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].