Using N-terminal targeting sequences, amino acid composition, and sequence motifs for predicting protein subcellular localization

Annette Höglund*, Pierre Dönnes*, Torsten Blum*,
Hans-Werner Adolph†, and Oliver Kohlbacher*

Abstract: Functional annotation of unknown proteins is a major goal in proteomics. A key step in this annotation process is the definition of a protein's subcellular localization. As a consequence, numerous prediction techniques for localization have been developed over the years. These methods typically focus on a single underlying biological aspect or predict a subset of all possible subcellular localizations. There is a clear need for new methods that utilize and represent available protein specific biological knowledge from several sources, in order to improve accuracy and localization coverage for a wide range of organisms.

Here we present a novel Support Vector Machine (SVM)-based approach for predicting protein subcellular localization, which integrates information about N-terminal targeting sequences, amino acid composition, and protein sequence motifs. An important step is taken towards emulating the protein sorting process by capturing and bringing together biologically relevant information. Our novel approach has been used to develop two new prediction methods, TargetLoc and MultiLoc. TargetLoc is restricted to analysis of proteins containing N-terminal targeting sequences, whereas MultiLoc covers all major eukaryotic subcellular localizations for animal, plant, and fungal proteins. Compared to similar methods, TargetLoc performs better than these. MultiLoc performs considerably better than comparable prediction methods predicting all major eukaryotic subcellular localizations, and shows better or comparable results to methods that are specialized on fewer localizations or for one organism.

1 Introduction

Assigning subcellular localization to a protein is an important step towards elucidating its interaction partners, function, and its potential role(s) in the cellular machinery [RLN+03]. Despite recent technological advancements, experimental determination of subcellular localization remains time-consuming and laborious. Hence, computational methods for assigning localization on a proteome-wide scale offer an attractive complement.

Free diffusion in the cell is prohibited by membranes, leading to the subdivision of specific biochemical microenvironments into several subcellular organelles. Intracellular transport

*Department for Simulation of Biological Systems, WSI/ZBIT, Eberhard Karls University Tübingen, Sand 14, D-72076 Tübingen, Germany, (contact: hoeglund@informatik.uni-tuebingen.de)

†Department of Biochemistry and Center for Bioinformatics, Saarland University, D-66041 Saarbrücken, Germany

of proteins into these organelles is a highly regulated and specific process [PR87]. Signals for protein sorting exist either in the form of primary sequences, usually N-terminal targeting sequences [RK95, ENBvH00] and internal sequence motifs [CNR00], or in the form of 3D protein-surface patches [HA01]. Proteins localized in the same organelle have been reported to show a similar overall amino acid composition and are thought to have evolved to function optimally in that specific environment [AOR98]. Furthermore, a functional domain can be specific to proteins in one organelle, e.g. DNA-binding domains are present almost exclusively in nuclear proteins [HDAK05].

Methods for predicting subcellular localization can be categorized according to the underlying theory, e.g. classification based on N-terminal targeting sequences, overall amino acid composition, and sequence homology [DH04]. TargetP [ENBvH00] uses neural networks for discriminating four localizations: chloroplast, mitochondrial, secretory pathway, and other proteins, based on their N-terminal sequence information. An alternative and comparable method, iPSORT [BTM$^+$02], uses biologically interpretable rules of the N-terminal sequences for assigning the same localizations as TargetP. A number of different computational approaches using the overall amino acid composition have presented, including neural networks [RH98], Hidden Markov Models (HMMs) [Yua99], Support Vector Machines (SVMs) [HS01, PK03], and nearest neighbours [CC03, YY04]. Marcotte *et al.* explored the possibility to assign subcellular localization based on the distribution of protein homologues and their phylogenetic profiles [MXvDBE00]). Recent studies tend to combine several sources of information for prediction. Cai and Chou presented a method using gene ontology and functional domain composition [CC04a]. ELSpred is an SVM-based method using PSI-BLAST scores and amino acid composition [BR04], which was recently complemented by HSLpred and PSLpred for predicting four human and prokaryotic localizations, respectively [GBR05, BGR05]. PSLT is based on InterPro motifs and specific membrane domains [STH04]. A recently published method, LOCtree [NR05], integrates various sequence and predicted structural properties. The prediction shows promising results, but covers a limited number of localizations. The most comprehensive prediction system reported so far, PSORT, is based on a collection of expert if-then rules originating from experimental and computational observations [NK92, NH99]. PSORT covers the main eleven eukaryotic organelles and their intraorganellar localizations.

In the development of new prediction methods, special attention should be given to sequence homology within the data set and the method chosen for performance evaluation [HDAK05]. Including too homologous sequences in the data set will lead to recognition of identical sequences rather than common general features. Leave-one-out cross-validation is useful when the data at hand is limited. However, it tends to overestimate the performance for larger data sets, which makes five-fold cross-validation a more appropriate alternative [HTF01].

Here we present, a new integrated approach for predicting subcellular localization. The approach considers N-terminal targeting sequences, amino acid composition, and the presence of specific protein sequence motifs obtained from established motif databases. These features form the input for a set of SVMs used for predicting the localization. This novel approach was used for developing two prediction systems TargetLoc and MultiLoc. TargetLoc predicts four plant and three non-plant localizations based on N-terminal targeting

sequences, whereas MultiLoc covers all eleven eukaryotic subcellular localizations. Both methods have been compared to the respective current state-of-the-art methods TargetP, iPSORT, and PSORT. TargetLoc shows an improved discrimination of chloroplast and mitochondrial proteins, which is a well-known challenge due to evolutionary similarity between the two [CRXM95]. MultiLoc shows an overall prediction accuracy of about 75% in a cross-validation test, which can be directly compared to the overall accuracy of slightly less than 60% obtained by the PSORT method. TargetLoc and MultiLoc are described in detail, followed by the results of the benchmark studies and predictions on two novel data sets. We show that our integrative approach, which utilizes several different protein-specific features, gives a robust and accurate prediction method. Both prediction systems have been implemented as web services and are accessible at: http://www-bs.informatik.uni-tuebingen.de/Services/Loc/.

2 Methods and Materials

The aim of our work was to investigate if the novel approach, including additional biologically relevant information, could significantly improve prediction of localization. First, our approach was tested on classification of N-terminal localization categories. We developed the novel method TargetLoc, which was compared to the two methods TargetP and iPSORT using the well-known TargetP data set [ENBvH00]. Secondly, we applied our ideas to a more challenging problem, predicting all main eukaryotic subcellular localizations. The new prediction method, MultiLoc, was evaluated using an extensive data set obtained from the Swiss-Prot database and finally compared to the PSORT method. The architecture of the TargetLoc and MultiLoc prediction systems will be outlined and illustrated in the following subsections. The individual building blocks of the two methods are described in detail, followed by information about the data sets used for training and testing, the machine learning procedures, and the measures used for performance evaluation.

2.1 TargetLoc architecture

Prediction of localization categories based on N-terminal targeting sequences is relatively reliable and directly connected to the underlying biological process. However, N-terminal-based discrimination is only possible for the mitochondrial (*mi*), chloroplast (*ch*), secretory pathway (*SP*), and other (*OT*) categories. TargetLoc was designed to integrate several biologically relevant features for generating a protein profile vector (PPV) representing each protein. The features in the PPV originate from a set of specialized methods designed to detect protein specific features. Novel in TargetLoc (compared to other methods based on N-terminal prediction) is that in addition to the prediction based on the N-terminal sequences (here performed by SVMTarget), information about the overall amino acid composition (SVMaac) and protein specific motifs are collected (MotifSearch) in the PPV. All methods mentioned here are described in detail in the following. The overall architecture of TargetLoc is illustrated in Fig. 1. A query sequence is processed by a first layer of

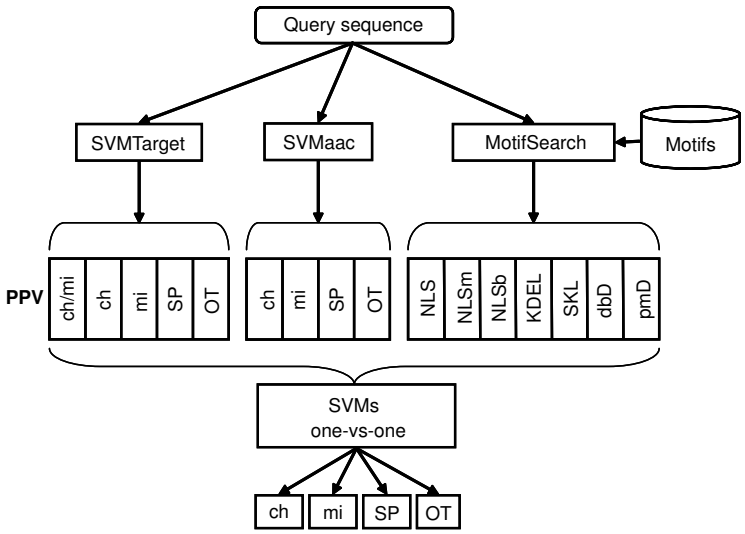

Figure 1: The architecture of the TargetLoc prediction system. A query sequence enters a first layer of prediction methods; SVMTarget, SVMaac, and MotifSearch. The information is collected in the protein profile vector (PPV). A set of one-versus-one SVMs are used by TargetLoc for the final classification according to the highest score using probability estimates.

three different methods for detecting protein features; SVMTarget, SVMaac, and MotifSearch. The results from these specific methods are collected in the PPV. The PPV is a representation of each protein and used as input for the final layer of SVMs in TargetLoc.

2.2 MultiLoc architecture

The basic idea to integrate additional biologically relevant information for improving the prediction of localization categories, was extended to enable discrimination between the main eleven eukaryotic subcellular localizations. Discrimination of all putative localizations is highly desirable from a biological point of view. However, it presents a challenging computational task for sparsely-populated localizations and for localizations for which no unique features exist. MultiLoc addresses this challenge by utilizing the same approach as was used in TargetLoc, but incorporating more information in order to enable extended prediction of further localizations. The architecture of our novel prediction method MultiLoc, is presented in Fig. 2. Several additional features have been incorporated into the first layer in order to facilitate for the extended number of localizations to be discriminated by MultiLoc. Furthermore, a method (SVMSA) for detecting signal anchors (SAs) has been implemented (described in detail below). MultiLoc was trained using a novel homology-reduced data set and the prediction performance of the three versions animal, fungal, and plant was compared to that of the PSORT method.

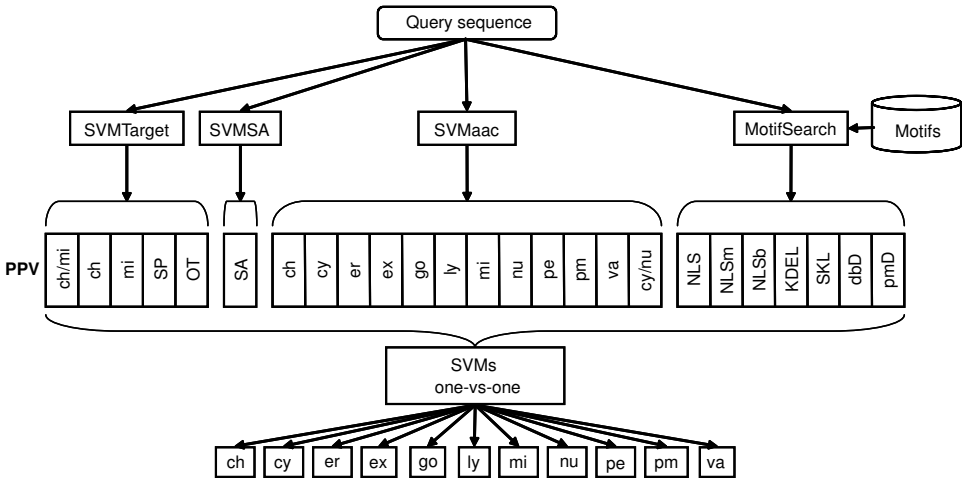

Figure 2: The architecture of the MultiLoc prediction system. A query sequence enters a first layer of prediction methods; SVMTarget, SVMSA, SVMaac, and MotifSearch. The characteristics of each protein is collected in the PPV, which is used to obtain probability estimates by a set of one-versus-one SVMs at the second layer.

2.3 Subprediction methods

The individual building blocks used in the development of TargetLoc and MultiLoc (illustrated in Fig. 1 and Fig. 2, respectively) are presented here.

SVMTarget

Similarly to TargetP, SVMTarget (plant and non-plant versions) predicts localization categories based on N-terminal targeting sequences. The plant version predicts four categories (*ch*, *mi*, *SP*, and *OT*) based on the type of targeting sequence: chloroplast transit peptides (cTP), mitochondrial transit peptides (mTP), targeting peptides of the secretory pathway (SP), and other proteins lacking N-terminal targeting sequences. The three non-plant categories *mi*, *SP*, and *OT* are predicted based on the recognition of the same types as above except for the cTP. The main differences between SVMTarget and TargetP are the encoding of the N-terminal sequences and the enhanced discrimination of cTPs and mTPs. TargetP uses the primary amino acid sequence, whereas SVMTarget uses the partial amino acid composition.

SVMTarget was constructed to reflect three basic biological observations, for a graphical illustration of the plant version, see Fig. 3. First, each type of N-terminal targeting sequence has a characteristic amino acid composition (rather than a similar primary amino acid sequence), which is different from the other N-terminal targeting sequences and the mature protein sequence. The second observation is that the different N-terminal targeting sequences have different average lengths. SPs are normally about 10 to 15 amino acids shorter than mTPs, and cTPs are usually significantly longer than mTPs. Finally, it has

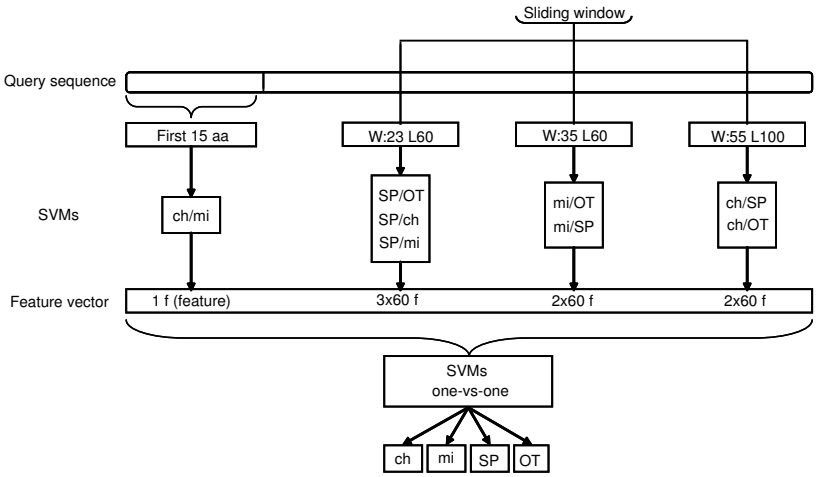

Figure 3: The architecture of the SVMTarget plant version is illustrated here. Sliding windows of width W over the first L number of N-terminal amino acid residues create the partial amino acid composition vectors, which are used as input for the first layer of SVMs. There are eight binary SVMs in the plant version and four in the non-plant version (not shown). The input for the *ch/mi* classifier is the amino acid composition of the first 15 N-terminal residues. The input for the second layer of SVMs consists of the output scores from the first, where a set of one-versus-one SVMs are used for the final classification using probability estimates.

been shown that the main difference between cTPs and mTPs is within the amino acid composition of the 15 most N-terminal residues. The first layer is a set of binary predictors, each specialized to recognize the differences in the amino acid composition between two types of N-terminal sequences. SVMs have been trained to recognize the first 60 (for mTPs and SPs) and the first 100 (for cTPs) N-terminal amino acids, by placing a window of length W around each amino acid in the targeting sequence (see Fig. 3). W is dependent on the length (L) of the experimentally observed targeting sequences and has been manually optimized for each localization (55 for cTPs, 35 for mTPs, and 23 for SPs). The cTP/mTP classification is not constructed using windows, instead the fact that mTPs tend to have a higher fraction of positive amino acids in their N-terminal end compared to cTPs [BTM$^+$02] is used. The second layer in the architecture is a set of SVMs, used for final classification based on the output of the first layer. The one-versus-one classification procedure uses probability estimates for determining the localization. In order to evaluate the performance of SVMTarget, the training and testing was done using a strict five-fold cross-validation procedure and the same data set as the TargetP method.

SVMaac

Overall amino acid composition contributes to the PPV through a set of binary localization specific SVMs (SVMaac), which is illustrated in Fig. 1 and 2, respectively. TargetLoc has four plant (*ch*, *mi*, *SP*, and *OT*) and three non-plant (*mi*, *SP*, and *OT*) binary classifiers for the overall amino acid composition. In MultiLoc the number of binary classifiers corresponds to the number of localizations, nine for animal and fungi, and ten for plant. Each

binary classifier discriminates between one localization and all others. Additionally there is a classifier that specifically discriminates cytoplasmic from nuclear proteins (*cy/nu*).

SVMSA

Membrane proteins of the secretory pathway can have a signal anchor (SA) sequence instead of the N-terminal targeting sequence. SAs are localized further away from the N-terminal end of the protein and they usually have a longer hydrophobic part compared to SPs. The SAs may escape detection by methods like SVMTarget and TargetP. To address this problem a novel classifier, SVMSA, specifically designed to recognize SAs was constructed. The basic architecture of SVMSA is similar to that of SVMTarget (which was presented in Fig. 3). The differences are that there is only one classifier at the first level and that the first 100 amino acids (L) and a window length (W) of 21 are used, compared to the first 60 amino acids and windows of length 23, 35, or 55 in SVMTarget.

Out of the 2595 proteins of the homology-reduced *SP* category, exactly 300 proteins contain an SA sequence instead of an SP sequence. This data set was used for training the SVMSA against an equal number of proteins lacking SAs and SPs (obtained from the *cy*, *ch*, *mi*, *nu*, and peroxisomal (*pe*) categories), using a five-fold cross-validation procedure. The recognition of SAs is very reliable, with an overall accuracy above 90%. The TargetP data set does not contain proteins with SAs, hence SVMSA was only used as an integrated part of the PPV in MultiLoc but not in TargetLoc.

MotifSearch

Sequence motifs and structural domains provide essential biological information about a protein. Detection of such information was facilitated through the development of MotifSearch, which has been integrated into both TargetLoc and MultiLoc and is illustrated in Fig. 1 and 2. MotifSearch relies on the information mainly from the PROSITE [BB94] and NLSdb [CNR00, NCR03] databases.

Most nuclear proteins carry a nuclear localization signal sequence (NLS), which can be recognized by nuclear import receptors. There are two major types of NLSs, the monopartite (*NLSm*) and the bipartite (*NLSb*). The NLSm:s are short (four to eight amino acids) and are rich in positively charged amino acids, whereas NLSb:s consist of two parts, each with a length of two to four amino acids that are connected by a spacer sequence. NLSdb, is a database containing experimentally known and potential NLSs [CNR00]. In addition to the specific NLSdb entries, the NLSm consensus pattern *K(K|R)X(K|R)* is detected. The PROSITE database contains information about protein sequence motifs, such as structural and functional domains. Four types of PROSITE motifs were found to have a high discrimination power on a scan of the MultiLoc homology-reduced data set. These motifs; the endoplasmic reticulum retention signal *KDEL*, the C-terminal targeting signal for the peroxisome *SKL*, 25 different DNA-binding domains (*dbD*), and 16 plasma membrane receptor domains (*pmD*), were included in the MotifSearch method.

2.4 Data sets

TargetP

The TargetP data sets were obtained from the TargetP web site and used for training and benchmarking TargetLoc against other comparable methods predicting N-terminal localization categories. These data sets contain a total of 3678 proteins representing four plant (*ch*, *mi*, *SP*, and *OT*) and three non-plant localizations (*mi*, *SP*, and *OT*). The *SP* category are proteins from the endoplasmic reticulum (*er*), extracellular space (*ex*), Golgi apparatus (*go*), lysosome (*ly*), plasma membrane (*pm*), and vacuole (*va*). Cytoplasmic (*cy*) and nuclear (*nu*) proteins belong to the category of *OT* proteins.

MultiLoc

The extensive data set was obtained by extracting all animal, fungal, and plant protein sequences from the Swiss-Prot [BA00] database release 42 (from 2003-2004), using the keywords *Metazoa*, *Fungi*, or *Viridiplantae*, respectively, in the OC (organism classification) field. These proteins were further assigned to one of eleven possible eukaryotic subcellular localizations, based on the annotation in the CC (comments) field. Plant proteins can be localized in the *ch*, *cy*, *er*, *ex*, *go*, *mi*, *nu*, *pe*, *pm*, and *va*. Fungal cells share the same subcellular localizations as plant cells, except that they lack the chloroplast. Finally, animal cells share all localizations with fungal cells, but have lysosomes instead of vacuoles.

A few localizations such as the cytoplasm, extracellular space, and the nucleus are densely populated, hence all proteins with annotations containing uncertainties such as ”potential”, ”by similarity”, or ”probable” were excluded. Mitochondrial and chloroplast proteins were only accepted if the keyword ”transit” followed by an annotated cleavage site was present in the FT (feature) field. Proteins of the *SP* category (*er*, *ex*, *go*, *ly*, *pm*, and *va*) were selected if the keywords ”signal” or ”signal-anchor” and annotated start and stop sites were present in the FT field. Plasma membrane proteins were required to have the keywords ”domain” and ”extracellular” as well as ”domain” and ”cytoplasmic” in the FT fields. These proteins were not accepted if the keywords ”domain” and ”lumenal” were present in one of the FT fields. A total of 9761 sequences were extracted, with no restrictions on the level of homology (All), see Table 1. Training the prediction models on data sets containing sequences with too high similarity will lead to recognition of nearly identical sequences, rather than general features. Hence, a homology-reduced data set was created using the ClustalW [THG94] algorithm, by removing proteins from the original data set until it contained no sequences with a pair-wise similarity higher than 80%, see Table 1.

2.5 SVM training and performance evaluation

The mathematical and technical details of SVMs are not explained here, but have been described in detail by Vapnik [Vap99]. In this study the LIBSVM (Chang, C.-C and Lin, C.-J, 2001, http://www.csie.ntu.edu.tx/∼cjlin/libsvm/) software was used. The Radial Basis Function kernel was used in all SVMs at all stages of the classification process, and

Table 1: The number of protein sequences listed for each data set according to localization. At the bottom of the table, the total number of sequences in the data set before and after homology reduction is displayed. The homology-reduced data set contains no proteins sequences with a sequence similarity higher than 80%, and was used for developing MultiLoc.

	Data set	
Localization	**All**	**Homology-reduced**
chloroplast (*ch*)	730	449
cytoplasm (*cy*)	2768	1411
endoplasmic reticulum (*er*)	328	198
extracellular space (*ex*)	1124	843
Golgi apparatus (*go*)	244	150
lysosome (*ly*)	164	103
mitochondria (*mi*)	872	510
nucleus (*nu*)	1040	837
peroxisome (*pe*)	278	157
plasma membrane (*pm*)	2115	1238
vacuole (*va*)	98	63
Total	**9761**	**5959**

optimized by tuning the c and g parameters. The one-versus-one (appropriate for multi-class classification) procedure was adopted throughout the SVM training, in favour of the one-versus-all procedure. The probability estimates by LIBSVM, were used for choosing the most probable classifications. Five-fold cross-validation was applied throughout the training and evaluation of both TargetLoc and MultiLoc. Special care was taken to ensure that no protein sequence used for training either SVMTarget, SVMaac, or SVMSA, was used in the evaluation neither of the TargetLoc, nor the MultiLoc performance. Five-fold cross-validation is a robust method for performance evaluation and used in favour of leave-one-out cross-validation when the data set is large enough, since it better avoids the danger of overfitting [HTF01].

The original data set split used for training and testing the TargetP method is not available. Hence, in addition to the five-fold cross-validation, a randomization process was performed by randomly splitting the data five times. This procedure delivers five parallel models for each classifier and enables a statistically sound and fair comparison between the TargetLoc, SVMTarget, and TargetP methods. Specificity (*SP*), sensitivity (*SE*), and Matthews correlation coefficient (*MCC*) [Mat75] (a measure capturing both *SP* and *SE*) were calculated for all prediction methods. Furthermore, the overall accuracy (correct[%]) and standard deviation of the performances were calculated.

3 Results

TargetLoc

The performances of the new method TargetLoc were compared against TargetP and iP-

Table 2: Performance comparison of TargetLoc against the TargetP and iPSORT methods, using the TargetP size non-equalized data set (940 plant and 2738 non-plant proteins). TargetLoc has been trained and evaluated using five-fold cross-validation. The performances presented in this table are averages from 25 different prediction models (originating from the cross-validation procedure, which was performed using five different data set splits). The standard deviations are shown in parenthesis and refer to the performance variations of the different cross-validation models. The TargetP and iPSORT values were presented by Emanuelsson *et al.* in [ENBvH00].

Version	Method	Category	*SE*	*SP*	*MCC*	correct[%]
Plant	**TargetLoc**	*ch*	0.88	0.76	0.78	**89.7 (±1.6)**
		mi	0.87	0.94	0.84	
		SP	0.93	0.97	0.93	
		OT	0.92	0.84	0.86	
	TargetP	*ch*	0.85	0.69	0.72	**85.3 (±3.5)**
		mi	0.82	0.90	0.77	
		SP	0.91	0.95	0.90	
		OT	0.85	0.78	0.77	
	iPSORT	*ch*	0.68	0.71	0.64	**83.4**
		mi	0.84	0.86	0.75	
		SP	0.91	0.98	0.92	
		OT	0.83	0.70	0.71	
Non-plant	**TargetLoc**	*mi*	0.91	0.77	0.81	**92.5 (±1.2)**
		SP	0.95	0.92	0.91	
		OT	0.91	0.97	0.86	
	TargetP	*mi*	0.89	0.67	0.73	**90.0 (±0.7)**
		SP	0.96	0.92	0.92	
		OT	0.88	0.97	0.82	
	iPSORT	*mi*	0.74	0.68	0.67	**88.5**
		SP	0.92	0.92	0.90	
		OT	0.90	0.92	0.78	

SORT using the TargetP data set, see Table 2. Five random splits of the data sets were used for training and evaluation, showing very low standard deviations. TargetP and iPSORT predict the four plant categories (*ch*, *mi*, *SP*, and *OT*) with an overall accuracy of 85.3%, and 83.4.%, respectively, whereas TargetLoc reaches an overall accuracy of 89.7%. A similar trend can be observed for the three non-plant categories (*mi*, *SP*, and *OT*), where TargetLoc reaches an overall accuracy of 92.5%. The corresponding performances for TargetP and iPSORT are 90.0% and 88.5%, respectively. The most important improvement by TargetLoc compared to previously reported methods, is in the discrimination between *ch* and *mi* proteins, and classification of the *OT* category. The *MCCs* for the *ch*, *mi*, *SP*, and *OT* categories are reported to be 0.72, 0.77, 0.90, and 0.77 for TargetP, which have been significantly improved to 0.78, 0.84, 0.93, and 0.86 by TargetLoc.

MultiLoc

The overall accuracies of the three MultiLoc versions; animal, fungal, and plant, reach approximately 75%. These results reflect a considerable improvement and should be compared to the corresponding values for PSORT ranging between 58-60%, see Table 3. The fungal version of MultiLoc can be compared to the yeast version of PSORT, since they predict the same localizations. The *SE*, *SP*, and *MCC* values for each version and localization

of both MultiLoc and PSORT are presented in detail in Table 3. Using the MultiLoc animal version the *MCC* ranges between 0.44 for peroxisomal and 0.83 for mitochondrial proteins and the other two versions show similar results. These results can be directly compared to those of PSORT in the same table. The PSORT performance on the homology-reduced data set is in agreement with earlier reports of PSORT performance, which was slightly below a 60% using a smaller data set [NK92, NH99]. The performance of the PSORT animal version varies widely with an *MCC* between 0.11 for proteins of the endoplasmic reticulum and 0.73 for plasma membrane proteins. The PSORT fungal version predicts the *go* and *va* localizations with a very low *MCC* values of 0.04 and 0.08, respectively. The corresponding *MCCs* for MultiLoc show a clear improvement to 0.60 and 0.42.

The overall accuracy of MultiLoc is significantly higher and less dependent on the localization category compared to PSORT. Low standard deviations indicate that the different prediction models are robust. The effect of bringing together different sources of information is even more prominent when predicting eleven localizations. MultiLoc has an overall accuracy higher than 74% for all three versions, which is a significant improvement when compared to the PSORT performance of less than 60%.

4 Discussion

We have shown that our new approach for predicting protein subcellular localization significantly improves the robustness and prediction reliability. In this approach several sources of biological information are integrated, thereby covering several aspects of the protein sorting process. Its successful application is exemplified through our new prediction methods TargetLoc and MultiLoc, which have been compared to other current state-of-the-art methods.

Predicting localization of proteins based on their N-terminal amino acid sequence is considered to be very reliable, as the overall prediction accuracy reached about 85% with the development of TargetP [ENBvH00]. The N-terminal amino acid sequences are important and highly characteristic for the precursor proteins of the *ch*, *mi*, and *SP* categories. Proteins of the *OT* category, on the other hand, lack this N-terminal precursor sequence and can easily be mixed up with mature proteins from one of the three first categories.

Our new integrative approach, TargetLoc, predicts the same localization categories as TargetP but differs in a number of fundamental ways. Three complementary sources of biologically relevant information are brought together, namely; N-terminal sequence information, overall amino acid composition, and protein specific sequence motifs. In SVMTarget the encoding of the N-terminal targeting sequence reflects the meaningful amino acid composition [CKH93], rather than the primary sequence. Proteins in the *ch* and *mi* categories are specifically discriminated by SVMTarget through the inclusion of the composition differences known to exist within the 15 most N-terminal amino acids [CRXM95]. The overall amino acid composition is used for capturing subtle differences between the different categories (SVMaac). MotifSearch identifies *SP* and *OT* proteins by their relatively high probability to carry one of the protein sequence motifs characteristic to the mixed group

Table 3: The performance of MultiLoc compared to PSORT using the homology-reduced data set. MultiLoc was trained and evaluated using five-fold cross-validation. The number of proteins (Nr) is listed next to each localization (Loc). Detailed values of the sensitivity (*SE*), specificity (*SP*), and Matthews correlation coefficient (*MCC*) are listed for each localization. The overall prediction accuracy (correct[%]) is listed for the animal, fungal, and plant versions. The standard deviations (in parenthesis) refer to the differences in the overall accuracies of the five cross-validation models (not availabel for PSORT).

			PSORT				**MultiLoc**			
Version	**Loc**	**Nr**	***SE***	***SP***	***MCC***	**correct[%]**	***SE***	***SP***	***MCC***	**correct[%]**
Animal	*cy*	1411	0.39	0.72	0.43	**59.9**	0.67	0.85	0.68	**74.6 (±1.0)**
	er	198	0.21	0.12	0.11		0.68	0.56	0.60	
	ex	843	0.81	0.72	0.72		0.79	0.83	0.77	
	go	150	0.02	0.14	0.04		0.71	0.43	0.53	
	ly	103	0.18	0.20	0.18		0.69	0.36	0.48	
	mi	510	0.70	0.56	0.58		0.88	0.82	0.83	
	nu	837	0.60	0.61	0.54		0.82	0.73	0.73	
	pe	157	0.48	0.17	0.25		0.71	0.31	0.44	
	pm	1238	0.83	0.76	0.73		0.73	0.90	0.76	
Fungal	*cy*	1411	0.39	0.73	0.43	**59.3**	0.68	0.85	0.69	**74.9 (±0.7)**
	er	198	0.23	0.13	0.13		0.71	0.59	0.63	
	ex	843	0.74	0.71	0.68		0.73	0.81	0.73	
	go	150	0.02	0.14	0.04		0.71	0.53	0.60	
	mi	510	0.70	0.56	0.58		0.88	0.82	0.83	
	nu	837	0.60	0.61	0.54		0.81	0.74	0.73	
	pe	157	0.48	0.17	0.25		0.68	0.30	0.43	
	pm	1238	0.83	0.77	0.73		0.76	0.89	0.78	
	va	63	0.13	0.07	0.08		0.76	0.24	0.42	
Plant	*ch*	449	0.49	0.58	0.50	**57.5**	0.88	0.85	0.85	**74.6 (±0.8)**
	cy	1411	0.40	0.70	0.42		0.68	0.85	0.70	
	er	198	0.21	0.11	0.11		0.72	0.54	0.61	
	ex	843	0.74	0.70	0.67		0.68	0.81	0.70	
	go	150	0.02	0.13	0.04		0.75	0.41	0.54	
	mi	510	0.65	0.53	0.54		0.85	0.81	0.81	
	nu	837	0.59	0.60	0.53		0.82	0.75	0.75	
	pe	157	0.47	0.16	0.24		0.71	0.34	0.47	
	pm	1238	0.81	0.75	0.72		0.74	0.89	0.77	
	va	63	0.13	0.06	0.07		0.70	0.20	0.36	

of proteins in these categories. TargetLoc has improved the overall accuracy compared to TargetP from 85.3% to 89.7% and from 90.0% to 92.5% for the plant and the non-plant versions, respectively.

Predicting these N-terminal categories is reliable and useful, nevertheless, suffer from two major drawbacks. First, predicting only four localizations in plant, when there are at least ten main localizations, and three for non-plant proteins, when there are nine main localizations. The further intraorganellar sorting is completely neglected. Second, precursor protein sequences of the *ch*, *mi*, and *SP* categories are not always available and the cleavage sites of the targeting sequences are not trivial to identify [FMG99]. MultiLoc was developed in order to meet the need of a fine-grained and reliable prediction system for protein subcellular localization. Utilizing biological knowledge for modeling the biological sorting process, an extensive homology-reduced data set, and the strong predictive power of SVMs proved successful. As strict cross-validation was used for evaluation, the performance of MultiLoc (75%) can be directly compared to that of PSORT (less than 60%). The low standard deviations indicate that MultiLoc is robust. The major improvements were made for the *cy*, *er*, *go*, *ly*, *pe*, and *va* localizations, which have been a major challenge so far.

Several biological aspects of protein sorting are yet to be understood. In the meantime it is useful to include as much information about each protein as possible when designing prediction models [DH04]. MultiLoc performs outstandingly well for a wide range of localizations, reaching levels of accuracy where it becomes interesting to investigate the incorrectly predicted proteins in greater detail. It is likely that some of these incorrectly predicted proteins are to be found in multiple localizations (multiplex localizations) [CRXM95, MMDS$^+$98, CC04b]. We have also taken an important step towards high reliability predictions of all eukaryotic subcellular localizations. Interesting computational challenges lie ahead, the main still being to mirror the biological events in the protein sorting process. The concept of a PPV makes the prediction method easily extendable. Near future improvements of both TargetLoc and MultiLoc, include careful selection of additional protein features for the PPV and extension of the prediction to cover the further intraorganellar sorting of mitochondria and chloroplast proteins. Integrating multiple sources of information and simulating the structure of the sorting process will probably be the key to inferring protein function from subcellular localization.

References

[AOR98] M A Andrade, S I O'Donoghue, and B Rost. Adaption of protein surfaces to subcellular location. *J Mol Biol.*, 276(2):517–525, 1998.

[BA00] A Bairoch and R Apweiler. The SWISS-PROT protein sequence database and its supplement in TrEMBL in 2000. *Nucleic Acids Res.*, 28(1):45–48, 2000.

[BB94] A Bairoch and P Bucher. PROSITE: recent developments. *Nucleic Acids Res.*, 22(17):3583–3589, 1994.

[BGR05] M Bhasin, A Garg, and GP Raghava. PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics*, Epub ahead of print, 2005.

[BR04] M Bhasin and G P Raghava. ELSpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acid Res.*, 32(Web Server issue):W414–W419, 2004.

[BTM+02] H Bannai, Y Tamada, O Maruyama, K Nakai, and S Miyano. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics.*, 18(2):298–305, 2002.

[CC03] Y D Cai and K C Chou. Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochem Biophys Res Commun.*, 305(2):407–411, 2003.

[CC04a] Y D Cai and K C Chou. Predicting subcellular localization of proteins in a hybridization space. *Bioinformatics*, 20(7):1151–1156, 2004.

[CC04b] YD Cai and KC Chou. Predicting 22 protein localizations in budding yeast. *Biochem Biophys Res Commun.*, 323(2):425–428, 2004.

[CKH93] S Clausmeyer, RB Klosgen, and RG Herrmann. Protein import into chloroplasts. The hydrophilic lumenal proteins exhibit unexpected import and sorting specificities in spite of structurally conserved transit peptides. *J Biol Chem.*, 268(19):13869–13876, 1993.

[CNR00] M Cokol, R Nair, and B Rost. Finding nuclear localization signals. *EMBO Rep.*, 1(5):411–415, 2000.

[CRXM95] G Creissen, H Reynolds, Y Xue, and P Mullineaux. Simultaneous targeting of pea glutathione reductase and of a bacterial fusion protein to chloroplasts and mitochondria in transgenic tobacco. *Plant J.*, 8(2):167–175, 1995.

[DH04] P Dönnes and A Höglund. Predicting Protein Subcellular Localization: Past, Present, and Future. *Genomics, Proteomics, and Bioinformatics*, 2(4), 2004.

[ENBvH00] 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(4):1005–1016, 2000.

[FMG99] D Frishman, A Mironov, and M Gelfand. Starts of bacterial genes: estimating the reliability of computer predictions. *Gene.*, 234(2):257–265, 1999.

[GBR05] A Garg, M Bhasin, and GP Raghava. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J Biol Chem.*, 280(15):14427–14432, 2005.

[HA01] A Helenius and M Aebi. Intracellular functions of N-linked glycans. *Science.*, 291(5512):2364–2369, 2001.

[HDAK05] A Höglund, P Dönnes, H Adolph, and O Kohlbacher. From prediction of subcellular localization to functional classification: Discrimination of DNA-packing and other nuclear proteins. *Online Journal of Bioinformatics*, 6(1):51–64, 2005.

[HS01] S Hua and Z Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics.*, 17(8):721–728, 2001.

[HTF01] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, USA, 2001.

[Mat75] B W Mathews. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta.*, 405:442–451, 1975.

[MMDS+98] B Menand, L Marechal-Drouard, W Sakamoto, A Dietrich, and H Wintz. A single gene of chloroplast origin codes for mitochondrial and chloroplastic methionyl-tRNA synthetase in Arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 95(18):11014–11019, 1998.

[MXvDBE00] E M Marcotte, I Xenarios, A M van Der Bliek, and D Eisenberg. Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci U S A.*, 97(22):12115–12120, 2000.

[NCR03] R Nair, P Carter, and B Rost. NLSdb: database of nuclear localization signals. *Nucleic Acids Res.*, 31(1):397–399, 2003.

[NH99] K Nakai and P Horton. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci.*, 24(1):34–36, 1999.

[NK92] K Nakai and M Kanehisa. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics.*, 14(4):897–911, 1992.

[NR05] R Nair and B Rost. Mimicking cellular sorting improves prediction of subcellular localization. *J Mol Biol.*, 348(1):85–100, 2005 2005.

[PK03] K-J Park and M Kanehisa. Prediction of protein subcellular location by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics.*, 19(13):1656–1663, 2003.

[PR87] S R Pfeffer and J E Rotheman. Biosynthetic transport and sorting by the endoplasmatic reticulum and Golgi. *Annu. Rev. Biochem.*, 56:829–852, 1987.

[RH98] A Reinhardt and T Hubbard. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, 26(9):2230–2236, 1998.

[RK95] S L Rusch and D A Kendall. Protein transport via amino-terminal targeting sequences: common themes in diverse systems. *Mol Membr Biol.*, 12(4):295–307, 1995.

[RLN+03] B Rost, J Liu, R Nair, KO Wrzeszczynski, and Y Ofran. Automatic prediction of protein function. *Cell Mol Life Sci.*, 60(12):2637–2650, 2003.

[STH04] MS Scott, DY Thomas, and MT Hallett. Predicting subcellular localization via protein motif co-occurrence. *Genome Res.*, 14(10A):1957–1966, 2004.

[THG94] J D Thompson, D G Higgins, and T J Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673–4680, 1994.

[Vap99] V N Vapnik. *The Nature of Statistical Learning Theory*. Wiley, New York, USA, 1999.

[Yua99] Z Yuan. Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.*, 451(1):23–26, 1999.

[YY04] H Ying and L Yanda. Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics*, 20(1):21–28, 2004.