YLoc—an interpretable web server for predicting subcellular localization

Sebastian Briesemeister1,*, Jörg Rahnenführer2 and Oliver Kohlbacher1

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

Received February 1, 2010; Revised May 2, 2010; Accepted May 13, 2010

ABSTRACT
Predicting subcellular localization has become a valuable alternative to time-consuming experimental methods. Major drawbacks of many of these predictors is their lack of interpretability and the fact that they do not provide an estimate of the confidence of an individual prediction. We present YLoc, an interpretable web server for predicting subcellular localization. YLoc uses natural language to explain why a prediction was made and which biological property of the protein was mainly responsible for it. In addition, YLoc estimates the reliability of its own predictions. YLoc can, thus, assist in understanding protein localization and in location engineering of proteins. The YLoc web server is available online at www.multiloc.org/YLoc.

INTRODUCTION
Protein sorting is a complex and still poorly understood process. It is crucial for a protein’s function as its location is often correlated with its molecular function. Thus, knowledge of protein localization can help biologists to infer the function of a protein. However, experimental methods for determining a protein’s location are expensive and time consuming. In contrast, computational predictions rely only on the protein sequence, and are fast, and fairly accurate. Over recent years, various prediction methods have been introduced. Most methods use sequence information, such as known sorting signals and amino acid composition (1–9). More advanced methods incorporate annotation information such as functional domains and motifs (10,11), homologous proteins (12,13), Gene Ontology (GO) terms (14) and textual information (15,16). Predictions based on annotated knowledge are often more accurate, but are less robust in cases where little is known about the protein. Hybrid prediction approaches combine the advantages of both information sources (17–21).

Although the prediction performance of subcellular localization predictors has increased significantly over recent years, their predictions are often not considered to be trustworthy. Very complex machine learning models of state-of-the-art prediction systems make it difficult to understand why a prediction was made. Consequently, the web interfaces of most methods are non-transparent and offer no explanation for a particular prediction. In addition, most methods do not offer confidence estimates for an individual prediction.

We present YLoc, an interpretable web server for predicting subcellular localization. Users are provided with the prediction itself, and also with an explanation why this prediction was made. The features contributing to the prediction are translated into natural language aiming at the most likely explanation of the localization. In addition, a confidence score helps the users to verify whether the prediction is reliable or not. YLoc is available in a low-resolution version, YLoc-LowRes, and a high-resolution version, YLoc-HighRes, covering 5 or 11 eukaryotic subcellular locations, respectively. YLoc+, the most general version, integrates multiple locations sites. All three predictors are available for animal, fungal and plant proteins.

METHODS AND MATERIALS
YLoc-LowRes was trained on the BaCelLo data set (6), which contains only globular proteins. The animal and fungal versions predict four locations: the nucleus (nu), cytoplasm (cy), mitochondrion (mi) and the secretory pathway (SP). The plant version additionally predicts the chloroplast (ch). YLoc-HighRes was trained on the Höglund data set (7). It covers 11 locations: nu, cy, mi, ch, endoplasmatic reticulum (er), Golgi apparatus (go), peroxisome (pe), plasma membrane (pm), extracellular space (ex), lysosome (ly) and vacuole (va). In the training of YLoc+, we used the Höglund data set and additional proteins with multiple locations from the DBMLoc database (22). The extracted 3054 proteins share <80% sequence similarity. Only dual locations

*To whom correspondence should be addressed. Tel: +49 7071 29 70462; Fax: +49 7071 29 5152; Email: briese@informatik.uni-tuebingen.de

© The Author(s) 2010. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
with more than 100 representative proteins were included: 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). To our knowledge, this is currently the largest data set of proteins from multiple locations.

We derived about 30 000 features from our protein sequences using amino acid composition and pseudo composition (3) as well as properties such as hydrophobicity, charge and volume of amino acids. In addition, we included PROSITE motifs and GO terms from close homologs. For more details, we refer to Briesemeister et al. (26). To guarantee interpretable predictions, we first reduced the number of features using a backward best first search together with correlation-based feature selection (23) implemented in the Weka machine learning library (24). For YLoc-LowRes, we obtained 20 features; for YLoc-HighRes and YLoc+, we obtained 30 features. However, a small number of features is only the first step toward interpretable predictions. To provide meaningful explanations, we manually annotated all selected features in biological terms. Unfortunately, not every feature can be easily mapped to a biological property. In such cases, we carefully inspected the initial feature set and transferred the biological meaning of a highly correlated feature. A list of all selected and annotated features can be found in the Supplementary Data.

YLoc uses naïve Bayes alongside entropy-based discretization (25) to make predictions. Given a set of features \( F = \{F_1, \ldots, F_k\} \) and a set of location classes \( C = \{C_1, \ldots, C_n\} \), the conditional distribution of class \( C_j \) can be expressed by:

\[
P(C_j | F) \propto P(C_j) \prod_{i=1}^{k} P(F_i | C_j). \tag{1}
\]

The final posterior probabilities \( P(C_j | F) \) are calculated by normalizing the right term of Equation (1) such that all posteriors sum up to one. Based on the feature likelihoods \( P(F_i | C) \), we calculate a discrimination score which provides a simple and transparent understanding of the influence of a feature on the prediction, for details see Briesemeister et al. (26). A positive score indicates that this feature is typical for the predicted location, whereas a negative score indicates that this feature alone would suggest a different location. Secondly, the discrimination score shows how strongly a feature influenced the prediction.

For multiple localization prediction, we assume that a protein present in multiple locations is equally distributed among those. Proteins labeled with two locations are assigned to a dual-location class, for example, \( C_{nu-cy} \). YLoc+ then evenly distributes the posterior probabilities of the dual-location classes onto the probabilities of the two individual locations. For example, \( P(C_{nu-cy} | F) / 2 \) is added to \( P(C_{nu} | F) \) and to \( P(C_{cy} | F) \). All locations with a probability above a threshold of 1/|\( C \)| are predicted, where |\( C \)| is the number of locations. If a location is less than half as probable as the next most probable one, this location and all less probable locations are not predicted.

The probability of the predicted location shows only how likely a protein is to be found in this location compared to the other locations. A confidence estimate, however, tells how likely it is that this prediction is to be correct. For this purpose, we analyze whether the protein is typical for the predicted class or whether YLoc already extrapolates. If a feature vector is more likely for proteins from the predicted location than for proteins from all locations, i.e. \( P(F | C_{pred}) > P(F | \bigcup C_j) \), we rate a prediction as being reliable. Since predicted locations with only a few training examples are often less reliable, we include the prior class probability in our confidence score:

\[
\frac{P(C_{pred}) P(F | C_{pred})}{P(C_{pred}) P(F | C_{pred}) + P(F | \bigcup C_j)}. \tag{2}
\]

Confidence scores range from zero for unreliable predictions to one for very confident predictions. For more details on the YLoc methodology, refer to Briesemeister et al. (26).

**EVALUATION**

We have tested the performance of YLoc on two independent data sets (IDSs). The BaCelLo IDS (27) consists of animal, fungal and plant proteins from the nu, cy, mi and SP which have at most 30% sequence identity to proteins in the BaCelLo data set. The Höglund IDS (20) contains animals proteins from remaining locations, the er, go, pe, pm, ex and ly, and was constructed with the same restrictions as the BaCelLo data set. In addition, proteins from the same location which align with an E-value \( \leq 10^{-5} \) are clustered and treated as one instance in the evaluation.

We compared the YLoc predictors with five other state-of-the-art subcellular localization predictors: MultiLoc2 (20), BaCelLo (6), LOCTree (4), WoLF PSORT (9) and Euk-mPloc (19). All methods are available as web servers. The individual prediction performance was evaluated using the overall accuracy (ACC), which is the percentage of correctly predicted instances, and the average \( F_1 \)-score \( (F_1) \), which is the average over the harmonic means of precision and recall of each location. Note that YLoc+, WoLF PSORT and Euk-mPloc are evaluated using the generalized ACC and \( F_1 \) from multilabel classification (28).

The evaluation results are summarized in Table 1. In our benchmark study, we observed that YLoc shows comparable performance to current state-of-the-art methods. For the BaCelLo IDS, YLoc-LowRes and MultiLoc2-LowRes perform best since they are specialized in distinguishing globular proteins. The high-resolution predictors YLoc-HighRes, YLoc+, MultiLoc2-HighRes, WoLF PSORT and Euk-mPloc perform slightly worse on this data set, since they are more general predictors. On the Höglund IDS, MultiLoc2-HighRes, YLoc-HighRes and YLoc+ show comparable performance, whereas WoLF PSORT and Euk-mPloc perform worse. MultiLoc2 shows very good accuracy throughout the study. However, its architecture is very complex and the
To show that users can benefit from the integrated confidence score, we analyzed the performance enrichment for high confidence scores. We reevaluated the performance of the YLoc predictors on the BaCelLo animals IDS by considering only proteins that could be predicted with a minimum confidence score. For statistical reasons, we excluded classes with less than five instances. The performance of YLoc for different minimum confidence scores is shown in Table 2. For the subset of proteins that can be predicted with high confidence, YLoc shows increased prediction performance. Consequently, predictions made with high confidence scores can be rated as more reliable.

We tested YLoc+’s ability to predict multiple localization sites in a nested 5-fold cross-validation scheme on the DBMLoc data set (22). We found that YLoc yields an ACC of 0.64 and an $F_1$ of 0.68 using multilabel measures. YLoc+ correctly identifies half of the proteins as multiple targeted and predicts both locations correctly in about one-third of the cases.

**WEB SERVER**

The YLoc web server requires protein sequences in FASTA format as input. It allows users to predict the location of at most 20 proteins. For large-scale predictions, users can access YLoc via SOAP or HTTP using the Python-based client scripts provided on the YLoc web site. Users can choose between three YLoc predictors, YLoc-LowRes, YLoc-HighRes and YLoc+, and three protein origins, animals, fungi and plants. In addition, they can switch off the use of GO term-based features. In this case, YLoc uses models in which the GO terms from close homologs are replaced by sequence-based features. Consequently, these YLoc models rely less on the presence of close homologous proteins. Every prediction will be assigned with a prediction ID that can be used to retrieve results later on. Alternatively, users can simply bookmark the waiting page or result page to obtain results later. Currently, predictions are saved for 2 weeks. The location prediction of a single protein takes 10–20 s, depending on the protein length.

Prediction results are displayed in three levels of details. The prediction summary presents the predicted localization(s), the probability of those and the confidence score for every query protein. The probability of a location is simply how likely the protein is located in this compartment. In contrast, the confidence score is a measure of reliability. A low confidence score implies the possibility that the real probability can differ considerably from the predicted probability. However, a high confidence score signifies that the predicted probability is close to the real probability for being located in the predicted location. Consequently, higher confidence scores imply a higher reliability of the prediction for the individual sequence. In addition, an explanation in natural language clarifies why the prediction has been made. This explanation includes the two most likely reasons for this localization, for example: ‘The most important reason for making this prediction is the strong SP sorting signal’ or ‘Moreover, it is a barely charged protein.’ This information can be very
important since it might already give a hint of the underlying mechanism for this protein localization.

The detailed prediction page provides more information on a particular protein prediction. For example, the probability distribution of the locations is provided. It is important to know the runner-up locations, especially for low confidence predictions, since rather ambiguous predictions should be inspected manually. YLoc also provides the most similar protein from Swiss-Prot 42.0 and associated GO terms. More details of how protein attributes influence the prediction are given in a large attribute table (Figure 1). The attributes are expressed in biological terms and ordered according to their absolute discrimination score, which corresponds to its influence on the prediction. A positive discrimination score implies that the attribute value is very typical for the predicted location, but atypical for some other location. In contrast, a negative discrimination score implies that the attribute value is more typical for some other location than the predicted one. A simple +/− encoding shows whether an attribute is typical for a location or not. By simply inspecting only the first lines of the table, it is sometimes already obvious which biological property lead to the prediction outcome and is likely to be responsible for the real localization of the protein. In addition, it gives hints of which parts of the protein should be considered for protein engineering.

How a particular biological attribute is calculated can be found on a detailed attribute page (Figure 2). For example, YLoc-LowRes (animal version) calculates the strength of the SP sorting signal using the ‘autocorrelation of every third hydrophobic amino acid within the first 20 amino acids in the N-terminus’. Knowing how the attribute value is calculated is essential to understand which particular amino acids and properties encode for possible sorting signal. Furthermore, the attribute is visualized. Embedded Javascript code displays the distribution of proteins from the different locations regarding this feature. The provided protein distributions are very helpful for understanding how proteins from different locations behave with respect to a biological property or sorting signal.

APPLICATION

The interpretable YLoc web service can be applied to numerous tasks that range from large-scale predictions to the identification of sorting signals. A very interesting application example is supervised protein engineering. YLoc can identify biological properties, e.g. example sorting signals that might be responsible for the localization. For example, human fumerate hydratase (FH, SwissProt AC P07954) is primarily located in the mi. The three YLoc predictors (animal version) detect the correct location and identify a mitochondrial targeting peptide (mTP). After truncating the leading 43 residues, FH lacks an mTP and shows a negatively charged N-terminus which is unfavorable for mitochondrial localization. Consequently, YLoc predicts FH to be cytoplasmic. In fact, the truncated FH protein is a known cytoplasmic isoform of FH encoded by the same gene (29). This example shows that YLoc can be valuable in location engineering of proteins.

YLoc VIA SOAP

For large-scale predictions, YLoc can be accessed via SOAP. The corresponding WSDL can be downloaded from the YLoc web site. In addition, we provide a Python-based script. Alternatively, YLoc can be accessed via an HTTP-based client that is also available for download.

CONCLUSION

As an interpretable web server for predicting subcellular localization of proteins, YLoc explains why a prediction was made and what features are likely to be responsible for the protein localization. This information can be very helpful to understand the localization of a protein and

![Figure 1.](http://nar.oxfordjournals.org/) The attribute table of the YLoc web service lists all attributes in order of their influence on the prediction outcome. All attributes are expressed in biological terms. The +(−) or −(+−) indicates whether that attribute value is (very) typical or (very) untypical for a location.
thus can assist in location engineering of proteins. Furthermore, a confidence score rates the reliability of a prediction. At the same time, it performs comparably with other state-of-the-art predictors. We believe that YLoc is a valuable alternative to experimental methods and current state-of-the-art predictors.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
We thank Jan Schulze for technical support and Nora Toussaint for comments on the manuscript.

FUNDING
LGFG Promotionsverbund ‘Pflanzliche Sensorhistidinkinasen’ of the University of Tübingen (S.B.). Funding for open access charge: LGFG Promotionsverbund ‘Pflanzliche Sensorhistidinkinasen’ of the University of Tübingen.

REFERENCES

Conflict of interest statement. None declared.


