

LITERATURE REVIEW OF PROTEIN LOCALIZATION PREDICTION

INTRODUCTION

Over millennia, eukaryotic cells have evolved specialized regions and compartments to complete specific biological tasks. These regions include the plasma membrane, the nucleus, the mitochondria, the endoplasmic reticulum and many more. To perform a certain function, each of the compartments must have task specific proteins that do not function when in the wrong compartment. For example, transcription factors bind to DNA to signal the start of mRNA transcription. Since DNA is only located in the nucleus, a transcription factor outside of the nucleus cannot function. Also, since the main activities in each cellular compartment are known, some guesses can be made about what the proteins localized to the area might do (i.e. A protein localized to the nucleus is likely somehow related to DNA.). There are many biological techniques to determine the sub-cellular localization of a protein, and without having an idea of the possible localizations, the scientist may do many extra experiments. Also, protein sequence data is being published so fast that researchers cannot keep up experimentally determining the function for each protein. Computational methods of predicting protein sub-cellular localization are important to give researchers an idea of a protein's function without doing any bench work. This review will cover eight computational tools that predict protein localization in eukaryotes.

SUBCELLULAR LOCALIZATION BACKGROUND

Before reviewing the different protein localization prediction tools, the biological basis of protein sorting and localization should be understood. After each mRNA is transported out of the nucleus, translation begins on cytosolic ribosomes. At this point, there are two possibilities:

one, proteins without an endoplasmic reticulum signal sequence continue being translated in the cytoplasm, or two, proteins with an endoplasmic reticulum signal sequence stop being translated until they are transported to the endoplasmic reticulum membrane (Figure 1) (Lodish 2004).

The proteins that complete translation in the cytosol have several possible final destinations; they could remain in the cytosol or they could be transported to one of several organelles: either the mitochondria, chloroplast, peroxisome, or nucleus. If a protein does not have any signal sequence at all, it will remain in the cytosol to do its job. This pathway could be called the default protein pathway. Many cytoskeletal proteins fall into this category (Figure 1, Table 1) (Lodish 2004).

The mitochondria, nicknamed “the powerhouse of the cell” are important for respiration, or the conversion of sugars into usable energy. Since the organelle likely resulted from a symbiotic relationship (the mitochondria originally lived inside the whole cell as its own organism) between the ancestral mitochondria and the ancestral cell, mitochondria have some of their own DNA and ribosomes. There are, however, some proteins important for mitochondria function that are encoded in the nucleus. These mRNAs do not have an ER signal peptide, so are translated in the cytosol, but they do have a mitochondria specific signal sequence. This sequence is always at the N-terminal region of the protein and contains three to five arginine and lysine residues mixed in with threonine and serine. The sequence never contains glutamic or aspartic acid residues. The uptake sequence is bound by a chaperone protein, which guides the protein to the mitochondria, where the signal sequence is cleaved (Figure 1 and Table 1) (Lodish 2004).

Chloroplasts are similar to mitochondria in many ways, except since they are involved in photosynthesis, they only occur in eukaryotic autotrophs (organisms that make their own food

from light). Again, these organelles likely resulted from a symbiotic relationship so they encode many of their own proteins. Like mitochondria, nuclear encoded chloroplast proteins have a specific signal sequence that is bound by a chaperone to bring the protein to the organelle. The sequence is then cleaved. This signal sequence is also located at the N-terminus, but there are no common sequence motifs. In general, the sequence has many serine, threonine and small hydrophobic residues and few glutamic and aspartic acid residues (Figure 1 and Table 1) (Lodish 2004).

Unlike mitochondria and chloroplasts, peroxisomes do not have any of their own DNA, so all their proteins must come from nuclear mRNAs. Peroxisomes have varying sizes and components, but all of them have enzymes that oxidize substrates using oxygen. For example, the protein catalase makes water from hydrogen peroxide. Peroxisomal proteins have a C-terminal signal sequence that is not cleaved when the protein arrives. Usually, a serine-lysine-leucine motif is bound by a chaperone to facilitate transfer into the peroxisome (Figure 1 and Table 1) (Lodish 2004).

The final location in the cell that does not require an ER signal sequence is the nucleus. Even nuclear proteins are translated in the cytosol and must be transported back into the nucleus to do their jobs. Once again, proteins need a nuclear localization signal (NLS) but this sequence can be located anywhere in the protein sequence. The signal is not cleaved and the transport mechanism is somewhat more complicated than the chaperone binding of the other organelle targeting sequences. The NLS is bound to an importin, which interacts with other proteins to enter the nucleus through a nuclear pore. The NLS is composed of a group of five basic amino acids or two small groups of basic amino acids separated by about 10 residues (Figure 1 and Table 1) (Lodish 2004).

The other main sorting pathway is for proteins that have an endoplasmic reticulum (ER) signal sequence. This signal is at the N-terminal portion of the protein and contains a positively charged N-terminus, a hydrophobic middle region, and a polar C-terminal region (Emanuelsson 2001). As soon as the signal sequence is translated, translation stops and the protein/ribosome complex is transported to the ER. There, translation is completed while the new protein is pushed into the ER lumen. Here, the signal peptide is cleaved (Figure 1 and Table 1) (Lodish 2004).

All proteins that enter the endoplasmic reticulum are transported to the Golgi apparatus for modifications unless the protein has a ER retention signal. These retention signals can vary, but they keep ER proteins from ending up in the Golgi. Similarly, Golgi specific proteins need a Golgi apparatus retention signal to prevent resident proteins from accidentally being secreted or sent to the lysosome. This Golgi retention signal's characteristics can also vary (Figure 1) (Lodish 2004).

Once the protein is done being modified in the Golgi, one of two things can happen. One, if mannose-6-phosphate is added to the protein in the Golgi, it will be sent to the lysosome, where the cell digests and breaks down a variety of molecules. If the protein contains the ER signal sequence only, the default localization is extracellular. These secreted proteins can be cell-cell signaling molecules, extracellular matrix proteins or other components needed outside of the cell. After being translated in the ER, they are modified and folded in the ER and Golgi apparatus before being released as secretory vesicles (Figure 1) (Lodish 2004).

The final point to make about cell localization is about membrane proteins. All membrane proteins have a ER signal sequence and are translated into the ER. They are inserted into the membrane while being translated. Then, the membrane protein must have the ER

retention signal to remain in the ER, the Golgi retention signal to remain in the Golgi, a mannose-6-phosphate modification to go to the lysosome, or no additional signal to end at the plasma membrane. Membrane proteins are never secreted because they are membrane anchored (Lodish 2004).

REVIEW OF SUBCELLULAR LOCALIZATION PREDICTION TOOLS

With an understanding of the biological principles of protein sorting, a better review can be made of protein localization prediction tools. In this review, only a subset of all the programs available are covered. The programs were chosen because all of them predict at least four cell localizations, each were mentioned in at least one published review paper (Emanuelsson 2001, Donnes 2004, Schneider 2004, Klee 2005), and all have high accuracy. Since all the prediction programs' papers reported high accuracy, used different methods to predict their accuracy, and have not all been benchmarked, this review will not attempt to pick one program over another. Instead, it will show the variety of prediction tools available, what they use to make predictions, how they make predictions, and how the programs have developed over the last few years.

There are four main approaches to predicting sub-cellular localization. One, the module can look for signal sequences unique to each compartment. Second, the program can look for subtle differences in overall sequence that are consistent in one compartment and different between compartments. The tool can use homology and evolutionary similarities to predict localization and lastly, it can use some combination of the mentioned methods to predict location (Emanuelsson 2001, Hogland 2006).

TargetP is one of the older prediction programs, first published in 2000. It stands for Target Peptide and predicts cellular destinations and cleavage sites based on N-terminal

sequence information only. Trained with eukaryotic sequences from SWISS-PROT with known localizations, this neural network based tool only predicts four localizations: mitochondria, chloroplast, secretory pathway, and other (Emanuelsson 2000).

In 2003, LOC3D, or LOCalization of proteins with 3D structure, was released as a program that used a progressive, four pronged approach to predict localization. It uses the predictNLS tool to look for nuclear localization signals; it uses LOChom to look for sequence homology to proteins with known localization; it uses LOCKey to look for SWISS-PROT keywords; and it uses LOC3Dini, which is a support vector machine and neural network tool to predict localization based on sequence. Of the four approaches, the one that gives the highest confidence score for the predicted localization is the final prediction. LOC3D is unique since it only used eukaryotic proteins from Protein Data Base (PDB) with known 3-dimensional structure and known localization to train the program. LOC3D predicts 10 possible locations in the cell: extracellular space, cytoplasm, nucleus, mitochondria, chloroplast, peroxisome, lysosome, endoplasmic reticulum, vacuoles, and Golgi (Nair 2003).

Another program that predicts localization based on N-terminal signal sequences is Predotar, which was started in 2004. Predotar, for PREdiction of Organelle TARgeting sequences, uses a neural network based approach to predict four sub-cellular locations: the plastid (plants), the mitochondria, the ER pathway, or other. It was again trained with SWISS-PROT sequences and is designed for plant, fungal or animal protein prediction (Small 2004).

PSLT, or Protein Sub-cellular Localization Tool was also released in 2004. It uses a Bayesian network trained with human sequences taken from SWISS-PROT. This program is nice since it predicts 9 localizations: the ER, Golgi, cytosol, nucleus, peroxisome, plasma membrane, lysosome, mitochondria, or extracellular, but is limited because it is only designed

for human proteins. PSLT's predictions are based on the presence or absence of InterPro motifs (a database of protein families, domains, functional sites, and post-translational modifications), signal peptides and transmembrane domains (Scott 2004).

A tool that predicts many sub-cellular localizations, but predicts them for all eukaryotes came out in 2005. LOCSVMPSI, which stands for LOCalization using Support Vector Machines and PSI-blast, used sequences from SWISS-PROT to train the program to predict 12 locations: the chloroplast, cytoplasm, cytoskeleton, ER, extracellular, Golgi, lysosomal, mitochondrial, nuclear, peroxisomal, plasma membrane, or vacuole. It uses evolutionary information gathered from a PSI-BLAST homology search to generate a position specific scoring matrix that is then entered into a support vector machine to make a prediction (Xie 2005).

The next localization prediction tool, also released in 2005, is pTarget. It predicts nine possible protein locations for animals, fungi and metazoans: cytoplasm, endoplasmic reticulum, extracellular, Golgi, lysosomes, mitochondria, nucleus, plasma membrane, and peroxisomes. This program looks for the presence of location specific domains using the pfam database and uses amino acid weight matrices to reflect differences in proteins' amino acid compositions in different compartments. Once again, this program was trained with SWISS-PROT sequences with known localization (Guda 2005).

The last two tools covered in this review were both published in 2006 and have only been reviewed in Carlos Sosa's unpublished work. The first one is BaCelLo, which stands for BALanced CELlular Localizations. It stores the entire sequence's composition and evolutionary information in an alignment profile which is entered in a support vector machine to generate a prediction. In order to balance out the bias from grouping all eukaryotes into one training set, it used three training sequence sets each for plants, animals and fungi from SWISS-PROT. This

tool only predicts five locations: secretory, cytoplasm, nucleus, mitochondria, and chloroplast (Pierleoni 2006).

The final program is MultiLoc, for MULTIPLE ways to predict LOCALization. It integrates N-terminal targeting sequences, amino acid composition and protein sequence motifs and inputs all this information into a support vector machine. Using SWISS-PROT training sequences from plants, animals and fungi, it can predict 11 localizations: chloroplast, cytoplasm, endoplasmic reticulum, extracellular, Golgi, mitochondria, nucleus, peroxisome, plasma membrane, vacuole (plants), and lysosomes (animals) (Hoglund 2006).

CONCLUDING REMARKS

There are many other prediction programs this review did not cover, many that were released in the last few years. It is impossible to even summarize all the programs available and it would be nearly impossible to objectively compare each program since they are so different. This review should help researchers appreciate the diversity and large number of protein prediction programs available as well as understand the basics of protein localization. Overall, protein sub-cellular localization predictions can be made using many different methods that all give a reported high accuracy; benchmark experiments must be done with all the programs to sort out which are best for predictions.

REFERENCES

- Donnes, P. and Hoglund, A. (2004) Predicting Protein Sub-cellular Localization: Past, Present, and Future. *Genome and Protein Bioinformatics*, **2**, 209-215.
- Emanuelsson, O. (2002) Predicting protein sub-cellular localization from amino acid sequence information. *Briefings in Bioinformatics*, **3**, 361-376.
- Emanuelsson, O. *et al.* (2000) Predicting Sub-cellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. *Journal of Molecular Biology*, **300**, 1005-1016.
- Emanuelsson, O. and von Hiejne, G. (2001) Prediction of organellar targeting signals. *Biochemica et Biophysica Acta*, **1541**, 114-119.
- Guda, C. and Subramaniam, S. (2005) Target: A new method for predicting protein sub-cellular localization in eukaryotes. *Bioinformatics*, **21**, 3963-3969.
- Hoglund, A. *et al.* (2006) MultiLoc: prediction of protein sub-cellular localization using N-terminal targeting sequences, sequence motifs, and amino acid composition. *Bioinformatics*, **22**, 1158-1165.
- Klee, E., and Ellis, L. (2005) Predicting eukaryotic secreted protein prediction. *BMC Bioinformatics*, **6**, 256-263.
- Lodish, H. *et al.* (2004) *Molecular Cell Biology: Fifth Edition*, W.H. Freeman and Company.
- Nair, R. and Rost, B. (2003) LOC3D: annotate sub-cellular localization for protein structures. *Nucleic Acids Research*, **31**, 3337-3340.
- Pierleoni, A. *et al.* (2006) BaCellLo: a balanced sub-cellular localization predictor. *Bioinformatics*, **14**, e406-e416.
- Schneider, G. and Fechner, U. (2004) Advances in the prediction of protein targeting signals. *Proteomics*, **4**, 1571-1580.

Scott, M. *et al.* (2004) Predicting Sub-cellular Localization via Protein Motif Co-Occurrence.

Genome Research, **14**, 1957-1966.

Small, I. *et al.* (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting

sequences. *Proteomics*, **4**, 1581-1590.

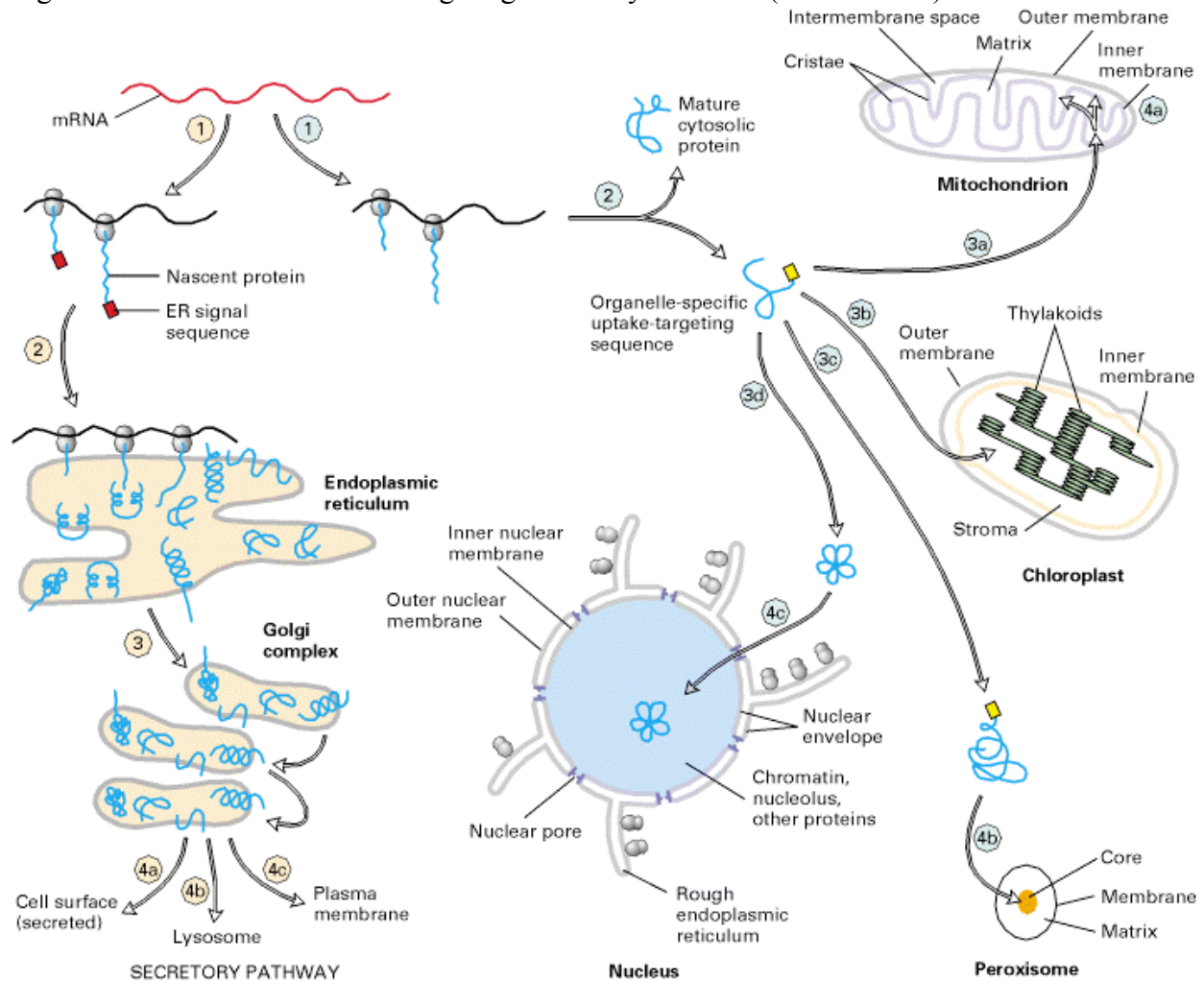
Xie, D. *et al.* (2005) LOCSVMPSI: a web server for sub-cellular localization of eukaryotic

proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Research*, **33**, W105-

W110.

FIGURES

Figure 1. Overview of Protein Targeting in Eukaryotic Cells (Lodish 2004)



TABLES

Table 1. Signal Sequences for Different Localizations

Sub-cellular Compartment	Signal Sequence Location	Signal Sequence Composition	Is Signal Cleaved?
Cytosol	No signal sequence	N/A	N/A
Mitochondria	N-terminus	3 – 5 arg and lys mixed with thr and ser, no glu or asp	Yes
Chloroplast	N-terminus	Many thr, ser, and small hydrophobic, few glu or asp	Yes
Peroxisome	Extreme C-terminus	Usually Ser-Lys-Leu	No
Nucleus	Anywhere	5 basic or two small groups of basic separated by about 10 residues	No
Endoplasmic reticulum	N-terminus	+ changed N-terminus, hydrophobic middle region, and a polar C-terminus	Yes

Table 2. Summary of Protein Sub-cellular Localization Prediction Programs

Program	Year	Algorithm	Organisms	Locations Predicted
TargetP	2000	Neural Network	Eukaryotes	4; mitochondria, chloroplast, secretory, other
LOC3D	2003	Combination	Eukaryotes	10; cytoplasm, mitochondria, chloroplast, peroxisome, nucleus, ER, Golgi, lysosome, extracellular, vacuoles
Predotar	2004	Neural Network	plant, animal, and fungi	4; mitochondria, secretory, plastid, or other
PSLT	2004	Bayesian Network	Human	9; cytoplasm, mitochondria, peroxisome, nucleus, ER, Golgi, lysosome, extracellular, plasma membrane
LOCSVMPSI	2005	Support Vector Machine and Position specific Matrix	Eukaryotes	12; cytoplasm, mitochondria, chloroplast, peroxisome, nucleus, ER, Golgi, lysosome, extracellular, plasma membrane, cytoskeleton, vacuole
pTarget	2005	Weight matrices	animals, fungi and metazoans	9; cytoplasm, mitochondria, peroxisome, nucleus, ER, Golgi, lysosome, extracellular, plasma membrane
BaCellLo	2006	Support Vector Machine	plant, animal, and fungi	5; cytoplasm, mitochondria, chloroplast, nucleus, secretory,
MultiLoc	2006	Support Vector Machine	plant, animal, and fungi	11; cytoplasm, mitochondria, chloroplast, peroxisome, nucleus, ER, Golgi, lysosome, extracellular, plasma membrane, vacuole

