

Mimicking Cellular Sorting Improves Prediction of Subcellular Localization

Rajesh Nair^{a, d, ✉, ✉} and Burkhard Rost^a

^{a, b, c}

^aCUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA

^bColumbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St. Nicholas Avenue, New York, NY 10032, USA

^cNorthEast Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA

^dDepartment of Physics, Columbia University, 538 West 120th Street, New York, NY 10027, USA

Received 27 December 2004; revised 8 February 2005; accepted 9 February 2005.
Edited by G. von Heijne. Available online 5 March 2005.

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

fluorescent protein; GFP-tagging, here used to refer to the large-scale experimental determination of localization through GFP tagging; GO, gene ontology; HSSP, database of protein structure-sequence alignments; IL, large scale localization of yeast proteins from high-throughput immuno-localization of epitope-tagged gene products; LOChom, predicting localization using annotation transfer through sequence homology; LOCKey, using SWISS-PROT keywords to predict subcellular localization; LOCTree, hierarchical system of SVMs introduced here; NLS, nuclear localization signal; NNPSL, neural networks predicting localization; PHD, profile based neural network prediction of secondary structure, solvent accessibility and transmembrane helices; PredictNLS, prediction of nuclear proteins through nuclear localization signals; PROFphd, advanced profile-based neural network prediction of secondary structure and solvent accessibility; PSORT, knowledge-based expert system using amino acid composition and sequence motifs; SGD, *Saccharomyces cerevisiae* genome database; SignalP, neural network system predicting signal peptides; SMART, simple modular architecture research tool; SVM, support vector machine; SWISS-PROT, data base of protein sequences; SubLoc, support-vector machine-based prediction of localization; TargetP, combined method predicting chloroplast (ChloroP), extra-cellular (SignalP), and mitochondrial proteins; TMS, identification of chloroplast proteins in *A. thaliana* using tandem mass spectroscopy; TrEMBL, translation of the EMBL-nucleotide database coding DNA to protein sequences

Article Outline

[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](#)

[As accurate as large-scale experiments?](#)
[Open tasks](#)
[Conclusion](#)
[Methods and Materials](#)
[Data sets used for development and evaluation](#)
[SWISS-PROT new set used for testing, only](#)
[HSSP-value to measure pair-wise sequence similarity](#)
[Increasing size of training set](#)
[Building evolutionary profiles](#)
[Hierarchical architecture and support vector machine training](#)
[Final decision through simple winner-takes-all](#)
[Cross-validation](#)
[Evaluating performance](#)
[Prediction methods](#)
[Estimate for composition in entire proteomes](#)
[Acknowledgements](#)
Appendix. [Supplementary data](#)
[References](#)

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^{1 and 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[†], and contribute about 413,000 protein sequences, i.e. about one-fourth of all currently known protein sequences.^{3, 4 and 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 *Saccharomyces cerevisiae* (baker's yeast).^{6 and 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 *Arabidopsis thaliana* (weed).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.⁹ In contrast, computational tools can provide fast and accurate localization predictions for any organism.^{10, 11, 12 and 13} Attempts to predict subcellular localization have increasingly become one of the central problems in bioinformatics/computational biology.^{14, 15, 16, 17, 18, 19 and 20}

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^{21, 22, 23, 24, 25 and 26} or nuclear localization signals (NLS)^{15, 27, 28 and 29} 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.^{30 and 31} 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 mitochondria 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.^{17, 24, 32 and 33} 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.²⁹ 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.^{34, 35, 36 and 37} 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.^{38 and 39} Furthermore, a particular problem for methods detecting N-terminal signals is that start codons are predicted with less than 70% accuracy by genome projects.^{1, 2 and 40} Overall, known and predicted sequence motifs enable annotating about 30% of the proteins in six entirely sequenced eukaryotic proteomes.^{4, 41 and 42} Other methods that can be reliably used to annotate localization but are not always applicable are annotation transfer from sequence homologues⁴³ and text analysis.^{44, 45, 46 and 47} A particular variant of homology-based predictions is the domain projection method that is based on similarity to SMART domains of known subcellular localization.⁴¹ 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,^{48 and 49} which is mostly due

to the altering of the protein surface in response to changing environmental conditions.⁵⁰ Methods using only amino acid composition to predict localization are *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.^{51, 52 and 53} Recently, we showed that incorporating structural and evolutionary information significantly improves prediction accuracy.⁵⁴ With the availability of many completely sequenced genomes, phylogenetic profiles have been employed to identify subcellular localization.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.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,^{57 and 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).^{59 and 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 *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.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 ([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.

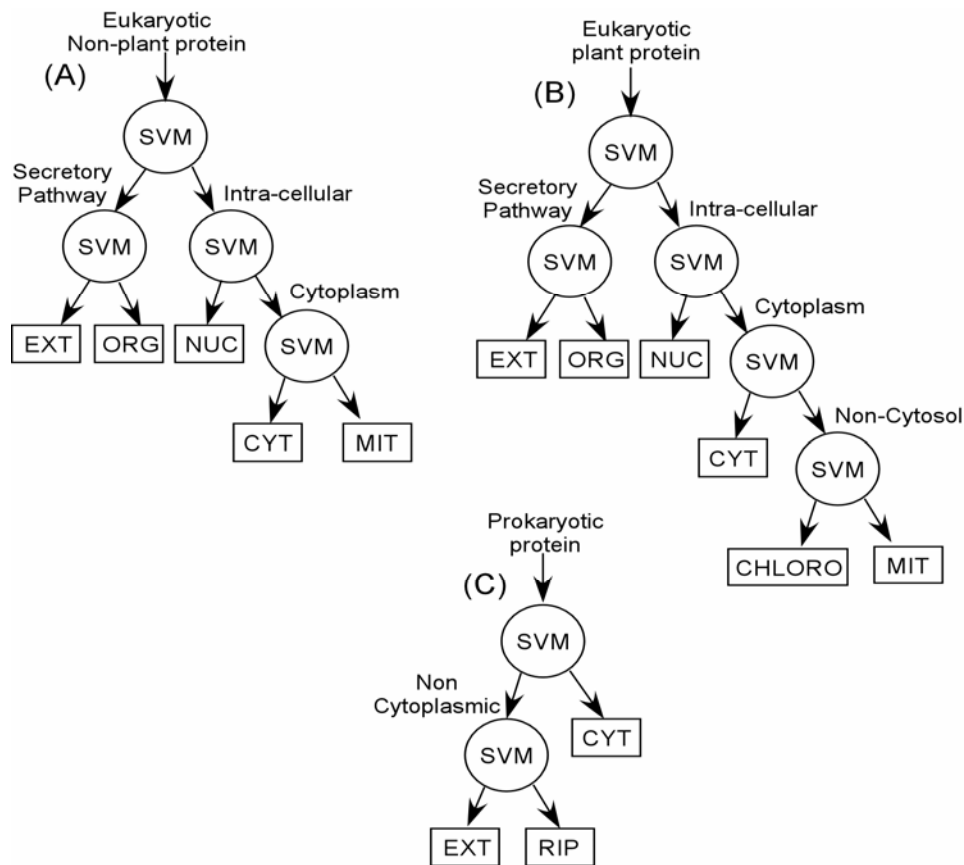


Figure 1. Hierarchical architecture of LOCTree. LOCTree uses specialized architecture to predict subcellular localization of proteins from different organisms: (a) architecture for eukaryotic non-plant proteins; (b) architecture for plant proteins; and (c) the architecture for prokaryotic proteins. At 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, peroxisomes, 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-PROT⁶¹ ([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.^{42, 62, 63, 64 and 65} 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 (LOChom⁴³) and keyword-based (LOCKey⁴⁵) 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.⁴⁵ Plants were treated separately since their compositional features differed significantly from non-plant eukaryotes (not shown). Using the SVM-light package,⁶⁶ 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.⁵⁴ 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

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

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).

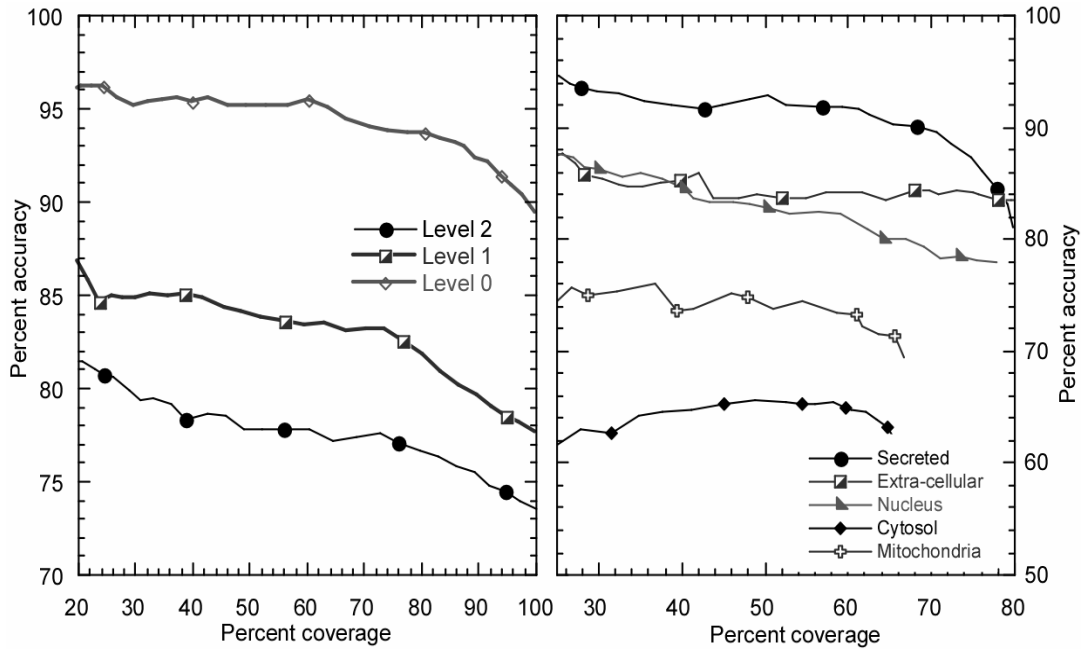


Figure 2. Reliability of LOctree. The curves show prediction accuracy of LOctree for eukaryotic animal sequences. (a) Overall performance: the prediction accuracy decreases as we descent the hierarchical tree (Figure 1(a)). The Level 2 accuracy shown includes the accuracy of all Level 1 leaves like the 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

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

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 $Q_5=74\%$ ([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

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

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,³³ SubLoc,⁶⁷ NNPSL,⁴⁰ and PSORT II.²⁵ 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, SubLoc⁶⁷ is also based on SVMs; NNPSL⁴⁰ is based on neural networks. SubLoc and NNPSL rely solely on amino acid composition while PSORT²⁵ 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 3.0,¹⁷ 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 3.0¹⁷ 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

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

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, 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 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,³³ about one percentage point; NNPSL,⁴⁰ about 2.5 percentage points; PSORT II,²⁵ about 3.5 percentage points. Since no error estimates were published for SubLoc,⁶⁷ 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.^{68, 69, 70 and 71} 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 group⁶ determined the localization for about 60%, while the O'Shea group⁷ 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.⁷² and 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

Method	Nuclear			Mitochondrial		
	Obs	Acc	Cov	Obs	Acc	Cov
GFP ⁷	586	82	68	418	83	51
IL ⁶	124	88	64	60	76	62

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;⁷ IL, large scale localization using high-throughput immunolocalization of epitope-tagged proteins.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 *et al.*⁸ recently identified 690 proteins localized in the chloroplast of *A. thaliana* (weed). Of these we predicted 190 to contain membrane helices. We compared LOCTree and TargetP³³ 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.⁷⁴

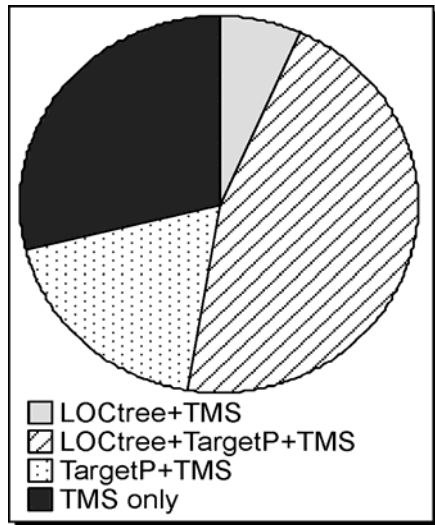


Figure 3. Benchmarking LOCTree using large-scale experimental data. Both LOCTree and TargetP³³ were used to predict the localization of nearly 500 chloroplast proteins from *A. thaliana* which were identified using tandem mass spectroscopy (TMS) by Kleffmann *et al.*⁸ LOCTree and TargetP both predicted 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),^{1 and 2} *A. thaliana* (weed),⁷⁵ and *S. cerevisiae* (yeast)¹⁹ (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.

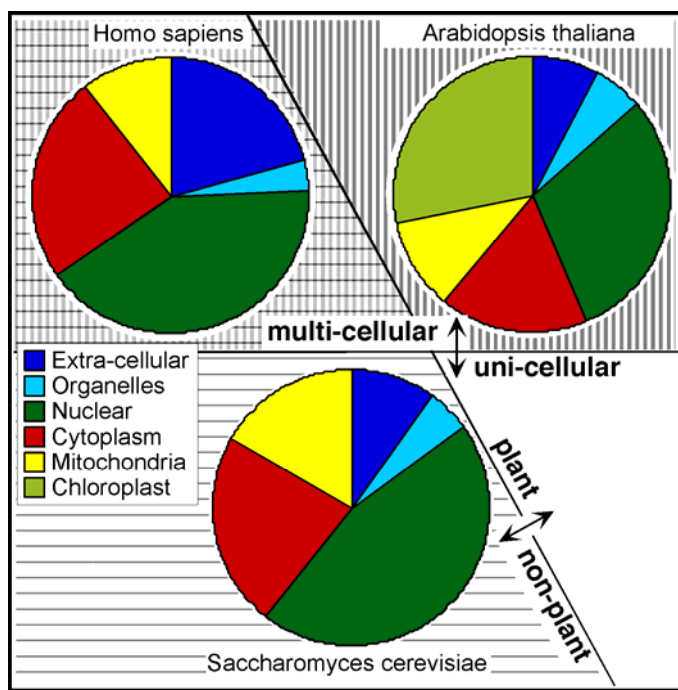


Figure 4. Composition of compartments in three representative proteomes. Note that 100% of the pie charts represents the number of proteins without transmembrane helices (predicted by PHDhtm^{98, 99 and 100} and taken from PEP³). The final estimates were corrected in order to account for our compartment-specific 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 ($\pm 9\%$), organelle ($\pm 34\%$), nuclear ($\pm 10\%$), cytosol ($\pm 20\%$), mitochondria ($\pm 25\%$) and for chloroplast ($\pm 17\%$).

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^{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, 50 and 77}. 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 other⁷⁸). 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 descends 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-analysis⁴⁵ 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 findings⁵⁴ 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 TargetP³³ (Table 4) that focuses on particular classes (secreted, mitochondria, chloroplast). TargetP³³ 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 NNPSL⁴⁰ 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, SubLoc⁶⁷ 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 *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 (LOChom⁴³) and text-analysis (LOCKey⁴⁵) 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 experiments⁷ 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 40.⁸⁷ 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.⁴³ Furthermore, the test set was redundancy reduced at $HVAL < 10$ using a simple greedy search.⁸⁸ 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^{89 and 90} 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,⁸⁹ 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.⁴³ The HVAL^{91 and 92} is given by:

$$HVAL = PID - \begin{cases} 100 & \text{for } L \leq 11 \\ 480 \cdot L^{-0.32 \cdot \{1 + \exp^{-L/1000}\}} & \text{for } L \leq 450 \\ 19.5 & \text{for } L > 450 \end{cases} \quad (1)$$

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,⁴⁵ 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,⁴³ 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 shown⁵⁴ 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.⁹³ 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.⁵⁴ 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.⁶⁶ 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 server²³ 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 γ 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:

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

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

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:

$$\text{gAv}(L) = \frac{1}{100} \cdot \sqrt{\text{Acc}(L) \cdot \text{Cov}(L)} \quad (4)$$

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 :

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

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:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \quad (6)$$

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

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

$$GC^2 = \frac{\sum_{ij} \frac{(z_{ij} - e_{ij})^2}{e_{ij}}}{N(K-1)} \quad (8)$$

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

$$e_{ij} = \frac{Acc_i \times Cov_j}{N} \quad (9)$$

For many classes the mutual information was modified to the information coefficient:⁹⁶

$$MI = \sum_i \left\{ -\frac{Cov_i}{N} \log\left(\frac{Cov_i}{N}\right) + \sum_j \frac{Z_{ij}}{N} \log\left(\frac{Z_{ij}}{Acc_j}\right) \right\} \quad (10)$$

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:

$$N_{\text{prd}}(L) = \frac{N_{\text{prot_notTMH}}}{N_{\text{prd_corrected_sum}}} \cdot N_{\text{prd_corrected}}(L) \quad (11)$$

with

$$N_{\text{prd_corrected}}(L) = \frac{Acc(L) \cdot N_{\text{prd_raw}}(L)}{Cov(L)} \quad \text{and} \quad N_{\text{prd_corrected_sum}} = \sum_{\forall L} N_{\text{prd_corrected}}(L)$$

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:

$$\frac{\sigma_{N(L)}}{N(L)} = \sqrt{\left(\frac{\sigma_{Acc(L)}}{Acc(L)}\right)^2 + \left(\frac{\sigma_{Cov(L)}}{Cov(L)}\right)^2} \quad (12)$$

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, not but 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

[1](#) J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural and G.G. Sutton *et al.*, The sequence of the human genome, *Science* **291** (2001), pp. 1304–1351. [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[2](#) E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody and J. Baldwin *et al.*, Initial sequencing and analysis of the human genome, *Nature* **409** (2001), pp. 860–921. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[3](#) P. Carter, J. Liu and B. Rost, PEP: predictions of entire proteomes, *NAR* **31** (2003), pp. 410–413. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[4](#) J. Liu and B. Rost, Target space for structural genomics revisited, *Bioinformatics* **18** (2002), pp. 922–933. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[5](#) M. Pruess, W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey and E. Kriventseva *et al.*, The Proteome Analysis database: a tool for the *in silico* analysis of whole proteomes, *Nucl. Acids Res.* **31** (2003), pp. 414–417. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[6](#) A. Kumar, S. Agarwal, J.A. Heyman, S. Matson, M. Heidtman and S. Piccirillo *et al.*, Subcellular localization of the yeast proteome, *Genes Dev.* **16** (2002), pp. 707–719. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[7](#) W.K. Huh, J.V. Falvo, L.C. Gerke, A.S. Carroll, R.W. Howson, J.S. Weissman and E.K. O'Shea, Global analysis of protein localization in budding yeast, *Nature* **425** (2003),

pp. 686–691. [Abstract-Compendex](#) | [Abstract-Elsevier BIOBASE](#) | [Full Text via CrossRef](#)

[8](#) T. Kleffmann, D. Russenberger, A. von Zychlinski, W. Christopher, K. Sjolander, W. Gruissem and S. Baginsky, The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions, *Curr. Biol.* **14** (2004), pp. 354–362. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(255 K\)](#)

[9](#) T.N. Davis, Protein localization in proteomics, *Curr. Opin. Chem. Biol.* **8** (2004), pp. 49–53. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(110 K\)](#)

[10](#) P. Bork and E.V. Koonin, Predicting functions from protein sequences—where are the bottlenecks?, *Nature Genet.* **18** (1998), pp. 313–318. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[11](#) E.V. Koonin, Bridging the gap between sequence and function, *Trends Genet.* **16** (2000), p. 16. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(59 K\)](#)

[12](#) K. Nakai, Prediction of *in vivo* fates of proteins in the era of genomics and proteomics, *J. Struct. Biol.* **134** (2001), pp. 103–116. [Abstract](#) | [Abstract + References](#) | [PDF \(96 K\)](#)

[13](#) B. Rost, J. Liu, R. Nair, K.O. Wrzeszczynski and Y. Ofran, Automatic prediction of protein function, *Cell. Mol. Life Sci.* **60** (2003), pp. 2637–2650. [Abstract-MEDLINE](#)

[14](#) F. Eisenhaber and P. Bork, Wanted: subcellular localization of proteins based on sequence, *Trends Cell Biol.* **8** (1998), pp. 169–170. [Abstract](#) | [PDF \(210 K\)](#)

[15](#) K. Nakai, Protein sorting signals and prediction of subcellular localization, *Advan. Protein Chem.* **54** (2000), pp. 277–344. [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#)

[16](#) G. Schneider and U. Fechner, Advances in the prediction of protein targeting signals, *Proteomics* **4** (2004), pp. 1571–1580. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[17](#) J.D. Bendtsen, H. Nielsen, G. von Heijne and S. Brunak, Improved prediction of signal peptides: SignalP 3.0, *J. Mol. Biol.* **340** (2004), pp. 783–795. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#)

[18](#) M. Rapp, D. Drew, D.O. Daley, J. Nilsson, T. Carvalho and K. Melen *et al.*, Experimentally based topology models for *E. coli* inner membrane proteins, *Protein Sci.* **13** (2004), pp. 937–945. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[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*

determined by experimentation and genome-wide prediction, *Plant Cell* **14** (2002), pp. 211–236. [Abstract-Elsevier BIOBASE](#) | [Full Text via CrossRef](#)

[20](#) L.J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames and C. Kesmir *et al.*, Prediction of human protein function from post-translational modifications and localization features, *J. Mol. Biol.* **319** (2002), pp. 1257–1265. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(431 K\)](#)

[21](#) G. von Heijne, Protein sorting signals: simple peptides with complex functions, *Exs* **73** (1995), pp. 67–76. [Abstract-MEDLINE](#)

[22](#) K.J. Nielsen, J.M. Hill, M.A. Anderson and D.J. Craik, Synthesis and structure determination by NMR of a putative vacuolar targeting peptide and model of a proteinase inhibitor from *Nicotiana glauca*, *Biochemistry* **35** (1996), pp. 369–378. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[23](#) H. Nielsen, J. Engelbrecht, S. Brunak and G. von Heijne, A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, *Int. J. Neural. Syst.* **8** (1997), pp. 581–599. [Abstract-INSPEC](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[24](#) O. Emanuelsson, H. Nielsen and G. von Heijne, ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites, *Protein Sci.* **8** (1999), pp. 978–984. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#)

[25](#) K. Nakai and P. Horton, PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, *Trends Biochem. Sci.* **24** (1999), pp. 34–36. [Abstract-MEDLINE](#)

[26](#) O. Emanuelsson and G. von Heijne, Prediction of organellar targeting signals, *Biochim. Biophys. Acta* **1541** (2001), pp. 114–119. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(146 K\)](#)

[27](#) M. Cokol, R. Nair and B. Rost, Finding nuclear localisation signals, *EMBO Rep.* **1** (2000), pp. 411–415. [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[28](#) T. La Cour, R. Gupta, K. Rapacki, K. Skriver, F.M. Poulsen and S. Brunak, NESbase version 1.0: a database of nuclear export signals, *Nucl. Acids Res.* **31** (2003), pp. 393–396. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[29](#) R. Nair, P. Carter and B. Rost, NLSdb: database of nuclear localization signals, *Nucl. Acids Res.* **31** (2003), pp. 397–399. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

- [30](#) S.L. Rusch and D.A. Kendall, Protein transport *via* amino-terminal targeting sequences: common themes in diverse systems, *Mol. Membr. Biol.* **12** (1995), pp. 295–307. [Abstract-MEDLINE](#)
- [31](#) G. Schatz and B. Dobberstein, Common principles of protein translocation across membranes, *Science* **271** (1996), pp. 1519–1526. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#)
- [32](#) H. Nielsen, J. Engelbrecht, S. Brunak and G. von Heijne, Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, *Protein Eng.* **10** (1997), pp. 1–6. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [33](#) O. Emanuelsson, H. Nielsen, S. Brunak and G. von Heijne, Predicting subcellular localization of proteins based on their N-terminal amino acid sequence, *J. Mol. Biol.* **300** (2000), pp. 1005–1016. [Abstract](#) | [Abstract + References](#) | [PDF \(189 K\)](#)
- [34](#) D. Rapaport, Finding the right organelle. Targeting signals in mitochondrial outer-membrane proteins, *EMBO Rep.* **4** (2003), pp. 948–952. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [35](#) N.J. Hoogenraad and M.T. Ryan, Translocation of proteins into mitochondria, *IUBMB Life* **51** (2001), pp. 345–350. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#)
- [36](#) S.A. Paschen and W. Neupert, Protein import into mitochondria, *IUBMB Life* **52** (2001), pp. 101–112. [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [37](#) D.J. Schnell, Functions and origins of the chloroplast protein-import machinery, *Essays Biochem.* **36** (2000), pp. 47–59. [Abstract-MEDLINE](#)
- [38](#) W. Nickel, The mystery of non-classical protein secretion. A current view on cargo proteins and potential export routes, *Eur. J. Biochem.* **270** (2003), pp. 2109–2119. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [39](#) J.D. Bendtsen, L.J. Jensen, N. Blom, G. Von Heijne and S. Brunak, Feature-based prediction of non-classical and leaderless protein secretion, *Protein Eng. Des. Sel.* **17** (2004), pp. 349–356. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Full Text via CrossRef](#)
- [40](#) A. Reinhardt and T. Hubbard, Using neural networks for prediction of the subcellular location of proteins, *Nucl. Acids Res.* **26** (1998), pp. 2230–2236. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

- [41](#) R. Mott, J. Schultz, P. Bork and C.P. Ponting, Predicting protein cellular localization using a domain projection method, *Genome Res.* **12** (2002), pp. 1168–1174. [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [42](#) J. Liu and B. Rost, Comparing function and structure between entire proteomes, *Protein Sci.* **10** (2001), pp. 1970–1979. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [43](#) R. Nair and B. Rost, Sequence conserved for subcellular localization, *Protein Sci.* **11** (2002), pp. 2836–2847. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [44](#) W. Fleischmann, S. Moller, A. Gateau and R. Apweiler, A novel method for automatic functional annotation of proteins, *Bioinformatics* **15** (1999), pp. 228–233. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [45](#) R. Nair and B. Rost, Inferring sub-cellular localization through automated lexical analysis, *Bioinformatics* **18** (2002), pp. S78–S86. [Abstract-MEDLINE](#)
- [46](#) E. Kretschmann, W. Fleischmann and R. Apweiler, Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT, *Bioinformatics* **17** (2001), pp. 920–926. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [47](#) Z. Lu, D. Szafron, R. Greiner, P. Lu, D.S. Wishart and B. Poulin *et al.*, Predicting subcellular localization of proteins using machine-learned classifiers, *Bioinformatics* **20** (2004), pp. 547–556. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [48](#) K. Nishikawa, Y. Kubota and T. Ooi, Classification of proteins into groups based on amino acid composition and other characters: I. Angular distribution, *J. Biochem.* **94** (1983), pp. 981–995. [Abstract-MEDLINE](#)
- [49](#) H. Nakashima and K. Nishikawa, Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies, *J. Mol. Biol.* **238** (1994), pp. 54–61. [Abstract](#) | [PDF \(407 K\)](#)
- [50](#) M.A. Andrade, S.I. O'Donoghue and B. Rost, Adaptation of protein surfaces to subcellular location, *J. Mol. Biol.* **276** (1998), pp. 517–525. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(510 K\)](#)
- [51](#) Y.D. Cai, X.J. Liu, X.B. Xu and K.C. Chou, Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect, *J. Cell Biochem.* **84** (2002), pp. 343–348. [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

- [52](#) K.C. Chou and Y.D. Cai, Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition, *J. Cell Biochem.* **90** (2003), pp. 1250–1260. [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [53](#) Y.X. Pan, Z.Z. Zhang, Z.M. Guo, G.Y. Feng, Z.D. Huang and L. He, Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach, *J. Protein Chem.* **22** (2003), pp. 395–402. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [54](#) R. Nair and B. Rost, Better prediction of sub-cellular localization by combining evolutionary and structural information, *Proteins* **53** (2003), pp. 917–930. [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [55](#) E.M. Marcotte, I. Xenarios, A.M. van Der Bliik and D. Eisenberg, Localizing proteins in the cell from their phylogenetic profiles, *Proc. Natl Acad Sci. USA* **97** (2000), pp. 12115–12120. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [56](#) A. Drawid and M. Gerstein, A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome, *J. Mol. Biol.* **301** (2000), pp. 1059–1075. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(538 K\)](#)
- [57](#) K. Nakai and M. Kanehisa, A knowledge base for predicting protein localization sites in eukaryotic cells, *Genomics* **14** (1992), pp. 897–911. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#)
- [58](#) K. Nakai and M. Kanehisa, Expert system for predicting protein localization sites in Gram-negative bacteria, *Proteins: Struct. Funct. Genet.* **11** (1991), pp. 95–110. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#)
- [59](#) M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler and J.M. Cherry *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nature Genet.* **25** (2000), pp. 25–29. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#)
- [60](#) S. Lewis, M. Ashburner and M.G. Reese, Annotating eukaryote genomes, *Curr. Opin. Struct. Biol.* **10** (2000), pp. 349–354. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(183 K\)](#)
- [61](#) B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher and E. Gasteiger *et al.*, The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003, *Nucl. Acids Res.* **31** (2003), pp. 365–370. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

- [62](#) B. Rost, R. Casadio and P. Fariselli, Topology prediction for helical transmembrane proteins at 86% accuracy, *Protein Sci.* **5** (1996), pp. 1704–1718. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#)
- [63](#) D.T. Jones, Do transmembrane protein superfolds exist?, *FEBS Letters* **423** (1998), pp. 281–285. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(126 K\)](#)
- [64](#) E. Wallin and G. von Heijne, Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms, *Protein Sci.* **7** (1998), pp. 1029–1038. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#)
- [65](#) K. Melen, A. Krogh and G. von Heijne, Reliability measures for membrane protein topology prediction algorithms, *J. Mol. Biol.* **327** (2003), pp. 735–744. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(291 K\)](#)
- [66](#) T. Joachims, Estimating the Generalization Performance of a SVM Efficiently, *Proceedings of the Seventeenth International Conference on Machine Learning 2000*, Morgan Kaufmann Publishers, Inc., San Francisco, CA (2000).
- [67](#) S. Hua and Z. Sun, Support vector machine approach for protein subcellular localization prediction, *Bioinformatics* **17** (2001), pp. 721–728. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [68](#) S.E. Brenner, Errors in genome annotation, *Trends Genet.* **15** (1999), pp. 132–133. [Abstract](#) | [PDF \(158 K\)](#)
- [69](#) E.V. Koonin, Computational genomics, *Curr. Biol.* **11** (2001), pp. R155–R158. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(136 K\)](#)
- [70](#) N.C. Kyrpides and C.A. Ouzounis, Errors in genome reviews, *Science* **281** (1998), p. 1457. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#)
- [71](#) T.J.P. Hubbard, New horizons in sequence analysis, *Curr. Opin. Struct. Biol.* **7** (1997), pp. 190–193. [Abstract](#) | [Abstract + References](#) | [PDF \(375 K\)](#)
- [72](#) S.A. Chervitz, E.T. Hester, C.A. Ball, K. Dolinski, S.S. Dwight and M.A. Harris *et al.*, Using the *Saccharomyces* Genome Database (SGD) for analysis of protein similarities and structure, *Nucl. Acids Res.* **27** (1999), pp. 74–78. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)
- [73](#) K.R. Christie, S. Weng, R. Balakrishnan, M.C. Costanzo, K. Dolinski and S.S. Dwight *et al.*, *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms, *Nucl. Acids Res.* **32** (2004), pp. D311–D314. [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#)

[74](#) E. Richly and D. Leister, An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice, *Gene* **329** (2004), pp. 11–16. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(194 K\)](#)

[75](#) Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, *Nature* **408** (2000), pp. 796–815.

[76](#) P. Horton and K. Nakai, A probabilistic classification system for predicting the cellular localization sites of proteins. In: D. States, P. Agarwal, T. Gaasterland, L. Hunter and R.F. Smith, Editors, *Fourth International Conference on Intelligent Systems for Molecular Biology, Fourth International Conference on Intelligent Systems for Molecular Biology, St Louis, MO, USA* vol. **1**, AAAI Press, St Louis, MO, USA (1996).

[77](#) M. Mann, R.C. Hendrickson and A. Pandey, Analysis of proteins and proteomes by mass spectrometry, *Annu. Rev. Biochem.* **70** (2001), pp. 437–473. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[78](#) Mika, S. & Rost, B. (2005). NMPdb: database of nuclear matrix proteins. *Nucl. Acids Res.*, **33**, D160–D163.

[79](#) Y.D. Cai and K.C. Chou, Using neural networks for prediction of subcellular location of prokaryotic and eukaryotic proteins, *Mol. Cell Biol. Res. Commun.* **4** (2000), pp. 172–173. [Abstract](#) | [Abstract + References](#) | [PDF \(29 K\)](#)

[80](#) J. Moult, K. Fidelis, A. Zemla and T. Hubbard, Critical assessment of methods of protein structure prediction (CASP)-round V, *Proteins: Struct. Funct. Genet.* **53** (2003), pp. 334–339. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[81](#) J. Moult, K. Fidelis, A. Zemla and T. Hubbard, Critical assessment of methods of protein structure prediction (CASP): round IV, *Proteins: Struct. Funct. Genet.* (2001), pp. 2–7. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[82](#) J. Moult, T. Hubbard, S.H. Bryant, K. Fidelis and J.T. Pedersen, Critical assessment of methods of protein structure prediction (CASP): round III, *Proteins: Struct. Funct. Genet.* (1999), pp. 2–6. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#)

[83](#) J. Moult, T. Hubbard, S.H. Bryant, K. Fidelis and J.T. Pedersen, Critical assessment of methods of protein structure prediction (CASP): round II, *Proteins: Struct. Funct. Genet.* **1** (1997), pp. 2–6. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[84](#) J. Moult, J.T. Pedersen, R. Judson and K. Fidelis, A large-scale experiment to assess protein structure prediction methods, *Proteins: Struct. Funct. Genet.* **23** (1995), pp. ii–iv.

[85](#) V. Eyrich, M.A. Martí-Renom, D. Przybylski, A. Fiser, F. Pazos and A. Valencia *et al.*, EVA: continuous automatic evaluation of protein structure prediction servers, *Bioinformatics* **17** (2001), pp. 1242–1243. [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[86](#) I.Y.Y. Koh, V.A. Eyrich, M.A. Marti-Renom, D. Przybylski, M.S. Madhusudhan and E. Narayanan *et al.*, EVA: evaluation of protein structure prediction servers, *Nucl. Acids Res.* **31** (2003), pp. 3311–3315. [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[87](#) A. Bairoch and R. Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucl. Acids Res.* **28** (2000), pp. 45–48. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[88](#) U. Hobohm, M. Scharf, R. Schneider and C. Sander, Selection of representative protein data sets, *Protein Sci.* **1** (1992), pp. 409–417. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#)

[89](#) H. Nielsen, J. Engelbrecht, G. von Heijne and S. Brunak, Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site, *Proteins: Struct. Funct. Genet.* **24** (1996), pp. 165–177. [Abstract-EMBASE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[90](#) B. Rost, Enzyme function less conserved than anticipated, *J. Mol. Biol.* **318** (2002), pp. 595–608. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(659 K\)](#)

[91](#) B. Rost, Twilight zone of protein sequence alignments, *Protein Eng.* **12** (1999), pp. 85–94. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

[92](#) C. Sander and R. Schneider, Database of homology-derived protein structures and the structural meaning of sequence alignment, *Proteins* **9** (1991), pp. 56–68. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#)

[93](#) C. Sander and R. Schneider, Database of homology-derived structures and the structural meaning of sequence alignment, *Proteins: Struct. Funct. Genet.* **9** (1991), pp. 56–68. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#)

[94](#) F.S. Mathews, The structure, function and evolution of cytochromes, *Prog. Biophys. Mol. Biol.* **45** (1985), pp. 1–56. [Abstract-INSPEC](#) | [Abstract-MEDLINE](#)

[95](#) B. Rost and C. Sander, Prediction of protein secondary structure at better than 70% accuracy, *J. Mol. Biol.* **232** (1993), pp. 584–599. [Abstract](#) | [PDF \(1053 K\)](#)

[96](#) P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen and H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, *Bioinformatics* **16** (2000), pp. 412–424. [Abstract-EMBASE](#) | [Full Text via CrossRef](#)

[97](#) P. Horton and K. Nakai, Better prediction of protein cellular localization sites with the k nearest neighbors classifier, *Ismb* **5** (1997), pp. 147–152. [Abstract-INSPEC](#) | [Abstract-MEDLINE](#)

[98](#) B. Rost, How to use protein 1D structure predicted by PROFphd. In: J.E. Walker, Editor, *The Proteomics Protocols Handbook*, Humana, Totowa, NJ (2005), pp. 879–908.

[99](#) B. Rost, PHD: predicting one-dimensional protein structure by profile based neural networks, *Methods Enzymol.* **266** (1996), pp. 525–539. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#)

[100](#) C.P. Chen, A. Kernytsky and B. Rost, Transmembrane helix predictions revisited, *Protein Sci.* **11** (2002), pp. 2774–2791. [Abstract-MEDLINE](#) | [Full Text via CrossRef](#)

Appendix. Supplementary data

Table 1_Mat: Comparison of LOctree to other publicly available predictors.*

Class ▼	Method Score →	Eukaryotic Non-plant					Plant	
		LOctree	TargetP ₁	SubLoc ₂	PSORT II ₃	NNPSL ₄	LOctree	TargetP ₁
Secretory Pathway	Acc	81	91	-	-	-	77	87
	Cov	80	70	-	-	-	79	73
	gAv	81	80	-	-	-	78	80
Ext	Acc	83	-	72	85	58	68	-
	Cov	81	-	51	34	61	68	-
	gAv	82	-	61	35	60	68	-
Nuc	Acc	78	-	69	55	69	70	-
	Cov	78	-	79	77	65	81	-
	gAv	78	-	74	66	67	75	-
Cyt	Acc	63	-	56	51	50	73	-
	Cov	66	-	66	48	43	74	-
	gAv	64	-	61	49	46	74	-
Mit	Acc	70	62	52	42	40	61	56
	Cov	67	64	64	51	70	74	39
	gAv	68	63	58	46	53	67	47

<i>Chloro</i>	<i>Acc</i>	-	-	-	-	-	77	83
	<i>Cov</i>	-	-	-	-	-	63	76
	<i>gAv</i>	-	-	-	-	-	70	79
<i>Overall Accuracy</i>		74		63	53	56	70	

* 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. *

<i>Method</i>		<i>LOCtree full length</i>	<i>LOCtree Rand</i>	<i>LOCtree Nterm</i>
<i>Sec Path</i>	<i>Acc</i>	87	79	59
	<i>Cov</i>	90	83	68
	<i>gAv</i>	88	81	63
<i>Ext</i>	<i>Acc</i>	86	72	55
	<i>Cov</i>	93	83	67
	<i>gAv</i>	89	77	61
<i>Nuc</i>	<i>Acc</i>	77	68	64
	<i>Cov</i>	85	80	66
	<i>gAv</i>	81	74	65
<i>Cyt</i>	<i>Acc</i>	82	72	54
	<i>Cov</i>	64	44	42
	<i>gAv</i>	72	56	48
<i>Mit</i>	<i>Acc</i>	73	54	40
	<i>Cov</i>	78	68	42
	<i>gAv</i>	75	61	41
<i>Overall Accuracy</i>		78	68	55

* 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. *

<i>Organism</i>	<i>Nprot</i>	<i>Secretory Pathway</i>			<i>Intra-cellular</i>		
		<i>Ext</i>	<i>Org</i>	<i>Nuc</i>	<i>Cyt</i>	<i>Mit</i>	<i>Chloro</i>
<i>Homo sapiens</i>	30371	20.9 (2)	3.5 (1)	40.8 (4)	24.3 (5)	10.5 (2)	0.0
<i>Saccharomyces cerevisiae</i>	5004	10.1 (1)	5.0 (2)	45.4 (5)	22.9 (5)	16.6 (4)	0.0
<i>Arabidopsis thaliana</i>	21085	7.8 (1)	5.9 (2)	30.0 (3)	17.2 (3)	10.7 (3)	28.4 (5)

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. *

<i>Sub-cellular localization</i>	<i>Eukaryotes</i>				<i>Prokaryotes</i>	
	<i>SWISS-PROT annotated</i>	<i>Non-plant unique</i>	<i>Plant unique</i>	<i>Non-plant new unique</i>	<i>SWISS-PROT annotated</i>	<i>Unique</i>
Nucleus	2673	562	32	168		
Cytosol	2137	330	77	117	12124	426
Extra-cellular	1936	363	22	121	449	117
Chloroplast	952	-	103	-		
Mitochondria	914	198	50	51		
Periplasmic Organelles	-	-	-	-	607	129
SUM	8979	1505	304	490	13180	672

* 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'

1. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* 300, 1005-16.
 2. Hua, S. & Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 17, 721-8.
 3. Nakai, K. & Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* 24, 34-6.
 4. Reinhardt, A. & Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res* 26, 2230-6.
-