Text-Based Prediction of Protein Subcellular Location

by

Scott Brady

A thesis submitted to the School of Computing

in conformity with the requirements for

the degree of Master of Science

Queen's University

Kingston, Ontario, Canada

April 2007

Abstract

The ability to efficiently and accurately determine the subcellular location of a protein is an active area of research in proteomics. The subcellular location of a protein can help to elucidate several of its characteristics, including its function, its role in biological processes, and its potential as a drug target. In this thesis we present a new system, called EpiLoc, for predicting subcellular location. EpiLoc represents proteins as term-vectors, where each component in a vector corresponds to a term that is correlated with a specific location. The system uses a method we refer to as *Z-Test* to identify the set of terms that is used to represent proteins, and we conduct an in-depth study comparing this method to other standard feature selection methods. For a given protein, the weight assigned to a term is based on its frequency of occurrence in text about the protein. To ensure that EpiLoc can predict the location of practically any protein, we develop several methods for associating text with proteins. The term-vectors are used to train a classifier to recognize term distributions that are indicative of specific locations. The performance of the EpiLoc system is examined both as a standalone classifier and as part of an integrated sequence- and text-based classifier called SherLoc (the latter developed in collaboration with a group from the University of Tübingen). Both systems are compared to several other state-of-the-art classifiers. The results demonstrate that as a standalone classifier EpiLoc performs at a level comparable to (and in some cases better than) other state-of-the-art systems. Moreover, the results from the integrated system suggest that the integration of text and sequence data can achieve a significant, quantitative improvement over a system that uses biological data alone.

Co-Authorship

Sections 3.3, 6.1, and 6.2 contain methods and results that were developed, obtained and published in collaboration with Professor Oliver Kohlbacher's group from the University of Tübingen, in the following papers:

1) Höglund, A., Blum T., Brady, S., Dönnes, P., Miguel, J.S., Rocheford, M., Kohlbacher, O., Shatkay, H.: Significantly Improved Prediction of Subcellular Localization by Integrating Text and Protein Sequence Data. In: *Proc. of the Pacific Symposium on Biocomputing* (*PSB*), 16-27, 2006.

2) Shatkay, H., Höglund, A., Brady, S., Blum, T., Dönnes, P., Kohlbacher, O.: SherLoc: High-Accuracy Prediction of Protein Subcellular Localization by integrating Text and Proteins Sequence Data. *Bioinformatics*, 2007.

Acknowledgements

I would like to express my sincerest thanks to my supervisor, Dr. Hagit Shatkay, who provided great knowledge and guidance throughout my thesis work. I would also like to thank her for all the support and insight she offered me throughout my education at Queen's University.

Special thanks to Torsten Blum for his time and help in designing the EpiLoc classifier. Without the data, information, and clarification he provided the results of this thesis would not have been possible.

My genuine gratitude goes to the entire group at the University of Tübingen: Annette Höglund, Torsten Blum, Pierre Dönnes, and Oliver Kohlbacher. This thesis would not have been the same without their productive collaboration.

I would like to thank my thesis examination committee, Dr. Abolmaesumi, Dr. Young, and Dr. Salomaa, for their insightful suggestions.

I extend my gratitude to the School of Computing at Queen's University for funding my studies.

Finally, I would like to express my sincerest appreciation to my parents for their love and support.

SCOTT BRADY

Queen's University at Kingston, Ontario
April 2007

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The subcellular location of a protein provides important information for deducing its role within the cell. Computational systems for predicting subcellular location are actively being developed and studied. To this end, we present in this thesis two new systems, *EpiLoc* and *SherLoc*, which utilize text data to predict location.

1.1 Motivation

Proteins are large molecules composed of *amino acids* arranged in a sequence. They control the behavior of the cell through the myriad of functions they perform, such as transporting molecules, acting as support structures, catalyzing reactions, digesting molecules, signaling to other cells, and defending against disease [52].

The set of proteins that an organism produces is determined by its *genes*; they specify how, and when, to create a protein. An organism's complete genetic information is called its *genome*. Recent methods allow researchers to sequence entire genomes, thereby making it possible to deduce the set of proteins found in organisms [52]. Several genomes, including the human genome [50, 25], have been determined. This breakthrough has resulted in vast amounts of protein data. The human genome alone encodes hundreds of thousands of proteins [37]. For the majority of proteins discovered through large-scale genomic efforts, very little is known about their role within the cell.

The goal of *proteomics*, the field of research concerned with the study of proteins, is to reliably annotate proteins with information regarding their structure and function. The reliable annotation of all known proteins is a long-term project that will take years to complete. It requires the development of systems that can quickly and reliably determine specific characteristics of a protein, including its structure, interaction partners, binding sites, and subcellular location. Computational methods that assign these characteristics to a protein are being developed [12]. In this thesis, we are specifically interested in methods that assign a subcellular location to a protein.

Subcellular location is a starting point for discerning other information about a protein. A protein's location is often directly related to its function. For example, proteins found in the mitochondria are often involved in metabolism. Moreover, knowing the location of a protein can help clarify its role in disease, or even indicate its suitability as a drug target [43]. Consequently, methods for determining subcellular location are being actively studied.

Experimental methods for determining a protein's location already exist [18, 26]. Such methods, however, are laborious and time consuming; applying them to the large number of proteins for which a subcellular location is still unknown would be almost impossible to accomplish in the near future. Therefore, other methods that can rapidly assign subcellular location are being developed.

The fastest methods for assigning a putative subcellular location are computer based. Many systems assign a location based on certain amino acid sequence characteristics, including the order of its amino acids [15, 4], its composition [41, 8, 23, 38], and its similarity to other protein sequences [33]. Other systems employ a subset of these criteria in order to assign subcellular location [20, 34]. However, none of these systems is completely accurate. Typically, systems that achieve high accuracy [14, 15] can only distinguish among proteins from a few locations. On the other hand, systems that assign proteins to a larger number of different subcellular locations [21, 8] usually achieve a low accuracy. The most advanced systems [38, 20] assign proteins to a large number of locations and achieve a relatively high accuracy, but still leave room for improvement.

To improve the reliability of prediction systems, other sources of protein information may need to be considered. Techniques that use textual information concerning a protein have already been introduced, although none have so far demonstrated a significant improvement over sequence-based classifiers. This thesis presents a new text-based method for the prediction of protein subcellular location, and demonstrates its use in a system that integrates both sequence and text information to assign protein location.

1.2 Thesis Objectives

The work described here aims to investigate the utility of using text data to predict the subcellular location of a protein. Our primary goal is to design, develop, and test a complete standalone text-based location prediction system that can also supplement and improve the performance of a sequence-based prediction system. Specifically, this work intends to:

- Produce a text-based system that predicts protein subcellular location, and compare our system's performance to that of state-of-the-art systems. To be comparable to other systems, ours must be able to assign a subcellular location to essentially any protein. This is a challenging task, as text relevant to each protein is not always available.
- Integrate our text-based system with a sequence-based system, compare the performance of the integrated system to that of state-of-the-art systems, and examine the effects of the integration.

We call our text-based classifier EpiLoc. Its development involves the following essential steps:

1. Assignment of text to a protein dataset.
2. Selection, from that text, of terms that will be used to represent the proteins.
3. Representation of each protein as a vector of term-weights, where each weight represents the significance of a term in the text specifically associated with the protein.
4. Training and testing of a classifier using the vectors.

We also build a system that combines text and sequence data, by integrating an early version of EpiLoc with the sequence-based system, MultiLoc [20]. The integrated system is called SherLoc. The design of both systems is described in detail in Chapter 3.

1.3 Thesis Contributions

The performance of both EpiLoc and SherLoc has been evaluated extensively, and suggest that both are viable tools for subcelluar location prediction. Specifically, the main contributions of this thesis are:

1. We demonstrate that our text-based prediction system, EpiLoc, is as effective as current state-of-the-art systems for protein subcellular location prediction. Moreover, we conduct an in-depth study of specific components of the EpiLoc system. In particular, we compare several different feature selection methods, and evaluate the effectiveness of several methods for associating text with proteins.
2. We demonstrate that the integrated sequence- and text-based classifier, SherLoc, significantly improves upon other state-of-the-art systems for predicting subcellular location.
3. We show that a system that uses biological and text data to address a biological problem can achieve a significant improvement over a system that uses biological data alone.

1.4 Thesis Outline

This thesis describes the development of the EpiLoc and the SherLoc systems, as well as the procedures used to test their effectiveness. Chapter 2 provides an introduction to the biology driving protein localization in organisms, and surveys current work on determining the subcellular location of a protein. Chapter 3 details the design of the EpiLoc and the SherLoc systems, and describes several methods for assigning text to proteins. In Chapter 4, we describe the approach used to test the effectiveness of EpiLoc and SherLoc, while in Chapter 5 we define and analyze our term selection method. The results of our experiments are presented in Chapter 6, while Chapter 7 concludes the thesis and proposes future work.

Chapter 2

Background

In this chapter we provide background about proteins and the methods used for inferring and predicting their subcellular location. We begin by presenting the basic biology pertaining to proteins, specifically details concerning subcellular localization. Such details are important, as they are often considered in the design of systems for subcellular location prediction.

We next survey experimental and computational techniques used to assign protein subcellular location. As laboratory techniques are slow and labour intensive, there is a clear need for quicker, computer-based approaches to predict subcellular location. Understanding existing approaches is imperative in order to understand where and how improvements may be made to the prediction process.

2.1 Proteins

Every organism is controlled by proteins and the functions they perform within its cells. Some proteins have mechanical or structural functions, while others play important roles in the production of cellular energy. Still others are involved in immune responses, cell signalling, digestion, and catalysis of biochemical reactions. Protein functions within a cell determine its role within an organism, thereby influencing the overall development and activity of the organism [52].

A protein is a macromolecule comprised of basic units, called *amino acids*, arranged in a sequence (Figure 2.1.1). There are millions of proteins, varying in length and in their amino-acid composition. The exact sequence of each protein is determined by its coding gene [52].

Genes are stored as part of the deoxyribonucleic acid (DNA), which is a long sequence of a particular type of molecules, known as *nucleotides*. Individual genes are located along sections of the DNA, and along each gene there are certain regions that encode a protein. These coding regions consist of sequences of *codons*, which are sequences of three nucleotides that specify an amino acid (Figure 2.1.1). Each codon encodes a single amino acid; the sequence of codons found in a gene determines the amino acid sequence of a protein. Therefore, the amino acid sequence of a protein can be derived directly from the nucleotide sequence of its corresponding gene [52].

The synthesis of a protein consists of two main steps, *transcription* and *translation*. During transcription, DNA is used to produce molecules called *mRNA* that carry the code for creating a protein from the nucleus, which contains the DNA, to the ribosomes, which form the part of the cell where proteins are formed. During translation, the ribosomes adjoin together, in order, the amino acids specified by the mRNA. Following translation, the newly formed amino acid sequence arranges itself into its final shape, or *conformation*, and protein synthesis is complete [52]. To perform its function, the fully formed protein is then transported to its final location within or outside the cell. Some proteins may be found in more than a single location, and some may even "shuttle" between multiple locations [6].

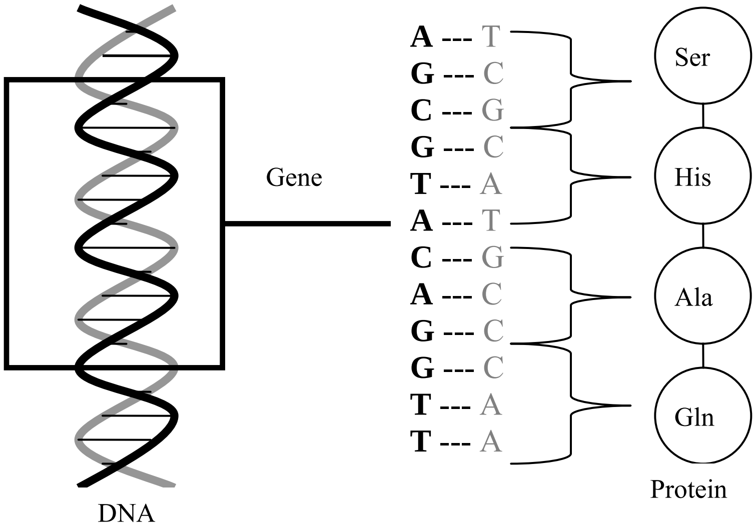


Figure 2.1.1: The correspondence between the nucleotide sequence of a gene and the amino acid sequence of a protein. A sequence of 3 nucleotides encodes one amino acid.

The nucleotide sequences of several genomes, including the human genome [50, 25], have been determined. Researchers are using these genomes to deduce the amino acid sequence of hundreds of thousands of proteins, many of which have yet to be experimentally produced and studied. Determining the function of these proteins is a major goal of proteomics, as a protein's function provides insight into its role in disease and healthy processes. However, function is often difficult to determine.

The subcellular location of a protein can often help to elucidate several of its properties, including its function [12]. For example, proteins found in the nucleus often interact with the DNA in processes such as transcription. In addition to providing cues about function, the location of a protein may also help to determine its interaction partners and its potential as a drug target [12, 20].

As knowing the subcellular location of a protein can help to advance its study, the ability to determine subcellular location has become an important objective of computational proteomics. Understanding the physical localization process of proteins within the cell may help guide the development of systems for computationally determining a protein's location. This process continues to be researched, and we describe it in the next section.

2.2 Cellular Sorting Processes

Proteins have evolved over time, during which their sequences have adapted to function optimally within specific subcellular locations [1]. As a result, the correct delivery of a protein to its final location is imperative to ensure its proper functioning [12].

The cells within organisms consist of several components (or locations), where different organisms have different sets of subcellular components. In this thesis we focus on subcellular locations found in either plant, animal, or fungal cells. These types of cells contain several membrane-bound components, known as *organelles*, in which proteins may be found. Each organelle performs a specific function within the cell (Table 2.1.1). The main organelles are the endoplasmic reticulum, Golgi apparatus, lysosome, mitochondria, nucleus, peroxisome, and plasma membrane. There are also organelles that are specific to certain organisms, such as the chloroplast (plant), lysosome (animal), and vacuole (plant and fungal). All organelles are surrounded by the cytoplasm, and the plasma membrane encapsulates the cell and separates its contents from the extracellular space [7]. Table 2.1.1 lists the organelles (also included are the extracellular space and the cytoplasm) along with their function and the type of organism in which they can be found.

Subcellular Location	Primary Function	Organism(s)
chloroplast	Generates energy for plant cells from sunlight	Plant
cytoplasm	Surrounds the organelles and maintains the shape of the cell	Animal, Plant, Fungal
Endoplasmic reticulum	Transports proteins to be attached to the plasma membrane or to be secreted to the extracellular space	Animal, Plant, Fungal
extracellular space	Surrounds the cell	Animal, Plant, Fungal
Golgi apparatus	Processes and packages molecules synthesized by the cell, including proteins	Animal, Plant, Fungal
lysosome	Primary site of digestion, breaks down proteins and other molecules within the cell before they are exported to the extracellular space	Animal
mitochondria	Creates energy for the cell by converting organic matter into energy that the cell can use	Animal, Plant, Fungal
nucleus	Stores the DNA and separates it from the cytoplasm	Animal, Plant, Fungal
peroxisome	Rids the cell of toxic materials	Animal, Plant, Fungal
plasma membrane	Controls the movement of molecules in and out of the cell	Animal, Plant, Fungal
vacuole	Stores - as well as exports - waste, water, and food	Plant, Fungal

Table 2.1.1: Subcellular locations and their primary functions. The type of organism in which each organelle can be found is listed in the rightmost column.

The translation of mRNA into proteins occurs in the cytoplasm. From there, proteins can enter the secretory pathway (the system that transports proteins and other molecules to the plasma membrane and the extracellular space), be routed to other locations that are not part of the secretory pathway, or remain in the cytoplasm [12]. The specific location of a protein within the cell depends in large part on its amino acid sequence and on certain subsequences within it (known as *motifs*) [1].

The secretory pathway is comprised of the Endoplasmic reticulum, extracellular space, Golgi apparatus, lysosome, plasma membrane, and vacuole. Proteins that are bound for the secretory pathway typically carry a targeting signal, called a *signal peptide*, as part of their amino acid sequence. The signal peptide is located on the N-terminal region, which is the end of a protein that has an unbound amino group (the other end is the C-terminal region, and has an unbound carboxyl group). Proteins bound for the secretory pathway may also contain a signal anchor, which is found further away from the N-terminal region [29]. In both cases, proteins are translated and simultaneously transported to the inside of the endoplasmic reticulum (ER). Once inside the ER, they may either remain there (if they contain an ER retention signal) or move to other regions of the secretory pathway [12].

Proteins bound for locations outside the secretory pathway are completely assembled in the cytoplasm, and then transported to their final destination. Chloroplast, mitochondrial, and lysosomal proteins usually have an N-terminal targeting sequence, or *transit peptide*, which is recognized by the trafficking system. Some intrinsic sequences guide proteins to the plasma membrane, while those going to the nucleus are generally identified by a nuclear localization signal, which consists of 4 to 8 amino acids. Peroxisomal proteins contain a short C-terminal signal sequence that allows them to cross the peroxisome’s membrane and pass into its interior. If a protein does not have any of the signals described above, it usually remains in the cytoplasm and carries out its function there [12].

The sorting signals described above are not the only ones to determine the protein's subcellular location. Factors such as conformation and amino acid composition also participate in protein localization. The conformation of a protein affects its localization, as conformation, to a large extent, determines the interactions a protein may undergo during localization. Moreover, a protein's amino acid composition, the relative frequency of each amino acid in the protein's sequence, is correlated with and influences subcellular localization [1]. For instance, plasma membrane proteins tend to contain a disproportionately large number of hydrophobic amino acids (amino acids that lack an affinity for water).

Methods for determining a protein's subcellular location are actively being investigated. These methods are generally either experimental or computational, and are described in the sections that follow.

2.3 Experimental Localization

The task of ascertaining the location of a protein within a cell was initially undertaken in a laboratory setting. Lab techniques involve attaching a marker to a protein to allow researchers to visually trace the path of the protein to its subcellular location. The two most common techniques are green fluorescent protein (GFP) localization [18] and immunolocalization [26].

Green fluorescent protein emits fluorescent light. Attaching a GFP to another protein makes the latter detectable by observing the light it emits. Tagging with a GFP involves fusing the GFP's corresponding gene sequence to the end of another gene. A region that can be used to promote transcription is also fused to the end of the gene. Transcription is induced, and the protein is synthesized. As the protein relocates to its subcellular location the fluorescent light emitted by the GFP tag makes it possible to observe and determine the location of the protein [18].

Immunolocalization is another experimental procedure that, similarly to the GFP tagging, allows a researcher to physically observe a protein. However, for immunolocalization the target protein is tagged with an epitope, which is the portion of a protein that other proteins recognize and to which they bind. A gene sequence encoding the epitope is fused to the end of the coding gene, and the resulting target protein thus contains the epitope. A protein that binds to the epitope and has been altered to emit fluorescent light is then added to the cell. The interaction between the target and the fluorescent-emitting proteins allows researchers to determine its location within the cell by observing the fluorescent light [26].

Immunolocalization and GFP tagging are both very accurate processes. Special consideration is required, however, when utilizing either method, as the required modification to the original gene sequence can affect the localization mechanism itself. For instance, fusing GFP at the N-terminal of a protein might prevent the trafficking system from interacting with the targeting signal that normally guides the protein to the mitochondria [26]. Another problem with using either method is that the inclusion of the extra protein portion may cause a conformational change in the protein that could either activate or deactivate a localization signal that may or may not be normally present in the native protein [18]. Finally, both processes are slow and labour intensive, and cannot handle the massive amounts of protein data currently available. As such, high-throughput computational techniques for the assignment of subcellular location have been proposed in recent years.

2.4 Classification using Sequence Information

In order to address the need for high-throughput methods for determining protein subcellular location, computational techniques have been developed. These techniques do not simply match localization signals against a protein sequence to determine subcellular location. Despite the knowledge that has been acquired regarding protein sorting, the final location of a protein cannot be fully determined from its sequence alone. Signal sequences are not always present or easy to identify, and proteins with similar amino acid composition may belong to different locations. Therefore, methods that attempt to predict subcellular location have been developed. Specifically, classification methods assign a location, (where the latter is viewed as a class tag), to a protein whose location is unknown.

One method for predicting subcellular location involves assigning a protein to the same location as that of a protein whose amino acid sequence is similar (known as a *homolog*), and whose location is already known. Often proteins that are highly similar share many other characteristics such as conformation, function, etc. [52, 33, 5]. There are several algorithms, such as BLASTP [2], that measure the similarity between two protein sequences. Nair and Rost [33] demonstrated that if the amino acid sequences of two proteins are similar enough, the subcellular location of one can be reliably assigned to its homolog. They found that in order to confidently assign a location, the protein whose location is unknown must be at least 70% identical to the protein whose location is known. As a result, proteins that do not have a homolog with a high enough similarity cannot be localized using this method, and expert systems or machine learning classifiers are used instead.

An expert system is a collection of if-then rules based on expert knowledge. Such systems have been used to derive the location from protein information [12, 34]. A different approach is to use a machine learning technique, such as hidden Markov Models (HMMs), artificial neural networks (ANN), Bayesian networks, K-nearest neighbours, or support vector machines (SVMs), to predict subcellular location [45]. To use such techniques, proteins are represented by information related to them, and a classifier is trained based on this representation to assign a subcellular location.

There are three common ways to represent a protein. The first is based on N-terminal targeting signals; the second global sequence features, most often amino acid composition. The third is based on combining several types of protein data that may come from N-terminal targeting signals, amino acid composition, sequence motifs, text annotations, or other sources of information about a protein [12]. These approaches vary in the number of possible locations into which they can classify proteins, the number of proteins they can represent (known as *coverage*), and most importantly – in their prediction accuracy[1]. The following sections provide more details about the different approaches and the systems using them.

2.4.1 N-Terminal Based Classifiers

Methods based on N-terminal targeting information classify proteins into four possible locations: *chloroplast, mitochondria, secretory pathway* (*SP*), and *Other*, where *Other* includes proteins belonging to all other possible locations. Plant proteins can be

[1] Accuracy is defined as the number of proteins correctly classified, divided by the total number of proteins. A formal definition of accuracy and other performance measures are provided in Section 3.2.2.

classified into all four locations, while animal proteins can only be classified into three of them, as they cannot be localized to the *chloroplast*. As a result of the broad classes into which N-terminal based classifiers may assign a protein (both *SP* and *Other* include proteins from multiple locations), they usually attain a high accuracy [20].

Two early attempts to classify proteins based on their N-terminal sequence were SignalP [36] and ChloroP [14]. SignalP classifies an amino acid sequence as belonging/not-belonging to a signal peptide, while ChloroP classifies a sequence as belonging/not-belonging to a chloroplast transit peptide. Both techniques are based on neural networks, and use a multi-layer network architecture. The first layer classifies each amino acid in the N-terminal sequence as belonging/not-belonging to a targeting signal. The output from the first layer is fed into a second layer, which classifies the overall protein according to whether or not it contains a targeting signal.

TargetP [15], which was based on SignalP and ChloroP, uses portions of its two predecessors to assign a subcellular location to a protein. The first layer of TargetP consists of the first layer classifier of SignalP and of ChloroP, joined by a third classifier that classifies each amino acid in a sequence as belonging/not-belonging to a mitochondrial targeting peptide. The output of all three classifiers is presented to a second layer neural network that classifies a protein as belonging to the *mitochondria*, *chloroplast*, *secretory pathway*, or *Other*. TargetP is currently considered the state-of-the-art of prediction systems among those based solely on N-terminal sequence information. It has demonstrated an overall prediction accuracy of 85% for plant proteins, and 90% for non-plant proteins.

A well-known problem with using neural networks for classification is that it is very difficult to understand why a class is assigned. This difficulty makes it impossible to gain any insight into the biological justification for the assigned location. The more recent iPSORT [4] system classifies proteins into the same categories as TargetP [15], but uses a series of knowledge-based rules that are easy to interpret. These rules capture features of the amino acids associated with certain locations, based on the AAindex database [17] of amino acid properties. While iPSORT's accuracy does not reach that of TargetP, the rules used to create the classifier are easy to understand and follow.

2.4.2 Amino Acid Composition Based Classifiers

ProtLoc [8] was one of the earliest systems based on amino acid composition. As has become standard for systems of this type, proteins are represented by a 20-dimensional feature vector. Each feature represents one of the 20 different amino acids, and holds the relative frequency of the corresponding amino acid in the protein sequence. ProtLoc predicts the location of a target protein by comparing its amino acid composition to that of proteins of known location. The target protein is assigned to the same location as the protein with the most similar composition vector.

Reinhardt and Hubbard [41] were the first to apply neural networks, trained on proteins represented by amino acid composition, to predicting protein subcellular location. Like TargetP, their system, NNPSL, classifies proteins into three or four subcellular locations, depending on the organism.

In their system, SubLoc, Hua and Sun [23] introduced support vector machines (SVMs) as a classification scheme for subcellular location. Trained on the same dataset as NNPSL, SubLoc was compared to neural networks and hidden Markov model systems, (both based on amino-acid composition), and outperformed them both. Hua and Sun [23] also used SubLoc to demonstrate the robustness of basing classifiers on amino acid composition compared with N-terminal sequence information. To do so, SubLoc was trained using protein sequences lacking a portion of their N-terminal sequence, in order to mimic the often incomplete amino acid sequence at the N-terminal region of proteins that are discovered through large sequencing projects. Despite the lack of an N-terminal sequence, the performance of the system remained essentially the same.

Park and Kanehisa followed in 2003 with the introduction of the PLOC [38] prediction system. PLOC uses SVMs to classify proteins into 12 different locations – the 11 described in Table 2.1.1, and the cytoskeleton, which is a support structure for the cell. PLOC represents a protein based on its on amino acid, amino acid pair, and gapped amino acid pair compositions. An amino acid pair is a pair of consecutive amino acids, whereas a gapped pair is two amino acids separated by one or more intervening amino acids. Park and Kanehisa considered gaps of up to three intervening amino acids. The motivation behind this representation was to capture the effects of order within the sequence. In particular, gapped amino acid pair compositions were included in order to detect periodic co-occurrences of certain amino acids. With an overall accuracy of 78.5% and 79.6% on plant proteins and animal proteins, respectively, PLOC remains the most successful prediction system based solely on composition data.

These accuracies were the best reported at the time, but still left room for improvement. Although amino acid composition influences protein localization, it cannot be used exclusively to determine a protein's location [8]. Several other factors play an important role in subcellular localization. As a result, techniques incorporating multiple forms of protein information have been developed.

2.4.3 Integrated Classifiers

PSORT [34], one of the earliest protein subcellular location prediction systems, was also one of the first to incorporate multiple types of protein information. Introduced in 1992, PSORT incorporates overall amino acid composition, N-terminal targeting sequence data, and sequence motif data into its prediction system. Using a knowledge base of if-then rules, which were derived either computationally or experimentally, PSORT classifies a protein into 14 animal and 17 plant subcellular locations. The increase in the number of locations, beyond the 11 listed in Table 2.1.1, is a result of the refined division of some locations into sub-subcellular locations; that is, specific areas within the location. For each of the chloroplast, ER, lysosome, mitochondria, and plasma membrane, finer sub-components were identified within them and considered as potential subcellular locations. Later versions of PSORT improved upon its accuracy by utilizing a probabilistic [21] and a K-nearest neighbour classifier [22].

More recently, the MultiLoc [20] system was introduced, and has been shown to outperform the PSORT and the TargetP classifiers. MultiLoc incorporates information about N-terminal targeting sequences, amino acid composition, and sequence motifs to make its predictions. It is constructed as an assembly of classifiers, where a protein sequence is presented to four different classifiers simultaneously. The output of each is fed into a final classifier that assigns a protein to one of the following eleven locations: chloroplast, cytoplasm, endoplasmic reticulum, extracellular space, Golgi apparatus,

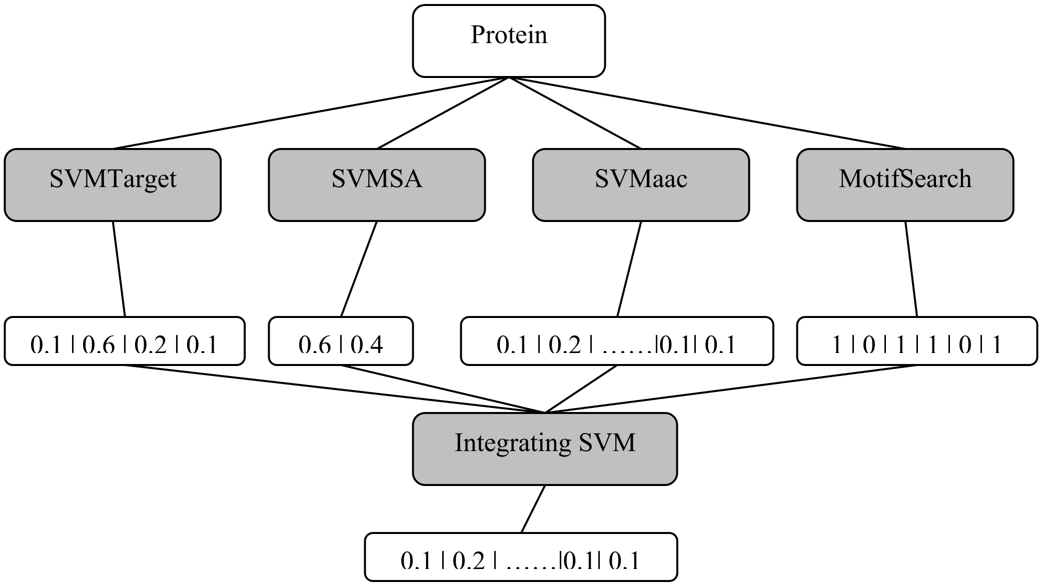


Figure 2.4.1: An overview of the MultiLoc classifier. White boxes represent the input/output of the classifiers. Grey boxes are the classifiers themselves. SVMTarget calculates the probability that a protein belongs to the *chloroplast*, *mitochondria*, *secretory pathway* and *other*. SVMSA calculates the probability that a protein contains a signal anchor. SVMaac calculates the probability that a protein belongs to each location in a dataset. MotifSearch checks for the presence of sequence motifs in a protein sequence. The Integrating SVM calculates the probability that a protein belongs to each location in a dataset.

lysosome, mitochondria, nucleus, peroxisome, plasma membrane, and vacuole. Figure 2.4.1 illustrates the architecture of the MultiLoc classifier.

The four different classifiers of the first layer are organized as follows. The first three classifiers are all SVM-based. SVMTarget is similar to TargetP [15] as it predicts location based on the N-terminal sequence. As with TargetP, SVMTarget classifies proteins from non-plant organisms into three locations, and proteins from plant organisms into four locations. The two classifiers differ in the machine learning methods and in the types of protein information that they are based on. SVMTarget is based on SVMs and the amino acid composition of the N-terminal sequence, while TargetP is based on neural networks and the primary amino acid sequence of the N-terminal sequence. The second classifier, SVMSA, calculates the probability that a protein sequence contains a signal anchor. The third, SVMaac, uses the amino acid composition of a protein to calculate its probability to occur in each of the locations associated with a dataset. SVMTarget, SVMSA, and SVMaac each produces a probability vector, in which each classified item is assigned an *n*-dimensional vector that denotes the item's probability to belong to each of the *n* locations associated with the classifier. The last classifier, MotifSearch, checks the protein sequence for the occurrence of certain motifs, defined in the PROSITE [24] and the NLSdb [31] databases. PROSITE is a database of motifs associated with certain protein families, while NLSdb is a database of nuclear localization signals. The output of MotifSearch is a binary vector in which 1 indicates the presence of a certain motif, and 0 indicates its absence. MultiLoc attains an overall accuracy of 74.6% for both plant and animal proteins, and is considered the state-of-the-art in subcellular location prediction systems based on integrating multiple forms of protein information. Although the results reported for MultiLoc are impressive, they still leave room for further improvement

One possible way to improve the accuracy of protein subcellular location prediction is to include information about the proteins other than their sequence data. In this work, we investigate the effectiveness of using text to train a standalone subcellular location prediction system, and of using a text-based classifier to supplement a sequence-based classifier. The use of text for both standalone and integrated subcellular location classifiers has previously been attempted, as described in the next section.

2.5 Classification using Text Information

An alternative approach to classifying proteins uses textual representations of the proteins. The approach is based on the idea that if there is a passage of text containing information relevant to a protein, then there is often enough information contained therein to deduce the protein's subcellular location. This deduction may be accomplished by recognizing specific words in the text that are indicative of location, or by recognizing the *jargon* that is commonly used when describing proteins from certain locations. Craven and Kumlien [10] demonstrated the possibility of *extracting* protein subcellular location from documents that specifically indicate subcellular location. Other groups have gone a step further to develop systems that predict a protein's subcellular location based on text about the protein, even if the text does not explicitly state its subcellular location.

Eisenhaber and Bork [13] suggested that the functional annotations associated with proteins in the Swiss-Prot [3] protein database can be used to determine the location of a protein. To show that, they created the Meta_A(nnotator) program, a database of if-then rules for classifying proteins into one of ten possible locations. The results of the system were not compared against those of other systems, and the validation was done by checking 4,000 proteins against their correct location assignments. The if-then rules were modified so that the location assignments for the entire set of 4,000 proteins were all correct.

The notion of using text as a means for predicting protein location has since been investigated by several groups. Instead of using rules, most of these groups have represented proteins as vectors of terms based on the text associated with them, and trained a classifier for assigning subcellular location. Systems that use this approach to representing proteins, called the "bag of words" [30] approach, generally differ from each other in the type of text associated with a protein, the terms chosen to represent the protein, and the method used to weight each term within the term-vector.

There are several possibilities for associating text with a set of proteins; any resource that contains documents related to proteins may be used as the source of text. However, when developing a location prediction system, it is important to select a text source that associates text with the majority of the proteins in the dataset, thereby allowing for the majority of the proteins to be represented (that is, achieve a high level of coverage). A system that cannot represent many proteins has little value as a predictive tool. Two resources for text are Swiss-Prot [3] and PubMed [35]. Swiss-Prot is a database containing information about hundreds of thousands of proteins, including their function, subcellular location, etc. PubMed is an online biomedical abstract database that contains the abstracts of millions of scientific articles. Both of the databases have previously been used by text-based systems to predict subcellular location [49, 32].

Once the text for a set of proteins has been gathered, terms must be selected from the text to represent the proteins. This process, known as *feature selection*, is commonly used in a variety of text classification tasks. The goal is to select only those terms that are useful for distinguishing between items from different classes. Feature selection reduces the computational expense of machine learning algorithms, and often improves classification accuracy [55]. Several methods for feature selection have previously been proposed, some of which are described in Chapter 5.

Following the feature selection step, a *weighting scheme* for representing each term within a protein's term-vector must be chosen. The simplest weighting scheme assigns a binary weight, where 1 indicates the presence of a term within the protein's associated text and 0 indicates its absence. Another scheme, term frequency (*tf*), assigns a weight based on the term's frequency of occurrence within the protein's associated text. A third, and the most common weighting scheme, is referred to as *tf·idf* (term frequency times inverse document frequency) [27]. This scheme measures how important a term is in the text associated with a protein. The importance increases according to the number of times it occurs in the text associated with the protein, and decreases according to the number of times it occurs in the text associated with the entire protein dataset.

Several recent studies have investigated text-based prediction of subcellular location. Stapley *et al.* [49] used SVMs to classify a set of yeast proteins into their respective locations. Text for each protein was chosen as the PubMed abstracts that contained the protein's gene name; this method of text association attained a high level of coverage. The vector for each protein was generated by using the *tf·idf* weighting scheme, without applying any feature selection. Stapley *et al.* compared their text-based system to an amino acid composition based system that they also trained, and found the former to perform better. They also compared their text system to a combined text- and sequence-based classifier, but did not find significant improvement over the text-based system alone. These results were not compared against those of any other state-of-the-art system, but did not appear to show an improvement over the state-of-the-art at that time.

Nair and Rost also used text for the classification of proteins, developing the LOCkey [32] classifier. They associated with each protein the functional keywords found in its corresponding Swiss-Prot entry [3]. A feature selection scheme, which is described in Chapter 5, was applied, and a binary weighing scheme was used to create the vector for each protein. Proteins were classified according to their vector's similarity to vectors associated with proteins of known location. Selecting only functional keywords to generate vectors greatly limited the coverage of the system, since many Swiss-Prot entries lack such keywords. For the proteins for which a vector could be generated, the results appeared to be compatible with the state-of-the-art at the time. The system, however, was not compared with any other system or dataset, making it difficult to assess its relative effectiveness.

Eskin and Agicthein [16] expanded on LOCkey [32] with a system that combined protein sequence and text information to create a classifier. Starting with a dataset of proteins of which only a small subset had known locations, they used a text-based classifier similar to LOCkey to increase the number of proteins with an assigned location. However, unlike Nair and Rost, who considered only functional keywords, Eskin and Agicthein incorporated all available textual annotations in a protein's Swiss-Prot entry. They then used this expanded dataset to train a joint sequence- and text-based SVM classifier. To represent each protein in the dataset, they used the spectrum method [28], which represents proteins as sets of their amino acid subsequences of a fixed length. The reported results did not demonstrate improvement over previous systems, nor did they indicate that integrating text with a sequence based classifier improves performance.

In this work, we introduce a new text-based subcellular location prediction system. It produces results comparable to those of other state-of-the-art systems, and when integrated with a sequence-based system (MultiLoc [20]), significantly improves on the current state-of-the-art. The next chapter describes in detail both our text-based system and the integrated text- and sequence-based system.

Chapter 3

EpiLoc and SherLoc

This chapter describes two new systems for predicting protein subcellular location. The first is the text-based system, named *EpiLoc*, and the second is the integrated system that uses both sequence and text data, named *SherLoc*. SherLoc is a combination of an early version of EpiLoc and the sequence-based classifier, MultiLoc [20, 19, 47].

We begin by fully describing the EpiLoc system, which includes two components: a primary method for representing a protein with text, and a machine learning method for predicting subcellular location. Next, we define the measures for evaluating the performance of EpiLoc and SherLoc. We then explain the method used to combine the early version of EpiLoc with MultiLoc, in order to produce SherLoc. Finally, we describe three methods used to assign text to a protein when the primary method of EpiLoc cannot do so.

3.1 Protein Representation

The common approach to representing a protein with text, as discussed in Section 2.5, is the "bag of words" approach. It involves creating a vector of terms to represent a protein. The terms used in the representation are referred to as *features*, and are selected from text associated with the set of proteins. The three main steps in this process are: the selection of text to associate with the set of proteins, the selection of important features from that text, and the weighting of these features with respect to each protein. The sections that follow describe in detail the text association process and the weighting scheme. The feature selection method is presented in Chapter 5.

3.1.1 Text Association

EpiLoc's primary method for associating text with a protein involves two steps: collecting text from a text source, and processing that text to produce a set of terms that are useful for classification. These two steps are discussed below.

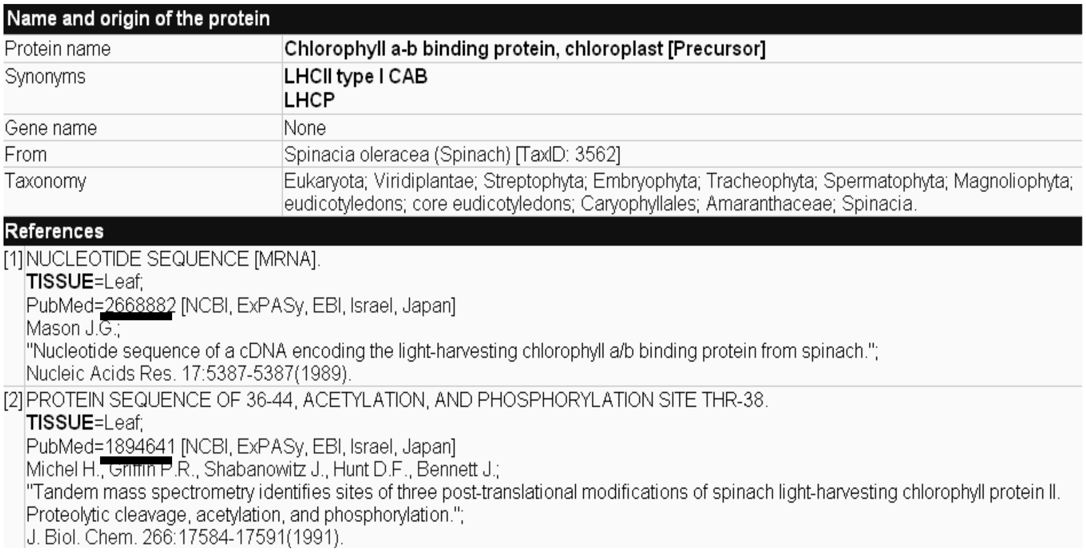
Text Source: PubMed via Swiss-Prot

Several sources of text information related to proteins are readily available, as discussed in Section 2.5. Depending on the source of the text associated with a set of proteins, the effectiveness of the representation of each protein may vary. Nair and Rost [32] associated with each protein the keywords of functional annotations found in a protein's Swiss-Prot [3] entry. However, the entry of many of the proteins in their dataset did not contain such keywords, and as a result these proteins could not be represented. Nair and Rost were able to represent less than 36% of their protein set. Stapley *et al*. [49] associated with each protein in their dataset the PubMed [35] abstracts that contained a protein's corresponding gene name. However, this approach to text association may incorporate abstracts that do not contain information pertinent to the proteins being represented. By selecting all abstracts that contain a protein's gene name, the resulting set may include abstracts that contain the gene name, but no information about the protein.

We attempt to select a large enough amount of text to represent the majority of the proteins in the dataset, while including only text that contains information pertinent to the proteins themselves. To this end, we select as the source of text for each protein the set of PubMed [35] abstracts referenced by its Swiss-Prot entry. Selecting these abstracts produces a set of *authoritative* abstracts for each protein, as determined by Swiss-Prot curators; the reference to a PubMed abstract by a Swiss-Prot entry implies that the document for which the abstract was written contains information directly relevant to the protein.

We note that the abstracts do not necessarily discuss localization – but rather are authoritative with respect to the protein in general. If proteins were only associated with abstracts that explicitly stated their location, the system would not have as much value, as it would only be able to provide the location of proteins for which a location is already known.

The set of abstracts for each protein is gathered by scanning Swiss-Prot [3] and extracting all PubMed [35] references. These references are given in the form of PubMed identifiers (PMIDs). Figure 3.1.1 illustrates a Swiss-Prot entry and the corresponding PMIDs associated with a protein.



Name and origin of the protein	
Protein name	**Chlorophyll a-b binding protein, chloroplast [Precursor]**
Synonyms	**LHCII type I CAB** **LHCP**
Gene name	None
From	Spinacia oleracea (Spinach) [TaxID: 3562]
Taxonomy	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; Caryophyllales; Amaranthaceae; Spinacia.

References

[1] NUCLEOTIDE SEQUENCE [MRNA].
TISSUE=Leaf;
PubMed=2668882 [NCBI, ExPASy, EBI, Israel, Japan]
Mason J.G.;
"Nucleotide sequence of a cDNA encoding the light-harvesting chlorophyll a/b binding protein from spinach.";
Nucleic Acids Res. 17:5387-5387(1989).

[2] PROTEIN SEQUENCE OF 36-44, ACETYLATION, AND PHOSPHORYLATION SITE THR-38.
TISSUE=Leaf;
PubMed=1894641 [NCBI, ExPASy, EBI, Israel, Japan]
Michel H., Griffin P.R., Shabanowitz J., Hunt D.F., Bennett J.;
"Tandem mass spectrometry identifies sites of three post-translational modifications of spinach light-harvesting chlorophyll protein II. Proteolytic cleavage, acetylation, and phosphorylation.";
J. Biol. Chem. 266:17584-17591(1991).

Figure 3.1.1: A Swiss-Prot entry [3]. The underlined numbers are the PMIDs that reference abstracts in the PubMed database.

Text Processing

We begin text processing by removing abstracts that are referenced by the Swiss-Prot entries of proteins from three or more locations. This removal is performed to facilitate the next step in the protein representation process, feature selection. The objective of feature selection is to select those terms that are useful for distinguishing one class from another. For subcellular location prediction, this involves selecting terms that distinguish between proteins from different locations. An abstract that is associated with many locations is not useful for obtaining terms that can characterize a single location. However, if we require each abstract to be associated with only a single location, we may be left with a set of abstracts that is too small to represent many of the proteins in the dataset. Furthermore, some proteins may actually be found in more than one location within a cell. Therefore, we associate with a protein those abstracts that are associated with at most two locations.

Once abstracts associated with three or more locations are removed, a local version of PubMed [35] is scanned to gather the remaining abstracts. For each PubMed entry, the title and the text of the abstract are retained. The abstract is parsed into a set of terms consisting of single words (unigrams) and pairs of consecutive words (bigrams). Following the weighting scheme used in PubMed's search engine, terms that occur in the title are counted twice: once as a part of the overall abstract, and once as a part of the title, practically assigning more weight to title terms, as they provide important information about the subject of a document. Additionally, a list of standard stop words, shown in Appendix A, is removed from the set of terms. The list consists of very common terms such as prepositions and articles, which have little value for distinguishing between proteins from different locations.

Porter Stemming [39] is applied next, to strip suffixes off terms so that different variations of the same term are coalesced into a single form. This is a standard step in many document classification systems, done in order to reduce the size of the feature space, as well as to expose connections between terms with similar semantics. For example, the two terms "connects" and "connecting" have similar semantics but slightly different forms. The use of stemming reduces them both to the form "connect", giving rise to a single semantic term.

Last, terms that occur in fewer than three abstracts or in more than 60% of all abstracts are removed; a term that occurs in fewer than three abstracts cannot be used to represent the majority of the proteins in the dataset, while terms that occur in more than 60% of all abstracts are likely to have little discriminative value.

Feature selection is applied to those terms that remain after the above term reduction steps. For this work, we use a feature selection method based on the Z-test [51]. The method used can greatly influence the classifier's performance [55, 44]. As such, we compare our feature selection method with several others, and present it, along with the results of the comparison, in Chapter 5.

3.1.2 Term Weighting

Our feature selection method selects a set of N terms, denoted T_N, that is helpful for *distinguishing* between different locations. Using these terms, each protein, p, is represented as a vector of length N, $< w_1^p \ldots w_N^p >$, where each value, w_i^p, in the vector denotes the weight of term, t_i. A weighting scheme was developed that represents each term by its significance, relative to the other distinguishing terms, within the abstracts associated with the protein. For a protein p, the weight w_i^p of term t_i at position i is defined as the probability of term t_i to occur in the abstracts associated with the protein p (the set of abstracts D_p). This probability, denoted $\Pr(t_i \mid D_p)$, is estimated as the ratio between the total number of occurrences of term t_i in D_p and the total number of occurrences of *all* distinguishing terms in D_p. Formally, each weight is calculated as:

$$W_t^{\,p} = \frac{\#\,\text{of times}\, t_i \,\text{occurs in}\, D_p}{\sum_{t_j \in T_N} (\#\,\text{of times}\, t_j \,\text{occurs in}\, D_p)},$$

where the sum is taken over all the terms t_j in the set of distinguishing terms T_N.

The above approach is used to represent the proteins in the dataset. In order to determine the effectiveness of this representation, we develop a classifier based on it and measure the classifier's performance.

3.2 Training and Testing a Classifier

Our classifier (also referred to as a *prediction system* or *predictor*) uses the LIBSVM [9] implementation of support vector machines (SVMs). LIBSVM supports soft, probabilistic categorization for *n*-class tasks [54], in which each classified item is assigned an *n*-dimensional vector that denotes the item's probability to belong to each of the *n* classes. Here *n* is the number of subcellular locations.

The classifier is trained on proteins that have been represented using the method described in Section 3.1. The performance of the predictor, evaluated through 5-fold cross-validation, is compared to that of several other state-of-the-art prediction systems. The cross-validation scheme and the choice of LIBSVM as the backbone of our classifier follow the design of the sequence-based classifier MultiLoc [20], as we developed our text-based classifier with the intent of integrating it with MultiLoc. The following sections describe the SVM method, the performance metrics, and the cross-validation scheme employed to develop and to test the EpiLoc classifier.

3.2.1 Support Vector Machines

Support vector machines are an example of a supervised learning method. In supervised learning, the dataset is split into a training set and a test set. The classifier is built based on the training set; the classifier "learns" the aspects of the data that will allow it to classify a sample. The test set is used to determine the effectiveness of the classifier; it predicts the class of the test samples, and the predicted classes are compared against the actual classes, in order to estimate the classifier's performance.

Support vector machines attempt to construct a hyperplane that separates two classes of vectors. To do so, they implicitly map a set of training vectors into a higher dimensional space defined through a *kernel* function. The kernel function denotes a similarity measure between vectors, which corresponds to a dot product between vectors in a higher-dimensional space, in which the calculated separating hyperplane is embedded. A vector whose class is unknown is classified according to its location relative to the hyperplane; it is assigned to the same class as that of the vectors on its side of the hyperplane.

There may be infinitely many hyperplanes that separate two classes of vectors. SVMs try to select the hyperplane whose distance to the nearest training vectors of each class is maximal. The distance between the selected hyperplane and the nearest vectors is called the *margin*, and the hyperplane itself is referred to as the *maximum margin hyperplane* (Figure 3.2.1). The vectors closest to the hyperplane are called *support vectors*, hence the name support vector machines. When two classes of vectors are not perfectly separable, there is a trade-off between the number of incorrectly classified vectors in the training set, and the size of the margin.

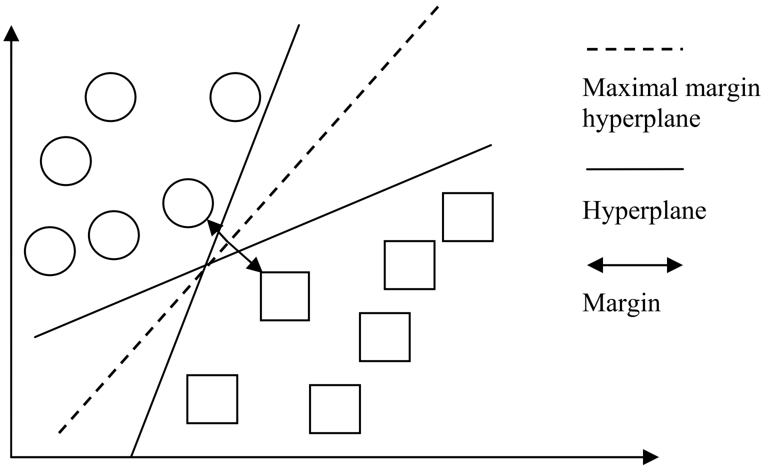


Figure 3.2.1: Two classes separated by three hyperplanes. The dashed line represents the maximum margin hyperplane.

The hyperplane created by an SVM separates only two classes. For a multi-class problem, the results from multiple binary classifiers must be combined. The one-vs-one approach to the multi-class case trains a binary classifier for each pair of classes. For a problem with *n* classes, this results in $\frac{n(n-1)}{2}$ different classifiers. When classifying an unknown sample, it is presented to all classifiers, and their results are combined to make a prediction. LIBSVM's [9] default approach to combining results uses a voting strategy: each binary classification is considered to be a vote for one of two classes, and a vector is assigned to the class that receives the most votes from the set of classifiers. An alternative approach implemented by LIBSVM, and used for EpiLoc, employs estimates of the probability of a vector to belong to a class (See publication by Wu *et al.* [54] for details).

For the EpiLoc classifier, we use the LIBSVM implementation of SVMs with the Radial Basis Function (RBF) kernel, a standard kernel used for classification tasks. The use of the RBF kernel requires the optimization of two parameters, C and γ. The C parameter controls the trade off between the number of errors made on the training data and the size of the margin. The γ parameter controls the width of the RBF kernel function [48]. As the kernel denotes a dot-product in high-dimensional space, when using the RBF kernel to calculate the dot product, increasing the width of the kernel increases the likelihood of two vectors to be considered similar.

3.2.2 Performance Metrics

The performance of previous systems has been measured using several different metrics. We calculate the same measures to allow a fair comparison to other systems. For each location the sensitivity (*Sens*), specificity (*Spec*), and Matthew's Correlation coefficient (*MCC*) [29] are calculated. These are formally defined as:

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

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

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. Sensitivity measures a classifier's ability to recognize all samples belonging to a class, while specificity assesses its ability to correctly identify *only* those samples belonging to a given class and exclude those from other classes; *MCC* is a combination of the two measures. We also measure the *overall accuracy* of a system, denoted *Acc*, defined as *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*, denoted *Avg*, over all locations, giving an equal weight to each location's sensitivity, regardless of the number of proteins associated with the location.

3.2.3 Cross-Validation

EpiLoc is tested on several different datasets using 5-fold cross-validation. In 5-fold cross-validation, the dataset is randomly partitioned into five equal subsets. This partitioning is referred to as a *split*. We use *stratified* cross-validation, where each subset maintains the same distribution of classes as is found in the whole dataset. The classes of the samples in each subset are predicted by a classifier trained on the remaining four subsets; the subset for which predictions are made is the test set, and the remaining four subsets comprise the training set. The predictions for each test set are combined and compared to their actual classes, to measure the system's performance on the entire dataset. This cross-validation process is implemented in the same manner for each dataset, as described next.

For a given dataset, each location's proteins are distributed uniformly at random among the five subsets. If the number of proteins associated with a certain location is not divisible by five, then the number of proteins equal to the remainder is excluded from the training process.

Once the dataset is partitioned, feature vectors are created for the proteins in each training set and its corresponding test set. To do so, a set of distinguishing terms is first selected from each training set. The distinguishing terms are used to represent each of the proteins in the training set and in the test set as feature vectors (as described in Section 3.1.3). After the feature vectors are created, SVM training begins.

The feature vectors are used to train the $\frac{n(n-1)}{2}$ classifiers required for the one-vs-one approach to multi-class classification. Each pair of locations has an associated classifier. For example, if there were only three locations, e.g. the chloroplast, mitochondria, and nucleus, three binary classifiers would be trained to distinguish between each of the possible pairs: chloroplast vs. mitochondria, chloroplast vs. nucleus, and nucleus vs. mitochondria.

We use an equal number of proteins from each location to train each binary classifier. The number of proteins associated with each location can vary greatly. If a classifier is trained over two locations, and one has a greater number of associated proteins, the classifier may be biased towards the more highly represented location. To prevent this bias, each binary classifier is trained with a balanced set of proteins. Proteins are randomly removed from each of the five subsets associated with the location with the largest number of proteins until these subsets are equal in size to the subset associated with the location with the fewest proteins. Once this equalization step is done, the binary classifiers are trained.

Training of a binary classifier first requires the optimization of the C and γ parameters of the SVM. To do so, a coarse grid search is performed, searching through the same values of C and γ examined when building the MultiLoc [20] classifier. These values are:

$$C = 0.01, 0.1, 0.5, 1, 2, 5, 8, 10, 20, 50, 100, 300, 500, 700, 1000, \text{and}$$

$$\gamma = 0.01, 0.1, 0.5, 1, 2, 5, 10, 100.$$

For all five pairs of training and test sets, a binary classifier is trained with each combination of C and γ values, and is used to predict the locations of the test vectors. The results of all five test sets for a single combination of C and γ are combined, and the *MCC* value is calculated over all test proteins. The values of C and γ that attain the highest *MCC* are selected, and a binary classifier is trained on each of the five training sets. Each binary classifier is thus trained using different values for C and γ.

Once trained, the collection of binary classifiers associated with each training set is assessed based on its ability to correctly classify the corresponding test proteins. Proteins excluded from each location during the initial partitioning prior to training are reintroduced and distributed uniformly at random among the five subsets in the test phase. Each binary classifier assigns to each test protein a probability to occur in each of the two locations associated with the classifier. The probabilities from each classifier in the collection are combined, as described by Wu *et al.* [54], to predict the location of the protein. The predicted classes of the proteins in all five test sets are compared to their actual classes to measure the overall performance of the system.

The EpiLoc system is also trained to classify proteins not included in the cross-validation datasets. The setup of the system is the same, but the training procedure is based on the entire protein set (as opposed to just 4/5 of it). The system is trained using the MultiLoc dataset, which is described in Chapter 4.

The EpiLoc system is capable of acting as a standalone system for predicting protein subcellular location. However, as discussed in Chapter 1, our goal is also to examine the integration of a sequence-based classifier and a text-based classifier. To do so, our EpiLoc classifier is combined with the MultiLoc classifier [20, 19, 47].

3.3 Combined Sequence and Text Classifier: SherLoc

We have discussed above our new text-based classifier for predicting subcellular location. In this section, we describe the integrated sequence- and text- based predictor called SherLoc [19, 47].

The sequence-based classifier, MultiLoc [20], was described in Section 2.4.3. It combines four classifiers, SVMTarget, SVMSA, SVMaac, and MotifSearch, to create one comprehensive classifier for protein location prediction. The SVM-based components produce probability vectors, while MotifSearch produces a vector of binary values. These vectors are concatenated to create a single vector that forms the input to a final classifier, the Integrating SVM, which predicts the location of a protein.

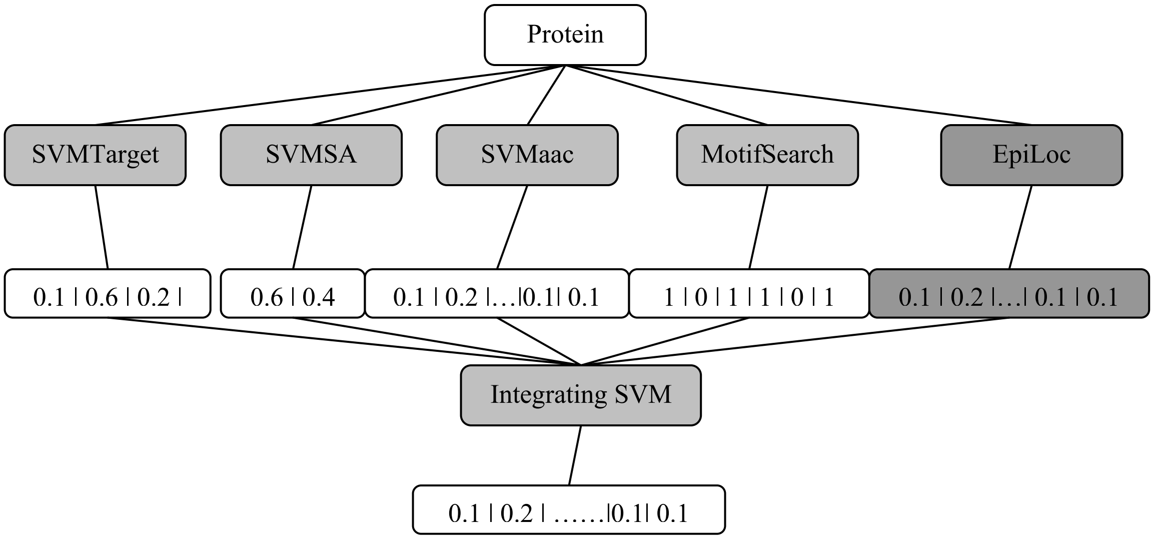


Figure 3.3.1: Overview of the SherLoc system. The output vector produced by EpiLoc is concatenated to the vectors produced by the MultiLoc classifiers forming a single vector. The dark grey boxes indicate the text components that we introduce into the system.

MultiLoc is then combined with an early version of the EpiLoc classifier, which we call *EarlyText*[2]. Combining MultiLoc with EarlyText entails concatenating the probability vector produced by EpiLoc to the vector output from the first four classifiers of MultiLoc, and training a new Integrating SVM to accept as input the combined vectors of the two systems. Figure 3.3.1 provides an overview of the SherLoc classifier [19, 47].

SherLoc was implemented early on in the development of EpiLoc. As a result, the version of EpiLoc incorporated in SherLoc, EarlyText, predates the version that we use to test EpiLoc as a standalone classifier. Changes have been made to EpiLoc since the testing of SherLoc was completed, all of which involve the feature selection process, and are discussed in Chapter 5.

[2] The integration of the two predictors, including the training of EarlyText from our term-vectors, was performed by the group that built the MultiLoc classifier [19].

Both SherLoc and EpiLoc should be able to predict the location of any protein. However, for some proteins, there may be insufficient text to represent them using the primary method of EpiLoc, described in Section 3.1. To make EpiLoc applicable to all proteins, it is necessary to develop additional methods for associating text with proteins. The next section presents three such methods.

3.4 Associating Text with Textless Proteins

Thus far, we have introduced one method for associating text with proteins. Namely, we have associated with each protein the PubMed abstracts referenced by its Swiss-Prot entry. However, this method is not applicable in four situations:

1. A protein has no Swiss-Prot [3] entry. While Swiss-Prot is a very large database of proteins, millions of proteins are still not included in it, either because their data has simply not been entered yet, or they have not been sufficiently studied. Since the text-selection procedure already described depends on the Swiss-Prot entry for references to PubMed, the lack of a Swiss-Prot entry prevents it from being applied.
2. A protein's Swiss-Prot entry contains no references to PubMed.
3. The Swiss-Prot entry does reference PubMed articles, but these references are all shared by other proteins from three or more different locations. In such cases the PubMed abstracts are not good representatives of any single location, and we do not associate them with the protein.
4. A protein's associated abstracts do not contain any of the selected distinguishing terms.

Situations 1-3 effectively result in a protein having no associated abstracts, while situation 4 results in a protein having no associated terms. These four situations produce what we call *textless* proteins, as the proteins involved effectively have no associated text to represent them. We next describe several methods that we have developed to associate text with such textless proteins.

3.4.1 HomoLoc

If a textless protein has close homologs that already have text associated with them, we use the text of these homologs to represent the protein. As stated in Section 2.4, homologs are proteins that share similar amino acid sequences. Many other characteristics are often shared between homologous proteins, such as structure, function, and subcellular location [52, 33, 5]. However, directly assigning a protein to the subcellular location of its homolog is not always reliable [33], and as a standalone method for location prediction it was shown less effective than other current methods [12]. Nevertheless, homologous proteins do share characteristics, and as such may serve as a good alternative source of textual information for textless proteins. Therefore, our first method, called *HomoLoc*, assigns to a textless protein the text of its homologs.

Homologous proteins are identified using a BLAST search [2]. BLAST is a search program that compares a biological query sequence against a specified sequence database and returns those entries in the database that share a similar subsequence, or region, with the query sequence. Specifically, we use BLASTP, which compares amino acid sequences, against the Swiss-Prot [3] database. A BLASTP search returns with each potential homolog an expectation value (E-Val). The E-Val indicates the expected number of proteins that may be returned by chance given the database being searched.

The lower the E-Val is, the more likely the proteins are to be true homologs. For this thesis, a protein is only considered a homolog if its E-Val is below 10 (the default value used by BLASTP to identify a potential homolog).

While BLAST is an effective tool for retrieving homologous proteins, the E-Val is not always an accurate indicator of protein similarity [33]. To further ensure that proteins are in fact homologs, we require that, for two proteins, the percent of their similar region that is identical, called *percent identity*, is at least 40%. This level of sequence similarity was chosen as a result of a study by Brenner et al. (1998) that suggested that percent identity of 40% or more between regions of similarity in two sequences usually implies the sharing of at least some characteristics between them. Thus, to accept a protein as a homolog from a BLASTP search, we require that the E-Val is below 10, and that the percent identity is at least 40%. We note that HomoLoc may have performed better had we required an E-Val lower than 10, or two sequences to share more than 40% identity over their *entire* length (as opposed to over their similar sequence regions only). Imposing more stringent requirements may have reduced the likelihood of identifying false homologs for a textless protein. However, as will be shown in Section 6.3, the current method used by HomoLoc is still very effective.

We combine the term-vectors of the three top homologs to produce the term-vector for a textless protein. The three vectors are selected as follows: The set of homologs with an assigned E-Val below 10 and a percent identity of at least 40% is gathered. From these homologs, the three with the lowest E-Vals that have an associated term-vector are selected (if there are fewer than three such homologs, then the 1-2 homologs that do meet the requirements are selected). If there are no homologous proteins with associated term-vectors meeting all the threshold requirements, a method other than homology must be used to obtain text for the 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, for a textless protein *p*, the modified weight used by HomoLoc is calculated for each term t_i as:

$$W_{t_i}^{p} = \frac{\sum_{h \in H} \left(\# \; of \;\; times \; t_i \;\; occurs \; in \; D_h\right) \cdot \left(\% \; identity \; of \;\; h\right)}{\sum_{h \in H} \sum_{t_j \in T_N} \left(\# \; of \;\; times \; t_j \;\; occurs \; in \; D_h\right) \cdot \left(\% \; identity \; of \;\; h\right)} \;\; ,$$

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

3.4.2 PubLoc

Proteins whose Swiss-Prot entries do not contain references to PubMed may still have abstracts in PubMed discussing them. PubLoc[3] uses a PubMed search to retrieve abstracts that may contain information about a textless protein. Abstracts that mention a protein may contain information about it; such abstracts may therefore be used as the text-source for producing the term-vector for the protein. To find abstracts relevant to a given protein, a query string is formed and posed to the PubMed database. The query string lists the protein’s name and its gene name, as found in the protein’s Swiss-Prot entry, separated by the “OR” Boolean operator. The five abstracts returned by the search

[3] We thank Annette Höglund of the University of Tübingen for suggesting this name.

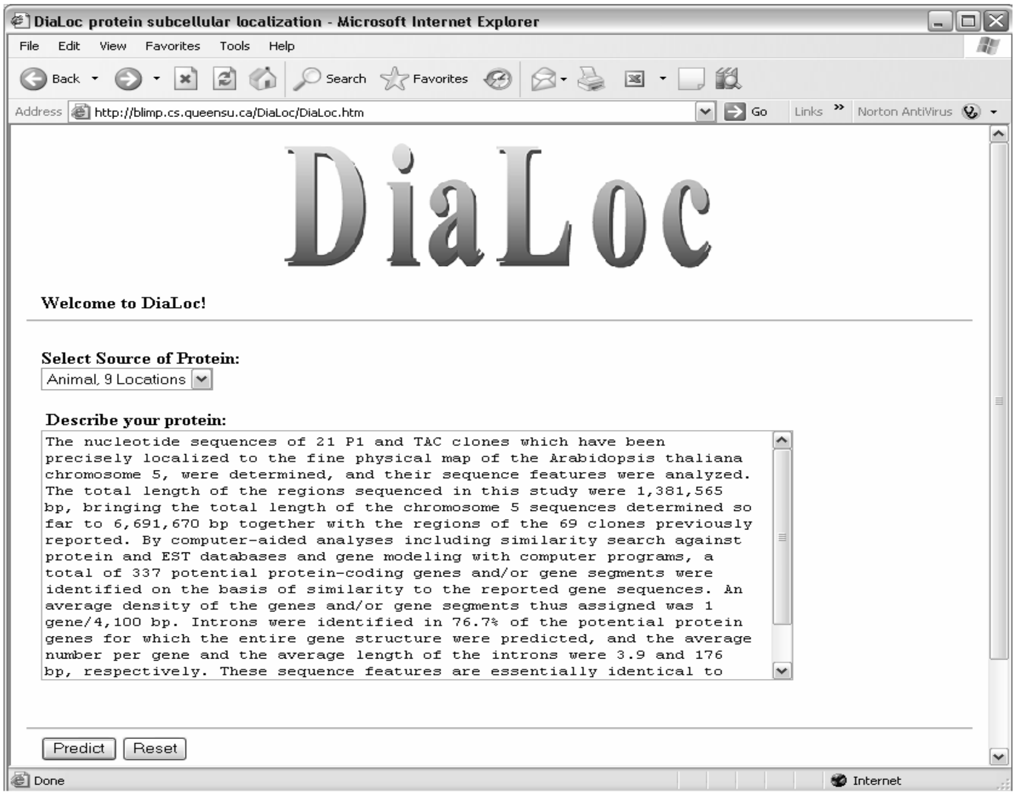
that were most recently entered into PubMed are assigned to the textless protein, and used to produce the protein's feature vector, as described in Section 3.1.

There are situations in which PubLoc cannot associate text with a protein. If a protein does not have a Swiss-Prot entry, or if the entry does not contain a gene or a protein name, no query string can be formed and used to search for text. Furthermore, if PubMed does not contain any abstracts satisfying the query, no text will be returned to be associated with the protein. Both situations result in a protein without a feature vector, and require the consideration of a different method for handling textless proteins.

3.4.3 DiaLoc

The last approach we developed for obtaining text for textless proteins is called DiaLoc. The text-association methods described so far may not be able to assign text to a protein in some situations. Such situations are most likely to occur when a protein has just recently been sequenced, as there is very little information about newly sequenced proteins in databases such as PubMed or Swiss-Prot. If there are no known close homologs for HomoLoc to be effective, the most reliable source of information pertaining to such textless proteins (and the one most likely to be interested in their subcellular location) may be the scientist researching the proteins.

Through a web-interface (Figure 3.4.1) created for this purpose, DiaLoc obtains textual information from a researcher studying a textless protein, and uses it to produce a feature vector to represent the protein. The researcher enters a description of the protein of at least 100 words, and those words take the place of a single abstract (without a title) associated with the protein. The same process described in Section 3.1 is then carried out


DiaLoc protein subcellular localization - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://blimp.cs.queensu.ca/DiaLoc/DiaLoc.htm Go Links Norton AntiVirus
DiaLoc
Predict
Reset
Done
Internet

Welcome to DiaLoc!

Select Source of Protein:
Animal, 9 Locations

Describe your protein:

The nucleotide sequences of 21 P1 and TAC clones which have been precisely localized to the fine physical map of the Arabidopsis thaliana chromosome 5, were determined, and their sequence features were analyzed. The total length of the regions sequenced in this study were 1,381,565 bp, bringing the total length of the chromosome 5 sequences determined so far to 6,691,670 bp together with the regions of the 69 clones previously reported. By computer-aided analyses including similarity search against protein and EST databases and gene modeling with computer programs, a total of 337 potential protein-coding genes and/or gene segments were identified on the basis of similarity to the reported gene sequences. An average density of the genes and/or gene segments thus assigned was 1 gene/4,100 bp. Introns were identified in 76.7% of the potential protein genes for which the entire gene structure were predicted, and the average number per gene and the average length of the introns were 3.9 and 176 bp, respectively. These sequence features are essentially identical to

Figure 3.4.1: The DiaLoc web-interface.

to produce a feature vector for the protein. The researcher may select the type of organism (plant, animal, or fungal) the protein comes from, and the EpiLoc system trained for the specified organism will assign the protein's feature vector to one of the organism's possible locations. DiaLoc is meant to be used as an interactive tool for laboratory research concerned with individual proteins, and not as a tool for large-scale annotation.

Together, HomoLoc, PubLoc, and DiaLoc should essentially allow any protein to be assigned text. The next chapter describes the experiments designed for evaluating our systems. Specifically, in Section 4.3 we describe experiments for testing the effectiveness of the HomoLoc and the PubLoc systems for handling textless proteins.

Chapter 4

Experimental Settings

This chapter presents a set of experiments that evaluate the performance of EpiLoc, SherLoc, and our methods for handling textless proteins. We compare SherLoc and EpiLoc to other state-of-the-art prediction systems using existing datasets. Moreover, we validate the performance of SherLoc and of EpiLoc by testing their performance on proteins outside of the cross-validation studies.

To test HomoLoc and PubLoc, we first compare their performance by applying them to the textless proteins of the MultiLoc dataset. We then select the method that performs best on these textless proteins, and apply it to the entire MultiLoc dataset (excluding the textless proteins), so that the best performing method may be compared to the primary method for associating text with proteins.

4.1 Systems Comparison

To determine the quality of EpiLoc and of SherLoc, the performance of each of the systems is compared to that of several other state-of-the-art classifiers using their respective datasets. We compare both EpiLoc and SherLoc to the MultiLoc [20], PLOC [38], and TargetP [15] subcellular location predictors (all of which were described in Chapter 2), using the same dataset and evaluation procedure (5-fold cross-validation) employed by the three systems. To evaluate EpiLoc, we do not use the same partitions as used to evaluate each of TargetP, PLOC, and MultiLoc, as these partitions include textless proteins, which we do not include in the evaluation of the primary method of EpiLoc. Therefore, for each dataset we randomized the data split five times (on top of the 5-fold cross-validation) to ensure the robustness of the evaluation. Results are averaged over the five different splits of each dataset. SherLoc is compared with PLOC and MultiLoc using the exact same partitions of their datasets. However, the split used to test TargetP was not available, and as such five sets of 5-fold cross-validation are used to compare SherLoc to TargetP. The performance of SherLoc and EpiLoc is compared to that of TargetP, PLOC, and MultiLoc, as reported in their corresponding publications. We next describe the TargetP, MultiLoc, and PLOC datasets used in the comparative study.

4.1.1 TargetP

A total of 3,415 proteins, of which 292 are textless, comprise the TargetP dataset [15]. Proteins are sorted into chloroplast (*ch*), mitochondria (*mi*), secretory pathway (*SP*), and *Other* (*OT*) classes for plant proteins, and mitochondria, *SP*, and *Other* classes for non-plant proteins. The *SP* class includes proteins from the endoplasmic reticulum (*er*), extracellular space (*ex*), Golgi apparatus (*go*), lysosome (*ly*), plasma membrane (*pm*), and vacuole (*va*). The *OT* class includes cytoplasm (*cy*) and nucleus (*nu*) proteins. Table 4.1.1 shows the number of proteins in each location for the TargetP dataset.

Location	**Number of Proteins**
Chloroplast (*ch*)	141
Mitochondria (*mi*)	477
Secretory pathway (*SP*)	983
Other (*OT*)	1,814
Total	**3,415**

Table 4.1.1: The number of proteins per location for the TargetP dataset.

4.1.2 MultiLoc

The MultiLoc [20] dataset consists of 5,959 proteins extracted from Swiss-Prot release 42.0 [3], and includes 614 textless proteins. Proteins originating in animal, fungal, and plant cells with annotated subcellular locations were collected and sorted into eleven different classes: *ch*, *cy*, *er*, *ex*, *go*, *ly*, *mi*, *nu*, *pe*, *pm*, and *va*. Homologous proteins with a sequence identity greater than 80% were excluded from the dataset, as were any proteins with a *SUBCELLULAR LOCATION* line in their Swiss-Prot entry's comment field that contained the words *by similarity*, *potential*, or *probable*. The latter were excluded so that only proteins whose location was certain were included. The number of proteins per location is shown in Table 4.1.2.

Location	Number of Proteins
Chloroplast (*ch*)	449
Cytoplasm (*cy*)	1411
Endoplasmic reticulum (*er*)	198
Extracellular space (*ex*)	843
Golgi apparatus (*go*)	150
Lysosome (*ly*)	103
Mitochondria (*mi*)	510
Nucleus (*nu*)	837
Peroxisome (*pe*)	157
Plasma membrane (*pm*)	1,238
Vacuole (*va*)	63
Total	**5,959**

Table 4.1.2: The number of proteins per location for the MultiLoc dataset.

4.1.3 PLOC

The dataset used to train the PLOC [38] classifier consists of 7,589 proteins (1,076 of which are textless) with a maximum sequence identity of 80%, extracted from Swiss-Prot release 39.0 [3]. The PLOC dataset is made up of the 11 locations comprising the MultiLoc dataset, as well as the cytoskeleton (*cs*) location. In contrast to the MultiLoc dataset, if the annotations *by similarity*, *potential*, or *probable* were included in a protein's *SUBCELLULAR LOCATION* line, the protein was still included in the dataset. Table 4.1.3 shows the number of proteins in each location for the PLOC dataset.

Location	**Number of Proteins**
Chloroplast (*ch*)	671
Cytoplasm (*cy*)	1245
Cytoskeleton (*cs*)	41
Endoplasmic reticulum (*er*)	114
Extracellular space (*ex*)	862
Golgi apparatus (*go*)	48
Lysosome (*ly*)	93
Mitochondria (*mi*)	727
Nucleus (*nu*)	1,932
Peroxisome (*pe*)	125
Plasma membrane (*pm*)	1,677
Vacuole (*va*)	54
Total	**7,589**

Table 4.1.3: The number of proteins per location for the PLOC dataset.

Each of the three datasets, TargetP, MultiLoc, and PLOC, is also divided into subsets that include only proteins from locations that are found in certain organisms. TargetP, as described above, has plant and non-plant subsets of proteins. PLOC and MultiLoc both include plant, animal, and fungal subsets. The plant subsets do not include *lysosomal* proteins, the animal subsets exclude the *chloroplast* and *vacuolar* proteins, and the fungal subsets do not include *lysosomal* or *chloroplast* proteins. The performance of EpiLoc and SherLoc is measured through stratified 5-fold cross-validation, for each of the organisms in each dataset. In the next section, we describe the extraction of a set of proteins that are used to test both EpiLoc's and SherLoc's performance on proteins outside of the cross-validation data.

4.2 *De Novo* Prediction

To further validate the predictive ability of EpiLoc and SherLoc, we use two new datasets [19, 47]. These datasets consist of proteins that were not included in the development of SherLoc and EpiLoc, or in the cross-validation studies. The first dataset, *Diff48*, consists of proteins that either had no assigned location, or were annotated as uncertain in Swiss-Prot release 42.0 (on which EpiLoc and SherLoc were trained) but have since been assigned a definite location in version 48.8[4]. The second dataset, *Unknown*, is formed of proteins with an uncertain or unknown location in Swiss-Prot version 48.8. The proteins in both datasets were required to have an associated PubMed reference, so that a text vector could be created for each protein using the primary method of EpiLoc[5].

The two datasets were created by first extracting from Swiss-Prot release 42.0 all proteins that either did not have a *SUBCELLULAR LOCATION* line, or contained the words *by similarity*, *potential*, or *probable* in that line. Only animal, plant, and fungal proteins were included in the new dataset, as indicated by the presence of the keywords *Metazoa*, *Fungi*, or *Viridiplantae*, respectively, in the Organism Classification (*OC*) field. Any protein that occurred in the MultiLoc dataset was removed from this new set, to ensure that no protein used for training EpiLoc or SherLoc was reused in the new evaluation. Swiss-Prot release 48.8 was then scanned for each of the remaining proteins. Those that were assigned with certainty to a location, as determined by the rules used to

[4] Swiss-Prot version 48.8 was the latest version available at the time of the dataset creation.

[5] SherLoc was developed and tested using an early version of EpiLoc, for which we had not yet developed methods for handling textless proteins.

build the MultiLoc dataset [20], were included in the *Diff48* dataset, for a total of 361 proteins. The rules used to construct the MultiLoc dataset are as follows:

- Proteins are assigned to a subcellular location only if the *SUBCELLULAR LOCATION* line contains the words *cyto*, *nucle*, *lyso*, *endopl*, *plasma*, *perox*i, *mitochon*, *golgi*, *secret* (or *extracellular*), *chlor*oplast, or *vacuol*, corresponding to *cy*, *nu*, *ly*, *er*, *pm*, *pe*, *mi*, *go*, *ex*, *ch*, and *va* locations, respectively.
- Mitochondrial and chloroplast proteins are required to have the keyword *transit*, followed by an annotated cleavage site, in their *FT* (feature) field.
- Secretory pathway proteins (*er*, *ex*, *go*, *ly*, *pm*, and *va*) are required to have the keywords *signal* or *signal-anchor* and annotated start and stop sites, in their *FT* field.
- Plasma membrane proteins are required to have the keywords *domain* and *extracellular* and *domain* and *cytoplasmic* in their *FT* fields. If the keywords *domain* and *luminal* are present in an *FT* fields, the protein is not accepted as a plasma membrane protein.
- The *SUBCELLULAR LOCATION* line cannot include the words *by similarity*, *potential*, or *probable*.

Table 4.2.1 displays the number of proteins in each location for the *Diff48* dataset. There are several locations for which no newly assigned proteins were found. These locations are left out of Table 4.2.1.

Location	**Number of Proteins**
Chloroplast (*ch*)	1
Cytoplasm (*cy*)	91
Endoplasmic reticulum (*er*)	3
Extracellular space (*ex*)	132
Mitochondria (*mi*)	21
Nucleus (*nu*)	111
Vacuole (*va*)	2
Total	**361**

Table 4.2.1: The number of proteins per location for the *Diff48* dataset. Only locations to which proteins had been newly assigned are shown.

The text for each protein in the *Diff48* set comes from the PubMed references listed within the protein's entry in Swiss-Prot release 45.0 or earlier, specifically, the last release in which the protein was not annotated with a location (all preceding release 48.8). Selecting text in this manner ensures that only text that was available before the protein was experimentally localized is used to represent it. To summarize, the *Diff48* proteins have the following characteristics: a) They were not included in any form in training EpiLoc; b) Their location was unknown in the version of the data used to train EpiLoc; c) The PubMed entries associated with the proteins *predate* the protein localization time; d) The location of these proteins is now known.

Proteins that remained either without an assigned location or with an uncertain location assignment (in Swiss-Prot release 48.8) form the *Unknown* dataset. This dataset contains 19,498 proteins, of which 14,890 have no known location, while the location of 4,608 proteins is annotated as uncertain.

The performance of both SherLoc and EpiLoc is measured on the *Diff48* set of proteins. SherLoc is also used to predict the location of the 19,498 proteins in the *Unknown* dataset; the accuracy of these predictions can only be validated once the proteins are experimentally localized in the lab.

4.3 Testing HomoLoc and PubLoc

In Section 3.4 we presented two methods, PubLoc and HomoLoc, for associating text with textless proteins. Our goal is for one of these methods to serve as the preferred method for handling textless proteins when large-scale annotation is required[6]. To select the preferred method, we compare the effectiveness of both HomoLoc and PubLoc at associating text with the textless proteins from the MultiLoc [20] dataset; the method shown most effective is selected. To assess how effective the preferred method is, we compare its performance to that of EpiLoc's primary method (the use of PubMed abstracts referenced by Swiss-Prot).

HomoLoc and PubLoc are compared using the textless proteins from the MultiLoc dataset. The number of textless proteins from each location of the MultiLoc dataset is displayed in Table 4.3.1. 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 these proteins using the trained system. We also compare the versions of HomoLoc and PubLoc described in Section 3.4 with simpler versions of these same methods in order to determine if the more complicated versions indeed improve performance. The HomoLoc method described in Section 3.4.1,

[6] DiaLoc is not meant for large-scale annotation.

Location	**Number of Proteins**
Chloroplast	101
Cytoplasm	121
ER	35
Extracellular space	22
Golgi Apparatus	10
Lysosome	5
Mitochondria	67
Nucleus	152
Peroxisome	22
Plasma Membrane	65
Vacuole	14
Total	**614**

Table 4.3.1: The number of textless proteins per location for the MultiLoc dataset.

which combines the vectors of the top three homologs of a protein to produce a term-vector, is compared against a version that uses only the single top homolog. The version of PubLoc described in Section 3.4.2, which uses the five most recent abstracts returned by a PubMed search to produce a term-vector, is compared against a version that uses only the three most recent abstracts returned. We refer to these simpler versions of HomoLoc and PubLoc as *SimpHom* and *SimpPub*, respectively. The method whose resulting representation of the textless proteins leads to the most accurate classification of the proteins is selected as the preferred method for handling textless proteins. As will be shown in Chapter 6, the method is HomoLoc.

We next compare the performance of HomoLoc to that of EpiLoc's primary method for associating text with proteins. We perform this comparison in order to determine if HomoLoc is indeed suitable for handling textless proteins. The comparison is made over the MultiLoc dataset using 5-fold cross-validation, and we employ the same splits of the data, and the same classifiers trained on those splits, that were originally used to measure the performance of the EpiLoc system. To test HomoLoc, we remove the text associated with the proteins in each of the five test subsets used for the cross-validation of EpiLoc. Each protein in each test subset is then assigned the text of its homologs by HomoLoc, without considering the protein's own Swiss-Prot entry. The pre-trained classifiers predict the location of the test proteins based on these representations. We compare the results of the predictions to those obtained for the test proteins represented using the primary method.

The results of the experiments described in this section, and of all other experiments described in this chapter, are presented in Chapter 6. Before presenting these results, we discuss our feature selection method in the next chapter.

Chapter 5

Feature Selection

We present here our feature selection method, which is based on the Z-test [51], and refer to it as the *Z-Test* method. We begin by fully defining the *Z-Test* method. Next, we describe an experiment designed to compare the *Z-Test* method with several other feature selection methods, and analyze the results of the experiment. Finally, we examine the differences between the feature selection method used in the mature version of EpiLoc and the earlier version of EpiLoc incorporated in the SherLoc classifier [19, 47].

5.1 The *Z-Test* Method

For text classification tasks, the goal of feature selection is to reduce the number of terms in a corpus while retaining those terms that best differentiate between classes. Including every feature may result in high-dimensional vectors that render many machine learning algorithms ineffective. Reducing the number of features alleviates this problem, and may even lead to improved performance [55].

In this work, we select what we call *distinguishing terms* as our features. A term is considered distinguishing for a location L if the likelihood of finding it in the abstracts associated with location L is significantly different from that of finding it in the abstracts associated with all other locations. In order to compare these likelihoods, terms are scored with respect to each subcellular location. The scoring method and the means for comparing scores are formalized in the following paragraphs.

Let t be a term, p a *protein*, L a location, and d an abstract. 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, and the set of abstracts associated with L, denoted D_L, is the set of all abstracts associated with the proteins that are localized to L. Formally, this set is defined as:

$$D_L = \bigcup_{p \in P_L} \{d \mid d \in D_p\} .$$

The number of abstracts in this set is denoted by $|D_L|$. The likelihood of the term t to occur in the abstracts associated with location L is represented by the probability of term t to be associated with L, denoted $\Pr(t \mid L)$. Formally, $\Pr(t \mid L)$ is the conditional probability of the term t to appear in an abstract, given that the abstract is associated with the location L, expressed as:

$$\Pr(t \mid L) = \Pr(t \in d \mid d \in D_L) . \tag{1}$$

A maximum likelihood estimate of this probability is the proportion of abstracts containing the term t out of all those associated with location L, calculated as:

$$\Pr(t \mid L) \approx \frac{\#\,\text{of abstracts}\; d \in D_L \quad s.t.\;\; t \in d}{|D_L|} \quad ,$$

where both the numerator and the denominator are estimated from the set of abstracts associated with location *L* in our dataset, denoted D_L. This probability estimate is calculated for each term *t* and each location *L*.

Based on the above formulation, a term *t* is deemed distinguishing for location *L*, if and only if its probability to occur in the abstracts associated with location *L* is significantly different from its probability to occur in the abstracts associated with any other location. To determine the significance of the difference between the two probabilities, a statistical test is employed that utilizes the Z-score [51]. The test scores the difference between two binomial probabilities; in this case, the probabilities of term *t* to occur in the abstracts associated with locations *L* and L', denoted $\Pr(t \mid L)$ and $\Pr(t \mid L')$, respectively. The Z-score is defined as:

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

The value of $\left|Z_{L,L'}^{t}\right|$ indicates the statistical significance of the difference between the two probabilities. For instance, if the Z-score calculated with respect to two probabilities is greater than 1.96 or less than -1.96, there is a confidence of 95% that the difference between the two probabilities is not arbitrary and can be considered statistically significant. Therefore, a term *t* is considered distinguishing for a location *L* if for any other location *L'*, $\left|Z_{L,L'}^{t}\right|$ is greater than a predetermined threshold. The precise thresholds we use for each dataset are presented in Section 5.2.4.

5.2 Comparison of Feature Selection Methods

The *Z-Test* method is one of several possible approaches to feature selection. As such, its performance needs to be compared with that of other feature selection techniques used in practice. We select a different feature set using each of the following methods: *odds ratio*, *Chi-squared*, *mutual information*, *information gain*, and *Entropy* – which is part of the LOCKEY classifier [32] discussed in Section 2.5. Each feature set is used to produce a representation of the proteins in the MultiLoc dataset (described in Section 4.1.2). A classifier is then trained and tested on each protein set representation, and the results from each classifier are compared to the results of a classifier that uses *Z-Test* for feature selection. The following sections describe in detail each of the feature selection methods, the comparison process, and the results of this comparison. We also examine adjusting the threshold for the *Z-Test* method when it is applied to different datasets.

5.2.1 Feature Selection Methods

For text classification tasks, feature selection methods score each term in a set such that terms with the highest scores are selected as features. We describe next four standard methods for feature selection: odds ratio, Chi-squared, mutual information, and information gain. We then describe the Entropy method.

Standard Scoring Methods

Each of the techniques discussed here captures an aspect of the term distributions that is useful for selecting distinguishing terms. Odds ratio (OR) measures the degree of association between two variables, in this case a term, *t*, and a location, *L*, while Chi-squared (χ^2) measures the amount of dependence between the two variables [55].

Information gain (IG) and mutual information (MI) incorporate ideas from Shannon's information theory to select terms. Information gain measures the amount of information gained about the location by knowing whether a term is present or absent in an abstract. Mutual information measures the amount of information added about one variable when the other is known, and vice versa. Each method scores a term *t* with respect to a location *L*, and is formally defined [44] in Table 5.2.1. The probabilities incorporated in the scoring functions are defined below. If any of the probabilities discussed throughout this chapter are calculated to be zero, they are set, instead, to 1 x 10 $^{-9}$. Using the latter value instead of zero reflects the fact that the calculated probabilities are estimates based on a limited set of data, as in reality these probabilities are not expected to be zero. We use the value 1 x 10 $^{-9}$ because it is sufficiently smaller than the value of any of the probability estimates that we calculate based on data, and is therefore guaranteed to represent the lowest possible probability in our system.

Function	Mathematical Form
Information Gain	$\Pr(t,L)\log\frac{\Pr(t,L)}{\Pr(L)\cdot\Pr(t)}+\Pr(t,L)\log\frac{\Pr(\bar{t},L)}{\Pr(L)\cdot\Pr(\bar{t})}$
Mutual Information	$\log\frac{\Pr(t,L)}{\Pr(t)\cdot\Pr(L)}$
Chi-squared	$\frac{\lvert Tr\rvert\cdot\left[\Pr(t,L)\cdot\Pr(\bar{t},\bar{L})-\Pr(t,\bar{L})\cdot\Pr(\bar{t},L)\right]^2}{\Pr(t)\cdot\Pr(\bar{t})\cdot\Pr(L)\cdot\Pr(\bar{L})}$
Odds Ratio	$\frac{\Pr(t\mid L)\cdot\left(1-\Pr(t\mid\bar{L})\right)}{\left(1-\Pr(t\mid L)\right)\cdot\Pr(t\mid\bar{L})}$

Table 5.2.1: The mathematical form of information gain, mutual information, Chi-squared, and odds ratio, defined through probabilities, where *t* is a term, *L* is a location, and *Tr* is the total number of terms [44].

A number of prior probabilities are estimated by the feature selection methods. The prior probability that an abstract contains the term *t*, denoted $\Pr(t)$, is estimated. A maximum likelihood estimate for this probability is the proportion of the number of abstracts containing term *t* among all abstracts, calculated as:

$$\Pr(t) \approx \frac{\#\, of\ \ abstracts\ containing\ t}{total\ \#\, of\ \ abstracts}\,.$$

The prior probability that an abstract does not contain term *t*, $\Pr(\bar{t})$, is calculated, as $1 - \Pr(t)$.

The prior probability of an abstract to be associated with a location, *L*, denoted $\Pr(L)$, is estimated as the proportion of abstracts associated with location *L* among all abstracts, and is defined as:

$$\Pr(L) = \frac{\#\, of\ \ abstracts\ associated\ with\ L}{total\ \#\, of\ \ abstracts}\,.$$

The prior probability of an abstract to not be associated with location *L*, $\Pr(\bar{L})$, is calculated as $1 - \Pr(L)$.

Each of the feature selection methods incorporates conditional probabilities in their scoring function. The probability of a term to occur (or not to occur) in an abstract is estimated under two conditions: given that the abstract *is associated with* location *L*, and given that the abstract *is not associated with* location *L*. The conditional probability of a term to appear in an abstract, given that the abstract is associated with location *L*, was already defined in Section 5.1 Eq. (1). It is the same probability used by the *Z-Test* method:

$$\Pr(t \mid L) = \Pr(t \in d \mid d \in D_L)\ ,$$

where t, d, and D_L denote a term, abstract, and set of abstracts associated with L, respectively. The probability is estimated by:

$$\Pr(t \mid L) \approx \frac{\#\,\text{of abstracts}\, d \in D_L \quad s.t.\ \ t \in d}{|D_L|}.$$

The conditional probability of a term not to occur in an abstract, given that the abstract is associated with location L, is formally defined as:

$$\Pr(\bar{t} \mid L) = \Pr(t \notin d \mid d \in D_L),$$

and is calculated as:

$$\Pr(\bar{t} \mid L) = 1 - \Pr(t \mid L).$$

The probability of a term to occur in an abstract given that the abstract is not associated with location L, and the probability of a term to not occur in an abstract given that the abstract is not associated with location L are defined as:

$$\Pr(t \mid \bar{L}) = \Pr(t \in d \mid d \notin D_L) \text{ and}$$

$$\Pr(\bar{t} \mid \bar{L}) = 1 - \Pr(t \mid \bar{L}),$$

respectively. The maximum likelihood estimate for $Pr(t \mid \bar{L})$ is:

$$\Pr(t \mid \bar{L}) \approx \frac{\#\,\text{of abstracts}\, d \in D_{\bar{L}} \quad s.t.\ \ t \in d}{|D_{\bar{L}}|}.$$

The definition of conditional probability can be used to calculate the joint probability of two events as:

$$Pr(A,B) = Pr(B) \cdot Pr(A \mid B),$$

where A and B are the two events. This formula, along with the prior and conditional probabilities defined above, allows for the calculation of the four joint probabilities

$\Pr(t, L)$, $\Pr(\bar{t}, L)$, $\Pr(t, \bar{L})$, $\Pr(\bar{t}, \bar{L})$. The prior, conditional, and joint probabilities are all that is required to calculate a score for a term with the IG, MI, χ^2, and OR scoring functions, as defined in Table 5.2.1.

The four functions calculate a score for each term with respect to each location. However, terms are selected based on *all* locations. Therefore, to score a term over all locations, we combine the term's scores from each location. We employ two different functions to calculate the single overall score: *SUM* and *MAX*. As their names imply, *SUM* takes the sum of a term's scores over all locations, while *MAX* takes the maximum score for a term with respect to all locations. Following previous evaluations [44, 55], we chose the *SUM* function to calculate the OR and IG scores, and the *MAX* function to calculate MI and χ^2 scores.

The Entropy Scoring Method

The Entropy method, developed by Nair and Rost [32], is also based on Shannon's Information [46]. For each term *t*, its Shannon Information (SI) is calculated as:

$$SI = -\sum_{i=1}^{n} A_i \, log \, A_i \ ,$$

where *n* is the number of different locations, and A_i is the probability of finding the term *t* in the abstracts associated with the *i*th location. This probability is estimated as the ratio of the number of proteins in the *i*th location whose associated abstracts contain the term to the total number of proteins whose associated abstracts contain the term. This estimate is calculated as:

$$A_i \approx \frac{\#\ of\ \ proteins\ \ p \in P_i\ \ \ s.t.\ \ t \in D_P}{\#\ of\ \ proteins\ \ p\ \ \ s.t.\ \ t \in D_P} \quad ,$$

where P_i is the set of proteins associated with the i^{th} location and D_p is the set of abstracts associated with protein p.

A normalized SI is also calculated for each term, defined as:

$$normSI = -\sum_{i=1}^{M} Z_i \log Z_i \quad , \text{ where } Z_i = X_i \Big/ \sum_{i=1}^{M} X_i \quad \text{and}$$

$$X_i = \frac{\#\ of\ \ proteins\ \ p \in P_i\ \ s.t.\ \ t \in D_p}{\#\ of\ \ proteins\ \ p \in P_i} \quad .$$

where M is the number of locations that have associated abstracts containing the term.

Finally, the percent of fractional change in both the Shannon Information (*fracSI*) and the normalized Shannon Information (*fracNormSI*) is calculated as:

$$fracSI = 100 \cdot \frac{max\,SI - SI}{max\,SI}$$

$$fracNormSI = 100 \cdot \frac{max\,NormSI - normSI}{max\,NormSI},$$

where *maxSI* is the maximum possible Shannon Information and *maxNormSI* is the maximum normalized Shannon Information. The *maxSI* is calculated as the log of the total number of locations, *n*, and *maxNormSI* is calculated as the log of *M*. A term is selected for the feature set if both its *fracSI* and *fracNormSI* exceed a predetermined threshold.

5.2.2 Comparison Procedure

All six feature selection methods were tested using the text processing and cross-validation scheme described in Sections 3.1.1 and 3.2.3, respectively. Each method was applied to the same random partitioning of proteins, in order for the results to be comparable. The MultiLoc dataset, described in Section 4.1.2, served for training and testing. Feature selection methods are compared on the accuracy of the classifier trained using vectors based on their respective feature sets, and across a number of different feature set sizes. To allow this comparison, each method scores the set of potential terms, and for several set-sizes *N*, the *N* terms with the highest scores are selected, as well as any term that has a score equal to the N^{th} term. The results of this comparison are presented next.

5.2.3 Results of the Feature Selection Comparison

The goal of the comparison described in this chapter is to verify that a classifier based on the *Z-Test* feature selection method produces results similar to, if not better than, those produced by classifiers based on the other methods.

Figure 5.2.1 shows the accuracy of each of the classifiers as a function of the *average* number of features used to represent the proteins; 5.2.1 a) displays the results of classifiers tested on plant proteins, while 5.2.1 b) shows the results of the classifiers tested on animal proteins.

We note that in the plot the X-axis denotes the *average* number of features, rather than simply the number of features. This is due to the setting of the cross-validation process: A set of features is selected for each of the five training sets. The number of features selected may be different for each training set, because not only the top *N* scoring terms are selected, but also any additional terms with the same score as the N^{th} term (the lowest scoring term among the top *N*). The number of terms with the same score as the N^{th} term may vary between training sets. Therefore, we average the number of terms over all five training sets, and plot accuracy against this average.

Figure 5.2.1 demonstrates that the performance of the *Z-Test*, IG, and χ^2 methods is almost equivalent; we could probably use any of these methods for our classifier and achieve similar results. We use the *Z-Test* method in the experiments described in this

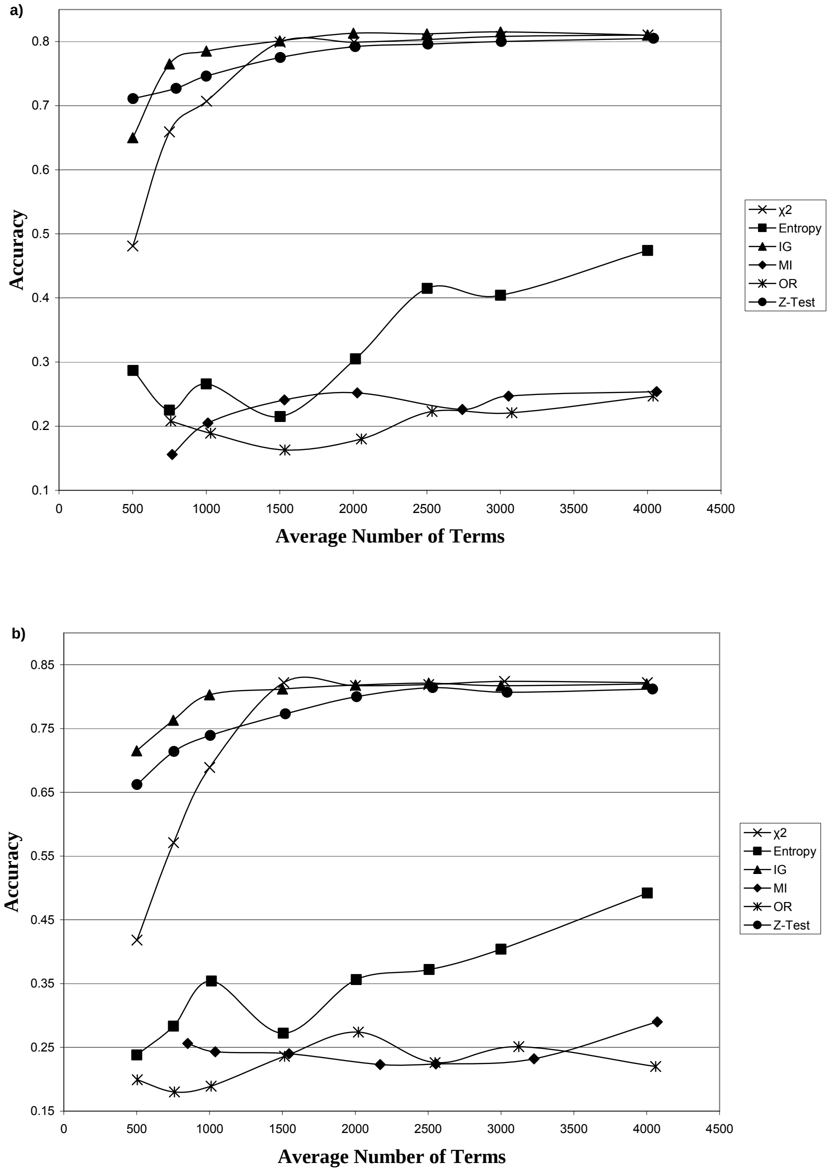


Figure 5.2.1: The accuracy of each classifier as a function of the average number of terms selected to represent the proteins. Figure a) displays the accuracy of a classifier trained with plant proteins. Figure b) shows the accuracy of a classifier trained with animal proteins.

thesis as *Z-Test* was our original approach and as it has a simple statistical interpretation. We cannot say conclusively which of these top-performing methods is best, as the differences between the three are very small, and the results were obtained from only one split of the MultiLoc [20] dataset (only one split was used because the amount of time required to perform a complete cross-validation experiment for every combination of feature number and selection method is very large). However, the results conclusively show that any of the three top-performing methods (*Z-Test*, χ^2 and IG) perform much better than the other three methods when used in conjunction with our classification scheme.

The poor performance of the classifiers based on mutual information is not surprising. Previous research has indicated that classification schemes that use mutual information for feature selection do not perform as well as those that use Chi-squared or information gain [55, 44]. As for the Entropy method, the poor performance may be attributed to the fact that it was developed to select features from a relatively small set of potential features compared to the set used here. Nair and Rost used the functional keywords in the Swiss-Prot entries of the proteins as potential features [32], whereas we use a much larger number of potential features.

Conversely, we did not expect to observe such poor results from the classifier that used odds-ratio for feature selection. According to previous publications, odds-ratio outperforms both Chi-squared and information gain for text categorization tasks [44]. The poor performance appears to stem from the formulation of the odds ratio function; it selects primarily terms associated with a single location, resulting in sparse term-vectors.

As was first presented in Section 5.2.1, the odds ratio function [44] is defined as:

$$\frac{\Pr(t \mid L) \cdot \left(1 - \Pr\left(t \mid \overline{L}\right)\right)}{\left(1 - \Pr(t \mid L)\right) \cdot \Pr\left(t \mid \overline{L}\right)} .$$

The probability $\Pr\left(t \mid \overline{L}\right)$ in the denominator is the probability of term *t* to occur in the abstracts associated with any location other than location *L*. If term *t* only occurs in abstracts associated with location *L*, then $\Pr\left(t \mid \overline{L}\right)$ is zero, and is thus replaced by the value 1 x 10 $^{-9}$. This leads to a very small denominator as compared to the numerator; $\left(1 - \Pr\left(t \mid \overline{L}\right)\right)$ is close to 1, since $\Pr\left(t \mid \overline{L}\right)$ is 1 x 10 $^{-9}$; the value $\Pr(t \mid L)$ is much larger than 1 x 10 $^{-9}$ because the term *t* has to have occurred in the abstracts associated with location *L* in this situation. Division by the small denominator results in a very high score for the term being evaluated. As a result, terms that occur only in abstracts associated with a single location are the first to be included in the set of distinguishing terms. Including primarily such terms causes feature vectors to be sparse; the only terms available to represent a given protein are those associated with its own location. If those terms are present in only a few abstracts, there may be insufficient terms associated with a location to represent many of its proteins. Figure 5.2.2, which plots the number of proteins represented against the number of distinguishing terms, illustrates this point, as the classifiers using odds ratio are able to represent fewer proteins as the number of features selected decreases. Indeed, for any size of feature set, the classifiers using odds ratio are unable to represent many of the proteins in the dataset. As the classifier is unable to represent a large portion of the dataset, it is unable to adequately “learn” a classification model, resulting in the poor performance of the classifier.

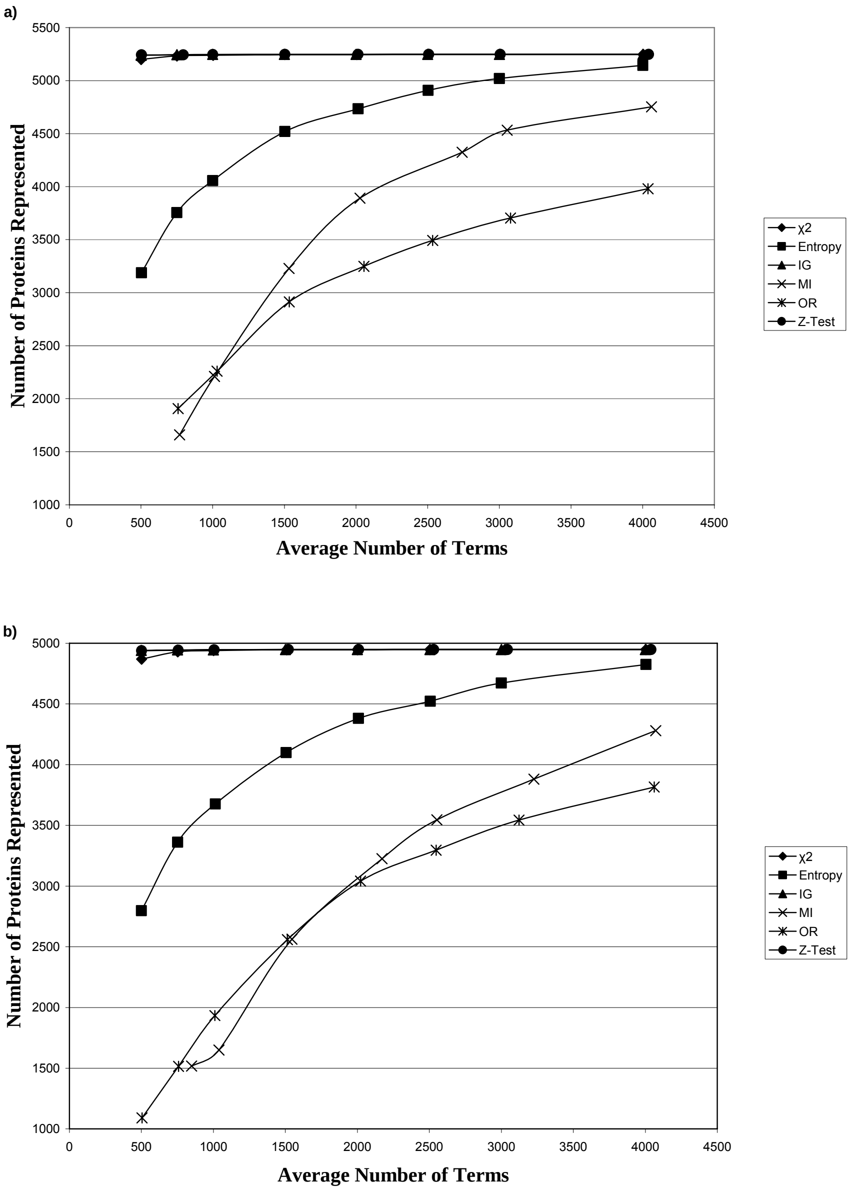


Figure 5.2.2: The number of proteins represented as a function of the average number of terms used to represent them, for each of the six feature selection methods. Figure a) displays the plot for plant proteins, b) for animal proteins.

The three feature selection methods that performed best with respect to accuracy are also the top performers with respect to coverage. Chi-squared, information gain, and *Z-Test* consistently represent the majority of the proteins, regardless of the number of terms selected. Even when using as few as about 500 terms, the three methods can represent almost all of the proteins in the dataset. Taken together, Figures 5.2.1 and 5.2.2 indicate that our *Z-Test* approach is an effective method for selecting features to represent proteins, and that we can confidently use it as a component in our system for predicting protein subcellular location.

5.2.4 Setting the Z-score Threshold for Different Datasets

As first mentioned in Section 5.1, a term *t* is considered distinguishing for a location *L* if for any other location *L'*, the absolute value of the Z-score, $\left|Z^{t}_{L,L'}\right|$ is greater than a set threshold. Based on the results shown in Figure 5.2.1, we decided to set a threshold that retains about 2,000 terms, as this number attains a balance between a computationally effective feature-space, and the accuracy of the classifier. As Figure 5.2.1 shows, the accuracy of the top methods does not significantly improve by including more than 2,000 features. In order to ensure that about 2,000 terms are indeed selected, we set a specific threshold for each dataset.

In order to select about 2,000 terms, for datasets that include locations that have only a small number of associated proteins we must set a lower threshold than for those datasets that do not include such locations. Locations that have a small number of associated proteins typically have only a few associated abstracts. The inclusion of a location *L*, with only a few associated abstracts, D_L, reduces the likelihood that, for a term *t*, $\left|Z^{t}_{L,L'}\right|$ will be above the threshold for *every* other location *L'*. Recall the formula for the Z-score:

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

All other values in the equation being equal, the inclusion of the location L will lead to a larger denominator in the Z-score expression, than if the same location has more associated abstracts. The larger denominator, which results from an increase in the value of $\left(\frac{1}{|D_L|}+\frac{1}{|D_{L'}|}\right)$, results in a lower $\left|Z^{t}_{L,L'}\right|$ than if location L has many associated abstracts.

For each dataset, we have two criteria for setting the threshold. First, the threshold must be sufficiently low to allow the selection of about 2,000 terms. Second, the threshold must be high enough to indicate that the probability of a term to be associated with one location is, in fact, statistically significantly different from its probability to be associated with all other locations. We describe next the process used to select a precise threshold.

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%]

Table 5.2.2: The threshold chosen for each organism and dataset.

To set the threshold for a dataset, we use a simple search process. We first partition the dataset as we do for 5-fold cross-validation. The *Z-Test* method then selects terms from each of the five resulting training sets, each consisting of 80% of the data, using thresholds of 2.576, 1.960, 1.645, 1.282, and 1.150. These thresholds correspond to confidence levels of 99%, 95%, 90%, 80%, and 75%, respectively, that the difference between the two probabilities, $\Pr(t \mid L)$ and $\Pr(t \mid L')$, is not arbitrary. Finally, for each threshold, we average the number of terms selected over all five training sets, and select the threshold whose average number of terms is closest to 2,000. Table 5.2.2 presents the threshold chosen for each organism in the datasets described in Chapter 4.

The thresholds shown in Table 5.2.2 are used for the standalone EpiLoc classifier. However, a threshold of 1.960 was set for the text-based portion of the SherLoc classifier. As stated in Section 3.3, there are differences between the feature selection method used for EpiLoc, and the feature selection method used for the text-based classifier integrated in SherLoc. These differences are discussed next.

5.3 Feature Selection for SherLoc

The SherLoc system was tested early in the development of EpiLoc, and as such, an earlier version of EpiLoc, which we call EarlyText, was incorporated in the SherLoc classifier. Since then, changes have been made to the feature selection process intended to improve EpiLoc's performance. As a result of these changes, EpiLoc and SherLoc differ with respect to the following three aspects of feature selection:

1) *The abstracts included for feature selection*: Feature selection for EpiLoc includes only those abstracts associated with proteins from locations found in the organism for which EpiLoc is being trained. For example, if EpiLoc is trained on plant proteins, lysosomal proteins are not included in the feature selection process. In contrast, EarlyText included in the feature selection process abstracts associated with proteins from *all* locations, regardless of the organism for which it was being trained.

2) *The level of the threshold set*: EpiLoc adjusts its threshold according to the dataset being considered. Conversely, EarlyText set a single Z-score threshold, 1.960, for all datasets.

3) *The point at which feature selection takes place relative to cross-validation*: For EpiLoc, feature selection is performed during the cross-validation process, to allow features to be chosen from each of the five training sets. Therefore, a different feature set, which is based on only 80% of the data, is used to represent each of the five training-test set pairs. For the text-based portion of SherLoc, feature selection took place before the protein sets were partitioned into five subsets. Thus, a single set of terms, selected from 100% of the data, was used as a basis for the feature vectors representing proteins for all the training and test sets.

The use of a single set of features selected from the entire dataset to represent all the proteins in EarlyText was revisited for the following reason. When using all the abstracts associated with the proteins to select a single feature list, the feature selection process involves both training and test proteins. While feature selection is often performed as a separate step from classification, and precedes the cross-validation process, we wanted to ensure that using the whole dataset for feature selection does not strongly affect the evaluation. To do so we conducted the following experiment.

We compared the performance of two classifiers on the MultiLoc dataset [20], one of which selects features from the entire dataset while the other selects features from only 80% of the dataset. For our experiment we used the cross-validation process described in Sections 3.2, and employ the exact same split of the data for each classifier.

The design of each classifier was intended to duplicate that of EarlyText; a threshold of 1.960 was applied, and all abstracts associated with all locations of the dataset were considered for term selection. However, the two classifiers differed with respect to the stage in which feature selection was performed. For one of the classifiers (denoted *Class100*), feature selection was performed on the whole dataset, before it had been partitioned for cross-validation. For the other classifier (denoted *Class80*), feature selection was performed after the dataset had been partitioned; therefore, features were

Loc		**Class80**			**Class100**	
			Plant (*Sens Spec MCC*)			
va	0.73	0.25	0.42	**0.83**	**0.31**	**0.50**
pe	0.88	0.77	0.82	0.88	0.77	0.82
go	0.81	0.46	0.60	**0.86**	**0.53**	**0.66**
er	0.58	0.60	0.57	0.57	**0.64**	**0.59**
ch	0.84	0.74	0.77	**0.88**	**0.76**	**0.80**
mi	0.81	0.80	0.79	**0.82**	0.81	**0.80**
nu	0.79	0.74	0.73	**0.83**	**0.78**	**0.78**
ex	0.61	0.74	0.61	**0.74**	**0.77**	**0.71**
pm	0.82	0.85	0.79	**0.84**	**0.88**	**0.82**
cy	0.60	0.66	0.51	**0.61**	**0.73**	**0.57**
Acc		0.72			**0.76**	
Avg		0.75			**0.79**	
			Animal Proteins (*Sens Spec MCC*)			
ly	0.80	0.31	0.48	**0.83**	**0.36**	**0.54**
pe	0.89	0.78	0.82	0.89	0.77	0.82
go	0.83	0.47	0.61	**0.87**	**0.54**	**0.67**
er	0.60	0.58	0.57	0.59	**0.62**	**0.59**
mi	0.81	0.82	0.80	**0.82**	0.82	**0.80**
nu	0.81	0.74	0.74	**0.84**	**0.78**	**0.78**
ex	0.61	0.74	0.61	**0.73**	**0.77**	**0.70**
pm	0.81	0.85	0.78	**0.83**	**0.88**	**0.81**
cy	0.61	0.69	0.54	**0.64**	**0.74**	**0.59**
Acc		0.72			**0.76**	
Avg		0.75			**0.78**	

Table 5.3.1: The results of the classifiers based on different feature selection methods, for plant and animal proteins. The mean values, over 200 measurements, of *Sens*, *Spec*, *MCC*, *Avg*, *Acc* are displayed. Values in bold indicate a statistically significant difference (p<0.01).

only selected from 80% of dataset, that is, from each of the four subsets comprising a training set.

We compare the two classifiers over 200 random splits of the data, for both plant and animal proteins. The 200 splits were performed so that a statistical test could be applied to determine if the differences in performance were statistically significant. For the two classifiers, we compared the distribution of their 200 results on the following measures: accuracy, average sensitivity, sensitivity, specificity, and *MCC*. The mean value of these measurements over the 200 iterations is shown in Table 5.3.1.

We note that the results displayed in Table 5.3.1 were produced by the SVMs that we trained, and not by the SVMs used in the EarlyText classifier (see Section 3.3). The parameter settings for the Class100 classifier are thus different from those used for the EarlyText classifier, and as such the results can be used to evaluate the difference between Class80 and Class100 - but do not reproduce (nor can they be compared to) the results of EarlyText as shown in Chapter 6.

We performed a Kolmogorov-Smirnov test [11], which is a statistical test used to determine if two datasets differ significantly, on the sensitivity, specificity, *MCC*, average sensitivity, and average accuracy distributions to determine if the differences between the results of the two classifiers are statistically significant. For almost all measurements, the test indicates the difference to be highly statistically significant ($p<0.01$). As is shown in Table 5.3.1, with the exception of peroxisomal (*Sens*, *Spec*, *MCC*), endoplasmic (*Sens*), and mitochondrial (*Spec*) proteins, the mean value of each distribution is higher for Class100 than for Class80. Therefore, it appears that the inclusion of the early version of the text-based system in the SherLoc classifier may somewhat affect its results. However, the difference in performance between Class80 and Class100 is less than or equal to 0.06 for almost all measurements, including accuracy and average sensitivity. Only the extracellular (*Sens*, *MCC*) and the vacuolar (*Sens*, *MCC*) proteins have associated measurements that differ, on average, by more than 0.06 between the two classifiers.

Loc	**Class100**	**Class80**	**% Change**
va	204.0	150.0	-26.5
ly	104.0	69.2	-33.5
pe	82.0	55.7	-32.1
go	86.0	57.5	-33.1
er	47.0	34.9	-25.7
ch	84.0	57.3	-31.8
mi	47.0	30.6	-34.9
nu	79.0	50.1	-36.6
ex	31.0	18.0	-41.9
pm	52.0	31.2	-40.0
cy	8.0	6.1	-23.8

Table 5.3.2: The number of terms per location for Class100 and Class80, and the percent change from Class100 to Class80. For Class80, the number of terms per location is averaged over the 200 random splits of the data.

For extracellular proteins, the reason for the large decrease in performance from Class100 to Class80 appears to be caused by the number of features that are selected based on the location. Table 5.3.2 shows the number of terms selected from the abstracts associated with each location by Class100 and Class80, where, for Class80, the number of terms is averaged over the 200 different splits (there is no need to average for Class100; the same number of terms are selected for each location for each split, as the terms are selected from 100% of the data). As is shown in Table 5.3.2, there is a 41.9% decrease in the number of terms selected from abstracts associated with extracellular proteins between Class100 and Class80, resulting in only 18 distinguishing terms for the extracellular space. Although the cytoplasm also has very few features associated with it, the relative change in the number of terms associated with the cytoplasm between Class80 and Class100 is much smaller; the small number of distinguishing terms associated with the cytoplasm appears to manifest itself in the relatively poor performance for the location in both the CLASS80 and CLASS100 experiments. Furthermore, while the plasma membrane also shows a large percentage decrease in the number of distinguishing terms between Class100 and Class80, there are still many more distinguishing terms for the plasma membrane (31.2), than there are for the extracellular space (18.0). The fact that the plasma membrane retains, relative to the extracellular space, a larger number of associated distinguishing terms, explains the relatively small decrease in performance on the plasma membrane proteins between Class100 to Class80.

The poor performance of Class80, relative to Class100, on the vacuolar proteins is likely due to a dearth of data associated with the location. The vacuole, after removal of textless proteins, has only 49 associated proteins in the MultiLoc dataset, half the number of the next smallest location, the lysosome. When performing feature selection for Class80, terms are selected from the abstracts associated with only 36 vacuolar proteins (80% of 45; 4 of the 49 proteins are removed so that each subset of the dataset receives an equal number of proteins). The number of abstracts associated with the vacuole is therefore very small, and the distinguishing terms found in these abstracts do not appear to effectively characterize the vacuole. Therefore the benefit of including 100% of the data is more apparent for the vacuole than for other locations, as Class100 is able to draw from a larger sample than Class80 when characterizing the vacuole.

The measurements in Table 5.3.1 suggest that although EarlyText, and therefore SherLoc, may have benefited from the selection of text features from 100% of the data, the resulting improvement is quite small compared with the overall improvement SherLoc achieved with respect to all earlier systems (see Section 6.1). Moreover, as we show in Chapter 6, EpiLoc – with its cross-validated feature selection process – outperforms EarlyText for *all* locations in the MultiLoc dataset. These results further suggest that the benefits from using the whole dataset for feature selection, as done in EarlyText, are marginal. In the next chapter, we present the performance of EarlyText, EpiLoc and SherLoc on several datasets, along with the results of the rest of the experiments described in Chapter 4.

Chapter 6

Results

In this chapter, we present the results of the experiments described in Chapter 4. The results of comparing EpiLoc and SherLoc to other state-of-the-art prediction systems, and of running SherLoc and EpiLoc on the *Diff48* dataset, are examined. We then look at the performance of HomoLoc and PubLoc, to determine if either is suitable for handling textless proteins for the EpiLoc system. Last, we demonstrate the use of DiaLoc.

6.1 Systems Comparison Results

This section focuses on the results of running EpiLoc and SherLoc [19, 47] on the three datasets described in Section 4.1: TargetP [15], PLOC [38], and MultiLoc [20]. For comparison, we present the results of the original TargetP, PLOC, and MultiLoc systems on their respective datasets, as reported in their corresponding publications. We also present the performance of the MultiLoc classifier on both the TargetP and PLOC datasets, in order to examine the effect of integrating MultiLoc with the early version of our text-based system, EarlyText. As described in Section 5.3, one of the ways in which EarlyText differs from EpiLoc is that EarlyText uses the entire dataset for feature selection. To further examine (beyond the examination done in Section 5.3) the possible effect of this difference on the evaluation of SherLoc, we compare EarlyText's results to those of EpiLoc for each of the three datasets. The results of both EarlyText and SherLoc, as reported in the following sections, have appeared in previous publications [19, 47]. For each system, on each dataset, we report the sensitivity (*Sens*), specificity (*Spec*), and Matthew's Correlation coefficient (*MCC*) with respect to each location, as well as the overall accuracy (*Acc*) and average sensitivity (*Avg*), all of which were defined in Section 3.2.2.

6.1.1 Results for the TargetP Dataset

We present the results of applying EpiLoc and SherLoc to the TargetP dataset, for both plant and non-plant proteins, in Table 6.1.1. The table also includes the results of TargetP, EarlyText, and MultiLoc on the dataset. The values in Table 6.1.1 suggest that EpiLoc, as a standalone classifier, performs at a level similar to TargetP, the state-of-the-art classifier based solely on N-terminal sequence data. Furthermore, the results indicate that among those systems shown in Table 6.1.1, SherLoc is the best performing system on the dataset.

For both plant and non-plant proteins, EpiLoc's accuracy and average sensitivity are slightly higher than those of TargetP. Moreover, for most location-specific performance metrics, EpiLoc and TargetP perform about the same. The only metrics for which TargetP significantly outperforms EpiLoc are specificity for chloroplast and for mitochondrial plant proteins.

EpiLoc's low specificity on chloroplast proteins appears to be caused by the small amount of data associated with the location. The 123 chloroplast proteins in the TargetP dataset classified by EpiLoc have a total of 181 associated PubMed abstracts. Conversely, the location with the most proteins, *Other*, on which EpiLoc performs very well, has 1,636 proteins with 3,782 associated abstracts. A common problem in machine learning is that it is difficult to "learn" a model from little data. The feature distribution that characterizes the underrepresented class in the training data is often different from the characteristic feature distribution for the same class in the test data [40].

TargetP Dataset					
Loc	**TargetP**	**EpiLoc**	**MultiLoc**	**EarlyText**	**SherLoc**
Plant (*Sens Spec MCC)*					
ch	0.85 0.69 0.72	0.92 0.53 0.68	0.88 0.76 0.78	0.78 0.74 0.72	**0.93 0.89 0.89**
mi	0.82 0.90 0.77	0.89 0.81 0.82	0.87 0.94 0.84	0.90 0.98 0.90	**0.95 0.99 0.95**
SP	0.91 0.95 0.90	0.89 0.84 0.80	0.93 0.97 0.93	0.77 0.84 0.74	**0.95 0.98 0.95**
OT	0.85 0.78 0.77	0.84 **0.95** 0.78	0.92 0.84 0.86	0.67 0.52 0.50	**0.95** 0.87 **0.89**
Acc	0.853 (±0.035)	0.862 (±0.004)	0.897 (±0.016)	0.812 (±0.026)	**0.947**(±0.015)
Avg	0.856 (n/a)	0.883 (±0.001)	0.902 (±0.02)	0.781 (±0.032)	**0.944** (±0.016)
Non-Plant (*Sens Spec MCC)*					
mi	0.89 0.67 0.73	0.92 0.84 0.86	0.91 0.77 0.81	0.91 0.78 0.81	**0.97 0.88 0.91**
SP	0.96 0.92 0.92	0.93 0.86 0.84	0.95 0.92 0.91	0.92 0.83 0.82	**0.98 0.96 0.96**
OT	0.88 0.97 0.82	0.88 0.95 0.81	0.91 0.97 0.86	0.87 0.95 0.79	**0.95 0.99 0.93**
Acc	0.900 (±0.007)	0.901 (±0.006)	0.925 (±0.012)	0.887 (±0.011)	**0.962** (±0.008)
Avg	0.907 (n/a)	0.908 (±0.003)	0.928 (±0.011)	0.898 (±0.016)	**0.967** (±0.009)

Table 6.1.1: Prediction performance of TargetP, EpiLoc, MultiLoc, EarlyText, and SherLoc on the TargetP dataset. Both location-specific (*Sens*, *Spec*, *MCC*) and overall results (*Acc* and *Avg*) are shown. Highest values appear in bold. Standard deviations (denoted ±) are provided for *Acc* and *Avg* where available.

Specifically, the small amount of data affects EpiLoc's performance on chloroplast proteins through the feature selection process. The benefit of including more data in the feature selection process is evident in EarlyText's superior specificity on proteins from the chloroplast. As discussed in Chapter 5, EarlyText differs from EpiLoc in that it selects features from the whole dataset, as opposed to just 80% as for EpiLoc.

In some cases EarlyText benefits from the inclusion of the additional data, as it is able to select more terms that better characterize proteins in the test set.

EarlyText also significantly outperforms EpiLoc with respect to specificity on mitochondrial plant proteins. This improved performance appears to be a result of EpiLoc's altered feature selection process (the setting of specific thresholds for each dataset, the inclusion of locations only from the organism being considered, and the use of 80% of the dataset) relative to EarlyText. As Table 6.1.1 shows, EarlyText's specificity on mitochondrial proteins is much higher for plant than for non-plant proteins; the inclusion of the chloroplast *improves* the classifier's specificity for the mitochondria. The dramatic drop in EarlyText's performance from plant to non-plant mitochondrial proteins suggests that many of the non-plant proteins that are misclassified as mitochondrial by EarlyText are instead misclassified as *Other* for plant proteins (as is reflected in EarlyText's very low specificity on *Other* proteins in Plant). Note that, with respect to all location-specific scores on *Other* proteins, the inclusion of the chloroplast proteins causes EarlyText's performance to decline drastically from non-plant to plant proteins. On the other hand, EpiLoc's performance on mitochondrial proteins and on *Other* proteins is similar for both plant and non-plant proteins. As the main difference between EpiLoc and EarlyText is the feature selection process, this is most likely the cause for the difference in performance on both mitochondrial and *Other* plant proteins between the two systems. Further research is required to determine exactly which difference in the feature selection process is causing EarlyText's improved specificity on mitochondrial plant proteins and its worsened performance on *Other* proteins.

Although EarlyText appears to benefit in a few cases from using extra data, overall, EpiLoc clearly outperforms EarlyText on the TargetP dataset. Two of the changes made to EpiLoc's feature selection process – the setting of specific thresholds for each dataset and the inclusion of locations only from the organism being considered – compensate for any advantage EarlyText may have over EpiLoc. EpiLoc's sensitivity, specificity, and *MCC* are at least equal to, and in most cases better than, those of EarlyText (excluding specificity for mitochondrial proteins and for chloroplast proteins). Moreover, EpiLoc's sensitivity is much higher than EarlyText's for chloroplast proteins, and the two systems' sensitivity is nearly the same – a 0.01 difference – for mitochondrial proteins. Finally, EpiLoc's accuracy and average sensitivity is slightly better than EarlyText's for non-plant proteins, and much better for plant proteins.

While EpiLoc's results are good, SherLoc's performance on the TargetP dataset is even better. SherLoc's accuracy and average sensitivity significantly exceed those produced by any other system to which it is compared for both plant and non-plant proteins. Moreover, with the exception of specificity for *Other* proteins, SherLoc's sensitivity, specificity, and *MCC* are higher than those of the other systems to which it is compared. SherLoc is clearly the best performing system on the TargetP dataset. SherLoc's results also demonstrate the usefulness of integrating text with sequence data for protein location prediction. The improved performance of the integrated system over both of its components, MultiLoc and EarlyText, demonstrates that text can be used to improve the prediction capability of a sequence-based classifier.

6.1.2 Results for the PLOC Dataset

The results of EpiLoc and of SherLoc on the PLOC dataset, along with those of MultiLoc, EarlyText and PLOC, are shown in Table 6.1.2[7]. EpiLoc's performance is comparable to that of the PLOC classifier, while SherLoc outperforms all other systems to which it is compared.

PLOC Dataset													
Loc	**PLOC**	**EpiLoc**			**MultiLoc**			**EarlyText**			**SherLoc**		
	Sens	**Plant (*Sens Spec MCC*)**											
go	0.15	0.76	0.38	0.53	0.55	0.08	0.20	**0.88**	**0.41**	**0.60**	0.81	0.34	0.52
cs	0.59	0.83	0.31	0.50	0.80	0.21	0.40	**0.86**	0.25	0.46	0.85	**0.34**	**0.53**
va	0.25	0.67	0.17	0.32	0.65	0.19	0.34	0.70	0.11	0.26	**0.83**	**0.28**	**0.48**
er	0.47	0.68	0.32	0.45	0.78	0.69	0.73	0.73	0.26	0.42	**0.84**	**0.73**	**0.78**
pe	0.25	0.82	0.58	0.68	0.72	0.29	0.44	0.75	0.50	0.61	**0.83**	**0.62**	**0.71**
ch	0.72	**0.88**	0.80	**0.82**	0.66	0.72	0.66	0.86	0.77	0.80	0.84	**0.83**	**0.82**
mi	0.57	0.76	**0.86**	0.79	0.67	0.65	0.62	0.76	**0.86**	0.79	**0.85**	0.84	**0.83**
ex	0.78	0.69	0.66	0.63	0.84	0.87	0.83	0.66	0.57	0.55	**0.87**	**0.92**	**0.88**
cy	0.72	0.55	0.61	0.50	0.60	0.68	0.57	0.40	0.54	0.37	**0.78**	**0.75**	**0.72**
pm	**0.92**	0.78	0.85	0.78	0.83	0.96	0.87	0.71	0.83	0.71	0.89	**0.98**	**0.92**
nu	**0.90**	0.80	0.91	0.80	0.75	0.88	0.75	0.77	0.90	0.77	0.88	**0.94**	**0.88**
Acc	0.782 (± 0.009)	0.743 (±0.005)			0.736 (±0.007)			0.687 (±0.007)			**0.851** (± 0.011)		
Avg	0.579 (± 0.021)	0.748 (±0.013)			0.713 (±0.028)			0.735 (±0.018)			**0.855** (± 0.012)		
		Animal (*Sens Spec MCC*)											
go	0.15	0.76	**0.51**	0.62	0.51	0.07	0.17	**0.88**	0.46	**0.64**	0.83	0.31	0.51
cs	0.59	0.84	**0.32**	**0.51**	0.75	0.25	0.43	**0.86**	0.27	0.48	0.80	0.22	0.41
ly	0.62	**0.89**	0.32	0.53	0.77	0.36	0.35	0.81	0.33	0.50	0.81	**0.52**	**0.64**
er	0.47	0.72	0.30	0.45	0.81	0.63	0.71	0.75	0.27	0.43	**0.88**	**0.69**	**0.78**
pe	0.25	**0.85**	0.55	0.68	0.74	0.31	0.46	0.79	0.50	0.62	0.81	**0.64**	**0.71**
mi	0.57	0.79	0.85	0.80	0.69	0.73	0.68	0.75	**0.86**	0.78	**0.86**	0.85	**0.83**
ex	0.78	0.74	0.68	0.66	0.83	0.87	0.83	0.70	0.58	0.57	**0.91**	**0.91**	**0.90**
cy	0.72	0.53	0.63	0.50	0.65	0.75	0.64	0.49	0.58	0.44	**0.80**	**0.79**	**0.75**
pm	**0.92**	0.79	0.85	0.78	0.81	0.97	0.86	0.73	0.83	0.72	0.89	**0.98**	**0.91**
nu	**0.90**	0.81	0.90	0.80	0.78	0.87	0.76	0.78	0.90	0.78	0.87	**0.95**	**0.88**
Acc	0.796 (± 0.009)	0.743 (±0.002)			0.760 (±0.007)			0.702 (±0.007)			**0.864** (± 0.008)		
Avg	0.579 (± 0.021)	0.773 (±0.0012)			0.736 (±0.039)			0.755 (±0.027)			**0.845** (± 0.036)		

Table 6.1.2: Prediction performance of PLOC, EpiLoc, MultiLoc, EarlyText, and SherLoc on the PLOC dataset. Overall results (*Acc* and *Avg*) are shown for all systems. For location-specific results, only sensitivity is presented for PLOC (as was done in its corresponding publication), while for all other systems sensitivity, specificity, and *MCC* are each displayed. As presented in its corresponding publication, PLOC's location-specific results are averaged over all three organisms (animal, plant, fungal). Highest values appear in bold. Standard deviations (denoted ±) are provided for *Acc* and *Avg* where available.

[7] Results for fungal proteins are reported in Table B.2 in Appendix B.

The performance of EpiLoc and of PLOC on the PLOC dataset is quite different, yet still indicates that EpiLoc is an effective method for predicting protein subcellular location. EpiLoc's overall accuracy is slightly lower than that of the PLOC system for both animal and plant proteins. However, EpiLoc's average sensitivity is significantly higher than PLOC's for both organisms. Furthermore, EpiLoc's sensitivity is higher than PLOC's for all locations except for the four with the largest number of associated proteins (*ex*, *cy*, *pm*, and *nu*)[8]. Whereas PLOC primarily works well on over-represented locations for which a large number of proteins is 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 have at most 125 proteins). The one location, for which EpiLoc's performance, with respect to sensitivity, is notably worse than its performance on other locations, is the cytoplasm.

EpiLoc's performance on cytoplasmic proteins is poor according to all measures (*Sens*, *Spec*, and *MCC*). It is better than that of EarlyText, but worse than that of both sequence based classifiers (MultiLoc and PLOC). The relatively poor performance of the text-based classifiers appears to stem from a lack of terms that characterize the location. Unlike organelles that have very specific functions, the cytoplasm has a less defined role, as it acts as a medium in which reactions that serve a variety of functions take place. The non-specific role of the cytoplasm is reflected in the abstracts associated with it, as very few distinguishing terms are selected from them; over the five different splits of the PLOC dataset, on average only 13 and 21 terms are associated with the cytoplasm for plant and animal proteins, respectively. This number of terms is the smallest compared to

[8] The PLOC publication does not report specificity or *MCC*.

all other locations, and comprises only 1.0% and 1.5%, respectively, of the total number of terms selected for representing plant and animal proteins. Moreover, the abstracts associated with the cytoplasm are also the source of the fewest distinguishing terms for the MultiLoc dataset, which is discussed in Section 6.1.3. With so few terms to characterize the cytoplasmic proteins, both EpiLoc and EarlyText perform poorly on them.

The specificity of both EpiLoc and EarlyText is also low for those locations that have fewer than 115 associated proteins (*go*, *cs*, *ly*, and *er*,). The poor specificity of the two systems is likely caused by the small number of proteins associated with the locations; the machine learning methods are not given enough information to adequately "learn" how to classify the proteins.

Overall, EpiLoc clearly outperforms EarlyText on the PLOC dataset. EpiLoc's accuracy and average sensitivity are higher than EarlyText's, and EpiLoc typically scores higher according to most measures with respect to individual locations (47 out of a total of 63 calculated scores). These results further suggest that EarlyText's benefit from including all the data during feature selection (as discussed in Section 5.3) is minimal. The only location for which EarlyText significantly outperforms EpiLoc is the Golgi apparatus, which has the fewest associated proteins, and the large improvement is limited to sensitivity. Aside from the Golgi apparatus, any significant benefit EarlyText has over EpiLoc for classifying proteins is negated by the changes we have made to the feature selection process of EpiLoc.

Although EpiLoc's performance is very good, once again, it is not as good as that of SherLoc. SherLoc's accuracy and average sensitivity exceed those of all the systems to which it is compared. Moreover, for almost all locations, its sensitivity, specificity, and *MCC* are the highest. SherLoc's results on the PLOC dataset are also better than those of its individual components, MultiLoc and EarlyText. This improvement further demonstrates that integrating text with sequence data improves prediction performance.

6.1.3 Results for the MultiLoc Dataset

Table 6.1.3 displays the results of MultiLoc, EpiLoc, EarlyText, and SherLoc when applied to the MultiLoc dataset[9]. Both SherLoc and EpiLoc improve upon MultiLoc on its own dataset, and SherLoc, again, performs better than the other systems.

Overall, EpiLoc performs better than MultiLoc when applied to the MultiLoc dataset. For both plant and animal proteins, EpiLoc's accuracy and average sensitivity are higher than MultiLoc's, and EpiLoc's sensitivity, specificity, and *MCC* are also higher for the majority of the locations.

EpiLoc's accuracy and average sensitivity significantly exceed those of EarlyText. Moreover, EpiLoc's sensitivity, specificity, and *MCC* are higher than EarlyText's for all but 3 of the 57 location-specific measurements. For the MultiLoc dataset, EpiLoc's results show similarities to its results on the two previous datasets, TargetP and PLOC. EpiLoc's performance on locations with very few associated proteins (the chloroplast and the lysosome in this case) is not as good as its performance on those locations with many associated proteins. Moreover, EpiLoc's results for the cytoplasm are not as good as for other locations with a similar number of associated proteins (for example, the plasma membrane).

[9] Results for fungal proteins are reported in Table B.1 in Appendix B.

MultiLoc Dataset				
Loc	**MultiLoc**	**EpiLoc**	**EarlyText**	**SherLoc**
	Plant (*Sens Spec MCC)*			
va	0.70 0.20 0.36	0.73 **0.29** 0.45	0.59 0.15 0.29	**0.83 0.29 0.48**
pe	0.71 0.34 0.47	**0.88 0.74 0.80**	**0.88** 0.71 0.79	0.85 0.59 0.70
go	0.75 0.41 0.54	**0.85** 0.55 0.67	0.82 0.42 0.57	0.84 **0.61 0.70**
er	0.72 0.54 0.61	0.72 **0.64** 0.67	0.73 0.55 0.62	**0.82** 0.63 **0.71**
ch	0.88 0.85 0.85	0.89 0.75 0.81	0.89 0.70 0.78	**0.94 0.91 0.92**
mi	0.85 0.79 0.80	0.82 0.81 0.80	0.80 0.80 0.78	**0.90 0.88 0.88**
nu	0.82 0.75 0.75	0.82 0.81 0.79	0.80 0.72 0.72	**0.89 0.85 0.85**
ex	0.68 0.81 0.70	0.84 0.82 0.80	0.74 0.80 0.73	**0.84 0.90 0.84**
pm	0.74 0.89 0.77	**0.85** 0.91 0.85	0.80 0.91 0.82	0.84 **0.96 0.87**
cy	0.68 0.85 0.70	0.64 0.78 0.63	0.53 0.75 0.54	**0.81 0.91 0.82**
Acc	0.746 (± 0.008)	0.790 (± 0.002)	0.731 (±0.011)	**0.851** (± 0.011)
Avg	0.752 (± 0.009)	0.805 (± 0.005)	0.760 (±0.023)	**0.855** (± 0.012)
	Animal (*Sens Spec MCC)*			
ly	0.69 0.36 0.48	**0.86** 0.39 0.57	0.75 0.32 0.47	**0.86 0.55 0.68**
pe	0.71 0.31 0.44	0.90 **0.77 0.82**	**0.93** 0.60 0.74	0.89 0.68 0.77
go	0.71 0.43 0.53	**0.88** 0.62 0.73	0.86 0.40 0.57	0.87 **0.65 0.74**
er	0.68 0.56 0.60	0.74 0.59 0.65	0.74 0.48 0.58	**0.82 0.67 0.73**
mi	0.88 0.82 0.83	0.82 0.82 0.80	0.80 0.79 0.77	**0.93 0.91 0.91**
nu	0.82 0.73 0.73	0.84 0.81 0.80	0.84 0.71 0.73	**0.89 0.83 0.84**
ex	0.79 0.83 0.77	0.80 0.82 0.77	0.76 0.78 0.72	**0.86 0.90 0.86**
pm	0.73 0.90 0.76	**0.85** 0.90 0.84	0.80 0.91 0.81	**0.85 0.95 0.87**
cy	0.67 0.85 0.68	0.68 0.79 0.65	0.51 0.77 0.53	**0.83 0.91 0.82**
Acc	0.746 (± 0.01)	0.792 (±0.008)	0.725 (±0.007)	**0.862** (± 0.009)
Avg	0.741 (± 0.025)	0.818 (±0.005)	0.775 (±0.015)	**0.868** (± 0.015)

Table 6.1.3: Prediction performance of MultiLoc, EpiLoc, EarlyText, and SherLoc on the MultiLoc dataset. Both location-specific (*Sens*, *Spec*, *MCC*) and overall results (*Acc* and *Avg*) are shown. Highest values appear in bold. Standard deviations (denoted ±) are provided for *Acc* and *Avg* where available.

SherLoc's results are also similar to those it displayed on the two other datasets; its results are better than those of any other classifier to which it is compared. In terms of overall accuracy, average sensitivity, and the vast majority of location-specific measurements, SherLoc produces the highest values for the MultiLoc dataset. SherLoc's performance also provides quantitative evidence as to the benefit of integrating text with sequence data. An independent t-test [51] indicates that the improved performance values of SherLoc, as compared to EarlyText and MultiLoc, are highly statistically significant (*p* << 0.05) for almost all subcellular locations. There are only two exceptions. The Golgi apparatus, where there is no significant difference in sensitivity with respect to EarlyText for plant and animal proteins, and the peroxisome, where EarlyText outperforms SherLoc for plant and animal proteins.

6.1.4 Systems Comparison Conclusion

Overall, the above results demonstrate that EpiLoc, as a standalone classifier, performs at a level similar to, and in some cases better than, that of other state-of-the-art classifiers. 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 PLOC dataset were not as strict as those employed for the MultiLoc and TargetP datasets (see Section 4.1), the resulting protein distribution among locations, and thus the distribution of the associated text, is quite different between the datasets. As such, a lower Z-score threshold, as shown in Table 5.2.2, was required to select a sufficient number of features (only about 1,250 were actually chosen) for the PLOC dataset. These terms are fewer and less distinguishing, and using them to represent the PLOC proteins results in EpiLoc's relatively lower performance on this dataset.

Furthermore, we have demonstrated, through EpiLoc's improved performance over EarlyText, that although there is some benefit to using the entire dataset for feature selection, the benefit is marginal and is minimized by making the changes to the feature selection process as is already done in EpiLoc. The fact that EpiLoc consistently outperforms its predecessor, EarlyText, also suggests that by integrating EpiLoc into the SherLoc system, we may not only produce results as good as those reported here for

SherLoc, but may also be able to further improve on these results (the new integration will be examined in future work).

Even before integrating EpiLoc (rather than EarlyText) into the combined system, the results shown in Tables 6.1.1 – 6.1.3 clearly demonstrate that the integrated classifier, SherLoc, significantly outperforms earlier prediction systems. The results demonstrate a significant improvement in the prediction of subcellular location through the integration of sequence- and text-based classifiers. For each dataset, SherLoc's performance is much better than that of either of its components, MultiLoc and EarlyText.

Our results clearly show the complementary nature of EarlyText and MultiLoc. The sequence-based method, MultiLoc, performs well on those proteins that rely on an N-terminal sequence for location prediction, such as proteins from the chloroplast and mitochondria. The text-based system, EarlyText, complements MultiLoc with its superior performance on proteins whose sequence localization signals are not as clear. Such proteins include those located in the peroxisome, and proteins from the secretory pathway locations. SherLoc's performance is clearly good on the commonly used datasets. To examine its performance in experiments that do not involve cross-validation, we applied it to proteins that have only recently been assigned a subcellular location. These proteins were represented, and their locations predicted, using text that precedes the determination of their location.

6.2 *Diff48* Results

We ran the two systems, EpiLoc and SherLoc [47], on the new dataset *Diff48*. We also ran SherLoc on the new dataset, *Unknown*. As the locations of the approximately 19,000 proteins in the *Unknown* set have not yet been experimentally determined, the predicted locations are not reported here but can be found at: *http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc/sherloc_information***.** In the section that follows, we focus on the results of running SherLoc and EpiLoc on the *Diff48* dataset.

As was shown in Table 4.2.1, the *Diff48* dataset does not uniformly represent all subcellular locations. Several locations are not represented at all within the dataset (*go*, *pe*, *ly*, *pm*), and a few have only 1-3 proteins each (*ch*, *va*, *er*). The results from the locations with 1-3 proteins are not shown here, as the sample size is too small to merit analysis. Instead, we concentrate on locations with a minimum of 20 proteins. It is important to note that the results reported here are obtained on a very small dataset with a very different data distribution from that used in the 5-fold cross-validation study. Consequently, the results shown in Table 6.2.1 cannot be directly compared to those shown in Tables 6.1.1-6.1.3.

The results from running both SherLoc and EpiLoc on the *Diff48* set, presented in Table 6.2.1, are very promising. Overall, SherLoc and EpiLoc predict the location of the proteins in *Diff48* with an accuracy of approximately 0.71 and 0.66, respectively. For

	Diff48 Dataset			
Loc	**SherLoc**		**EpiLoc**	
	(*Sens Spec)*			
ex	0.79	0.99	0.87	0.97
mi	0.95	0.75	0.67	0.64
cy	0.79	0.59	0.59	0.56
nu	0.56	0.85	0.47	0.74
Acc	0.71		0.66	

Table 6.2.1: Results from running SherLoc and EpiLoc on the Diff48 dataset. Results are only shown for locations that have more than 20 associated proteins in the dataset.

both systems, the accuracy is lower than their respective accuracy on the cross-validation data.

SherLoc performs very well on the extracellular and the mitochondrial proteins, but less well on the cytoplasmic and the nuclear proteins. SherLoc's specificity on extracellular proteins exceeds all previously reported predictive results, including its own, on cross-validation data. The system's sensitivity on extracellular proteins is only slightly lower than the same measure on cross-validation data. For mitochondrial proteins, SherLoc's performance compares favorably to its performance on cross-validation data. The classifier predicts the location of these proteins with a sensitivity that exceeds any sensitivity reported by a classifier that assigns proteins to more than four locations. SherLoc's specificity on mitochondrial proteins is slightly lower than its demonstrated specificity on cross-validation data, but still remains high.

SherLoc's sensitivity on cytoplasmic proteins is similar to its performance on cross-validation data, as is its specificity on nuclear proteins. However, SherLoc's specificity on cytoplasmic proteins, and its sensitivity on nuclear proteins, is considerably lower than that reported for cross-validation data. The lower performance on the two measurements is caused by a well-known problem in distinguishing between proteins from the two locations. The majority of misclassified nuclear proteins are classified as cytoplasmic, and vice versa. Specifically, 50 of the 54 misclassified nuclear proteins are classified as cytoplasmic, and 10 of the 19 misclassified cytoplasmic proteins are classified as nuclear. If we were to view the cytoplasm and the nucleus as a single class, as TargetP does, the sensitivity and specificity of SherLoc for the location would rise to above 0.90. However, our goal for SherLoc is to predict as specific a location as possible for each protein, and therefore we do not consider such a combination.

EpiLoc's results, while not as good as those obtained on cross-validation data, are still quite promising. The text-based system's specificity and sensitivity on extracellular proteins are higher than those obtained on cross-validation data. For the mitochondria, EpiLoc's results are quite a bit lower than those for cross-validation data. However, as the mitochondria has so few associated proteins, these results are unlikely to represent the performance for a larger set of proteins. EpiLoc's performance on nuclear and cytoplasmic proteins shares a similar pattern with SherLoc; EpiLoc's sensitivity for the cytoplasm and specificity for the nucleus are similar to its cross-validation results, while the system's specificity for the former and sensitivity for the latter are quite a bit lower. EpiLoc, too, performs poorly when distinguishing between proteins from the two locations; 33 of the 59 misclassified cytoplasmic proteins are classified as nuclear, and 15 of the 37 misclassified nuclear proteins are classified as cytoplasmic. If, like TargetP, we were to consider the two locations as one, EpiLoc's sensitivity and its specificity for the single location would rise to 0.76 and 0.93, respectively.

The results of running SherLoc on the *Diff48* set indicate that we can expect SherLoc to perform well on proteins not included in the training set. They further suggest that EpiLoc, too, should yield quite good results when used to predict a protein's subcellular location. Although some of the results of the two systems on *Diff48* are not as good as those reported on cross-validation data (which is expected given the small dataset), these results still confirm that the systems can be applied successfully to data outside the cross-validation setting. We expect the actual prediction performance to be closer to the one obtained on the cross-validation data when the systems are applied to a larger set of proteins. For EpiLoc, these results also show that by using text that predates a confirmed subcellular location, location can indeed be predicted.

To allow SherLoc and EpiLoc to classify the proteins with which they were presented, textless proteins were excluded from the *Diff48* dataset. However, in order to act as a fully functional prediction system, the text-based component of both systems must be able to predict the location of textless proteins. In the next section, we examine the effectiveness of using HomoLoc and PubLoc to represent textless proteins.

6.3 HomoLoc, PubLoc, and DiaLoc Results

In this section, we evaluate the possibility of using HomoLoc and PubLoc to handle textless proteins for the EpiLoc classifier, as described in Section 4.3. Although we have yet to extensively test DiaLoc, we also provide an example of its usage.

6.3.1 HomoLoc and PubLoc

In order to select a method for handling textless proteins, we first apply HomoLoc and PubLoc to the 499 animal proteins and the 609 plant proteins[10] in the MultiLoc dataset that are textless. We present the results of predicting the location of these proteins, which are represented by each of the two methods, using an EpiLoc classifier trained on proteins in the MultiLoc dataset that *do* have associated text. For comparison, we also show the results of assigning text to the proteins using the simpler versions of HomoLoc and PubLoc, denoted SimpHom and SimpPub, respectively. SimpHom takes text only from the top homolog of a protein, instead of from the top three as is done in HomoLoc,

[10] Results for fungal proteins are reported in Table B.3 in Appendix B.

and SimpPub takes text from the three most recent abstracts returned by a PubMed search for a protein, as opposed to taking the five most recent as is done in PubLoc. Table 6.3.1 displays the results for the four different methods.

.

MultiLoc Textless Proteins												
Loc	**HomoLoc**			**SimpHom**			**PubLoc**			**SimpPub**		
	Plant (*Sens Spec MCC)*											
va	**0.64**	**0.09**	0.20	0.50	0.07	0.14	0.54	0.18	**0.29**	0.39	0.14	0.20
pe	0.46	**1.00**	**0.67**	0.36	0.73	0.50	0.52	0.65	0.57	**0.62**	0.65	0.62
go	**0.60**	**1.00**	**0.77**	0.50	0.83	0.64	0.25	0.40	0.31	0.00	0.00	0.00
er	**0.66**	**0.92**	**0.77**	0.63	0.71	0.65	0.61	0.48	0.51	0.55	0.45	0.46
ch	**0.88**	**0.80**	**0.80**	0.83	0.79	0.77	0.29	0.78	0.42	0.30	0.73	0.41
mi	**0.84**	**0.97**	**0.89**	0.63	0.89	0.72	0.59	0.64	0.57	0.56	0.69	0.58
nu	0.75	**0.93**	**0.79**	**0.76**	0.81	0.71	0.70	0.92	0.76	0.68	0.85	0.70
ex	**0.68**	**0.52**	**0.58**	0.59	0.37	0.44	0.50	0.23	0.30	0.46	0.20	0.26
pm	**0.74**	**0.98**	**0.84**	0.66	0.90	0.75	0.59	0.78	0.64	0.56	0.71	0.59
cy	**0.62**	**0.77**	**0.63**	0.51	**0.77**	0.56	0.70	0.47	0.43	0.64	0.45	0.38
Acc	**0.731**			0.658			0.574			0.544		
Avg	**0.686**			0.597			0.529			0.475		
	Animal (*Sens Spec MCC)*											
ly	0.80	0.57	**0.67**	0.60	**0.60**	0.60	**1.00**	0.36	0.60	**1.00**	0.31	0.55
pe	0.46	**1.00**	**0.67**	0.36	0.89	0.56	0.52	0.61	0.55	**0.67**	0.61	0.62
go	**0.90**	0.11	0.29	0.80	0.10	0.25	0.38	**0.38**	**0.36**	0.25	0.50	0.35
er	**0.69**	**0.73**	**0.69**	0.66	0.52	0.55	0.61	0.59	0.57	0.52	0.57	0.51
mi	**0.91**	**0.94**	**0.91**	0.67	0.92	0.76	0.59	0.70	0.59	0.56	0.70	0.57
nu	**0.80**	**0.95**	**0.82**	0.76	0.83	0.71	0.73	0.91	0.75	0.73	0.86	0.73
ex	**0.73**	**0.70**	**0.70**	0.59	0.59	0.57	0.41	0.21	0.24	0.41	0.20	0.24
pm	**0.68**	**0.94**	**0.77**	0.63	0.93	0.74	0.59	0.76	0.63	0.56	0.63	0.54
cy	**0.74**	**0.86**	**0.74**	0.71	0.81	0.69	0.70	0.60	0.51	0.63	0.58	0.46
Acc	**0.762**			0.685			0.644			0.618		
Avg	**0.745**			0.643			0.614			0.592		

Table 6.3.1: Prediction performance of HomoLoc, SimpHom, PubLoc, and SimpPub on the textless proteins of the MultiLoc dataset (499 animal proteins and 609 plant proteins). Both location-specific (*Sens*, *Spec*, *MCC*) and overall results (*Acc* and *Avg*) are shown. Highest values appear in bold.

The results shown in Tables 6.3.1 clearly indicate that HomoLoc produces the best results among the four methods. Its accuracy and average sensitivity greatly exceed those produced by SimpHom, PubLoc, and SimpPub. Moreover, HomoLoc generates the highest scores for 47 out of the 57 location-specific plant and animal scores.

PubLoc does not perform as well as HomoLoc for associating text with textless proteins. We note that the current version of PubLoc uses a simple criterion to select abstracts to be associated with a protein – the five abstracts most recently entered into PubMed that are returned by a search for the protein's name and its corresponding gene's name separated by the "OR" Boolean operator. A more complex method of scoring returned abstracts might improve PubLoc's performance. One approach might be to rank the returned abstracts according to the number of times each of the two names occurs in an abstract. While the current version of PubLoc does improve on SimpPub, the results of classifying proteins using PubLoc are not good enough to justify using it instead of HomoLoc to handle textless proteins. For the EpiLoc system, PubLoc would only be used if no suitable homologs with associated text could be found for a protein.

HomoLoc is the best performing method, among the methods we have tried, on the MultiLoc textless proteins. Therefore, we select it as our preferred method for handling textless proteins, and compare its performance to that of EpiLoc. To do so, we use HomoLoc to classify the same set of proteins from the MultiLoc dataset as classified by EpiLoc; that is, proteins that do have associated text in their Swiss-Prot entry. To apply HomoLoc to these proteins, we ignore the text in a protein's Swiss-Prot entry, and instead use the text associated with the protein's homologs. The results of classifying these proteins, after using both HomoLoc and EpiLoc to associate text with them, are shown in Table 6.3.2.

HomoLoc's performance on the MultiLoc dataset is very similar to that of EpiLoc. The accuracy of HomoLoc is slightly higher than that of EpiLoc for both plant and animal proteins. The systems' average sensitivities are identical on plant proteins, and nearly the same on animal proteins. Moreover, for six of the locations (*va*, *mi*, *nu*, *ex*, *pm*, *cy*), the location-specific scores of EpiLoc and of HomoLoc all differ by less than 0.05. As for the other five locations (*pe, ly, ch, er*, *go*), no location-specific measure differs by more than 0.11 between EpiLoc and HomoLoc, and several of the measures

MultiLoc Dataset		
Loc	**HomoLoc**	**EpiLoc**
	Plant (*Sens Spec MCC)*	
va	0.77 0.27 0.45	0.73 0.29 0.45
pe	0.77 0.68 0.72	0.88 0.74 0.80
go	0.87 0.64 0.74	0.85 0.55 0.67
er	0.79 0.71 0.74	0.72 0.64 0.67
ch	0.81 0.79 0.79	0.89 0.75 0.81
mi	0.78 0.83 0.79	0.82 0.81 0.80
nu	0.86 0.82 0.82	0.82 0.81 0.79
ex	0.85 0.81 0.80	0.84 0.82 0.80
pm	0.89 0.91 0.87	0.85 0.91 0.85
cy	0.66 0.79 0.65	0.64 0.78 0.63
Acc	0.803 (± 0.005)	0.790 (± 0.002)
Avg	0.805 (± 0.005)	0.805 (± 0.005)
	Animal (*Sens Spec MCC)*	
ly	0.84 0.49 0.63	0.86 0.39 0.57
pe	0.80 0.69 0.74	0.90 0.77 0.82
go	0.90 0.72 0.80	0.88 0.62 0.73
er	0.77 0.67 0.71	0.74 0.59 0.65
mi	0.79 0.84 0.80	0.82 0.82 0.80
nu	0.87 0.84 0.83	0.84 0.81 0.80
ex	0.83 0.83 0.79	0.80 0.82 0.77
pm	0.89 0.91 0.87	0.85 0.90 0.84
cy	0.72 0.80 0.67	0.68 0.79 0.65
Acc	0.812 (± 0.010)	0.792 (±0.008)
Avg	0.822 (± 0.005)	0.818 (±0.005)

Table 6.3.2: Prediction performance of HomoLoc and EpiLoc on the MultiLoc dataset. Both location-specific (*Sens*, *Spec*, *MCC*) and overall results (*Acc* and *Avg*) are shown. Standard deviations (denoted ±) are provided for *Acc* and *Avg*.

differ by less than 0.05.

Overall, HomoLoc's performance is, in fact, slightly better than EpiLoc's. HomoLoc's improved performance is likely to be a result of the large amount of text that the method associates with each protein. HomoLoc utilizes the abstracts associated with a protein's three most similar homologs, whereas EpiLoc uses only the abstracts associated with the protein itself. Having more abstracts, originating from the three close homologs of the protein, 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.

As stated in Section 4.1, our evaluation of EpiLoc does not include the textless proteins from each of the TargetP, PLOC, and MultiLoc datasets. Consequently, when applied to the three datasets, EpiLoc predicts the location of 91.4%, 85.8%, and 89.7% of the proteins, respectively. We note that if we apply HomoLoc (as described in Section 3.5.1) to assign text to the textless proteins, EpiLoc predicts the location of 100% of the proteins, while maintaining its high accuracy (for example, an overall accuracy of *0.81* on the MultiLoc dataset).

Based on the above results, we believe that HomoLoc is an effective method for handling textless proteins. The results indicate that using HomoLoc will lead to subcellular location predictions that are likely to be as reliable as those made when using the primary method of EpiLoc to associate text with proteins. In fact, given that for the testing of HomoLoc we excluded text directly associated with a test protein when representing it, HomoLoc may be even more effective than reported here. Therefore, HomoLoc may even warrant consideration as the primary method for associating text with *all* proteins.

Together, PubLoc and HomoLoc can support text-based representation for most proteins. HomoLoc is the preferred method to be used to associate text with a textless protein, while PubLoc should only be used if no suitable homologs with associated text exist. In situations that do not involve the large-scale annotation of proteins, DiaLoc may be used to obtain text for a textless protein.

6.3.2 DiaLoc Example

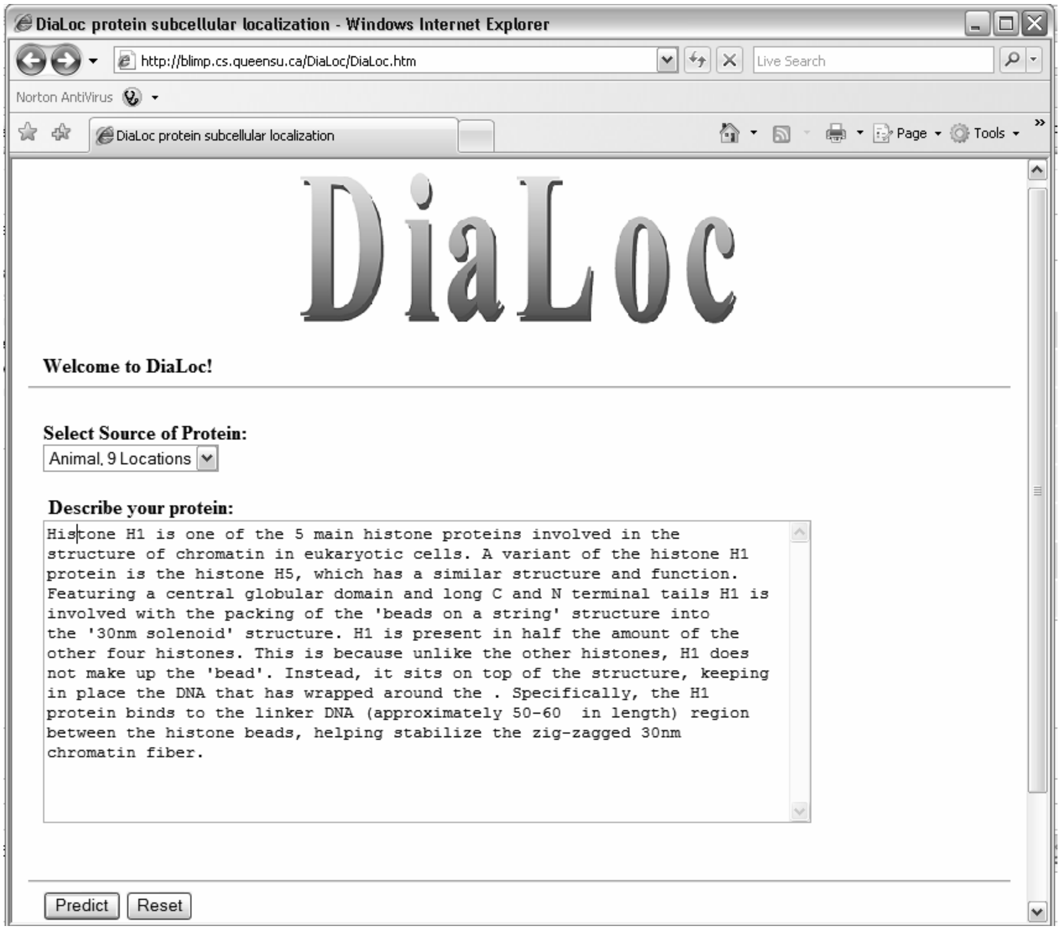
A proper evaluation of DiaLoc requires a study over a prolonged period of time, in which researchers will use the web-interface to enter text and evaluate the results. Thus we have not yet quantitatively tested the performance of DiaLoc for assigning text to textless proteins. Here we only demonstrate DiaLoc by example, and do not formally evaluate it.

For our example we use DiaLoc to predict the location of the protein histone H1, a nuclear protein involved in the structure of DNA. For the "expert" text describing the protein, we use the description of the protein found at the Wikipedia website [53]. Such a Wikipedia entry has the high-level description we expect to obtain from an expert who has some knowledge about the protein, but is still searching for more details. From the Wikipedia [53] website[11], we obtain the following text to describe the protein:

> Histone H1 is one of the 5 main histone proteins involved in the structure of chromatin in eukaryotic cells. A variant of the histone H1 protein is the histone H5, which has a similar structure and function. Featuring a central globular domain and long C and N terminal tails H1 is involved with the packing of the 'beads on a string' structure into the '30nm solenoid' structure. H1 is present in half the amount of the other four histones. This is because unlike the other histones, H1 does not make up the nucleosome 'bead'. Instead, it sits on top of the structure, keeping in place the DNA that has wrapped around the nucleosome. Specifically, the H1 protein binds to the linker DNA (approximately 50-60 nucleotides in length) region between the histone beads, helping stabilize the zig-zagged 30nm chromatin fiber.

[11] Downloaded December 18, 2006 from: http://en.wikipedia.org/wiki/Histone_H1.

We remove from this text any word whose first five letters are *nucle*. This removal ensures that these words, which may be viewed as indicators of a nuclear protein, are not included in the protein's text vector (note that terms such as "nucleotide" are thus also removed). We then enter the text in the DiaLoc web-interface, and select *Animal* as our source of the protein, as shown in Figure 6.3.1.


DiaLoc protein subcellular localization - Windows Internet Explorer
http://blimp.cs.queensu.ca/DiaLoc/DiaLoc.htm
Live Search
Norton AntiVirus
DiaLoc protein subcellular localization
Page
Tools
DiaLoc
Welcome to DiaLoc!
Predict
Reset

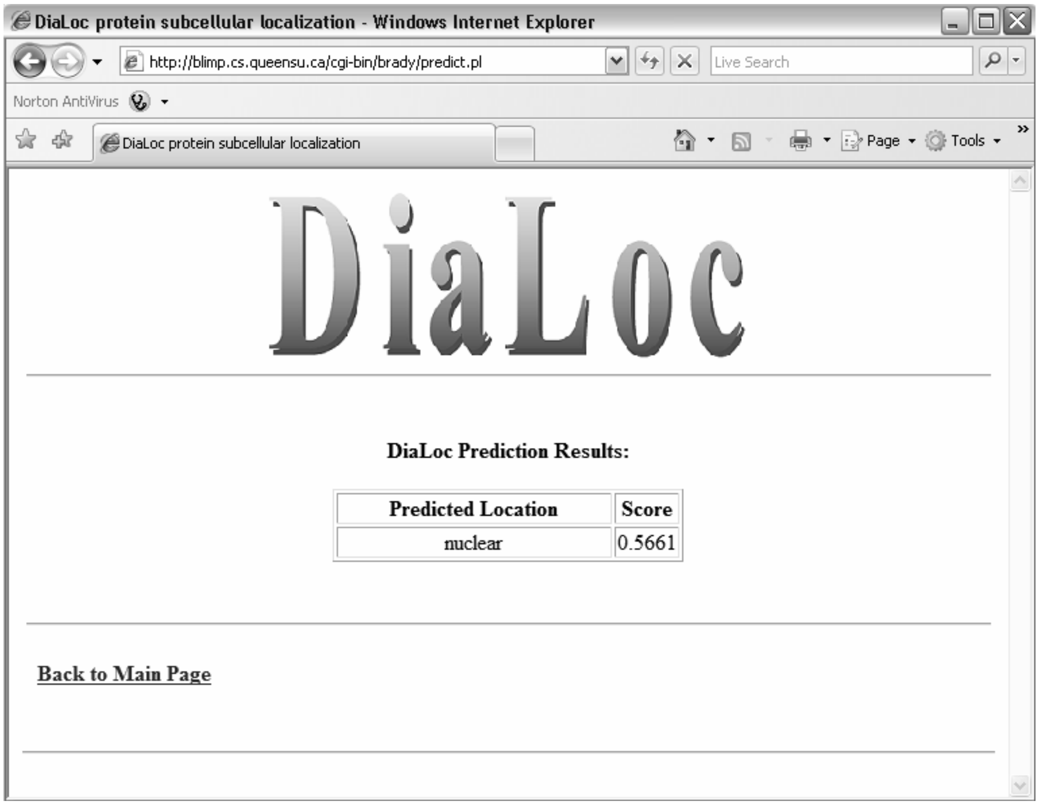
Select Source of Protein:
Animal, 9 Locations

Describe your protein:

```
Histone H1 is one of the 5 main histone proteins involved in the
structure of chromatin in eukaryotic cells. A variant of the histone H1
protein is the histone H5, which has a similar structure and function.
Featuring a central globular domain and long C and N terminal tails H1 is
involved with the packing of the 'beads on a string' structure into
the '30nm solenoid' structure. H1 is present in half the amount of the
other four histones. This is because unlike the other histones, H1 does
not make up the 'bead'. Instead, it sits on top of the structure, keeping
in place the DNA that has wrapped around the . Specifically, the H1
protein binds to the linker DNA (approximately 50-60  in length) region
between the histone beads, helping stabilize the zig-zagged 30nm
chromatin fiber.
```

Figure 6.3.1: The first page of the DiaLoc web-interface. The text from the Wikipedia website pertaining to the histone H1 protein has been entered, and any words with the first five letters *nucle* have been removed.

Upon pressing the predict button, the DiaLoc web server presents the page shown in Figure 6.3.2. As the figure shows, DiaLoc correctly predicts the location of the protein as the nucleus with a probability of 0.5661; this is a strong prediction, as there are nine different locations the probability distribution is divided amongst. This example demonstrates that the DiaLoc web server is a functional program and can be used to predict a protein's subcellular location.


DiaLoc protein subcellular localization - Windows Internet Explorer
http://blimp.cs.queensu.ca/cgi-bin/brady/predict.pl
Live Search
Norton AntiVirus
DiaLoc protein subcellular localization
Page
Tools
DiaLoc
Back to Main Page

DiaLoc Prediction Results:

Predicted Location	Score
nuclear	0.5661

Figure 6.3.2: The prediction page of the DiaLoc web-interface. DiaLoc assigns Histone H1 to the nucleus with a probability of 0.5661.

By using the three modules HomoLoc, PubLoc, and DiaLoc, to handle textless proteins for the EpiLoc system, we create a text-based prediction system that not only produces accurate results, but can also provide a subcellular location prediction for almost any protein. Based on the results presented in this thesis, we believe that the predictions of subcellular location for newly discovered proteins made by EpiLoc will be quite reliable. These predictions should be able to serve as a guide for researchers, in order to speed up and improve the research of other protein properties.

Chapter 7

Conclusion and Future Work

We introduced in this thesis a new text-based system, EpiLoc, for predicting protein subcellular location. We also described, SherLoc [19, 47], an integration of an early version of the EpiLoc system with the previously developed sequence-based classifier, MultiLoc [20]. Both EpiLoc and SherLoc have been compared to other state-of-the-art classifiers using their respective datasets. Moreover, we have reported the results of applying both of these classifiers to the newly formed dataset, *Diff48*, to test their effectiveness outside of the cross-validation setting. For the EpiLoc classifier we have also developed three alternative approaches for assigning text to textless proteins. For two of these methods, PubLoc and HomoLoc, we have performed experiments measuring their reliability when applied to textless proteins. For the third method, DiaLoc, we have demonstrated its utility through an example.

7.1 Summary of Contributions

The work presented in this thesis demonstrates the following contributions to the prediction of protein subcellular location:

1. We have produced a text-based system that predicts subcellular location as effectively as, and often better than, other state-of-the-art systems. Moreover, we have demonstrated that EpiLoc may be effectively applied to proteins not included in cross-validation studies. Furthermore, HomoLoc has been shown to be as effective as the primary method of EpiLoc for assigning text to proteins. By using HomoLoc, PubLoc and DiaLoc, our system can associate text with practically any protein, and predict its location.
2. In collaboration with the group that developed the MultiLoc classifier [20, 19, 47], we have demonstrated that SherLoc significantly outperforms all previous state-of-the-art prediction systems. SherLoc was compared with the MultiLoc, PLOC [38], and TargetP [15] systems using their own datasets, and produced the best results. Additionally, the performance of SherLoc was validated by its application to the set of proteins with newly assigned locations, *Diff48*. Overall, SherLoc demonstrated unprecedented performance for predicting a protein's subcellular location.
3. We have demonstrated that a text-based system can be used to improve the performance of a sequence-based system. SherLoc showed a statistically significant improvement in performance over each of its components. Our results demonstrate, for the first time, that an integrated text- and sequence-based

approach to a biological problem can achieve a quantitative and significant improvement over a system that uses biological data alone.

7.2 Future Work

There are several natural extensions to this work. The first is the integration of the current version of EpiLoc into the SherLoc system. As was discussed in Section 6.1, EpiLoc outperforms its predecessor, EarlyText. Incorporating EpiLoc into SherLoc has the potential to further improve upon SherLoc's performance.

SherLoc's performance, with or without the integration of EpiLoc, should be validated further. We have predicted the location for the *Unknown* set of proteins, which currently have no assigned location. Experimentally determining the location of these proteins should be used to validate their predicted location. This could provide further evidence concerning the reliability and usability of the SherLoc system. In the interim, the current predictions can serve as clues for researchers interested in discovering a protein's functional characteristics.

Experiments should also be performed to determine the effectiveness of DiaLoc. This will require the involvement of a biologist who is working with a set of proteins of unknown location. The biologist will need to enter information concerning each protein into DiaLoc, and then determine if the predicted location is correct. This should be done on proteins of unknown location, so that no bias towards a certain location appears in the biologist's description of each protein.

A further extension to this work is the expansion of EpiLoc and SherLoc to predict the intraorganelle location of a protein. Several locations within the cell can be divided into subcompartments. For instance, the mitochondria consists of an inner and an outer membrane, the area that the inner membrane surrounds (the matrix), and the cristae (folds within the inner membrane). Expanding the system to offer more precise location predictions would provide users with further insight regarding the protein's function.

The good performance of both EpiLoc and SherLoc has already been demonstrated. Undertaking further experimentation, as described above, is expected to improve and validate the systems as useful prediction tools.

Bibliography

[1] Andrade, M.A., O'Donoghue, S.I., Rost, B.: Adaptation of Protein Surfaces to Subcellular Location. *Journal of Molecular Biology*, **276**, 517-525, 1998.

[2] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic Local Alignment Search Tool. *Journal of Molecular Biology*, **215**, 403-410, 1990.

[3] Bairoch, A., Apweiler, R.: The SWISS-PROT protein sequence database and its supplement in TrEMBL in 2000. *Nucleic Acids Research*, **28**, 45-48, 2000.

[4] Bannai, H., Tamada, Y., Maruyama, O., Nakai, K. , Miyano, S.: Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298-305, 2002.

[5] Brenner, S.B., Chothia, C., Hubbard, T.J.P.: Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. In*: Proc. of the National Academy of Sciences of the United States of America* (*PNAS*), **95**, 6073-6078, 1998.

[6] Cai, Y.D., Chou, K.C.: Predicting 22 protein localization in budding yeast. *Biochemical and Biophysical Research Communications*, **323**, 425-429, 2004.

[7] Campbell, N.A., Reece, J.B., Mitchell, L.G. *Biology* (5th ed.), Benjamin-Cummings, NY, 1999.

[8] Cedano, J., Aloy, P., Perez-Pons, J.A., Querol, E.: Relation Between Amino Acid Composition and Cellular Location of Proteins. *Journal of Molecular Biology*, **266**, 594-600, 1997.

[9] Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines, 2001. *http://www.csie.ntu.edu.tw/~clin/libsvm/*.

[10] Craven, M., Kumlien, J.: Constructing Biological Knowledge Bases by Extracting Information from Text Sources. In: *Proc. of the International Conference on Intelligent Systems for Molecular Biology* (*ISMB*), 77-86, 1999.

[11] Degroot, M.H. Probability and Statistics, Addison-Wesley, MA, 1991.

[12] Dönnes, P., Höglund, A.: Predicting Protein Subcellular Localization: Past, Present, and Future. *Genomics, Proteomics, and Bioinformatics*, **2**, 209-215, 2004.

[13] Eisenhaber, F., Bork, P.: Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics*, **15**, 528-525, 1999.

[14] Emanuelsson, O., Nielsen, H., von Heijne, G.: ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science*, **8**, 978-984, 1999.

[15] Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G.: Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, **300**, 1005-1016, 2000.

[16] Eskin, E., Agichtein, E.: Combining text mining and sequence analysis to discover protein functional regions. In: *Proc. of the Pacific Symposium on Biocomputing* (*PSB*), 288-299, 2004.

[17] GenomeNet Japan: AAindex Amino acid indices and similarity matrices, 2005. *http://www.genome.ad.jp/aaindex/*.

[18] Hanson, M.R., Köhler, R.H.: GFP imaging: Methodology and application to investigate cellular compartmentation in plants. *Journal of Experimental Botany*, **52**, 529-539, 2001.

[19] Höglund, A., Blum T., Brady, S., Donnes, P., Miguel, J.S., Rocheford, M., Kohlbacher, O., Shatkay, H.: Significantly Improved Prediction of Subcellular Localization by Integrating Text and Protein Sequence Data. In: *Proc. of the Pacific Symposium on Biocomputing* (*PSB*), 16-27, 2006.

[20] Höglund, A., Dönnes, P., Blum, T., Adolph, H., Kohlbacher, O.: MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22**, 1158-1165, 2006.

[21] Horton, P., Nakai, K.: A probabilistic classification system for predicting the cellular localization of proteins. In: *Proc. of the International Conference on Intelligent Systems for Molecular Biology* (*ISMB*), 109-115, 1996.

[22] Horton, P., Nakai, K.: Better prediction of protein cellular localization sites with the k nearest neighbors classifier. In: *Proc. Of the International Conference on Intelligent Systems for Molecular Biology* (*ISMB*), 147-152, 1997.

[23] Hua, S., Sun, Z.: Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721-728, 2001.

[24] Hulo, N., Sigrist, C.J.A., Le Saux, V., Lagendijk, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., Bairoch, A.: Recent improvements to the PROSITE database. *Nucleic Acids Research*, **32**, D134-D137, 2004.

[25] The International Human Genome Mapping Consortium: A physical map of the human genome. *Nature*, **409**, 934-941, 2001.

[26] Kumar, A., Agarwal, S., Heyman, J.A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid A., Jansen, R., Yang, L., Cheung, K.H., Miller, P., Gerstein, M., Roeder, G.S., Snyder, M.: Subcellular localization of the yeast proteome. *Genes & Development*, **16**, 707-719, 2002.

[27] Lan, M., Tan, C.L., Low, H.B., Sun, S.Y.: A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines. *International World Wide Web Conference*, 1032-1033, 2005.

[28] Leslie, C., Eskin, E., Noble, W.S.: The spectrum kernel: A string kernel for SVM protein classification. In: *Proc. of the Pacific Symposium on Biocomputing* (*PSB*), 566-575, 2002.

[29] Lipp, J., Dobberstein, B.: Signal and Membrane Anchor Functions Overlap in the Type II Membrane Protein IγCAT. *The Journal of Cell Biology*, **106**, 1813-1820, 1988.

[29] Matthews, B.W.: Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, **405**, 442-451, 1975.

[30] Mitchell, T.M.: Machine Learning, McGraw-Hill Education, NY, 1997.

[31] Nair, R., Carter, P., Rost, B.: NLSdb: database of nuclear localization signals. *Nucleic Acids Research*, **31**, 397-399, 2003.

[32] Nair, R., Rost, B.: Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, **18**, S78-S86, 2002.

[33] Nair, R., Rost, B.: Sequence conserved for subcellular localization. *Protein Science*, **11**, 2836-2847, 2002.

[34] Nakai, K., Kanehisa, M.: A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897-911, 1992.

[35] NCBI: PubMed Overview, 2006. *http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html*.

[36] Nielsen, H., Engelbrecht, J., Brunak., S., von Heijne, G.: Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Proteins Engineering*, **10**, 1-6, 1997.

[37] Pennisi, E.: Why Do Humans Have So Few Genes? *Science*, **309**, 80, 2005.

[38] Park, K.J., 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.

[39] Porter, M.: An algorithm for suffix stripping. *Program*, **14**, 130-137, 1980.

[40] Provost, F.: Learning with Imbalanced Data Sets 101. Invited paper for the American Association for Artificial Intelligence 2000 Workshop on Imbalanced Data Sets, 2000.

[41] Reinhardt, A., Hubbard, T.: Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Research*, **26**, 2230-2236, 1998.

[42] Rothman, K.J.: Writing for Epidemiology. *Epidemiology*, **9**, 333-337, 1998.

[43] Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O., Ofran, Y.: Automatic prediction of protein function. *Cellular and Molecular Life Sciences*, **60**, 2637-2650, 2003.

[44] Sebastiani, F.: Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, **34**, 1-47, 1999.

[45] Schneider, G., Fechner, U.: Advances in the prediction of protein targeting signals. *Proteomics*, **4**, 1571-1580, 2004.

[46] Shannon, C.E.: Prediction and entropy of printed English. *Bell System Technical Journal*, **30**, 50-64, 1951.

[47] Shatkay, H., Höglund, A., Brady, S., Blum, T., Dönnes, P., Kohlbacher, O.: SherLoc: High-Accuracy Prediction of Protein Subcellular Localization by integrating Text and Proteins Sequence Data. *Bioinformatics*, 2007.

[48] Soares, C., Brazdil, P.B., Kuba, P.: A Meta-Learning Method to Select the Kernel Width in Support Vector Regression. *Machine Learning*, **54**, 195-209, 2004.

[49] Stapley, B.J., Kelley, L.A., Sternberg, M.J.E.: Predicting the subcellular location of proteins from text using support vector machines. In: *Proc. of the Pacific Symposium on Biocomputing* (*PSB*), 374-385, 2002.

[50] Venter, J.C., Adams, M.D., Myers, E.W. et al.: The sequence of the human genome. Science, **291**, 1304-1351, 2001.

[51] Walpole, R.E., Myers, R.H., Myers, S.L. Probability and Statistics for Engineers and Scientists, Prentice-Hall, 235-335, 1998.

[52] Weaver, R.F. Molecular Biology (2nd ed.) McGraw-Hill Education, NY, 2002.

[53] Wikipedia contributors: Histone H1. Wikipedia, The Free Encyclopedia, 2006.

[54] Wu, T.-F., Lin, C.-J., Weng, R.C.: Probability Estimates for Multi-Class Classification by Pairwise Coupling. *Journal of Machine Learning Research*, **5**, 975-1005, 2004.

[55] Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: *Proc. Of ICML-97, 14th International Conference on Machine Learning* (*ICML*), 412-420, 1997.

Glossary

EarlyText Early version of the EpiLoc predictor that is integrated with MultiLoc to form the SherLoc classifier. Presented in this thesis.

EpiLoc Text-based predictor of subcellular location presented in this thesis.

DiaLoc Method for handling textless proteins; text about a textless protein is obtained from the scientist researching it. Presented in this thesis.

HomoLoc Method for handling textless proteins. Associates with a textless protein the abstracts of its three closest homologs. Presented in this thesis.

MultiLoc Sequence-based predictor of subcellular location that uses amino acid composition data, N-terminal sequence data, and sequence motif data to represent proteins [20].

PLOC Sequence-based predictor of subcellular location that uses amino acid composition data to represent proteins [38].

PubLoc Method for handling textless proteins; uses a PubMed search to retrieve five abstracts to be associated with a textless protein. Presented in this thesis.

SherLoc Integrated text- and sequence based predictor of subcellular location. SherLoc integrates the MultiLoc classifier with the EarlyText classifier [19, 47]. Presented in this thesis and in co-authored earlier publications.

SimpHom Simple version of HomoLoc, associates with a textless protein the abstracts of its single closest homolog, as opposed to its three closest homologs. Presented in this thesis.

SimpPub Simple version of PubLoc, retrieves three abstracts to be associated with a textless protein as opposed to five abstracts. Presented in this thesis.

TargetP Sequence-based predictor of subcellular location that uses N-terminal sequence data to represent proteins [15].

Textless protein A protein for which a term-vector cannot be made when using the primary method of EpiLoc.

***Z-Test* method** Feature selection method that uses the Z-test [51] to determine if the probability of a term being associated with one location is statistically significantly different from the probability of the term being associated with any other location.

Appendix A

Stop Word List

a
about
above
across
after
afterwards
again
against
al
all
almost
alone
along
already
also
although
always
am
among
amongst
an
analyze
and
another
any
anyhow
anyone
anything
anywhere
applicable
apply
are
around
as
assume
at
be
became
because
become
becomes
becoming
been
before
beforehand
being
below
beside
besides
between
beyond
both
but
by
came
cannot
cc
cm
come
compare
could
de
dealing
department
depend
did
discover
dl
do
does
during
each
ec
ed
effected
eg
either
else
elsewhere
enough
et
etc
ever
every
everyone
everything
everywhere
except
find
for
found
from
further
get
give
go
gov
had
has
have
he
hence
her
here
hereafter
hereby
herein
hereupon
hers
herself
him
himself
his
how
however
hr
ie
if
ii
iii
in
inc
incl
indeed
into
investigate
is
it
its
itself
j
jour
journal
just
kg
last
latter
latterly
lb
ld
letter
like
ltd
made
make
many
may
me
meanwhile
mg
might
ml
mm
mo
more
moreover
most
mostly
mr
much
must
my
myself
namely
neither
never
nevertheless
next
no
nobody
noone
nor
not
nothing
now
nowhere
of
off
often
on
only
onto
or
other
others
otherwise
our
ours
ourselves
out
over
own
oz
per
perhaps
pm
precede
presently
previously
pt
rather
regarding
relate
said
same
seem
seemed
seeming
seems
seriously
several
she
should
show
showed
shown
since
so
some
somehow
someone
something
sometime
sometimes
somewhere
still
studied
sub
such
take
tell
th
than
that
the
their
them
themselves
then
thence
there
thereafter
thereby
therefore
therein
thereupon
these
they
this
thorough
those
though
through
throughout
thru
thus
to
together
too
toward
towards
try
type
ug
under
unless
until
up
upon
us
used
using
various
very
via
was
we
were
what
whatever
when
whence
whenever
where
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
whoever
whom
whose
why
will
with
within
without
wk
would
wt
yet
you
your
yours
yourself
yourselves
yr

Table A.1: The set of stop words removed during text processing.

Appendix B

Fungal Protein Results

MultiLoc Dataset												
Loc	**MultiLoc**			**EpiLoc**			**EarlyText**			**SherLoc**		
	Fungal (*Sens Spec MCC)*											
va	0.76	0.24	0.42	0.75	0.25	0.42	0.66	0.13	0.27	**0.78**	**0.26**	**0.44**
pe	0.68	0.30	0.43	0.87	**0.77**	**0.81**	**0.90**	0.69	0.78	0.88	0.73	0.79
go	0.71	0.53	0.60	0.86	**0.58**	**0.70**	0.81	0.41	0.56	**0.87**	0.57	**0.70**
er	0.71	0.59	0.63	0.72	0.60	0.65	0.72	0.55	0.61	**0.80**	**0.69**	**0.74**
mi	0.88	0.82	0.83	0.83	0.82	0.80	0.81	0.80	0.79	**0.95**	**0.90**	**0.92**
nu	0.81	0.74	0.73	0.83	0.79	0.78	0.81	0.72	0.72	**0.90**	**0.82**	**0.84**
ex	0.73	0.81	0.73	**0.85**	0.82	0.80	0.76	0.78	0.72	0.82	**0.88**	**0.82**
pm	0.76	0.89	0.78	**0.86**	0.91	0.85	0.80	0.91	0.81	0.84	**0.96**	**0.87**
cy	0.68	0.85	0.69	0.66	0.79	0.63	0.54	0.75	0.54	**0.82**	**0.92**	**0.82**
Acc	0.749 (± 0.007)			0.790 (±0.007)			0.738 (±0.016)			**0.849** (± 0.008)		
Avg	0.747 (± 0.01)			0.802 (±0.005)			0.750 (±0.014)			**0.850** (± 0.014)		

Table B.1: Prediction performance of MultiLoc, EpiLoc, EarlyText, and SherLoc on the fungal proteins of the MultiLoc dataset. Both location-specific (*Sens*, *Spec*, *MCC*) and overall results (*Acc* and *Avg*) are shown. Highest values appear in bold. Standard deviations (denoted ±) are provided for *Acc* and *Avg* where available.

PLOC Dataset													
Loc	**PLOC**	**EpiLoc**			**MultiLoc**			**EarlyText**			**SherLoc**		
	Sens	**Fungal (*Sens Spec MCC*)**											
go	0.15	0.78	0.35	0.52	0.53	0.07	0.18	**0.88**	**0.41**	**0.60**	0.81	0.24	0.44
cs	0.59	0.83	**0.29**	**0.48**	0.75	**0.29**	0.46	**0.86**	0.25	0.46	0.83	0.23	0.43
va	0.25	0.78	0.18	0.37	0.69	0.19	0.35	0.67	0.10	0.25	**0.83**	**0.28**	**0.48**
er	0.47	0.74	0.25	0.41	0.78	**0.72**	0.74	0.71	0.25	0.41	**0.86**	**0.71**	**0.78**
pe	0.25	**0.85**	0.51	0.65	0.74	0.30	0.46	0.77	0.51	0.62	0.80	**0.63**	**0.70**
mi	0.57	0.77	0.82	0.77	0.69	0.72	0.67	0.76	**0.86**	0.79	**0.85**	**0.86**	**0.84**
ex	0.78	0.68	0.58	0.57	0.83	0.88	0.83	0.66	0.57	0.55	**0.86**	**0.91**	**0.87**
cy	0.72	0.36	0.59	0.36	0.64	0.74	0.63	0.43	0.55	0.39	**0.78**	**0.80**	**0.75**
pm	**0.92**	0.77	0.80	0.74	0.83	0.97	0.87	0.71	0.83	0.71	0.88	**0.98**	**0.91**
nu	**0.90**	0.79	0.87	0.76	0.78	0.87	0.76	0.78	0.90	0.77	0.88	**0.94**	**0.87**
Acc	0.795 (± 0.009)	0.687 (±0.011)			0.758 (±0.008)			0.678 (±0.005)			**0.854** (± 0.008)		
Avg	0.568 (± 0.019)	0.735 (±0.013)			0.725 (±0.025)			0.724 (±0.016)			**0.838** (± 0.028)		

Table B.2: Prediction performance of PLOC, EpiLoc, MultiLoc, EarlyText, and SherLoc on the fungal proteins of PLOC dataset. Overall results (*Acc* and *Avg*) are shown for all systems. For location-specific results, only sensitivity is presented for PLOC, while for all other systems sensitivity, specificity, and *MCC* are each displayed. As presented in its corresponding publication, PLOC's location-specific results are averaged over all three organisms (animal, plant, fungal). Highest values appear in bold. Standard deviations (denoted ±) are provided for *Acc* and *Avg* where available.

MultiLoc Textless Proteins												
Loc	**HomoLoc**			**SimpHom**			**PubLoc**			**SimpPub**		
	Fungal (*Sens Spec* Mcc)											
va	**0.71**	0.09	0.20	0.57	0.07	0.15	0.46	**0.33**	**0.37**	0.39	0.29	0.32
pe	0.41	**0.90**	0.60	0.41	0.69	0.52	**0.52**	0.73	**0.61**	0.48	0.63	0.53
go	**0.70**	**0.78**	**0.73**	0.50	0.50	0.49	0.38	0.50	0.43	0.00	0.00	0.00
er	**0.63**	**0.92**	**0.75**	0.57	0.77	0.64	0.58	0.53	0.52	0.55	0.56	0.52
mi	**0.82**	**0.98**	**0.88**	0.63	0.93	0.74	0.68	0.75	0.67	0.59	0.75	0.62
nu	**0.76**	**0.94**	**0.79**	0.72	0.79	0.66	0.65	0.89	0.69	0.68	0.76	0.63
ex	**0.68**	**0.50**	**0.56**	0.59	0.48	0.51	0.46	0.26	0.30	0.36	0.18	0.20
pm	**0.71**	**0.96**	**0.80**	0.63	0.85	0.70	0.66	0.84	0.71	0.61	0.74	0.62
cy	0.65	**0.83**	**0.67**	0.59	0.78	0.59	**0.71**	0.55	0.47	0.60	0.51	0.38
Acc	**0.707**			0.628			0.640			0.582		
Avg	**0.675**			0.579			0.565			0.472		

Table B.3: Prediction performance of HomoLoc, SimpHom, PubLoc, and SimpPub on the textless fungal proteins of the MultiLoc dataset. Both location-specific (*Sens*, *Spec*, *MCC*) and overall results (*Acc* and *Avg*) are shown. Highest values appear in bold.