BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 11 2007, pages 1410–1417
doi:10.1093/bioinformatics/btm115

Data and text mining

SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data

Hagit Shatkay[1,*], Annette Höglund[2], Scott Brady[1], Torsten Blum[2], Pierre Dönnes[2] and Oliver Kohlbacher[2]

[1]School of Computing, Queen's University, Kingston, Ontario, Canada and [2]Division for Simulation of Biological Systems, ZBIT/WSI, University of Tübingen, Germany

Received on September 11, 2006; revised and accepted on March 17, 2007

Advance Access publication March 28, 2007

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Knowing the localization of a protein within the cell helps elucidate its role in biological processes, its function and its potential as a drug target. Thus, subcellular localization prediction is an active research area. Numerous localization prediction systems are described in the literature; some focus on specific localizations or organisms, while others attempt to cover a wide range of localizations.

Results: We introduce SherLoc, a new comprehensive system for predicting the localization of eukaryotic proteins. It integrates several types of sequence and text-based features. While applying the widely used support vector machines (SVMs), SherLoc's main novelty lies in the way in which it selects its text sources and features, and integrates those with sequence-based features. We test SherLoc on previously used datasets, as well as on a new set devised specifically to test its predictive power, and show that SherLoc consistently improves on previous reported results. We also report the results of applying SherLoc to a large set of yet-unlocalized proteins.

Availability: SherLoc, along with Supplementary Information, is available at: http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc/

Contact: shatkay@cs.queensu.ca

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

1 INTRODUCTION

Protein subcellular localization prediction is an important and well-studied problem in bioinformatics (Dönnes and Höglund, 2004; Schneider and Fechner, 2004). Knowing a protein's localization helps elucidate its function, its role in biological processes, and its potential use as a drug target. Experimental methods for protein localization range from immunolocalization (Burns *et al.*, 1994) to tagging of proteins using green fluorescent protein (Hanson and Köhler, 2001) and isotopes (Dunkley *et al.*, 2004). Such methods are accurate, but slow and labor-intensive compared to high-throughput computational methods. Computationally predicting subcellular localization enables a proteome-wide initial 'triage'. Moreover, computational methods provide information that is not otherwise attainable, e.g. for proteins that are hard to isolate, produce or locate experimentally, but whose amino acid sequence may be determined from the genomic sequence.

Much progress in computational prediction of protein subcellular localization using sequence-based information has been reported in recent years. Nakai and Kanehisa (1991, 1992) introduced a rule-based expert system, PSORT, which was later improved upon using machine-learning methods for classification (Horton and Nakai, 1997). Other prominent systems, TargetP (Emanuelsson *et al.*, 2000) and ChloroP (Emanuelsson *et al.*, 1999), based on artificial neural networks, have demonstrated a high accuracy when applied to a limited set of subcellular localizations in either plant (ChloroP) or animal cells (TargetP). Other recent systems apply a variety of machine-learning techniques. Most focus on a few subcellular localizations and improve – or just meet – the prediction accuracy of earlier systems (Bannai *et al.*, 2002; Cai and Chou, 2004; Gardy *et al.*, 2003; Nair and Rost, 2005). The best performing comprehensive sequence-based systems reported to date, which were extensively tested and compared to previous systems, are PLOC (Park and Kanehisa, 2003) and, more recently, MultiLoc (Höglund *et al.*, 2006a). While they report the best accuracy so far on a broad range of organisms and localizations, there is still room for improvement.

SherLoc is a new system that computationally assigns proteins to their respective subcellular localization. It integrates several types of sequence-derived and text-based information, and performs very well in terms of sensitivity, specificity and overall accuracy. The system is applicable to – and retains its good performance across – a wide variety of eukaryotic organisms and subcellular localizations.

SherLoc uses text to obtain features for representation and classification, as is done in information retrieval, and does not apply traditional text mining or natural language processing methods. That is, unlike Craven and Kumlien (1999) we do not try to discover localization statements in the literature. While finding such statements may help in surveying the literature, it does not support localization prediction for yet unlocalized proteins. In contrast, we use the text to obtain a set of features

*To whom correspondence should be addressed.

that are *correlated* with location, without necessarily stating or directly indicating it. The idea of using a collection of indirect weak features is well rooted in both machine learning (Hastie *et al*., 2001) and biology. For example, Jensen *et al*. (2002) predict protein function using neural networks applied to a wide range of protein features, including sequence length and isoelectric point. While such features do not biologically explain the function, the collection of feature values is correlated with function, and can help predict the function when used as a basis for a machine-learning classification method. Following the same general approach, we use a set of text-based features (which we call *distinguishing terms*) that are correlated with – but not necessarily biologically indicative of – specific subcellular locations.

Our hypothesis is that distinguishing terms, derived from text, provide features that can characterize localizations – and can thus be used to represent proteins associated with these localizations. Introducing such features into the classification process improves the ability to distinguish among proteins from different locations, and thus improves prediction performance. Underlying this hypothesis is the idea that scientists working with proteins write differently about proteins that may end up, for instance, in the nucleus, as opposed to proteins that will end up in the peroxisome. This is because such proteins are likely to be studied in different processes, related to different small molecules, be associated with diverse functions – and as such require a different language to be reported – long before their localization is determined. Put simply, the 'jargon' of the articles discussing different proteins can be used to provide cues about their localization, even when their localization is still unknown.

Several groups have already explored the use of text features to characterize biological entities and proteins in particular. Chang *et al*. (2001) use the classification of text that accompanies protein sequences to enhance homology search by PSI-BLAST. More recently, Glenisson *et al*. (2003) integrated text data into the clustering of gene expression profiles, which extends another early work that characterized and clustered genes based on text (Shatkay *et al*., 2000). Several recent publications have examined the use of text to support subcellular localization annotations. Specifically, Stapley *et al*. (2002) represented yeast proteins as vectors of weighted terms taken from all the PubMed abstracts mentioning their respective genes. The protein-text vectors were used to train support vector machines (SVM) to distinguish among subcellular localizations. The results reported were favorable in comparison to using the amino acid composition alone. However, the system was not compared against any state-of-the-art prediction method, and combining the two data sources did not show better performance than the text-based classifier alone. Nair and Rost (2002) used text obtained from curated Swiss-Prot annotations to represent proteins with known localization, and trained a prediction model using this representation. While the method just met the state-of-the-art at that time, it is limited to a few localizations and does not integrate text with other types of data. Eskin and Agichtein (2004) extended this idea, using amino acid subsequences as some of the terms considered in the text representation. The system was not compared to existing systems and the results do not suggest improvement over previous methods.

In addition to testing SherLoc on publicly available and previously used data, in the current study we introduce two new datasets and demonstrate SherLoc's predictive ability. To this end, as discussed in Section 3, we apply SherLoc to proteins whose localization was *unknown* at the time the training set was extracted and the system was trained, but has become known since, as well as to proteins whose localization is still undetermined. Finally, along with this article we present SherLoc as a publicly available server for predicting the subcellular localization of proteins.

In the next section, we outline the methods used for constructing the integrated prediction system. Sections 3 and 4 present the experimental settings and demonstrate the performance of SherLoc. Section 5 concludes the article and outlines future work.

2 METHODS

SherLoc uses localization predictions from four different sequence-based classifiers and from one text-based classifier, and integrates them to produce an improved prediction of the subcellular localization of the input protein. The four sequence-based classifiers originate from the MultiLoc prediction system, which has been described in detail elsewhere (Höeglund *et al*., 2006a). We provide here a brief description of these classifiers as well as of the text-based classifiers. Section 2.3 explains how these classifiers are combined to form the integrated prediction system, as depicted in Figure 1.

Four of the five classifiers are based on SVMs and have been implemented using the libSVM package (Chang and Lin, 2003). This implementation supports soft and probabilistic n-class categorization (Wu *et al*., 2004), in which an n-dimensional vector denoting the probability of belonging to each of the n classes is assigned to each item. Radial basis function (RBF) kernels and 5-fold cross-validation were used throughout this study. Further details are given below.

2.1 Sequence-based methods

The four sequence-based classifiers utilize four types of biological features which are known to play an important role in intracellular sorting of proteins. Namely, N-terminal targeting peptides, internal signal anchors, overall amino acid composition and certain sorting sequence motifs. SVM classifiers are used to identify the first three types of features. The fourth classifier scans the protein sequences for pre-specified short motifs indicative of structure and function (for details, see (Höglund *et al*., 2006)).

SVMTarget uses the N-terminal targeting peptide to predict chloroplast (*ch*), mitochondria (*mi*), secretory pathway (*SP*) and other (*OT*) localizations in plant cells, and only *mi*, *SP*, and *OT* in non-plant cells. Targeting peptides are represented by their partial amino acid composition. Given an input protein, the classifier outputs a 4-dimensional (3-dimensional for non-plant) vector with the probabilities of each localization.

SVMSA is a binary classifier reporting the probability that the query sequence contains a signal anchor (SA). Signal anchors are located further from the N-terminus and have a longer hydrophobic region than targeting peptides. They characterize a few secretory pathway proteins that lack an N-terminal targeting peptide and may escape detection by SVMTarget.

SVMaac is based on overall amino acid composition (aac). It is a collection of binary classifiers, one for each localization, and an additional classifier trained to separate cytosolic (*cy*) from

nuclear (*nu*) proteins. The output is a vector containing the respective probability for each localization.

MotifSearch produces a vector of binary features indicating the presence (or absence) of 43 sequence motifs in the query sequence. These motifs[1] were obtained from the PROSITE (Bairoch and Bucher, 1994) and from the NLSdb (Cokol *et al*., 2000; Nair *et al*., 2003) databases.

2.2 Text-based methods

The text-based classifier relies on the idea of representing each protein as a vector of weighted text features. While text-based protein classification has been suggested before (Nair and Rost, 2002; Stapley *et al*., 2002), our approach differs in several ways from previous research, specifically in the text source used, the feature selection and the term weighting scheme, as described below.

2.2.1 Text sources For each protein, the primary text source is the set of PubMed abstracts assigned to it by its Swiss-Prot entry. The titles and abstracts[2] of the PubMed articles referenced from Swiss-Prot are obtained for each protein. This choice of text is different from that of Stapley *et al*. (2002), who use all PubMed abstracts mentioning a certain gene's name, and from that of Nair and Rost (2002), who use Swiss-Prot annotation text rather than abstracts. The selected abstracts are tokenized into a set of terms consisting of singletons (unigrams) and pairs of consecutive words (bigrams), with standard stop words excluded from consideration. Porter stemming (Porter, 1997) was applied to all the words in the final set of terms.

An important point to note is that some proteins may not have PubMed identifiers associated with their Swiss-Prot entry, while others – newly discovered proteins – may not even be included in Swiss-Prot yet. We refer to such proteins as 'textless'. If such a protein has close homologs, which already have text associated with them, we use the text of the homologs.[3] Note that this is different from assigning the localization directly through homology, as all of the other components of the integrative system still use the original query sequence of the 'textless' protein itself. Results based on this strategy are reported in Section 4. While homology search does have limitations, our results suggest that this strategy is quite effective.

2.2.2 Term selection While the text associated with proteins may contain numerous terms, we select only a subset of *distinguishing terms* for representing proteins. The selection is done by scoring terms with respect to each subcellular localization, such that the score reflects the term's probability to occur in abstracts associated with proteins of this particular localization. Essentially, a term is *distinguishing* for a localization L if it is much more likely to occur in abstracts associated with localization L than in abstracts associated with any other localization. This idea is formalized in the following paragraphs.

Let t be a term, L a subcellular localization and p a protein. We define the following sets:

D_p is the set of all abstracts associated with p;

P_L is the set of all proteins known to be localized to L;

D_L is the set of abstracts that are associated with a localization L, defined as: $D_L = \bigcup_{p \in P_L} \{d \mid d \in D_P\}$. The number of abstracts in D_L is denoted $|D_L|$.

The probability of a term t to be associated with localization L, (denoted Pr^t_L) is expressed as the conditional probability of the term to be included in a document, given that the document is associated with the localization: $Pr^t_L = Pr(t \in d \mid d \in D_L)$. For each term t and localization L, this probability is easily estimated as the proportion of documents containing t among all those associated with the localization: $Pr^t_L \approx$ (Number of documents $d \in D_L$ *s.t.* $t \in d$)$/|D_L|$.

Based on the probability Pr^t_L, a term t is called *distinguishing* for localization L, if and only if its probability to occur in localization L, Pr^t_L, is significantly different from its probability to occur in any other localization L', $Pr^t_{L'}$. The significance is measured using a statistical test that evaluates the difference between the probabilities, Pr^t_L and $Pr^t_{L'}$, based on a Z-score (Walpole *et al*., 1998) (see Höglund *et al*., 2006b for details). When the Z-score is greater than a certain threshold (1.96, in our case), the hypothesis that the two probabilities Pr^t_L and $Pr^t_{L'}$ are indeed different is accepted with confidence greater than 95%. In this case, the term t is considered *distinguishing for localization* L, and is included in the set of distinguishing terms. Our text-based method uses only *distinguishing terms*, (of which there are about 550), for representing proteins as term vectors. The table below gives examples of some of the distinguishing terms for several localizations.

Localization	Examples of distinguishing terms
nu	*bind, control, dna, histon, transcript*
mi	*coa (CoA), cytochrom, dehydrogenas, oxidas*
go	*acceptor, catalyt domain, fucosyltransferas*
er	*calcium, chaperon, disulfid isomeras, lumen*

The distinguishing terms do not necessarily include the name of the organelle that they represent. This is an important feature, as it supports our hypothesis that documents discussing proteins of a certain localization demonstrate an over-abundance of specific terms, (which we view as the 'localization jargon'). These terms may not state the localization itself, but can be used as cues to identify the localization. These cues are helpful not because they say directly what the organelle is, but because they tend to occur in documents discussing proteins localized to that particular organelle. The text classifier uses these cue terms to associate the protein with the respective organelle.

2.2.3 Term weighting Once the collection of N distinguishing terms, denoted T_N, is established, each protein p is represented as an N-dimensional vector, where the weight $W^p_{t_i}$ at position i, $(1 \le i \le N)$, is the conditional probability of the term t_i to appear in the abstracts associated with protein p, given the set of all abstracts associated with 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 of protein p and the total number of occurrences of all distinguishing terms in the same abstracts. Formally, it is calculated as:

$$W^p_{t_i} = \frac{\sum_{d \in D_p, \text{ such that } t_i \in d} (\text{Number of times } t_i \text{ occurs in } d)}{\sum_{d \in D_p} \sum_{t_j \in T_N} (\text{Number of times } t_j \text{ occurs in } d)}, \quad (1)$$

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

The weighted term vectors representing the proteins were partitioned into training and test sets for each subcellular localization. As was the case with sequence-based feature vectors, an SVM was trained to classify these protein feature vectors into their most probable localization.

[1]Statistical analysis was conducted to obtain the set of motifs with discriminative power with respect to the localizations.
[2]Without the MeSH (Medical Subject Headings) terms.
[3]BLAST, *http://www.ncbi.nlm.nih.gov/BLAST/*, is used for identifying homologs.

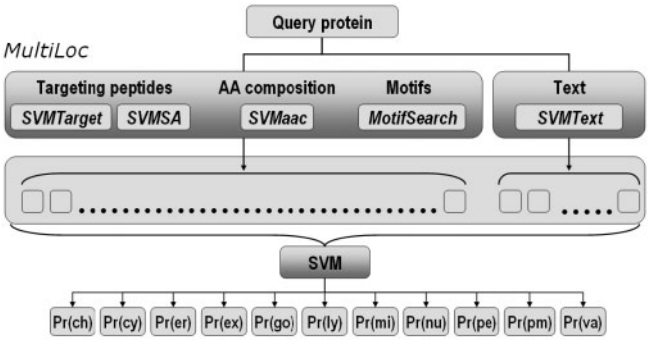

Fig. 1. SherLoc's architecture comprises four sequence-based and one text-based classifiers, and a final integrating classifier. All classifiers whose name includes the *SVM* acronym are based on SVMs. SVMTarget uses N-terminus targeting peptides, SVMSA identifies the presence of a signal anchor, SVMaac uses amino acid composition as its features. The output from the final classifier is the probability of the query protein to be localized to each of the possible 11 localizations (*ch:* chloroplast, *cy:* cytoplasm, *er:* endoplasmic reticulum, *ex:* extracellular, *go:* Golgi, *ly:* lysosome, *mi:* mitochondria, *nu:* nucleus, *pe:* peroxisome, *pm:* plasma membrane, *va:* vacuole). The most probable localization is selected as the predicted localization for the query protein.

2.3 An integrated system: SherLoc

Given a query protein as input, each of the above classifiers produces a subcellular localization prediction for this protein. Each prediction is typically a vector indicating for each localization the probability of the protein to be associated with it (an exception is MotifSearch, whose output is a binary vector indicating the presence/absence of motifs in the protein).

As illustrated in Figure 1, the output vectors of the five classifiers are combined to form the input for the final SVM classifier. This last classifier produces a vector denoting the probability of the protein to belong to each of the possible localizations. The localization with the highest probability is the one assigned to the protein as the final prediction. As proteins may belong to more than one localization, our system can easily be adjusted to output several top-ranking localizations along with their probabilities instead of a single predicted localization.

This combination of classifiers creates an integrated prediction method, utilizing both protein sequence and text data. Training and evaluation were executed using strict 5-fold cross-validation, in accordance with the practice recommended in statistical machine-learning literature (Hastie *et al*., 2001). Thus, no test protein was used to train any of the classifiers.

3 EXPERIMENTS

SherLoc was tested extensively, using at first several large sets of proteins of known localization, partitioned into training and test sets for conducting 5-fold cross-validation. Additionally, we created a novel test set by extracting proteins whose localizations were unknown at the time the training sets were created, but have become known since. Finally, we applied SherLoc to the proteins in Swiss-Prot whose localization is still unknown, providing *de novo* prediction as a basis for future laboratory experiments.

Three existing datasets, namely those used for training and testing TargetP, MultiLoc and PLOC, were used for training and for testing SherLoc using 5-fold cross-validation. As these sets were used in previous studies (Emanuelsson *et al*., 2000; Höglund *et al*., 2006a; Park and Kanehisa, 2003) they provide the basis for an extensive and sound performance comparison.

Two additional datasets were newly created, based on a recent Swiss-Prot release 48.8 (released January 2006). The proteins in these sets were not used in any way to train SherLoc, as they were not yet localized in release 42.0 (the 2003 version used to create the dataset for training and testing MultiLoc and SherLoc). The first set contains proteins that were not yet localized when SherLoc was trained but were localized in release 48.8. The second is a set of proteins whose localization is still undetermined (either uncertain or unknown) as of release 48.8. The datasets, the evaluation procedure and the results are described throughout this section.

3.1 Experimental setting

3.1.1 A comparative study The three datasets used in our comparative experiments are the following:

TP: This dataset was used for training TargetP (Emanuelsson *et al*., 2000) and contains a total of 3415 proteins from four plant (*ch*, *mi*, *SP* and *OT*) and three non-plant (*ch* excluded) localizations. The *SP* category includes proteins from all localizations in the secretory pathway: endoplasmic reticulum (*er*), extracellular space (*ex*), Golgi apparatus (*go*), lysosome (*ly*), plasma membrane (*pm*) and vacuole (*va*). The *OT* (Other) category includes cytoplasmic (*cy*) and nucleus (*nu*) proteins.

ML: A total of 5959 proteins extracted from Swiss-Prot release 42.0 (Bairoch and Apweiler, 2000) form the MultiLoc dataset (Höglund *et al*., 2006a). It covers 11 eukaryotic localizations (*cy*, *ch*, *er*, *ex*, *go*, *ly*, *mi*, *nu*, *pe*, *pm*, *va*), from animal, fungus and plant.

PL: The PLOC dataset (Park and Kanehisa, 2003) consists of 7579 proteins from Swiss-Prot release 39 covering 12 localizations with a maximum sequence identity of 80%. In contrast to MultiLoc, this dataset introduces the additional cytoskeleton (*cs*) localization within the cytoplasm.

Using these three datasets, the performance of the new system, SherLoc, is compared to that of TargetP, PLOC and MultiLoc.[4] In addition, we compare SherLoc's performance to that of an SVM classifier applied to the text data alone. Following previous evaluations (Emanuelsson *et al*., 2000; Park and Kanehisa, 2003) we consistently employ strict 5-fold cross-validation. For comparison with the PLOC dataset, we use the same split as the one used by Park and Kanehisa (2003). For the TargetP data, since Emanuelsson *et al*. (2000) do not provide the split they have used, we randomize the data split five times (on top of the 5-fold cross-validation) to ensure the robustness of the evaluations.

In order to test the performance of SherLoc on textless proteins, we conduct an experiment in which the text associated with test data was removed, and each protein in the test data was assigned the term vector associated with its closest homolog from the training data. We recall that there are no

[4]Comparison to PSORT (Nakai and Kanehisa, 1992) is not included here, since MultiLoc has already demonstrated a higher prediction accuracy when compared to it (Höglund *et al*. 2006a).

two proteins with homology exceeding 80% in the set, therefore this is a stringent test, resulting in what we view as a lower bound on the performance of SherLoc on textless proteins. In practice, we expect to assign 'textless' proteins with the text of multiple homologs with higher than 80% identity, thus expecting results that even exceed the ones reported here.

For each system and dataset, the performance is measured in terms of the sensitivity (Sens), specificity (Spec), and the Matthews correlation coefficient (MCC), defined as:

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

where *TP*, *FP*, *TN* and *FN* denote the number of true positives, false positives, true negatives and false negatives, respectively, for a given localization. Furthermore, the overall accuracy (*Acc*) is provided for each dataset, as well as the average sensitivity (*Avg*), (also known as average localization-specific *accuracy*).

3.1.2 De novo prediction Once SherLOC has been built and trained, its performance on data outside the cross-validation setting is evaluated on two new datasets: the first, *Diff48*, consists of proteins not included in the training of SherLoc, as their localization was either unknown or annotated as uncertain in release 42.0. We further limit this set to contain only proteins for which SherLoc's predictions can actually be checked, by keeping in it only proteins that were assigned a definite localization in a recent Swiss-Prot release, 48.8. Having such proteins allows us to faithfully simulate a true prediction scenario, as we predict location using a text-representation based on old abstracts and a system trained on data from several years ago, while verifying the prediction against new localization data that became available only recently. Thus, *Diff48* includes proteins that adhere to three criteria: (1) They did not participate in any way in the training of SherLoc; (2) PubMed abstracts are associated with them through Swiss-Prot entries that *predate* their localization, and were not used in the training; and (3) They now have a validated localization, indicated in a recent Swiss-Prot release, allowing us to validate SherLoc's predictions.

The second set, called *Unknown*, contains all the proteins whose localization is still uncertain or unknown in Swiss-Prot release 48.8 and have a PubMed reference associated with them.

Similar to the sets used in the comparative study, both new sets contain only animal, fungal and plant proteins, as indicated by the presence of the keywords Metazoa, Fungi or Viridiplantae, respectively, in the OC (organism classification) field. To generate the sets, we first scanned Swiss-Prot release 42.0 for all proteins that either did not have a SUBCELLULAR LOCATION line in their comment field, or contained the keywords *potential, probable* or *by similarity* in it. Any protein that occurred in the MultiLoc dataset was removed from the set to ensure that no protein used for training SherLoc is reused in the new evaluation.

As stated above, an important property of the *Diff48* set, is that it contains only proteins whose localization, as predicted by SherLoc, can be checked and verified. Thus, to build the *Diff48* set, a second step was taken in which the latest release of Swiss-Prot (48.8) was scanned for each of the proteins in the unknown/uncertain set generated above. Proteins that were now *localized with certainty* were included in the *Diff48* set. At the end of this procedure, there were 361 proteins in the *Diff48* set.

To represent the proteins and perform localization prediction, the text for each of the *Diff48* proteins comes from PubMed abstracts referenced in an earlier Swiss-Prot release, specifically, in which the protein *was still not localized*. The text for each protein is then represented as a weighted term vector using the same *distinguishing terms* that were selected as discussed in Section 2.2.2 and used in the experiments discussed in Section 3.1.1. SherLoc uses this text representation, along with protein sequence features, to predict localization. Thus the prediction does not use any information that became available after the true localization was determined. The predictions are then compared against the true locations as curated in Swiss-Prot release 48.8.

The second dataset, called *Unknown*, consists of proteins that are still unlocalized (or localized without complete confidence, as indicated by the annotation: *potential, probable* or *by similarity*), which we use for *de novo* prediction. There are ~19000 proteins in the *Unknown* set, of which ~15000 have no known localization, while the localization of the remaining 4000 is uncertain. We note that in contrast to the *Diff48* set, the predictions in this set are not presently verified.

We examine SherLoc's performance on the set *Diff48* of proteins whose localization was recently determined. As the set is quite small, some localizations are not represented at all, while other have only 1–2 associated proteins, the results – although quite good – are not always conclusive. We also applied SherLoc to predict the localizations of the *Unknown* proteins. Given the very good performance of our system on cross-validation data as well as on the newly localized proteins, we believe these predictions will prove useful. They are available from the SherLoc web site. Future laboratory experiments are, of course, necessary to validate such predictions.

4 RESULTS

Table 1 shows the total and average accuracies (*Acc* and *Avg*, respectively) over the localizations, using the TargetP (TP) and PLOC (PL) datasets. Results obtained from SherLoc, as well as from each of its individual components (MultiLoc and the text-bases classifier), are shown. The TargetP set separates plant and non-plant proteins, while PLOC distinguishes among animal, fungal and plant proteins. For comparison, we also list the results obtained by TargetP and PLOC on their respective datasets, taken directly from their respective original publications (Emanuelsson *et al*., 2000; Park and Kanehisa, 2003).[5]

A detailed comparison of our four approaches – MultiLoc (sequence only), Text (text only), SherLoc (integrated) and Homology (SherLoc, but using homologous proteins for text

[5]A comparison of prediction results for individual localizations with respect to TargetP and PLOC, is shown in Tables T1 and T2 of the Supplementary Material.

Table 1. Summary of the prediction results using the datasets previously used by TargetP (TP) and by PLOC (PL). Both the total accuracy (*Acc*) and the average sensitivity (*Avg*) are shown for the original method using this dataset (TargetP or PLOC, respectively), as well as for MultiLoc (sequence features only), Text (text features only) and SherLoc (integrating sequence and text features). The highest values are shown in bold. Standard deviations (denoted ±) are provided where available.

Dataset	Method	*Acc* (± Standard Deviation)/*Avg* (±Standard Deviation)		
TP		Plant	Non-Plant	
	TargetP	0.853 (±0.035) / 0.856 (n/a)	0.90 (±0.007) / 0.907 (n/a)	
	MultiLoc	0.897(±0.016) / 0.902 (±0.02)	0.925 (±0.012) / 0.928 (±0.011)	
	Text	0.812 (±0.026) / 0.781 (±0.032)	0.887 (±0.011) / 0.898 (±0.016)	
	SherLoc	**0.947** (±0.015) / **0.944** (±0.016)	**0.962** (±0.008) / **0.967** (±0.009)	
PL		Plant	Animal	Fungal
	PLOC	0.782 (±0.009) / 0.579 (±0.021)	0.796 (±0.009) / 0.599(±0.033)	0.795 (±0.009) / 0.568 (±0.019)
	MultiLoc	0.736 (±0.007) / 0.713 (±0.028)	0.76 (±0.007) / 0.736 (±0.039)	0.758 (±0.008) / 0.725 (±0.025)
	Text	0.687 (±0.007) / 0.735 (±0.018)	0.702 (±0.007) / 0.755 (±0.027)	0.678 (±0.005) / 0.724 (±0.026)
	SherLoc	**0.853** (±0.012) / **0.842** (±0.024)	**0.864** (±0.008) / **0.845** (±0.036)	**0.854** (±0.008) / **0.838** (±0.028)

Table 2. Prediction results, per localization, on the MultiLoc dataset, for the three systems – MultiLoc (sequence), Text (text), SherLoc (integrated) – as well as from a version of SherLoc where the text is taken from a homologous protein (*Homology*). Localization-specific (*sensitivity*, *specificity*, *MCC*) measures as well as overall results (percent accuracy (*Acc*) and average percent sensitivity (*Avg*) with standard deviations) are shown for animal and plant proteins. The results for the fungal proteins are similar to those of animal proteins and can be found in Table T4 of the Supplementary Material.

Localization	Plant (*Sens Spec MCC*)				Animal (*Sens Spec MCC*)			
	MultiLoc	Text	SherLoc	Homology	MultiLoc	Text	SherLoc	Homology
ch	0.88 0.85 0.85	0.89 0.70 0.78	**0.94 0.91 0.92**	0.91 0.87 0.88	–	–	–	–
cy	0.68 0.85 0.70	0.53 0.75 0.54	**0.81 0.91 0.82**	0.77 0.88 0.78	0.67 0.85 0.68	0.51 0.77 0.53	**0.83 0.91 0.82**	0.79 0.88 0.78
er	0.72 0.54 0.61	0.73 0.55 0.62	**0.82** 0.63 **0.71**	0.79 **0.66 0.71**	0.68 0.56 0.60	0.74 0.48 0.58	**0.82 0.67 0.73**	0.80 **0.67** 0.72
ex	0.68 0.81 0.70	0.74 0.80 0.73	**0.84 0.90 0.84**	0.80 0.89 0.82	0.79 0.83 0.77	0.76 0.78 0.72	**0.86 0.90 0.86**	0.84 0.88 0.84
ly	–	–	–	–	0.69 0.36 0.48	0.75 0.32 0.47	**0.86 0.55 0.68**	0.82 **0.55** 0.66
go	0.75 0.41 0.54	0.82 0.42 0.57	**0.84 0.61 0.70**	0.84 0.58 0.69	0.71 0.43 0.53	0.86 0.40 0.57	**0.87 0.65 0.74**	0.83 0.61 0.70
mi	0.85 0.79 0.80	0.80 0.80 0.78	**0.90 0.88 0.88**	0.86 0.84 0.83	0.88 0.82 0.83	0.80 0.79 0.77	**0.93 0.91 0.91**	0.89 0.86 0.86
nu	0.82 0.75 0.75	0.80 0.72 0.72	**0.89 0.85 0.85**	0.86 0.82 0.82	0.82 0.73 0.73	0.84 0.71 0.73	**0.89 0.83 0.84**	0.86 0.80 0.80
pe	0.71 0.34 0.47	0.88 **0.71** 0.79	**0.85 0.59 0.70**	0.82 0.57 0.67	0.71 0.31 0.44	**0.93** 0.60 0.74	0.89 **0.68 0.77**	0.84 0.61 0.70
pm	0.74 0.89 0.77	0.80 0.91 0.82	**0.84 0.96 0.87**	**0.84** 0.94 0.86	0.73 0.90 0.76	0.80 0.91 0.81	0.85 **0.95 0.87**	**0.86** 0.93 0.86
va	0.70 0.20 0.36	0.59 0.15 0.29	**0.83 0.29 0.48**	0.76 0.25 0.43	–	–	–	–
Acc [%]	74.6 (± 0.8)	73.1 (± 1.1)	**85.1 (± 1.1)**	82.6 (± 0.9)	74.6 (± 1.0)	72.5 (± 0.7)	**86.2 (± 0.9)**	83.7 (± 0.7)
Avg [%]	75.2 (± 0.9)	76.0 (± 2.3)	**85.5 (± 1.2)**	82.6 (± 0.9)	74.1 (± 2.5)	77.5 (± 1.5)	**86.8 (± 1.5)**	**83.6** (± 0.8)

assignment) – is summarized in Table 2, listing the sensitivity (*Sens*), specificity (*Spec*) and Matthew's correlation coefficient (*MCC*) for animal and plant proteins. Results for fungal proteins are shown in Table T4 of the Supplementary Material. The results were obtained using strict 5-fold cross-validation on the MultiLoc (ML) dataset, repeated five times with five different randomized splits.

Notably, the table also demonstrates the ability of our system to handle 'textless' proteins by obtaining text through homology. In this case, the test proteins were stripped of their original text, and the text from a homologous protein in the training data was used instead. We observe that the results, as shown in the Homology columns of Table 2, are still almost as good as those of SherLoc, in which we used the Swiss-Prot curated abstracts for the protein.

The results in Tables 1 and 2 clearly show that the combined classifier, SherLoc, integrating text and sequence data, outperforms earlier prediction methods, as well as its own individual components.

The Homology column in Table 2 shows an overall performance that is still better than the individual components and of previous systems, and on average is only 3% inferior to SherLoc's performance. These results strongly support the idea that in the absence of curated text for a protein, 'borrowing' text from a relatively remote homolog (less than 80% identity), yields a very good prediction when integrated with sequence data. It affirms homology-based text recovery as an effective way to handle proteins that have no curated text when using the integrative framework. Again, we stress that homology is used here *only* to obtain text for a protein that may not have it. It is

not used for assigning the localization of a homolog to an unknown protein.

Finally, we ran our systems on the two new datasets *Unknown* and *Diff48*. As the *Unknown* set contains ~19000 proteins, the results are not given here but provided on the SherLoc web site.

As for *Diff48*, this relatively small set does not uniformly represent all localizations (*go, pe, ly* and *pm* are not represented at all, and *ch, va* and *er* have between 1 and 3 proteins each). Overall, SherLoc predicted the localization of these newly localized proteins with an accuracy of about 71%, but performance per localization varies. (See Table T3 of the Supplementary Material.) For instance, SherLoc predicts the 132 *extracellular* proteins with 79% sensitivity and 99% specificity, exceeding all previously reported predictive results (including SherLoc's own specificity) on cross-validation data. SherLoc's sensitivity on the newly localized extracellular proteins is only slightly lower than its demonstrated extracellular sensitivity on cross-validation data. For the 21 newly localized *mitochondrial* proteins, SherLoc demonstrated a similarly high performance (75% specificity; 95% sensitivity). On the 91 newly localized cytoplasmic proteins, SherLoc's sensitivity was 79%, similar to its own performance and exceeding all previously reported methods on cross-validation data. Its specificity was much lower (59%), as it predicted ~50 of the 111 nucleus proteins as cytoplasmic proteins. This obviously also lowered SherLoc's performance on the nucleus proteins. This demonstrates a well-known problem in distinguishing nucleus from cytoplasmic proteins. TargetP handles it by simply grouping nucleus and cytoplasmic proteins together as one set. If we were to do the same, the performance soars to above 90% in both sensitivity and specificity for the combined set. However, we aim to predict the most specific localization possible rather than to combine localizations into sets. As for the localizations with 1–3 proteins, while SherLoc does correctly predict some of them, this sample size is not sufficient to merit analysis.

Overall, the above results demonstrate excellent performance using the standard measures over all available datasets when using the widely accepted scheme of 5-fold cross-validation. Moreover, we show almost equally good results on the data available for newly localized proteins that were not part of the training/testing procedure. Finally, we provide *de novo* prediction for a set of ~15 000 proteins whose localization is unknown to date, and for additional 4000 proteins whose localization is still uncertain.

5 DISCUSSION AND OUTLOOK

We introduced SherLoc, a new comprehensive system for predicting subcellular localization through integration of text and sequence data. SherLoc, similarly to the system reported by Nair and Rost (2002), uses Swiss-Prot as its primary text source. However, we do not use the curated annotation text, but rather the PubMed abstracts referenced in Swiss-Prot. Stapley *et al*. (2002) use *every* abstract that contains the gene name for the protein. In contrast, we use *only* abstracts that are referenced by Swiss-Prot. Moreover, rather than using all the terms with a standard (TF*IDF)[6] weighting, as done by Stapley *et al*., we select terms discriminatively as described in Section 2.2.2, and apply a probabilistic weighting scheme (Section 2.2.3). Moreover, we suggest a way to support text-based localization even when the protein entry in Swiss-Prot does not contain any PubMed abstracts (or when proteins may not even have a Swiss-Prot entry).

The methods, experiments and results presented here clearly demonstrate that SherLoc achieves a significantly improved prediction of eukaryotic protein subcellular localization. Table 2 in particular demonstrates that the use of text and sequence data distinctly complement each other. MultiLoc, which uses sequence data, typically performs well predicting localizations that are directed by N-terminal signals, such as the mitochondria and the chloroplast. Text information complements it, and its contribution is particularly noticeable for localizations whose sequence-based signal is not as overt, including those related to the secretory pathway, such as the Golgi apparatus and the endoplasmic reticulum.

When relying on curated text for prediction, it is often pointed out that such text is not always available. We introduce one solution to this problem by using text that is associated with a homologous protein. We have shown that using the text of related proteins (less than 80% identity) instead of the protein's own text is indeed effective, as the performance is still significantly better than that of a system relying on sequence data alone. Furthermore, this performance is only slightly lower than that obtained by SherLoc when using the protein's own curated abstracts.

In addition, we have created and introduced two new datasets and applied SherLoc to them, demonstrating SherLoc's ability to predict the localization of completely new proteins, outside the dataset on which it was trained and tested through 5-fold cross-validation. The set *Diff48* allows us to evaluate SherLoc's predictions realistically. By applying SherLoc to proteins whose localization was not yet determined in SherLoc's training data, but became known in a new Swiss-Prot release, we can validate predictions made outside the standard test-and-train cross-validation setting. Our results showed that while for several localizations (mitochondria, extracellular) prediction was actually better than expected from the cross-validation studies, for nuclear and cytoplasmic proteins the cross-validation studies showed better results than those observed over new data. This serves as a reminder that cross-validation studies do have their limitations when the characteristics of yet unknown proteins are not necessarily the same as those of well-studied and well-known ones.

By predicting localization for the *Unknown* set of 19 000 proteins, whose localization is either unknown or uncertain, we provided new tentative localizations for 15 000 proteins that currently do not have any associated localization. These predictions can serve as putative annotations and should be experimentally validated. In the meantime, they provide preliminary clues for experimentalists.

From the text-mining perspective, we note that biological text mining has been an active research area for about a

[6]An acronym for Term Frequency times Inverse Document Frequency.

decade now, and much work has focused on identifying entities and relations in text. A Nature news feature titled *Biology's Name Game* (Pearson, 2001) pointed out the difficulties in identifying protein and gene names in biomedical text. This challenging problem is the center of active and fruitful research (Hanisch *et al*., 2003; Hirschman *et al*., 2005; Tanabe and Wilbur, 2002), under the assumption that the best way for Biology to utilize the literature is by first accurately identifying biological entities in it. So far, the progress made in biological named entity recognition has not translated into a quantitative improvement with respect to any specific biological problem. Here we use a very different approach. We do not try to 'play' biology's name game, but rather use text as just another source of features to characterize proteins. Our results demonstrate, for the first time, the definite utility of text, by achieving a measurable and significant improvement in accuracy in predicting protein subcellular localization.

We are expanding the system further to allow the localization of proteins by integrating text sources other than PubMed abstracts. Specifically, we are working on ways to combine text from several homologous proteins instead of just one, as well as on using other text sources such as user-provided summaries. Experimental validation of SherLoc's predictions as well as computational prediction of intraorganelle localizations are additional lines of ongoing work.

ACKNOWLEDGEMENTS

H.S. gratefully acknowledges the support of NSERC Discovery Grant 298292-04 and CFI New Opportunities Award 10437. We thank Nora Toussaint for suggesting the name SherLoc.

Conflict of Interest: none declared.

REFERENCES

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

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

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

Burns,N. *et al*. (1994) Large-scale analysis of gene expression, protein localization and gene disruption in Saccharomyces cerevisiae. *Genes Dev*., **8**, 1087–1105.

Cai,Y.D. and Chou,K.C. (2004) Predicting 22 protein localizations in budding yeast. *Biochem. Biophys. Res. Commun*., **323**, 425–428.

Chang,C.C. and Lin,C.J. (2003) LIBSVM: a library for support vector machines. *http://www.csie.ntu.edu.tw/~cjlin/libsvm/*.

Chang,J.T. *et al*. (2001) Including biological literature improves homology search. In *Pac. Symp. Biocomp. (PSB'01)*, pp. 364–383.

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

Craven,M. and Kumlien,J. (1999) Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Int. Conf. on Int. Systems for Mol. Bio. (ISMB'99)*, pp. 77–86.

Dönnes,P. and Höglund,A. (2004) Predicting protein subcellular localization: past, present, and future. *Genomics Proteomics Bioinformatics*, **2**(4), 209–215.

Dunkley,T. *et al*. (2004) Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics*, **3**, 1128–1134.

Emanuelsson,O. *et al*. (1999) Chlorop, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci*., **8**, 978–984.

Emanuelsson,O. *et al*. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol*., **300**, 1005–1016.

Eskin,E. and Agichtein,E. (2004) Combining text mining and sequence analysis to discover protein functional regions. In *Pac. Symp. Biocomput. (PSB'04)*, pp. 288–299.

Gardy,J.L. *et al*. (2003) PSORT-B: improving protein subcellular localization prediction for gram-negative bacteria. *Nucleic Acids Res*., **31**, 137–140.

Glenisson,P. *et al*. (2003) Meta-clustering of gene expression data and literature-extracted information. *ACM SIG KDD Explorations*, **5**(2), 101–112.

Hanisch,D. *et al*. (2003) Playing biology's name game: identifying protein names in scientific text. In *Proceedings of the Pac. Symp. Biocomp. (PSB)*, pp. 403–411.

Hanson,M.R. and Köhler,R.H. (2001) GFP imaging: methodology and application to investigate cellular compartmentation in plants. *J. Exp. Bot*., **52**.

Hastie,T. *et al*. (2001) *The Elements of Statistical Learning*. Springer-Verlag, New York.

Hirschman,L. *et al*. (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, **6**.

Höglund,A. *et al*. (2006a). Multiloc: prediction of protein localization using n-terminal targeting sequences, sequence motifs and amino acid compositions. *Bioinformatics*, **22**, 1158–1165.

Höglund,A. *et al*. (2006b) Significantly improved prediction of subcellular localization by integrating text and protein sequence data. In *Pac. Symp. Biocomp. (PSB'06)*, pp. 16–27.

Horton,P. and Nakai,K. (1997) Better prediction of protein cellular localization sites with the k nearest neighbors classifier. In *Proceedings of Int. Conf. Intell. Syst. Mol. Biol. (ISMB'97)*, pp. 147–152.

Jensen,LJ. *et al*. (2002). Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol*., **319**, 1257–1265.

Nair,R. *et al*. (2003) NLSdb: database of nuclear localization signals. *Nucleic Acids Res*., **31**, 397–399.

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

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

Nakai,K. and Kanehisa,M. (1991) Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins: Structure, Function and Genetics*, **11**, 95–110.

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

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

Pearson,H. (2001) Biology's name game. *Nature*, **411**, 631–632.

Porter,M.F. (1997) An algorithm for suffix stripping (reprint). In *Readings in Information Retrieval*. Morgan Kaufmann, San Francisco.

Schneider,G. and Fechner,U. (2004) Advances in the prediction of protein targeting signals. *Proteomics*, **4**, 1571–1580.

Shatkay,H. (2005) Hairpins in bookstacks: information retrieval from biomedical text. *Briefings in Bioinformatics*, **6**, 222–238.

Shatkay,H. *et al*. (2000) Genes, themes and microarrays: using information retrieval for large-scale gene analysis. In *Proceedings of the Int. Conf. on Int. Systems for Mol. Bio. (ISMB'00)*, pp. 317–328.

Stapley,B.J. *et al*. (2002) Predicting the subcellular location of proteins from text using support vector machines. In *Pac. Symp. Biocomp. (PSB'02)*, pp. 374–385.

Tanabe,L. and Wilbur,W.J. (2002) Tagging gene and protein names in full text articles. In *Proceedings of the Workshop on Nat. Lan. Proc. in the Biomed. Domain*, pp. 9–13.

Walpole,R.E. *et al*. (1998) *Probability and Statistics for Engineers and Scientists*. Prentice Hall, College Div. pp. 235–335.

Wu,T.-F. *et al*. (2004) Probability estimates for multi-class classification by pairwise coupling. *J. Mach. Learn. Res*., **5**, 975–1005.