Significantly Improved Prediction of Subcellular Localization by Integrating Text and Protein Sequence Data

Annette Hoglund, Torsten Blum, Scott Brady, Pierre Donnes, John San Miguel, Matthew Rocheford, Oliver Kohlbacher, and Hagit Shatkay

Pacific Symposium on Biocomputing 11:16-27(2006)

SIGNIFICANTLY IMPROVED PREDICTION OF SUBCELLULAR LOCALIZATION BY INTEGRATING TEXT AND PROTEIN SEQUENCE DATA

ANNETTE HÖGLUND†, TORSTEN BLUM†, SCOTT BRADY‡,
PIERRE DÖNNES†, JOHN SAN MIGUEL‡, MATTHEW ROCHEFORD‡,
OLIVER KOHLBACHER†, HAGIT SHATKAY‡*

† *Div. for Simulation of Biological Systems, ZBIT/WSI, University of Tübingen, Sand 14, D-72076 Tübingen, Germany*

‡ *School of Computing, Queen's University, Kingston, Ontario, Canada K7L 3N6*

Computational prediction of protein subcellular localization is a challenging problem. Several approaches have been presented during the past few years; some attempt to cover a wide variety of localizations, while others focus on a small number of localizations and on specific organisms. We present a comprehensive system, integrating protein sequence-derived data and text-based information. It is tested on three large data sets, previously used by leading prediction methods. The results demonstrate that our system performs significantly better than previously reported results, for a wide range of eukaryotic subcellular localizations.

1. Introduction

In this paper we introduce a new system for computationally assigning proteins to their subcellular localization. By integrating several types of sequence-derived features and text-based information, the achieved performance is the best reported so far, in terms of sensitivity, specificity, and overall accuracy. Unlike several recent systems which focus on a few subcellular localizations or on a specific organism[1,2,3,4], our system is applicable to – and retains its good performance across – a wide variety of organisms and subcellular localizations. Moreover, we show that the integrated system, which combines sequence and text, performs significantly better than its individual components, based on each data source alone.

The task of protein subcellular localization prediction is important and well-studied[5,6]. Knowing a protein's localization helps elucidate its function, its role in both healthy processes and in the onset of disease, and its potential use as a drug target. Experimental methods for protein localization range from immunolocalization[7] to tagging of proteins using green fluorescent protein (GFP)[8]

*To whom correspondence should be addressed: shatkay@cs.queensu.ca. HS is supported by NSERC Discovery grant 298292-04.

and isotopes[9]. Such methods are accurate but, even at their best, are slow and labor-intensive compared with large-scale computational methods. Computational tools for predicting localization are useful for a large-scale initial "triage", especially for proteins whose amino acid sequence may be determined from the genomic sequence, but are hard to produce, isolate, or locate experimentally.

The past decade, and most notably the last five years, has seen much progress in computational prediction of protein localization from sequence data. Nakai and Kanehisa[10,11] introduced PSort, a rule-based expert system, which was later improved upon by a probabilistic[12] and by a K-nearest neighbor[13] classifier. Another pair of prominent systems, TargetP[1] and ChloroP[14], based on artificial neural networks, demonstrated a significantly higher accuracy when applied to a limited set of subcellular localizations in plant and animal cells. Other recent systems use a variety of machine learning techniques. Most of them focus on a few subcellular localizations and improve upon – or just meet – the state of the art on those[15,3,16].

Several recent publications have examined the possibility of using text to support subcellular localization. Specifically, Stapley *et al.*[17] represented yeast proteins as vectors of weighted terms from all the PubMed articles mentioning their respective genes. They then trained a support vector machine (SVM) on protein-text-vectors, to distinguish among subcellular localizations. The performance was favorable when compared to a classifier trained on amino acid composition alone, but it was not compared against any state-of-the-art localization system, and the reported results do not suggest an improvement over earlier systems. Moreover, while their text-based classifier performed better than an amino acid composition classifier, combining the two forms of data did not significantly improve performance with respect to the text-based classifier alone.

Nair and Rost[2] used the text taken from Swiss-Prot annotations of proteins to represent these proteins, and trained a subcellular classifier using this representation. They concentrate on a few subcellular localizations, and report results that are compatible – but do not improve upon – the state of the art at that time. Their work was elaborated upon by Eskin and Agichtein[18], who added subsequences from the protein's amino acid sequence as part of the terms considered in the text representation. The system was not tested against existing systems or data sets, and the reported results do not indicate improvement over previous systems.

The best performing comprehensive systems reported so far, which were tested on a large set of proteins, are PLOC[19] and, more recently, MultiLoc[20]. While they report the best accuracy until now, on a broad range of organisms and localizations, there is still room for improvement.

The work reported here, similarly to that reported by Nair and Rost[2], uses Swiss-Prot as a text source. Unlike them though, we use the PubMed abstracts

referenced by Swiss-Prot, rather than the annotation text placed by Swiss-Prot curators. Furthermore, unlike Stapley *et al*. who use all abstracts that contain the gene name for the protein, we use only abstracts that are referenced by Swiss-Prot, and moreover, rather than use all the terms in them with a standard (TF*IDF[a]) weighting, as done by Stapley *et al*., we select terms based on a *distinguishing* criterion described in Section 2, and apply a probability-based weighting scheme. We train an SVM as a text-based classifier, and combine it with a sequence-based classifier, to produce a comprehensive subcellular categorizer. Our integrated system is tested on a number of publicly available, extensive, homology-reduced, data sets which were used for evaluating earlier systems (TargetP, PLOC, and MultiLoc). For each system, we first conduct a comparison using the same data and the same subcellular localizations as reported in the paper published about that system. We then conduct a test using all the proteins in Swiss-Prot for which a subcellular annotation is assigned, among the 11 localizations: chloroplast, cytoplasm, endoplasmic reticulum, extracellular space, Golgi apparatus, lysosome, mitochondria, nucleus, peroxisome, plasma membrane, and vacuole. On each of the data sets our system performs better than the state-of-the-art systems in terms of overall prediction accuracy, and other standard measures.

The next section outlines the methods used, while in Section 3 we demonstrate the performance of our system. Section 4 concludes and outlines future work.

2. Methods

Our system combines five separate classifiers, four sequence-based and one text-based. Their output is integrated through a sixth classifier to produce an improved prediction of protein subcellular localization. The sequence-based classifiers have been successfully used before by the MultiLoc system[20] and are briefly described below. Section 2.2 then presents the novel text-based method, while Section 2.3 explains how all these classifiers are combined to form an integrated prediction system. Four of the five classifiers are based on support vector machines (SVMs), using the LIBSVM implementation[21]. The latter supports soft, probabilistic categorization for n-class tasks[22], assigning to each classified item an n-dimensional vector denoting its probability to belong to each of the n classes. Radial Basis Function kernels were used throughout this study. Further details are given below.

2.1. *Sequence-based methods*

Each of the sequence-based classifiers utilizes a different approach to derive biologically informative features that can be used to predict localization, and classifies the input protein sequence to its respective localization using these features.

[a] An acronym for Term Frequency, Times Inverse Document Frequency.

Three of these classifiers are SVM-based. The fourth scans the protein sequences for short sequence motifs indicative of structure and function. The four classifiers are briefly described below (see the MultiLoc paper[20] for further details).

SVMTarget – This classifier uses the N-terminal targeting peptide (TP) to predict a few subcellular categories. It distinguishes among four plant (chloroplast (*ch*), mitochondria (*mi*), secretory pathway (*SP*), and other (*OT*)) and three non-plant (*mi, SP, OT*) localizations. The targeting peptides are represented by their partial amino acid composition, motivated by the observation that TPs for specific localizations have a similar amino acid composition while their actual sequence may differ. Given an input protein, the classifier outputs a three-dimensional vector (four-dimensional for plant) of class probabilities. SVMTarget alone demonstrated a slightly better performance than TargetP[1] in a comparative study[20].

SVMSA – Some proteins of the secretory pathway carry a signal anchor (SA) that, unlike the targeting peptide, is usually located further away from the N-terminus and contains a longer hydrophobic component. SVMSA can predict secretory pathway (*SP*) proteins that are hard to detect using SVMTarget. It is a binary classifier, trained to distinguish proteins carrying SA from those that do not. It outputs, given an input sequence, its probability to contain a signal anchor.

SVMaac – This method uses the whole protein amino acid composition (*aac*), and categorizes proteins into any of the possible localizations. It combines a collection of binary classifiers, each trained to distinguish one class from all others, although one classifier in the collection was especially trained to distinguish cytosolic (*cy*) from nuclear (*nu*) proteins, as these are hard to separate using the one-against-all approach. Given an input protein, p, with n possible localizations, the classifier outputs an n-dimensional probability vector containing p's probability to belong to each localization.

MotifSearch – Proteins from several subcellular localizations can be characterized by a few types of short sequence motifs, such as Nuclear Localization Signal and DNA-binding domains. The motifs were obtained from the PROSITE[23] and from the NLSdb[24,25] databases. This classifier outputs a discrete, binary vector, representing the presence (1) or the absence (0) of each type of motif in the query protein sequence.

2.2. *Text-based method*

The idea underlying the text-based classifier is the representation of each protein as a vector of weighted text features. While text-based localization has been presented before[2,17], the key differences between the current work and previous ones is in the text source used, the feature selection, and the term weighting scheme.

First, for each protein the text comes from the abstracts curated for the protein

in its Swiss-Prot entry. We used a script that scanned each protein in Swiss-Prot for all the PubMed identifiers occurring in its Swiss-Prot entry, and obtained the respective title and abstract[b] from PubMed. Each protein is thus assigned a set of PubMed abstracts, based on Swiss-Prot. This choice of abstracts is different from that of Stapley *et al.*[17] who used all the PubMed abstracts mentioning the gene's name, and from that of Nair and Rost[2] – who use Swiss-Prot annotation text rather than PubMed abstracts. The assigned abstracts are then tokenized into a set of terms, consisting of singleton and pairs of consecutive words, with a list of standard stop words excluded from consideration. The results reported here also include the application of Porter stemming[26] to all the words in the terms.

Second, from all the extracted terms, we select a subset of *distinguishing terms*. This is done by scoring each term with respect to each subcellular localization, where the score reflects the probability of the term to occur in abstracts that are associated with proteins of this certain localization. Intuitively, a term is *distinguishing* for a localization L, if it is much more likely to occur in abstracts associated with localization L than with abstracts associated with all other localizations. We formalize this idea in the following paragraphs.

Let t be a term, L a localization, and p a protein. If protein p is known to be localized in L, we denote this $p \in L$. We also define the following sets:

- The set of all PubMed abstracts associated with protein p according to Swiss-Prot, denoted D_p ;
- The set of all proteins known to be localized at L, denoted P_L ;
- The set of abstracts that are associated with a localization L, denoted D_L, is defined as: $D_L = \bigcup_{p \in P_L} \{d \,|\, d \in D_P\}$. It is the set of all the abstracts associated with the proteins that are in localization L. The number of documents in this set is denoted $|D_L|$.

The probability of a term t to be associated with a localization L, denoted Pr_L^t, is defined as the conditional probability of the term to appear in a document, given that the document is associated with the localization: $Pr_L^t = Pr(t \in d \,|\, d \in D_L)$. A maximum likelihood estimate for this probability is simply the proportion of documents containing t among all those associated with the localization: $Pr_L^t \approx (\#\text{ of documents } d \in D_L \ s.t.\ t \in d)/|D_L|$. For each term t and each localization L, the estimate for the probability Pr_L^t is calculated.

Based on this probability, a term t is called *distinguishing* for localization L, if and only if its probability to occur in localization L, Pr_L^t, is significantly different from its probability to occur in any other localization L', $Pr_{L'}^t$. The statistical test applied, uses the Z-score[27], which evaluates the difference between two binomial

[b] Without using any of the MeSH terms.

Table 1. Examples of distinguishing stemmed terms for several localizations

Localization	Example Terms
Nucleus	*bind, control, dna, histon, nuclear, promot, transcript*
Mitochondria	*coa (CoA), complex, cytochrom, dehydrogenas, mitochondri, oxidas, respiratori*
Golgi Apparatus	*acceptor, catalyt domain, fucosyltransferas, galactos, glycosyltransferas, golgi, transferas*
Endoplasmic Reticulum	*calcium, chaperon, disulfid isomeras, endoplasm, lumen, microsom, transmembran*

probabilities, Pr^t_L, and $Pr^t_{L'}$, as follows:

$$Z^t_{L,L'} = \frac{Pr^t_L - Pr^t_{L'}}{\sqrt{\overline{P}\cdot(1-\overline{P})\cdot\left(\frac{1}{|D_L|}+\frac{1}{|D_{L'}|}\right)}}\ , \text{ where } \overline{P} = \frac{|D_L|\cdot Pr^t_L + |D_{L'}|\cdot Pr^t_{L'}}{|D_L|+|D_{L'}|}\ .$$

When $|Z^t_{L,L'}| \geq 1.96$, the hypothesis that the two probabilities Pr^t_L, $Pr^t_{L'}$ are different is accepted with a confidence level greater than 95%. Therefore, if the term t has a localization L such that for any other localization L' $|Z^t_{L,L'}| \geq 1.96$, t is considered *distinguishing for localization* L, and is included in the set of distinguishing terms. In our representation of proteins as term vectors, we use only *distinguishing terms*. In the experiments described in Section 3, using several different proteins sets, the average number of PubMed abstracts is on the order of $10,000$, while that of distinguishing terms is about 800. Some examples of distinguishing terms for several localizations are shown in Table 1.

Finally, once the collection of N distinguishing terms, denoted as T_N, was established, each protein p is represented as an N-dimensional vector, where the weight $W^p_{t_i}$ at position i, (where $1 \leq i \leq N$), is the conditional probability of the term t_i to appear in the abstracts associated with the protein p, given all the PubMed abstracts related to the protein, (the set D_p) This probability is estimated as the ratio between the total number of times the term t_i occurs in the abstracts associated with the protein p and the total number of all the occurrences of distinguishing terms in these same abstracts. Formally it is calculated as:

$$W^p_{t_i} = \frac{\sum_{d\in D_p,\, s.t.\ t_i\in d} (\#\text{ of times } t_i \text{ occurs in } d)}{\sum_{d\in D_p}\sum_{t_j\in T_N} (\#\text{ of times } t_j \text{ occurs in } d)}\ ,$$

where the sums are taken over all the abstracts d in the set of abstracts associated with the protein p, D_p.

The representation of proteins as weighted term vectors, is then partitioned into training and test sets for each subcellular localization, and as before, an SVM is trained to classify these protein vectors into their respective localization. This classifier, like SVMacc described above, produces an n-dimensional probability vector denoting the probability of the protein to be in each of the n localizations.

2.3. *Integrated method*

The output from the five classifiers above, is a set of four probability vectors and one binary-valued vector (resulting from MotifSearch). These are all concatenated to form one integrated feature vector for each protein. Again, an SVM classifier is trained on these feature vectors to produce a prediction. This classifier consists of a set of one-against-one classifiers (each of which distinguishes between a pair of localizations) and its output, yet again, is a probabilistic vector, holding for each localization the probability of the protein to belong to it. Based on this final classification step, a protein is assigned to the localization with the highest probability value in the last output vector. The training and evaluation procedure uses strict five-fold cross-validation, where no test protein was used to train any of the classifiers comprising the system.

3. Experiments and Results

To train and to evaluate our integrated system, we used three different data sets, namely those used for training and testing TargetP, MutliLoc, and PLOC. These sets provide the basis for an extensive and sound comparison. The data sets, the evaluation procedure, and the results are described throughout this section.

3.1. *Experimental setting*

The data sets used in our experiments are the following:

TargetP – This data set[1] contains a total of 3,415 distinct proteins representing four plant (*ch*, *mi*, *SP*, and *OT*) and three non-plant (*mi*, *SP*, and *OT*) localizations. Homologs were removed from it by the TargetP authors. The *SP* category includes proteins from several localizations in the secretory pathway: endoplasmic reticulum (*er*), extracellular space (*ex*), Golgi apparatus (*go*), lysosome (*ly*), plasma membrane (*pm*), and vacuole (*va*). The *OT* category includes *cy* and *nu* proteins.

MultiLoc – The MultiLoc data set[20] contains a total of 5,959 protein sequences, which were extracted from the Swiss-Prot database release 42.0[28]. Animal, fungal, and plant proteins with an annotated subcellular localization[c] were grouped into eleven eukaryotic localizations: *cy*, *ch*, *er*, *ex*, *go*, *ly*, *mi*, *nu*, peroxisome (*pe*), *pm*, *va*. In the experiments reported here homologous proteins with identity higher than 80%, (the same threshold used by PLOC[19]), were excluded from the set, to avoid the occurrence of highly similar sequences in both the training and the test sets[d]. Further details about the data set extraction and the implications of homology reduction are available in the MultiLoc publication[20].

[c]Excluding proteins whose annotation was commented *by similarity* or *potential*.

[d]We also conducted experiments with a more lenient and more stringent homology constraints, of 90% and 40% identity, respectively (data not shown).

PLOC – The PLOC data set was used by Park and Kanehisa[19] and consists of proteins extracted from Swiss-Prot release 39.0, covering 12 localizations. In contrast to MultiLoc, (aside for the older Swiss-Prot version), this data set introduces an additional category within the *cy* proteins, namely, the cytoskeleton (*cs*). There are 41 *cs* proteins, compared to 1,245 *cy* proteins. The total number of sequences is 7,579 (max. sequence identity 80%). This set is larger than the MultiLoc data set due to a less restrictive data extraction, assigning proteins to localization even when the localization annotation includes the words "potential" or "by similarity".

Using these three data sets, the performance of our integrated system is compared to that of TargetP, PLOC, and MultiLoc[e]. In addition, we also compare the performance of the integrated system to that of an SVM classifier applied to the text data alone. Following previous evaluations[1,19], we consistently employ five-fold cross-validation. For comparison against the PLOC data set we use the same split as the one used by Park and Kanehisa[19]. For the TargetP data, as the split used by Emanuelsson *et al.*[1] was not provided, we ran the five-fold cross-validation procedure five times, each using a different randomized five-way split, to ensure robustness. The reported results are averaged over all the 5 folds, and over the 5 randomized splits when those are used.

Since the performance of previous systems[1,19] was evaluated using several different metrics, for a fair comparison we calculated these same performance measures. Thus, for each system and data set the performance is measured, for each localization, in terms of the sensitivity (*Sens*), specificity (*Spec*), and Matthews correlation coefficient (*MCC*)[29]. These are defined as:

$$Sens = \frac{TP}{TP+FN}, \quad Spec = \frac{TP}{TP+FP}, \quad \text{and}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN) \cdot (TP+FP) \cdot (TN+FN) \cdot (TN+FP)}} \ ,$$

where TP, TN, FP, FN denote the number of true positives, true negatives, false positives, and false negatives, respectively, with respect to a given localization. Like Park and Kanehisa[19] we also measure the *overall accuracy*, namely, $Acc = C/N$, where C is the number of correctly classified proteins over all the localizations, and N is the total number of classified proteins. They also measured the *average sensitivity*, over all the localizations, a metric they call *local accuracy*, which we calculate as well. This last measure, which we denote as *Avg*, gives an equal weight to the categorization performance on each localization, regardless of the number of proteins known to be associated with it.

[e]Comparison to PSort[11] is not included here, since MultiLoc has already demonstrated a higher prediction accuracy compared to this method[20].

3.2. *Results*

We present the results of running the sequence-based system, MultiLoc, the text-based classifier alone (denoted *Text*), and the integrated system (denoted *MultiLocText*), on all the three data sets. For completeness, we also present the results reported by the authors of PLOC[19] and of TargetP[1] on the respective data sets. These numbers were directly taken from the respective publications.

Table 2 summarizes the results, showing the overall accuracy (*Acc*) and the average local accuracy (*Avg*) for both the TargetP and the PLOC data sets. For TargetP the results are shown for plant and non-plant proteins, while for PLOC results are shown for plant, animal, and fungal proteins. Table 3 compares the performance of TargetP and PLOC with our integrated system, with respect to the individual subcellular localizations.

Table 2. An overview of the prediction results using the TargetP and PLOC data sets. Both the total (*Acc*) and the average (*Avg*) prediction accuracies are shown for all the methods. The highest values appear in bold. Standard deviations, (denoted ±) are provided where available.

Data set	Method	*Acc* [%] (± Standard Deviation) / *Avg* [%] (± Standard Deviation)		
TargetP		**Plant**	**Non-Plant**	
	TargetP	85.3 (±3.5) / 85.6 (n/a)	90.0 (±0.7) / 90.7 (n/a)	
	MultiLoc	89.7 (±1.6) / 90.2 (±2.0)	92.5 (±1.2)/ 92.8 (±1.1)	
	Text	81.2 (±2.6) / 78.1 (±3.2)	88.7 (±1.1)/ 89.8 (±1.6)	
	MultiLocText	**94.7** (±1.5) / **94.4** (±1.6)	**96.2** (±0.8) / **96.7** (±0.9)	
PLOC		**Plant**	**Animal**	**Fungal**
	PLOC	78.2 (±0.9)/ 57.9 (±2.1)	79.6 (±0.9)/ 59.9 (±3.3)	79.5 (±0.9)/ 56.8 (±1.9)
	MultiLoc	73.6 (±0.7) / 71.3 (±2.8)	76.0 (±0.7) / 73.6 (±3.9)	75.8 (±0.8) / 72.5 (±2.5)
	Text	68.7 (±0.7) / 73.5 (±1.8)	70.2 (±0.7) / 75.5 (±2.7)	67.8 (±0.5) / 72.4 (±2.6)
	MultiLocText	**85.3** (±1.2) / **84.2** (±2.4)	**86.4** (±0.8) / **84.5** (±3.6)	**85.4** (±0.8) / **83.8** (±2.8)

Table 3. Localization specific results using the TargetP (left), and the PLOC (right) data sets. For both sets, the results reported in the respective papers are compared to results of our integrated system (MultiLocText). As PLOC localization-specific results are averaged over all three organisms, we show such averaged results for our system as well. Specificity and MCC values were not available for PLOC, hence only its *Sensitivity* is listed and compared with our sensitivity values. The highest compared values for each data set are shown in bold.

TargetP Data Set			**PLOC Data Set**		
Loc	**TargetP**	**MultiLocText**	**Loc**	**PLOC**	**MultiLocText**
	Plant (*Sens Spec MCC*)			**Avg. *Sens***	**Avg. (*Sens Spec MCC*)**
ch	0.85 0.69 0.72	**0.93 0.89 0.89**	*ch*	0.72	**0.84** 0.83 0.82
mi	0.82 0.90 0.77	**0.95 0.99 0.95**	*mi*	0.57	**0.85** 0.85 0.83
OT	0.85 0.78 0.77	**0.95 0.87 0.89**	*cs*	0.59	**0.83** 0.26 0.46
SP	0.91 0.95 0.90	**0.95 0.98 0.95**	*cy*	0.72	**0.79** 0.78 0.74
	Non-Plant (*Sens Spec MCC*)		*er*	0.47	**0.86** 0.71 0.78
mi	0.89 0.67 0.73	**0.97 0.88 0.91**	*ex*	0.78	**0.88** 0.91 0.88
OT	0.88 0.97 0.82	**0.95 0.99 0.93**	*go*	0.15	**0.82** 0.30 0.49
SP	0.96 0.92 0.92	**0.98 0.96 0.96**	*nu*	**0.90**	0.88 0.94 0.88
			pe	0.25	**0.81** 0.63 0.71
			pm	**0.92**	0.89 0.98 0.91
			va	0.25	**0.83** 0.28 0.48
			ly	0.62	**0.81** 0.52 0.64

A comparison of the performance of our three systems (MultiLoc alone, Text alone, and the integrated MultiLocText) using five-fold cross-validation over the

5,959 proteins of the MultiLoc data set, is presented in Table 4. The sensitivity (*Sens*), specificity (*Spec*), and Matthews *MCC* values for the plant and animal versions are listed. (Similar results were obtained for the fungal version, and are not shown here due to space limitation).

The results in Tables 2, 3, and 4 clearly show that the combined classifier, which integrates text and sequence data, outperforms earlier prediction methods. It also outperforms its own text-based (Text) and sequence-based (MultiLoc) components, if taken separately. A significance test was performed to evaluate the differences between the values obtained from MultiLocText and those obtained from each of MultiLoc and Text alone, (Table 4). The improved performance values of MultiLocText are highly statistically significant ($p \ll 0.05$), for almost all the subcellular localizations. The only exceptions are the Golgi (*go*, animal and plant), where there is no significant difference in sensitivity with respect to text-alone, as well as the peroxisome predictions (*pe*, animal and plant), where MultiLocText does not outperform the text-alone system.

4. Discussion and Conclusion

The methods, experiments, and results presented here clearly demonstrate a significant improvement in the prediction of protein subcellular localization through the integration of sequence- and text-based methods. Table 4 shows that the two

Table 4. Prediction performance of MultiLoc, Text, and MultiLocText on the MultiLoc data set. Both localization-specific values (*sens, spec,* MCC) and overall results (*Acc* and *Avg*) are shown. Highest values appear in bold.

Loc	MultiLoc	Text	MultiLocText
	Plant (*Sens Spec MCC*)		
ch	0.88 0.85 0.85	0.89 0.70 0.78	**0.94 0.91 0.92**
cy	0.68 0.85 0.70	0.53 0.75 0.54	**0.81 0.91 0.82**
er	0.72 0.54 0.61	0.73 0.55 0.62	**0.82 0.63 0.71**
ex	0.68 0.81 0.70	0.74 0.80 0.73	**0.84 0.90 0.84**
go	0.75 0.41 0.54	0.82 0.42 0.57	**0.84 0.61 0.70**
mi	0.85 0.81 0.80	0.80 0.80 0.78	**0.90 0.88 0.88**
nu	0.82 0.75 0.75	0.80 0.72 0.72	**0.89 0.85 0.85**
pe	0.71 0.34 0.47	**0.88 0.71 0.79**	0.85 0.59 0.70
pm	0.74 0.89 0.77	0.80 0.91 0.82	**0.84 0.96 0.87**
va	0.70 0.20 0.36	0.59 0.15 0.29	**0.83 0.29 0.48**
***Acc* [%]**	74.6	73.1	**85.1**
***Avg* [%]**	75.2	76.0	**85.5**
	Animal (*Sens Spec MCC*)		
cy	0.67 0.85 0.68	0.51 0.77 0.53	**0.83 0.91 0.82**
er	0.68 0.56 0.60	0.74 0.48 0.58	**0.82 0.67 0.73**
ex	0.79 0.83 0.77	0.76 0.78 0.72	**0.86 0.90 0.86**
go	0.71 0.43 0.53	0.86 0.40 0.57	**0.87 0.65 0.74**
ly	0.69 0.36 0.48	0.75 0.32 0.47	**0.86 0.55 0.68**
mi	0.88 0.82 0.83	0.80 0.79 0.77	**0.93 0.91 0.91**
nu	0.82 0.73 0.73	0.84 0.71 0.73	**0.89 0.83 0.84**
pe	0.71 0.31 0.44	**0.93** 0.60 0.74	0.89 **0.68 0.77**
pm	0.73 0.90 0.76	0.80 0.91 0.81	**0.85 0.95 0.87**
***Acc* [%]**	74.6	72.5	**86.2**
***Avg* [%]**	74.1	77.5	**86.8**

types of methods distinctly complement each other. MultiLoc, which is based on sequence data, typically performs well predicting protein localizations that are directed by N-terminal signals such as the mitochondria and the chloroplast. The use of text information complements and significantly boosts its performance for localizations whose sequence-based signal is not as overt, including the peroxisome and localizations related to the secretory pathway such as the Golgi apparatus and the endoplasmic reticulum.

In this work we have demonstrated, using five-fold cross-validation, that our system can reproduce, with unprecedented sensitivity and specificity, localizations of proteins which were already annotated in Swiss-Prot. A natural next step is to apply the method to yet un-localized proteins. We are developing the means to predict subcellular localization of proteins for which PubMed reference exist in Swiss-Prot but no localization assigned, as well as for those with no curated PubMed reference. Our current use of "raw text" from PubMed abstracts (in contrast, for instance, to the use of Swiss-Prot annotation text as was done before[2]), is expected to make our approach amenable to such extensions. We are also investigating methods for the localization of proteins with no PubMed references, through the use of alternative data sources.

References

1. Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G.: Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. **300** (2000) 1005–1016
2. Nair, R., Rost, B.: Inferring sub-cellular localization through automated lexical analysis. Bioinformatics **18** (2002) S78–S86
3. Gardy, J.L., Spencer, C., Wang, K. *el al.*: PSORT-B: Improving protein subcellular localization prediction for gram-negative bacteria. Nucleic Acids Research **31** (2003) 137–140
4. Cai, Y.D., Chou, K.C.: Predicting 22 protein localizations in budding yeast. Biochem Biophys Res Commun. **323** (2004) 425–428
5. Schneider, G., Fechner, U.: Advances in the prediction of protein targeting signals. Proteomics **4** (2004) 1571–1580
6. Dönnes, P., Höglund, A.: Predicting Protein Subcellular Localization: Past, Present, and Future. Genomics, Proteomics, and Bioinformatics **2** (2004)
7. Burns, N., Grimwade, B., Ross-Macdonald, P., Choi, E., Finberg, K., GS, R., M, S.: Large-scale analysis of gene expression, protein localization and gene disruption in Saccharomyces cerevisiae. Genes and Development **8** (1994) 1087–1105
8. Hanson, M.R., Köhler, R.H.: GFP imaging: Methodology and application to investigate cellular compartmentation in plants. Journal of Experimental Botany **52** (2001)
9. Dunkley, T., Watson, R., Griffin, J., Dupree, P., Lilley, K.: Localization of organelle proteins by isotope tagging (LOPIT). Molecular and Cellular Proteomics **3** (2004)
10. Nakai, K., Kanehisa, M.: Expert system for predicting protein localization sites in gram-negative bacteria. Proteins: Structure, Function and Genetics **11** (1991) 95–110

11. Nakai, K., Kanehisa, M.: A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics. **14** (1992) 897–911
12. Horton, P., Nakai, K.: A probabilistic classification system for predicting the cellular localization of proteins. In: Proc. of the Int. Conf. on Intelligent Systems for Molecular Biology (ISMB). (1996)
13. Horton, P., Nakai, K.: Better prediction of protein cellular localization sites with the k nearest neighbors classifier. In: Proc. of the Int. Conf. on Intelligent Systems for Molecular Biology (ISMB). (1997)
14. Emanuelsson, O., Nielsen, H., von Heijne, G.: Chlorop, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Science **8** (1999) 978–984
15. Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S.: Extensive feature detection of N-terminal protein sorting signals. Bioinformatics. **18** (2002) 298–305
16. Nair, R., Rost, B.: Mimicking cellular sorting improves prediction of subcellular localization. J Mol Biol. **348** (2005) 85–100
17. Stapley, B.J., Kelley, L.A., Sternberg, M.J.E.: Predicting the subcellular location of proteins from text using support vector machines. In: Proc. of the Pacific Symposium on Biocomputing (PSB). (2002) 374–385
18. Eskin, E., Agichtein, E.: Combining text mining and sequence analysis to discover protein functional regions. In: Proc. of the 9th Pacific Symposium on Biocomputing (PSB). (2004) 288–299
19. Park, K.J., Kanehisa, M.: Prediction of protein subcellular location by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics. **19** (2003) 1656–1663
20. Höglund, A., Dönnes, P., Blum, T., Adolph, H., Kohlbacher, O.: Using N-terminal targeting sequences, amino acid composition, and sequence motifs for predicting protein subcellular localization. German Conference on Bioinformatics (GCB) 2005.
21. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines (2003) *http://www.csie.ntu.edu.tw/∼clin/libsvm/*.
22. Wu, T.F., Linand, C.J., Weng, R.C.: Probability Estimates for Multi-class Classification by Pairwise Coupling. Journal of Machine Learning Research **5** (2004) 975–1005
23. Bairoch, A., Bucher, P.: PROSITE: recent developments. Nucleic Acids Res. **22** (1994) 3583–3589
24. Cokol, M., Nair, R., Rost, B.: Finding nuclear localization signals. EMBO Rep. **1** (2000) 411–415
25. Nair, R., Carter, P., Rost, B.: NLSdb: database of nuclear localization signals. Nucleic Acids Res. **31** (2003) 397–399
26. Porter, M.F.: An Algorithm for Suffix Stripping (Reprint). In: Readings in Information Retrieval. Morgan Kaufmann (1997) *http://www.tartarus.org/∼martin/PorterStemmer/*.
27. Walpole, R.E., Myers, R.H., Myers, S.L. In: One- and Two-Sample Tests of Hypotheses. (1998) 235–335
28. Bairoch, A., Apweiler, R.: The SWISS-PROT protein sequence database and its supplement in TrEMBL in 2000. Nucleic Acids Res. **28** (2000) 45–48
29. Matthews, B.W.: Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. **405** (1975) 442–451