Pacific Symposium on Biocomputing 13:604-615(2008)

EPILOC: A (WORKING) TEXT-BASED SYSTEM FOR PREDICTING PROTEIN SUBCELLULAR LOCATION

SCOTT BRADY AND HAGIT SHATKAY

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

Motivation: Predicting the subcellular location of proteins is an active research area, as a protein's location within the cell provides meaningful cues about its function. Several previous experiments in utilizing text for protein subcellular location prediction, varied in methods, applicability and performance level. In an earlier work we have used a preliminary text classification system and focused on the integration of text features into a sequence-based classifier to improve location prediction performance.
Results: Here the focus shifts to the text-based component itself. We introduce *EpiLoc*, a comprehensive text-based localization system. We provide an in-depth study of text-feature selection, and study several new ways to associate text with proteins, so that text-based location prediction can be performed for practically any protein. We show that EpiLoc's performance is comparable to (and may even exceed) that of state-of-the-art sequence-based systems. EpiLoc is available at: *http://epiloc.cs.queensu.ca*.

1. Introduction

Knowing the location of proteins within the cell is an important step toward understanding their function and their role in biological processes. Several experimental methods, such as those based on green fluorescent proteins or on immunolocalization, can identify the location of proteins. Such methods are accurate, but slow and labour-intensive, and are only effective for proteins that can be readily expressed and produced within the cell.

Given the large number of proteins about which little is known, and that many of these proteins may not even be expressed under regular conditions – it is important to be able to computationally infer protein location based on readily available data (e.g. amino acid sequence). Once effective information is computationally elucidated outside the lab, well-targeted lab experiments can be judicially performed. For well over a decade many computational location-prediction methods were suggested and used, typically relying on features derived from sequence data[7,9,12,13].

Another type of information that can assist in location prediction is derived from text. One option is to explicitly extract location statements from the literature[6]. While this approach offers a way to access pre-existing knowledge, it does not support prediction. An alternative *predictive* approach is to employ classifiers using text-features that are derived from literature discussing the proteins. These features may not state the location, but their relative frequency in the text associated with a certain protein is often correlated with the protein's location. Examples of this approach include work by Nair and Rost[11] and by

Stapley *et al*[17]. They represent proteins using text-features taken from annotations[11] or from PubMed abstracts in which the protein's name occur[17], and train classifiers to distinguish among proteins from different locations. The main limitations of this earlier work are: a) It was not shown to meet or improve upon the performance of state-of-the-art systems. b) The systems depended on an explicit source of text; in its absence many proteins cannot be localized.

In an earlier work[8,16] we studied the integration of text features into a sequence-based classifier[9], showing significant improvement over state-of-the-art location prediction systems. The text component was a preliminary one, and was not studied in detail. Here we provide an in-depth study and description of a new and complete text-based system, *EpiLoc*. We compare several text-feature selection methods, and extensively compare the performance of this system to other location prediction systems. Moreover, we introduce several alternative ways to associate text with proteins, making the system applicable to practically any protein, even when text is not available from the preferred primary source. Further details about the differences between the preliminary version[8,16] and EpiLoc are given in the complete report of the work[3].

While our work focuses on protein subcellular localization, the ideas and methods, including the study of feature selection and of ways for associating text with biological entities, are applicable to other text-related biological enquiries.

In Section 2 we introduce the methods for associating text with proteins, and the way in which text is used to represent proteins. Section 3 focuses on feature selection methods, while Sections 4 and 5 describe our experiments and results, demonstrating the effectiveness of the proposed methods.

2. Data and Methods

EpiLoc is based on the representation of each protein as an *N*-dimensional vector of weighted text features, $< w_1^p \ldots w_N^p >$. Each position in the vector represents a term from the literature associated with the proteins. As not all terms are useful for predicting subcellular location, and to save time and space, feature selection is employed to obtain *N* terms, as discussed in Section 3. Here we describe our primary method for associating text with individual proteins and our term-weighting scheme. We also present three alternative methods that assign text to proteins when the primary method cannot do so.

Primary Text Source: The literature associated with the whole protein dataset is the collection of text related to the individual proteins. For training EpiLoc, text per protein is taken from the set of PubMed abstracts referenced by the protein's Swiss-Prot[2] entry. Abstracts associated with proteins from three or more subcellular locations are excluded, as their terms are unlikely to effectively characterize a single location. Each protein is thus associated with a set of

authoritative abstracts, as determined by Swiss-Prot curators. As we noted before[16], the abstracts do not typically discuss localization – but rather are authoritative with respect to the protein in general. This choice of text is more specific than that of Stapley *et al.*[17], who used *all* abstracts containing a protein's gene name. Moreover, unlike Nair and Rost[11], who used Swiss-Prot *annotation text* rather than referenced abstracts, our choice is general enough to assign text to the majority of proteins, allowing the method to be broadly applicable.

The text in each abstract is tokenized into a set of terms, consisting of singletons and pairs of consecutive words; a list of standard stop words[a] is removed, and Porter stemming[14] is then applied to all the words in this set. Last, terms occurring in fewer than three abstracts or in over 60% of all abstracts are removed; very rare terms cannot be used to represent the majority of the proteins in a dataset, while overly frequent terms are unlikely to have a discriminative value. The resulting term set typically contains more than 20,000 terms, and is reduced through a feature selection step (see Section 3). The feature-selection process produces a set of *distinguishing terms* for each location, that is, terms that are more likely to be associated with proteins within a certain location than with proteins from other locations. The combined set of all distinguishing terms forms the set of terms that we use to represent proteins, as discussed next.

Term Weighting: Given the set of N distinguishing terms, each protein p, is represented as an N-dimensional weight-vector, where the weight $W_{t_i}^{p}$ at position i, ($1 \leq i \leq N$), is the probability of the distinguishing term t_i to appear in the set of abstracts known to be associated with protein p, denoted D_p. This probability is estimated as the total number of occurrences of term t_i in D_p divided by the total number of occurrences of *all* distinguishing terms in D_p. Formally $W_{t_i}^{p}$ is calculated as: $W_{t_i}^{p} = (\#\text{ of times } t_i \text{ occurs in } D_p) / \Sigma_j (\#\text{ of times } t_j \text{ occurs in } D_p)$, where the sum in the denominator is taken over all terms t_j in the set of distinguishing terms T_N.

Once all the proteins in a set have been represented as weighted term vectors, the proteins from each subcellular location are partitioned into training and test sets, and a classifier is trained to assign each protein to its respective location. Our classifier is based on the LIBSVM[5] implementation of support vector machines (SVMs). LIBSVM supports soft, probabilistic categorization for n-class tasks, where each classified item is assigned an n-dimensional vector denoting the item's probability to belong to each of the n classes. Here n is the number of subcellular locations.

Alternative Text Sources: As pointed out by Nair and Rost[11], the text needed to represent a protein is not always readily available. In our case, some proteins

[a] Stop words are terms that occur frequently in text but typically do not bear content, such as prepositions.

may not have PubMed identifiers in their Swiss-Prot entry, and others – newly discovered proteins – may not even have a Swiss-Prot entry. We refer to such proteins as textless, and propose three methods to assign them with text.

HomoLoc **–** In previous work[16], if a textless protein had a homolog with associated text, we used the text of the homolog to represent the textless protein. Homoloc extends this idea to consider multiple homologs and re-weight terms accordingly. A BLAST[1] search identifies the set of homologs, and we retain those that share at least 40% sequence identity with the textless protein. (This level of similarity was chosen based on a study by Brenner et al.[4,3]). The retained homologs are then ranked in ascending order according to their E-value, and the set of abstracts associated with the top three homologs are associated with the textless protein. To reflect the degree of homology in the term vector representation, a modified weighting scheme is used where the number of times each term occurs in the abstracts associated with a homolog is multiplied by the percent identity between the homolog and the textless protein. Formally, the modified weight is calculated as:

$$W_{t_i}^{\ p} = \frac{\sum_{h \in H} (\#\ \text{of occurences of}\ t_i\ \text{in}\ D_h) \cdot (\%\ \text{identity of}\ h)}{\sum_{h \in H} \sum_{t_j \in T_N} (\#\ \text{of occurences of}\ t_j\ \text{in}\ D_h) \cdot (\%\ \text{identity of}\ h)},$$

where h is a homolog, D_h is the set of abstracts associated with h, and a sum is taken over all the homologs in the set of homologs H.

DiaLoc – Proteins are most likely to be textless when they have just recently been sequenced/identified, as little information about them exists in databases such as PubMed or Swiss-Prot. When no close homologs with assigned text are known, HomoLoc cannot be used. The most reliable source of information for such proteins (and the one most likely to be interested in their localization) is the scientist researching the proteins. A user interface (shown in Fig. 2), allows a researcher to type her own short description of the protein based on the current state of knowledge. This description is used as the text associated with the textless protein. DiaLoc is meant to be used as an interactive tool for researchers concerned with individual proteins, and not as a large-scale annotation tool.

PubLoc[b] – Proteins whose Swiss-Prot entries do not contain reference to PubMed may still have PubMed abstracts discussing them. To check if such abstracts exist, the name of the textless protein and its gene are extracted from the Swiss-Prot entry. A query consisting of an *OR*-delimited list of these names is posed to PubMed. The five most recent abstracts returned are used as the protein's text source. This is a simple selection criterion and can be further improved upon.

[b] We thank Annette Höglund for suggesting this name.

To select the preferred method for handling textless proteins for large-scale annotation, we compared HomoLoc's and PubLoc's performance on the 614 textless proteins of the MultiLoc dataset (see Section 4). A complete discussion of these experiments is beyond the scope of this paper and is provided elsewhere[3]; we briefly summarize them here. We trained EpiLoc on all the proteins in the MultiLoc dataset that *do* have associated text. We then represented the remaining textless proteins using both PubLoc and HomoLoc, and classified them using the trained system. The overall accuracy obtained (for these 614 proteins) using HomoLoc is 73% for plant and 76% for animal. Using PubLoc the accuracy dropped to 57% and 64%, respectively[c]. As PubLoc is clearly less effective than HomoLoc, it is only applied in cases where neither HomoLoc nor DiaLoc can be used. HomoLoc is thus our method of choice for handling textless proteins, and is further discussed in Section 4.

3. Feature Selection

As stated in Section 2, each protein is represented as a weight-vector defined with respect to a set of *distinguishing terms*. Using a set of selected features can improve performance (even when SVMs are used) and reduces computational time and space. Intuitively, a term t is distinguishing for a location L, if its likelihood to occur in text associated with location L is significantly different from that of occurring in text associated with all other locations. To compare these likelihoods, for each location we assign to each term a score reflecting its probability to occur in the abstracts associated with the location. We formalize this method, referred to as the *Z-Test* method, in Section 3.1, and compare it with several alternatives in Section 3.2.

3.1. *The Z-Test Method*

Let t be a term, p a protein, and L a location. A protein, p, localized to L, is denoted $p \in L$ and has a set of associated abstracts, denoted D_p. The set of all proteins known to be localized to L is denoted P_L. We denote by D_L the set of abstracts associated with location L, (i.e. all abstracts associated with the proteins localized to L). Formally, this set is defined as: $D_L = \bigcup_{p \in P_L} \{d | d \in D_p\}$, and the number of abstracts in this set is denoted $|D_L|$. The probability of term t to be associated with location L, denoted $Pr(t|L)$, is defined as the conditional probability of t to appear in an abstract d, given that d is associated with location L. This probability is expressed as: $Pr(t|L) = Pr(t \in d | d \in D_L)$. Its maximum likelihood estimate is the proportion of abstracts containing the term t among all abstracts associated with L: $Pr(t|L) \approx$ *(# of abstracts* $d \in D_L$ *such that* $t \in d$*)* $/ |D_L|$. We calculate

[c] We also tested simpler versions of these methods (including the single-homolog method we tried in the past[16]); these were not as effective as the methods presented here[3].

the probability *Pr(t|L)* for each term *t* and location *L*.

Based on the above formulation, a term *t* is considered *distinguishing* for location *L*, if and only if its probability to occur in abstracts associated with *L*, *Pr(t|L)*, is significantly different from its probability to occur in abstracts associated with any other location *L'*, Pr*(t|L')*. To determine the significance of the difference between the two probabilities, a statistical test is employed that utilizes a *Z*-score[18]. The test evaluates the difference between two binomial probabilities, Pr*(t|L)* and Pr*(t|L')*, by calculating the following statistics:

$$Z^{t}_{L,L'} = \frac{\Pr(t|L) - \Pr(t|L')}{\sqrt{\overline{P}\cdot(1-\overline{P})\cdot\left(\frac{1}{|D_L|}+\frac{1}{|D_{L'}|}\right)}}, \quad \text{where } \overline{P} = \frac{|D_L|\cdot\Pr(t|L) + |D_{L'}|\cdot\Pr(t|L')}{|D_L| + |D_{L'}|}$$

The higher the absolute value $\left|Z^{t}_{L,L'}\right|$, the greater is the confidence level that the difference between Pr*(t|L)* and Pr*(t|L')* is statistically significant. Therefore, we consider a term *t* as distinguishing for location *L* if for any other location *L'*, the score $\left|Z^{t}_{L,L'}\right|$ is greater than a predetermined threshold. Table 1 shows examples of distinguishing terms for several locations; note that the terms do not necessarily state the location, but are merely correlated with it. The precise threshold selected was based on the experiment described next.

3.2. *Feature Selection Comparison*

To determine the effectiveness of the *Z-Test* method, we compare it to four standard feature selection methods: *odds ratio* (*OR*), *Chi-squared* (χ^2), *mutual information* (*MI*), and *information gain* (*IG*)[15]. We also compare it to the *Entropy* method, used by Nair and Rost[11]. Each of the four standard methods attempts to quantify how well a term represents a location by scoring a term *t* with respect to a location *L*. The total score for a term is then calculated as a combination of its location-specific scores. Following previous evaluations[15,20], to calculate the total *OR* and the *IG* scores we *sum* the term's scores over all locations, and to calculate the *MI* and χ^2 scores we take the *maximum* score for the term with respect to all locations. The *Entropy* method[11] scores terms with respect to locations, based on the difference between their Shannon information and the maximum attainable information.

To compare among the different feature selection methods we calculated the overall accuracy achieved by classifiers based on each method, on both plant and animal proteins of the MultiLoc dataset. For each of the methods, we used the same text pre-processing and partitioning of the data for five-fold cross-validation. Each of the six methods was evaluated based on its performance over a range of possible number of selected terms (ranging from 500 to 4,000).

Figure 1 shows the overall location prediction accuracy as a function of the number of selected terms for plant proteins. Similar results were obtained for

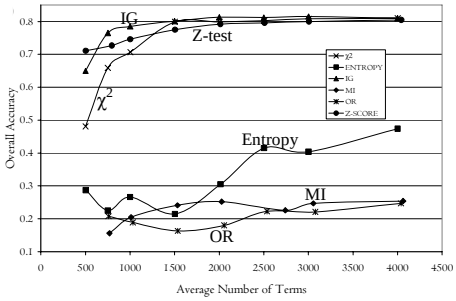

Figure 1. Accuracy of the classifiers (for plant proteins), based on different feature selection methods, as a function of the average number of selected terms (features).

Table 1. Stemmed Distinguishing terms.

Loc.	Example Terms
nu	*bind, base pair, chromatin, DNA*
mi	*acyl coa, cytochrom, electron transport*
go	*acceptor, galactos, golgi, transferase*
Er	*chaperon, disulfid isomeras, endoplasm*

Table 2. The threshold (and confidence level) chosen for each organism and dataset.

Dataset	Organism	**Threshold** [Confidence]
TargetP	Plant	1.645 [90%]
	Non-Plant	2.576 [99%]
PLOC	Plant	1.150 [75%]
	Animal	1.150 [75%]
MultiLoc	Plant	1.282 [80%]
	Animal	1.645 [90%]

animal proteins[3]. The figure demonstrates that the performance of the *Z-Test*, *IG*, and χ^2 methods is almost equivalent, and any of them could have been used by our classifier with similar results. We use the *Z-Test* in our experiments as this was our original approach[8,16] and it has a simple statistical interpretation. In contrast, the performance of the *MI*, *OR*, and *Entropy* methods is not as good. *MI*'s poor performance relative to that of both *IG* and χ^2 was expected, as it has been noted in previous research[20]. The *Entropy* method was originally developed to select features from a relatively small set of potential features compared to the set used here; Nair and Rost used only the *functional keywords in Swiss-Prot annotations* of the proteins, whereas we use a much larger number of potential features. As such, the relatively poor performance of the *Entropy* method shown here is not surprising. Conversely, we expected better results from *OR*. Its poor performance appears to be the result of its preferential selection of terms that occur in the abstracts associated with only a single location, leading to very sparse term vector representations for most proteins (a detailed discussion is provided elsewhere[3]). As mentioned above, we used this experiment as a guide for setting the threshold on the *Z*-score. For each dataset, we place a lower bound of 1.15 on the threshold, and set it to retain about 2,000 terms, as this number attains a balance between a computationally effective feature-space, and classification accuracy. As Figure 1 shows, the accuracy of the top methods does not significantly improve by including over 2,000 features. Table 2 shows the *Z*-score threshold used for each organism in each of the datasets described below.

4. Experimental Setting

EpiLoc was extensively evaluated, and compared to three state-of-the-art prediction systems – TargetP, PLOC, and MultiLoc – using the respective datasets that were used to train and test these systems. HomoLoc's performance is evaluated on the MultiLoc dataset. The datasets and evaluation procedures are

described throughout this section.

The following three datasets are used in our comparative study:

TargetP[7] – A total of 3,415 proteins, sorted into four plant (*ch*, *mi*, *SP*, and *OT*) and three non-plant (mi, *SP*, and *OT*) locations. The *SP* (Secretory Pathway) class includes proteins from the endoplasmic reticulum (*er*), extracellular space (*ex*), Golgi apparatus (*go*), lysosome (*ly*), plasma membrane (*pm*), and vacuole (*va*); the *OT* (Other) class includes cytoplasmic (cy) and nuclear (nu) proteins.

MultiLoc[9] – The MultiLoc dataset consists of 5,959 proteins extracted from Swiss-Prot release 42.0. Animal, fungal, and plant proteins with annotated subcellular locations were collected and sorted into eleven locations: *ch*, *cy*, *er*, *ex*, *go*, *ly*, *mi*, *nu*, *pe*, *pm*, and *va*. Proteins with a sequence identity greater than 80% were excluded from the dataset, as were any proteins whose subcellular location annotation included the words *by similarity*, *potential*, or *probable*.

PLOC[13] – This dataset consists of 7,579 proteins with a maximum sequence identity of 80%, extracted from Swiss-Prot release 39.0. In addition to the 11 locations covered by the MultiLoc dataset, proteins from the cytoskeleton (*cs*) are also included. This set is larger than the MultiLoc dataset, due to the inclusion of proteins whose subcellular location line in Swiss-Prot included the words *by similarity*, *potential*, or *probable*.

Using these three datasets, we compare the performance of EpiLoc to that of TargetP, PLOC, and MultiLoc. Following previous evaluations[7,9,13] we use strict, stratified, five-fold cross-validation. We do not use the same partitions as used to evaluate each of TargetP, PLOC, and MultiLoc, as these partitions include textless proteins, which are not included in the evaluation of the primary EpiLoc method, (the TargetP, PLOC, and MultiLoc datasets contain 292, 1076, and 614 textless proteins, respectively). Therefore, for each dataset we perform five sets of five-fold cross-validation runs to ensure the robustness of the evaluations.

The metrics used here for performance evaluation are those used for evaluating previous systems[7,9,13]. For each dataset, and each location, performance is measured in terms of sensitivity (*Sens*), specificity (*Spec*), and Matthew's Correlation coefficient (*MCC*)[10]. These are formally defined as:

$$Sens = \frac{TP}{TP + FN}, \quad Spec = \frac{TP}{TP + FP}, \text{ and } MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN) \cdot (TP + FP) \cdot (TN + FN) \cdot (TN + FP)}},$$

where *TP*, *TN*, *FP*, and *FN* represent the number of true positives, true negatives, false positives, and false negatives, respectively, with respect to a given location. We also measure the *overall accuracy*, *Acc* = *C*/*N*, where *C* is the total number of correctly classified proteins and *N* is the total number of classified proteins. Finally, we calculate the *average sensitivity*, *Avg*, over all locations.

To evaluate HomoLoc's performance, we conducted an experiment in which the text associated with the proteins in each of the five test subsets used for the

cross-validation of MultiLoc was removed. Each protein in each test subset was then assigned the text of its homologs by HomoLoc, without including the text associated with the protein itself.

5. Results and Discussion

Tables 3, 4, and 5 show the results of running EpiLoc on the TargetP, PLOC and MultiLoc datasets, respectively. For comparison, we also list the results reported by the authors of TargetP[7], PLOC[13], and MultiLoc[9] on their corresponding datasets, taken from the respective publications. Table 5 also shows earlier results of applying our basic text-based system[8,16] (denoted here *EarlyText*) to the MultiLoc dataset, demonstrating EpiLoc's improvement relative to the early system. Each table shows the overall accuracy (*Acc*), average sensitivity (*Avg*), and location-specific results. The highest values for each measure appear in bold, and standard deviations (denoted ±) are provided where available.

The results in Tables 3, 4, and 5 clearly indicate that the EpiLoc classifier performs at a level similar to earlier prediction systems. EpiLoc's overall accuracy and average sensitivity slightly exceed those of TargetP (Table 3), while each of the two systems scores higher than the other on some of the location-specific measures. On the MultiLoc dataset (Table 5), EpiLoc's overall accuracy, average sensitivity, and almost all location-specific scores are higher than those of the MultiLoc classifier.

On the PLOC dataset (Table 4) PLOC's overall accuracy is higher than EpiLoc's, while EpiLoc's average sensitivity is much higher than PLOC's. EpiLoc's sensitivity is actually higher for most locations. Whereas PLOC works well primarily on over-represented locations for which a large number of proteins are known (*ex*, *cy*, *pm*, *nu*, all have at least 860 proteins), EpiLoc performs well even for locations with relatively few associated proteins (*pe*, *er*, *ly*, *cs*, *go*, all with at most 125 proteins). These results all demonstrate that EpiLoc's performance is comparable to state-of-the-art prediction systems.

We note that EpiLoc's performance on both the TargetP and the MultiLoc datasets is better than it is on the PLOC set. As the criteria used for selecting proteins for the MultiLoc and TargetP datasets were stricter than those employed for the PLOC dataset (see Section 4), the resulting protein distribution among locations, and thus the distribution of associated text, is quite different among the datasets. As such, a lower *Z*-score threshold, as shown in Table 2, was needed to select a sufficient number of features (only about 1,250 actually chosen) for the PLOC set. As these terms are fewer and less distinguishing, using them to represent the PLOC dataset results in EpiLoc's lower performance.

As stated in Section 4, our evaluation of EpiLoc does not include the textless proteins from each of the three datasets. Consequently, when applied to the

Table 3. Prediction performance of TargetP and EpiLoc on the TargetP dataset, for both plant and non-plant proteins.

Loc.	TargetP	EpiLoc	TargetP	EpiLoc
	Plant (*Sens Spec MCC*)		**Non-Plant** (*Sens Spec MCC*)	
ch	0.85 **0.69 0.72**	**0.92** 0.53 0.68	N/A	
mi	0.82 **0.90** 0.77	**0.89** 0.81 **0.82**	0.89 0.67 0.73	**0.92 0.84 0.86**
SP	**0.91 0.95 0.90**	0.89 0.84 0.80	**0.96 0.92 0.92**	0.93 0.86 0.84
OT	**0.85** 0.78 0.77	0.84 **0.95 0.78**	**0.88 0.97 0.82**	**0.88** 0.95 0.81
Acc	0.853 (±0.035)	**0.862 (±0.004)**	0.900 (±0.007)	**0.901 (±0.006)**
Avg	0.856 (n/a)	**0.883 (±0.001)**	0.907 (n/a)	**0.908 (±0.003)**

Table 4. Prediction performance of PLOC and EpiLoc on the animal proteins of the PLOC dataset. Specificity and MCC values were not available for PLOC, hence only its sensitivity is listed and compared with our sensitivity values.

		PLOC Dataset (Animal)										
	Loc.	***go***	***cs***	***ly***	***er***	***pe***	***Mi***	***ex***	***cy***	***pm***	***nu***	**Acc/Avg**
PLOC	(*Sens*)	0.15	0.59	0.62	0.47	0.25	0.57	**0.78**	**0.72**	**0.92**	**0.90**	**0.796 (± 0.009)**/ 0.579 (± 0.021)
EpiLoc	(*Sens*)	**0.76**	**0.84**	**0.89**	**0.72**	**0.85**	**0.79**	0.74	0.53	0.79	0.81	0.743 (±0.002)/ **0.773 (±0.0012)**
	(*Spec*)	0.51	0.32	0.32	0.30	0.55	0.85	0.68	0.63	0.85	0.90	
	(*MCC*)	0.62	0.51	0.53	0.45	0.68	0.80	0.66	0.50	0.78	0.80	

Table 5. Prediction performance of MultiLoc, EarlyText (our basic text-based system used in earlier work[8,16]), EpiLoc and HomoLoc on the animal[d] proteins of the MultiLoc dataset.

	MultiLoc Dataset (Animal)			
Loc.	**MultiLoc**	**EarlyText**	**EpiLoc**	**HomoLoc**
	(*Sens Spec MCC*)			
go	0.71 0.43 0.53	0.86 0.40 0.57	0.88 0.62 0.73	**0.90 0.72 0.80**
ly	0.69 0.36 0.48	0.75 0.32 0.47	**0.86** 0.39 0.57	0.85 **0.49 0.63**
er	0.68 0.56 0.60	0.74 0.48 0.58	0.74 0.59 0.65	**0.77 0.67 0.71**
pe	0.71 0.31 0.44	**0.93** 0.60 0.74	0.90 **0.77 0.82**	0.80 0.69 0.74
mi	**0.88** 0.82 **0.83**	0.80 0.79 0.77	0.82 0.82 0.80	0.79 **0.84** 0.80
ex	0.79 **0.83** 0.77	0.76 0.78 0.72	0.80 0.82 0.77	**0.83 0.83 0.79**
cy	0.67 **0.85** 0.68	0.51 0.77 0.53	0.68 0.79 0.65	**0.72** 0.80 **0.67**
pm	0.73 0.90 0.76	0.80 0.91 0.81	0.85 0.90 0.84	**0.89 0.91 0.87**
nu	0.82 0.73 0.73	0.84 0.71 0.73	0.84 0.81 0.80	**0.87 0.84 0.83**
Acc	0.746 (± 0.01)	0.725 (±0.007)	0.792 (±0.008)	**0.812 (± 0.010)**
Avg	0.741 (± 0.025)	0.775 (±0.015)	0.818 (±0.005)	**0.822 (± 0.005)**

TargetP, PLOC, and MultiLoc datasets, EpiLoc predicts the location of 91.4%, 85.8%, and 89.7% of the proteins, respectively. We note that if HomoLoc (as described in Section 2) is used to assign text to the textless proteins, EpiLoc predicts the location of 100% of the proteins, while maintaining its high accuracy (e.g. overall accuracy of *0.81* on the MultiLoc dataset).

Table 5 shows the performance of HomoLoc on the MultiLoc dataset. HomoLoc's overall accuracy actually exceeds EpiLoc's, and its average sensitivity is at least as high. Moreover, HomoLoc produces many of the highest location-specific results. HomoLoc's improved performance on the MultiLoc

[d] Similar results were obtained for plant and fungus proteins.

dataset is most likely the result of the large amount of text that it associates with each protein. Having more abstracts, originating from the three close homologs, provides a larger sample of representative terms for the protein than the single set of abstracts referenced by the protein's single Swiss-Prot entry.

HomoLoc's performance on the MultiLoc dataset clearly demonstrates its utility for handling textless proteins. These results strongly support the idea that in the absence of curated text for a protein, using the text of its homologs to represent the protein yields a very good prediction.

Finally, we demonstrate by example the use of the DiaLoc method. Its proper evaluation requires a study over a prolonged period of time, in which researchers will use the web-interface to enter text and assess the results. Thus no formal evaluation is given here. Our example is the histone H1, a nuclear protein involved in the structure of DNA. For the "expert" text describing the protein, we use the description of H1 given by Wikipedia[19]. This choice of example is reasonable as it provides the high-level description we expect to obtain from an expert who has some knowledge of the protein, but is still searching for more details. Any word starting with the letters *nucle*, which might be viewed as a hint for a nuclear protein, was removed from the text. The resulting text is the input to the DiaLoc web server (Fig. 2), and the output is a location prediction. DiaLoc correctly assigns H1 to the nucleus with a probability of 0.5661, (a high value within a multinomial distribution over *9* possible locations). Although this example clearly does not test DiaLoc's overall predictive ability, it demonstrates DiaLoc as a working tool. As the prediction engine used by DiaLoc is the same one used by EpiLoc, given the same PubMed abstracts as were used for testing EpiLoc, DiaLoc's performance is the same as EpiLoc's. DiaLoc's strength lies in its ability to serve as an interactive tool for researchers.

Figure 2. User interface for DiaLoc.

6. Conclusion and Future Directions

The work presented here clearly demonstrates that EpiLoc can predict the subcellular location of proteins as reliably as other state-of-the-art systems. Moreover, we have demonstrated that the HomoLoc method is an effective way to represent proteins for location prediction. By using HomoLoc, PubLoc and DiaLoc, our system can associate text with practically any protein, and predict its location. DiaLoc is expected to be a useful tool for lab scientists, while EpiLoc and HomoLoc are primarily large-scale annotation tools.

In an earlier study[8,16] we showed that the integration of a relatively basic text-based system with the sequence-based MultiLoc system[9] produced a much

improved prediction performance with respect to the state-of-the-art. While the work presented here focuses on EpiLoc as a text based system, we expect that its integration with MultiLoc will further improve the overall performance. We plan to study such integration in the near future. Other future directions include a thorough evaluation of DiaLoc, and the extension of EpiLoc to predict sub-subcellular locations of proteins. EpiLoc and DiaLoc are available online at: *http://epiloc.cs.queensu.ca* and *http:// epiloc.cs.queensu.ca/DiaLoc.html.*

Acknowledgments

Many thanks to Oliver Kohlbacher's group at Tübingen, and particularly to Annette Höglund and Torsten Blum, for working with us on the early integration of text-features into their Multiloc system. The research is supported by CFI award #10437 and NSERC Discovery grant #298292-04.

References

1. Altschul SF, *et al. Basic Local Alignment Search Tool*. J. Mol. Biol., **215,** 403–410, 1990.
2. Bairoch A, Apweiler R. *The SWISS-PROT protein sequence database and its supplement in TrEMBL in 2000.* Nucleic Acids Res., **28,** 45–48, 2000.
3. Brady S. *Improved Prediction of Protein Subcellular Location through a Text-based Classifier*. M.Sc. Thesis, Queen's University, *http://www.cs.queensu.ca/~shatkay/papers/ScottBradyThesis.pdf*, 2007.
4. Brenner SB, *et al. Assessing sequence comparison methods with reliablestructurally identified distant evolutionary relationships. PNAS*, **95**, 6073-6078, 1998.
5. Chang CC, Lin CJ. *LIBSVM: A library for support vector machines*. 2003. *http://www.csie.ntu.edu.tw/~clin/libsvm/.*
6. Craven M, Kumlien J. *Constructing Biological Knowledge Bases by Extracting Information from Text Sources*. Proc. of the ISMB, 77–86, 1999.
7. Emanuelsson O *et al*. *Predicting subcellular localization of proteins based on their N-terminal amino acid sequence*. J. Mol. Biol., **300,** 1005–1016, 2000.
8. Höglund A *et al*. *Significantly Improved Prediction of Subcellular Localization by Integrating Text and Protein Sequence Data*. Proc. of the Pacific Symp. on Biocomput. (PSB), 16–27, 2006.
9. Höglund A *et al*. *MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition***.** Bioinformatics, **22**, 1158–1165, 2006.
10. Matthews, BW. *Comparison of predicted and observed secondary structure of T4 phage lysozyme*. Biochim. Biophys. Acta., **405,** 442–451, 1975.
11. Nair R, Rost B. *Inferring sub-cellular localization through automated lexical analysis*. Bioinformatics, **18,** S78–S86, 2002.
12. Nakai, K and Kanehisa, M. *A knowledge base for predicting protein localization sites in eukaryotic cells*. Genomics, **14**, 897–911, 1992.
13. Park, KJ, Kanehisa, M. *Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs*. Bioinformatics, **19,** 1656–1663, 2003.
14. Porter MF. *An Algorithm for Suffix Stripping* (Reprint). In: Readings in Information Retrieval, Morgan Kaufmann, 1997. *http://www.tartarus.org/~martin/PorterStemmer/.*
15. Sebastiani F. *Machine Learning in Automated Text Categorization*. ACM Computing Surveys, **34,** 1–47, 1999.
16. Shatkay H *et al*. *SherLoc: High-Accuracy Prediction of Protein Subcellular Localization by integrating Text and Proteins Sequence Data*. Bioinformatics, 23, 1410–1417, 2007.
17. Stapley *et al*. *Predicting the sub-cellular location of proteins from text using support vector machines*. Proc. of the Pacific Symp. On Biocomputing. (PSB), 374–385, 2004.
18. Walpole RE *et al*. Probability and Statistics for Engineers and Scientists, Prentice-Hall, 235–335, 1998.
19. Wikipedia contributors. *Histone H1*. Wikipedia, The Free Encyclopedia.
20. Yang Y, Pedersen JO. *A Comparative Study on Feature Selection in Text Categorization*. Proc. of International Conference on Machine Learning (ICML), 1997.