Mimicking Cellular Sorting Improves Prediction of Subcellular Localization

Rajesh Nair, Burkhard Rost

CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA
Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St. Nicholas Avenue, New York, NY 10032, USA
NorthEast Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA
Department of Physics, Columbia University, 538 West 120th Street, New York, NY 10027, USA


Predicting the native subcellular compartment of a protein is an important step toward elucidating its function. Here we introduce LOCtree, a hierarchical system combining support vector machines (SVMs) and other prediction methods. LOCtree predicts the subcellular compartment of a protein by mimicking the mechanism of cellular sorting and exploiting a variety of sequence and predicted structural features in its input. Currently LOCtree does not predict localization for membrane proteins, since the compositional properties of membrane proteins significantly differ from those of non-membrane proteins. While any information about function can be used by the system, we present estimates of performance that are valid when only the amino acid sequence of a protein is known. When evaluated on a non-redundant test set, LOCtree achieved sustained levels of 74% accuracy for non-plant eukaryotes, 70% for plants, and 84% for prokaryotes. We rigorously benchmarked LOCtree in comparison to the best alternative methods for localization prediction. LOCtree outperformed all other methods in nearly all benchmarks. Localization assignments using LOCtree agreed quite well with data from recent large-scale experiments. Our preliminary analysis of a few entirely sequenced organisms, namely human (Homo sapiens), yeast (Saccharomyces cerevisiae), and weed (Arabidopsis thaliana) suggested that over 35% of all non-membrane proteins are nuclear, about 20% are retained in the cytosol, and that every fifth protein in the weed resides in the chloroplast.

Keywords: protein subcellular localization prediction; support vector machines; hierarchical ontology; sequence alignment; database search

Abbreviations used: 1D structure, one-dimensional structure (e.g. sequence or string of secondary structure, or solvent accessibility); ER, endoplasmic reticulum; GFP, green
Assignment and prediction of subcellular localization indispensable
Most reliable predictions cover less than 50% of all proteins
De novo predictions of localization restricted by limited “biophysical reality”
Results
Data sets and cross-validation results
More data with noise better than less data with less noise
Very accurate distinction between secretory pathway proteins and all others
Overall accuracy of 74% for non-plants
Accurate distinction of three prokaryotic classes
Comparison with other methods using additional test sets
Other methods tested on new data set
LOCtree over 20 percentage points more accurate than other general servers
Performance better than existing methods even for incorrect sequences
About 80% agreement between predictions and large-scale experiments in yeast
Chloroplasts: experimental data supported by predictions
Application to representative proteomes
Discussion
Tree-based system provided additional advantages to boosting performance
Over 20 percentage points improvement over existing generalized methods
Many estimates for performance had a rather short-life span
Assignment and prediction of subcellular localization indispensable

The sequencing of the genomes, i.e. all DNA sequences, of over 260 organisms (February 2005), including the human genome\textsuperscript{1} and \textsuperscript{2} has been completed. For over 200 of the entirely sequenced organisms, the protein sequences are publicly available; 105 have been analyzed in the PEP database\textsuperscript{4}, and contribute about 413,000 protein sequences, i.e. about one-fourth of all currently known protein sequences.\textsuperscript{3, 4, 5} With this explosion of genome sequences, the major challenge in modern biology is to follow suit in advancing the knowledge of the expression, regulation, and function of the entire set of proteins encoded by an organism, i.e. its proteome. This information will be invaluable for understanding how complex biological processes occur at a molecular level, how they differ in various cell types, and how they are altered in disease states. Proteins must be localized in the same subcellular compartment to cooperate towards a common function. Therefore, experimentally unraveling the native compartment of a protein constitutes one step on the long way to determining its role. Using experimental high-throughput methods for epitope and green fluorescent protein (GFP) tagging, two groups have recently reported localization data for most proteins in \textit{Saccharomyces cerevisiae} (baker's yeast).\textsuperscript{6, 7} So far, the majority of large-scale experimental efforts to predict localization have been restricted to yeast, or to particular compartments, such as a recent analysis of chloroplast proteins in \textit{Arabidopsis thaliana} (weed).\textsuperscript{8} As of now, these large-scale experiments cannot be repeated for mammalian or other higher eukaryotic proteomes. One major obstacle is that large scale production of a collection of cell lines each with a
defined gene chromosomally tagged at the 3′ end is not yet possible. In contrast, computational tools can provide fast and accurate localization predictions for any organism. Attempts to predict subcellular localization have increasingly become one of the central problems in bioinformatics/computational biology.

Most reliable predictions cover less than 50% of all proteins

A number of methods predict localization by identifying short sequence motifs, such as signal peptides or nuclear localization signals (NLS) that are responsible for protein targeting. Most proteins destined for the secretory pathway, the mitochondria and the chloroplast contain N-terminal peptides that are recognized by the translocation machinery. The term “signal peptide” is used to describe the peptides in secreted proteins that are cleaved in the endoplasmic reticulum (ER) by signal peptidases; the peptides responsible for targeting proteins to the mitochondrial and chloroplast are referred to as “transit peptides”. Signal and transit peptides can be recognized by generic prediction methods, that by the detection of these peptides also predict subcellular localization. Many proteins destined for the nucleus contain NLS motifs that may occur anywhere in the sequence. Recently, we have collected a data set of experimental and potential NLS motifs as an aid to predicting nuclear localization. However, the vast majority of nuclear proteins have no known motif. For mitochondria and chloroplast, a number of alternative targeting pathways have also been discovered recently. Additionally, proteins such as fibroblast growth factors are targeted to the extra-cellular space via non-classical secretory pathways, i.e. they do not possess N-terminal signal peptides. Furthermore, a particular problem for methods detecting N-terminal signals is that start codons are predicted with less than 70% accuracy by genome projects. Overall, known and predicted sequence motifs enable annotating about 30% of the proteins in six entirely sequenced eukaryotic proteomes. Other methods that can be reliably used to annotate localization but are not always applicable are annotation transfer from sequence homologues and text analysis. A particular variant of homology-based predictions is the domain projection method that is based on similarity to SMART domains of known subcellular localization. Despite recent high-throughput experiments, the most reliable prediction methods together cover less than 50% of entirely sequenced multi-cellular proteomes.

De novo predictions of localization restricted by limited “biophysical reality”

In the near future, the only hope of assigning compartments to the remaining half of all multi-cellular proteins is using methods that predict localization from features other than known import/export motifs. The most promising approach is to exploit the correlation between localization and amino acid composition of a protein, which is mostly due
to the altering of the protein surface in response to changing environmental conditions.\textsuperscript{50}

Methods using only amino acid composition to predict localization are \textit{de novo} methods; they predict localization without any explicit experimental knowledge of the protein under investigation. In particular, they are as accurate if any information about function is available for a target as when the target is merely a “hypothetical protein”. Higher-order residue correlations (between residues \(i\) and \(i+n\), for \(n=2,3,4\)) have been accounted for by using pseudo-amino acid composition.\textsuperscript{51, 52} and \textsuperscript{53} Recently, we showed that incorporating structural and evolutionary information significantly improves prediction accuracy.\textsuperscript{54}

With the availability of many completely sequenced genomes, phylogenetic profiles have been employed to identify subcellular localization.\textsuperscript{55} So far, this approach has been much less accurate than methods based solely on composition. Drawid & Gerstein have proposed a Bayesian system, based on a diverse range of 30 different features, to predict the localization of yeast proteins.\textsuperscript{56} The problem with all these methods is that they are based on sequence features that may reveal localization but are not the reason why proteins are transported, such as signal and transit peptides and nuclear localization signals. Furthermore, all general methods, with the exception of PSORT,\textsuperscript{57} and \textsuperscript{58} implicitly assume that all localizations are equidistant, i.e. if a method predicts a nuclear protein to be cytoplasmic it makes the same mistake as another method which predicts this protein to be extra-cellular. In reality, however, some compartments are more similar to each other than others, e.g. ER is closer to extra-cellular than to nuclear due to the proximity in the space of the biological sorting machinery.

Here, we describe a novel system of support vector machines (SVMs) that predict subcellular localization by incorporating a hierarchical ontology of localization classes modeled onto biological processing pathways. By construction, the system penalizes confusions of classes along the same pathway (e.g. ER instead of extra-cellular) less than confusions between classes from different pathways (e.g. ER instead of nuclear). The biological similarities are incorporated from the description of cellular components in the gene ontology (GO).\textsuperscript{59} and \textsuperscript{60} We simplified and tailored the GO definitions to the problem of protein sorting. For example, in GO both the ER and the Golgi apparatus are subcategories of the cytoplasm. However, proteins destined for the extra-cellular space, the ER, the Golgi, endosomes and lysosomes are targeted \textit{via} the same secretory pathway. By this criterion, proteins from the secretory pathway are more similar to each other than they are to other intra-cellular proteins.\textsuperscript{30} Hence, in our classification scheme these compartments are grouped together and are designated as belonging to the secretory pathway. Technically, we incorporated the ontology through a decision tree with SVMs as the nodes (\textbf{Figure 1}). We favored SVMs over neural networks due their improved performance (data not shown). The final system, LOCtree, was extremely successful at learning evolutionary similarities among subcellular localization classes and was significantly more accurate than other traditional networks at predicting subcellular localization.
Figure 1. Hierarchical architecture of LOCtree. LOCtree uses specialized architecture to predict subcellular localization of proteins from different organisms: (a) architecture for eukaryotic non-plant proteins; (b) architecture for plant proteins; and (c) the architecture for prokaryotic proteins. At each branch point a support vector machine (SVM) is used to accomplish a binary classification (either protein belongs to localization class L or does not belong to L). The hierarchical architecture has been designed to mimic the biological protein sorting mechanism as closely as possible. The branches of the tree represent intermediate stages in the sorting machinery while the nodes represent the decision points in the sorting machinery. The different levels of SVMs in the hierarchical tree are labeled Level 0, Level 1, etc. For example, Level 0 represents the top node SVM which discriminates between secretory pathway proteins and other intra-cellular proteins ((a) and (b)) or proteins which remain in the cytoplasm from the rest (c). The intermediate node SVMs in the next level are represented as Level 1, and are responsible for separating extra-cellular proteins from proteins sorted to the organelles and nuclear proteins from cytoplasmic proteins ((a) and (b)). For the prokaryotic architecture (c), Level 1 is the terminal level for Gram-negative bacteria and separates extra-cellular proteins from periplasmic proteins. In addition, Level 1 also contains the cytoplasmic leaf which is propagated without branching from Level 0. For Gram-positive bacteria, Level 0 is the terminal level and separates cytoplasmic proteins from extra-cellular proteins (non-cytoplasmic branch). The leaves of the tree, represented by rectangular boxes represent the final localization classes for which prediction is made. If a leaf has a depth smaller than the overall depth of the tree it is propagated without branching for the remainder of
the tree. Level 2 is the terminal level for the eukaryotic non-plant architecture (a) and is responsible for sorting proteins into one of five subcellular classes (mitochondria and cytosol plus the three leaves from Level 1), while Level 3 is the terminal level for the plant architecture (c) and separates proteins into one of six classes (mitochondria and chloroplast plus the four leaves from Level 2). The prediction accuracy of the parent nodes is higher than the child nodes leading to a significantly improved prediction accuracy for the intermediate localization states. Abbreviations: EXT, extra-cellular; NUC, nucleus; CYT, cytosol; MIT, mitochondria; CHLORO, chloroplast; RIP, periplasm; and ORG, organelle. Organelles are the endoplasmic reticulum, Golgi apparatus, peroxysomes, lysosomes, and vacuolar compartments.

We have applied LOCtree to analyze the subcellular localization of complete genomes of a number of eukaryotic and prokaryotic organisms. The LOCtree subcellular localization prediction server and the results of our localization annotations for entire proteomes are available.†

Results

Data sets and cross-validation results

More data with noise better than less data with less noise

Proteins with experimentally annotated subcellular localization were extracted from SWISS-PROT (Methods and Materials). For this study, we excluded membrane proteins, i.e. all our results are valid for a subset of 75–80% of all proteins. In total, we had 8980 eukaryotic and 13,186 prokaryotic non-membrane proteins with explicit experimental annotations (Methods and Materials). Training and test sets were constructed by partitioning the data such that test sequences had less than 25% sequence identity to any sequence in the training set over an alignment length of 250 residues (HVAL=5; equation (1) Methods and Materials). To avoid overestimating performance, we reduced redundancy such that our final sequence-unique test set contained 1505 non-redundant eukaryotic non-plant sequences, 304 plant and 672 prokaryotic test sequences. All results of the methods described here were based on sixfold cross-validation experiments, i.e. we cycled six times through the entire sequence-unique data such that each protein was used for testing once. To increase the size of the training set, we included homology-based (LOChom) and keyword-based (LOCkey) predictions in the training data. While adding these noisy predictions, we ascertained that no homologues to any of the test proteins were included. This procedure almost quadrupled the training data; it increased prediction accuracy by nearly seven percentage points (data not shown). The major improvement resulted from the addition of keyword-based annotations using LOCkey. Plants were treated separately since their compositional features differed significantly from non-plant eukaryotes (not shown). Using the SVM-light package, we found the radial basis function (RBF) kernel to perform better than linear and polynomial kernels. This result was obtained on a small subset of all proteins without cross-validation. In particular, we did not optimize this solution for the final test set.
Very accurate distinction between secretory pathway proteins and all others

To predict the localization of an unknown eukaryotic protein, LOCtree first determines if it is sorted using the secretory pathway. The SVM that makes this distinction achieved an overall prediction accuracy around 90% for both eukaryotic non-plant (Table 1; Figure 2(a)) and plant proteins (Table 2). Using the signal peptide prediction of SignalP23 as one input to the SVM improved accuracy by over one percentage point (data not shown). We also confirmed our previous observation that using overall composition in conjunction with N-terminal composition improved performance.24 Our methods also distinguished intra-cellular proteins very accurately from those entering the secretory pathway (>90% accuracy; Table 1 and Table 2).

Table 1.
LOCtree on non-redundant test set of eukaryotic non-plant proteins

<table>
<thead>
<tr>
<th>Hierarchy level</th>
<th>Class</th>
<th>Nprot</th>
<th>Acc</th>
<th>Cov</th>
<th>GAv</th>
<th>Q (StdDev)</th>
<th>MCC</th>
<th>MI</th>
<th>Nstates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>Secretory pathway</td>
<td>415</td>
<td>81</td>
<td>80</td>
<td>81</td>
<td>89 (2)</td>
<td>0.73</td>
<td>0.44</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Intra-cellular</td>
<td>1090</td>
<td>92</td>
<td>93</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>Extra-cellular</td>
<td>363</td>
<td>83</td>
<td>81</td>
<td>82</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organelles</td>
<td>52</td>
<td>51</td>
<td>52</td>
<td>52</td>
<td>78 (4)</td>
<td>0.55</td>
<td>0.40</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Nuclear</td>
<td>562</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cytoplasm</td>
<td>528</td>
<td>76</td>
<td>78</td>
<td>77</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Cytosol</td>
<td>330</td>
<td>63</td>
<td>66</td>
<td>64</td>
<td>74 (6)</td>
<td>0.55</td>
<td>0.39</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Mitochondria</td>
<td>198</td>
<td>70</td>
<td>67</td>
<td>68</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations used: hierarchy level and class are as illustrated in Figure 1; Nprot, number of proteins in sequence-unique test set with a given localization; Nstates, number of effective states predicted at given level (note that Level 1 contains four states while Level 2 contains five states, namely the Level 1 leaves (extra-cellular, organelles and nuclear) + cytosol + mitochondria). Performance measures: Acc, accuracy or specificity (equation (2)); Cov, coverage or selectivity (equation (3)); gAv, geometric average between Acc and Cov (equation (4)); Q, overall prediction accuracy for a given level in the hierarchy (equation (5); note depending on the level this is a two-state, four-state, or five-state value); MCC, Mathews correlation coefficient (equations Figs. (6) and (8)); MI, mutual information (equations Figs. (7) and (10)). Note 1: Q=74% at Level 2 is the overall accuracy for classification into one of five localization classes (extra-cellular, organelles, nuclear, cytosol or mitochondria).
Figure 2. Reliability of LOCtree. The curves show prediction accuracy of LOCtree for eukaryotic animal sequences. (a) Overall performance: the prediction accuracy decreases as we descent the hierarchical tree (Figure 1(a)). The Level 2 accuracy shown includes the accuracy of all Level 1 leaves like the extra-cellular, organelle and nuclear classes (Figure 1(a)), and represents the accuracy of classifying the protein into one of five subcellular classes. At 75% coverage the prediction accuracy is around 94% for Level 0, dropping to 84% for Level 1 and 77% for Level 2. The ability of the hierarchical system to predict intermediate localization states at a significantly higher accuracy is evident from the 17% difference in prediction accuracy between Level 0 and Level 2. Level 1 separates proteins into one of four subcellular classes and is over 7% more accurate than Level 2, which separates proteins into one of five classes. (b) Class-wise performance: LOCtree is best at discriminating secretory pathway proteins from all other proteins (91% accuracy at 50% coverage). Prediction of nuclear and extra-cellular proteins was only slightly less accurate (84% accuracy at 50% coverage) while performance was significantly worse for cytosolic proteins with only 64% correctly predicted. The standard deviation in the prediction accuracy for each of the localization classes was roughly 7%.

Table 2.

LOCtree on non-redundant test set of plant proteins
<table>
<thead>
<tr>
<th>Hierarchy level</th>
<th>Class</th>
<th>Nprot</th>
<th>Acc</th>
<th>Cov</th>
<th>gAv</th>
<th>Q (StdDev)</th>
<th>MCC</th>
<th>MI</th>
<th>Nstates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>Secretory pathway</td>
<td>42</td>
<td>77</td>
<td>79</td>
<td>78</td>
<td>94 (4)</td>
<td>0.74</td>
<td>0.49</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Intra-cellular</td>
<td>262</td>
<td>97</td>
<td>96</td>
<td>97</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>Extra-cellular</td>
<td>22</td>
<td>68</td>
<td>68</td>
<td>68</td>
<td>88 (5)</td>
<td>0.58</td>
<td>0.49</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Organelles</td>
<td>20</td>
<td>57</td>
<td>60</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nuclear</td>
<td>32</td>
<td>70</td>
<td>81</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cytoplasm</td>
<td>230</td>
<td>95</td>
<td>93</td>
<td>94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Cytosol</td>
<td>77</td>
<td>73</td>
<td>74</td>
<td>74</td>
<td>77 (7)</td>
<td>0.59</td>
<td>0.44</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Non-cytosol</td>
<td>153</td>
<td>84</td>
<td>80</td>
<td>82</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>Mitochondria</td>
<td>50</td>
<td>61</td>
<td>74</td>
<td>67</td>
<td>70 (3)</td>
<td>0.58</td>
<td>0.42</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Chloroplast</td>
<td>103</td>
<td>77</td>
<td>63</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations used as for Table 1.

**Overall accuracy of 74% for non-plants**

If a protein is predicted as belonging to the secretory pathway it is further sub-classified into extra-cellular or not (Figure 1(a)). The non extra-cellular proteins belong to either of the following organelles: endoplasmic reticulum (ER), Golgi apparatus, peroxisome, lysosome, or vacuole. Proteins native in one of these organelles were predicted at levels around 50% accuracy and 52% coverage (values were higher for plant proteins; Table 2); these values were much lower than the averages for all other classes. The sub-classification of intra-cellular proteins into nucleus and cytoplasm was less accurate; however, levels of accuracy and coverage were still above 76% (Table 1 and Table 2). Of the final localization classes (the leaves in Figure 1(a)), the extra-cellular class was predicted most accurately while the nuclear class was predicted only slightly less accurately (Figure 2(b)). Cytosolic proteins were, as expected, predicted with the lowest accuracy by the non-plant method (Table 1 and Figure 2(b)), while mitochondrial proteins were the least accurately predicted by the plant specialist (Table 2). Predictions for all classes of the non-plant system were extremely balanced between accuracy and coverage (Table 1). For reporting prediction accuracy, if a leaf (terminal node) has a smaller depth than the overall tree (the extra-cellular leaf has a depth of 1 while the overall tree depth is 2), the predictions for this leaf are propagated without branching for the depth of the tree (Figure 1(a)). Hence, the quoted Level 2 accuracy of Q_s=74% (Table 1) is the overall accuracy for classification into one of five localization classes (extra-cellular, organelles, nuclear, cytosol or mitochondrial). This was over seven percentage
points more accurate than another system that used the same input with the traditional pair-wise SVMs (data not shown). The unique feature of our method is that it predicts “intermediate” localizations such as intra-cellular and cytoplasm (Table 1 and Table 2). These intermediate localizations are predicted with a much higher accuracy as is evident from the progressive decrease in prediction accuracy as we descent the hierarchical tree (Figure 2(a)).

Accurate distinction of three prokaryotic classes

For prokaryotic proteins, LOCtree first determines if the protein is cytoplasmic or not. The SVM discriminating between these localizations reached an overall accuracy of 90% (Table 3). Prediction accuracy did not differ significantly between Gram-positive and Gram-negative bacteria. For Gram-negative bacteria, the non-cytoplasmic proteins are further classified into periplasmic and extra-cellular. The overall three class (cytoplasmic, periplasmic or extra-cellular) prediction accuracy for Gram-negative bacteria was $Q_3=83\%$; the two class (cytoplasmic or extra-cellular) accuracy for Gram-positive bacteria was $Q_2=90\%$. For Gram-negative bacteria, the distinction between periplasmic and extra-cellular proteins was at a much lower accuracy than cytoplasmic proteins.

Table 3.

LOCtree on non-redundant test set of prokaryotic proteins

<table>
<thead>
<tr>
<th>Hierarchy level</th>
<th>Class</th>
<th>Nprot</th>
<th>Acc</th>
<th>Cov</th>
<th>gAv</th>
<th>$Q$ (StdDev)</th>
<th>MCC</th>
<th>MI</th>
<th>Nstates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>Cytoplasm</td>
<td>426</td>
<td>89</td>
<td>97</td>
<td>93</td>
<td>90 (4)</td>
<td>0.79</td>
<td>0.52</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Non-cytoplasm</td>
<td>246</td>
<td>93</td>
<td>80</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>Periplasmic</td>
<td>125</td>
<td>86</td>
<td>62</td>
<td>73</td>
<td>83 (2)</td>
<td>0.55</td>
<td>0.45</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Extra-cellular</td>
<td>42</td>
<td>59</td>
<td>74</td>
<td>66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations used as for Table 1. Note 1: Level 1 is applicable to Gram-negative bacteria only. For Gram-positive bacteria, the system performs a two-state classification with non-cytoplasmic proteins being classified as extra-cellular. For Gram-negative bacteria, non-cytoplasmic proteins are further separated into periplasmic and extra-cellular proteins. Note 2: the Level 0 prediction accuracy did not differ significantly between Gram-positive and Gram-negative bacteria. The prediction accuracy reported above is the combined prediction accuracy for Gram-positive and Gram-negative bacteria. The overall two class prediction accuracy for Gram-positive bacteria was $Q_2=90\%$ while the three class prediction accuracy for Gram-negative bacteria was $Q_3=83\%$. 


Comparison with other methods using additional test sets

Other methods tested on new data set

We compared our method to the following publicly available methods: TargetP, SubLoc, NNPSL, and PSORT II. In contrast to all other methods, TargetP focuses exclusively on N-terminal sorting signals (secreted, chloroplast, mitochondria); it does not predict proteins targeted using other mechanisms or in any other compartment, such as in the nucleus or cytoplasm. Of the other servers that predict at least four classes, SubLoc is also based on SVMs; NNPSL is based on neural networks. SubLoc and NNPSL rely solely on amino acid composition while PSORT combines information from local sequence motifs and a neural network based method. All publicly available methods were tested on smaller data sets than LOCtree, and on data sets with little mutual overlap. We could run all servers on our non-redundant test set from the cross-validation experiments (Table 1, Table 2 and Table 3). However, most of the proteins in our data set had been used to develop those servers and we could not cross-validate any method other than our own. A benchmark with our data set would, therefore, have very limited value. For completeness, we reported the results of this test which as expected, over-estimated the public servers significantly (Supplementary Data, Table 1). The most meaningful comparison of prediction methods is based on a significantly sized, sequence-unique data set of proteins that have neither been used for the development of any of the methods tested, nor have significant sequence similarity to any of the methods tested.

Unfortunately, such sets are often difficult to get. If we ignored the most recent improvement of one of the components of TargetP, namely SignalP 3.0, we could find such a data set in proteins added between SWISS-PROT version 40 and 41 (Methods and Materials). While SignalP 3.0 has been developed after release 41, all the methods that we compared have been developed before release 41. Note that we deliberately restricted our development of LOCtree to proteins available in version 40 so that we could carry out this comparison. In many ways, our evaluation was also informative of the sustained performance of the methods tested, some of which had fallen prey to severe over-estimates of performance in their original publications. Note that PSORT, TargetP and the different versions of SignalP stood out in that their authors had correctly estimated the sustained performance all along.

LOCtree over 20 percentage points more accurate than other general servers

In the benchmark of proteins that had not been used for the development of any method, LOCtree outperformed all other servers (Table 4). TargetP was more accurate at predicting proteins targeted via the secretory pathway but its coverage was lower than that of LOCtree. The reason was that TargetP slightly under-predicted the secretory pathway (imbalance in gAv (equation (4)), i.e. the geometric average over Acc and Cov; Table 4). On our data set, we found the accuracy of SignalP 3.0 to be slightly lower than that of TargetP, since the difference was not significant, SignalP 3.0 was not shown separately in order to simplify. PSORT II was the most accurate server for predicting extra-cellular proteins; however, this was achieved at the cost of an extremely low level of coverage; in the geometric average between accuracy and coverage, PSORT II was
more than 30 percentage points lower than LOCtree. As shown in our cross-validation experiments (Table 1, Table 2 and Table 3), LOCtree was very balanced in its compromise between accuracy and coverage, i.e. between under and over-prediction, for all classes, and it was much more balanced than any other server. In terms of the overall four-state accuracy ($Q_4$; equation (5)), LOCtree scored 21 percentage points higher than its best competitor SubLoc (Table 4).

Table 4.

Comparison on identical sequence-unique set of new SWISS-PROT non-plant eukaryotic proteins

<table>
<thead>
<tr>
<th>Class</th>
<th>Method</th>
<th>LOCtree here</th>
<th>TargetP 23; 24; 33</th>
<th>SubLoc 67</th>
<th>PSORT 25; 57; 58; 97</th>
<th>NNPSL 40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secretory Pathway</td>
<td>Acc</td>
<td>87</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cov</td>
<td>90</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>gAv</td>
<td>88</td>
<td>82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ext</td>
<td>Acc</td>
<td>86</td>
<td>73</td>
<td>91</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cov</td>
<td>93</td>
<td>53</td>
<td>32</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gAv</td>
<td>89</td>
<td>62</td>
<td>54</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>Nuc</td>
<td>Acc</td>
<td>77</td>
<td>64</td>
<td>56</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cov</td>
<td>85</td>
<td>71</td>
<td>75</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gAv</td>
<td>81</td>
<td>67</td>
<td>65</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>Cyt</td>
<td>Acc</td>
<td>82</td>
<td>43</td>
<td>47</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cov</td>
<td>64</td>
<td>56</td>
<td>47</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gAv</td>
<td>72</td>
<td>49</td>
<td>47</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Mit</td>
<td>Acc</td>
<td>73</td>
<td>54</td>
<td>48</td>
<td>46</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Cov</td>
<td>78</td>
<td>75</td>
<td>59</td>
<td>59</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>gAv</td>
<td>75</td>
<td>64</td>
<td>53</td>
<td>52</td>
<td>45</td>
</tr>
<tr>
<td>Overall accuracy $Q_4$</td>
<td>78</td>
<td>57</td>
<td>51</td>
<td>52</td>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations used as for Table 1, with the following exceptions. Data set: all sequence-unique eukaryotic non-plant proteins added between release 41 and 40 of SWISS-PROT (Non-plant new unique in Table 4 of the Supplementary Data). Note that none of the proteins in this set had significant sequence similarity to any of the proteins that had annotations about localization in SWISS-PROT at the time of development of the
prediction methods for which results are shown. In this sense, our test set could also provide an independent and likely more accurate estimate for the sustained performance than some of the original publications for some of the methods. Localization: Ext, extracellular; Nuc, nuclear; Cyt, cytosolic; Mit, mitochondria; Chloro, chloroplast. Methods: Predictions from methods other than LOCtree, introduced here, were taken from their public Internet servers (Methods and Materials), except for PSORT II that was run locally; numbers in square brackets under methods refer to the original publication (References). Numbers in bold: in each row, the best method(s) is (are) marked in bold letters; methods are grouped according to significant differences (below), i.e. all values that are statistically indistinguishable are marked as one best group. Significant differences: For LOCtree, the standard deviation in the five-state accuracy was roughly six percentage points. The following estimates for standard deviations were published: TargetP, about one percentage point; NNPSL, about 2.5 percentage points; PSORT II, about 3.5 percentage points. Since no error estimates were published for SubLoc, we used 2.5 percentage points as the mean over the other three.

**Performance better than existing methods even for incorrect sequences**

LOCtree explicitly used information from the first 50 residues (N termini) and compositions from the entire protein. Both these values are likely to be wrong for many proteins taken from large-scale sequencing projects. We tried to estimate the effect of such mistakes through two different “models”: (1) we cleaved off 30 N-terminal residues for all proteins; and (2) we randomly picked positions to remove one-third of the sequence for each protein. These tests constituted worst-case scenarios in the sense that they all over-estimated sequencing errors substantially. We found that the overall prediction accuracy of LOCtree on the randomly cleaved fragments was 68% (Supplementary Data, Table 2), 10% less than what was obtained using the full protein sequence. For the N-term cleaved sequences, the accuracy further dropped to 55% due to the explicit dependence of LOCtree on N-terminal sequence information. This is still accurate enough to provide reliable first estimates of localization for genomic sequences.

**About 80% agreement between predictions and large-scale experiments in yeast**

Over the last years the large-scale experimental determination of subcellular localization for a substantial fraction of all yeast proteins has become increasingly accurate. Using high-throughput immuno-localization (IL) of epitope-tagged gene products, the Snyder group determined the localization for about 60%, while the O'Shea group exploited high-throughput GFP tagging to cover about 66% of all yeast proteins. Both studies did not distinguish between membrane and non-membrane proteins, and both did not capture secreted proteins. Many proteins were experimentally associated to more than one single compartment: 35% for Snyder et al. and 31% for O'Shea et al. We compared the LOCtree predictions of all proteins predicted to not contain membrane helices, and observed to be nuclear or mitochondrial in the two large-scale experiments. Proteins observed to be in the cytoplasm in the two large-scale studies were excluded from our analysis, since a large fraction (43% for Snyder et al. and 55% for O'Shea et al.) of cytosolic proteins were also observed in alternative compartments. Similarly, we also excluded all other proteins experimentally associated with more than one compartment. This filtering left
over 1000 proteins from the GFP data and about 200 proteins from the IL data. For both these data sets, about 80% of the predictions from LOCtree were identical with the experimental results (Table 5). This is comparable to the 80% agreement between the GFP data and traditional non-high-throughput results previously annotated in SGD.\textsuperscript{72} and \textsuperscript{73} The agreement between GFP and IL is about 75% for the subset of 146 proteins found to be nuclear or mitochondrial by GFP that were also found in the IL data set. The agreement between GFP annotations and yeast proteins annotated using homology to SWISS-PROT proteins was 79%. This is for five subcellular classes and using an HVAL>10 (equation (1)) for homology annotations. For the IL data the agreement with SWISS-PROT was 72% (note due to the small data set this number is a very inaccurate estimate).

Table 5.
Performance of LOCtree based on large scale yeast localization data

<table>
<thead>
<tr>
<th>Method</th>
<th>Nuclear</th>
<th></th>
<th>Mitochondrial</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs</td>
<td>Acc</td>
<td>Cov</td>
<td>Obs</td>
</tr>
<tr>
<td>GFP\textsuperscript{2}</td>
<td>586</td>
<td>82</td>
<td>68</td>
<td>418</td>
</tr>
<tr>
<td>IL\textsuperscript{6}</td>
<td>124</td>
<td>88</td>
<td>64</td>
<td>60</td>
</tr>
</tbody>
</table>

Abbreviations used: Methods: Experimental subcellular localization data for proteins in yeast were obtained from two methods: GFP, large scale localization using green fluorescent protein tagging;\textsuperscript{2} IL, large scale localization using high-throughput immunolocalization of epitope-tagged proteins.\textsuperscript{6} Data: Obs, number of non-membrane proteins for which localization was predicted in this compartment by the respective large-scale experiment. All proteins observed to be in multiple compartments by the large-scale methods were excluded from our analysis. LOCtree was used to predict localization of the remaining proteins. The prediction accuracy (Acc) and coverage (Cov) of LOCtree was calculated by assuming that the localization observed in the large-scale experiment represents the true localization of the protein. Note 1: cytoplasmic proteins were excluded from our analysis, since a large fraction (45–70%) of all proteins was observed to be in the cytoplasm in the two large-scale experiments. Nearly half of all cytoplasmic proteins were observed to be associated with more than one compartment and many are likely to be further sorted to other compartments.

**Chloroplasts: experimental data supported by predictions**

Using tandem mass spectroscopy (TMS), Kleffmann \textit{et al.}\textsuperscript{8} recently identified 690 proteins localized in the chloroplast of \textit{A. thaliana} (weed). Of these we predicted 190 to contain membrane helices. We compared LOCtree and TargetP\textsuperscript{33} for the remaining 500 proteins (Figure 3). The following results stood out: (1) less than half of these proteins were identified by all three methods; (2) a considerable fraction (29%) of the 500 proteins
was only identified by TMS; and (3) when comparing the chloroplast predictions for all weed proteins, we found that LOCtree and TargetP agreed in about 87% of their predictions. TargetP, however, predicts more chloroplast proteins than LOCtree. This could be due to the over-prediction of chloroplast proteins by TargetP which has been reported by an independent group.  

Figure 3. Benchmarking LOCtree using large-scale experimental data. Both LOCtree and TargetP were used to predict the localization of nearly 500 chloroplast proteins from *A. thaliana* which were identified using tandem mass spectroscopy (TMS) by Kleffmann *et al.* LOCtree and TargetP both predicted over 45% of these proteins to be localized in the chloroplast lending strong support to the large-scale experimental data using TMS. Over 70% of the proteins were predicted to be in the chloroplast by at least one server. TargetP showed a high degree of agreement with LOCtree, agreeing with over 87% of the predictions using LOCtree.

**Application to representative proteomes**

We used LOCtree to annotate the subcellular localization for all non-membrane proteins in the entire proteomes of *Homo sapiens* (human), *A. thaliana* (weed), and *S. cerevisiae* (yeast) (Figure 4). The results of our proteome annotations can be queried (downloaded) from the LOCtree website. We estimated that over 60% of all non-plant and over 50% of all plant proteins are nuclear or remain in the cytosol (Figure 4 and Table 3 of Supplementary Data). While over 75% of the non-membrane proteins in all genomes appeared intra-cellular, the fraction of secreted proteins varied substantially between 8% and 20%, with plants having fewer than 10% extra-cellular proteins and the number exceeding 20% in human. Nuclear proteins were overabundant in yeast. In general, the unicellular yeast was somewhere in between human and weed in its composition of compartments.
Figure 4. Composition of compartments in three representative proteomes. Note that 100% of the pie charts represents the number of proteins without transmembrane helices (predicted by PHDhtm\textsuperscript{98, 99, and 100} and taken from PEP\textsuperscript{3}). The final estimates were corrected in order to account for our compartment-specific estimates of accuracy and coverage (equation (11)). For all three proteomes the nucleus appeared to take the lion's share of all proteins, only the chloroplast came near this value for the plant representative weed (A. thaliana). Human (H. sapiens) has significantly more secreted proteins than do weed and yeast (S. cerevisiae). Yeast appeared to have the highest fraction of mitochondrial proteins. For the proteomes the percent fractional error for the estimates of the different compartments are given by: extra-cellular (±9%), organelle (±34%), nuclear (±10%), cytosol (±20%), mitochondria (±25%) and for chloroplast (±17%).

**Discussion**

**Tree-based system provided additional advantages to boosting performance**

Our results demonstrated how the prediction of subcellular localization can be substantially improved by mimicking the biological protein trafficking mechanism as closely as possible through, LOCtree, a hierarchical tree of SVMs (Figure 1 and Table 1, Table 2 and Table 3). PSORT\textsuperscript{25, 57, 58 and 76} is based on an implementation of a reasoning tree that is conceptually most similar to LOCtree. However, unlike LOCtree, PSORT is not based on an explicit “ontology” of subcellular localization. Instead, the nodes of the PSORT reasoning tree assign a probabilistic value to the presence/absence of a single
feature and have no intrinsic meaning. In contrast, the nodes of LOCtree separate proteins belonging to different cellular sorting pathways. In addition to being more accurate, our machine learning system provided two added benefits. The first was the prediction of “intermediate stages”. The prediction of intermediate stages, such as secretory pathway, was achieved at much higher levels of accuracy than the native compartments. This is not too surprising given that large-scale experiments using IL and GFP-tagging\(^6\) and \(^7\) suggest that 30% of all proteins are ambiguous, i.e. spend a considerable portion of their life-time in more than one native compartment. These experimental data might strike many biologists as “expected,” since the vast majority of proteins travel through the cell, e.g. most extra-cellular proteins “visit” at least three other compartments (ER, Golgi, vesicles) before they eventually are secreted. However, the very fact that our system reaches levels above 74% accuracy for the distinction of proteins into one of five compartments (extra-cellular, cytoplasmic, mitochondrial, nuclear, and organellar) suggested that most proteins have very strong preferences for one single compartment imprinted onto their sequences. The importance of post-translational modifications in altering these sequence signals as a means to increase the “fitness” for other compartments might explain the difference between these two extreme opposite perceptions of proteins as native to “one native compartment” and as “frequent travelers between compartments”\(^12\), \(^20\), and \(^72\). The second advantage of our hierarchical system might appear to be of more technical nature, namely that our modular architecture allows the addition of more fine-grained modules at later stages (e.g. the split of nuclear into nuclear-matrix and other\(^78\)). However, this seemingly technical detail actually once again was borne out of the advantage of mimicking the actual sorting system. We observed that as one descent's the hierarchical tree the prediction accuracy progressively decreases, since the classification task becomes increasingly complex and the SVM has to discriminate between increasingly similar proteins. One problem with our decision tree-like implementation was that a prediction mistake at a top node could not be corrected at nodes lower in the hierarchy. The appropriate choice of the evolutionary hierarchy was, therefore, crucial. The fact that we cannot correct mistakes from higher levels was by no means a feature of the design: we tried to recover from higher-level sorting mistakes by predicting localization of a protein at all nodes and averaging over the prediction strengths over all higher level nodes. Sometimes this worked, however, most of the time such an alteration introduced new mistakes.

**Over 20 percentage points improvement over existing generalized methods**

We also showed that the increase in the size of the training set through the addition of noisy predictions was more relevant than the noise added from the mistakes in these annotations obtained through text-analysis\(^45\) and homology-transfer.\(^43\) This surprising finding suggested that prediction methods might improve even more through the continued addition of large-scale experimental tackling of localization. Finally, we confirmed our previous findings\(^54\) that predicted structure and evolutionary profiles contain information relevant for the prediction of localization. All these data combined with our hierarchical tree-based system improved the overall accuracy over 20 percentage points over the best competitor that generically predicted localization in four states (extra-cellular, nuclear, cytoplasmic, mitochondrial; Table 4). The only method that
performed at a similarly high level as LOCtree was TargetP$^{33}$ (Table 4) that focuses on particular classes (secreted, mitochondria, chloroplast). TargetP$^{33}$ was also the only method that appeared significantly better at predicting one particular compartment, namely, chloroplast proteins (Table 1 of Supplementary Data; note, however, that in this test we did compare our method in cross-validation mode to TargetP in not-cross-validation mode, i.e. were likely to have over-estimated the performance of TargetP). Predictions using LOCtree had the added advantage of being extremely balanced between accuracy and coverage (Table 1, Table 2, Table 3 and Table 4). In contrast, methods such as SignalP$^{23}$ the secreted/not-secreted component of TargetP, are either very prone to over (high coverage, low accuracy) or under-prediction (low coverage, high accuracy; e.g. TargetP, PSORT II). The only other general method that was as well balanced as LOCtree was NNPSL$^{40}$ that had an overall performance of 27 percentage points below LOCtree (Table 4).

Many estimates for performance had a rather short-life span

Another problem that we noticed was that only TargetP and PSORT had published estimates for performance that were close to our results on a “never-seen-before” set of sequence-unique proteins. In particular, SubLoc$^{67}$ was estimated to achieve an overall accuracy of $Q_4=79\%$, while it reached only 57% on our data (Table 4). The differences may be explained by the fact that up to 90% pair-wise sequence identity was allowed between testing and training set for the original publication of SubLoc and NNPSL$^{40}$ Cai \textit{et al.}$^{79}$ also claim very high level of accuracy (73%). That value was more difficult to compare because their methods are not available as servers and because their publications did not rigorously describe protocols for removing redundancy. In fact, it appears that only proteins identical between training and testing set were excluded. Furthermore, the accuracy was compiled on a different partition of the prediction goal. More recently this group published even higher estimates using similar data sets with unspecified sequence similarity between testing and training$^{51}$ and$^{52}$ In general, the problem of correctly estimating performance is a very difficult one as illustrated by the bi-annual meetings for the critical assessment of structure prediction (CASP$^{80, 81, 82, 83}$ and$^{84}$) and by servers that evaluate the performance of servers such as EVA$^{85}$ and$^{86}$ The task is particularly difficult in a field in which we have too few and no continuous resource of experimental data.

As accurate as large-scale experiments?

Numerically, our predictions from LOCtree agreed as much with traditional “small-scale” biochemical determinations of subcellular localization as did the recent large-scale experiments$^{6, 7}$ and$^{8}$ (Table 4 and Table 5). Interestingly, our predictions reached a similar level of performance as large-scale experiments (GFP: 79%, IL: 72%, and LOCtree: 74–78%) if analyzed against more careful traditional approaches as the standard-of-truth. This by no means implies that we aimed at the replacement of experiments. Rather, we see predictions from LOCtree as a reasonable, cheap starting point for careful experiments and as a complement for the interpretation of large-scale results. Furthermore, our prediction method had slightly different potential than the large-scale
experiments, e.g. while we could identify secreted proteins, we currently could not clearly distinguish between Golgi and vesicles, nor did we include membrane proteins.

Open tasks

Our current system marked in some ways the end of a very long series of methods addressed at predicting localization. While methods using homology-transfer (LOChom\textsuperscript{43}) and text-analysis (LOCkey\textsuperscript{45}) were crucial for achieving our new state-of-the-art level of performance we will have to tie some loose ends, in particular, we currently exclude membrane proteins, treat each protein as one without any experimental annotations (LOChom and LOCkey are used for training and for our prediction server; however, they are not generically integrated into LOCtree), and do not distinguish between proteins that are generically native to more than one compartment and those which are not. Furthermore, the task of annotating more than a few representative proteomes remains.

Conclusion

Previous attempts at predicting subcellular localization have implemented machine-learning algorithms using the standard parallel architecture as is common practice in computer science and have focused on improving prediction by incorporating additional sequence features that are correlated with localization. Here we have shown that prediction accuracy can be significantly improved by using a hierarchical architecture of support vector machines to mimic the protein sorting mechanism. This result is likely to hold for other aspects of protein function and can significantly aid the development of more accurate predictors of protein function. The ability of many proteins to function in more than one native subcellular compartment makes the prediction task especially difficult. In fact, over 30% of the more than 4000 yeast proteins for which localization has been determined using high-throughput experiments\textsuperscript{7} are associated with more than one compartment. The hierarchical architecture of LOCtree can better incorporate proteins which spend time in more than one native compartment by predicting “intermediate” localization states, which span multiple subcellular classes, at a much higher accuracy. The fact that the system achieved an overall five-state prediction accuracy of 74% seems to indicate that the native subcellular localization is imprinted somehow onto the protein sequence and that a majority of proteins carry only one strong sequence signal for one particular compartment. In future, it should be possible to further extend the abilities of LOCtree by adding modules that can make fine-grained distinctions such as discriminating among the different organelles and various substructures like the nucleolus.

Methods and Materials

Data sets used for development and evaluation
We selected all eukaryotic and prokaryotic proteins with explicit annotations about subcellular localization in SWISS-PROT release 40. We excluded proteins annotated as MEMBRANE, POSSIBLE, PROBABLE, SPECIFIC PERIODS or BY SIMILARITY. We also excluded proteins annotated with multiple localizations. This left about 9000 eukaryotic proteins and 13,000 prokaryotic proteins in our trusted set of experimentally annotated localization (“SWISS-PROT annotated” set; Table 4 of Supplementary Data). Training and test sets were constructed from this set such that no pair of proteins from any two sets had sequence similarity levels corresponding to HVAL>5 (equation (1)). We picked this value, since below this threshold assigning subcellular localization based solely on homology leads to significant errors. Furthermore, the test set was redundancy reduced at HVAL<10 using a simple greedy search. This ensured that no two proteins in the test set had greater than 25% sequence identity over more than 250 residues (number of sequence unique proteins given in Table 4 of Supplementary Data). The reason for this reduction was to find a balance between biased data known to yield over-estimates and between too small data sets likely to yield incorrect estimates. We did not have to define thresholds for significant sequence similarity between motifs such as signal peptides, since we never explicitly used this information, rather we used the entire protein information.

**SWISS-PROT new set used for testing, only**

After we completed the development of all our methods, we used an additional data set to re-examine performance, namely, we collected all proteins that had been added to SWISS-PROT between release 40 and 41 (results presented in Table 4). We excluded all new proteins that had HVAL>5 (equation (1)) to any previously used protein and found the sequence-unique subset of the new proteins (Table 4 of Supplementary Data). We never used any of these proteins for development, and it is rather unlikely that any of the other methods tested (Table 4) used any of these, since all methods were developed based on SWISS-PROT releases <41.

**HSSP-value to measure pair-wise sequence similarity**

The simplest way to measure sequence similarity is percentage pair-wise sequence identity (PIDE), i.e. the percentage of residues identical between two proteins (not counting gaps). Another measure is the statistical expectation values as reported by BLAST. Here, we used a third measure, namely the HSSP-value (HVAL) because it more accurately allowed the separation between proteins pairs for which similarity in localization is recognizable from sequence than the other two. The HVAL is given by:

\[
HVAL = \text{PID} - \begin{cases}
100 & \text{for } L \leq 11 \\
480 \cdot \left(1 + \exp^{-L/1000}\right) & \text{for } L \leq 450 \\
19.5 & \text{for } L > 450
\end{cases}
\]  (1)
where \( L \) was the number of residues aligned between two proteins, PID the percentage of pair-wise identical residues.

**Increasing size of training set**

Preliminary results suggested that a larger training set improved SVM performance through increased coverage of the sequence space. Another source of improvement was using a sequence redundant set of proteins to train the SVM. Two strategies were used to increase the size of the training set. (1) SWISS-PROT keyword-based annotations: using LOCkey, we first annotated localization for all proteins in the SWISS-PROT database for which adequate keyword functional information was present in the database. Next, proteins with HVAL>5 to proteins in the test set were excluded. The remaining proteins were added to the training set. (2) Homology-based annotations: using LOChom, we annotated localization for all sequence homologues in the SWISS-PROT database of proteins in the training set. Using both procedures increased the size of the training set by almost a factor of 4.

**Building evolutionary profiles**

We have shown that using evolutionary information in the form of sequence profiles significantly improves prediction accuracy. Profiles were built by aligning the sequences against the SWISS-PROT+TrEMBL database using the MaxHom dynamic programming algorithm. The aligned sequences were filtered for redundancy at 95% pair-wise sequence identity, i.e. pairs exceeding this limit were removed. Finally, we included only those proteins that had HVAL>5 and PID>50% with respect to the guide sequence. These thresholds were previously found to be optimal for a rather different prediction method. Finally, the profile composition was calculated by replacing each amino acid residue in the protein by the residue frequencies in the profile. All composition information was input to the SVM in the form of profile-based composition.

**Hierarchical architecture and support vector machine training**

Each decision node in the hierarchical architecture of LOCtree (Figure 1) was implemented using a support vector machine (SVM). The SVMs were implemented using the SVM-light package. We used the following input information: amino acid composition (20 units), composition of the 50 N-terminal residues (20 units), and amino acid composition in the three secondary structure states (60 units). For the eukaryotic plant and non-plant systems, raw output from the SignalP server was used as additional input to the SVM at the top node which determined whether a protein is sorted through the secretory pathway or not. Altogether 100 variables (104 for the top-level SVM in the eukaryotic architecture) were used as input to the SVMs. Each SVM was trained using the radial basis function (RBF) kernel. The \( \gamma \) parameter for the RBF-kernel and the \( C \) parameter for the trade-off between training error and margin were determined by optimization on small subset of the training data. We used a constant value of \( \gamma = 15 \) and \( C = 500 \) for all SVMs. The predictions from the SVMs were observed to be quite resilient to changes in kernel parameters.
**Final decision through simple winner-takes-all**

We experimented with the following two methods for determining the localization of a protein. (1) Decision tree: at each node in the hierarchical architecture (Figure 1), a simple yes/no decision was made based only on the SVM output at that node to determine which branch of the localization tree the protein belongs to. (2) Summing over prediction strengths: the branch of the localization tree the protein is sorted through at each node was determined by summing the prediction strength's over all previous nodes. The simple decision tree architecture was finally used, since it outperformed the summing architecture by over one percentage point.

**Cross-validation**

The “SWISS-PROT annotated” data was partitioned into six sets of equal size using the HVAL criteria described above: five of these sets were combined to give the training set and the sixth one was used for testing. Finally, we rotated through the test set such that each protein was used for testing exactly once. We never used any information from the test set to optimize parameters.

**Evaluating performance**

As a simple measure for performance we used the percentage accuracy (equation (5)). The accuracy/specificity and coverage/sensitivity of the two-state networks were measured using four ratios derived from $TP$ (true positives, i.e. the number of proteins predicted to be in localization $L$ and experimentally observed to be in localization $L$), $FP$ (false positives, i.e. the number of proteins predicted to be in localization $L$ and observed in not-$L$) and $FN$ (false negatives, i.e. the number of proteins predicted not to be in not-$L$ and observed to be in $L$). We used:

$$\text{Acc}(L) = 100 \times \frac{TP}{TP + FP}$$  \hspace{1cm} (2)

$$\text{Cov}(L) = 100 \times \frac{TP}{TP + FN}$$  \hspace{1cm} (3)

In other words, $\text{Acc}(L)$ is the accuracy/specificity in predicting localization $L$, and $\text{Cov}(L)$ is the corresponding coverage/selectivity. We combined these two numbers through the geometric average:

$$\text{gAv}(L) = \frac{1}{100} \cdot \sqrt{\text{Acc}(L) \cdot \text{Cov}(L)}$$  \hspace{1cm} (4)
We omitted the specific qualifier \( L \) in text and Tables whenever the localization that we referred to was obvious. The overall accuracy was measured by \( Q \):

\[
Q = 100 \times \frac{TP}{TP + FN} = 100 \times \frac{\text{number correctly predicted}}{\text{number of proteins in data set}}
\] (5)

Note that the values of \( Q \) we presented in text and Tables referred to different number of states, depending on which level in the hierarchy (Figure 1) we monitored the overall accuracy. Where needed, we indicated the number of states through subscripts, e.g. \( Q_5 \) is the overall five-state accuracy that measures the accuracy in predicting one of five classes of localization for non-plant eukaryotes.

For the two class predictions we also reported the Mathews correlation coefficient (MCC)\(^{94}\) and the normalized mutual information (MI):\(^{95}\)

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}
\] (6)

\[
\text{MI} = -\frac{TP + FN}{N} \cdot \log \left( \frac{TP + FN}{N} \right) - \frac{TN + FP}{N} \log \left( \frac{TN + FP}{N} \right)
\] (7)

The Mathews correlation coefficient and the mutual information as defined above are only applicable for two classes. For more than two classes, the Mathews correlation coefficient was modified to the generalized correlation coefficient:\(^{96}\)

\[
\text{GC}^2 = \frac{\sum_{ij} (z_{ij} - e_{ij})^2}{e_{ij}}
\] (8)

where \( N \) was the number of proteins, \( K \) the number of localization classes, \( z_{ij} \) the confusion matrix and \( e_{ij} \) was given by:

\[
e_{ij} = \frac{\text{Acc}_i \times \text{Cov}_j}{N}
\] (9)

For many classes the mutual information was modified to the information coefficient:\(^{96}\)
\[ MI = \sum_i \left\{ \frac{\text{Cov}_i}{N} \log \left( \frac{\text{Cov}_i}{N} \right) + \sum_j \frac{Z_{ij}}{N} \log \left( \frac{Z_{ij}}{\text{Acc}_j} \right) \right\} \]  

(10)

**Prediction methods**

The prediction accuracy of four publicly available methods was evaluated using the “Non-plant unique”, “Plant unique” and the “Non-plant new unique” sets (Table 4 of Supplementary Data). The four methods were: (1) TargetP: neural network based tool predicting localization based on N-terminal sequence information,23,24 and 33 (2) SubLoc: support vector machine prediction of localization from amino acid composition,67 (3) PSORT II: integrated method based on detecting sorting signals and predictions from other methods like NNPSL and SignalP;25,58 and 97 and (4) NNPSL: neural network based tool predicting localization from amino acid composition.40 All methods were run with default parameter settings.

**Estimate for composition in entire proteomes**

Three different values determined our final estimates for the percentage of proteins in localization \( L \) in entirely sequenced proteomes (Figure 3), namely the number of proteins actually predicted by LOCtree to be in localization \( L \), our estimate for the accuracy (equation (2)) and coverage (equation (3)) of that prediction. In detail:

\[ N_{\text{prd}}(L) = \frac{N_{\text{prot notTMH}}}{N_{\text{prd corrected sum}}} \cdot N_{\text{prd corrected}}(L) \]  

(11)

with

\[ N_{\text{prd corrected}}(L) = \frac{\text{Acc}(L) \cdot N_{\text{prd raw}}(L)}{\text{Cov}(L)} \]  

\[ N_{\text{prd corrected sum}} = \sum_{\forall L} N_{\text{prd corrected}}(L) \]

where \( N_{\text{prd raw}}(L) \) was the number of proteins predicted directly by LOCtree to be in localization \( L \), Acc(L) and Cov(L) the estimates for prediction accuracy and coverage, respectively, and \( N_{\text{prot notTMH}} \) the number of proteins without membrane helices. The error in the number of proteins predicted to be in localization \( L \) was estimated by:

\[ \frac{\sigma_N(L)}{N(L)} = \sqrt{\left( \frac{\sigma_{\text{Acc}(L)}}{\text{Acc}(L)} \right)^2 + \left( \frac{\sigma_{\text{Cov}(L)}}{\text{Cov}(L)} \right)^2} \]  

(12)

**Acknowledgements**
Thanks to Jinfeng Liu (Columbia University) for computer assistance and the collection of genome data sets and to Kazimierz Wrzeszczynski (Columbia University) for proof reading the manuscript. Thanks to Astrid Reinhardt (Baylor College of Medicine, Texas), Tim Hubbard (Sanger Centre, Hinxton), Sujun Hua (Tsinghua University), Zhirong Hun (Tsinghua University), Olof Emanuelsson (Stockholm University), Henrik Nielsen (CBS, Copenhagen), Søren Brunak (CBS, Copenhagen), Gunnar von Heijne (Stockholm University), Kenta Nakai (Tokyo University) and Paul Horton (UCB) for access to their prediction methods. Thanks to both anonymous referees for important corrections. Last, but not least, thanks to Amos Bairoch (SIB, Geneva), Rolf Apweiler (EBI, Hinxton), Phil Bourne (UCSD), and their crews for maintaining excellent databases and to all experimentalists who enabled this analysis by making their data publicly available.

References


19 J.B. Peltier, O. Emanuelsson, D.E. Kalume, J. Ytterberg, G. Friso and A. Rudella et al., Central functions of the lumenal and peripheral thylakoid proteome of *Arabidopsis*


---

**Appendix. Supplementary data**

<table>
<thead>
<tr>
<th>Class</th>
<th>Method</th>
<th>Eukaryotic Non-plant</th>
<th>Plant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOCtree</td>
<td>TargetP 1</td>
<td>SubLoc 2</td>
</tr>
<tr>
<td>Secr</td>
<td><strong>Acc</strong></td>
<td>81</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td><strong>Cov</strong></td>
<td>80</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>gAv</td>
<td>81</td>
<td>80</td>
</tr>
<tr>
<td>Ext</td>
<td><strong>Acc</strong></td>
<td>83</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>Cov</strong></td>
<td>81</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>gAv</td>
<td>82</td>
<td>-</td>
</tr>
<tr>
<td>Nuc</td>
<td><strong>Acc</strong></td>
<td>78</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>Cov</strong></td>
<td>78</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>gAv</td>
<td>78</td>
<td>-</td>
</tr>
<tr>
<td>Cyt</td>
<td><strong>Acc</strong></td>
<td>63</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>Cov</strong></td>
<td>66</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>gAv</td>
<td>64</td>
<td>-</td>
</tr>
<tr>
<td>Mit</td>
<td><strong>Acc</strong></td>
<td>70</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td><strong>Cov</strong></td>
<td>67</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>gAv</td>
<td>68</td>
<td>63</td>
</tr>
<tr>
<td>Chloro</td>
<td>Acc</td>
<td>Cov</td>
<td>gAv</td>
</tr>
<tr>
<td>--------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>74</td>
<td>63</td>
<td>53</td>
</tr>
</tbody>
</table>

Abbreviations used as in Table 1, with the following exceptions:

**Data set**: all sequence-unique eukaryotic non-plant proteins added between release 41 and 40 of SWISS-PROT (*Non-plant new unique* in Table 4_Mat). Note that none of the proteins in this set had significant sequence similarity to any of the proteins that had annotations about localization in SWISS-PROT at the time of development of the prediction methods for which results are shown. In this sense, our test also could provide an independent and likely more accurate estimate for the sustained performance than some of the original publications for some of the methods.

**Localization**: Ext, extra-cellular; Nuc, nuclear; Cyt, cytosolic; Mit, mitochondria; Chloro, chloroplast.

**Methods**: Predictions from methods other than LOCtree - introduced here - were taken from their public Internet servers (Methods), except for PSORT II that was run locally; numbers under methods refer to original publication (References).

Numbers in bold: in each row, the best method(s) is (are) marked in bold letters; methods are grouped according to significant differences (below), i.e. all values that are statistically indistinguishable are marked as one best group.

**Significant differences**: For LOCtree, the standard deviation in the five-state accuracy was roughly six percentage points. The following estimates for standard deviations were published: TargetP ¹, about one percentage point; NNPSL ⁴, about 2.5 percentage points; PSORT II ³, about 3.5 percentage points. Since no error estimates were published for SubLoc, we used 2.5 percentage points as the mean over the other three.

**Performance measures**: Acc: accuracy or specificity (Eqn. 2); Cov: coverage or selectivity (Eqn. 3); gAv: geometric average between Acc and Cov (Eqn. 4); Q: overall prediction accuracy for a given level in the hierarchy (Eqn. 5, note this is a 5-state value for non-plants and 6-state value for plants).

Note: Our testing procedure overestimates the prediction accuracy of the public methods since some of the test sequences could have been used during training for these methods.
### Table 2. Mat: Estimating the influence of missing N-termini and sequencing mistakes. *

<table>
<thead>
<tr>
<th>Method</th>
<th>LOCtree full length</th>
<th>LOCtree Rand</th>
<th>LOCtree Nterm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sec Path</td>
<td>Acc</td>
<td>87</td>
<td>79</td>
</tr>
<tr>
<td>Cov</td>
<td>90</td>
<td>83</td>
<td>68</td>
</tr>
<tr>
<td>gAv</td>
<td>88</td>
<td>81</td>
<td>63</td>
</tr>
<tr>
<td>Nuc</td>
<td>Acc</td>
<td>86</td>
<td>72</td>
</tr>
<tr>
<td>Cov</td>
<td>93</td>
<td>83</td>
<td>67</td>
</tr>
<tr>
<td>gAv</td>
<td>89</td>
<td>77</td>
<td>61</td>
</tr>
<tr>
<td>Ext</td>
<td>Acc</td>
<td>77</td>
<td>68</td>
</tr>
<tr>
<td>Cov</td>
<td>85</td>
<td>80</td>
<td>66</td>
</tr>
<tr>
<td>gAv</td>
<td>81</td>
<td>74</td>
<td>65</td>
</tr>
<tr>
<td>Cyt</td>
<td>Acc</td>
<td>82</td>
<td>72</td>
</tr>
<tr>
<td>Cov</td>
<td>64</td>
<td>44</td>
<td>42</td>
</tr>
<tr>
<td>gAv</td>
<td>72</td>
<td>56</td>
<td>48</td>
</tr>
<tr>
<td>Mit</td>
<td>Acc</td>
<td>73</td>
<td>54</td>
</tr>
<tr>
<td>Cov</td>
<td>78</td>
<td>68</td>
<td>42</td>
</tr>
<tr>
<td>gAv</td>
<td>75</td>
<td>61</td>
<td>41</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td></td>
<td>78</td>
<td>68</td>
</tr>
</tbody>
</table>

* Abbreviations used as in Table 1_Mat, with the following exceptions:

**Methods**: We estimated the effect of sequencing errors on the performance of LOCtree in two different ways: **LOCtree Rand**, we randomly picked positions within the amino acid sequence and cleaved off one third of the protein sequence. Input to LOCtree was calculated based on the remaining two-thirds of the protein sequence. **LOCtree Nterm**, the first thirty N-terminal residues were cleaved off for all proteins. **LOCtree full length**, is the performance of LOCtree on full length protein sequences and is shown here for comparison.
Table 3 Mat: LOCtree estimate of localizations for entire proteomes. *

<table>
<thead>
<tr>
<th>Organism</th>
<th>Nprot</th>
<th>Secretory Pathway</th>
<th>Intra-cellular</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext</td>
<td>Org</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>30371</td>
<td>20.9 (2)</td>
<td>3.5 (1)</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>5004</td>
<td>10.1 (1)</td>
<td>5.0 (2)</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>21085</td>
<td>7.8 (1)</td>
<td>5.9 (2)</td>
</tr>
</tbody>
</table>

**Data**: Nprot: number of non-membrane proteins in the genome; membrane proteins were identified by PHDhtm. The standard deviation for the estimated percentages of the different compartments is shown in brackets.

**Localization**: Ext: extra-cellular; Nuc: nuclear; Cyt: cytoplasm; Mit: mitochondria; Chloro: chloroplast; Org: organellar proteins, i.e. proteins localized in either the endoplasmic reticulum, Golgi apparatus, peroxysome, lysosome or vacuole.
Table 4_Mat: Number of proteins in data set. *

<table>
<thead>
<tr>
<th>Sub-cellular localization</th>
<th>Eukaryotes</th>
<th>Prokaryotes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SWISS-PROT annotated</td>
<td>Non-plant unique</td>
</tr>
<tr>
<td>Nucleus</td>
<td>2673</td>
<td>562</td>
</tr>
<tr>
<td>Cytosol</td>
<td>2137</td>
<td>330</td>
</tr>
<tr>
<td>Extra-cellular</td>
<td>1936</td>
<td>363</td>
</tr>
<tr>
<td>Chloroplast</td>
<td>952</td>
<td>-</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>914</td>
<td>198</td>
</tr>
<tr>
<td>Periplasmic</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Organelles</td>
<td>367</td>
<td>52</td>
</tr>
<tr>
<td>SUM</td>
<td>8979</td>
<td>1505</td>
</tr>
</tbody>
</table>

* Data sets:
- **SWISS-PROT annotated**: number of all non-membrane proteins with annotated experimentally determined sub-cellular localization taken from SWISS-PROT release 40 (Methods);
- **Non-plant unique**: number of non-plant proteins in sequence unique subset of eukaryotic proteins (Methods);
- **Non-plant new unique**: number of non-plant eukaryotic proteins in sequence-unique subset of all proteins found in SWISS-PROT release 41 and not in release 40 (chosen by same procedure as SWISS-PROT unique).
References for 'Supporting online material'


