ribosomes) domain. The rough endoplasmic reticulum takes part in protein processing, while the smooth in lipid metabolism. The space inside is called the lumen.

Most of the proteins directed to the endoplasmic reticulum in mammalian cells use the co-translational targeting while yeast can use both post- and co-translational targeting. Some proteins are destined to endoplasmic reticulum and float inside, while others are anchored in the membrane [1].

Proteins targeted to the endoplasmic reticulum have a signal sequence at the N-terminus of the growing polypeptide chain known as signal peptide. The signal peptide has a short positively charged
N-terminal segment, a central hydrophobic region and a polar C-terminal segment (Figure 2.1) [8].

Some proteins do not follow the secretory pathway, but they are retained and function within the endoplasmic reticulum. These proteins have the so-called retention signal that consists of amino acids such as Lysine, Asparagine, Glutamic Acid and Leucine at the C-terminus[1].
Golgi apparatus

The Golgi apparatus is adjacent to the endoplasmic reticulum and
is an organelle made of flat discs of membranes, which is of extreme
importance in processing proteins destined for secretion [1].

Golgi apparatus proteins do not have a specific sorting signal.
Thus, the protein transport takes place throughout vesicles from the
endoplasmic reticulum that travel to Golgi cis face, fuse with it, and
empty their contents into the lumen of the Golgi apparatus. Before
arriving at their final destination, the lipids and proteins in the trans-
port vesicles are sorted, packaged, and tagged so that they end up in
the right location [1].

Lysosome

The lysosome is a membrane-enclosed organelle that contains digest-
ive enzymes and acts as the organelle-recycling facility in the cell.
The lysosome uses enzymes to break down molecules such as carbo-
hydrates, lipids, nucleic acid and proteins. They also digest material
derived from autophagy and phagocytosis [1].

Lysosomes form when transport vesicles from the Golgi apparatus
fuse with the late endosome. To be targeted to the late endosome,
which is at the intersection between the secretory and the endocytic
pathway, lysosomal proteins must have a signal containing mannose-
6-phosphate residue [1].

Cell membrane

The cell membrane separates the cytoplasm from the extracellular
space and controls every interaction with the outside. The cell mem-
brane is composed of carbohydrates and a phospholipid bilayer with
embedded proteins. The lipids are essential to give the structure to
the cell membrane [1].

Two classes of membrane proteins exist: peripheral and integral
proteins. Peripheral proteins are embedded to the phospholipid bi-
layer or are attached to integral proteins. In contrast, integral pro-
teins are always attached to the membrane.

A significant class of integral membrane proteins are transmembrane proteins. Transmembrane proteins can cross the membrane one or multiple times. Four major types of single-spanning membrane proteins exist. Type I membrane-bound proteins are targeted to the secretory pathway using the signal peptide. Then type I proteins are anchored to the membrane by a subsequent stop-transfer segment of 20 hydrophobic residues. This stop sequence acts as a transmembrane anchor by halting the further translocation of the polypeptide chain [11]. The majority of type II proteins use their first transmembrane domain, which is similar to a signal peptide but is not cleaved and can be positioned internally within the protein. This signal peptide, known as signal anchor, consists of 18-25 mostly apolar amino acids. The signal anchor is responsible for insertion and anchoring. Moreover, it induces the translocation of its C-terminal end across the membrane [12]. Type III proteins are the opposite. The signal anchor translocates its N-terminal end across the membrane. The insertion of Type I, II and III membrane proteins take place by the same machinery involving the SRP, SRP receptor and the Sec61 translocon [13]. Besides, the fourth class of single-spanning proteins use a C-terminal signal sequence, which is exposed to the cytosol and anchored to the membrane [11].

In multi-spanning membrane proteins, the first transmembrane domain targets the protein to the endoplasmic reticulum. From this point, the protein can initiate translocation and membrane insertion. Multi-spanning proteins can be of type I, II, or III. They are part of one class based on whether the most N-terminal apolar sequence is cleaved by signal peptidase or spans the membrane with a Ncyt/Cexo or Nexo/Ccyt orientation, respectively [14].

The final signal orientation is dependent on the distribution of charged residues near the hydrophobic core of signal and transmembrane segments. This principle is known as the “positive-inside rule”. According to this, positive residues are four times more abundant in cytoplasmic than in periplasmic loops [15].
2.2.2 Post-translational sorting

Proteins targeted to chloroplasts, mitochondria, nucleus and peroxisomes use post-translational sorting (Figure 2.3) [1]. These types of signals are present at the N-termini of the mitochondrial, chloroplast and peroxisomal (with the peroxisomal targeting signal 2 = PST2) proteins. In contrast, the peroxisomal targeting signal 1 (PST1) is present at the C-termini of the protein. Uniquely, in nuclear proteins, the signal can be located either at the N-/C-termini or in the central region of the sequence. The signal is a short linear motif.

![Diagram of post-translational targeting](Figure 2.3)

Figure 2.3: Scheme of the post-translation targeting pathway. The figure is inspired from www.khanacademy.org/science/biology/gene-expression-central-dogma/translation-polypeptides/a/protein-targeting-and-traffic and originally adapted from [9]
Chloroplast

The chloroplast is a type of plastid exclusively found in plant and green algae formed by several parts. The unique feature of the chloroplast is the presence of chlorophyll pigments, which give the typical green colour to the cells. Chloroplasts are the site of photosynthesis.

Chloroplastic proteins have a specific signal, the chloroplast transit peptides (cTP), which guide the proteins to the chloroplast stroma. The cTP consists of three regions: an uncharged N-terminal region, a central region lacking acidic amino acids but enriched in Serine and Threonine and a C-terminal region enriched in Arginine that forms an amphiphilic β strand [16, 17].

Chloroplastic proteins have a bipartite pre-sequence structure to reach the thylakoid lumen. Once the cTP is cleaved, the protein enters the stroma, a luminal transit peptide (lTP) is recognised, and the protein arrives to the thylakoid, where the lTP is cleaved. The lTP is similar to a bacterial SP, and the thylakoidal processing peptidase belongs to the family of signal peptidases [18, 17].

Mitochondria

Mitochondria are membrane-bound organelles organized in several parts. Mitochondria have the vital function to generate energy, and are considered the power plant of eukaryotic cells.

The vast majority of mitochondrial proteins (around 500 in Saccharomyces cerevisiae and 1,000 in mammals) are encoded by nuclear genes, translated on cytosolic ribosomes, and post-translationally transported to the mitochondria [19]. Proteins targeted to the mitochondrial matrix have a specific signal of about 20-60 residues called mitochondrial transit peptides (mTP). mTPs are usually enriched in positively charged and hydroxylated amino acids (Arginine, Leucine and Serine). mTPs tend to form an amphiphilic helical structure to interact with the import receptor on the mitochondrial membrane [20, 17]. Proteins targeted to the inner mitochondrial membrane, or the intermembrane space often have a bipartite mTP, where the second part is similar to an SP [21, 17].
Three peptidases are responsible for the proteolytic processing of both nuclear and mitochondrial-encoded precursor polypeptides targeted to the various sub-compartments of the mitochondria. The mitochondrial processing peptidase (MPP) cleaves the vast majority of mitochondrial proteins, while the inner membrane peptidase (IMP) and the mitochondrial intermediate peptidase (MIP) process specific subsets of precursor polypeptides.

Peroxisome

The peroxisome is an important membrane-bound organelle used by the cell in the oxidation of biomolecules and the biosynthesis of membrane lipids [1].

Proteins targeted to the peroxisome can have two peroxisomal targeting peptides: PTS1 and PTS2. PTS1 contains a chain of residues such as Serine-Lysine-Leucine at the C-terminus. PTS2 is an N-terminal bi-partite extension with a typical consensus sequence ([RK]-[LVI]-x-x-x-x-x-[HQ]-[LA]) [9].

Nucleus

The nucleus is the compartment that differentiates eukaryotic from prokaryotic cells. The nucleus is the venue of many activities of the cell, including transcription, gene regulation, metabolism and growth [1].

Two main signals govern the nucleo-cytoplasmic transport through the nuclear pores, the nuclear export signal (NES) and the nuclear localisation signal (NLS).

NLS directs proteins to the nucleus and is made from basic amino acid residues (e.g. Lysine and Arginine). The canonical nuclear localisation signal can be adjacent (monopartite) or separated by a stretch of amino acids (bipartite). There are also additional signals, which are not recognised by the transport receptors (importins).

NES directs proteins outside the nucleus using specific receptors called exportins. It is, in general, constituted by hydrophobic amino acids (e.g. Leucine).
Some nuclear proteins do not contain these signals within their sequences, or they are not necessarily unique (i.e. proteins that moves in and out the nucleus). The presence of both import and export signal(s) in the target protein, makes it difficult to assess how the protein is imported and exported into and from the nucleus [22].

The internal structure of the Nucleus

The nucleus contains several functionally distinct parts including the nuclear matrix, chromosomal localised proteins and several membrane-less organelles (MLOs).

MLOs are distinct functional units without defined lipid membranes, which exhibit self-organizing properties where proteins and RNA assemble into liquid-like subcellular compartments [23, 24, 25]. These micro environments contain specific RNAs, intrinsically disordered and multivalent hub proteins (Figure 2.4) but the underlying reasons for their formation still remain unclear.

Figure 2.4: The figure shows the internal organisation of the nucleus, which contains several functionally distinct parts, including several MLOs. Adapted from [26]
Nuclear bodies are a prominent sub-class of MLOs, which includes Cajal bodies, nuclear splicing speckles, PML (promyelocytic leukaemia) nuclear bodies, and other droplets [27, 28].

For example, Cajal bodies are highly present in proliferative cells that have a high demand for splicing activity. They are active small nuclear loci of small nuclear RNA (snRNA) transcription [1].

Nucleoli instead control most of the ribosomal biogenesis since they are in proximity of actively transcribed rRNA [1].

Nuclear splicing speckles are enriched in spliceosomal small nuclear ribonucleoprotein particles (snRNPs) and components of the pre-mRNA splicing machinery but are not considered as sites of pre-mRNA splicing. They might constitute storage sites of splicing factors to supply them to adjacent active transcription sites according to need [27, 28].

PML bodies are involved in transcriptional regulation, DNA damage response and apoptosis. PML bodies are considered to be storage sites and hubs for protein modifications [27, 28].
Chapter 3

Approaches to study subcellular localisation of proteins

3.1 Experimental techniques and databases for determining localisation sites

For a long time, to determine the subcellular localisation, experimental methods had been the most used. Among these methods, studies involving antibodies that specifically bind to the target protein in combination with imaging techniques were the most common.

An additional contribution was made with the advent of LOPIT (Localisation of Organelle Proteins by Isotope Tagging) [29] and later of the hyperplexed LOPIT or hyperLOPIT [30], which are both mass-spectrometry based methods.

In addition to these two methods, many complete eukaryotic proteomes with relative annotation are available in UniProt [31]. Although only a tiny fraction of the millions of proteins have been studied experimentally, UniProt [31] remains the most comprehensive source of information. It provides subcellular localisation, annotations of signal and targeting peptides together with the correspond-
3.1.1 Immunofluorescence Microscopy and the “Cell” Human Protein Atlas

Immunofluorescence Microscopy is a method to determine the localisation of a target protein using a specific antibody. In Immunofluorescence Microscopy, the cell is treated with two antibodies: the first one binds to the target protein, while the second antibody is tagged with a fluorescent marker and binds to the immunoglobulin epitope of the first one. During fluorescing, specific reference markers can emit light at a different wavelength. This property is used so as to have a full site of the target protein. Consequently, if the spatial pattern of the two wavelengths of light coincides, most probably the target protein has the same localisation as the reference marker. While having a second antibody that does not depend on the target protein is an advantage, developing accurate and reliable antibodies is an arduous task.

The Cell Atlas in the Human Protein Atlas (HPA) project [32] provides one of the most extensive human datasets. The annotation of subcellular localisation is generated using antibody-based profiling by immunofluorescence confocal microscopy [33]. The study includes data derived from 64 cell lines of various cell populations in different organs and tissues of the human body stained with antibodies and reference markers [33]. All the resulting images are manually annotated regarding subcellular localisation and assigned to one of the following reliability scores: enhanced, supported, approved, uncertain.

From the documentation on the website, “enhanced” proteins have no contradicting data. For these proteins, there is no experimental evidence in the literature describing the different location for the same protein. Plus, for this class, one or more antibodies are further validated (www.proteinatlas.org/about/assays+annotation). In the “supported” group, the annotated localisation is reported in the literature. However, there is no enhanced validation of the used antibody.
The class “approved” contains proteins for which the localisation was detected by only one antibody without additional antibody validation. Moreover, the localisation of the protein has not been previously described. Proteins annotated as “uncertain” have the antibody-staining pattern that can contradict the experimental data.

3.1.2 Localisation of Organelle Proteins by Isotope Tagging

The LOPIT method is a quantitative mass spectrometry workflow for determining protein subcellular localisation presented for the first time in 2004 [29]. The LOPIT method identifies and quantifies differential subcellular enrichment and their respective protein complements. The LOPIT method separates the organelles using centrifugation and determines the protein distribution in each organelle by pairwise comparisons of gradient fractions [29].

The hyperplexed LOPIT (hyperLOPIT) is an advance of the original LOPIT method, which combines biochemical fractionation techniques and shotgun proteomics with isobaric tag quantification. The hyperLOPIT method applies gradient-based ultra-centrifugation to separate the complete cell content. The ultra-centrifugation steps allow a continuous separation of the entire cell content as a function of its density. The relative protein abundances within the fractions represent different organelle-specific distributions amongst partially enriched fractions [30].

Both LOPIT and hyperLOPIT determine the localisation of novel proteins using multivariate data analysis techniques to match the distributions with known proteins located in specific organelles [34].

To generate data, the mass spectrometry community has dedicated an extensive effort over the years. However, spatial proteomics still extensively relies on reliable organelle markers to infer proteome-wide localisation [35]. Several of these datasets are publicly available under the pRloc [36] package in Bioconductor and can also be visualized using the prolocGUI (www.bioconductor.org/packages/proloc or prolocGUI).
3.1.3 UniProt

UniProt [31] is the reference database of protein sequence and functional information. Three leading institutions collect the data: the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR).

UniProt has two main parts: the manually annotated Swissprot and the automatically annotated TrEMBL with more than 560,000 and 157,000,000 proteins, respectively (version 2019_06). Uniprot contains a large amount of biological information about proteins derived from several resources. Over the years, UniProt has integrated information from other resources as well, including Pfam [37] and the Protein Data Bank (PDB) [38].

UniProt also includes information about subcellular localisation, annotations of signal and targeting peptides together with the corresponding cleavage site for many eukaryotic organisms. For subcellular localisation, as for other types of information, UniProt provides “evidence tags” represented by a code from the Evidence Codes Ontology (ECO) which describe the source of the data. The source of information is usually a database record. In the website, evidence coloured in gold corresponds to the manual assertions, while those coloured blue come from the automatic assertions.

In this thesis, I used only manually curated information with published experimental evidence. In UniProt, proteins with this annotation have the code ‘ECO:0000269’.

3.2 Computational techniques for determining localisation sites

The limitations of experimental methods in terms of costs and time have increased the necessity of alternative approaches, including computational ones. For more than 30 years, experimental and computational methods have worked hand-in-hand to improve the identification of the sorting signal and consequently of the subcellular localisation. In general, predictions are faster and cheaper than using
experimental wet-lab methods and sometimes are necessary to annotate entire genomes.

Although the computational approach is still challenging, and our knowledge is still incomplete, computational methods have become an efficient alternative \[7\]. Computational prediction of protein subcellular localisation using sequence-based information was introduced 33 years ago in the case of signal peptides \[39\]. Since then, the prediction of protein subcellular localisation starting from the amino acid sequence has become more frequent and efficient.

Better classifiers make it possible to predict the localisation of non-annotated proteins and systematically increase our understanding of protein sorting. Today, most of the computational tools use machine learning methods including Hidden Markov Models, Support Vector Machines, Artificial Neural Networks and Random Forests \[40\].

3.2.1 Supervised machine learning

Machine learning algorithms are computer programs that learn the best way to solve a problem. There are three main types of machine learning procedures (unsupervised, semi-supervised and supervised) and two general applications (classification and regression). Regression is used to predict a continuous target variable based on input data, while classification is used to predict a discrete variable.

In this thesis, only supervised machine learning algorithms for classification problems have been used and thus presented in more detailed.

Supervised machine learning uses training examples with associated known labels to learn the relationship between X (the examples) and Y (the labels). In this way, the ML algorithm can learn a function that can be used to classify unknown and unseen examples.

For example, when learning to classify protein localisations, we use thousands of cases along with labels containing the correct localisation each protein has. The algorithm then determines the relationship between the protein sequence and its associated label. It applies the learned relationship to classify protein localisation (without labels) that the machine has not seen before. If trained correctly, a good
model should be able to generalise well when new examples X (protein sequences) are given without knowing Y (the localisations).

### 3.2.2 Decision Trees

A decision tree is a widely used algorithm for classification or regression problems that can handle both categorical and numerical data. No domain knowledge or parameter setting is required to construct a decision tree. Often, a decision tree is used to explore the properties of the data [41].

Decision trees are defined as a “white-box model” that is simple to understand and interpret [41]. However, if a small change occurs in the data, a decision tree can be relatively inaccurate, unstable and prone to a significant difference in the optimal structure [42].

In the purest form, decision tree classifies items based on their features, using the logic of yes or no. The questions are encoded as a hierarchical tree, where a node corresponds to a question. Every internal node points to one child node for each possible answer [43].

For instance, a decision tree can be used to classify whether a protein is nuclear or not. The decision is based on specific features such as the type of targeting pathway (post-translation - yes vs no) or amino acid properties (charged vs uncharged) (Figure 3.1). Decision trees have also been applied to other problems such as assigning protein function and predicting splice sites [43].

### 3.2.3 Random Forests

The Random forest (RF) is a supervised machine learning algorithm that can be used for classification and also to compute feature importance. The building blocks of a RF are decision trees, which are randomly created [44].

The algorithm consists of splitting a node among a random subset of features to create specific numbers of trees. The final prediction is the result of the majority vote of the generated decision trees. In this way, the RF algorithm creates weak and uncorrelated sub-models
Figure 3.1: Examples of decision trees for the classification of nuclear proteins. The decision is based on specific features such as the type of targeting pathway (post-translation - yes vs no) or amino acid properties (charged vs uncharged).

for the ensemble, eliminating the overfitting issue that is typical of decision trees [44],[45].

Randomness is both the strong and weak point in any RF model. First, randomness can remove overfitting by not using all the features of the dataset together, but randomly taking a subset of them. Indeed, every tree in the forest is unique. The user can control many parameters (i.e. the number of trees, maximum depth, number of max features), but not the way each tree is created. Therefore, it is hard to know how RF makes the prediction. Moreover, a disadvantage of the RF is that the outcome is predicted from a previously encountered data, meaning that the final score of the prediction is based on the previous one.

RF has been used to predict gene expression in relation with age in the brain [46], but also in the prediction of contacts in proteins [47] and for classification of amino acid substitutions in human proteins [48].

3.2.4 Artificial Neural Networks

Artificial neural networks (ANNs) are algorithms inspired by how the animal brain works [49, 50, 51]. ANNs perform tasks by considering examples. Like the synapses in the brain, each connection in an ANN can transmit a signal from one node (neuron) to another.
A prevalent and widely used type of ANN is the feed-forward neural network (Figure 3.2). Feed-forward neural networks (FFNs) consist of at least three layers (input, hidden and output). The connections between nodes (neurons) of different layers are called “edge”, which have weights adjusted internally while learning through the “backpropagation” function.

In ANNs, backpropagation is used to learn and thus to update the weights at each node [52]. The training process works by optimizing a loss function, which measures the difference between the network predictions and the actual values of the labels.

The basic idea behind the ANN is that, with a given input, a non-linear activation function (i.e. tanh or sigmoid) is applied to provide an output, which is the sum of its inputs [53, 54].

### 3.2.5 Deep Neural Networks

The simplest “Deep” neural network consists of an input layer, at least three hidden layers and an output layer (Figure 3.3). However, this definition is shallow compared to the complexity of the architecture we
can use today. From 2006, when Hinton described the layer-by-layer training of deep belief network began a new era of neural network. Only six years later, thanks to faster numerical computations using GPU cards, new activation functions such as ReLU and procedures like batch normalization, deep neural networks have started to be usable and forceful. From 2012, the AlexNet (ImageNet) [55] and next the GoogLeNet [56], ResNet [57] and DenseNet [58] architectures have improved the state-of-the-art in several machine learning domains including object detection, speech recognition, machine translation as well as bioinformatics.

The principles behind these architectures are the same. They transform the representation at one level into a representation at a higher and more abstract level, by constructing simple non-linear modules and using layers of features not designed by a human [53].

Figure 3.3: Example of a deep neural network

Deep neural networks are generally hard to train. Moreover, to prevent overfitting, deep learning methods require a large number of verified data points and extensive parameter optimization for training.

In this section, I introduce the one-dimensional Convolutional Neural Network (1D CNN) and Long Short-Term memory cell (LSTM). I present some essential properties of these neural networks, and the concept of attention mechanism applied together with LSTM. These are the concepts of deep learning I have used the most in my thesis projects.
Convolutional Neural Network (1D)

Convolutional neural network (CNN) (Figure 3.4) is a type of artificial neural network in which the convolution operation corresponds to a sliding window function applied to a matrix. CNNs automatically and adaptively learn spatial hierarchies of features through backpropagation [59].

![Diagram of a 1D Convolutional neural network used for sequence classification.](image)

Figure 3.4: Example of a 1D Convolutional neural network used for sequence classification. Three filter regions of sizes 2, 3 and 4 with two filters each are used. Each filter performs convolution on the 7*5 matrix and generates feature maps. Max pooling is performed over each map, and a feature vector is generated from all six maps. These six features are concatenated to form a feature vector for the penultimate layer. The final softmax layer then receives this feature vector as input and uses it to classify the sequence (two possible output states). Image adapted from [60].
CNNs consist of multiple building blocks containing several convolutions, pooling, and fully connected layers with non-linear activation functions (Figure 3.4) [59].

CNNs are comparable to an FFN repeatedly used once at each location on its input. In this way, the network becomes much smaller but increases in depth. In a feed-forward network (FNN), each input neuron is connected to an output neuron in the next layer. CNNs use local connections and each region of the input is connected to a neuron in the output. Different filters are applied on each layer, and the results are combined later. In a CNN, a filter corresponds to the number of nodes for each part of the matrix. Notably, CNN learns the values of its filters based on the task to perform.

Convolutional neural networks are often associated with computer vision, but they are also used in natural language processing (NLP). NLP is the branch of deep learning which includes speech recognition, natural language understanding and generation and aims to teach computers to process natural language data.

The principle of NLP can also be used in biological sequence analysis using 1D CNN, as shown by Sonderby et al. [61] and DeepLoc [62]. Unlike 2D CNN, in 1D CNNs, forward and backpropagation require simple array operations rather than matrix operations. As a consequence, the computational complexity of 1D CNNs is significantly lower than 2D CNNs [63].

**Recurrent Neural Network and Long Short Term Memory Cell**

In a traditional neural network, all inputs and outputs are independent of each other. Recurrent Neural Networks (RNNs) perform the same task for every element of a sequence, and the output depends on the previous computations. RNNs make use of sequential information, and they are ideal for processing sequences (Figure 3.5). For example, predicting the next word in a sentence is easier after knowing which words came before it. RNNs have a memory capturing the information already calculated, which, theoretically, concede to process arbitrarily long sequences.
Regular RNNs are challenging to train due to the so-called vanishing gradient problem [64]. They often fail to capture dependencies far apart in the sequence [64], and they cannot hold information from multiple steps back [17]. A variant of the RNN cell, the Long Short-Term Memory (LSTM) (Figure 3.5), solves this problem using "computer memory cell" that holds information for multiple steps.

Several methods have taken advantage of this type of network in bioinformatics [65, 66] and have been successfully applied to the prediction of signal and mitochondrial targeting peptide [67, 68]. They make use of bidirectional RNNs (BiRNN), in which one RNN processes the sequence forwards and another processes the sequence backwards. For example, in biological sequence analysis, this type of network can model the context around each amino acid. The forward RNN processes all the amino acids from the N-terminus up to one position and the backward RNN processes all the amino acids from the C-terminus up to the same position [17].

![Diagram of Bi-Direction LSTM](image)

**Figure 3.5: Example of a Bi-Direction LSTM.** Bidirectional LSTM process the input forwards and backwards and then both outputs are merged. This structure provides the output layer with complete "past" and "future" information. It models the context around each position and can capture long-range dependencies.
**Attention mechanism**

Attention is a mechanism often combined with the RNN. Attention mechanism allows the network to focus on relevant parts of the input sequence when predicting a specific section of the output sequence.

This strategy was presented for the first time in 2014 by Bahdanau et al and enabled more natural and higher quality learning [69]. This was applied in machine translation for combining hidden state information from an encoder-decoder RNN [69].

In the paper by Sonderby et al, they used a similar approach for biological sequence analysis. They implemented a new function that assigns importance to each hidden state from a bidirectional LSTM [61]. The purpose is to give a different score to each position in the sequence. The function can provide a high score to more essential and lower rating to the less important parts of the sequence.

Today, this concept is further applied to biological problems such as n-terminal sorting signals and subcellular localisation prediction [62] to improve the performance by focusing on relevant biological signatures.

### 3.3 Practical machine learning techniques

From the perspective of a machine learning user, there are two essential steps after collected and cleaned all the relevant data. First, we need to encode the data in the best possible way. Second, we use Cross-Validation to assess how accurately a machine learning model will perform in practice. Consequently, we need to use the correct metrics to evaluate the machine learning model. In this section, I describe these essential steps.

#### 3.3.1 Data Encoding in Bioinformatics

The representation of the data influences the model performance. Thus, it is crucial to represent biological sequence data accurately. In bioinformatics, there are three common ways to represent a se-
sequence: one-hot-encoding, encoding using BLOSUM matrix [70] and encoding using sequence profiles.

In one-hot-encoding, we describe an amino acid by a vector of length 20, containing a single one and 19 zeros [71].

An alternative is to encode an amino acid using its corresponding row in the BLOSUM matrix [70, 72]. One advantage of this encoding is that the BLOSUM matrix captures information about which pairs of amino acids are easily interchangeable during evolution [71]. Without knowing which amino acid positions are highly conserved and which are variable, we exclude a priori important evolutionary information on a protein family [71].

Evolutionary information can also be included using sequence profiles, which are generally created by running PSI-BLAST [73] against a reference database [71]. A bottleneck with a sequence profile is the speed. They are shown to be helpful in some applications (secondary structure prediction [74] on top) but to the expense of the speed.

3.3.2 Cross-Validation

Cross-Validation (CV) is a statistical method used to evaluate and compare machine-learning algorithms [75]. In a typical CV procedure, we divide the data into N parts (at least two but generally more). In the simplest scenario, where we divide our dataset into two parts, one part is used to train a model, and the other is used to validate the model. The training and validation sets are used in successive rounds. In this way, each data point has a chance of being validated against but never used for training and validation at the same time [75].

A common type of cross-validation is the k-fold cross-validation, where K is equal to the number of splits (generally from 3 to 10). We first partition the data into k groups. Then, we perform k iterations of training and validation such that we use k-1 folds for learning. Here, a different subset of the data is held-out for validation in each iteration [75].

The k-fold nested cross-validation is a variant of cross-validation. If we adopt 5-fold for instance, four groups are used to train/validate the model and to identify the best set of hyperparameters. We use the
fifth group only to assess the performance. We repeat this procedure using all five subsets as the test set. Although we have to train 5x4=20 models, we obtain an unbiased test set performance on the whole dataset.

### 3.3.3 Evaluation of the performance

To obtain a comprehensive evaluation of the performance, we can use several measurements. All these measurements are generally expressed in terms of “tp” = true positive, “tn” = true negative, “fp” = false positive, “fn” = false negative.

Single class performance is often evaluated using precision and recall.

\[
\text{precision} = \frac{tp}{tp + fp} \quad (3.1)
\]

\[
\text{recall} = \frac{tp}{tp + fn} \quad (3.2)
\]

However, due to its ability to separate one type of class from all other class, the Matthews Correlation Coefficient (MCC) is a better metric [76].

\[
\text{MCC} = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fn)(tp + fp)(tn + fp)(tn + fn)}}. \quad (3.3)
\]

Methods for evaluating multiple classifications are not as well established. Among the most used metrics, there is the weighted F\(_1\) score. In the multi-class case, the F\(_1\) score corresponds to the weighted average of the F\(_1\) score for each class, where the average is weighted by the number of true instances for each label.

\[
F_1 = 2 \left( \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \right) \quad (3.4)
\]

We can use the overall accuracy (Q) as well. The overall accuracy is computed as the average of the fraction of correct predictions in each
Another measure is the Generalized Squared Correlation ($GC^2$) [76].

\[
GC^2 = \frac{\sum_{ij} (z_{ij} - e_{ij})^2}{e_{ij}}\frac{1}{N(K-1)}
\]  

(3.6)

where

\[
e_{ij} = \frac{x_iy_j}{N}
\]

is the expected number of data in row $i$ and column $j$ in the contingency table under the null hypothesis assumption [76, 78]. In the $GC^2$, the null hypothesis implies no correlation between observations and predictions [76, 78].

The True Positive Rate versus the False Positive Rate is also commonly used. In multi-class problems, first we assess the performance for a single class considering pairwise comparisons (one class versus all other classes). Then, the resulting sets are combined and used to evaluate the overall performance [79, 78].
Chapter 4

History and concepts of subcellular localisation prediction

From the first attempt to predict the signal peptide (SP) over 30 years ago, the field has progressed in steps [80]. Since then, prediction of sorting signals and cleavage site positions as well as of the subcellular localisation of proteins have gained massive popularity in bioinformatics. The approaches to predict the sorting signals or the subcellular localisation of proteins from the protein sequence alone are mainly three: signal-, global property- and homology-based [81]. Moreover, tools based on deep neural network architecture are becoming popular in the field.

4.1 Signal-based

Signal-based methods aim to predict the subcellular localisation site of a protein by recognising the actual sorting signals, as the cellular machinery does. A popular method using this approach is TargetP, which discriminates between proteins destined for the mitochondrion, the chloroplast, the secretory pathway, and other localizations [40].
The same approach was used in ChloroP for predicting chloroplast transit peptides [82] and in LumenP for the prediction of proteins targeted to the thylakoid lumen [83] and of their cleavage sites.

4.2 Global property-based

Global property-based methods identify the subcellular localisation of a protein by looking at the physico-chemical properties and composition of the amino acids. Generally, these methods use a fixed number of parameters to count the occurrences of amino acid pairs, either adjacent or separated by a small distance up to five positions. When used, this distance has shown an improvement in predictive performance. A drawback of this approach lies in excluding important information such as possible signatures of actual sorting signals from the actual prediction.

In the beginning, global property-based methods were mainly used to discriminate between intracellular and extracellular proteins [81]. Today, examples of such methods are CELLO 2.5 [84], WoLF PSORT [85] and Mitofates [86].

CELLO 2.5 [84] incorporates physico-chemical properties such as amino acid composition, dipeptide composition or partitioned amino acid composition and sequence composition into a multi-class SVM classification system to predict the subcellular localisation of a protein.

WoLF PSORT assigns localisations based on sorting signals, amino acid composition and functional motifs using a k\textsuperscript{th}-nearest neighbours algorithm [85].

Mitofates combines amino acid composition and physico-chemical properties with positively-charged amphiphilicity, pre-sequence motifs, and position weight matrices [86] to predict the presence of mitochondrial targeting peptide.
4.3 Homology-based

The homology-based approach starts from the assumption that a protein tends to stay in the same compartment in the course of evolution. However, it is not trivial to determine how similar a pair of proteins has to be to infer the possible subcellular localisation. The natural consequence of this is to use the ability of sequence alignment programs such as BLAST [73].

Using BLAST [73], it is possible to transfer the subcellular localisation annotation from the best hit to the query. Tools like LocTree3 [87] and Locnuclei [88] follow this approach.

Another option, demonstrated in SubNucPred [89] involves checking the presence or absence of unique Pfam [37] domains in the query sequence.

Other methods, including MultiLoc2 [90], Sherloc2 [91] and YLoc [92] use a different approach. They derive annotation from keywords or functional descriptions, or titles and abstracts from the literature references. In essence, this means to infer subcellular localisation from the annotation of homologues, which do not necessarily have experimentally determined subcellular localisation [81].

4.4 Deep learning-based

Computational methods, used in the prediction of subcellular localisation and sorting signals, exploit different machine learning methods including Hidden Markov Models, Artificial Neural Networks [40], Support Vector Machine [91], $k$th-nearest neighbours and naïve Bayes algorithm. However, today, thanks to more extensive and more accurate experimental datasets, tools exploiting deep neural network architecture are becoming popular and start to dominate the field. These tools have better performance compared to their ancestors.

In 2017, DeepLoc opened a new era in the prediction of subcellular localisation by using convolutional neural networks and long-short memory cells with attention mechanism [62]. Unlike previously developed methods, DeepLoc learns the subcellular localisation of the
protein starting from the amino acid sequence without including homology information, physico-chemical properties, or amino acids composition. Here, the network determines the subcellular localisation by focusing on the relevant part of the sequence.

From 2017, other methods using deep neural network have been developed. DeepSig uses convolutional neural networks to predict the presence of signal peptides [93], while DeepMito to identify protein sub-mitochondrial localisation [94].

The fifth version of SignalP uses long-short memory cells and conditional random field to distinguish between various types of signal peptides across all domains of life and between three kinds of prokaryotic signal peptides (Sec/SPI, Sec/SPII and Tat/SPI) [68].
Chapter 5

Results and Discussion

The objective of this thesis is to develop better computational methods for the classification of the subcellular localisation of eukaryotic proteins.

In the first paper, I describe how we can improve the performance of previous methods by using Random Forest classification. The ensemble method (SubCons) exploited the strengths of four methods and was trained with highly-accurate human experimental protein data.

Next, I present the web-server of SubCons and its performance when applied to a different organism (*Mus musculus* and *Drosophila melanogaster*). An advantage of deep neural networks is their ability to extract and learn features without being explicitly told. As shown in paper III, if properly trained, deep neural networks allow us to focus on important biological features in the data otherwise omitted.

Finally, in paper IV, I demonstrate that we can improve the training of deep neural networks by adding highly homologous proteins to the small number of experimentally verified proteins.
5.1 SubCons: a new ensemble method for improved human subcellular localisation predictions (Paper I).

In this paper, we try to improve the prediction of human protein subcellular localisation using an ensemble method that combines four predictors using a Random Forest classifier (Figure 5.1).

Figure 5.1: The figure shows the SubCons workflow. SubCons is a Random Forest classifier, which combines predictions from four independent predictors. The input of SubCons is a fasta sequence(s) and the corresponding profile from the multiple sequence alignment. The scores of the predicted localizations are used as an input for a Random forest classifier. The figure is taken from [78].

To train and validate SubCons, we used data from two large-scale studies and the experimentally verified annotation in UniProt. Additionally, from this data, we assembled a test set ("Golden dataset"), where two independent methods confirm the subcellular localisation. Using the "Golden dataset", we compared the performance of state-of-the-art methods and SubCons.
In a total of nine subcellular localisations, SubCons achieves an $F_1$-Score of 0.79 compared to 0.70 of the second-best method. Furthermore, at an FPR of 1% the true positive rate (TPR) is over 58% for SubCons compared to less than 50% for the best individual predictor [78].

5.2 The SubCons web-server: A user-friendly web interface for state-of-the-art subcellular localisation prediction (Paper II).

In this paper, we present the user-friendly web-interface implementation of SubCons [95]. The server predicts the subcellular localisation of an individual protein starting from the sequence. The server can handle batch submissions of multiple proteins either by uploading the files or programmatically by using a command-line WSDL API script. Therefore, SubCons is ideal for proteome-wide analyses since the user can scan a whole proteome in a few days.

Additionally, we evaluate the performance of SubCons and LocTree3 [87] using protein annotation from mass-spectrometry studies in two eukaryotic organisms (Mus musculus and Drosophila melanogaster). We also provide pre-calculated predictions for several eukaryotic organisms that are possible to download.


TargetP is a state-of-the-art method to identify N-terminal sorting signals, which direct proteins to the secretory pathway, mitochondria and chloroplasts or other plastids. The first version of TargetP was based on two layers of feed-forward neural networks to predict N-terminal sorting signals and cleavage site position of proteins with Signal Peptide (SP), mitochondrial targeting peptide (mTP), chloro-
plast targeting peptide (cTP)) and proteins without a targeting peptide (noTP). After 19 years, we updated TargetP [40] using bidirectional long short-term memory cell and a multi-head attention mechanism (Figure 5.2). Unlike the first version, TargetP 2.0 is also able to predict thylakoid targeting peptides (ITP).

Figure 5.2: The figure shows the TargetP 2.0 architecture. TargetP 2.0 consists of a bidirectional LSTM and a multi-attention mechanism. TargetP 2.0 predicts both the type of peptide and the position of the cleavage site. The figure is taken from [17].

Next, we show that deep neural networks could also provide new insights into the underlying biology. By examining the output values from the attention layer in the network, we find that the second residue, the one following the initial Methionine, has a strong influence on the classification. Looking at chloroplast and thylakoid targeting peptides, we observe that the frequency of Alanine in position two is very high (two-thirds compared to other amino acids) (Figure 5.3). In Fungi and single-celled eukaryotes, we note that less than 30% of the targeting peptides have an amino acid that allows the removal of the Methionine in position one. In contrast, 60% of the proteins without targeting peptide have an amino acid that enables
the removal of the N-terminal Methionine.

Figure 5.3: The figure shows that the second residue, the one following the initial Methionine, has a strong influence on the classification. In the figure, the proteins are divided into SP, mTP, cTP, ITP and proteins without a targeting peptide (noTP). Additionally, the proteins are grouped according to Fungi (F), Metazoa (M), Plantae (P) and other eukaryotic organisms (O). In the figure, the height of each letter corresponds to the frequency of the short side-chained amino acids that allow the cleavage of the N-terminal Methionine. The figure is taken from [17].
5.4 Improved sub-nuclear prediction by Deep Learning using an augmented dataset (Paper IV).

In this paper, we present a new deep neural network-based method (Figure 5.4) able to accurately predict sub-nuclear localisation of a protein as well as nucleo-cytoplasmic shuttling proteins only from the sequence.

Figure 5.4: Description of the architecture of NudeeP. The model consists of blocks of parallel convolutional layers with filters of different sizes. The resulting output layers are concatenated, and a dense layer is finally used to predict the types of class.
Next, we generated an augmented dataset by adding highly homologous proteins to the small number of experimentally verified proteins. Using this dataset for training, we show that the fraction of correct predictions is increased from 18% to 51%. In comparison, earlier methods predict 18-29% of the proteins correctly.

By using only sequence feature, NudeeP has a few advantages compared to earlier methods. First, it is faster. Second, it solves the problem of accurate prediction of proteins without closer homologs. Therefore, it makes it possible to predict accurately unknown protein without directly relying on homology-based inference. Moreover, it would be possible to use the power of convolution neural networks to discover the intrinsic properties of these type of proteins without being explicitly programmed. We believe that NudeeP can be a step towards understanding the biological functions of sub-nuclear and nucleo-cytoplasmic shuttling proteins in the cell.
Sammanfattning


Den eukaryotiska cellen är väl organiserad med flertalet membranomslutna organeler. Dessa organeler är interna divisioner av cytoplasman i flera olika delar, såsom endoplasmic reticulum, golgiapparaten, lysosomer, mitokondrier, kärnan, peroxisomen och kloroplaster (som endast existerar i plantor). Det finns även organiserade delar utan ett tydligt omslutande membran som nucleoli, cajal kroppar, nukleära fläckar och PML kroppar. Alla dessa delar är omsluten av cytoplasman.

Proteiner har specifika funktioner beroende på var de är lokaliserade i cellen. Deras transport till rätt plats är därför livsviktig för cellens överlevnad. Specifika signaler i proteinets aminosyrasekvens dirigerar proteiner till olika subcellulära delar.

Det är ytterst viktigt att korrekt identifiera och annotera proteiners lokalisering med både experimentella och databaserade metoder. Utvecklandet av både experimentella och beräkningsmetoder för att klassificera proteiners subcellulära lokalisering har i över 30 år varit...
ett viktigt forskningsområde.

Acknowledgements

The past five years in Stockholm have flown. However, the entire period would not have been the same without the support of extraordinary people around me.

This journey would not have been started without the support of my supervisor, Arne. Your patience, guidance, ideas, criticism and encouragements helped me move forward as a person and as a researcher. It was fun to discuss both scientific and non-scientific things, and your hints have always been helpful. It was also lovely to spend a few moments together during our trips to India and Greece. I had a great time in Bilbao together with you and Susanna while learning about Basque culture and eating pintxos. It has been a great pleasure being part of your group and learn from you. I will always be grateful to you.

Many thanks go to my mentor, Gunnar, for all the knowledge you shared, not only for the TargetP project. During, the few times we had the chance to talk, I learnt a lot, and your endless experience is of inspiration. It has been great to learn from you.

The same thought goes to Henrik. I learnt a lot in the course of the TargetP project, and your immense expertise in the field has been of great help. My short visit to DTU was fantastic, and I enjoyed discussing scientific and danish matters.

Many thanks Ole. Listening to your extensive expertise in machine learning has been of great help. I am looking forward to continuing when I will come to Copenhagen.
Nanjiang without you, none of our tools/servers would be running. Thanks for the vast work you do/did, for the help and the thoughts when we were publishing SubCons. I can say that you taught me how to be a more organized computational scientist.

David, we started and finished almost at the same time. I had an excellent time when we went to Bilbao and India together. Thank you also for having shared your knowledge about deep and transfer learning.

Sudha thank you for always having a nice thought to me and for being a very nice person to my family and me. Both you and Narayanan are genuine and lovable people. I wish Tharu, Narayanan and you all the best.

Saman thank you for being a very charming colleague. I enjoyed the time discussing politics, religion, food and science.

Claudio, thank you for being very calm even when I gave you a hard time for your half Austro-Bavarian-German nationality. Despite the jokes, I had a pleasant time with you in the lab, during our barbecues, beers, work trips and chats. Although we have a completely different view about politics, it was always pleasing and constructive to discuss with you. The only thing I hope is that you will learn how to cook a proper carbonara!

Gabriele, thank you for being one of the loudest and most positive people I have ever met. Our after-work commutes home are unforgettable (probably, now all Stockholm knows about how to make proper burgers and how to train deep neural networks). The time spent together with you and Melania was enjoyable. Next challenge for you will be to cook arrosticini properly!

My previous experience as a football player taught me that you always learn and improve when surrounded by several people. The same was during my time in this group. I want to thank also the new members Patrick and Julie to have brought fresh air and new insights in the group, and all the past and in some way associate group members; Sara, Minttu, Per, Karolis, Oxana, Alba, Rafael, Alok, Petras, Erik, Samuel, and the bachelor, summer, visiting and master students.
Thank you Jose. During your external and my short DTU period, I learnt a lot about deep learning from you. It was fun and great to collaborate with you. We also had an excellent time together in Stockholm, Copenhagen and Paris.

Thank you Michael for being always gentle with me and for organising the trip to India. My visit to Chennai and at the IIT Madras was a great experience.

Thanks Marie. The conversations with you about Swedish matters and not, were always fascinating and your positivity is contagious.

Thanks also to Christian, Magnus and Björn and to all the people in Alpha 6 to make a nice and stimulating environment.

I want to thank also a few other people that were or are at DBB/SciLifeLab; Roger Salvatori, Christoph Ogris, Daniel Morgan and Dimitri Guala.

During my PhD, I also had the chance to be part of the PhD council of SciLifeLab. It was great to organize events with you Alba, Dörte, Karen and Marianna.

Thank you to all the fantastic people at DBB. A special thanks to Alexander Tulling and Maria Sallander for being always kind, professional and extremely helpful.

Mikael Oliverberg, Daniel Daley and IngMarie Nilsson, thank you for allowing me to be teaching assistant in your courses together with the amazing PhD students Riccardo, Theresa and Sarah. It has been a great experience and I have met awesome students.

During the various courses in which I was teaching assistant, I have been lucky to meet amazing students. I wish all the best in your future careers; Sarah, Urska, Linnea and Helen.

On a more personal level, I want to thank also other people that I had the pleasure to meet during the past five years.

Walter, you have been my first contact with Sweden. I still remember the call we had to try to find a project for my master thesis. Although we did not find one, we managed to collaborate in some successful projects during our doctoral studies. Your passion for computational-related things was contagious, and in the beginning, I
learnt a lot from you. You also introduced me in most of the Swedish social/working/cultural related matters and all your hints were extremely helpful. For this, I will be indebted to you. Thank you for the lovely time we spent together outside the lab and to have let me know your adorable daughter.

Thanks, **Alfredo** to share so many notions and knowledge with me. Our days at SciLifeLab and our Skype calls were always exciting and plenty of fruitful discussions with some pure bullshit moments in between. One day, I will come to California to continue.

**Mirco**, you provided me with insightful notions, and you were incredibly calm even when I gave you a hard time for cluster related matters. Thanks for being a fantastic colleague and for the help during my PhD. I enjoyed the time we spent outside the lab both at our and your home. I wish Elli, Lia, Nina and you the best in your lives.

**Axel, Christoph** and **Stefan** thanks for all the conversations about thousand of different topics during our lunches, dinners, barbecues, pubs and beers. It was great to spent time together at SciLifeLab and outside.

Thanks a lot, **Anirudh** for many fun days spent discussing various topics from cuisine, culture and football. Last but not least, thanks for the hospitality of you and your family during my trip to India. It was a great time, and I will never forget it. I wish you all the best in your life.

Thanks, **Carsten** to always have time for me, for sometimes acting as a mentor when I came to your office and for the delightful dinners and barbeques, that you and your wife **Hanna** organised. Thanks to nicely open the door of your home to me as I was part of your group. The invitation for presenting my projects at the Riken-KI symposium allowed me to meet many interesting people, and partially take the way to the next step to Copenhagen. I will always be grateful to you.

Also, thanks to the **Carsten Daub group** for welcoming me every time I came to visit and to invite me to many dinners. Thanks **Matthias**, it was nice to discuss many topics with you over cakes
and dinners, especially when you were telling about your bike trips.

Thanks Kadir for many entertaining afternoon/nights spent playing football, bowling, pool and discussing various topics. Although you never won against me in any matches (smiling), you were always a sincere and good friend to Enrichetta and me. I wish you all the best in your life.

Thanks, Niyaz, it was great to make jokes about Kadir every time we played and ate together, with the support of Shalina in the last period. You guys have always brought happiness and smiles in our dinners. I wish you all the best in your lives.

Thanks, Serena, for all the coffees together and for many fun lunches, dinners and evenings spent discussing various topics and listening to your fun adventures. Both Enrichetta and I will wait for you in Copenhagen to continue. I wish you all the best in your life.

The past years would not have been the same without a few special people I have the luck and the honour to meet. We often say that “friends are the family you choose”. This is the case with you Kostas. From the first day, you helped (moving furniture, eating, smiling and many more things) and showed your big heart. When you became older, it is hard to make new friends. This is for sure wrong when I think about you and Stephanie. I always had the feeling that I could count on you for everything. Thanks for being genuine and honest friends. We had and will still have great moments together. I wish you all the best in your life together, and I hope to keep being part of it. PS: your videos when I told you about Luce and when she was born are masterpieces.

John you always said that “you need to show respect to your colleagues, but you do not have to befriend with them necessarily”. Despite being a colleague, we also became friends. I enjoyed our early morning conversations about all possible topics, our trips and various excellent drinks. Thanks for having translated all Swedish matters, introduced me in the whisky world and for pushing me to train again properly. It was great to spend the last three years together. Both you and Faye are and have been excellent and honest friends to Enrichetta and me. You opened the door of your home to us, and we
spent many great days with your kids. I wish you all the best in your lives.

Thanks Alessandro and Ana Maria for being special friends and always have a thought for me and Enrichetta. The great lunches, dinners and days together are unforgettable (I will miss those, but perhaps we can replicate in Copenhagen...who knows). You are/were always kind and supportive, and I know I can count on you every time. It is not easy to find people like you that are so respectful, direct and honest. I wish you all the best in your lives.

Un ringraziamento speciale va a voi, mamma e papá. Il vostro supporto incondizionato ha contribuito in maniera indelebile all’uomo che sono oggi. La vostra voglia innata di migliorare e fare qualsiasi cosa al meglio delle possibilità mi hanno sempre spornato anche nel raggiungimento di questo traguardo. Sono fortunato, orgoglioso e grato di avere due genitori come voi.

Grazie a te Andrea. Tu sai di essere molto più di un fratello. Devo a te in parte quello che sono oggi come uomo e ricercatore. Mi hai insegnato molto e soprattutto all’inizio della mia carriera, la tua conoscenza informatica mi è stata di grande aiuto. Grazie a te e Michela per aver contribuito con risate e consigli al mio percorso in Svezia e dato alla luce il mio fantastico nipotino Niccolò.

Grazie a Tio Diogenes Vicente. Come sempre, anche in questi ultimi cinque anni hai dimostrato di essere uno zio meraviglioso. Sei venuto a trovarmi, mi hai ascoltato, scritto, supportato e mai fatto mancare parole incoraggianti. Alcune volte la tua curiosità innata mi ha aiutato a vedere i problemi con un occhio alternativo. Sono grato ed fortunato ad avere uno zio come te.

Grazie anche a tutta la famiglia Salvatore per avere fatto festa ogni volta che tornavo e farmi sentire come se non fossi mai andato via. Grazie soprattutto a zio Fausto e Valeria per tutte le volte che avete avuto un pensiero e organizzato una festicciola per me ed Enrichetta nonostante le tante difficoltà. Grazie anche a zia Elisa, Stefano e Luca che con i loro messaggi mi sono stati vicini in tutto questo periodo.

Grazie al mio miglior amico Marco per essermi stato accanto in
tutti questi anni. Anche se non sei mai venuto a trovarmi a Stoccolma (spero almeno tu venga a Copenaghen!), ho sempre avuto il tuo supporto e la sensazione che fossi qui con me. Tutti dovrebbero avere un amico come te e io sono estremamente fortunato e grato per questo.

Grazie anche ad Italo, Marina e Nella per il vostrro costante supporto. Specialmente a te, Nella, che ci hai aiutato e sopportato soprattutto durante l’ultimo periodo e mi hai permesso di finire di scrivere questa tesi in estrema tranquillità.

Enrichetta grazie di essere la persona meravigliosa e speciale che sei. “Dietro ogni grande uomo c’è sempre una grande donna”. Non esiste frase più adatta per descriverti. Tutto questo non sarebbe mai potuto accadere senza il tuo costante supporto, aiuto e amore. In questi dodici anni abbiamo sofferto, lottato, riso, gioito, viaggiato e conosciuto tanto insieme. Ci siamo spinti in diversi continenti per migliorarci e crescere...ma tu sei sempre stata al mio fianco forte e decisa. Non solo una mamma eccezionale, sei anche la persona più brillante che io conosca. Il tuo aiuto è stato fondamentale in molti aspetti e non di meno in quello scientifico. Devo a te molto del mio dottorato. Tu e il nostro amore Luce rendete i miei giorni straordinari. Grazie di tutto. Siete la mia vita e vi amo con tutto me stesso.
References


58


