JEPSLD: A JUDGMENTAL EUKARYOTIC PROTEIN SUBCELLULAR LOCATION DATABASE

by

Sanjeev Patra

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics and Computational Biology

Spring 2013

JEPSLD: A JUDGMENTAL EUKARYOTIC PROTEIN SUBCELLULAR LOCATION DATABASE

by

Sanjeev Patra

Approved: _____
Hagit Shatkay, PhD
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____
Errol L. Lloyd, PhD
Chair of the Department of Computer & Information Sciences

Approved: _____
Babatunde Ogunnaike, PhD
Dean of the College of Engineering

Approved: _____
James G. Richards, Ph.D.
Vice Provost for Graduate and Professional Education

ACKNOWLEDGMENTS

I would like to express my sincerest thanks to my thesis advisor, Professor Hagit Shatkay, who provided great knowledge and guidance throughout my thesis. I would like to thank her for all the support and advice she offered me throughout my Master's degree at the University of Delaware.

Special thanks to Juilee Patankar for her time and help in processing the data and implementing the JEPSLD database.

My genuine gratitude goes to all members of the Computational Biomedicine Lab at the University of Delaware, especially to Ramunuja Simha for helping me develop the user interface. This thesis would not have been the same without the constant encouragement and support from them.

I would like to thank Dr. Cathy Wu and Dr. Li Liao for serving on my thesis committee and providing their insightful suggestions. I also thank all the CBCB staff and faculty members.

Finally, I would express my sincerest thanks to my parents and friends for their love and support.

SANJEEV PATRA

University of Delaware, Delaware

April 2013

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Proteins play a key role in every biological process. Moreover, they localize to specific compartments or organelles within the cell. Knowledge of the subcellular location of proteins can help elucidate their function, interaction partners and role in disease. Several protein databases have been developed to store information relevant to the subcellular location of proteins in certain organisms. However, there is no current central database that stores all available information about the subcellular location of eukaryotic proteins.

In the work presented here, we surveyed protein databases that store the location of eukaryotic proteins, and selected those that are up-to-date and contain experimental evidence pertaining to the proteins' subcellular location. Five protein data sources were selected, and the relevant data were obtained from them and stored in a new database that we have developed, namely, the Judgmental Eukaryotic Protein Subcellular Location Database (JEPSLD). The JEPSLD database also stores information about proteins that localize to multiple compartments. In addition, for each protein entry, the database contains information about tissue, cell line, cell type, images and PubMed IDs, when available. A user interface allows users to query the database using gene name, protein name, protein accession numbers such as UniProtKB/SwissProt IDs and gene identifiers such as Entrez gene ID and Ensembl gene ID.

Overall, the JEPSLD database is the largest eukaryotic protein subcellular location database to date, storing information about 113,525 eukaryotic proteins. The database will be publicly available and the entire dataset can be downloaded from the JEPSLD database in XML format.

Chapter 1

INTRODUCTION

Proteins are macromolecules made up of long chains of amino acids. They play an essential role in all biological processes and perform specific functions. Some proteins are involved in the enzymatic catalysis of metabolic reactions while others may help in the transportation of molecules across the cell or provide a defense mechanism to the body from diseases [61]. For proteins whose functions are unknown, other properties such as their subcellular location and structure can help provide insight about their function.

The living cell comprises several compartments or subcellular locations, and each compartment is associated with specific biochemical processes. Hence, knowledge about the subcellular location of proteins can help elucidate their metabolic or biochemical functions as well as help understand their role in certain diseases and indicate if they can serve as potential drug targets [26, 67]. Moreover, since proteins localize to specific compartments within the cell, information about their location can help identify their interaction partners. It is also important to study proteins that localize to multiple compartments since they may perform different functions in each compartment. Proteins may behave differently based on different conditions such as tissue, cell line, disease and abnormalities; hence information about their subcellular location under such conditions is also valuable.

Our main objective is to extract all available information about the subcellular location of eukaryotic proteins from reliable data sources and store it in a new resource that we are creating, the Judgmental Eukaryotic Protein Subcellular Location Database (JEPSLD). To obtain reliable information about proteins, we survey several data sources and select those that store experimental information about their subcellular location. We also aim to store information about proteins that localize to multiple compartments.

Throughout the rest of this chapter we provide the motivation for this work, discuss the main contributions and present an outline of the remaining chapters.

1.1 Motivation

Our work is motivated by the need to create a database that stores all available and reliable information about the subcellular location of eukaryotic proteins. Several databases have been developed to store information about the subcellular location of proteins identified using experimental methods or predicted using similarity search and machine learning methods [59]. UniProtKB/SwissProt [11, 82] stores information about 175,158 eukaryotic proteins, as of March 2013, but only 90,475 of them contain experimental evidence for their subcellular location. As another example, the database LOCATE [26, 74], stores the experimental subcellular location of 14,659 mouse proteins and 17,666 human proteins. There are other databases that store the subcellular location of proteins as well as information about the tissue, cell line and cell type from the model organisms [81, 60]. But, there is no central database

currently that stores all available information about the location of eukaryotic proteins. Moreover, very few databases store information about proteins that localize to multiple compartments.

We discuss the databases providing reliable information about the subcellular location of eukaryotic proteins, extract the information from those repositories and store it in our own database, the Judgmental Eukaryotic Protein Subcellular Location Database (JEPSLD). We also aim to gather information about proteins that localize to multiple compartments since it would help researchers study how they function in each location. Additional information such as tissue, cell line, cell type and images are also stored in our database. Since JEPSLD provides experimental information about the subcellular location of proteins, incorporation of the JEPSLD data into the training set of single and multiple location predictors may improve their performance.

1.2 Thesis Contributions

Specifically, the main contributions of this work are:

- We developed the database JEPSLD, a repository that stores all available information about eukaryotic proteins that localize to single or multiple locations.
- We also stored additional information about proteins when available, such as tissue, cell type, cell line, images and PubMed IDs, in the JEPSLD database. This information can be helpful for understanding how proteins may behave differently under different conditions.

- We developed a user interface, which allows users to query the database using protein name, gene name, and protein accession numbers such as UniProt/SwissProt ID or gene identifiers such as Ensembl gene ID.

1.3 Thesis Outline

The rest of the thesis is organized as follows: Chapter 2 provides a background about proteins and their subcellular location, and surveys databases that store the location of proteins. Chapter 3 discusses the selection criteria used to select the data sources whose information is stored in our database, the JEPSLD. We then discuss in detail the data sources that satisfy the selection criteria, the methods they use to collect and store the data and examine whether they provide any additional information such as tissue, cell type, cell line or images. In Chapter 4, we describe in detail the database design, development and the web interface used to query the database. Chapter 5 concludes the thesis and proposes future work.

Chapter 2

BACKGROUND

In this chapter we provide background about eukaryotic proteins and the importance of knowing their location, and survey a few databases that store information about eukaryotic proteins. We briefly introduce the eukaryotic cell, its organelles and the proteins within it. Eukaryotic proteins perform specific functions based on their location; moreover, proteins that localize to the same compartment tend to have similar functions. Hence, knowledge about the subcellular location of a protein can provide insight about its function and interaction partners.

Our main objective is to gather reliable information about the subcellular location of eukaryotic proteins. Hence in this chapter, we also survey protein databases that store information about proteins' locations to understand their data, methods used to extract/generate the data and the reliability of the methods. The subcellular location of a protein is either determined experimentally (e.g. Green Fluorescent Protein tagging, Immunofluorescence) or predicted using computational methods (similarity search or machine learning methods). Finally, we discuss localization of proteins to multiple compartments.

2.1 The Eukaryotic Cell and its Organelles

The cell is the basic unit of life. The eukaryotic cell has an enclosed compartment containing most of the cell's DNA called the *nucleus*. In contrast, prokaryotic cells do not have a nucleus. Additional membrane bound complex organelles such as mitochondria, endoplasmic reticulum, Golgi apparatus and lysosomes play different roles in the cell. Specifically: the mitochondria generates energy (ATP) and regulates cellular metabolism; lysosomes break down and recycle organelles and macromolecules; and the endoplasmic reticulum synthesizes the lipids and proteins of cell membranes and also helps transport other materials throughout the cell; the Golgi apparatus is responsible for the transport of proteins, lipids and other macromolecules within the cell; the cytoplasm is a semi-fluid substance found inside the cell, surrounding and protecting all the organelles. Although each organelle in the eukaryotic cell performs a specific function, they work together in an integrated fashion to meet the overall needs of the cell [1, 24].

Eukaryotic cells contain larger quantities of genetic material (DNA) than prokaryotes. For instance, the genetic material in human cells is about 1000 times more than that in bacterial cells [24]. DNA is made up of a sequence of four nucleotides: adenine, cytosine, guanine and thymine. Genes are smaller subsequences of the DNA and they code for proteins. A protein sequence is made up of long chains of subunits called amino acids. The amino acid sequence is unique for every protein and it is specified by the nucleotide sequence of the gene encoding that protein [24]. The Gene Ontology (GO) contains a hierarchy of terms used for indexing and

retrieving information about genes and gene products (proteins). GO is represented as an acyclic graph where each entry or GO term forms a node and the nodes are connected with one another using arcs that define the relationships between the nodes [78].

2.2 Proteins

Proteins are biological macromolecules that play a fundamental role in every process within the living cell. The production of proteins from DNA involves two major steps. The first step is transcription during which the double stranded DNA sequence gets copied into a single stranded mRNA. In the next step, known as translation, the ribosome helps in the synthesis of proteins from the mRNA. In eukaryotes mRNA is produced in the nucleus and then transported across the nuclear membrane into the cytoplasm for translation [20]. After translation, proteins are transported to different subcellular locations, where they perform their specific roles in various biological processes. The function of proteins may range from enzymatic catalysis of biochemical reactions to the maintenance of the electrochemical potential across the cell membrane. Knowledge about the function of a protein is essential to understand its importance in regulating the biological pathways and the working of a living cell [19].

Apart from knowing the function of a protein it is also important to know its three-dimensional structure and the subcellular location. *In vivo* studies have shown that proteins localize themselves to their compartments very specifically. Hence, knowledge about the subcellular location of proteins can help understand their

function [74]. Experimental methods such as Green Fluorescent Protein (GFP) tagging and Immunofluorescence have helped in the identification of proteins' locations.

In GFP tagging, the GFP protein is bound to the target protein and emits green fluorescence [31]. In Immunofluorescence, antibodies tagged by fluorescent chemicals bind to specific proteins. The fluorescent chemicals emit fluorescence when exposed to light and helps identify the location of a protein [43]. There are two types of immunofluorescence methods: direct and indirect. In Direct Immunofluorescence a single antibody tagged by a fluorophore binds to the target protein and the fluorophore is then detected by microscopy. On the other hand, in Indirect Immunofluorescence, two antibodies are used where only one of them is tagged to a fluorophore. First the untagged antibody binds to the target protein, followed by the tagged antibody binding to the untagged antibody. Using fluorescence microscopy the location of the target protein is studied [43].

Different microscopic techniques are used to visualize and capture images of the proteins tagged using GFP or Immunofluorescence [77]. Other methods to identify/study the subcellular location of a protein are discussed in the next section.

2.3 Protein Subcellular Location Databases

The information about the subcellular location of proteins can help explain other properties of a cell, such as its function. Several protein databases have been developed to store information about the location of proteins in various organisms. In this section we survey a few databases that store information about proteins' locations.

UniProt/SwissProt [11, 82] is a canonical protein database that has served as a reliable source for collecting, curating and storing information about proteins from a wide range of organisms. UniProt/SwissProt references about 100 other databases and stores for each protein, information about its function, its sequence, location and several other features. The March 2013 release of UniProt/SwissProt contains 90,475 proteins having experimental evidence for their subcellular location. Several other databases have been developed to store experimental information about proteins for specific organisms [26, 74, 84].

There is no central database currently that stores all available and reliable information about the subcellular location of eukaryotic proteins. Hence, in this thesis we are interested in collecting and storing all eukaryotic proteins, whose subcellular location is known, into our database JEPSLD (A Judgmental Eukaryotic Protein Subcellular Location Database). To gather data from reliable sources it is important to understand the information stored in these databases and the reliability of the data. Therefore, we surveyed several databases that store information about protein subcellular location obtained either experimentally or by using computational methods. We discuss how each database tries to provide additional information beyond what is currently available in UniProt/SwissProt.

A similar approach had been used by the researchers at the Murphy Lab, Carnegie Mellon University, who developed the Waldo framework [64] by obtaining information about proteins' location from all available resources. They used the data to assign locations to proteins that shared similar sequence and structure. In this work,

we do not make any predictions and our goal is to collect and store data about eukaryotic proteins whose subcellular locations have been identified experimentally. In addition, we are also interested in collecting data about tissue, cell line, cell type, images and PubMed IDs.

2.3.1 Databases Storing Subcellular Location Obtained Experimentally

The fundamental approach to determine the location of a protein has been through experimental methods. As explained in section 2.2, the main experimental methods used for identifying locations include GFP tagging and Immunofluorescence. These methods have helped in identifying the subcellular location of a large number of proteins in UniProtKB/Swissprot, ranging from 46% of the *S.cerevisiae* proteome to less than 10% of the proteins of *A.thaliana* and of *C.elegans* [78]. In this section we discuss two databases, Organelle DB [84] and LOCATE [26, 74], that store information about protein subcellular location identified using experimental methods.

Organelle DB

Organelle DB [84] is a database that stores experimental information about the subcellular location of eukaryotic proteins. Characteristics of the protein, its genes and Gene Ontology (GO) [78] annotation are also stored. The dataset in Organelle DB is obtained from the experiments conducted at the Kumar Lab, Life Sciences Institute, University of Michigan as well as by extracting data from other reliable sources. Protein subcellular location data was extracted from five model organism databases: the *Arabidopsis* Information Resource TAIR [66], the *Caenorhabditis elegans* database WormBase [70], the *Drosophila melanogaster* database FlyBase [27], the

Mouse Genome Database MGD [9], and the *Saccharomyces* Genome Database SGD [35]. UniProtKb/SwissProt [11, 82] was used to collect the location data for human proteins and other proteins outside of the standard model organisms. Furthermore, for additional yeast-related information not available in SGD, the curators of Organelle DB also collected information obtained from large scale and systematic studies in the budding yeast *Saccharomyces cerevisiae* at the University of Michigan.

The experimental data in Organelle DB was obtained by using epitope-tagging and Indirect Immunofluorescence. In epitope-tagging the gene encoding the target protein is tagged with an epitope (HA or V5). The subcellular location of the protein translated from the epitope tagging gene is identified using Indirect Immunofluorescence [43]. Two fluorescent micrograph images were obtained for every protein using fluorescent microcopy. One image shows the yeast cells stained with the DNA-binding dye DAPI that highlights the nucleus and mitochondria while the other image shows the same cells stained with monoclonal antibodies directed against the epitope-tagged protein. The fluorescent micrographs can help confirm if a protein is localized to the nucleus, the mitochondria or to other subcellular compartments. For example, Figure 2.1 shows the fluorescent micrograph image of the yeast gene YNL052 (or COX5A). From the staining results it is concluded that the protein cytochrome oxidase chain Va is localized to the mitochondrial inner membrane [84].

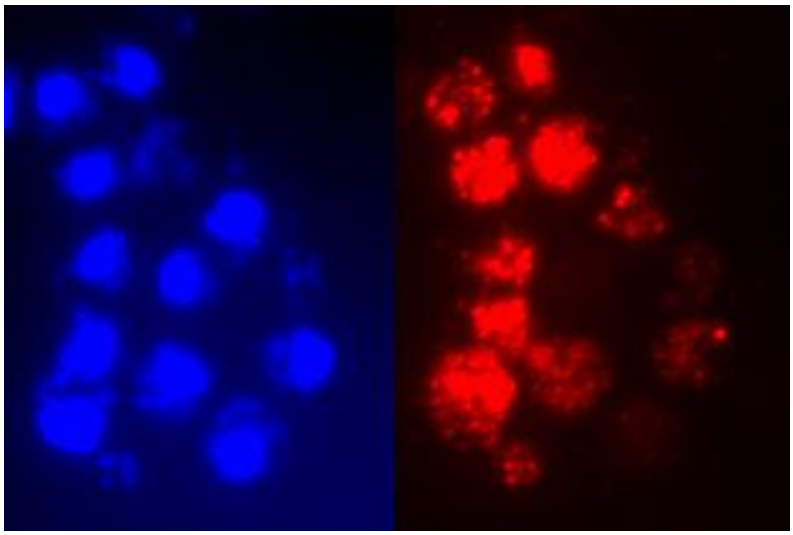

Figure 2.1: Fluorescent micrograph image of the yeast gene YNL052W (or COX5A). The image on the left shows to the yeast cells stained with the DNA-binding dye DAPI, highlighting the nucleus and mitochondria. The image on the right shows the same cells stained with monoclonal antibody directed against the epitope-tagged protein. Image source: Organelle DB [84].

Organelle DB stores 34,633 proteins from 154 eukaryotic organisms. The database includes about 1,500 fluorescent micrographs of yeast cells visualized with antibodies directed against epitope-tagged proteins (Indirect Immunofluorescence) from their own studies on protein subcellular localization in *S.cerevisiae*. Table 2.1 shows the number of proteins from each organism in Organelle DB that localize to different subcellular compartments. From Table 2.1 it is observed that Organelle DB contains a total of 34,633 proteins among which 29,425 are classified as localized to the nucleus, mitochondria, ER and membrane. The column 'Membrane proteins' lists all of the proteins that localize to the plasma membrane or to the membranes of any other organelles in eukaryotes. The column 'Protein complex' contains those proteins that form complexes such as respiratory chain complex, RNA polymerase complex,

Succinate dehydrogenase complex. 'Miscellaneous Proteins' denote proteins that localize to smaller sub-cellular structures such as the basal body, bud tip, bud neck, cytoskeleton and inclusion bodies and not to the main subcellular compartments.

Table 2.1: Protein location records in Organelle DB (Last updated September 2006) [84]

Organism	**Subcellular locations**						
	Nucleus	**Mitochondria**	**ER**	**Membrane proteins**	**Protein complex**	**Miscellaneous**	**Total**
Saccharomyces cerevisiae	2,223	989	359	889	623	435	5,518
Arabidopsis thaliana	1,168	596	70	764	558	381	3,537
Drosophila melanogaster	1,383	498	101	863	554	260	3,659
Caenorhabditis elegans	140	280	7	80	7	29	543
Mus musculus	1,443	455	156	1,225	201	346	3,826
Homo sapiens	1,691	320	227	1,911	265	438	4,852
Others (132 in total)	1,847	8,178	253	1,309	630	481	12,698
Total records	**9895**	**11316**	**1173**	**7041**	**2838**	**2370**	**34633**

LOCATE

The LOCATE [26, 74] database contains information about the subcellular location of human and mouse proteins. The data for LOCATE was obtained from the mouse and human proteome FANTOM3 Isoform Protein Sequence set (IPS8), generated by the RIKEN FANTOM Consortium [14]. The FANTOM3 Isoform Protein Sequence set comprises protein sequences encoded from specific transcript sequences. Transcripts are mRNA molecules that encode a protein. A transcription

unit (TU) is defined as the group of transcripts arising from a single genomic locus. Different transcription units (TU) encode for different forms of the same protein produced from related genes, known as protein isoforms. The curators of LOCATE found that in the mouse proteome 29,682 TUs encode for 58,128 unique protein isoforms whereas in the human proteome 26,583 TUs encode for 64,637 unique protein isoforms.

The transfection process, which introduces foreign proteins into the cell, was used to introduce the mouse proteins into the HeLa cells. The HeLa cells are the oldest and most commonly used human cell line and were initially derived from a female patient suffering from cervical cancer. The proteins expressed inside the HeLa cell line were detected using Indirect Immunofluorescence [3]. The fluorescence was observed using microscopy and the representative images were gathered and studied to identify the location of the proteins. Figure 2.2 shows the microscopic image of a mitochondrial protein in mouse, identified using Immunofluorescence (LOCATE ID: 05089_50). Experimental methods helped in identifying the subcellular location of 2068 mouse protein isoforms. It is important to note that the experiments to identify the subcellular location were conducted only in mouse proteins.

Apart from the experimentally generated data, LOCATE also stores protein location data extracted by manual curation from literature. To mine the literature the list of organelles were divided into two groups: primary and secondary cell. The primary cell consists of 15 main organelles such as nucleus, cytoplasm, mitochondria and endoplasmic reticulum while the secondary cell comprises 15 smaller

compartments such as centrosome, transport vesicles and cytoskeleton. Manual curation of the literature helped select publications that provided experimental information about the location of mouse and human proteins. Currently the database stores 9,268 proteins whose subcellular location has been extracted from the literature.

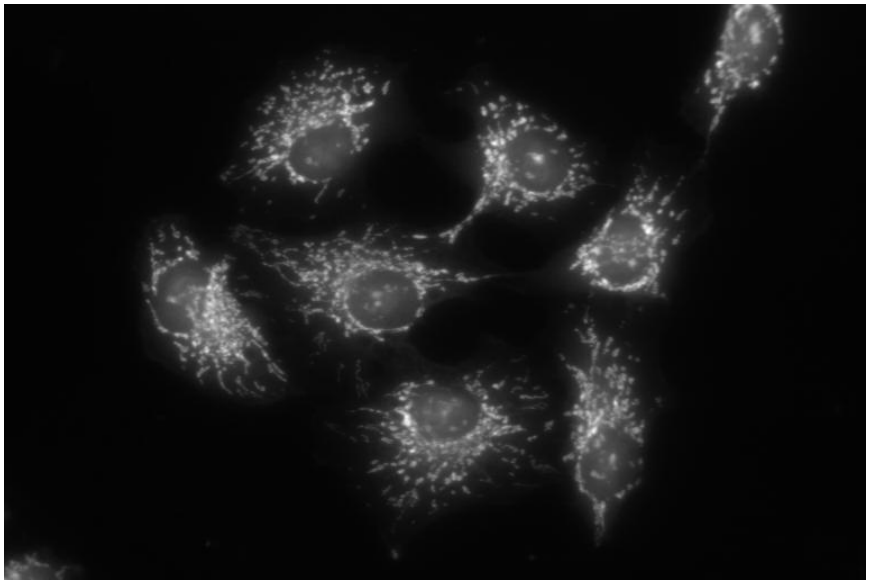

Figure 2.2: Image of a mitochondrial protein in mouse identified using Immunofluorescence (LOCATE ID: 050809_50). Image Source: LOCATE [26]

2.3.2 Databases Storing Data Obtained from Similarity Search

Although GFP tagging and Immunofluorescence are both reliable methods for identifying a protein's location, these methods are time-consuming and laborious. Experimental methods cannot be simultaneously applied to a large number of proteins and it is very difficult to experimentally determine the subcellular location of all eukaryotic proteins.

Besides using experimental methods to determine a protein's subcellular location, similarity search based methods have been used to try and assign a plausible location for unknown proteins that have at least 70% sequential identity with proteins whose subcellular location is known.

Inferring the subcellular location of a protein using similarity search is not considered to be highly reliable. A very early study performed on only 7405 proteins showed that for proteins that have greater than 50% similarity as measured by BLAST [2], their subcellular location can be assigned using similarity search with 90% accuracy. However, the accuracy of similarity search goes down for sequences that have lower similarity as measured by BLAST (tool for similarity search) [52]. Proteins having high sequential similarity do not necessarily localize to the same compartment and proteins that localize to the same compartment are often not similar. For example, homologous beta oxidation enzymes are targeted to mitochondria in humans and to the peroxisomes in yeast [72]. The similarity search approach also fails for divergent and novel proteins as they might have less than 70% sequential identity with all known proteins. Hence, similarity search is not a reliable method for assigning a plausible location to proteins whose subcellular location is unknown.

DBSubLoc

DBSubLoc [29] was one of the earliest and most comprehensive databases that collected and managed information about a protein's subcellular location. DBSubLoc contains annotations from two protein sequence databases: SwissProt/TrEMBL [11] and the Protein Information Resource [85]. Furthermore, it also contains annotations

from four model organism databases: SGD (*Saccharomyces cerevisiae*), TAIR (*Arabidopsis thaliana*), FlyBase (*Drosophila melanogaster*) and MGD (*Mus musculus*). Only proteins with known subcellular location are stored in the DBSubLoc database. Repetitive and short protein sequences having less than 20 amino acid residues are excluded. DBSubLoc also stores the GO terms for each protein, obtained from the Gene Ontology database.

DBSubLoc contains two separate datasets: full dataset and non-redundant dataset. The full dataset comprises all proteins that were extracted from UniProtKB/SwissProt and four model organism databases. The non-redundant dataset is a subset of the full dataset in DBSubLoc and it consists of only those proteins that share less than 60% similarity among themselves. To create the non-redundant dataset all protein sequences were compared with each other using BLAST and grouped based on their sequence similarity. DBSubLoc allows users to query the database using protein name and UniProtKB/SwissProt accession number. To obtain a plausible location for a protein whose location is unknown (query protein), the similarity of the protein to all proteins in the non-redundant dataset is measured by BLAST. The query protein is assigned the location of the protein in the non-redundant dataset that attains the maximum similarity score.

Table 2.2 shows that DBSubLoc stores the subcellular location of 18,143 Eukaryotic non-redundant proteins. The dataset consists of proteins that localize to seven major subcellular compartments: nucleus, cytoplasm, membrane (transmembrane as well as membrane of all other organelles), extracellular,

mitochondrion, chloroplast and the ribosome. The category “Other” denotes proteins that lack strong experimental evidence and may localize to other smaller compartments inside the cell.

Table 2.2: Statistics about the non-redundant (NR) dataset in DBSubLoc database [29]

NR data set	Fungi	Plants	Animals	Viruses	Archaea	Total
Total	4,201	1,367	12,575	1,182	1,231	18,143
Nucleus	802	181	2,066	73	83	3,049
Cytoplasm	327	25	629	0	186	1,203
Membrane	1,247	304	4,517	363	412	6,223
Extracellular	61	76	1,096	10	0	1,281
Mitochondrion	291	91	581	0	0	963
Chloroplast	0	331	0	0	0	331
Ribosome	116	78	182	0	287	1,305
Other	1,357	281	3,504	736	263	8,691

2.3.3 Databases Storing Data Predicted Using Machine Learning Methods

Since both experimental and similarity based approaches are unable to provide information about the subcellular location of all eukaryotic proteins, computational methods have been developed to predict the location of proteins based on certain features. Unique characteristics of the protein, such as its N-terminal amino acid sequence, sorting signal, amino acid composition and Gene Ontology (GO) terms are used to build the features for the computational methods. The N-terminal amino acid sequence refers to the small amino acid sequence at the start of the protein that helps in determining whether the protein localizes to the nucleus, cytoplasm and secretory

pathway or to other organelles in the cell [36]. Sorting signals are short sequences of amino acids that guide a protein to localize to specific compartments. For example, the signal peptide is found in proteins that localize to the secretory pathway. Similarly, the amino acid composition provides essential information about a protein's location by estimating the relative frequency of each amino acid in the protein's sequence [36, 15].

Techniques such as support vector machines, hidden Markov models, artificial neural networks, k-nearest neighbor and Bayesian networks have been used to build classifiers for predicting a protein's location [42]. The prerequisite for these machine learning methods is that they should have a well defined training set that covers almost all possible locations and contains sequences sharing very low sequence similarity [72].

eSLDB

The eSLDB [4] stores the subcellular location of eukaryotic proteins from five different proteomes: *H. sapiens*, *M. musculus*, *S. cerevisiae*, *C. elegans* and *A. thaliana*. Figure 2.3 describes the pipeline used by eSLDB where membrane proteins are discriminated with Spep [25] and ENSEMBLE [47]. Spep is a neural network based classifier that predicts the presence of a signal peptide, while ENSEMBLE is a classifier that predicts the topology of all-alpha transmembrane proteins based on Hidden Markov Models (HMMs) and Neural Networks. Once the Spep classifier identifies a signal peptide, the signal peptide is cleaved from the sequence before identifying the presence and the location of the transmembrane helices. Furthermore,

when no transmembrane helix is found, the uncleaved sequence is analyzed using BaCelLo [58], a classifier for eukaryotic subcellular location prediction.

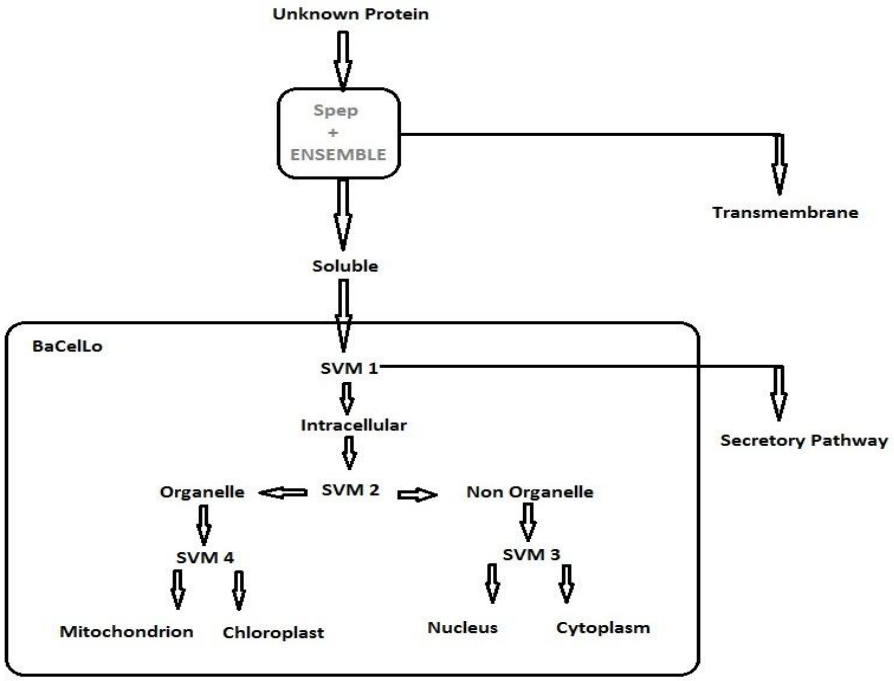

Figure 2.3: Flow chart of the prediction pipeline adopted in eSLDB. Spep and ENSEMBLE help predict the presence of a signal peptide and the topology of all-alpha transmembrane proteins respectively. The predictor BaCelLo contains four SVMs organized into a decision tree structure.

BaCelLo classifies data using four support vector machines (SVMs). The first SVM classifies proteins as intracellular or extracellular; the second SVM separates them as organelle vs non-organelle. The third SVM classifies the organelle proteins into mitochondrion and chloroplast, while the fourth SVM further classifies the non-organelle proteins into nucleus and cytoplasm. The BaCelLo classifier predicts four locations in Metazoa and Fungi (cytoplasm, extracellular, mitochondria and nucleus)

and five subcellular locations in Viridiplantae (cytoplasm, extracellular, mitochondria, nucleus and chloroplast). Table 2.3 shows the number of sequences from the five model organisms whose location was predicted by eSLDB based on machine learning methods.

Table 2.3: Number of sequences predicted to be localized to six subcellular locations, by eSLDB [59]

Subcellular location	***H.sapiens***	***M.musculus***	***C.elegans***	***S.cereisiae***	***A.thaliana***
Transmembrane	10,229	7,750	6,593	1,657	8,079
Secretory	7,816	4,971	5,172	348	3,001
Nucleus	12,358	6,820	4,733	1,717	7,649
Cytoplasm	14,720	9,356	6,280	1,710	6,033
Mitochondrion	3,630	2,326	1,454	1,112	963
Chloroplast	-	-	-	-	4,875

Computational methods can predict the subcellular location for a large number of proteins, but it is difficult to achieve high precision and recall values. The reliability of predicted location remains much lower than that of locations identified through experimental based methods. Hence, databases storing experimental information provide the most reliable dataset. We are interested in extracting all available information about proteins from reliable sources. Such databases will be discussed in detail in the next chapter.

2.4 Proteins Located in Multiple Subcellular Compartments

Proteins may simultaneously locate to or move between different subcellular compartments. Even though the same protein may be found in multiple compartments it may perform a different function in each location. For example, the Human Protein Reference Database [60] stores evidence about the protein NAD(P)H dehydrogenase, quinone 1 (HPRD ID: 00518) in humans whose primary location is the mitochondrion, while alternative locations include the microsome, plasma membrane, cytoplasm and nucleus. Proteins that localize to multiple compartments usually have more biological functions than those localizing to a single location [68].

There are very few databases that store information about specific eukaryotic organisms that localize to multiple subcellular compartments. Databases such as the Human Protein Atlas [81] and the Human Protein Reference Database [60] store information about human proteins, LOCATE [26, 74] stores information about both human and mouse proteins, whereas SUBA3 [77] stores the location of Arabidopsis proteins. As we aim to gather all available and reliable information about proteins that localize to single as well as multiple compartments, we select only those databases that store such information. In the next chapter, we discuss the selection criterion used to select the data sources for our database, the JEPSLD.

Chapter 3

DATA SOURCES

In this chapter, we begin by briefly discussing databases that store information about the sequence, structure as well as location of proteins. We then survey protein databases that specifically store information about subcellular location. Most of the data stored in these databases is either gathered experimentally or manually curated from the literature. Some of these databases also contain information about the location of proteins predicted using computational methods (similarity search or machine learning methods).

To create our database, we are interested in those databases that provide either experimental or manually curated literature-based information. Experimental methods such as GFP tagging and Immunofluorescence provide highly reliable data. Similarly, manual curation of the literature by biologists obtained reliable information about location of proteins. Therefore, we designed a selection criterion to decide which repositories should be considered for obtaining reliable information about the location of proteins.

Databases that are up-to-date and store information about eukaryotic proteins are considered. Furthermore, those databases that provide information about proteins that localize to multiple subcellular compartments are given higher preference, since

such data is scarce. It is important to note the subcellular location of proteins under different conditions such as cell line, tissue, disease and abnormalities since based on these conditions some proteins may behave differently while others may play the same role under all conditions. Hence information about tissue name, cell type, cell line as well as images representing the locations of proteins are all valuable and can help distinguish proteins' role under different conditions. In this chapter we discuss the sources of information used to gather reliable information about the location of proteins for our database.

3.1 Databases Storing Information about Proteins' Location

As described in Section 2.3, there are several public databases that store specific information about proteins such as their sequence, structure or subcellular location. In particular UniProtKB/SwissProt contains non-redundant, experimental or manually curated proteins from various organisms, giving detailed information about their sequence, function, gene name, tissue and PubMed references, when available. The database includes information from other protein databases maintained by SIB (Swiss Institute of Bioinformatics), PIR (Protein Information Resource) [6] and EBI (European Bioinformatics Institute). UniProtKB/SwissProt is maintained by the UniProt Consortium, which also maintains the UniProtKB/TrEMBL [11], a database containing proteins that were translated computationally from the coding sequences extracted from EMBL/GenBank/DDBJ nucleotide sequence databases.

UniProt/SwissProt is the central database for protein sequence information and contains 539,616 protein sequence entries from Prokaryotes and Eukaryotes, as of

March 2013. On the other hand, the Protein Data Bank (PDB) [7] is the largest repository for protein structural information and stores the three-dimensional structures of 89,003 Prokaryotic and Eukaryotic proteins (as of March 2013). In this thesis, we are interested only in repositories that provide information about the subcellular location of eukaryotic proteins. Table 3.1 lists databases that store information about eukaryotic proteins that localize to one or more subcellular compartments. The table briefly describes the data stored in the databases and the methods used to identify or predict the information they provide.

For instance, the LOCtarget database [54], developed by Nair and Rost in 2004, stores information about the location of proteins obtained either from literature search using specific keywords or predicted using similarity search and neural network algorithms. DBSubLoc [29] stores all proteins from SwissProt/PIR that contain location information with experimental evidence. The LOCATE database [26, 74] contains experimental information about mouse and human proteins, identified using immunofluorescence. Several other databases such as eSLDB [59] and LocDB [65] store information about location of proteins that were either identified by using experimental methods, manually curated from the literature, or predicted using machine-learning methods.

Table 3.1: Databases storing location of eukaryotic proteins

Database	Authors & Year of Publication	Description	Type of Cell
LOC3D [53]	Nair and Rost (2003)	8,700 eukaryotic proteins taken from PDB, whose locations were either retrieved from literature computationally using specific keywords or predicted using similarity search and machine learning methods (neural networks and support vector machines).	Eukaryotic
LOCtarget [54]	Nair and Rost (2004)	4,691 eukaryotic and 45,076 prokaryotic proteins taken from TargetDB [65], which is the central database for structural genomics targets, whose locations were retrieved from TargetDB or predicted using LOCnet [6]. LOCnet predicts the subcellular location of proteins from their sequence using neural networks.	Eukaryotic and Prokaryotic
DBSubLoc [29]	Guo et al (2004)	38,275 eukaryotic and 19664 prokaryotic proteins taken from SwissProt, PIR or the model organism databases, whose locations have been experimentally verified.	Eukaryotic and Prokaryotic
AMPDB [34]	Heazlewood and Millar (2005)	4, 416 Arabidopsis proteins obtained from The Institute of Genomic Research (TIGR), whose locations were identified experimentally or predicted computationally as mitochondrial.	Eukaryotic

Database	Authors & Year of Publication	Description	Type of Cell
Organelle DB [83]	Wiwattwatana and Kumar (2005)	25,000 eukaryotic proteins whose locations were either identified experimentally by the researchers at University of Michigan or retrieved from other model organism databases.	Eukaryotic
PA-GOSUB [45]	Lu et al (2005)	107,684 proteins taken from SwissProt and 10 model organism databases, whose locations were either retrieved from the reference databases or predicted using machine learning methods.	Eukaryotic and Prokaryotic
PDB_TM [79]	Tusnady et al (2005)	1,827 proteins taken from the Protein Data Bank (PDB), localized to the transmembrane. The TMDET algorithm [80] developed by the researchers at the Institute of Enzymology, Hungary helps distinguish between transmembrane and globular proteins, using certain structural features of the proteins.	Eukaryotic and Prokaryotic
LOCATE [26, 74]	Sprenger et al (2007); Fink et al (2006)	2,068 mouse proteins, whose locations were identified experimentally by the researchers at The University of Queensland, Australia. Images relevant to the proteins were also stored in the database. 9,268 proteins from mouse and humans, whose locations were manually curated from the literature.	Eukaryotic

Database	Authors & Year of Publication	Description	Type of Cell
FTFLP Database [44]	Li et al (2006)	1,300 Arabidopsis proteins from The Arabidopsis Information Resource (TAIR), whose locations were, identified experimentally using GFP tagging by the researchers at the Carnegie Institute, Stanford, California.	Eukaryotic
SUBA [77]	Heazlewood et al (2007)	9,024 Arabidopsis proteins from TAIR, whose locations were extracted from the literature by manual curation. Mass spectrometry and GFP tagging were the methods used to identify the location for most of the proteins in the literature.	Eukaryotic
eSLDB [59]	Pierleoni et al (2007)	21,373 eukaryotic proteins taken from five model organism databases, whose subcellular locations were obtained from the UniProtKB/SwissProt database. 78,835 eukaryotic proteins whose subcellular locations were assigned using similarity search. Furthermore, machine learning methods were used to predict the location for all eukaryotic proteins whose location was unknown.	Eukaryotic
DBMLoc [86]	Zhang et al (2008)	10,470 proteins from primary protein databases and subcellular location databases, that localize to multiple subcellular locations were extracted by manual curation of the literature or retrieved	Eukaryotic
LocDB [65]	Rastogi and Rost (2011)	6,262 Arabidopsis proteins and 13,342 human proteins, whose subcellular locations were retrieved from other databases such as UniprotKB/SwissProt, LOCATE and Suba2.	Eukaryotic

3.2 Selection Criteria

Among all the databases shown in Table 3.1, we focus on those that are up-to-date and provide reliable information about the location of eukaryotic proteins. In chapter 2, we surveyed methods used to identify or predict the location of proteins, and noted that experimental data is considered the most reliable evidence for location. Hence, we focus on repositories that store location information based on experiments, and record the number of subcellular compartments covered by them. Some of these databases provide additional information such as images, tissue, cell type or cell line information. Moreover, we focus on databases that store reliable information about proteins that localize to multiple subcellular compartments. In most databases that provide information about multi-localized proteins, the locations are classified as primary or main location vs secondary or alternate (other) location. The primary location represents the predominant subcellular location of that protein whose experimental evidence is strong, whereas the secondary location is assigned when the experimental evidence does not provide a distinct location and hence the reliability is moderate or weak.

It is important that the databases are publicly available and the data is freely downloadable. Based on all these factors we decided to gather the data from five protein databases: SUBA3 [77], the Human Protein Reference Database (HPRD) [60], the Human Protein Atlas (HPA) [81], LOCATE [26, 74] and UniprotKB/SwissProt [11, 82], as further discussed below.

3.3 Data Sources for JEPSLD

In this section we discuss the five data sources that satisfied the selection criteria. Data from these databases was collected and stored in our database. The information we store consists of protein name, gene name, subcellular location and when available, other relevant information such as tissue, cell line, cell type and images. We next discuss each database, its data sources, the information provided by the database, and any unique information stored in it.

3.3.1 SUBA3: A Database for Integrating Experimentation and Prediction to define the SUBcellular Location of Proteins in Arabidopsis thaliana

SUBA3 [77] provides detailed information about Arabidopsis proteins. The information about Arabidopsis proteins was retrieved from The Arabidopsis Information Resource (TAIR) [75] and all information relevant to each protein's subcellular location was extracted. SUBA3 stores data about Arabidopsis proteins localized to 11 subcellular locations and contains 9,024 proteins whose locations were determined experimentally.

SUBA3 also stores experimental data about each protein's location, manually curated from literature sources by curators at Centre of Excellence in Computational Systems Biology, The University of Western Australia. To obtain the relevant publications, the PubMed database was searched using keywords such as 'MS', '*Arabidopsis*', 'GFP', 'fluorescent protein', 'CFP', 'YFP', or 'RFP'. The full-text articles were then downloaded and read by the curators to identify relevant information about the location of Arabidopsis proteins. A total of 122 publications

were retrieved providing information about the location of 7,685 Arabidopsis proteins identified using mass spectrometry. Similarly 1,074 articles were retrieved providing information about the location of 2,477 proteins determined using GFP tagging.

In addition to the experimental information obtained from the literature, SUBA3 stores the location for Arabidopsis proteins predicted by 22 eukaryotic protein location predictors: AdaBoost [56], ATP [50], BaCelLo [58], ChloroP 1.1 [25], EpiLoc [12], iPSORT [4], MitoPred [28], MitoProt [17], MultiLoc2 [10], Nucleo [33], PCLR [69], Plant-mPLoc [16], PProwler 1.2 [32], Predotar v1.03 [73], PredSL [57], PTS1 [55], SLPFA [76], SLP-Local [48], SubLoc [39], TargetP 1.1 [22], WoLF PSORT [38] and YLoc [13]. Furthermore, SUBA3 also stores a consensus location that was generated using SUBAcon, a Naïve Bayes classifier that determines each protein's location from the experimental and predicted data available. SUBA3 is up-to-date and the downloadable data file contains 2,477 proteins. For all proteins stored, SUBA3, also stores their gene name, protein description, SwissProt ID, SwissProt Location, TAIR ID, TAIR Location and PubMed References, when available.

3.3.2 The Human Protein Reference Database (HPRD)

The Human Protein Reference Database [60] currently stores information about 30,047 human proteins and provides information about the subcellular location for 22,490 of them. The data was obtained through manual literature curation by biologists at the Institute of Bioinformatics, Bangalore and Department of Biological Chemistry and Depart of Pathology and Oncology, Johns Hopkins University. In addition to the protein location, the database also provides information about the tissue

and cell line (Karpas, K-562, HeLa, Ramos and Jurkat) for a few protein entries. Moreover, the database also contains information about multi localized proteins. The database references other databases such as OMIM [30], RefSeq [62], Entrez [46] and SwissProt [11].

A protein distributed annotation system called the Human Proteinpedia [40] was also integrated into HPRD. The Human Proteinpedia collects and stores experimental data relevant to human proteins submitted by scientists worldwide. Currently the Human Proteinpedia contains over 15,000 protein entries submitted by 249 research laboratories. Experimental methods such as mass spectrometry, fluorescence microscopy and protein-microarrays were used to study the proteins. The Human Proteinpedia stores the subcellular location of 2,900 proteins. The data in the Human Proteinpedia is integrated into the HPRD database and is publicly available. The HPRD database is maintained and is constantly growing and new protein entries that are experimentally identified are added to the database on a regular basis through the Human Proteinpedia. The downloadable file contains, for each protein entry, information consisting of the protein name, gene name, gene symbol, SwissProt ID, OMIM ID, RefSeq ID and the organism name. For proteins having multiple locations the database stores the primary location as a main location and the alternate locations as other location.

3.3.3 The Human Protein Atlas (HPA)

The Human Protein Atlas [81] was constructed to store the expression patterns of human proteins in normal and cancer tissues. HPA stores data obtained from experimental analysis of 48 normal tissues and 20 cancer tissues in humans. Antibodies specific to the target proteins were used to perform a functional study of the human proteome. The Human Protein Atlas stores information about human tissue, cell type and cell line as well as the location of human proteins. The database also provides images obtained from the expression studies using immunofluorescence. For example, Figure 3.1 shows the image of the protein Mitogen-activated kinase kinase kinase 4 (MAP3K4), using Immunofluorescence, indicating that the protein localizes to the nucleus but not to the cytoplasm or nucleoli.

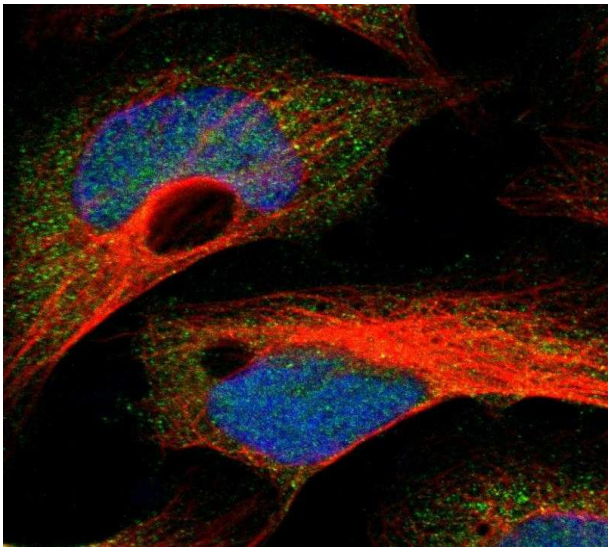

Figure 3.1: Immunofluorescent staining of the protein mitogen-activated protein kinase kinase kinase 4 (MAP3K4) in human cell line U-2 OS where the red color represents microtubules, blue color represents nucleus and the green dots represent the antibodies. We infer from the image that the protein localizes to the nucleus but not to the nucleoli and the cytoplasm. Image Source: HPA [81]

PA provides information about the subcellular location of 11,354 human proteins obtained through confocal microscopy, a technique where the contrast and optical resolution of a microscopic image is increased by using point illumination and capturing the image on a focal plane [81]. These images were analyzed by certified pathologists and the locations of proteins were identified.

Proteins that localize to multiple locations are also stored in the HPA. Originally the HPA contained data from three cell lines; U-2 OS, A-431 and U-251 MG. In the 2012 update, eight more cell lines were included: A-549, CACO-2, HEK 293, HeLa, Hep-G2, MCF-7, PC-3, and RT-4. The downloadable files provide gene name, Ensembl ID, SwissProt ID, main location and other location.

3.3.4 LOCATE: A Mouse and Mammalian Protein Subcellular Localization Database

The LOCATE database [26, 74] was created in 2006 to store information about the subcellular location of mouse proteins. The researchers at the Institute of Molecular Bioscience, the University of Queensland obtained the data by either conducting experiments or through manual curation of the literature. Immunofluorescence was used to experimentally determine the location of 417 proteins and the images obtained are stored in the database. Manual curation of over 1,700 peer-reviewed publications that provided experimental descriptions, helped in identifying the location of 1,752 proteins. Furthermore, information about the location of mouse proteins was collected from other databases such as the Mouse Genome Informatics [63], LIFEdb [5], UniProt [82] and RefSeq [13].

The 2007 update of LOCATE provided information about the subcellular location of both human and mouse proteins. Currently, LOCATE stores information about the subcellular location of 2,068 proteins obtained using experimental methods and 9,268 proteins obtained by manual curation of literature. The database contains information about protein function, organism, class, location, source name and PubMed references. The LOCATE database also stores information about proteins that localize to multiple compartments and classifies the subcellular locations into either primary location or other location.

3.3.5 UniProtKB/SwissProt

As discussed in section 3.1, the UniProtKB/SwissProt [11, 82] is the most comprehensive, publicly available protein database, maintained by the UniProt consortium. The database references about 100 other databases and includes information about proteins from various organisms. Most of the data stored in the UniProtKB/SwissProt database is validated experimentally or curated from the literature. The database is updated regularly every two weeks and minimizes redundancy by combining separate entries for the same gene product into a single protein entry. The database provides information about protein name, gene name, organism, sequence, function, subcellular location and references for each protein entry, when available.

For proteins that have not been studied experimentally, three non-experimental qualifiers are used. The qualifiers are “Potential”, “Probable” and “By Similarity”. The “Potential” qualifier indicates that the associated information was computationally

predicted. The “Probable” qualifier indicates that the information is based on stronger evidence than “Potential” and might be supported by indirect experimental information. The “By Similarity” qualifier, associated with information of a particular protein indicates that experimental evidence supports information associated with proteins homologous to that protein, and the information is thus transferred to the protein itself. Proteins in UniprotKB/SwissProt, whose locations are identified non-experimentally (Potential, Probable or by Similarity), are not included in our database. Overall, the UniProtKB/SwissProt has a wide coverage and includes data from most other protein databases, but still there are proteins that have either not been stored or do not have experimental information about their subcellular location in UniProtKB/SwissProt.

In the next chapter we discuss in detail how the data from these five databases was processed and stored in our database, JEPSLD. We provide details about processing the data, creating the database, and developing a user interface.

Chapter 4

IMPLEMENTATION OF THE JEPSLD DATABASE

In this chapter we discuss the steps involved in implementing the Judgmental Eukaryotic Protein Subcellular Location Database. We begin by describing the method used to process the data obtained from the five selected databases discussed in Section 3.3. Once all the files were downloaded and converted to the same file format, we examined each file and extracted all the relevant information. The Python programming language was used to process the data, populate the database, and develop the user interface.

We developed a relational database schema and created the database JEPSLD using MySQL [51]. The database schema has six tables: *protein_info*, *cell_line*, *cell_type*, *location*, *tissue* and *publication*. We created the tables using the Structured Query Language (SQL). Furthermore, for proteins that have experimental evidence supporting the information about their subcellular location, we extracted information including their location, gene name, protein name and other fields, as described in Section 4.1, from the five data sources and stored it in the MySQL database using a Python script.

We used the web development framework, Django [18] to create a user interface that enables users to query the JEPSLD database using gene name, protein

name, protein accession IDs such as UniProtKB/SwissProt ID or gene identifiers such as the Entrez gene ID and the Ensembl gene ID.

In the next section we discuss the methods used to process the data obtained from the five data sources. The following section discusses the database construction and gives details about the number of proteins in JEPSLD. Finally, we discuss the user interface that allows users to query the database and extract relevant information.

4.1 Obtaining and Processing the Data

We obtain information about the subcellular location of eukaryotic proteins from the five data sources discussed in Section 3.3. They are: SUBA3 [77], the Human Protein Reference Database (HPRD) [60], the Human Protein Atlas (HPA) [8], LOCATE [26, 74] and UniProtKB/SwissProt [11, 82]. The data from these sources are available for download in either XML or CSV file format.

The Extensible Markup Language (XML) uses a format that is both human-readable and machine-readable and is designed to structure, store and transport information [23]. On the other hand, the Comma-Separated Values (CSV) file format, stores tabular data in plain-text format, where each record in the file consists of fields that are separated by some designated character or a string, most commonly a comma or a tab [71].

The data from HPRD, HPA, LOCATE and UniProtKB/SwissProt were available for download in XML format, whereas the SUBA3 data was available in CSV format. The LOCATE database consists of two separate XML files, one for mouse and one for human proteins. Since most of the files were in XML format, we

decided to convert the CSV file for SUBA3 into XML in order to have a standard file format while inserting the data into our database.

We used the Python (version 2.7.3) programming language to process the XML files and extract all relevant information. Python is a freely available, interpreted object-oriented programming language used for scripting and developing applications. We used the *xml.etree.ElementTree* module [49] in Python, which enables reading an entire XML document and creating objects to represent its elements, attributes and content. The Python script traverses the entire XML file and searches for tags and attributes that meet a specific criterion. All relevant information about gene name, protein name, protein accession number, gene reference IDs, location, PubMed IDs, tissue, cell line and cell type are extracted from these nodes.

Table 4.1 summarizes the information extracted from the XML files of the five selected data sources. Certain tags, such as tissue, have attributes such as *normal* and *cancer* that indicate whether the tissue information has been obtained from normal or from cancer patients. In the UniProtKB/SwissProt file, we extract information only from proteins having *Eukaryota* as their taxon. Moreover, we do not consider proteins whose subcellular locations are tagged with non-experimental identifiers (*by similarity*, *probable*, *potential*).

Some of the protein entries in the XML files of the source databases do not contain some of the information such as gene name or UniProt/SwissProtID or Ensembl gene ID. To obtain all the missing information about gene name, Ensembl gene ID, Ensembl protein ID we used the Synergizer service [8] to convert the

identifiers from one naming scheme to another, and the BioMart Central Portal [41] to obtain information about the gene name using the Ensembl gene or proteins IDs.

Table 4.1: Information extracted from XML files of the five selected data sources

Database Name	**Information extracted**
SUBA3	gene name, protein name, UniProtKB/SwissProt ID, main location , PubMed ID, organism
HPA	gene name, Ensembl gene ID, UniProt/SwissProt ID, tissue name, cell type, main location, other location, cell line, organism
HPRD	gene name, protein name, OMIM ID, RefSeq ID, main location, other location, Entrez gene, organism
LOCATE	UniProt/SwissProt ID, RefSeq ID, Ensembl gene ID, Ensembl protein ID, Entrez gene, cell line, main location, PubMed ID, organism, image
UniProtKB/SwissProt	gene name, protein name, organism, main location, UniProtKB/SwissProt ID, RefSeq ID, Ensembl gene ID

4.2 Database Construction and Content

This section describes the steps involved in the construction of the JEPSLD database. Figure 4.1 shows the relational database schema of the database and defines the tables, attributes and relationships. The database includes six tables: *protein_info*, *location*, *cell_line*, *cell_type*, *publication* and *tissue*. Each of the attributes JEPSLD_ID, location_ID, cell_line_ID, cell_type_ID, publication_ID and tissue_ID serves as the primary key of their respective tables. The JEPSLD_ID is automatically incremented for each protein entry obtained from the data sources and inserted into the

JEPSLD database. The other primary keys are incremented whenever any information is added to their respective tables. The *protein_info* table stores for each protein the protein name, gene name, resource database, UniProt/SwissProt ID and gene reference IDs such as RefSeq ID and Ensembl gene ID. The tables, *cell_line*, *cell_type*, *tissue* and *location* store information about the cell line, cell type, tissue and the subcellular location of each protein, respectively. The table *publication* stores for each protein, PubMed IDs of all PubMed abstracts referencing that protein. The attributes *gene name*, *resource database* and *ensemble_protein_ID* from the *protein_info* table serve as the foreign key for the *cell_line*, *cell_type*, *tissue*, *location* and *publication* tables.

MySQL is a freely available, relational database, and is one of the most popular databases for use in web applications [51]. The MySQL server (version 5.1.50) discussed in this work has been installed on a Linux (Ubuntu) platform. The host name for our MySQL server is *db.eecis.udel.edu* and is maintained by the Computer and Information Sciences Dept, University of Delaware. We created JEPSLD as a MySQL database, and used SQL queries to create all the tables as discussed in Figure 4.1. Once the tables were created, a Python script was used to extract relevant information from the XML files and store it in the database.

Figure 4.1: Relational Database Schema for the Judgmental Eukaryotic Protein Subcellular Location Database (JEPSLD)

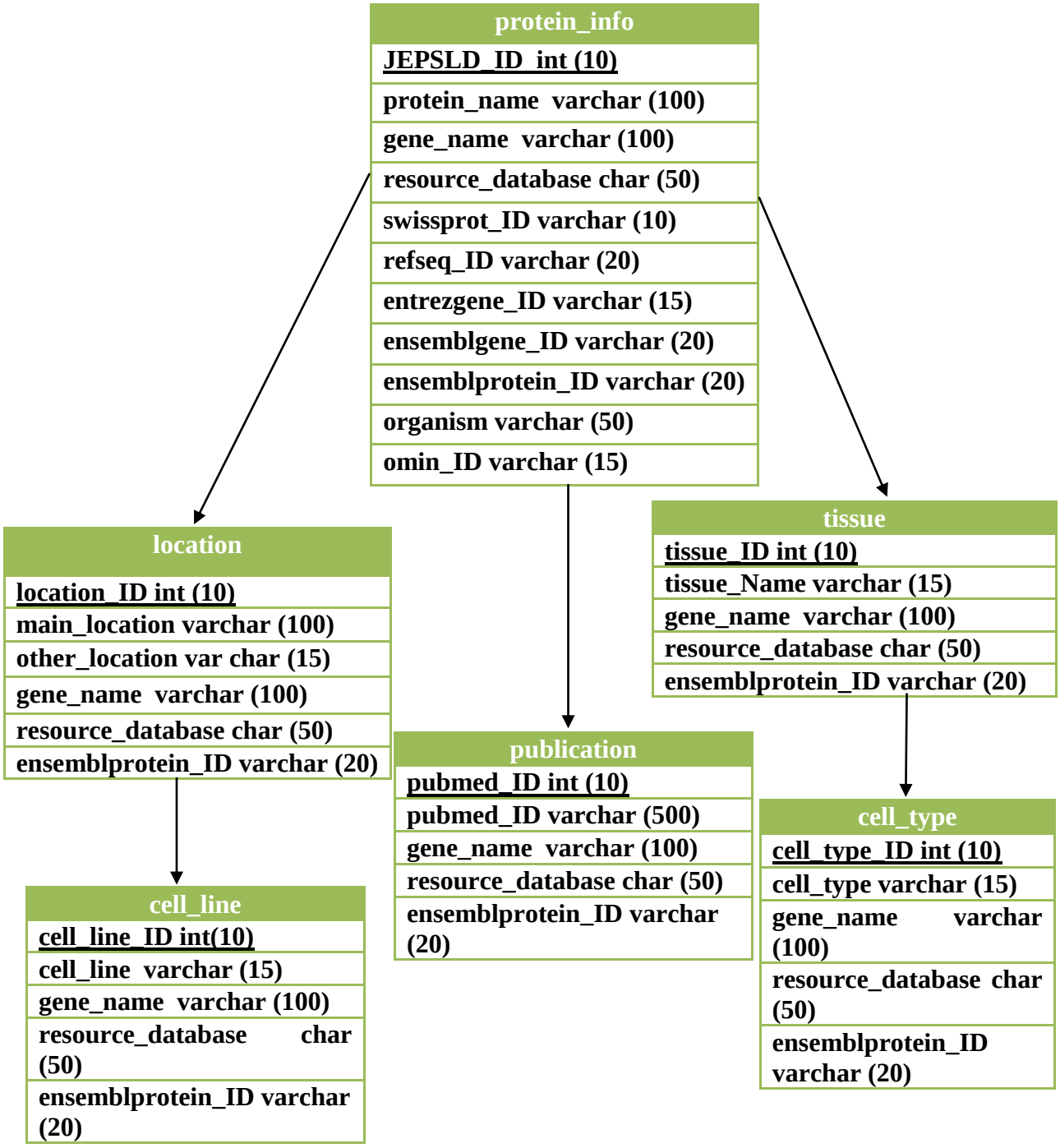

Table 4.2 lists the total number of eukaryotic proteins obtained from the five selected data sources. It also provides the number of proteins that have information about their subcellular location identified experimentally, the number of proteins that localize to multiple compartments, and the number of proteins with cell line information from each database.

Table 4.2: The total number of eukaryotic proteins, number of proteins having experimental information about their subcellular location, the number of proteins that localize to multiple locations and the number of proteins with cell line information in each of the five selected databases.

Database	**Total # of eukaryotic proteins**	**Eukaryotic proteins with location**	**Eukaryotic proteins with multiple locations**	**Eukaryotic proteins with cell line information**
SUBA3	2,477	2,477	1,138	-
HPRD	30,046	21,466	255	-
HPA	20,115	12,819	4,087	9,911
LOCATE	122,765	32,325	11,852	32,325
UniProtKB/ SwissProt	175,158	90,475	32,772	-
JEPSLD	113,525	113,525	50,104	42,236

4.3 User Interface

Our next step was to develop a user interface that would accept input queries, hand them back to the back-end relational database and display the results. The user interface was developed using Django (Official release version 1.5.1) [18, 37], a freely

available web application framework that is implemented in Python. The Django web application framework provides libraries for database access and HTML templates.

Figure 4.2 shows the architecture of the Django user interface to the JEPSLD database. The user may query the JEPSLD database by specifying information such as gene name, protein name, protein accession numbers or gene identifiers, thus sending a request to the web server hosting the site. The Django web server accepts the request and uses the Django framework to access the JEPSLD database and obtain the desired data. To perform these functions, the Django framework has three components: *Model, View and Controller.* The *Model* is the data access layer that contains information about the data, relationships between the data and methods to access and validate it. This layer is responsible for querying the database and retrieving the desired data. The *View* takes the information from the Model layer and passes it as response to the user's browser for display in HTML format. The *Controller* receives and manages inputs to update the *Model* layer and also updates the elements for the *View* layer, as necessary. Once the web server receives the results from the database, we further process the data using a Python script to remove any redundancies, and information relevant to unique protein entries are sent as response. The web server is continuously running to accept users' queries and respond to them.

The user interface, as shown in Figure 4.3, allows users to query the JEPSLD database. In addition to protein accession numbers and gene identifiers, users may also search for proteins using their subcellular location, tissue name, cell line and organism.

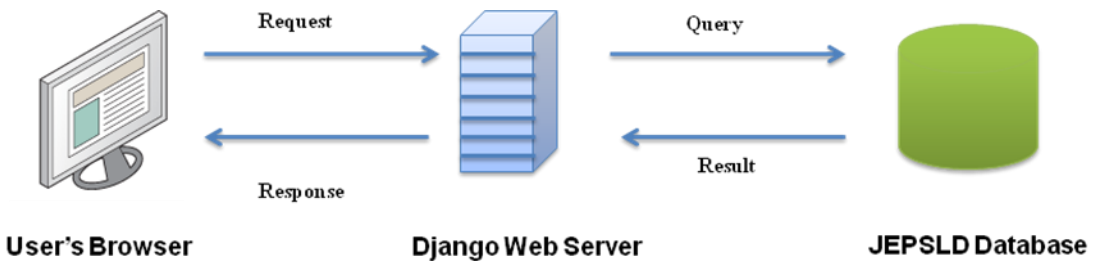

Figure 4.2: Architecture of the Django user interface and the JEPSLD database

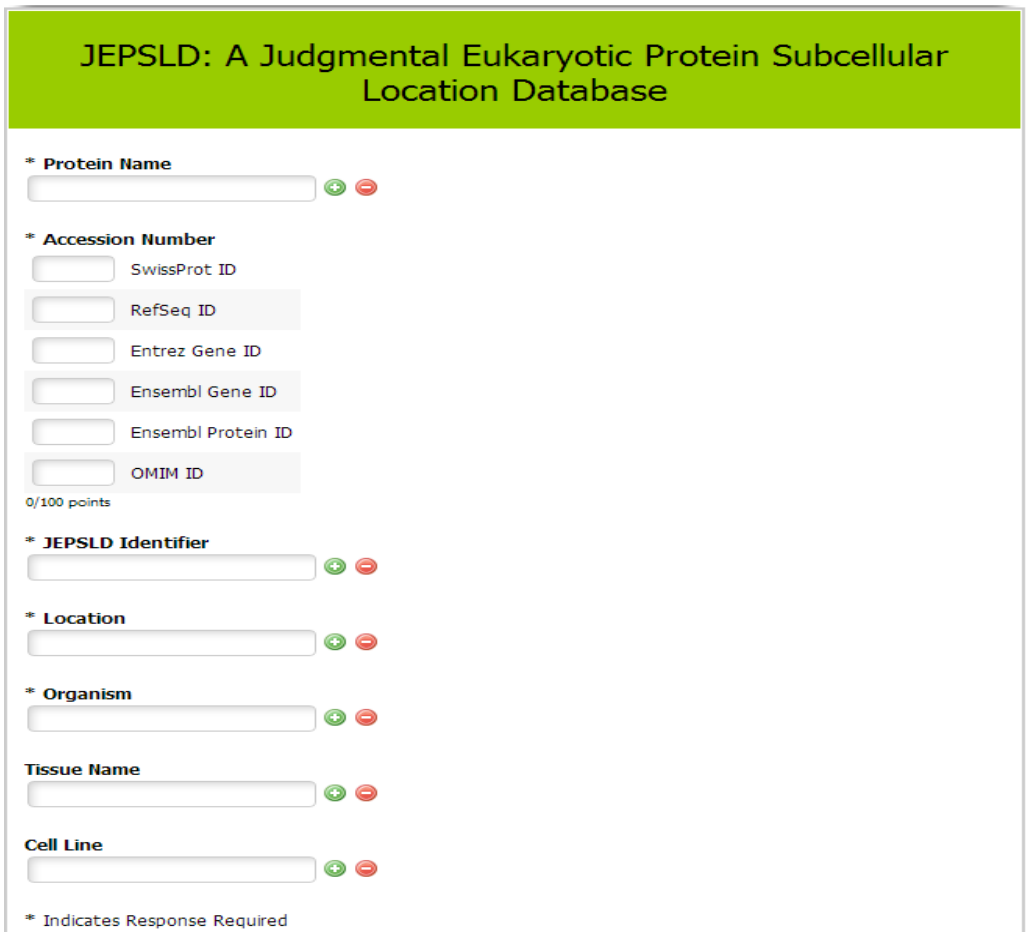

Figure 4.3: Entry retrieval and keyword search. From the JEPSLD database, the user can input protein name, accession numbers, location, organism, tissue name and cell line to query the database.

The user interface and the web server are currently at the development stage, and can only be accessed through a local machine. We intend to set up a web server at the Dept. of Computer and Information Sciences, University of Delaware that would allow JEPSLD to be publicly accessible in the near future.

In the next Chapter, we conclude the thesis and briefly discuss the main contributions and future prospects of this work.

Chapter 5

CONCLUSION

We have presented the design and development of the Judgmental Eukaryotic Protein Subcellular Location Database (JEPSLD) that stores the subcellular location of eukaryotic proteins, extracted from five selected data sources. Since there is no central database currently that stores all available information about the subcellular location of eukaryotic proteins, our objective was to create such a database. Moreover, we also aimed to collect information about proteins that localize to multiple compartments, as well as additional information including tissue, cell type, cell line and images.

We surveyed several protein databases and defined certain criteria to select five data sources that are up-to-date and store reliable information about the subcellular location of proteins. The data from these five sources was processed and only proteins, whose subcellular locations were identified experimentally, were stored in the JEPSLD database. We next provide a summary of the contributions of this work and discuss future work.

5.1 Summary of Contributions

The work presented in this thesis makes the following contributions:

- We constructed the Judgmental Eukaryotic Protein Subcellular Location Database, which is the largest eukaryotic protein subcellular location database. It stores information about 113,525 eukaryotic proteins, for which there is experimental evidence pertaining to their subcellular location.
- The JEPSLD database stores 50,104 proteins that localize to multiple compartments. This information is valuable for studying how proteins localizing to multiple locations may perform the same or different functions in each location.
- The database also provides additional information such as tissue, cell type, cell line, images and PubMed Ids, when available. This information is valuable for studying the role of proteins under different conditions.

5.2 Future Directions

The JEPSLD database is currently hosted on the server *db.eecis.udel.edu* at the Dept. of Computer and Information Sciences, University of Delaware, and can be accessed locally. Our immediate next step is to make the JEPSLD publicly available. We shall also modify the user interface to support batch queries using gene names, protein names and protein accession numbers. The user interface will also allow users to download the entire data from the JEPSLD database in XML format. We also aim

to allow users to specifically query and download proteins that localize to multiple locations.

The current database stores only information about eukaryotic proteins. A similar approach can be used to create a database for prokaryotic proteins, to store experimental information about their subcellular location.

REFERENCES

[1] Alberts, B., Bray D., Lewis J., Raff M., Roberts, K. and Watson, J. D. *Molecular Biology of the cell*. 3rd ed.; 17-23.

[2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol*, **215:** 403–410.

[3] Aturaliya, R. N., Fink, J. L., Davis, M. J., Teasdale, M. S., Hanson, K. A., Miranda, K. C., Forrest, A. R., Grimmond, S. M., Suzuki, H. et al. (2006) Subcellular localization of mammalian type II membrane proteins. *Traffic,* **7**: 613–625.

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

[5] Bannasch, D., Mehrie, A., Glatting, K. H., Pepperkok, R., Poustka, A. and Wiemann, S. (2004). LIFEdb: a database for functional genomics experiments integrating information from external sources, and serving as a sample tracking system. *Nucleic Acids Research*; **32**: D505-D508.

[6] Barker, W. C., Garavelli, J. S., Huang, H., McGarvey, P. B., Orcutt, B. C., Srinivasarao, G. Y., Xiao, C., Yeh, L. L., Ledley, R. S., Janda, J. F., Pfeiffer, F., Mewes, H. W., Tsugita, A., Wu, C. (2000). The Protein Information Resource (PIR). *Nucleic Acids Res;* **28**(1): 41-44.

[7] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*; **28**: 235-242.

[8] Berriz, G., Roth., F. (2008). The synergizer service for translating gene, protein, and other biological identifiers. *Bioinformatics*. **24**(19): 2019-2027.

[9] Blake, J. A., Eppig, J. T., Bult, C. J., Kadin, J. A., Richardson, J. E. and Group, M.G.D. (2006) The mouse genome database (MGD): updates and enhancements. *Nucleic Acids Res*., **34:** D562–D567.

[10] Blum T, Briesemeister S, Kohlbacher O. (2009). MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics.* ;**10**:274.

[11] Boeckmann, B., Bairoch, A., Apweller, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K. *et al.* (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research;* **31**(1): 365-370.

[12] Brady S, Shatkay H. (2008).EpiLoc: a (working) text-based system for predicting protein subcellular location. *Pac. Symp. Biocomput.* 604–615.

[13] Briesemeister S, Rahnenfuhrer J, Kohlbacher O. 2010. YLoc–an interpretable web server for predicting subcellular localization. *Nucleic Acids Res.*; **38**:W497–W502.

[14] Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B. et al. (2005). The transcriptional landscape of the mammalian genome. *Science*, **309**: 1559–1563.

[15] Cedano, J., Aloy, P., Perez-Pons, J. A., Querol, E. (1997). Relation between amino acid composition and cellular location of proteins. *Journal of Molecular Biology*; **266**: 594-600.

[16] Chou KC, Shen HB. 2010. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS One.*;**5**:e11335.

[17] Claros MG, Vincens P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.*; **241**:779–786.

[18] Django documentation. *https://docs.djangoproject.com/en/1.5/* Last accessed: 04/07/2013

[19] Dobson, C. M., Sali, A. and Karplus, M. (1998). Protein Folding: A perspective from Theory and Experiment. *Angew. Chem. Int. Ed*., **37:**868-893.

[20] Donnes, P., Hoglund, A. (2004). Predicting protein subcellular localization: Past, present and future. *Genomics, Proteomics and Bioinformatics*; **2**: 209-215.

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

[22] Emanuelsson O, Nielsen H, Brunak S, von Heijne G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*;**300**: 1005–1016.

[23] Extensible Markup Language (XML) 1.0; 5th ed. *http://www.w3.org/TR/2008/REC-xml-20081126/#*charsets Last accessed: 04/07/2013

[24] Eukaryotic cells. *Scitable-Nature education*. *http://www.nature.com/scitable/topicpage/eukaryotic-cells-14023963* Last accessed: 03/09/13

[25] Fariselli, P., Finocchiaro, G. and Casadio, R. (2003) SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics*, **19**: 2498–2499.

[26] Fink, J. L., Aturaliya, R. N., Davis, M. J. Zhang, F., Hanson, K., Teasdale, M. S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. and Teasdale R. D. (2006). LOCATE: a mouse protein subcellular localization database. *Nucleic Acids Research*, **34**: D213-D217.

[27] Grumbling, G. and Strelets, V. (2006) FlyBase: anatomical data, images and queries. *Nucleic Acids Res*., **34**: D484–D488.

[28] Guda C, Guda P, Fahy E, Subramaniam S. (2004). MITOPRED: a web server for the prediction of mitochondrial proteins. *Nucleic Acids Res.*;**32**:W372–W374.

[29] Guo, T., Hua, S., Ji, X. and Sun, Z. (2004). DBSubLoc: database of protein subcellular localization. *Nucleic Acids Research*, **32**: D122-D124.

[30] Hamosh, A. (2002). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*; **30**: 52-55.

[31] Hanson, M. R. and Kohler, R. H. (2001). GFP imaging: Methodology and application to investigate cellular compartmentation in plants. *J. of Experimental Botany*; **52**: 529-539.

[32] Hawkins J, Boden M. (2006). Detecting and sorting targeting peptides with neural networks and support vector machines. *J. Bioinform. Comput. Biol.* ;**4:**1–18.

[33] Hawkins J, Davis L, Boden M. (2007). Predicting nuclear localization. *J. Proteome Res.* ;**6**:1402–1409.

[34] Heazlewood, J. L. and Millar, A. H. (2005). AMPDB: the Arabidopsis Mitochondrial Protein Database. *Nucleic Acids Research*; **33**: D605-D610.

[35] Hirschman, J. E., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hong, E. L., Livstone, M. S., Nash, R.et al. (2006) Genome Snapshot: a new resource at the Saccharomyces Genome

Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome. *Nucleic Acids Res.*, **34**: D442–D445.

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

[37] Holovaty, A. and Kaplan-Moss J. The Django Book. *http://www.djangobook.com/en/2.0/license.html* Last accessed: 04/10/2013

[38] Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. 2007. WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*; **35**:W585–W587.

[39] Hua S, Sun Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics.* ;**17:**721–728.

[40] Kandaswamy, K., Keerthikumar, S., Goel, R., Mathivanan, S., Patankar, N., Shafreen, B., Renuse, S., Pawar, H., ramachandra, Y. L., Acharya, P. K. *et al.* (2008). Human Proteinpedia: a unified discovery resource for proteomics research. *Nucleic Acids Research*; **37**: D773-D781.

[41] Kasprzysk, A. (2011). BioMart: driving a paradigm change in biological data management. *Database*; bar049**.**

[42] Kotsiantis, S. B. (2007). Supervised Machine Learning: A review of Classification Techniques. *Informatica;* **31**: 249-268.

[43] Kumar, A., Agarwal, S., Heyman, J. A., Matson, S., Heidman, M., *et al.* (2002). Subcellular localization of the yeast genome. *Genes & Development;***16**: 707-719.

[44] Li, S., Ehrhardt, D. W., and Rhee, S. Y. (2006). Systematic analysis of Arabidopsis organelles and a protein localization database for facilitating fluorescent tagging of full-length Arabidopsis proteins. *Plant Physiol;* **141**(2): 527-39.

[45] Lu, P., Szafron, D., Greiner, R., Wishart, D. S., Fyshe, A., Pearcy, B., poulin, B., Eisner, R., Ngo, D. and Lamb, N. (2005). PA-GOSUB: a searchable database of model organism protein sequences with their predicted Gene Ontology molecular function and subcellular localization. *Nucleic Acids Research*; **33**: D147-D153.

[46] Maglott, D., Ostell, J., Pruitt, K. D. and Tatusova. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*; **33**: D54-D58.

[47] Martelli, P. L., Fariselli, P. and Casadio, R. (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics,* **19**: i205–i211.

[48] Matsuda S, Vert JP, Saigo H, Ueda N, Toh H, Akutsu T. 2005. A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci*.;**14**:2804–2813.

[49] Mitchell, L. M. Bioinformatics Programming Using Python. 1st ed., 2009.

[50] Mitschke J, Fuss J, Blum T, Hoglund A, Reski R, Kohlbacher O, Rensing SA. 2009. Prediction of dual protein targeting to plant organelles. *New Phytol.* ;**183**:224–235.

[51] MySQL. *www.mysql.com* Last accessed: 04/08/2013

[52] Nair R, Rost B. Sequence conserved for subcellular localization. Protein Sci. 2002; 11(12):2836–2847.

[53] Nair, R. and Rost, B. (2003). LOC3D: annotate sub-cellular localization for protein structures. *Nucleic Acids Research*; **31**(13): 3337-40.

[54] Nair, R. and Rost, B. (2004). LOCnet and LOCtarget: subcellular localization for structural genomic targets. *Nucleic Acids Research*; **32**: W517-21.

[55] Neuberger G, Maurer-Stroh S, Eisenhaber B, Hartig A, Eisenhaber F. 2003. Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol.*;**328:**581–592.

[56] Niu B, Jin YH, Feng KY, Lu WC, Cai YD, Li GZ. 2008. Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. *Mol. Divers*. ;**12**:41–45.

[57] Petsalaki EI, Bagos PG, Litou ZI, Hamodrakas SJ. 2006. PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genomics Proteomics Bioinformatics*.;**4**:48–55.

[58] Pierleoni, A., Martelli, P. L., Fariselli, P. and Casadio, R. (2006) BaCelLo: a Balanced subCellular Localization predictor. *Bioinformatics*, **22**:408–e416.

[59] Pierleoni, A., Martelli, L. Fariselli, P and Casadio, R. (2007). eSLDB: eukaryotic subcellular localization database. *Nucleic Acids Research*, **35**: D208-D212.

[60] Prasad, K. T. S., Goel, R., Kandaswamy, K., Keerthikumar, S., Kumar, S. *et al.* (2009). Human protein reference database- 2009 update. *Nucleic Acids Research;* **37**: D767-72.

[61] Protein function. *About.com*. *http://biology.about.com/od/molecularbiology/a/aa101904a.htm* Last accessed: 04/02/13

[62] Pruitt, K. D., Tatusova, T., Brown, G. R. and Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*; **40**: D130-D135.

[63] Qi, D., Blake, J. A., Kadin, J. A., Richardson, J. E., Ringwald, M., Eppig, J. T. and Bult, C. J. (2005). Data integration in the mouse genome informatics (MGI) database. *Computational Systems Bioinformatics Conference IEEE*; 37-38.

[64] Quinn, S., Murphy, R. F., Singh, A. and Shatkay, H. (2010). A Framework for Inferring Protein Location as a Function of Condition. *CMU-CB-10-101:* 1-18.

[65] Rastogi, S. and Rost, B. (2011). LocDB: experimental annotations of localization for *Homo Sapiens* and *Arabidopsis thaliana*. *Nucleic Acids Research,* **39**: D230-D234.

[66] Rhee, S., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M. et al. (2003). The Arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**: 224–228.

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

[68] Rozengurt, E. (2011). Protein kinase D signaling: multiple biological functions in health and disease. *Physiology*; **26**(1): 23-33.

[69] Schein AI, Kissinger JC, Ungar LH. 2001. Chloroplast transit peptide prediction: a peek inside the black box. *Nucleic Acids Res.*;**29:**E82.

[70] Schwarz, E. M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Canaran, P., Chan, J., Chen, N., Chen, W. J., Davis, P. et al. (2006) WormBase: better software, richer content. *Nucleic Acids Res.*, **34:** D475–D478.

[71] Shafranovich, Y. Common format and MIME type for comma-separated values (CSV) files. Last accessed: 04/07/2013 http://tools.ietf.org/html/rfc4180

[72] Shen YQ, Burger G. Plasticity of a key metabolic pathway in fungi. Funct Integr Genomics. 2009;9(2):145–151.

[73] Small I, Peeters N, Legeai F, Lurin C. 2004. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences.*Proteomics.* ;**4:**1581–1590.

[74] Sprenger, J., Fink, J. L., Karunaratne, S., Hanson, K., Hamilton, N. A. and Teasdale, R. D. (2008). LOCATE: a mammalian protein subcellular localization database. *Nucleic Acids Research*, **36**: D230-D233.

[75] Swarbeck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Hemandez, M. G., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P. and Huala, E. (2007). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research*; **36**: D1009-D1014.

[76] Tamura T, Akutsu T. 2007. Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition. *BMC Bioinformatics.*;**8:**466.

[77] Tanz, S. K., Castleden, I., Hooper, C. M., Vacher, M., Small, I. and Millar, H. A. (2013). SUBA3: a database for integrating experimentation and prediction to define the subcellular location of proteins in Arabidopsis. *Nucleic Acids Research;* **41**: D1185-91.

[78] The Gene Ontology. *http://www.geneontology.org/* Last accessed: 04/12/13

[79] Tusnady, G. E., Dosztanyi, Z. and Simon, I. (2005). PDB_TM: selection and membrane organization of transmembrane proteins in the protein data bank. *Nucleic Acids Research*; **33**: D275-D278.

[80] Tusnady, G. E., Dosztanyi, Z. and Simon, I. (2005). TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*; **21**(7): 1276-1277.

[81] Uhlen, M., Bjorling, E., Agaton, C., Szigyarto, C. A., Amini, B. *et al.* (2005). A human protein atlas for normal and cancer tissues on antibody proteomics. *Mol Cell Proteomics;* **4**(12): 1920-32.

[82] Uniprot, C. (2010). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research;* **39**: D214-D219.

[83] Wiwatwattana, N. and Kumar, A. (2005). Organelle DB: a cross-species database of protein localization and function. *Nucleic Acids Research*; **33**: D598-D604.

[84] Wiwatwattana, N., Landau, C. M., Cope, J., Harp, G. A. and Kumar, A. (2007). Organelle DB: an updated resource of eukaryotic protein localization and function. *Nucleic Acids Research*, **35**: D810-D814.

[85] Wu, C. H., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z. Z., Ledley, R. S., Lewis, K. C., Mewes, H. W., Orcutt, B. C. et al. (2002) The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.*, **30**: 35-37.

[86] Zhang, S., Xia, X., Shen, J., Zhou, Y. and Sun, Z. (2008). DBMLoc: A database of proteins with multiple subcellular localizations. *BMC Bioinformatics;* **9**: 127.

Appendix A

LIST OF FILES AND PYTHON SCRIPTS USED

The Python scripts, data source files and the Django project and models are available in the directory "/*eecis/shatkay/Projects/JEPSLD*" on the Redtape server maintained by the Dept. of Computer and Information Sciences at the University of Delaware.

The descriptions of the files are as follows:

1) README: This file gives a brief description about all the files stored in the JEPSLD folder
2) proteinatlas.xml: This is the XML file downloaded from the Human Protein Atlas
3) Suba3-04-06-2013.xml: This is the XML file that was obtained after converting the CSV file downloaded from the SUBA3 database into XML format.
4) uniprot_sprot.xml: This is the XML file downloaded from UniProtKB/SwissProt database.
5) LOCATE_mouse_v6_20081121.xml: This is the XML file downloaded from the LOCATE database and contains information about mouse proteins.
6) LOCATE_human_v6_20081121.xml: This is the XML file downloaded from the LOCATE database and contains information about human proteins.

7) HPRDXML: This folder contains all the XML files downloaded from the Human Protein Reference Database
8) JEPSLD Database README: The JEPSLD database is hosted on the server db.eecis.udel.edu. This file contains details about creating the database, the tables, description of the tables and list of SQL queries to query the database
9) processProtatlas.py: This python script was used to obtain all relevant information from the XML file, obtained from the Human Protein Atlas, and store them in the JEPSLD database
10) procesSuba.py: This python script was used to obtain all relevant information from the XML file, obtained from the SUBA3 database, and store them in the JEPSLD database
11) processSwissprot.py: This python script was used to obtain all relevant information about eukaryotic proteins from the XML file, obtained from the UniProtKb/SwissProt database, and store them in the JEPSLD database
12) processLocatemouse.py: This python script was used to obtain all relevant information about mouse proteins from the XML file, obtained from the LOCATE database, and store them in the JEPSLD database
13) processLocatehuman.py: This python script was used to obtain all relevant information about humans from the XML file, obtained from the Locate human database, and store them in the JEPSLD database
14) processHprd.py: This python script was used to obtain all relevant information from the XML file, obtained from the HPRD database, and store them in the JEPSLD database

15) JEPSLD_backend: This python script was used to create a command line user interface and an admin interface. The user can only access and query the database whereas the admin has the privileges to insert, update and delete any specific entry in the database.
16) JEPSLD_frontend_README: This file gives a detailed description about installing Django, the Django-Python and Python-MySQL packages, creating the project JEPSLD and steps involved in updating the *models.py, settings.py, views.py* and *urls.py* files in the JEPSLD project. It also gives a detailed description about the purpose of each file in the Django project.
17) JEPSLD: This folder is a Django project that contains the data model *JEPSLD*, another folder *Templates* and the file *models.py*. The data model *JEPSLD* contains the files *settings.py, views.py* and *urls.py*. All these files have been discussed in detail in JEPSLD_frontend_README.