Computational Prediction of Protein Subcellular Locations in Eukaryotes: an Experience Report

John Meinken 1, Xiang Jia Min 2

1. Department of Computer Science and Information Systems, Youngstown State University, Youngstown, OH 44555, USA
2. Center for Applied Chemical Biology, Department of Biological Sciences, Youngstown State University, Youngstown, OH 44555, USA

Corresponding Author email: xmin@ysu.edu; Author

Abstract Computational prediction of protein subcellular locations in eukaryotes facilitates experimental design and proteome analysis. We provide a short review on recent development of computational tools and our experience in evaluating some of these tools. Classical secretomes can be relatively accurately predicted using computational tools to predict existence of a secretory signal peptide and to remove transmembrane proteins and endomembrane reticulum (ER) proteins. The protocols of differentially combining SignalP, Phobius, WoLFPSORT, and TargetP for identifying a secretory signal peptide in different kingdom of eukaryotes, with TMHMM for removing transmembrane proteins and PS-Scan for removing ER proteins significantly improve the secretome prediction accuracies. Our evaluation showed that current computational tools for predicting other subcellular locations, including mitochondrial or chloroplast localization, still need to be improved.

Keywords Eukaryotes; Protein subcellular location; Secretome; Computational prediction

Introduction

Eukaryotic cells have a complex endomembrane system, in addition to independent organelle structures such as mitochondria and chloroplasts. These subcellular structures include nucleus, endoplasmic reticulum (ER), Golgi apparatus, lysosome, peroxisome, vacuole, cytoskeleton, cytosol, mitochondrion, chloroplast, and plasma membrane. For a subcellular membrane enclosed structure, these structures consist of membrane and internal space such as ER lumen. Outside the plasma membrane, the cell wall and extracellular matrix and space are also important sites for cellular activities.

Eukaryotic cells synthesize thousands of different proteins. For example, Saccharomyces cerevisiae, commonly known as baker’s yeast, with a relative small genome size of 12 Mb, encodes approximately 5000–6000 different proteins. The proteins encoded by a nuclear genome are synthesized on ribosomes in cytosol or ribosomes attached to rough ER. However, these proteins need to be translocated to one or more than one specific subcellular location(s) in order to play their biological roles, a process called protein targeting or sorting. Experimental approaches for identifying protein subcellular locations are widely exploited including isolation of organelles, green fluorescence tagging proteins, etc (Heazlewood et al., 2005). Some signal targeting peptides determining protein subcellular locations have been identified experimentally (Blobel and Dobberstein, 1975). Thus protein targeting is believed to be determined by the protein’s physical and chemical properties of targeting domains that could be identified from its amino acid sequence. A number of computational tools have been recently developed to predict the subcellular locations of eukaryotic proteins. Nakai and Horton (2007) comprehensively reviewed computation methods and tools for subcellular location.
prediction. We provide a short review about more recent progress in this area and discuss the challenges remaining for future development based on our research experiences.

**Secretory Signal Peptide and Secretome Prediction**

The term secretome is used to refer to a whole set of proteins that are secreted outside a cell—including cell wall, extracellular matrix and extracellular space—in an organism. Recently many efforts have been made to identify secretomes as these proteins have both potential applications in environmental industry and biomedicine (Lum and Min, 2011; Makridakis and Vlahou, 2010). For example, fungal secretomes often contain secreted extracellular enzymes to break down biopolymers that have potential applications in biofuel production (Lum and Min, 2011), and the human secretome plays important biological roles, such as insulin, and provides useful information for the discovery of novel biomarkers such as for cancer diagnosis (Makridakis and Vlahou, 2010).

A secretome consists of two types of proteins: classical secreted proteins and nonclassical secreted proteins. A typical classical secreted protein contains a secretory signal peptide located on its N-terminus and does not contain other targeting signals in its sequence (Emanuelsson et al., 2007). The secretory signal peptide directs the ribosome to the rough ER to complete the synthesis of a signal peptide containing protein. The secretory signal peptide, typically 15–30 amino acids long, is cleaved off during translocation across the membrane (von Heijne, 1990). One basic concept that should be made clear is that not all secretory signal peptide containing proteins are secreted. A number of papers reported secretomes solely based on existence of secretory signal peptide prediction, resulting in an overestimation of the number of secreted proteins in a proteome. Secreted proteins make up only a fraction of the proteins that enter the ER secretory pathway, as proteins that contain a signal peptide and enter the ER also include residents of the rough ER, smooth ER, Golgi complex, lysosomes, endosomes, and plasma membrane.

Nowadays the commonly used tools for secretory signal peptide prediction include SignalP 3.0 (Bendtsen et al., 2004b), SignalP 4.0 (Petersen et al., 2011), Phobius (Käll et al., 2004 and 2007), TargetP (Emanuelsson et al., 2000), and PrediSi (http://www.predi.si.de/) (Hiller et al., 2004). In addition, WoLFPSORT and MultiLoc2 can also be used for secreted protein prediction (Horton et al., 2007; Blum et al., 2009). The accuracy of SignalP 4.0 was improved over SignalP 3.0 with a higher specificity by intergrading transmembrane prediction (Petersen et al., 2011). However, SignalP 3.0 is more accurate than SignalP 4.0 in prediction of the cleavage site of the signal peptide. The default length of N-terminal peptide is 70 residues in both SignalP 3.0/4.0 and PrediSi and thus proteins sequences having a long signal peptide (>70 amino acids) cannot be predicted when using the default truncation parameter. Phobius was also a relatively accurate signal peptide predictor that combined transmembrane topology and signal peptide prediction.

Overall the accuracy for signal peptide prediction of the tools mentioned above was acceptable for general use. However, our recent evaluation of these tools showed that the accuracy for classical secretome prediction could be significantly improved by combining multiple tools, mainly due to an increase in prediction specificity (Min, 2010). In addition, adding TMHMM for removing transmembrane proteins and PS-Scan (a standalone version downloaded from Scan-Prosite) for removing ER resident proteins (Prosite: PS00014, Endoplasmic reticulum targeting sequence) significantly improved the accuracy for secretome prediction (Min, 2010). Our evaluation also showed different tools have different strength in processing protein data generated from different kingdoms of eukaryotic organisms. We proposed the following protocols for secretome prediction in different kingdoms of eukaryote: SignalP/WoLFPSORT/Phobius in fungi, Phobius/ WoLFPSORT/TargetP in animals, SignalP/Phobius/ TargetP in plants, and SignalP/Phobius/TargetP/ WoLFPSORT in protists. The specificity for signal peptide prediction is significantly increased when two or more tools are used. In addition, TMHMM and
PS-Scan should be used for all eukaryotic secretome predictions (Min, 2010).

There is only one tool, SecretomeP, available for nonclassical secretome prediction (http://www.cbs.dtu.dk/services/SecretomeP/) (Bendtsen et al., 2004a) of mammalian and bacterial organisms. As about 50% of secreted proteins in plants were estimated to be non-classical, i.e. leaderless secreted proteins (LSPs) (Agrawal et al., 2010), certainly a plant specific trained tool or method is needed for prediction of plant specific non-classical secretomes.

**Predictors for Multiple Subcellular Locations**

TargetP was implemented to predict subcellular locations of eukaryotic proteins by discriminating between chloroplast transit peptide (cTP, in plants), mitochondrial targeting peptide (mTP) and secretory pathway signal peptide (Emanuelsson et al., 2007). Except for fungal protein data sets, combining TargetP with SignalP, TMHMM and PS-Scan increased secretome prediction accuracy in other eukaryotic protein data sets (Min, 2010). Other widely used tools for predicting multiple subcellular locations are WoLFPSORT and MultiLoc2. WoLFPSORT predicts 12 subcellular locations including chloroplast, cytosol, cytoskeleton, ER, extracellular, Golgi apparatus, lysosome, mitochondria, nuclear, peroxisome, plasma membrane, and vacuolar membrane (Horton et al., 2007). MultiLoc2 predicts 9 subcellular locations for animals and fungi and 10 subcellular locations for plants (Blum et al., 2009). Chou and Shen (2008) developed a package of web servers, called Cell-PLoc, which includes 6 different servers for predicting up to 22 subcellular locations of proteins in various organisms including viruses, bacteria, plants, humans, or general eukaryotes (http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/). However, the servers in the Cell-PLoc package can only process a single sequence per submission, and no stand-alone tools are available, which prevented us for further evaluating the accuracies of these tools.

While developing the plant secretome knowledge-base (PlantSecKB), which is now publicly available (http://proteomics.ysu.edu/secretomes/plant.php), we compared the prediction accuracies of TargetP, WoLFPSORT, and MultiLoc2 using a set of plant proteins retrieved from the UniProtKB Swiss-Prot data set. Proteins having multiple subcellular locations or labeled as “fragment”, or having a term of “by similarity” or “probable” or “predicted” in subcellular location annotation were removed. A total of 6908 proteins having annotated subcellular locations were selected. The results are shown in Table 1. If we ignore the subcellular location predictions with less than 100 positive entries, our evaluation showed prediction of secreted proteins was relatively more accurate than prediction of other subcellular locations by all three tools. TargetP was significantly more accurate than the other two tools in predicting secreted proteins. The Mathews’ correlation coefficient (MCC) (Matthews, 1975) values for prediction of all other subcellular locations by all three tools were lower than 50%. Thus, an improvement in prediction accuracies for these subcellular locations of plant proteins is really needed. Overall prediction accuracies between WoLFPSORT and MultiLoc2 using its sequence-based prediction method did not show significant differences. MultiLoc2 incorporated phylogenetic profiles and Gene Ontology terms and was reported to perform considerably better than other methods for animal and plant proteins (Blum et al., 2009). However, its accuracy cannot be fairly tested as our data all had Gene Ontology annotation. In addition, we also found that MultiLoc2 was about 500 times slower in data processing than WoLFPSORT, which prevented us from using MultiLoc2 in data processing for our database development.

**Other Computational Tools**

Table 2 lists a collection of subcellular localization prediction tools and their related publications. The weblinks for all these tools can be found at our webservice (http://proteomics.ysu.edu/tools/subcell.html). This is not an exhaustive list, but focuses on the tools discussed in this paper as well as more recent tools published since 2008. Our knowledgebases currently collect predictions from SignalP 3.0, SignalP 4.0, TMHMM, Phobius, TargetP, WoLFPSORT, PS-Scan and FragAnchor as discussed above.
Table 2 A collection of published protein subcellular localization prediction tools

<table>
<thead>
<tr>
<th>Tool name</th>
<th>Locations or protein features predicted</th>
<th>Organism categories</th>
<th>Publication citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TargetP</td>
<td>Extracellular, mitochondrial, chloroplast</td>
<td>Non-plant, plant</td>
<td>Emanuelsson et al., 2000</td>
</tr>
<tr>
<td>TMHMM</td>
<td>Transmembrane helices</td>
<td>Any</td>
<td>Krogh et al., 2001</td>
</tr>
<tr>
<td>ScanProsite (PS-Scan)</td>
<td>ER retention signal</td>
<td>Any</td>
<td>de Castro et al., 2006</td>
</tr>
<tr>
<td>SecretomeP 2.0</td>
<td>Non-classical i.e. not signal peptide triggered protein secretion</td>
<td>Gram-positive bacteria, gram-negative bacteria, mammal</td>
<td>Bendtsen et al., 2004a</td>
</tr>
<tr>
<td>Phobias</td>
<td>Signal peptide and transmembrane topology</td>
<td>Any</td>
<td>Käll et al., 2007</td>
</tr>
<tr>
<td>WoLF PSORT</td>
<td>&gt; 5 locations</td>
<td>Fungi, plant, animal</td>
<td>Horton et al., 2007</td>
</tr>
<tr>
<td>PRED-LIPO</td>
<td>Lipoprotein signal peptides</td>
<td>Gram-positive bacteria</td>
<td>Bagos et al., 2008</td>
</tr>
<tr>
<td>ProLoc-GO</td>
<td>&gt; 5 locations</td>
<td>Human and eukaryotes</td>
<td>Huang et al., 2008</td>
</tr>
<tr>
<td>KnowPredsite</td>
<td>&gt; 5 locations</td>
<td>Prokaryotes and eukaryotes</td>
<td>Lin et al., 2009</td>
</tr>
<tr>
<td>MultiLoc2</td>
<td>&gt; 5 locations</td>
<td>Animal, fungal, plant</td>
<td>Blum et al., 2009</td>
</tr>
<tr>
<td>PRED-SIGNAL</td>
<td>Signal peptides</td>
<td>Archaea</td>
<td>Bagos et al., 2009</td>
</tr>
<tr>
<td>RSLPred</td>
<td>Chloroplast, cytoplasm, mitochondria, nucleus</td>
<td>Oryza sativa only</td>
<td>Kaundal and Raghava, 2009</td>
</tr>
<tr>
<td>SherLoc2</td>
<td>&gt; 5 locations</td>
<td>Animal, fungi, plant</td>
<td>Briesemeister et al., 2009</td>
</tr>
<tr>
<td>Cell-PLoc 2.0</td>
<td>&gt; 5 locations</td>
<td>Eukaryotic, human, plant, virus, gram-positive bacteria, gram-negative bacteria</td>
<td>Chou and Shen, 2010</td>
</tr>
<tr>
<td>CoBaltDB</td>
<td>&gt; 5 locations</td>
<td>Prokaryotes</td>
<td>Goudenège et al., 2010</td>
</tr>
<tr>
<td>PSORTb</td>
<td>&gt; 5 locations</td>
<td>Gram-positive and gram-negative bacteria</td>
<td>Yu et al., 2010</td>
</tr>
<tr>
<td>SCLPred</td>
<td>Cytoplasm, mitochondrion, nucleus, secretory, chloroplast</td>
<td>Animals, plants, fungi</td>
<td>Mooney et al., 2011</td>
</tr>
<tr>
<td>SignalP 4.0</td>
<td>Signal peptide</td>
<td>Eukaryotic, gram-positive bacteria, gram-negative bacteria</td>
<td>Petersen et al., 2011</td>
</tr>
<tr>
<td>SlocX</td>
<td>&gt; 5 locations</td>
<td>Aribdopsis thaliana only</td>
<td>Ryngajillo et al., 2011</td>
</tr>
</tbody>
</table>

Some tools make predictions for only a single subcellular location or identify the presence of a single protein feature (such as a signal peptide). Then there are more comprehensive tools that can make predictions for many locations, and may employ a combination of multiple computational methods as well. The trend in recent years seems to be toward more comprehensive tools. Of the tools we collected that were published since 2008, twelve out of fifteen contain predictions for four or more subcellular localizations.

With the emergence of so many tools that can already predict a variety of subcellular locations, one might ask if our approach of combining analysis results from multiple tools into a database is still relevant. We believe our work can make several valuable contributions in this area. Firstly, a combination of data from multiple predictions often produces more accurate results than the individual predictions. This principle has been demonstrated in our specific work with secretomes (Min, 2010) and is also a widely recognized statistical concept. Also, a database can be used in ways that a prediction tool cannot. For most of the prediction tools, analysis is performed at the time of request. The user must know which protein(s) they are interested in before they can get analysis results. With our database, the user can work in the other direction as well. They can start with a subcellular location and species they are interested in and get a list of proteins that meet those criteria.
In addition, the development of so many tools that can perform the same task creates a dilemma for researchers, who must choose which tool(s) they will use. There is a need for testing that compares different tools and identifies their relative strengths and weaknesses. Perhaps some tools perform better for plants while others perform better for bacteria. Some tools may have better specificity for a certain subcellular location while others may have better sensitivity. Our knowledgebases can serve as a rich dataset for performing such comparisons. In this work, we compare the prediction accuracies for plant proteins using TargetP, WoLFPSORT and MultiLoc2. Much more work is needed to continue these types of comparative studies for improving the prediction accuracy of proteome-wide protein subcellular location in the future.

Acknowledgments
The work was supported by the Ohio Plant Biotechnology Consortium and Youngstown State University (YSU) Research Council to XJM.

References
http://dx.doi.org/10.1104/pp.105.065532
PMid:16219920 PMCid:PMC1255979


http://dx.doi.org/10.1186/1471-2105-9-80
PMid:18241343 PMCid:PMC2262056

http://dx.doi.org/10.1016/j.jmb.2004.03.016
PMid:15111065


Kaundal R., and Raghava G.P.S., 2009, RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information, Proteomics, 9(9): 2324-2342
http://dx.doi.org/10.1002/pmic.200700597
PMid:19402042

http://dx.doi.org/10.1006/jmbi.2000.4315
PMid:11152613

http://dx.doi.org/10.1186/1471-2105-10-S15-S8

http://dx.doi.org/10.1093/database/bar001

http://dx.doi.org/10.1016/j.jprot.2010.07.001
PMid:20637910

Matthews B.W., 1975, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta, 405: 442-451
http://dx.doi.org/10.1016/0005-2795(75)90109-9


http://dx.doi.org/10.1093/bioinformatics/btr494
PMid:21873639

http://dx.doi.org/10.1007/978-1-59745-466-7_29
PMid:17951705

http://dx.doi.org/10.1038/nmeth.1701
PMid:21959131


http://dx.doi.org/10.1007/BF01868635
PMid:2197415

http://dx.doi.org/10.1093/bioinformatics/btq249
PMid:20472543 PMCid:PMC2887053
<table>
<thead>
<tr>
<th>Subcellular Location</th>
<th>Total positives</th>
<th>Total negatives</th>
<th>TargetP Sn (%)</th>
<th>TargetP Sp (%)</th>
<th>TargetP MCC (%)</th>
<th>WoLF PSORT Sn (%)</th>
<th>WoLF PSORT Sp (%)</th>
<th>WoLF PSORT MCC (%)</th>
<th>MultiLoc2 Sn (%)</th>
<th>MultiLoc2 Sp (%)</th>
<th>MultiLoc2 MCC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secreted</td>
<td>263</td>
<td>6645</td>
<td>76.8</td>
<td>98.7</td>
<td>72.3</td>
<td>43.7</td>
<td>99.6</td>
<td>58.0</td>
<td>51.3</td>
<td>98.6</td>
<td>53.7</td>
</tr>
<tr>
<td>Mitochondrial</td>
<td>402</td>
<td>6506</td>
<td>61.4</td>
<td>77.5</td>
<td>21.1</td>
<td>33.3</td>
<td>96.2</td>
<td>30.4</td>
<td>60.9</td>
<td>84.2</td>
<td>27.3</td>
</tr>
<tr>
<td>Chloroplast</td>
<td>4918</td>
<td>1990</td>
<td>28.2</td>
<td>90.7</td>
<td>20.4</td>
<td>28.2</td>
<td>83.9</td>
<td>12.7</td>
<td>19.2</td>
<td>98.6</td>
<td>23.2</td>
</tr>
<tr>
<td>ER</td>
<td>87</td>
<td>6821</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9.2</td>
<td>99.9</td>
<td>24.5</td>
<td>29.9</td>
<td>98.4</td>
<td>23.0</td>
</tr>
<tr>
<td>Cytosol</td>
<td>23</td>
<td>6885</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>52.2</td>
<td>65.6</td>
<td>2.2</td>
<td>95.7</td>
<td>51.5</td>
<td>5.4</td>
</tr>
<tr>
<td>Golgi Apparatus</td>
<td>54</td>
<td>6854</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0</td>
<td>99.9</td>
<td>-0.2</td>
<td>24.1</td>
<td>98.5</td>
<td>15.5</td>
</tr>
<tr>
<td>Peroxisome</td>
<td>52</td>
<td>6856</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>15.4</td>
<td>99.3</td>
<td>14.3</td>
<td>42.3</td>
<td>98.1</td>
<td>23.9</td>
</tr>
<tr>
<td>Nucleus</td>
<td>788</td>
<td>6120</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>80.3</td>
<td>78.1</td>
<td>41.1</td>
<td>26.8</td>
<td>93.3</td>
<td>22.4</td>
</tr>
<tr>
<td>Plasma Membrane</td>
<td>14</td>
<td>6894</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>21.4</td>
<td>97.9</td>
<td>6.1</td>
<td>7.1</td>
<td>99.4</td>
<td>3.9</td>
</tr>
<tr>
<td>Vacuole</td>
<td>121</td>
<td>6787</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>14.0</td>
<td>99.5</td>
<td>21.2</td>
<td>5.0</td>
<td>99.8</td>
<td>11.6</td>
</tr>
</tbody>
</table>

Note: Sn: Sensitivity; Sp: Specificity; MCC: Mathews' correlation coefficient