BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 9 2010, pages 1232–1238
doi:10.1093/bioinformatics/btq115

Systems biology

Advance Access publication March 17, 2010

Going from where to why—interpretable prediction of protein subcellular localization

Sebastian Briesemeister[1,*], Jörg Rahnenführer[2] and Oliver Kohlbacher[1]

[1]Division for Simulation of Biological Systems, Universität Tübingen, Tübingen and [2]Department of Statistics, TU Dortmund University, Dortmund, Germany

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Protein subcellular localization is pivotal in understanding a protein's function. Computational prediction of subcellular localization has become a viable alternative to experimental approaches. While current machine learning-based methods yield good prediction accuracy, most of them suffer from two key problems: lack of interpretability and dealing with multiple locations.

Results: We present YLoc, a novel method for predicting protein subcellular localization that addresses these issues. Due to its simple architecture, YLoc can identify the relevant features of a protein sequence contributing to its subcellular localization, e.g. localization signals or motifs relevant to protein sorting. We present several example applications where YLoc identifies the sequence features responsible for protein localization, and thus reveals not only to which location a protein is transported to, but also why it is transported there. YLoc also provides a confidence estimate for the prediction. Thus, the user can decide what level of error is acceptable for a prediction. Due to a probabilistic approach and the use of several thousands of dual-targeted proteins, YLoc is able to predict multiple locations per protein. YLoc was benchmarked using several independent datasets for protein subcellular localization and performs on par with other state-of-the-art predictors. Disregarding low-confidence predictions, YLoc can achieve prediction accuracies of over 90%. Moreover, we show that YLoc is able to reliably predict multiple locations and outperforms the best predictors in this area.

Availability: www.multiloc.org/YLoc

Contact: briese@informatik.uni-tuebingen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 2, 2009; revised on February 25, 2010; accepted on March 14, 2010

1 INTRODUCTION

Subcellular protein localization is a key process in many eukaryotic cells, and hence a major research topic in biology. After being synthesized, proteins are transported into different compartments depending on their molecular role within the cell. Some proteins are even transported to multiple sites. Protein localization is often mediated by sorting signals or sorting patches. However, the process of protein sorting is not fully understood yet. The subcellular localization of a protein is highly correlated with its function and is thus used to draw conclusions about its cellular role, interaction partners and function in biological processes. During the past decade, huge number of novel proteins were discovered in the context of large-scale sequencing projects. Unfortunately, for a majority of these proteins their subcellular localization is unknown and experimentally determining the localization of a protein is expensive and time-consuming.

Computational prediction methods that predict subcellular localization from the amino acid sequence represent an attractive alternative to experimental methods. Over the past few years, numerous prediction methods have been developed. We distinguish between sequence- and annotation-based methods. Sequence-based predictors make use of sequence-coded sorting signals (Bannai *et al.*, 2002; Boden and Hawkins, 2005; Cokol *et al.*, 2000; Emanuelsson *et al.*, 2007; Fujiwara and Asogawa, 2001; Petsalaki *et al.*, 2006; Small *et al.*, 2004), amino acid composition information (Cedano *et al.*, 1997; Chou and Cai, 2003b; Cui *et al.*, 2004; Guo and Lin, 2006; Hua and Sun, 2001; King and Guda, 2007; Nair and Rost, 2005; Park and Kanehisa, 2003; Pierleoni *et al.*, 2006; Reinhardt and Hubbard, 1998; Xie *et al.*, 2005) or even both information sources (Garg *et al.*, 2009; Höglund *et al.*, 2006; Horton *et al.*, 2007). Annotation-based predictors use information about functional domains and motifs (Chou and Cai, 2002; Scott *et al.*, 2004), protein–protein interaction (Lee *et al.*, 2009; Shin *et al.*, 2009), homologous proteins (Garg and Raghava, 2008; Lin *et al.*, 2009), annotated Gene Ontology (GO) terms (Huang *et al.*, 2008; Lei and Dai, 2006; Lu and Hunter, 2005) and textual information from Swiss-Prot keywords (Lu *et al.*, 2004; Nair and Rost, 2002a) or PubMed abstracts (Brady and Shatkay, 2008; Fyshe *et al.*, 2008). Since proteins with sufficiently similar protein sequences are usually located in the same compartment (Nair and Rost, 2002b), missing annotation information might also be transferred from close homologues. Annotation-based predictors often show higher accuracies than pure sequence-based predictors although they are less robust for novel proteins without known close homologues. Hybrid prediction approaches (Blum *et al.*, 2009; Briesemeister *et al.*, 2009; Chou and Cai, 2003a; Chou *et al.*, 2007; Scott *et al.*, 2005) take advantage of both information sources.

Although there is an evidence that more than one-third of all eukaryotic proteins are transported to multiple compartments (Zhang *et al.*, 2008), multiple targeting of proteins has only rarely been considered by prediction methods. As one of the first groups, Scott *et al.* (2004) introduced a method for multiple localization prediction based on about 500 multiple localized proteins. More recent predictors such as WoLF PSORT (Horton *et al.*, 2007),

*To whom correspondence should be addressed.

Euk-mPloc (Chou *et al.*, 2007), ngLoc (King and Guda, 2007) and KnowPred (Lin *et al.*, 2009) use even up to 2200 multiply targeted proteins for their predictions. Although there has been recent development on multiple localization prediction, we believe that there is still room for improvement.

State-of-the-art methods show high prediction performance that has significantly improved over the years. Unfortunately, the machine learning models behind high-accuracy predictors are often very complex making it difficult to understand why a particular prediction was made. Moreover, most predictors do not provide a confidence estimate. Consequently, predictions cannot be verified with regard to their significance and reliability.

In this work, we present YLoc, a novel method for predicting subcellular localization of proteins. YLoc is based on the simple naive Bayes classifier. It combines various feature types for its predictions ranging from simple amino acid composition to annotation information like PROSITE domains and GO terms from close homologues. Most importantly, it uses at most 30 of these features. The small number of features as well as the simple architecture guarantee interpretable predictions. YLoc is able to elucidate why a prediction was made and what attributes of the protein contributed most to this prediction. In addition, it returns confidence estimates that rate predictions as reliable or not. YLoc is available in three versions. The low-resolution version, YLoc-LowRes, is specialized in distinguishing the localization of globular proteins and predicts up to five locations. The high-resolution version, YLoc-HighRes, covers 11 main eukaryotic subcellular locations. YLoc$^+$ is the most general predictor. It covers 11 main eukaryotic locations while integrating multiple localization sites. All three predictors are available for animal, fungal and plant proteins.

We compared YLoc against other state-of-the-art protein subcellular localization predictors using two recently published independent datasets (IDS; Blum *et al.*, 2009; Casadio *et al.*, 2008). The results confirm that YLoc, even though its architecture is very simple, performs comparably to current state-of-the-art predictors. For instances predicted with high confidence, YLoc yields an even better prediction performance. For proteins with multiple localizations, YLoc shows an outstanding accuracy compared to existing methods. In an example prediction, we show that YLoc prediction outputs can be easily interpreted. Moreover, we illustrate that YLoc can be applied to explain localization changes caused by mutations in experiments.

2 METHODS

2.1 Features

In the past, various types of features were studied in the context of subcellular localization. However, in many cases only one or two of feature types were included in one predictor. In our study, we included numerous types of features and properties.

First, we make use of sequence-derived features. These include amino acid composition, normalized amino acid composition and pseudo-amino acid composition (Chou, 2001). In addition to counting simple amino acids, we use the compositions of certain amino acid types such as hydrophobic, positively charged, negatively charged, aromatic and small. Moreover, we calculate sum and autocorrelation of properties such as hydrophobicity, charge, and volume of the amino acids. The autocorrelation measures the correlation of a signal with itself and can be used to identify periodic patterns. For a given distance j, we calculate $\sum_n x_n x_{n-j}$ and normalize it by the length of the sequence. All features are calculated over the whole sequence length, as well as for subsequences of various lengths in the N-terminus (10–200), C-terminus (10–100) and middle part of the protein. In all cases, we omit the first residue to avoid the bias caused by methionine. In addition, various known sorting signals such as mono nuclear localization signal (NLS), bipartite NLS, nuclear export signal (NES), peroxisomal targeting signal, mitochondrial targeting signal, chloroplast targeting signal, secretory pathways (SPs) signal and endoplasmatic reticulum retention signal are considered.

Second, we make use of annotation-based features such as PROSITE patterns and GO terms. PROSITE patterns describe protein domains, families, as well as functional sites. A PROSITE pattern feature is assigned a value of one if the pattern is found in the protein sequence using PROSITE scan. In addition, we calculate a feature for each location which is defined by PROSITE patterns that are typical for this location. A PROSITE pattern is typical for a location if >80% of all proteins in the training dataset containing this pattern are present in this particular location. The resulting feature is assigned a value of one if at least one typical PROSITE pattern of this location is present in the protein or zero otherwise. Finally, we use GO terms from close homologues from Swiss-Prot release 42.0. A GO-term feature equals one if at least one protein that locally aligns with the query sequence with a maximal E-value 10^{-10} and a sequence identity of >30% is annotated with this GO term. Using these alignment conditions, we are able to transfer GO terms from known proteins that share domains with the query protein. Similarly to PROSITE patterns, we create a feature for each location that indicates whether the protein is likely to be annotated with a GO term that is typical for this location. A GO term is typical for a location if >95% of all proteins containing this GO term are located there. We use a higher threshold due to the fact that GO terms naturally contain more noise since they were inferred from sequences that do not neccessarily have to be orthologues, or even homologues. An additional feature indicates the location for which the most typical GO terms could be transfered. The overall number of features is about 30000.

2.2 Feature selection

Because of the limited number of learning examples, learning with a small number of features often leads to a better generalization of machine learning algorithms (Occam's razor). Moreover, interpreting predictions is possible if the number of features is very small. Since we could not observe a significant improvement in a nested cross-validation of our method (data not shown), we decided that 30 features are sufficient for our predictors. For YLoc-LowRes, even 20 features are sufficient due to the reduced number of locations.

To find the set of the most important features, we started a large-scale feature selection using a correlation-based feature selection (CFS) approach (Hall, 2000). It favors a feature set that shows high correlation with the class variable but low redundancy among the features in the set. The following heuristic expresses the merit of a feature subset I of size k

$$\frac{\sum_{i\in I} r_i}{\sqrt{k+\sum_{i,j\in I} r_{ij}}}, \tag{1}$$

where r_i is the correlation between feature i and the class variable and r_{ij} the correlation between two features i,j in the subset. Both correlations are calculated using the symmetric uncertainty coefficient after the data were discretized. To avoid large feature subsets, we assign a value of zero if k is >30.

We use CFS together with a backward best-first search, which continually catches the best 100 subsets and terminates after 50 backtracking steps The search algorithm as well as CFS are implemented in the Weka machine learning library (Whitten and Frank, 2005). The average running time for the feature selection was ∼2 h (data not shown).

All selected features are manually described in biological terms. For some features, a biological explanation is not obvious. In these cases, we transferred the biological meaning from a strongly correlated feature.

2.3 Naive Bayes classification

YLoc uses naive Bayes, a very simple and robust classification model, to make predictions. It assumes features to be independent and, thus, allows a straightforward decomposition of a prediction into the individual contributions of each feature. It has been shown that naive Bayes is still surprisingly effective in cases where the independency assumption is violated (Rish, 2001). Given a set of features $F=\{F_1,\ldots,F_n\}$, a set of locations $L=\{L_1,\ldots,L_k\}$ and a set of corresponding classes $C=\{C_{L_1},\ldots,C_{L_k}\}$, it estimates the posterior probability by

$$P(C_{L_k}|F)\propto P(C_{L_k})\prod_{i=1}^{n}P(F_i|C_{L_k}). \quad (2)$$

The class priors and the feature probability distributions are estimated using previous discretized training data. For our purposes, we use the entropy-based supervised discretization (Fayyad and Irani, 1993). The final probabilities are obtained by normalizing the posteriori such that the sum of all posteriori is one.

Since features are treated independently, we can easily assess the influence of a single feature F_i on the prediction. The probability of observing feature F_i ranges from $\min_k P(F_i|C_k)$ to $\max_k P(F_i|C_k)$ over the given classes C_k. Let $C_{\max}=\arg\max_{C_k}P(C_k|F)$ be the predicted class. We define

$$\log\frac{P(F_i|C_{\max})}{\min_k P(F_i|C_k)} \quad \text{and} \quad \log\frac{P(F_i|C_{\max})}{\max_k P(F_i|C_k)} \quad (3)$$

to be the support and the opposition score, respectively. A large support score originates from a high probability for the observed feature value in the predicted class, compared to the class where this feature value is least likely. In contrast, the opposition score is always negative. Given a very low opposition score, it is more likely to observe this feature in a class that was not predicted. Hence, a prediction based on the feature alone would lead to a different decision than using all features. We merge both values in the discrimination score (DS). If the support for $C_{\max}$ is stronger than the opposition, i.e. the sum of the scores is >0, the DS equals the support score and vice versa. We use the absolute value of the DS to order the features according to their influence on the prediction.

To predict multiple localizations with YLoc$^+$, we transform our multi-label data into single-label data. For proteins labeled with multiple locations L_i and L_k, we create a new class, $C_{L_i\wedge L_k}$. When inferring predictions, the probability output of the naive Bayes classifier is transformed as follows:

$$P(L_i|F)=\sum_{\{C_x|C_x\in C\wedge L_i\in\alpha(C_x)\}}P(C_x|F)\frac{1}{|\alpha(C_x)|}, \quad (4)$$

where $\alpha(C_x)$ is the set of labels of class C_x. This transformation is based on the assumption that proteins present in multiple locations are equally concentrated in these compartments. Obviously, this does not hold for all proteins with multiple localizations. However, given only qualitative data, this is the best assumption we can make. To report only relevant locations, YLoc employs a simple heuristic. After sorting the locations by probability, YLoc reports the locations with probability better than chance, i.e. $P(L_i|F)>1/|L|$, where L is the set of locations. To report only relevant locations with reasonable probability, YLoc stops reporting locations if a location is less than half as probable as the preceding location. Transforming the probabilities as above yields the advantage that label combinations not present in the training data can also be predicted.

2.4 Confidence estimates

To provide users with an estimate of how reliable a prediction is, YLoc computes confidence estimates. The estimate is based on the fact that proteins can be predicted more reliably if the corresponding feature vector is very typical for the predicted classes and less typical for any other class. Given the feature vector F of a protein, we calculate $P(F|\bigcup_{C_i\in C}C_i)$, the probability of observing F, given our training dataset. On the other hand, we calculate $P(F|C_{\max})$, the probability of F, given the most probable class $C_{\max}$. Since F should be more typical for the predicted class $C_{\max}$ than for the set of all proteins, $P(F|C_{\max})$ should be greater than $P(F|\bigcup_{C_i\in C}C_i)$, the baseline probability of observing F. For our final confidence score, we calculate the fraction of both probabilities and additionally weight classes with few training examples as less reliable by multiplying the class probability $P(C_{\max})$. The final confidence score is calculated as follows:

$$\text{conf}=\frac{P(C_{\max})P(F|C_{\max})}{P(C_{\max})P(F|C_{\max})+P(F|\bigcup_{C_i\in C}C_i)}. \quad (5)$$

A confidence score close to one indicates a reliable prediction, whereas a score close to zero indicates that YLoc is less confident about the given prediction. Note that if we assume $P(F|\bigcup_{C_i\in C}C_i)=P(F)$, the presented confidence score would be a monotone transformation of $P(C_{\max}|F)$, given by $\text{conf}=1/(1+\frac{1}{P(C_{\max}|F)})$.

2.5 Datasets

2.5.1 BaCelLo For training the YLoc-LowRes predictor, we used the BaCelLo training dataset (Pierleoni *et al.*, 2006). The homology reduced dataset extracted from Swiss-Prot release 48 contains 2597 animal, 1198 fungal and 491 plant proteins, resulting in three versions of YLoc-LowRes. Only globular proteins were considered in the annotation. Animal and fungal proteins originate from four locations: nucleus (nu), cytoplasm (cy), mitochondrion (mi) and the SP. Plant proteins originate from five locations: nu, cy, mi, SP and chloroplast (ch). The BaCelLo IDS (Casadio *et al.*, 2008) contains proteins added to Swiss-Prot between release 49 and 54 with at most 30% sequence identity to proteins in the BaCelLo dataset. Moreover, proteins from the same location that align with an E-value $<10^{-3}$ using BLAST are clustered, resulting in 432 animal, 418 fungi and 132 plant groups.

2.5.2 Höglund For training YLoc-HighRes and YLoc$^+$, we used the Höglund training dataset (Höglund *et al.*, 2006). The 5959 eukaryotic proteins extracted from Swiss-Prot release 42 covering 11 locations: nu, cy, mi, ch, endoplasmic reticulum (er), golgi apparatus (go), peroxisome (pe), plasma membrane (pm), extracellular space (ex), lysosome (ly) and vacuole (va). The Höglund IDS was constructed with proteins from Swiss-Prot release 55.3 and covers the locations er, go, pe, pm, ex, ly and va. Proteins that share >30% sequence identity with proteins from the original Höglund dataset were excluded. In this study, we only make use of the animal Höglund IDS, since it contains sufficient amount of proteins (198). By clustering proteins from the same location with >40% sequence identity, 158 animal groups were obtained.

2.5.3 DBMLoc In addition to proteins from the Höglund dataset, YLoc$^+$ was trained using proteins from the DBMLoc database (Zhang *et al.*, 2008). The DBMLoc database contains >10 000 proteins with multiple subcellular localization, which were experimentally determined or extracted from the literature. We extracted proteins that share <80% sequence similarity with each other from DBMLoc. Most proteins in DBMLoc are present in two subcellular locations. Still, there is a small portion of proteins with three or more localizations. However, for training we selected only multiple locations with >100 representative proteins: cy and nu (cy_nu), ex and pm (ex_pm), cy and pm (cy_pm), cy and mi (cy_mi), nu and mit (nu_mi), er and ex (er_ex), and ex and nu (ex_nu). Due to the limited number of training examples for some localizations, we could not use a lower sequence similarity threshold. More details concerning the 3054 proteins with multiple localization can be found in the Supplementary Material.

2.6 Training and evaluation

We implemented YLoc using Python, the machine learning library Weka (Whitten and Frank, 2005), BLAST and PROSITE scan. Each YLoc predictor is available as an animal, fungi or plant version.

To evaluate the prediction performance, we use the overall accuracy (ACC), the percentage of correctly predicted instances and the average

F_1-score (F_1), which is the harmonic mean of recall (REC) and precision (PRE), defined as follows:

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad \mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{6}$$

$$\mathrm{F1} = \frac{2\ \mathrm{REC}\ \mathrm{PRE}}{\mathrm{REC}+\mathrm{PRE}}. \tag{7}$$

We think that the F_1 is better suited than the ACC as an evaluation measure. Especially for unbalanced datasets, the ACC biases towards an overrepresented class. Thus, if all instances are predicted to belong to this class, the ACC is still rather high.

The ACC and F_1 can be easily generalized using measures from multi-label classification (Tsoumakas and Katakis, 2007). Let D denote a dataset with n instances. Further, let Y_i and Z_i be the set of correct labels and the set of predicted labels of instance $i \in D$, respectively. Consequently, we can define the ACC, REC and PRE for label k as follows:

$$\mathrm{ACC} = \sum_{\{i|i\in D\}} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{8}$$

$$\mathrm{REC}_k = \sum_{\{i|i\in D \wedge k\in Y_i\}} \frac{|Y_i \cap Z_i|}{|Y_i|} \tag{9}$$

$$\mathrm{PRE}_k = \sum_{\{i|i\in D \wedge k\in Z_i\}} \frac{|Y_i \cap Z_i|}{|Z_i|}. \tag{10}$$

Using multi-label measures, we can rate predictions as 'half-right' when only a portion of the correct labels were recovered or more labels than the correct ones were predicted.

3 DISCUSSION

3.1 Benchmark study using two IDSs

To show that YLoc is well-suited to predict the localization of novel proteins, we carried out a benchmark study using two recently published IDSs, the BaCelLo IDS (Casadio *et al.*, 2008) and the Höglund IDS (Blum *et al.*, 2009). We compared YLoc against six other state-of-the-art subcellular localization predictors, MultiLoc2 (Blum *et al.*, 2009), BaCelLo (Pierleoni *et al.*, 2006), LOCTree (Nair and Rost, 2005), WoLF PSORT (Horton *et al.*, 2007), Euk-mPloc (Chou *et al.*, 2007) and KnowPred (Lin *et al.*, 2009). These predictors were chosen because they are quite recent and are available as online or as stand-alone version. In the case of the BaCelLo IDS, we grouped predicted locations from the secretory pathway into the class SP to deal with predictors that distinguish between these locations. In contrast, for the Höglund IDS we excluded predictors that cannot distinguish between the secretory pathway locations. To predict multiple locations with KnowPred, we defined a threshold of 30 for the multi-localized confidence score (see Supplementary Material). As mentioned before, very similar proteins from the same location in the IDS are clusterd. Instead of evaluating the performance based on one representative of each cluster, we re-weight instances such that the weight of all instances within one cluster sums to one. The results are summarized in Table 1.

We observed that YLoc-LowRes and MultiLoc2-LowRes yield the best overall performance on the BaCelLO IDS. This is due to the fact that both predictors are specialized in distinguishing globular proteins. Among the high-resolution predictors, MultiLoc2-HighRes and KnowPred perform best, followed by YLoc-HighRes. Although YLoc$^+$ was designed to predict multiple localizations, it performs comparably to Euk-mPloc and WoLF PSORT. Clearly, the prediction performance depends on the origin of the proteins. In particular, the YLoc predictors are less accurate for fungal proteins, but yield good performance for animal and plant proteins. In contrast, Euk-mPloc performs well for fungal proteins but poorly for animal and plant proteins. Note that KnowPred does not predict chloroplasts and, thus, performs poorly on plant proteins. Most interestingly, the YLoc predictors perform comparably to the other predictors in the benchmark study, even though they have a very simple architecture and use at most 30 features. Similar results were observed for the animal Höglund IDS. MultiLoc2-HighRes performs best among the high-resolution predictors, followed by YLoc$^+$, YLoc-HighRes and KnowPred. Euk-mPloc and WoLF PSORT, the other high-resolution predictors in this study, yield a poor F1 and ACC. In general, the performance of all predictors is comparably low for this dataset. This is due to the limited amount of available training data for the peroxisome and the secretory pathway locations. Since the number of protein sequences of the animal Höglund IDS is comparably low, the performance results should be seen as a trend.

Using YLoc$^+$ has an advantage. Predictions can be borderline due to weak and noisy sorting signals. Hence, predicting all top-ranked locations leads to an increased recall. Moreover, it can help users to identify real multiple localization of proteins.

We also tested YLoc without transferring information from homologous proteins by excluding GO-term features from the feature selection. The resulting predictors show only slightly reduced prediction performance on the IDSs (see Supplementary Material). Additional predictors not using homology information can be helpful to analyze whether a prediction outcome would change if we were restricted to sequence information only.

Table 1. Performance comparison using two IDSs

Dataset	YLoc-LowRes	YLoc-HighRes	YLoc$^+$	MultiLoc2-LowRes	MultiLoc2-HighRes	BaCelLo	LOCTree	WoLF PSORT	Euk-mPloc	KnowPred
B Animals	0.75 (**0.79**)	0.69 (0.74)	0.67 (0.58)	**0.76** (0.73)	0.71 (0.68)	0.66 (0.64)	0.58 (0.62)	0.67 (0.70)	0.54 (0.61)	0.69 (0.75)
B Fungi	**0.61** (0.56)	0.51 (0.56)	0.51 (0.48)	**0.61** (0.60)	0.58 (0.53)	0.60 (0.57)	0.43 (0.47)	0.51 (0.50)	0.56 (0.60)	0.56 (**0.66**)
B Plants	0.58 (0.71)	0.54 (0.58)	0.49 (0.53)	**0.64** (**0.76**)	0.54 (0.62)	0.56 (0.69)	0.58 (0.70)	0.46 (0.57)	0.37 (0.46)	0.23 (0.29)
H Animals	– (–)	0.34 (0.56)	0.37 (0.53)	– (–)	**0.41** (**0.57**)	– (–)	– (–)	0.18 (0.36)	0.24 (0.27)	0.37 (0.49)

Performance of the YLoc predictors and other state-of-the-art predictors using the Bacello (B) IDS and the Höglund (H) IDS concerning F1 and ACC (in brackets). The performance of YLoc$^+$, WoLF PSORT, Euk-mPloc and KnowPred was measured using the generalized F1 and ACC. The highest-ranking method regarding each measure is highlighted in bold. Note that the WoLF PSORT results differ slightly from those obtained in Blum *et al.* (2009) due to some changes in the underlying dataset. Also note that KnowPred does not predict chloroplasts.

Table 2. Performance of YLoc using the BaCelLo animal IDS for different minimum confidence levels

Predictor	Measure	0.00	0.20	0.40	0.60	0.80	0.90
YLoc-LowRes	F1	0.75	0.76	0.78	0.80	0.84	0.95
	ACC	0.79	0.79	0.81	0.86	0.91	0.93
	No. Inst.	576	467	395	299	189	118
YLoc-HighRes	F1	0.69	0.74	0.76	0.76	0.77	0.77
	ACC	0.74	0.78	0.80	0.82	0.83	0.84
	No. Inst.	576	507	470	428	391	354
YLoc$^+$	F1	0.67	0.69	0.72	0.77	0.76	0.81
	ACC	0.58	0.60	0.62	0.65	0.65	0.69
	No. Inst.	576	494	423	324	219	142

For each minimum confidence score the prediction performance is given using F1 and ACC as well as the number of instances that can be predicted with at least this score. The performance of YLoc$^+$ was measured using the generalized F1 and ACC.

3.2 Confidence estimates

To prove that YLoc highly benefits from confidence scores, we analyzed the influence of the confidence score on the prediction performance. Following our benchmark study from above, we analyzed the performance of YLoc by considering only proteins that could be predicted with a given minimum confidence score. We excluded classes that had less than five instances left. The performance of YLoc on the animal BaCelLo IDS for different minimum confidence scores is given in Table 2. The ACC and F1 of all predictors increase with an increasing minimum confidence score. The F1 and ACC of YLoc-HighRes increase by at least 4% given a minimum score of 0.2 and by at least 8% given a confidence threshold of 0.90. YLoc-LowRes and YLoc$^+$ show an even higher enrichment for high confidence scores. For example, YLoc-LowRes achieves an F1 of 0.84 and an ACC of 0.91 for a minimum confidence score 0.8. Thus, YLoc-LowRes could correctly predict the location for 91% of the 189 proteins, which have a confidence score of at least 0.8. We got similar results for fungi and plant proteins (see Supplementary Material). Although only a certain portion of proteins can be predicted with high confidence, their predicted locations are much more likely to be correct.

3.3 Evaluation of multiple-localization prediction

In a last benchmark study, we compared YLoc$^+$, WoLF PSORT, Euk-mPloc and KnowPred regarding their ability to predict multiple localization sites. The locations for all proteins in the DBMLoc dataset were predicted by WoLF PSORT, Euk-mPloc and KnowPred by considering this dataset as an IDS. For YLoc$^+$, we evaluated the predictions of the DBMLoc proteins using the 5-fold nested cross-validation results. We compared all predictors using single-label as well as multi-label measures. The results are shown in Table 3. YLoc$^+$ is superior to WoLF PSORT and Euk-mPloc in this study in terms of ACC as well as F1. While predicting at least one location correctly for many proteins, Euk-mPloc and WoLF PSORT are only able to predict 5% of the correct multiple locations. In contrast, YLoc$^+$ and KnowPred are able to recover more than one-third of the multiple locations correctly. In a similar study, we are able to show that the performance of all predictors remains almost unchanged if we use a cutoff of 40% in the homology reduction of the DBMLoc dataset. For more details see Supplementary Materials.

Table 3. Performance comparison using the DBMLoc dataset

Measures	YLoc$^+$	Euk-mPloc	WoLF PSORT	KnowPred
Single-label	**0.31** (0.35)	0.04 (0.05)	0.03 (0.05)	0.28 (**0.36**)
Multi-label	**0.68** (**0.64**)	0.44 (0.41)	0.52 (0.43)	0.66 (0.63)

The performance was measures using F1 and ACC (in brackets). For YLoc$^+$ and WoLF PSORT, only the best-performing version is shown. The highest-ranking method regarding each measure is highlighted in bold.

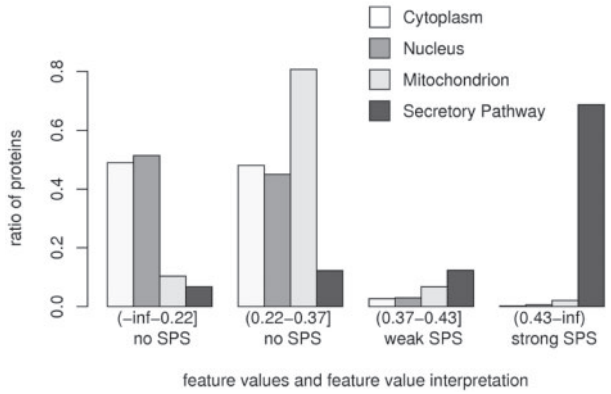

Fig. 1. The distribution of proteins regarding the secretory pathway signal (SPS) feature of YLoc-LowRes (animal version) is shown. For every discretization interval, the interval borders and an interpretation is given.

3.4 Understanding predictions

To show how YLoc elucidates a subcellular localization prediction, we provide an interpretable example prediction output. The example protein *Neurotoxin magi-12* (U13-HXTX) with Swiss-Prot AC Q75WG7, obtained from the animal BaCelLo IDS, was predicted to be located in the SP by YLoc-LowRes with a probability 99.99% and a confidence score 0.99. Hence, users can be very confident that the prediction is correct. U13-HXTX is known to be secreted into the extracellular space. YLoc found that U13-HXTX contains a strong secretory pathway signal, which is known to mediate the transport into the SP. Moreover, YLoc identified this feature to be the most discriminating, since 69% of all proteins in the SP have a similar secretory pathway signal, whereas only 0%, 2% and 1% of all proteins present in the cy, mi and nu, respectively, have the same kind of feature. Figure 1 shows the distribution of proteins from different locations concerning this particular feature. In addition, YLoc identified other features that highly influenced the prediction, such as the low charge of the protein and the lack of a mono-nuclear localization signal. Table 4 shows an example output of YLoc for the six most discriminating attributes. Given this output, it is easy to understand why this prediction was made and what features were responsible for it.

3.5 Understanding localization changes

A key step in understanding the localization process of proteins is to elucidate why proteins localize to different compartments when undergoing mutation. In the following, we show some examples where YLoc could have been helpful to understand the underlying localization processes.

Table 4. YLoc output of an example prediction

Sequence feature	DS	Nu	Cy	Mi	SP
Strong secretory pathway sorting signal (high hydrophobic autocorrelation within first 20 amino acids)	5.72	0.01	0.00	0.02	0.69
Barely charged (low overall charge autocorrelation)	2.89	0.10	0.16	0.02	0.28
No mono NLS sorting signal	2.89	0.04	0.12	0.02	0.26
Strong putative mitochondrial or secretory pathway sorting signal (large weighted sum of amino acids typical for mi and SP (Nakai and Kanehisa, 1992))	1.68	0.58	0.62	0.16	0.84
Very hydrophobic protein [high pseudo-amino acid count of hydrophobic amino acids (CITVWY)]	2.32	0.08	0.13	0.04	0.36
Very hydrophonic N-terminus (high pseudo-amino acid count of very hydrophobic residues within the first 90 amino acids)	2.06	0.09	0.05	0.08	0.41

The six most discriminating protein features are displayed in order of their absolute DS. The features are manually annotated with a biological property. A more detailed description of each feature is given in italics. For each location the ratio of proteins having this particular feature is shown.

Takada *et al.* (1990) showed that human glyoxylate aminotransferase 1 (AGT1), located in the peroxisome, is likely to have lost its mitochondrial targeting peptide (mTP) by point mutation. In fact, the mTP of AGT1 of rat, located in the mitochondrion, shares 74% sequence identity with the upstream region of human AGT1. If we corrected the single point mutation, we would extend human AGT1 by 22 residues. YLoc-HighRes (animal version) is able to predict a localization shift from the peroxisome to the mitochondrion. In addition, it recognizes the appearance of an weak mTP. According to YLoc$^+$, the extended AGT1 is very likely in the mitochondrion.

In 1982, Carlson and Botstein (1982) found two isoforms of glycosylated invertase in yeast, which is encoded by the SUC2 gene. The extracellular isoform is regulated by glucose repression, whereas the N-terminal truncated cytosolic isoform is constitutively expressed. YLoc-LowRes (fungi version) is able to predict the localization change of this truncation, although it still recognizes associated GO terms that indicate a secreted localization. In addition, the truncation of the signal peptide was recognized by YLoc. Four years later, Kaiser and Botstein (1986) examined the signal peptide of the same protein by inducing multiple mutations in the signal peptide region ranging from short deletions up to long substitutions. Five of the the 10 functional mutants lack extracellular invertase activity and show only cytoplasmic activity. Three of these cases could be validated by YLoc-LowRes. In one case, YLoc predicted a localization change, but not to the cytoplasm. In all five cases, YLoc confirms the loss of a signal peptide. In addition, YLoc reproduced the residual of the five remaining mutants in the secretory pathway.

The GLR1 gene of yeast encodes two different isoforms of glutathione reductase: a longer, mitochondrial isoform and a shorter, cytoplasmic isoform (Outten and Culotta, 2004). The two different isoforms very likely arise from leaky ribosomal scanning. YLoc-LowRes (fungi version) predicted GLR1 as mitochondrial and identified an mTP within the first 20 amino acids. The truncated isoform is still predicted to be located in the mitochondrion but with a decreased probability. Moreover, YLoc observed the loss of the mTP. Both YLoc-HighRes and YLoc$^+$ reproduced the location shift and observed a change in mTP.

4 CONCLUSION

Understanding protein subcellular localization is crucial for functional annotation of proteins. In contrast to many prediction methods, predictions made by YLoc are highly interpretable. YLoc explains why a prediction was made and shows which particular attributes contributed most and in which direction. Explaining why a subcellular localization prediction was made clearly influences the trust in the results. A user might find a prediction reasonable but might also find attributes indicating a different localization that are more convincing to him. In addition, a users can identify properties of their proteins that are typical or atypical for a certain cell organelle. Thus, YLoc can be helpful to understand the localization of novel proteins that have not been annotated before.

Our benchmark results suggest that using complex computational models is less important than using highly discriminating features with different biological background. When considering only proteins that can be predicted with a certain confidence score, the prediction performance increases considerably. We believe that a confidence estimate is of great interest since it increases the trust in prediction results. When predicting proteins from multiple locations, YLoc yields often better prediction quality than current state-of-the-art predictors. Moreover, YLoc's flexible probability transformation allows predicting novel location combinations that are not part of the training data.

We showed several examples where YLoc predicted experimentally validated changes of localization sites and known sorting signals caused by mutations. This is a key step toward understanding subcellular localization processes without conducting expensive experiments. However, single amino acid substitutions that will not change important physiochemical properties are not likely to cause a change of the predicted location.

In the future, we hope to increase both performance and interpretability of YLoc by integrating further biologically relevant features. Improvement will rely on traditional biology and computational biology proceeding hand in hand. Discovering novel protein sorting signals can improve the performance of YLoc, whereas an improved predictor can help biologists to elucidate the localization of novel proteins. Since we applied YLoc successfully on proteins with alternative isoforms that differ in localization, it seems promising to include alternative transcription and translation sites as features for YLoc$^+$. In addition, qualitative distribution data of multiply targeted proteins will help to improve the prediction quality of YLoc$^+$. The YLoc web service is available at www.multiloc.org/YLoc.

ACKNOWLEDGEMENT

The authors thank Nico Pfeifer for comments on the manuscript.

Funding: S.B. gratefully acknowledges financial support from LGFG Promotionsverbund 'Pflanzliche Sensorhistidinkinasen' of the Universtity of Tübingen.

Conflict of Interest: none declared.

REFERENCES

Bannai,H. *et al*. (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305.

Blum,T. *et al*. (2009) MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics*, **10**, 274.

Boden,M. and Hawkins,J. (2005) Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics*, **21**, 2279–2286.

Brady,S. and Shatkay,H. (2008) EpiLoc: a (working) text-based system for predicting protein subcellular location. In *Pacific Symposium on Biocomputing.* World Scientific, pp. 604–615.

Briesemeister,S. *et al*. (2009) SherLoc2: a high-accuracy hybrid method for predicting protein subcellular localization. *J. Proteome Res.*, **8**, 5363–5366.

Carlson,M. and Botstein,D. (1982) Two differentially regulated mRNAs with different 5′ ends encode secreted with intracellular forms of yeast invertase. *Cell*, **28**, 145–154.

Casadio,R. *et al*. (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief Funct. Genomic Proteomic*, **7**, 63–67.

Cedano,J. *et al*. (1997) Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.*, **266**, 594–600.

Chou,K. and Cai,Y. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem*., **277**, 45765–45769.

Chou,K. and Cai,Y. (2003a) A new hybrid approach to predict subcellular localization of proteins by incorporating Gene Ontology. *Biochem. Biophys. Res. Commun.*, **311**, 743–747.

Chou,K. and Cai,Y. (2003b) Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J. Cell Biochem.*, **90**, 1250–1260.

Chou,K. *et al.* (2007) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.*, **6**, 1728–1734.

Chou, K. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct. Funct. Genet.*, **43**, 246–255.

Cokol,M. *et al*. (2000) Finding nuclear localization signals. *EMBO Rep.*, **1**, 411–415.

Cui,Q. *et al*. (2004) Esub 8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. *BMC Bioinformatics*, **5**, 66.

Emanuelsson,O. *et al*. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–971.

Fayyad,U.M. and Irani,K. (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufman Publishers, San Fransisco, pp. 1022–1027.

Fujiwara,Y. and Asogawa,M. (2001) Prediction of subcellular localizations using amino acid composition and order. *Genome Inform.*, **12**, 103–112.

Fyshe,A. *et al*. (2008). Improving subcellular localization prediction using text classification and the Gene Ontology. *Bioinformatics*, **24**, 2512–2517.

Garg,A. and Raghava,G. (2008) ESLpred 2: improved method for predicting subcellular localization of eukaryotic proteins. *BMC Bioinformatics*, **9**, 503.

Garg,P. *et al*. (2009) SubCellProt: predicting protein subcellular localization using machine learning approaches. *In Silico Biol.*, **9**, 35–44.

Guo,J. and Lin,Y. (2006) TSSub: eukaryotic protein subcellular localization by extracting features from profiles. *Bioinformatics*, **22**, 1784–1785.

Hall,M. (2000) Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., pp. 359–366.

Höglund,A. *et al*. (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22**, 1158–1165.

Horton,P. *et al*. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res*., **35**, W585–W587.

Huang,W. *et al*. (2008) ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinformatics*, **9**, 80.

Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Kaiser,C. and Botstein,D. (1986) Secretion-defective mutations in the signal sequence for Saccharomyces cerevisiae invertase. *Mol. Cell. Biol.*, **6**, 2382–2391.

King,B. and Guda,C. (2007) ngLOC: an n-gram-based Bayesian method for estimating the subcellular proteomes of eukaryotes. *Genome Biol.*, **8**, R68.

Lee,K. *et al*. (2009) Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic Acids Res.*, **36**, e136.

Lei,Z. and Dai,Y. (2006) Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinformatics*, **7**, 491.

Lin,H. *et al*. (2009) Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. *BMC Bioinformatics*, **10**, S8.

Lu,Z. and Hunter,L. (2005) GO molecular function terms are predictive of subcellular localization. In *Proceedings of Pacific Symposium on Biocomputing*, World Scientific, pp. 151–161.

Lu,Z. *et al*. (2004) Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, **20**, 547–556.

Nair,R. and Rost,B. (2002a) Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, **18**(Suppl. 1), S78–S86.

Nair,R. and Rost,B. (2002b) Sequence conserved for subcellular localization. *Protein Sci.*, **11**, 2836–2847.

Nair,R. and Rost,B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.*, **348**, 85–100.

Nakai,K. and Kanehisa,M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897–911.

Outten,C. and Culotta,V. (2004) Alternative start sites in the Saccharomyces cerevisiae GLR1 gene are responsible for mitochondrial and cytosolic isoforms of glutathione reductase. *J. Biol. Chem.*., **279**, 7785–7791.

Park,K. and Kanehisa,M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19**, 1656–1663.

Petsalaki,E. *et al*. (2006) PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genomics Proteomics Bioinformatics*, **4**, 48–55.

Pierleoni,A. *et al*. (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.

Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, **26**, 2230–2236.

Rish,I. (2001) An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, Morgan Kaufmann, pp. 41–46.

Scott,M. *et al.* (2004) Predicting subcellular localization via protein motif co-occurrence. *Genome Res.*, **14**, 1957–1966.

Scott,M. *et al*. (2005) Refining protein subcellular localization. *PLoS Comput. Biol.*, **1**, e66.

Shin,C.J. *et al*. (2009) Protein-protein interaction as a predictor of subcellular location. *BMC Syst. Biol*., **3**, 28.

Small,I. *et al*. (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.

Takada,Y. *et al*. (1990) Human peroxisomal L-alanine: glyoxylate aminotransferase. *Biochem. J.*, **268**, 517–520.

Tsoumakas,G. and Katakis,I. (2007) Multi-label classification: an overview. *Int. J. Data Warehousing Min.*, **3**, 1–13.

Whitten,I. and Frank,E. (2005) *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufman Publishers, San Fransisco.

Xie,D. *et al*. (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res.*, **33**, W105–W110.

Zhang,S. *et al*. (2008) DBMLoc: a database of proteins with multiple subcellular localizations. *BMC Bioinformatics*, **9**, 127.