ANALYSES AND WEB INTERFACES FOR PROTEIN SUBCELLULAR LOCALIZATION AND GENE EXPRESSION DATA

A THESIS

SUBMITTED TO THE DEPARTMENT OF MOLECULAR BIOLOGY

AND GENETICS

AND THE INSTITUTE OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Biter Bilen

January, 2007

What does not kill me, makes me stronger.

Friedrich Nietzsche, Twilight of the Idols, 1888

Beni öldürmeyen şey, beni güçlendirir.

Friedrich Nietzsche, Putların Alacakaranlığı, 1888

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Rengül Çetin-Atalay (Supervisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Özlen Konu

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Volkan Atalay

Approved for the Institute of Engineering and Science:

Prof. Dr. Mehmet B. Baray
Director of the Institute Engineering and Science

ABSTRACT

ANALYSES AND WEB INTERFACES FOR PROTEIN SUBCELLULAR LOCALIZATION AND GENE EXPRESSION DATA

Biter Bilen
M.S. in Molecular Biology and Genetics
Supervisor: Assist. Prof. Dr. Rengül Çetin-Atalay
January, 2007

In order to benefit maximally from large scale molecular biology data generated by recent developments, it is important to proceed in an organized manner by developing databases, interfaces, data visualization and data interpretation tools. Protein subcellular localization and microarray gene expression are two of such fields that require immense computational effort before being used as a roadmap for the experimental biologist. Protein subcellular localization is important for elucidating protein function. We developed an automatically updated searchable and downloadable system called model organisms proteome subcellular localization database (MEP2SL) that hosts predicted localizations and known experimental localizations for nine eukaryotes. MEP2SL localizations highly correlated with high throughput localization experiments in yeast and were shown to have superior accuracies when compared with four other localization prediction tools based on two different datasets. Hence, MEP2SL system may serve as a reference source for protein subcellular localization information with its interface that provides various search and download options together with links and utilities for further annotations. Microarray gene expression technology enables monitoring of whole genome simultaneously. We developed an online installable searchable open source system called differentially expressed genes (DEG) that includes analysis and retrieval interfaces for Affymetrix HG-U133 Plus 2.0 arrays. DEG provides permanent data storage capabilities with its integration into a database and being an installable online tool and is valuable for groups who are not willing to submit their data on public servers.

Keywords: protein subcellular localization prediction, microarray gene expression, eukaryotic model organisms, web interface and database, proteome.

ÖZET

PROTEİN HÜCRE İÇİ YERLEŞİM VE GEN İFADESİ VERİLERİ İÇİN ANALİZLER VE ÖRÜN ARAYÜZLERİ

Biter Bilen
Moleküler Biyoloji ve Genetik, Yüksek Lisans
Tez Yöneticisi: Yard. Doç. Dr. Rengül Çetin-Atalay
Ocak, 2007

Moleküler biyolojideki son gelişmelerle ortaya çıkan büyük ölçekli verilerden en yüksek oranda yararlanabilmek için bunlarla organize şekilde ilgilenmek; veritabanları, arayüzler, veri görüntüleme ve yorumlama araçları geliştirmek gerekmektedir. Protein hücre içi yerleşimi ve mikrodizi gen anlatım ifadesi deneysel biyolojici için bir yol haritası olmadan önce yoğun hesaplamalar gerektiren iki alandır. Protein hücre içi yerleşimi protein işlevini açıklamak açışından önemlidir. Bu çalışmada, MEP2SL (model organisms proteome subcellular localization database) adında, model organizmaların tüm proteinleri için kendini güncelleyen aranabilir ve verileri bilgisayara aktarılabilir bir veritabanı yapılmıştır. Bu veritabanı, dokuz çokhücreli organizma için bilinen deneysel yerleşim bilgisinin yanısıra tahmine dayalı yerleşim bilgilerini barındırmaktadır. MEP2SL tahmine dayalı yerleşim sonuçları yüksek verimli deneysel maya yerleşim bilgileriyle uyumluluk göstermektedir. Ayrıca iki farklı veri kümesinde dört farklı yerleşim tahmin aracı doğruluk oranlarına göre daha iyi sonuçlar vermektedir. Bu bulgular göz önüne alındığında MEP2SL sistemi pek çok arama, verileri bilgisayara aktarma seçeneği yanısıra daha fazla bilgiye yönelik araçları ve bağlantılarıyla beraber protein hücre içi yerleşim bilgisi için bir referans kaynağı olabilecek niteliktedir. Mikrodizi teknolojisi tüm genomun aynı anda incelenmesi için uygun bir ortam hazırlamaktadir. Bu çalışmada Affymetrix HG-U133 Plus 2.0 dizileri için DEG (differentially expressed genes) adında, analiz ve veri geri aktarımı arayüzlerine sahip, örün üzerinde kurulabilen ve açık kaynak kodlu ayrımsal gen ifadeleri veritabanı kurulmuştur. DEG, veritabanı ile tamamlanması sonucu sürekli veri depolamaya imkan sağlar. Ayrıca örün üzerine kurulabilme özelliğiyle verilerini ortak erişime açık sunuculara göndermek istemeyen kullanıcılar için yararlı bir araçtır.

Anahtar sözcükler: protein hücre içi yerleşimi öngörüsü, mikrodizi gen ifadesi, çok hücreli model organizmalar, örün arayüzü ve veritabanı, proteom.

Acknowledgement

I am indebted to my thesis advisor Dr. Rengül Çetin-Atalay for her enthusiasm, patience and invaluable guidance during my thesis project. I am also deeply thankful to Dr. Volkan Atalay, Dr. Özlen Konu, and Dr. Mehmet Öztürk for their valuable comments and discussions. I would like to thank to all my associates and friends at Bilkent Molecular Biology and Genetics Department and Middle East Technical University Computer Engineering Department for their help, friendship, and share of knowledge. Finally, special thanks go to my family for their support and love.

This work was supported by the Turkish Academy of Sciences to R.C.A. This study is partially supported by TÜBİTAK under EEEAG-105E035 and TBAG-2268.

Abbreviations

BH	Benjamini Hochberg
BLAST	Basic Local Alignment Search Tool
BLASTp	Protein BLAST
BY	Benjamini Yekutieli
cDNA	Complementary Deoxyribonucleic Acid
CGI	Common Gateway Interface
CYGD	The Comprehensive Yeast Genome Database
DEG	Differentially Expressed Genes System
DNA	Deoxyribonucleic Acid
EC	Enzyme Commission
ER	Endoplasmic Reticulum
FDR	False Discovery Rate
GO	Gene Ontology
HBV	Hepatitis B Virus
HCC	Hepatocellular Carcinoma
HPRD	Human Protein Reference Database
KEGG	Kyoto Encyclopedia of Genes and Genomes
MEP2SL	Model Organisms Proteome Subcellular Localizations System
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NUSE	Normalized Unscaled Standard Errors
OMIM	Online Mendelian Inheritance in Man
PLM	Probe-level Linear Models
RLE	Relative Log Expression
RMA	Robust Multiple-array Average
RNA	Ribonucleic Acid
SOM	Self Organizing Map
SQL	Structured Query Language
SVM	Support Vector Machine
UNIX	Uniplexed Information and Computing System

Contents

1 Introduction 1

1.1 Motivation 1

1.2 Organization of the Thesis 4

2 Materials and Methods 5

2.1 MEP2SL 5

2.1.1 Dataset 5

2.1.2 MEP2SL Infrastructure 7

2.1.3 Protein Localization Predictor Evaluation 11

2.2 DEG 13

2.2.1 Dataset 13

2.2.2 DEG Infrastructure 13

3 Results 20

3.1 MEP2SL 20

3.1.1 Query Specification Interface of MEP2SL 21

3.2 Protein Subcellular Localization Analysis 23

3.3 Protein Subcellular Localization Predictor Comparison 23

3.4 DEG 34

3.4.1 Interface 34

3.4.2 Case Study Using DEG Interface 35

4 Conclusions and Future Work 49

A Localization Labeling for Predictor Evaluation 58

A.1 Localization Labels of CYGD Dataset 58

A.2 Localization Labels of HPRD Dataset 59

A.3 Localization Labels of Prediction Tools 60

List of Figures

2.1 Internal structure of MEP2SL. 6

2.2 Phenodata file and CEL archive file for DEG upload. 14

2.3 Internal structure of DEG. 15

3.1 Scaled color-coded Venn diagram for protein subcellular localization distribution in nine model organisms. 24

3.2 Color-coded Venn diagram for human proteome subcellular localization distribution. 25

3.3 Color-coded Venn diagram for mouse proteome subcellular localization distribution. 26

3.4 Color-coded Venn diagram for rat proteome subcellular localization distribution. 27

3.5 Color-coded Venn diagram for fruit fly proteome subcellular localization distribution. 28

3.6 Color-coded Venn diagram for zebrafish proteome subcellular localization distribution. 29

3.7 Color-coded Venn diagram for yeast proteome subcellular localization distribution. 30

3.8 Color-coded Venn diagram for frog proteome subcellular localization distribution. 31

3.9 Color-coded Venn diagram for slime mold proteome subcellular localization distribution. 32

3.10 Color-coded Venn diagram for worm proteome subcellular localization distribution. 33

3.11 RNA degradation plot. 37

3.12 Histogram plot. 38

3.13 Pre-normalization boxplot. 39

3.14 M versus A plot. 40

3.15 PLM residuals image. 41

3.16 PLM RLE plot. 42

3.17 PLM NUSE plot. 43

3.18 Post-normalization boxplots. 45

3.19 DEG retrieve interface. 47

3.20 DEG merge interface. 48

List of Tables

2.1 MEP2SL localization types. 8

2.2 MEP2SL database table field names. 9

2.3 MEP2SL database table field names. 11

2.4 DEG data table field names. 17

2.5 DEG HGUmetaData table field names. 18

2.6 DEG HGUmetaDataPair table field names. 18

2.7 DEG HGUmetaDataProc table field names. 18

3.1 Proteome subcellular localization distributions. 22

3.2 Evaluation of subcellular localization tools on CYGD dataset. . . 25

3.3 Evaluation of subcellular localization tools on HPRD dataset. . . 26

3.4 Significant probe numbers of rma & gcrma normalization methods. 46

A.1 CYGD dataset protein subcellular localization labeling. 58

A.2 HPRD dataset protein subcellular localization labeling. 59

A.3 Protein subcellular localization labeling of prediction tools. . . . 60

Chapter 1

Introduction

Recent developments in molecular biology require *in silico* analysis of the large scale genome and proteome data prior to laboratory studies. In order to benefit maximally from this vast amount of data, one must deal with data in an organized way: this implies establishing, sustaining and distributing databases, providing user friendly interfaces, and state-of-the-art visualization and data interpretation tools [41]. Only by these means experimentalists could get the roadmap they need to analyze their data. To fulfill this need, we aimed to analyze large scale biological data and constructed online analysis interfaces for protein subcellular localization and microarray gene expression data analysis.

1.1 Motivation

We first have chosen protein subcellular localization analysis since functional annotation of thousands of gene products produced out of an experiment is a challenging task for understanding the biological behavior of a system. Investigation of the subcellular localization of a set of proteins is invaluable in terms of better representation of cellular machinery with respect to the site of protein action and the pathways in which these proteins are involved since each compartment and its vicinity contain functionally linked proteins associated with them [42].

Eventually, studying proteome wide subcellular localization may have its implication in recent advances regarding systems biology for better representation of cellular machinery. Subcellular localization of a protein can be experimentally determined through *in vivo* techniques. However, the number of experimentally obtained data is very limited since all of these experimental methods are time consuming and costly. In addition, a protein may have more than one site of localization. Therefore, *in silico* analysis of protein subcellular localization is required through computational prediction techniques.

There are various methods with comparable accuracy for subcellular localization prediction based on the existence of signal peptide cleavage sites on protein sequences (TargetP [12], PSORT [29], and SignalP [31]). In addition, machine learning methods that cover extensive biological knowledge, such as amino acid composition, protein sequence homology, and protein and literature database text analysis, have been applied to achieve a better accuracy of prediction (SortPred [13], pTARGET [17], LOC3D [28], SubLoc [19], PASUB [26], and P2SL [8]). Based on the accuracy rates of these localization prediction tools, 90% sorting precision achievement among primary localizations does not seem unlikely. However, multi-functional proteins with more than one acting-site hinder the development of near-perfect prediction tools [32]. Hence, the prediction of subcellular localization can be considered as a tool that gives the molecular biologist an initial opinion for the experimental design. Motivated by this fact, we compared the accuracy of five protein subcellular localization tools with two different protein datasets. Among these tools, P2SL is a hybrid machine learning based, subcellular localization prediction tool founded on implicit motif distribution which employs local subsequence features together with several amino acid similarity schemes. We selected P2SL, which gave comparable accuracy results and constructed an automatically updated, downloadable, and searchable web interface called MEP2SL (http://www.i-cancer.org/mep2sl) for the prediction and representation of multi-compartmental protein subcellular localizations in nine eukaryotic model organism proteomes: human, mouse, rat, fruit fly, zebrafish, yeast, frog, slime mold, and worm.

Second, we chose microarray gene expression data analysis which enables monitoring the whole genome simultaneously in a single DNA microarray chip. Microarray technology gives a global view since the genes in a living organism function collaboratively. The microarray technology has two variants in terms of the property of arrayed DNA sequence with known identity: In the first technology, probe cDNA (500-5 000 bases long) is immobilized to a solid surface such as glass using robot spotting and exposed to a set of targets either separately or in a mixture [11]. In the second one, an array of oligonucleotide (20-80-mer oligos) or peptide nucleic acid probes is synthesized either *in situ* (on-chip) or by conventional synthesis followed by on-chip immobilization. The array is exposed to labeled sample DNA, hybridized, and complementary sequences are determined.

The analysis and interpretation of the large amount of data produced out of a microarray experiment is not possible without the integration of statistical analysis and appropriate visualization and annotation tools. Recently, Bioconductor project, which is based on the statistical programming language R, (http://www.R-project.org) has been a reference tool for the analysis and interpretation of these experiments; however, using R is not an easy task for novice programmers. This brings the need for a user-friendly graphical array analysis application. However, microarray analysis is not technically performable on a standard computer due to large memory requirement. Hence, a powerful machine is required on which to run an analysis. There are numerous analysis pipelines for both cDNA and oligo array analysis including web-based tools like GEPAS [27], ArrayPipe [23], MIDAW [10], RACE [35], or CARMAweb [37]. To the best of our knowledge, there is not an installable integrated web based software and database which brings a simple and user-friendly analysis pipeline for gene expression data analysis for microarray research laboratories who need to perform their analysis on their own without submitting to a generic web site. Therefore, we constructed a simple but comprehensive web based application that is online installable searchable open source web based analysis suite for Affymetrix GeneChip arrays.

1.2 Organization of the Thesis

The thesis is organized as follows. Chapter 2 described the datasets and methods used to construct the mentioned web interfaces together with multicategory protein localization analysis method. In Chapter 3, the web-based interfaces together with multicategory protein localization evaluation results and a case study for gene expression data analysis have been given. Finally, Chapter 4 presented conclusions and future perspectives of the study.

Chapter 2

Materials and Methods

2.1 MEP2SL

MEP2SL is an automatically updated downloadable and searchable system for predicted protein subcellular localization information. MEP2SL runs on a Linux operating system. It is developed and implemented using the MySQL relational database system and Perl-CGI for server side scripting language. We use standalone BLAST [7] and its specific database constructor tool, `formatdb`, for sequence alignments. UNIX utilities, `wget` and `cron` are used to implement automatic updating feature.

2.1.1 Dataset

Version releases in UniRef100 database are checked periodically, once a week. If a new version exists, the eukaryotic protein sequence data is downloaded from UniRef100 database and the sequences from human, mouse, rat, fruit fly, zebrafish, yeast, frog, slime mold, and worm based on the model organism classification of National Institutes of Health (NIH) [3] are extracted. After additional processing as mentioned in Section 2.1.2, the current dataset is composed of 217 102 protein sequences for UniRef100 v.9.2 data.

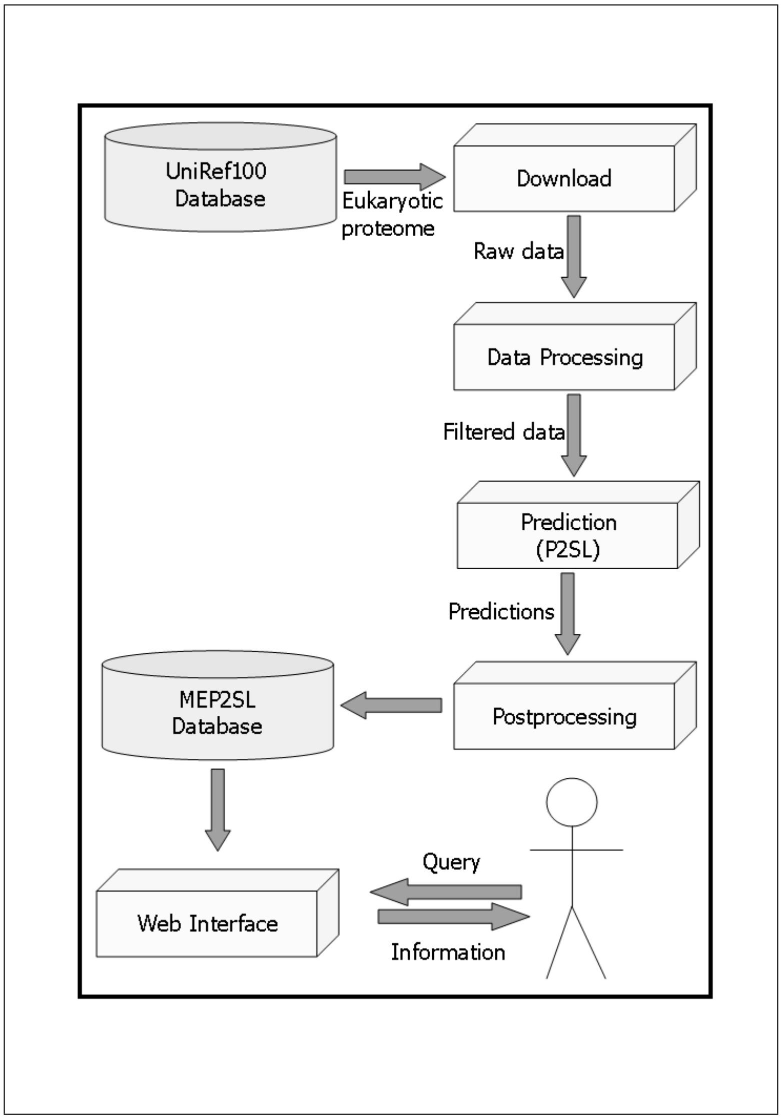


Figure 2.1: Internal structure of MEP2SL. Five modules are represented in boxes.

2.1.2 MEP2SL Infrastructure

MEP2SL is implemented with five sequential modules: Download, Data Processing, Prediction, Postprocessing, and Web Interface along with MEP2SL database as shown in Figure 2.1. These modules are controlled by a main process, scheduled to work once a week by a system scheduling event (`cron`). Hence, cyclic execution of the modules fulfills the automatic update feature of the system.

2.1.2.1 Download

Comparing the previously used UniRef100 Database release file with the current UniRef100 release file, if the current release number of the UniRef100 is greater than the one used in MEP2SL system, this module downloads the eukaryotic proteome data from UniRef100 Database as xml and fasta formatted files along with the UniRef100 release information file from UniRef100 site with a network downloader (`wget`).

2.1.2.2 Data Processing

Data processing module is responsible for filtering and formatting of the proteome data to be processed by the prediction module. It extracts the sequences of selected model organisms. Protein sequences containing less than 50 amino acid residues or containing one of X, Z, U, and B amino acid codes are excluded from above mentioned nine model organism sequences due to the prediction tool restrictions. The organism sequence files are formatted for the usage of the prediction module as one-line protein sequence files.

2.1.2.3 Prediction

P2SL tool is used in the prediction module which determines the frequency distribution of protein subsequences over nuclear, cytosolic, mitochondrial and endoplasmic reticulum (ER) targeted subcellular localization classes and then uses this

Table 2.1: MEP2SL localization types. i.e. `3/3 Nuclear & 2/3 Cytosolic` represents a protein that localizes to the nucleus with 3/3 possibility and to the cytosol with 2/3 possibility.

Localization Type
3/3 Nuclear
3/3 Cytosolic
3/3 ER Targeted
3/3 Mitochondrial
3/3 Nuclear & 2/3 Cytosolic
3/3 Nuclear & 2/3 ER Targeted
3/3 Nuclear & 2/3 Mitochondrial
3/3 Cytosolic & 2/3 Nuclear
3/3 Cytosolic & 2/3 Mitochondrial
3/3 Cytosolic & 2/3 ER Targeted
3/3 Mitochondrial & 2/3 Nuclear
3/3 Mitochondrial & 2/3 Cytosolic
3/3 Mitochondrial & 2/3 ER Targeted
3/3 ER Targeted & 2/3 Nuclear
3/3 ER Targeted & 2/3 Mitochondrial
3/3 ER Targeted & 2/3 Cytosolic
2/3 Cytosolic & 2/3 Nuclear
2/3 Cytosolic & 2/3 Mitochondrial
2/3 ER Targeted & 2/3 Cytosolic
2/3 ER Targeted & 2/3 Nuclear
2/3 ER Targeted 2/3 Mitochondrial
2/3 Mitochondrial & 2/3 Nuclear
2/3 Cytosolic & 2/3 Mitochondrial & 2/3 Nuclear
2/3 ER Targeted & 2/3 Mitochondrial & 2/3 Nuclear
2/3 ER Targeted & 2/3 Cytosolic & 2/3 Nuclear
2/3 ER Targeted & 2/3 Cytosolic & 2/3 Mitochondrial

Table 2.2: MEP2SL database table field names.

Field	Type	Key	Default
id	varchar(30)	PRI	
loc	varchar(8)		
des	varchar(255)		
seq	text		
exp	text		
met	varchar(10)		p2sl

distribution as a feature for classification. Localization class probability distributions are represented by samples of subsequence distributions over self-organizing maps. The following binary support vector machine (SVM) classifiers are then used for the classification:

- ER versus Cytosolic,
- ER versus Mitochondrial,
- ER versus Nuclear,
- Mitochondrial versus Cytosolic,
- Mitochondrial versus Nuclear,
- Nuclear versus Cytosolic.

Each class is voted over three classifiers. Then, majority voting gives the predicted localization class(es). Compartments gaining more than one vote are considered as significant and others are treated as insignificant. Hence, there exists twenty six significant localization types for the predictions as given in Table 2.1. These results were also presented in a color-coded Venn diagram as shown in Figure 3.2.

2.1.2.4 Post Processing

In postprocessing module, localization results and the sequences were stored in a relational database and in a local BLAST database. The relational database is for standard queries and contains nine database tables for each of the model organism. These tables have the same six fields; including UniRef100 database id (id), predicted localization (loc), sequence description (des), the protein sequence (seq), UniProt Knowledgebase protein subcellular localization annotation (exp) and prediction method (met) as given in Table 2.2. The local BLAST database is for sequence queries to perform pairwise sequence alignment and it contains the same information as the relational database tables but it is structured differently by the built-in BLAST database construction tool, `formatdb`. In addition, this module is responsible for reflecting the changes to web site interface including the generation of the protein subcellular localization distribution table as given in Table 3.1 in download interface and the color-coded Venn diagram images for each model organism as shown in Figure 3.2, Figure 3.3, Figure 3.4, Figure 3.5, Figure 3.6, Figure 3.7, Figure 3.8, Figure 3.9, and Figure 3.10. Venn Diagram images are generated with offscreen rendering library of MESA (`libOSMesa`). Furthermore, partial and whole downloadable files of prediction results are made into archive in this module. These files have tabularly separated plain text format composed of five columns:

1. UniRef100 id,
2. Predicted subcellular localization distribution of the sequence,
3. Sequence description,
4. Sequence,
5. Annotated subcellular localization from UniProt Knowledgebase.

2.1.2.5 Web Interface

We supply information through web interface module when a user requests it through download and search interfaces. Users may download the prediction results either as complete or as partial data for each organism and localization class. Protein localization distributions for each organism are observable via the color-coded Venn diagrams. The search interface consists of standard queries and sequence query in the MEP2SL database. Keyword standard query matches

Table 2.3: MEP2SL database table field names.

Query	Database	Database Field
Keyword	mySQL database	des
Localization	mySQL database	loc
Localization Compartment	mySQL database	loc
Database Id	mySQL database	id
Sequence	BLAST database	seq

to descriptions of sequences using logical operators AND and OR. Database Id standard query matches to UniRef100 sequence ids. Localization standard query exact matches to localization distributions. Localization Compartment standard query partial matches to localizations. Finally, Sequence query matches to sequences by BLASTp with the chosen expectation value (E-Value) as given in Table 2.3.

2.1.3 Protein Localization Predictor Evaluation

We used five protein subcellular localization tools:

- PA-SUB [26],
- P2SL [8],
- PSORT2 [29],
- pTARGET [17],
- TargetP [12]

for comparison of the subcellular localization prediction on two annotated datasets from HPRD [33] (Human Protein Reference Database) from [2] and CYGD [18] (Comprehensive Yeast Genome Database) from [1], Initially, CYGD (updated on 14-11-2005) and HPRD v.6 datasets consisted of 18 841 and 6 736 protein sequences, respectively. After extraction of proteins having subcellular localization information, we ended up with 4 692 proteins in CYGD and 11 557 proteins in HPRD before using these with the mentioned predictors. Predictions are done on web servers of pTARGET (last updated on 03-02-2006) at [5], TargetP v.1.1 at [6], and PA-SUB v.2.5 at [4]. However, PSORT (last revised on 01-12-1998) and P2SL v.0.1 predictions are done in house. Prediction evaluation are done with our multi-category accuracy evaluation criteria that is explained in Section 2.1.3.2.

2.1.3.1 Category Mapping in Actual and Predicted Sets

Every predictor we used predicts over varying number of categories and assigns different reliability scores, probabilities, etc. for the categories they predict over. However, we did not consider the prediction scores of categories and considered only the existence of a category in predicted set and labeled each category with a unified scheme as given in Table A.3. By labeling the actual set categories with the same labels we chose to label subcellular localization tools as indicated in Table A.1 and Table A.2, we had a universal label set. Over this universal set, we provided set intersection and coverage operations to assign prediction accuracy as mentioned in Section 2.1.3.2.

2.1.3.2 Multi-category Accuracy Evaluation Criteria

For every protein sequence, we have an actual set of compartments set by the dataset and a predicted set of compartments predicted by a multi-category localization predictor. Handling these two sets, we should give an accuracy score for the performance of the mentioned five prediction tools in a test dataset. However, assigning a generalized accuracy criterion in multi-category predictions is not a straightforward task especially when the predictors produce a range of outputs [32]. We produce a one-to-one mapping between actual set as and predicted set as mentioned in Section 2.1.3.1. Using these mappings, we defined a rough accuracy range with worst and best case criteria which we consider the precise accuracy of a multi-category prediction tool should be in between. The best case accuracy criterion assigns a prediction true whenever an intersection set between actual and predicted sets exists. However, the worst case accuracy assigns true whenever predicted set covers actual set. Otherwise, a prediction is considered as false both in worst and best case criteria. For evaluating the accuracy of a tool on a test set, we sum the number of true predictions for worst and best case criteria separately and present altogether.

2.2 DEG

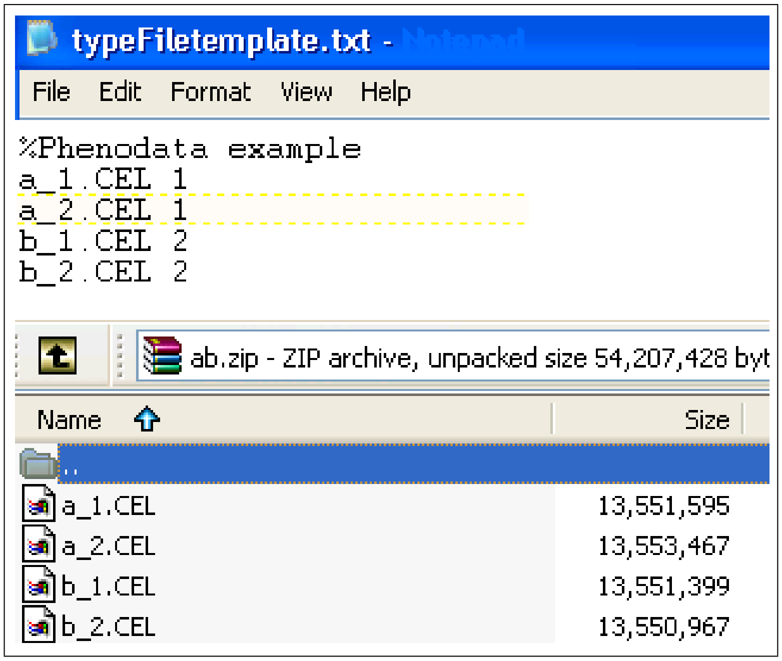
Differentially Expressed Genes (DEG) is an installable, downloadable, and open source analysis suite for Affymetrix HG-U133 Plus 2.0 array. DEG runs on a Linux operating system. It is developed and implemented using the MySQL relational database system, Perl-CGI for server side scripting language and R statistical programming language [36] and R Bioconductor packages RColorBrewer [30], affyPLM [9], affy [21], gcrma [43], multtest [34], siggenes [38], genefilter [14], annaffy [40], hgu133plus2 [25] for calculations, visualizations and annotations.

2.2.1 Dataset

Archive of Affymetrix HG-U133 Plus 2.0 array CEL files together with user specified phenodata file is required as shown in Figure 2.2. The phenodata file has a tabular plain text format. It is composed of two columns where first column includes the name of the CEL file, and the second contains the phenotype of that CEL file. The phenotype of a CEL file should be a 1-2 digit integer number and the maximum number of CEL files in an archive is not restricted; however it should be considered according to the server machine memory and processor capabilities.

2.2.2 DEG Infrastructure

DEG has two main interfaces. The first interface is used for CEL file analysis which needs to be performed once and consists of CEL file upload, normalization, significance analysis, annotation and loading of data into DEG Database.


typeFiletemplate.txt - Notepad
File
Edit
Format
View
Help

```
%Phenodata example
a_1.CEL 1
a_2.CEL 1
b_1.CEL 2
b_2.CEL 2
```

ab.zip - ZIP archive, unpacked size 54,207,428 byt

Name	Size
..	
a_1.CEL	13,551,595
a_2.CEL	13,553,467
b_1.CEL	13,551,399
b_2.CEL	13,550,967

Figure 2.2: Phenodata file and CEL archive file for DEG upload. `typeFileTemplate.txt` is the phenodata file and `ab.zip` is the compressed CEL file.

The second interface is for the retrieval and merging together of the previously performed analyses as shown in Figure 2.3.

2.2.2.1 CEL File Analysis

The evaluation of this part consists of four modules and takes long execution times.

2.2.2.1.1 Upload and Quality Control File uploading is the initial step of the CEL File Analysis interface. The user specified phenodata and compressed CEL files are downloaded to the server. In addition, user specified quality control images are produced by using R Bioconductor packages `RColorBrewer`, `affy`, and `affyPLM` which are as below:

- Boxplot,

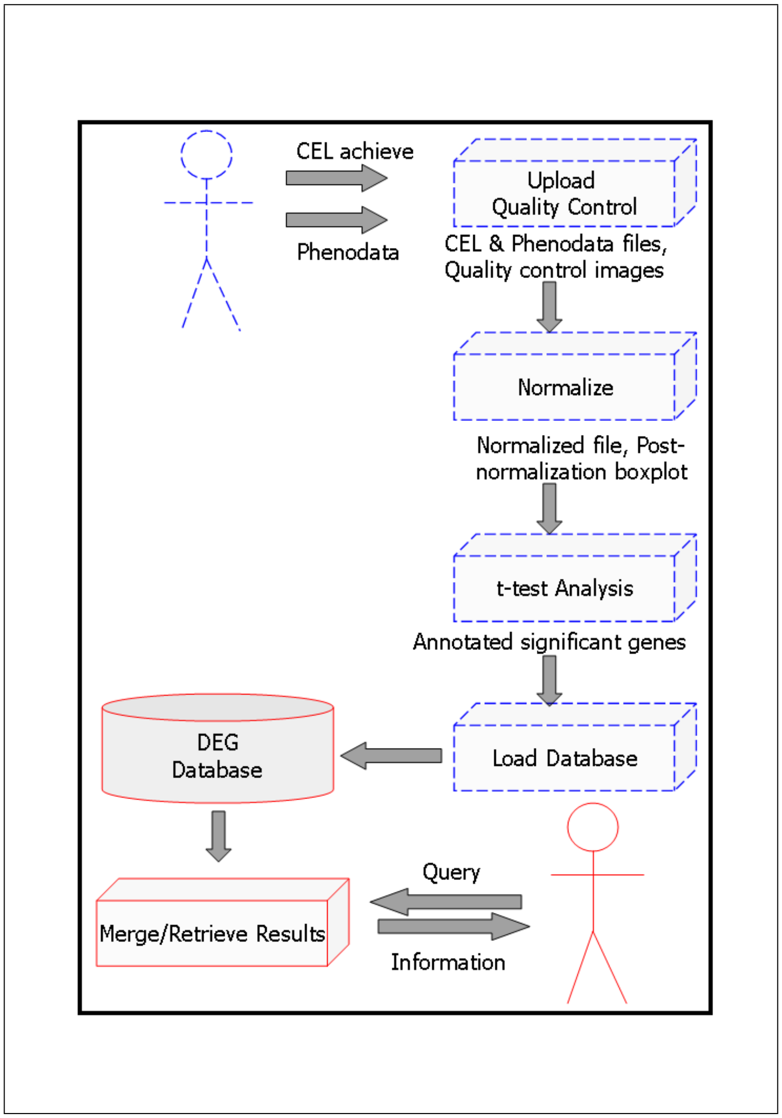


Figure 2.3: Internal structure of DEG. CEL file Analysis and Retrieval interfaces are represented with dashed blue and continuous red lines, respectively.

- Histogram,
- MAplot,
- RNA Degradation,
- PLM Residuals Image,
- PLM RLE (Relative Log Expression),
- PLM NUSE (Normalized Unscaled Satandard Error).

2.2.2.1.2 Normalize The files in the CEL file archive that also exist in the first column of the phenodata file are renamed with their phenotype information. These files are normalized according to the user selected normalization method of either gcrma or rma with R Bioconductor `affy` and `gcrma` packages. In addition, post normalization boxplots are produced.

2.2.2.1.3 t-test Analysis The user is fronted with all pair combinations of CEL file phenotypes that exist in the normalized file. Upon selecting some of these pairs, the specified t-test analysis (equal/unequal variance, paired/unpaired samples, two/one tailed) for obtaining the differentially expressed genes is performed. Here, the user may restrict the expression values that are included in the t-test analysis with `expression value limit` and select multiple hypothesis correction methods among `BH`, `BY`, `Bonferroni`, `Hochberg`, `Holm`, `SidakSD`, and `SidakSS`. These calculations are performed with R Bioconductor multtest, siggenes, genefilter packages. Upon finding the differentially expressed genes, they are annotated with raw and adjusted p-values, up/down regulation information, Gene Symbol, GenBank Accession Number, Chromosomal Location, Chromosome, Entrez Gene Id, Enzyme Commission (EC) Id, Gene Gene Ontology (GO), Cytogenetic Maps, OMIM Id, KEGG Pathway, PubMed Id, RefSeq Id, UniGene Cluster Id. The annotations are performed with R Bioconductor `affy`, `annaffy`, and `hgu133plus2` packages. At the end, an analysis Id is supplied to the user to be used in the retrieval and merge interfaces as mentioned in Section 2.2.2.2.

[1] User selected multiple hypothesis selection procedures (ADJP) among `BH`, `BY`, `Bonferroni`, `Hochberg`, `Holm`, `SidakSD`, and `SidakSS` for keeping adjusted p-values.

[2] Phenodata-CEL file names (CELFILENAME) for keeping probe intensity values.

Table 2.4: DEG data table field names.

Field	Type	Key	Default
TYPE	varchar(30)	PRI	
ID	varchar(30)	PRI	
REG	char(2)		
SYMBOL	varchar(60)		
ACCNUM	varchar(60)		
CHRLOC	text		0
CHR	int(2)		
ENTREZID	varchar(60)		
ENZYME	text		
GENENAME	text		
GO	text		
MAP	text		
OMIM	text		
PATH	text		
PMID	text		
REFSEQ	text		
UNIGENE	text		
RAWP	decimal(11,10)		0
ADJP1	decimal(11,10)		0
ADJP2	decimal(11,10)		0
..[1]	decimal(11,10)		0
ADJP7	decimal(11,10)		0
CELFILENAME1	decimal(11,10)		0
CELFILENAME2	decimal(11,10)		0
..[2]	decimal(11,10)		0
CELFILENAMEn	decimal(11,10)		0

2.2.2.1.4 Load Database The annotation fields together with the expression values of a particular probe is stored in a database table as given in Table 2.4 In order to maintain the dynamic content of the web interface for `CEL File Analysis Retrieval Interface` and `Merging Interface` as mentioned in Section 2.2.2.2, the analysis parameters are stored in three meta tables as given in Table 2.5, Table 2.6, and Table 2.7.

Table 2.5: DEG HGUmetaData table field names.

Field	Type	Key	Default
aid	int(5)	PRI	
time	timestamp		
norMet	varchar(255)		
expLim	text		
ttest	varchar(10)		p2sl

Table 2.6: DEG HGUmetaDataPair table field names.

Field	Type	Key	Default
aid	int(5)		0
proc	varchar(15)		

Table 2.7: DEG HGUmetaDataProc table field names.

Field	Type	Key	Default
aid	int(5)	PRI	0
pairvalue	varchar(30)	PRI	
pairlabel	text		

2.2.2.2 CEL File Analysis Retrieval and Merging

This part is for the quick retrieval of a previously performed microarray analysis. It has two functionalities; one is for retrieval of single t-test analysis, and the other is the merging of the two t-test analyses. User specifies either gene symbol or probe id based retrieval or merging. User specifies FDR, gene regulation, annotation fields among Gene Symbol, GenBank Accession Number, Chromosomal Location, Chromosome, Entrez Gene Id, EC Id, GO, Cytogenetic Maps, OMIM Id, KEGG Pathway, PubMed Id, RefSeq Id, UniGene Cluster Id.

Chapter 3

Results

After recent advances in the information technology, individual groups developed applications for their own use. However, there is a great need for integration of information. Here, we present two such information integration approaches. One is for protein subcellular localization information and the other is for the determination and annotation of differentially expressed genes. For the global interpretation of protein subcellular localization information across proteomes, we constructed an database called MEP2SL and additionally confirmed our prediction method by yeast high throughput experimental localization information and prediction results of other tools. In expression data analysis, we constructed an online analysis suite called DEG and presented a case study for the usage and interpretation of it.

3.1 MEP2SL

MEP2SL is an automatically updated downloadable and searchable system housing predicted and existing existing experimental subcellular localization information of nine model organisms: human (*H. sapiens*), mouse (*M. musculus*), rat (*R. norvegicus*), fruit fly (*D. melanogaster*), zebrafish (*D. rerio*), yeast (*S. cerevisiae*), frog (*X. tropicalis*), slime mold (*D. discoideum*), and worm (*C. elegans*).

The predictions are made with a machine learning tool, P2SL. P2SL is a multi-class subcellular localization tool and gives protein localization probabilities over ER targeted, cytosolic, nuclear and mitochondrial cellular compartments. Considering some votes as insignificant, we come up with twenty-six different protein localization distribution types. This data is downloadable through a web interface as whole or single download of protein localization distribution files. The possible queries are presented in the next section.

3.1.1 Query Specification Interface of MEP2SL

Four standard searches (keyword, database id, localization, and localization compartment) can be performed which extract information from the relational database. Matched sequences are represented in a table from which users may access to a detailed page for the specific sequence. The detail page represents subcellular localization distribution possibility, and UniProt Knowledgebase subcellular localization along with a UniProt link to get additional biological features, and an NCBI BLAST link to find homologous sequences.

In addition to the standard search results, users may have the pairwise alignment of the matched sequence to the queried sequence using the sequence search which is supported by a local BLAST in the local BLAST database. The sequence search option may be used for experimentally designed peptide localization prediction. A user may construct an arbitrary peptide and test its localization by the MEP2SL sequence search option on local BLAST.

Table 3.1: Proteome subcellular localization distributions by P2SL for UniRef100 v.9.2.

Protein Localization Distribution Type	Zebrafish	Worm	Slime mold	Fruit fly	Human	Mouse	Rat	Yeast	Frog
3/3 Nuclear	129	99	17	281	649	499	110	29	42
3/3 Cytosolic	171	209	25	193	495	436	121	66	57
3/3 ER-Targeted	319	784	48	638	1166	1037	304	116	100
3/3 Mitochondrial	157	149	5	216	799	643	161	81	64
3/3 Nuclear and 2/3 Cytosolic	3702	3524	653	6315	13927	11809	2511	1171	1279
3/3 Nuclear and 2/3 ER-Targeted	73	85	19	160	287	275	64	23	22
3/3 Nuclear and 2/3 Mitochondrial	285	211	13	780	1819	1444	298	95	74
3/3 Cytosolic and 2/3 ER-Targeted	190	297	43	221	1171	586	143	86	84
3/3 Cytosolic and 2/3 Mitochondrial	350	473	24	486	1242	1133	290	190	123
3/3 Cytosolic and 2/3 Nuclear	6152	8442	1014	7418	18892	16177	4089	3122	2387
3/3 Mitochondrial and 2/3 Cytosolic	639	688	26	1092	3124	2692	723	345	324
3/3 Mitochondrial and 2/3 ER-Targeted	163	161	7	213	1072	751	237	65	74
3/3 Mitochondrial and 2/3 Nuclear	174	167	10	581	1909	1364	242	82	82
3/3 ER-Targeted and 2/3 Cytosolic	1312	2598	236	1606	3002	3208	936	381	423
3/3 ER-Targeted and 2/3 Mitochondrial	1584	2840	101	2631	7430	6398	1988	452	590
3/3 ER-Targeted and 2/3 Nuclear	252	675	79	650	1053	941	282	69	76
2/3 Cytosolic and 2/3 Mitochondrial	132	212	6	221	536	465	111	64	61
2/3 Cytosolic and 2/3 Nuclear	285	427	30	435	1247	921	206	155	89
2/3 ER-Targeted and 2/3 Cytosolic	138	249	26	190	523	418	105	73	43
2/3 ER-Targeted and 2/3 Mitochondrial	113	178	2	178	718	539	110	59	45
2/3 ER-Targeted and 2/3 Nuclear	44	97	9	122	333	222	51	30	22
2/3 Mitochondrial and 2/3 Nuclear	92	110	7	155	500	344	61	40	27
2/3 Cytosolic and 2/3 Mitochondrial and 2/3 Nuclear	405	509	18	748	2169	1699	321	203	154
2/3 ER-Targeted and 2/3 Cytosolic and 2/3 Mitochondrial	69	102	7	108	281	227	50	36	23
2/3 ER-Targeted and 2/3 Cytosolic and 2/3 Nuclear	253	413	43	351	783	753	156	108	85
2/3 ER-Targeted and 2/3 Mitochondrial and 2/3 Nuclear	29	23	2	67	179	144	32	9	9
Total	17212	23722	2470	26056	65306	55125	13702	7150	6359

3.2 Protein Subcellular Localization Analysis

MEP2SL contained a total of 217 102 protein sequences from the nine model organisms from UniRef100 v.9.2. The human proteome constituted the largest set with 65 529 sequences while slime mold proteome was the smallest with 3 393 sequences as given in Table 3.1. Subcellular localization distributions for all organisms were visualized in detail in a color-coded Venn diagram where similar distribution patterns can be observed as shown in Figure 3.1. Venn diagram representation of subcellular localizations clearly demonstrates that proteins are not single site acting molecules. For instance, in human proteome, only 3 154 over 65 529 protein sequences (4.81%) were predicted to be located or acting in a single compartment; yet more than half of the human proteins (52.15%) were predicted to localize both in nucleus and cytosol.

Similar percentile distributions were also observed in other organisms. From each organism analyzed in this study, between 28 to 44% of the protein sequences from different proteomes were predicted to be 3/3 Cytosolic & 2/3 Nuclear, meaning that proteins localize to the cytosol with 3/3 possibility and to the nucleus with 2/3 possibility. Between 14 to 27% of all proteins were in 3/3 Nuclear & 2/3 Cytosolic distribution type. Therefore in general, the majority of proteins are distributed between cytosol and nucleus indicating that these proteins may have roles in both or either compartment. This phenomenon is a good demonstration of how cell signaling system works such that 3/3 Cytosolic & 2/3 Nuclear or 3/3 Nuclear & 2/3 Cytosolic proteins interact with signaling proteins in the cytosol in order to be localized to the nucleus upon simulation by an external signal or when they are done with their duty in the nucleus they are shuttled back to the cytosol [16].

3.3 Protein Subcellular Localization Predictor Comparison

We compared the accuracy of five protein subcellular localization tools on one human dataset and one yeast dataset. We calculated the accuracy of five protein

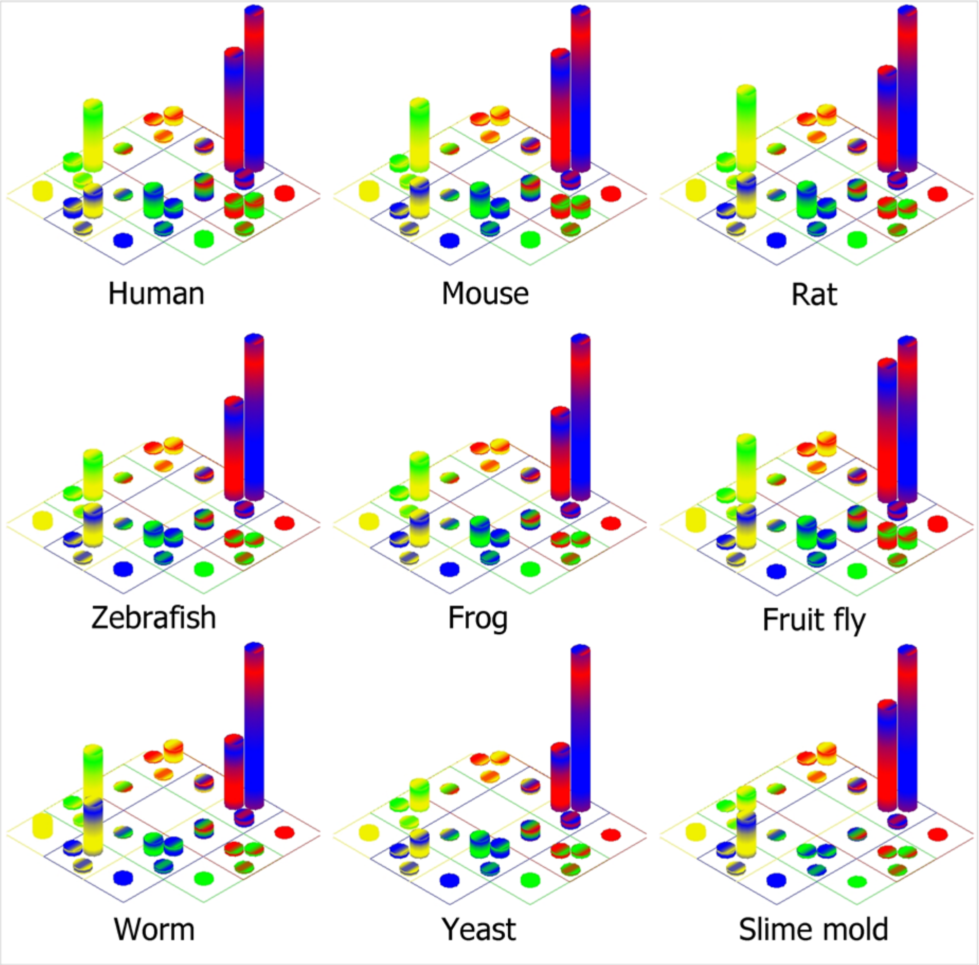


Figure 3.1: Scaled color-coded Venn diagram for protein subcellular localization distribution in nine model organisms. Protein subcellular localization distribution is represented with twenty-six columns over nuclear (red), cytosolic (blue), mitochondrial (green), and ER targeted (yellow) subcellular localizations. Thickness of the colored bands indicates the prediction votes such that thinner band is for two votes and thicker one is for three votes. In each organism, the distribution pattern of the localizations are similar others.

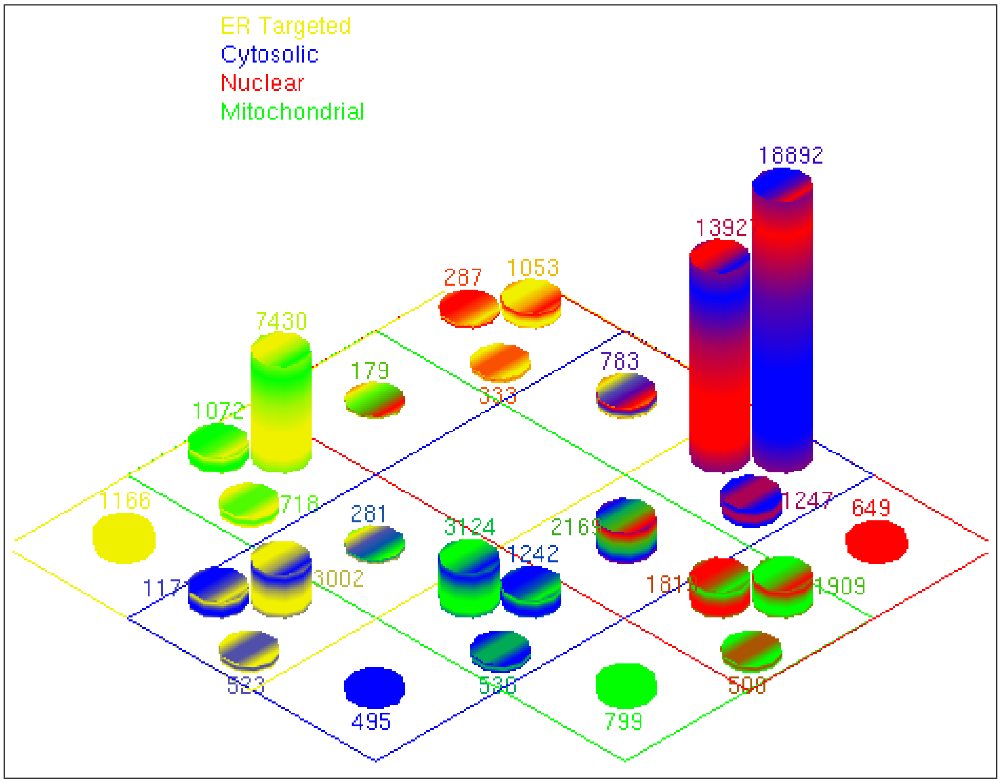


Figure 3.2: Color-coded Venn diagram for human proteome subcellular localization distribution. Protein subcellular localization distribution is represented with twenty-six columns over nuclear (red), cytosolic (blue), mitochondrial (green), and ER targeted (yellow) subcellular localizations. Thickness of the colored bands indicates the prediction votes such that thinner band is for two votes and thicker one is for three votes. The number of sequences is indicated for each column.

Table 3.2: Evaluation of subcellular localization tools on CYGD dataset with 4 692 yeast proteins.

Tool	Coverage	Best Case Accuracy (Number-Percent)	Worst Case Accuracy (Number-percent)
P2SL	4690	3904 - 83.24	3052 - 65.07
PA-SUB	3366	2863 - 85.06	1625 - 48.28
PSORTII	4692	4236 - 90.28	3445 - 73.42
pTARGET	4692	2729 - 58.16	1263 - 26.92
TargetP	4690	3711 - 79.13	2929 - 62.45

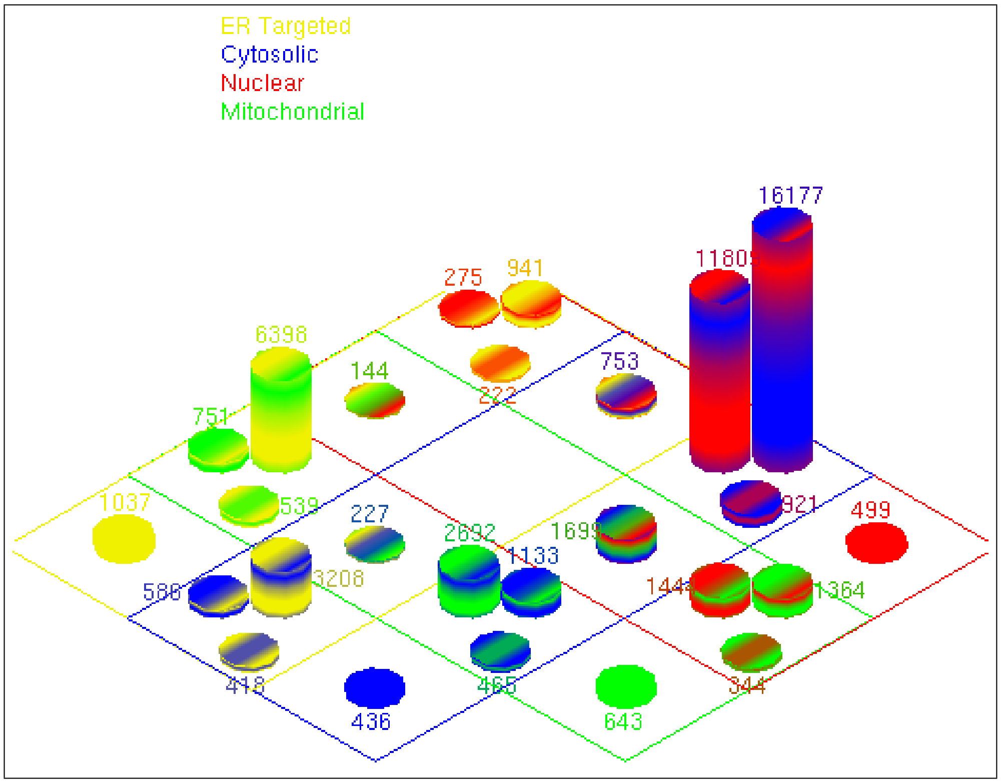


Figure 3.3: Color-coded Venn diagram for mouse proteome subcellular localization distribution. Protein subcellular localization distribution is represented with twenty-six columns over nuclear (red), cytosolic (blue), mitochondrial (green), and ER targeted (yellow) subcellular localizations. Thickness of the colored bands indicates the prediction votes such that thinner band is for two votes and thicker one is for three votes. The number of sequences is indicated for each column.

Table 3.3: Evaluation of subcellular localization tools on HPRD dataset with 11 557 yeast proteins.

Tool	Coverage	Best Case Accuracy (Number-Percent)	Worst Case Accuracy (Number-percent)
P2SL	11550	9429 - 81.64	7755 - 67.14
PA-SUB	9327	7286 -78.12	4919 - 52.74
PSORTII	11557	8539 - 73.89	6873 - 59.47
pTARGET	11557	7447 - 64.44	4993 - 43.20
TargetP	10732	8389 - 78.17	6938 - 64.65

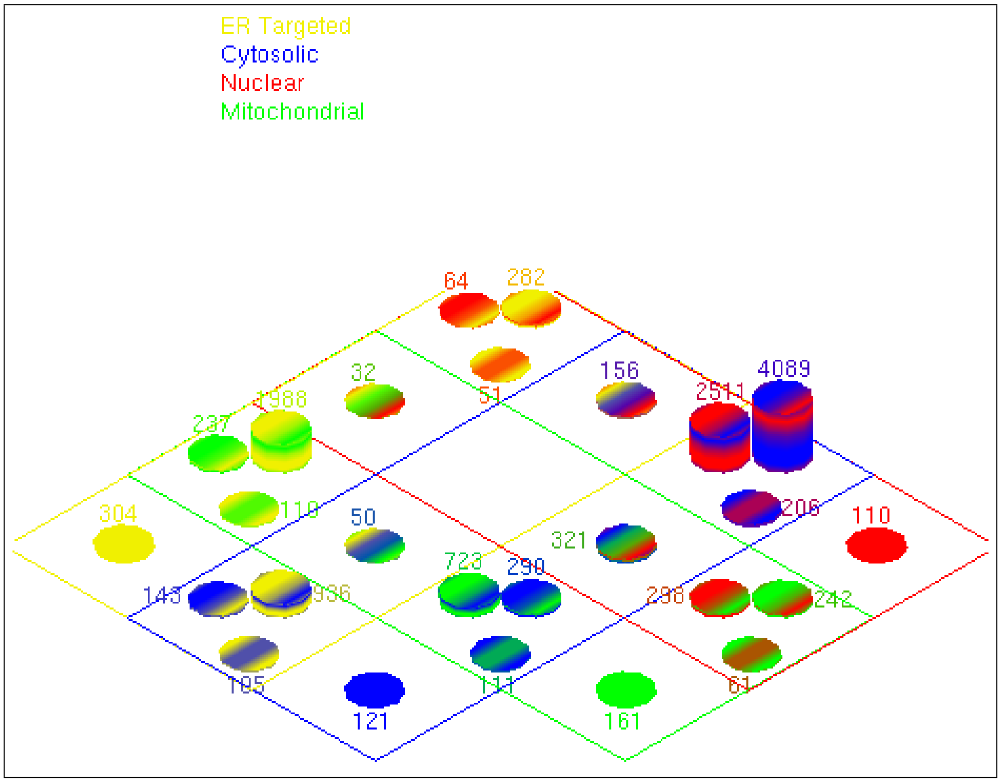


Figure 3.4: Color-coded Venn diagram for rat proteome subcellular localization distribution. Protein subcellular localization distribution is represented with twenty-six columns over nuclear (red), cytosolic (blue), mitochondrial (green), and ER targeted (yellow) subcellular localizations. Thickness of the colored bands indicates the prediction votes such that thinner band is for two votes and thicker one is for three votes. The number of sequences is indicated for each column.

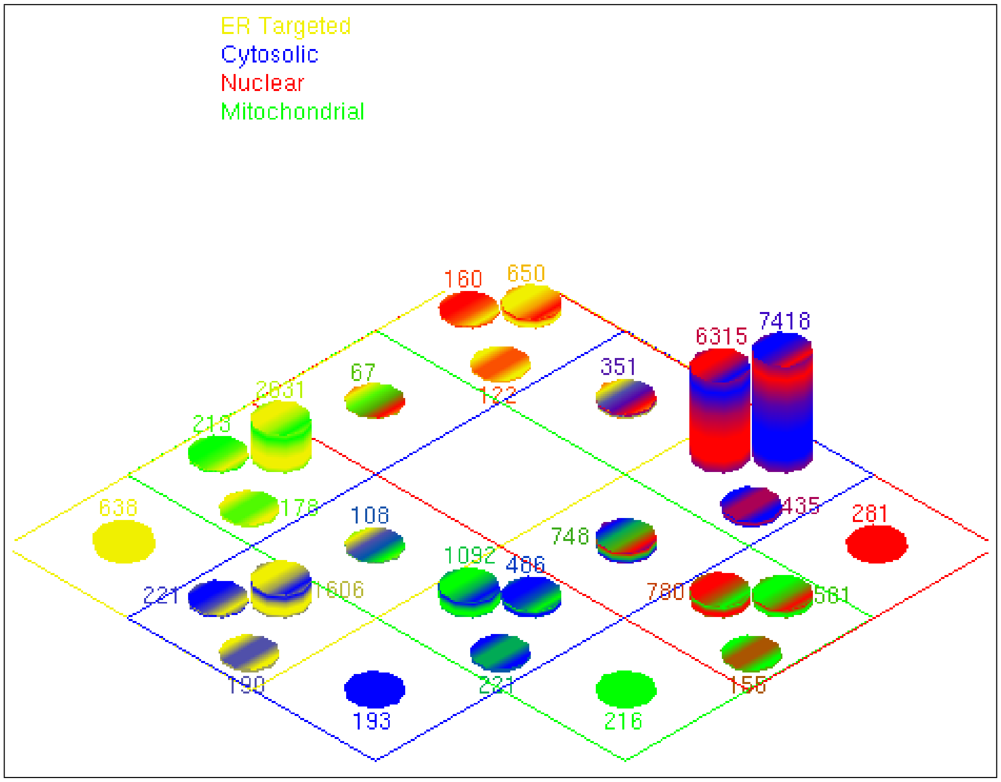


Figure 3.5: Color-coded Venn diagram for fruit fly proteome subcellular localization distribution. Protein subcellular localization distribution is represented with twenty-six columns over nuclear (red), cytosolic (blue), mitochondrial (green), and ER targeted (yellow) subcellular localizations. Thickness of the colored bands indicates the prediction votes such that thinner band is for two votes and thicker one is for three votes. The number of sequences is indicated for each column.

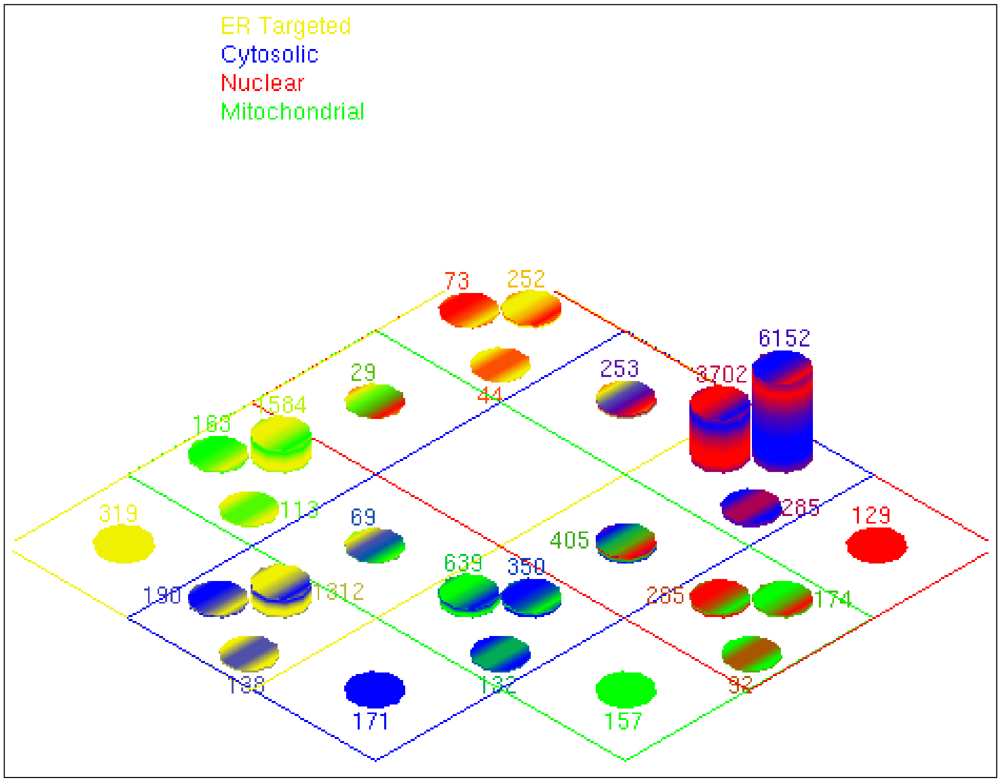


Figure 3.6: Color-coded Venn diagram for zebrafish proteome subcellular localization distribution. Protein subcellular localization distribution is represented with twenty-six columns over nuclear (red), cytosolic (blue), mitochondrial (green), and ER targeted (yellow) subcellular localizations. Thickness of the colored bands indicates the prediction votes such that thinner band is for two votes and thicker one is for three votes. The number of sequences is indicated for each column.

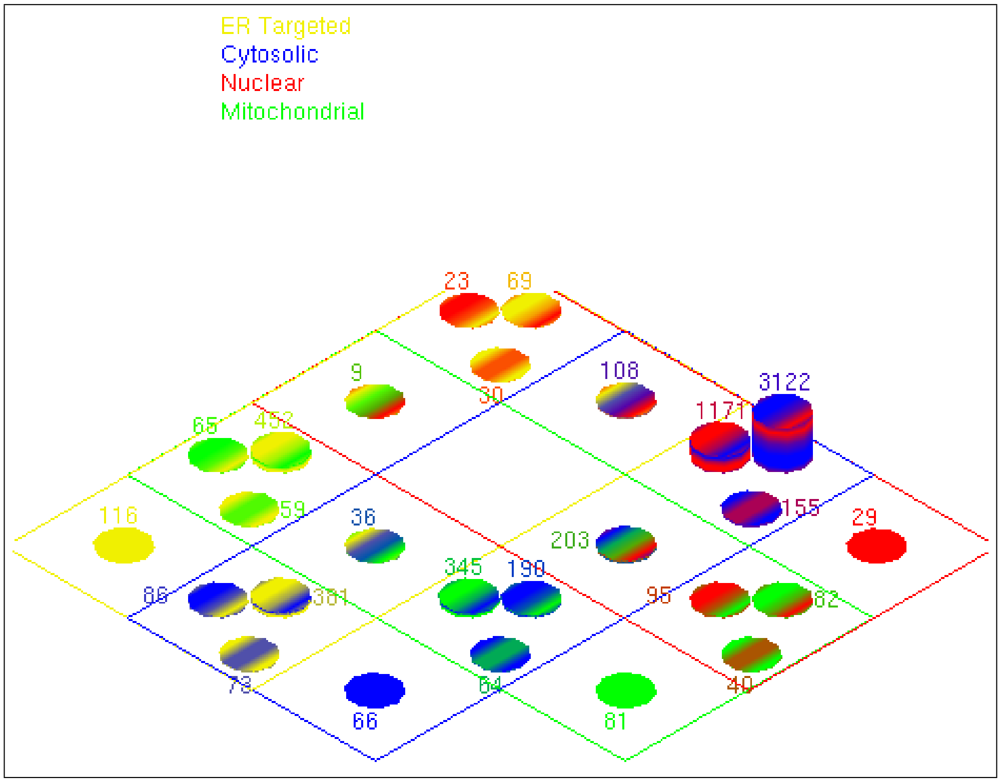


Figure 3.7: Color-coded Venn diagram for yeast proteome subcellular localization distribution. Protein subcellular localization distribution is represented with twenty-six columns over nuclear (red), cytosolic (blue), mitochondrial (green), and ER targeted (yellow) subcellular localizations. Thickness of the colored bands indicates the prediction votes such that thinner band is for two votes and thicker one is for three votes. The number of sequences is indicated for each column.

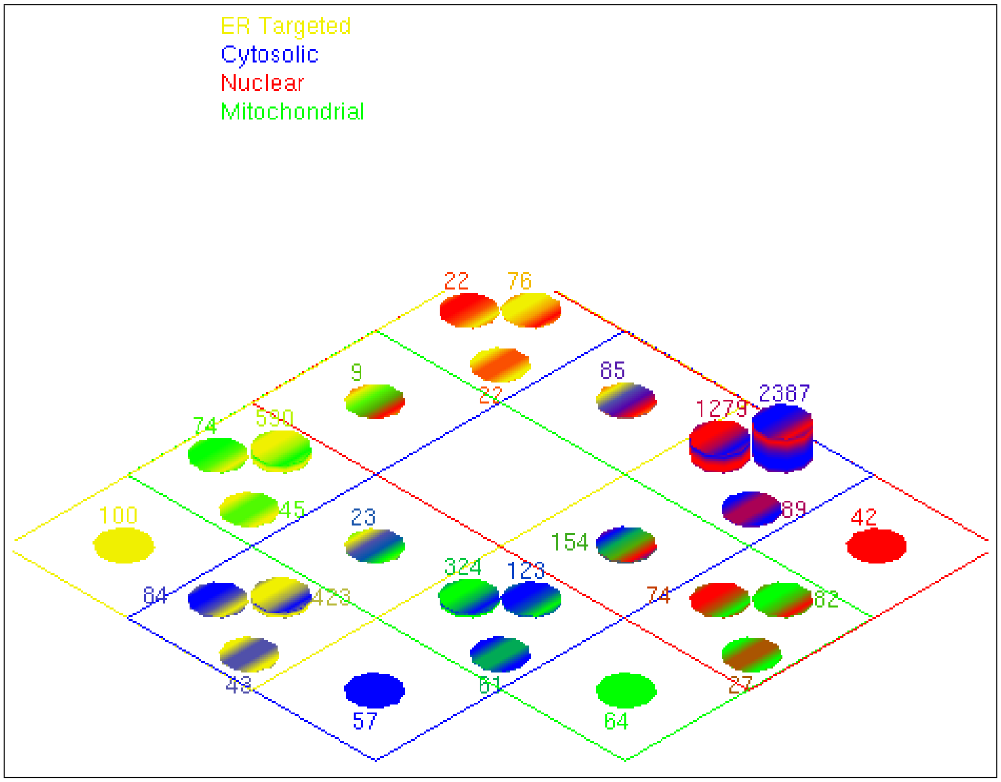


Figure 3.8: Color-coded Venn diagram for frog proteome subcellular localization distribution. Protein subcellular localization distribution is represented with twenty-six columns over nuclear (red), cytosolic (blue), mitochondrial (green), and ER targeted (yellow) subcellular localizations. Thickness of the colored bands indicates the prediction votes such that thinner band is for two votes and thicker one is for three votes. The number of sequences is indicated for each column.

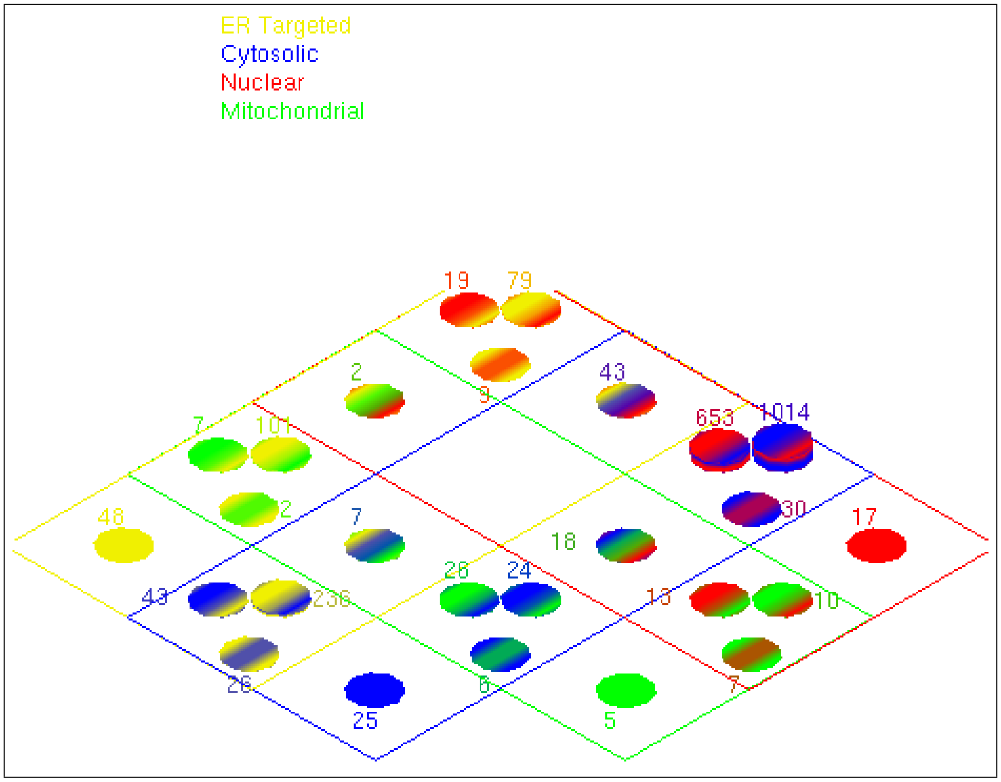


Figure 3.9: Color-coded Venn diagram for slime mold proteome subcellular localization distribution. Protein subcellular localization distribution is represented with twenty-six columns over nuclear (red), cytosolic (blue), mitochondrial (green), and ER targeted (yellow) subcellular localizations. Thickness of the colored bands indicates the prediction votes such that thinner band is for two votes and thicker one is for three votes. The number of sequences is indicated for each column.

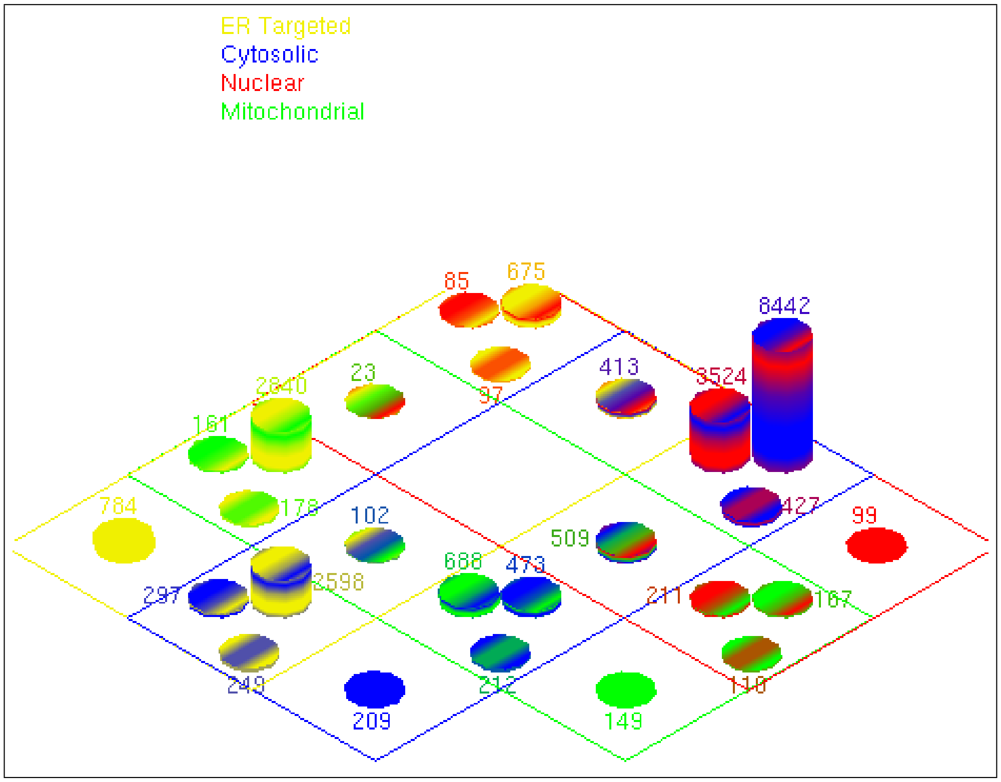


Figure 3.10: Color-coded Venn diagram for worm proteome subcellular localization distribution. Protein subcellular localization distribution is represented with twenty-six columns over nuclear (red), cytosolic (blue), mitochondrial (green), and ER targeted (yellow) subcellular localizations. Thickness of the colored bands indicates the prediction votes such that thinner band is for two votes and thicker one is for three votes. The number of sequences is indicated for each column.

subcellular localization prediction tools on two different test sets. P2SL, among PA-SUB, PSORTII, TargetP, and pTARGET, gave the most accurate predictions 67.14% for the worst case and 81.64% for the best case in HPRD dataset comprised of 11 557 sequences. CYGD dataset consisted of 4 692 sequences yeast *S. cerevisiae* for which PSORTII gave the most accurate results (73.42% for the worst case and 90.28% for the best case). These results may be related with the training sets of the predictors; since PSORTII is trained with a set of yeast sequences and the dominating organism in P2SL training set is is human. The pTARGET tool gave the worst performance on both datasets. This may be due to the multi-categorial nature of the tested data and single category prediction method of the tool. In addition, coverage of PA-SUB is least in both datasets as given in Table 3.2 and Table 3.3.

3.4 DEG

DEG is an online installable searchable and open source analysis suite for Affymetrix HG-U133 Plus 2.0 array. It has two main interfaces, one is for CEL file significantly modulated gene analysis, and other is for the retrieval and merging of previously performed analyses.

3.4.1 Interface

User supplies a .zip achieve of CEL files and a phenodata file. The phenodata is a two column file where the first column is the name of the CEL files and the second column is the sample type of the CEL files. User may specify array quality control plots among `RNA degradation plot`, `pre-normalization boxplot`, `histogram`, `MAplot`, and PLM quality control plots such as `residuals image`, `RLE plot`, `NUSE plot`. After uploading these files, user selects a normalization method among gcrma (gcrma function of R gcrma package) and rma (justrma function of R affy package) and the files specified in the phenodata first column and existing in the CEL archive are normalized with the selected method. User may download the normalized comma separated values file and post-normalization boxplot. After normalization, the user is fronted with a set of t-test analysis options interface where one can specify `equal/unequal variance`, `unpaired`, `two-tailed` t-test parameters along with all possible t-test pair combinations. The user may also filter the expression values that are all below the specified `expression value limit` value. Methods among `BH`, `BY`, `Bonferroni`, `Hochberg`, `Holm`, `SidakSD`, and `SidakSS` are selectable for multiple hypothesis correction procedure. After the analysis, the annotated files are downloadable; thus selected ones are loaded into the database. Once loaded into the database, the user is fronted with an analysis number for future retrieving and merging of the information. The information retrieval and merging interfaces refers to the already existing data in the database. The user may then select among Gene Symbol, GenBank Accession Number, Chromosomal Location, Chromosome, Entrez Gene Id, EC Id, GO, Cytogenetic Maps, OMIM Id, KEGG Pathway, PubMed Id, RefSeq Id, and UniGene Cluster Id annotation fields. Gene Symbol or probe id based, FDR restricted analysis results are fronted within a table like structure.

3.4.2 Case Study Using DEG Interface

We applied our tool on an experimental data obtained from Selenium deficiency induced oxidative stress on HCC derived parental HepG2 cells and HBV-transfected 2.2.15 clone of HepG2 cells, designated as HepG2-2.2.15. HepG2-2.2.15 cell line has been produced by stably transfected HepG2 cells with four tandem copies of the HBV genome. HepG2 and HepG2-2.2.15 cells were cultivated in selenium adequate and selenium deficient medium for 3 days in plates and as duplicates in different times. Under selenium deficient conditions the responses of the two isogenic cell lines were completely different. Parental HepG2 cells were dying due to oxidative stress, while HBV-positive HepG2-2.2.15 cells were still alive under selenium deficient conditions. Cells were collected from 5 plates and RNA extracted and pooled in order to analyze differential gene expression on Affymetrix platform in day 1-2-3.

We constructed an analysis approach considering selenium treatment existence without considering it as time course data. Hence, the phenotype of selenium adequate HepG2-2.2.15 cell lines is `HP` and selenium deficient ones are `HN`. Similarly, phenotype of selenium adequate HBV-transfected 2.2.15 clone of HepG2 cell lines is `2P` and selenium deficient ones are `2N`. We sequentially applied quality control, normalization, t-test analysis on this data. Finally, we uploaded the data in the database for further analysis in a later time.

3.4.2.1 Quality Control Plots

We plotted and analyzed pre-normalization boxplot, histogram, MAplot, RNA degradation plot, PLM residuals image, PLM RLE plot, and PLM NUSE plot. Boxplot is shown in Figure 3.13, histogram is shown in Figure 3.12, MAplot is shown in Figure 3.14, RNA degradation plot is shown in Figure 3.11, PLM residuals image is shown in Figure 3.15, PLM RLE plot is shown in Figure 3.16, and PLM NUSE plot is shown in Figure 3.17.

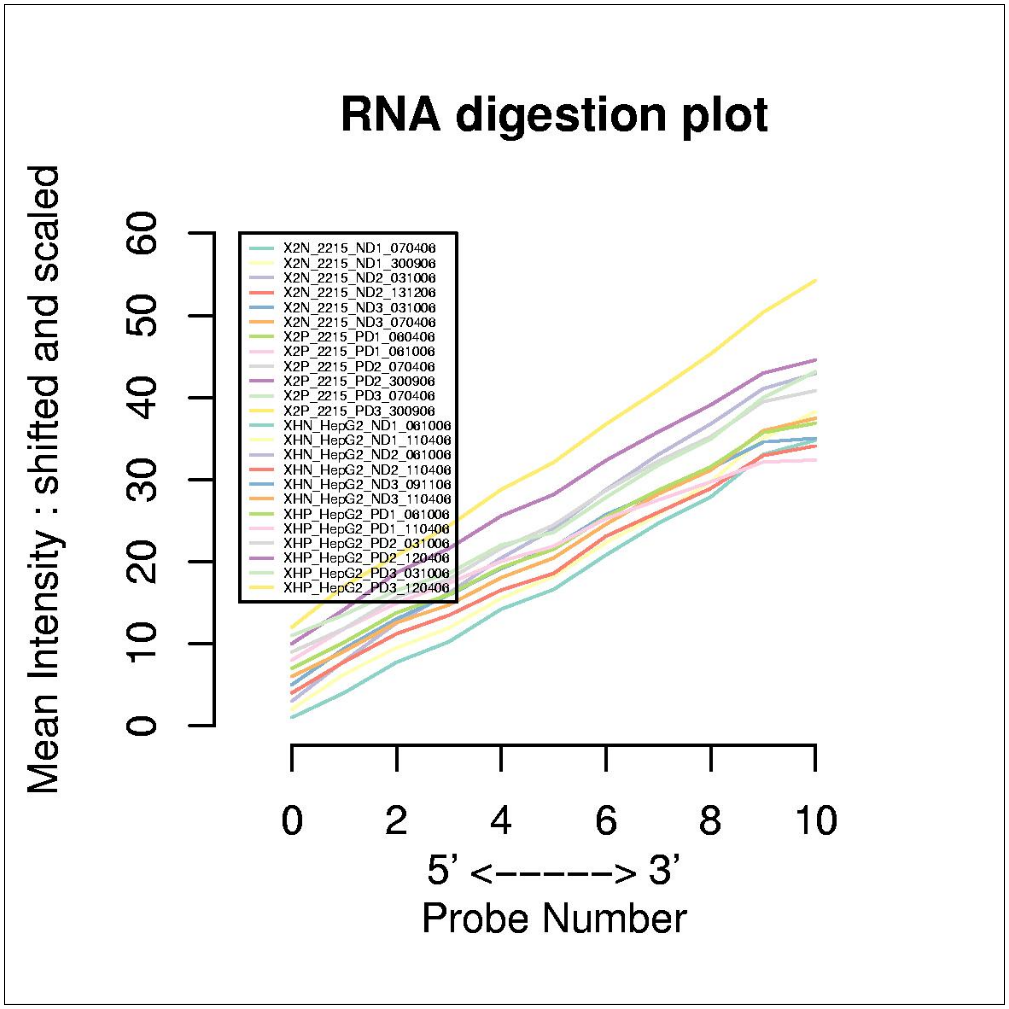


Figure 3.11: RNA degradation plot. Individual probes in each probe set are ordered by location relative to the 5' end of the targeted mRNA molecule. We also know that RNA degradation typically starts at the 5' end, so we would expect probe intensities to be lower near the 5' end than near the 3' end. The ratios should differ for each chip type; we should suspect RNA degradation if slopes are greater than three for HG-U133 Plus 2.0 arrays [15].

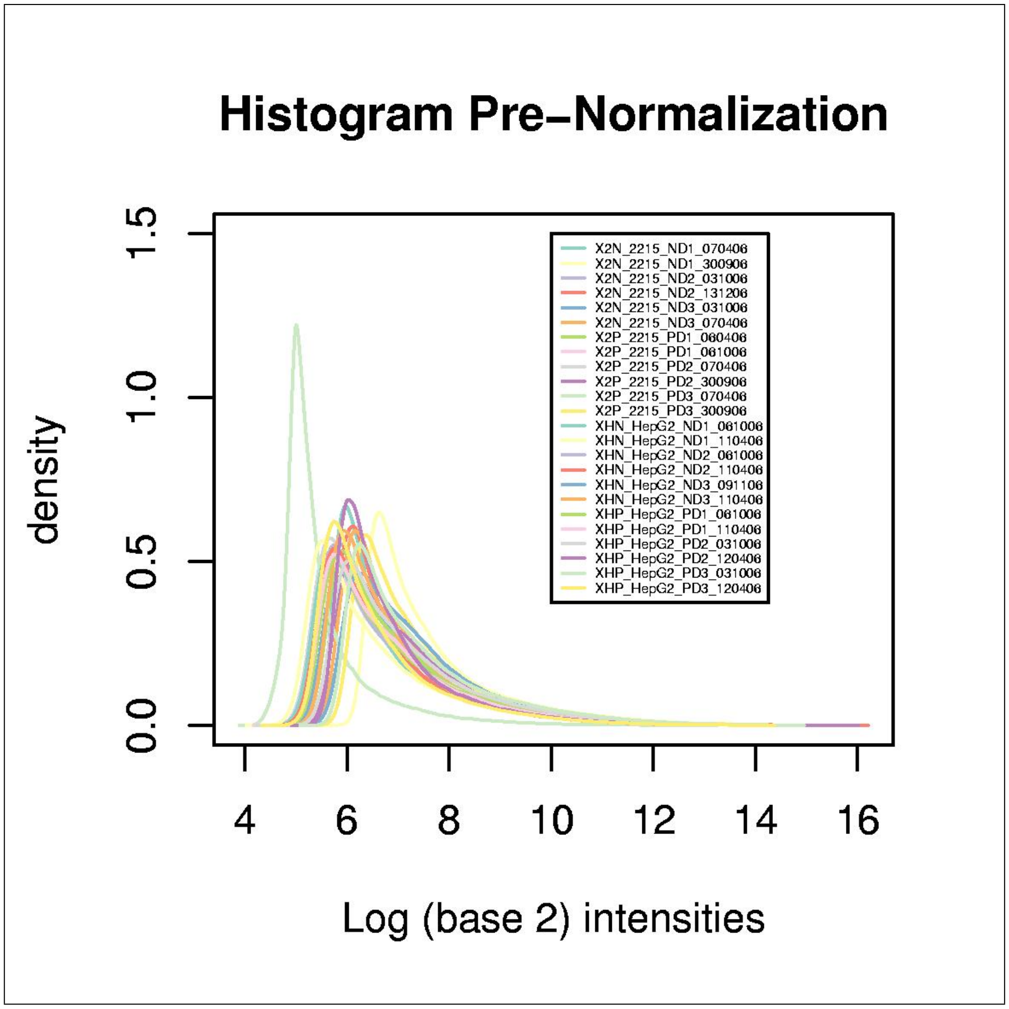


Figure 3.12: Histogram plot. Histograms is a good visualization tool for the identification of saturation, which can be seen as an additional peak at the highest log intensity in the plot.

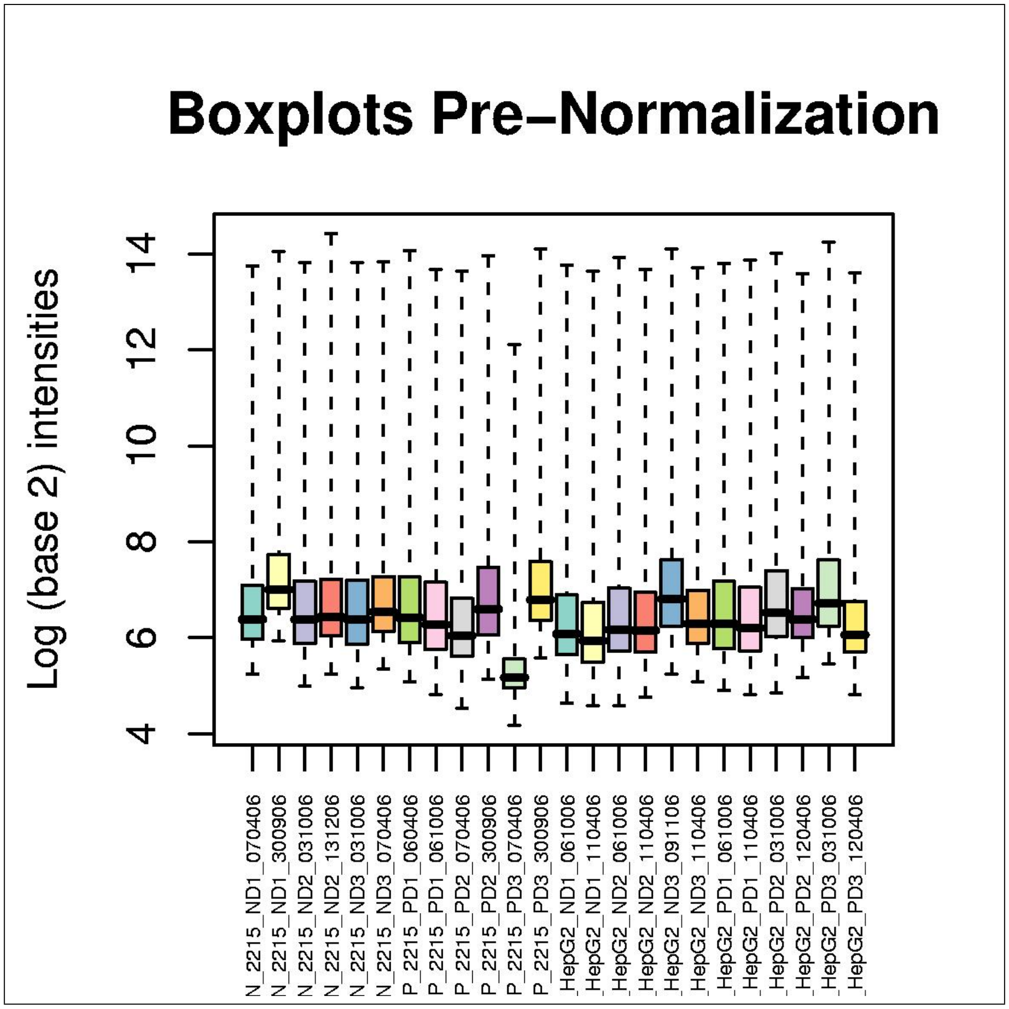


Figure 3.13: Pre-normalization boxplot. Box plot is also a good visualization tool for analyzing the overall intensities of all probes across the array. The box is drawn from the 25th and 75th percentiles in the distribution of intensities. The median, or 50th percentile, is drawn inside the box. The whiskers describe the spread of the data.

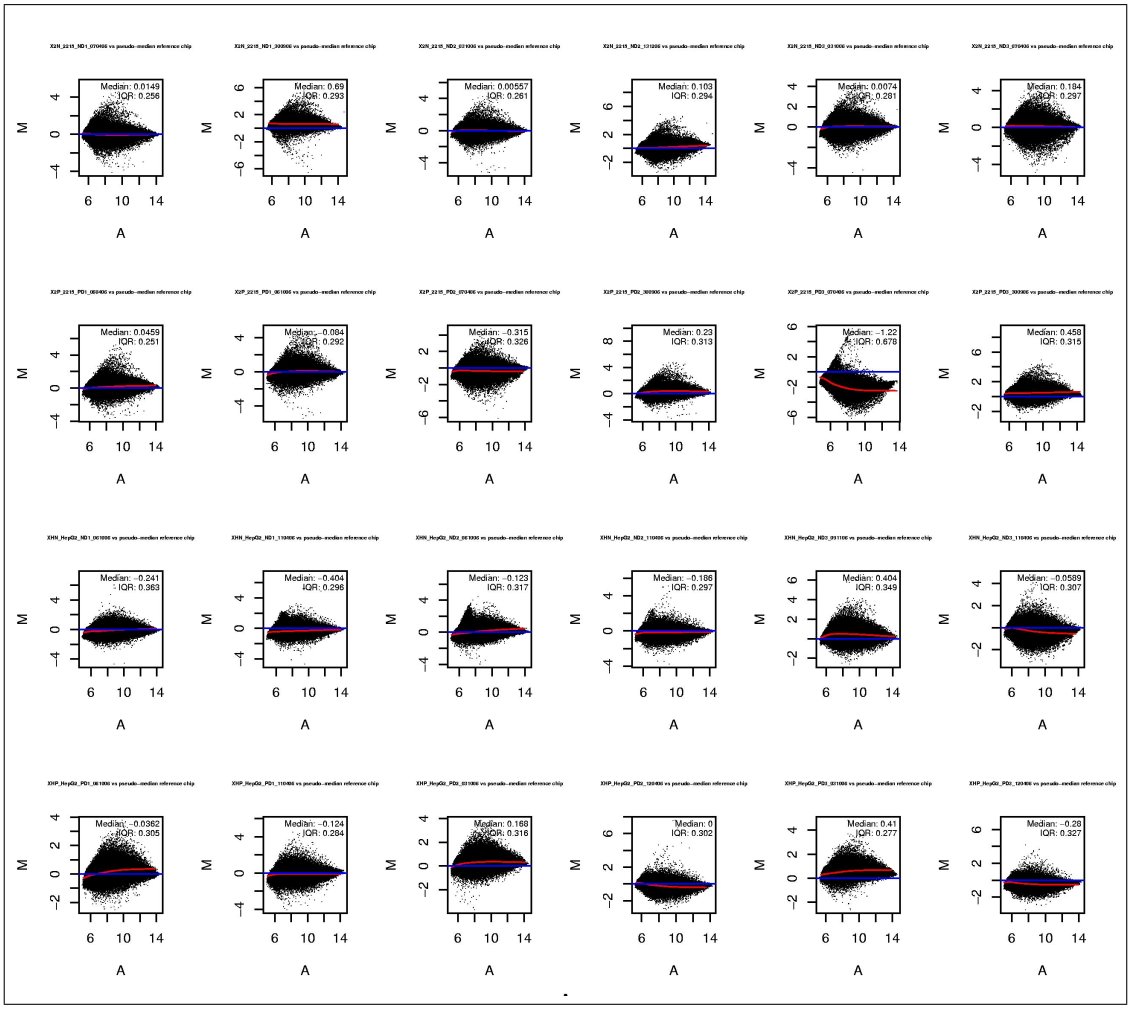


Figure 3.14: M versus A plot (MAplot). An MAplot is a scatter plot used to compare two arrays. The y-axis is the log-fold change and the x-axis is the average log intensity between the two arrays. Each array is compared to a pseudo-reference array. The reference array in the following graphs is the median intensities across all arrays. Again, the expectation is a random scatter plot, centered about the zero horizontal line. Loess curve fitted to the scatter plot, indicated with red, summarizes the nonlinearities. Oscillating loess smoothers indicate quality problems.

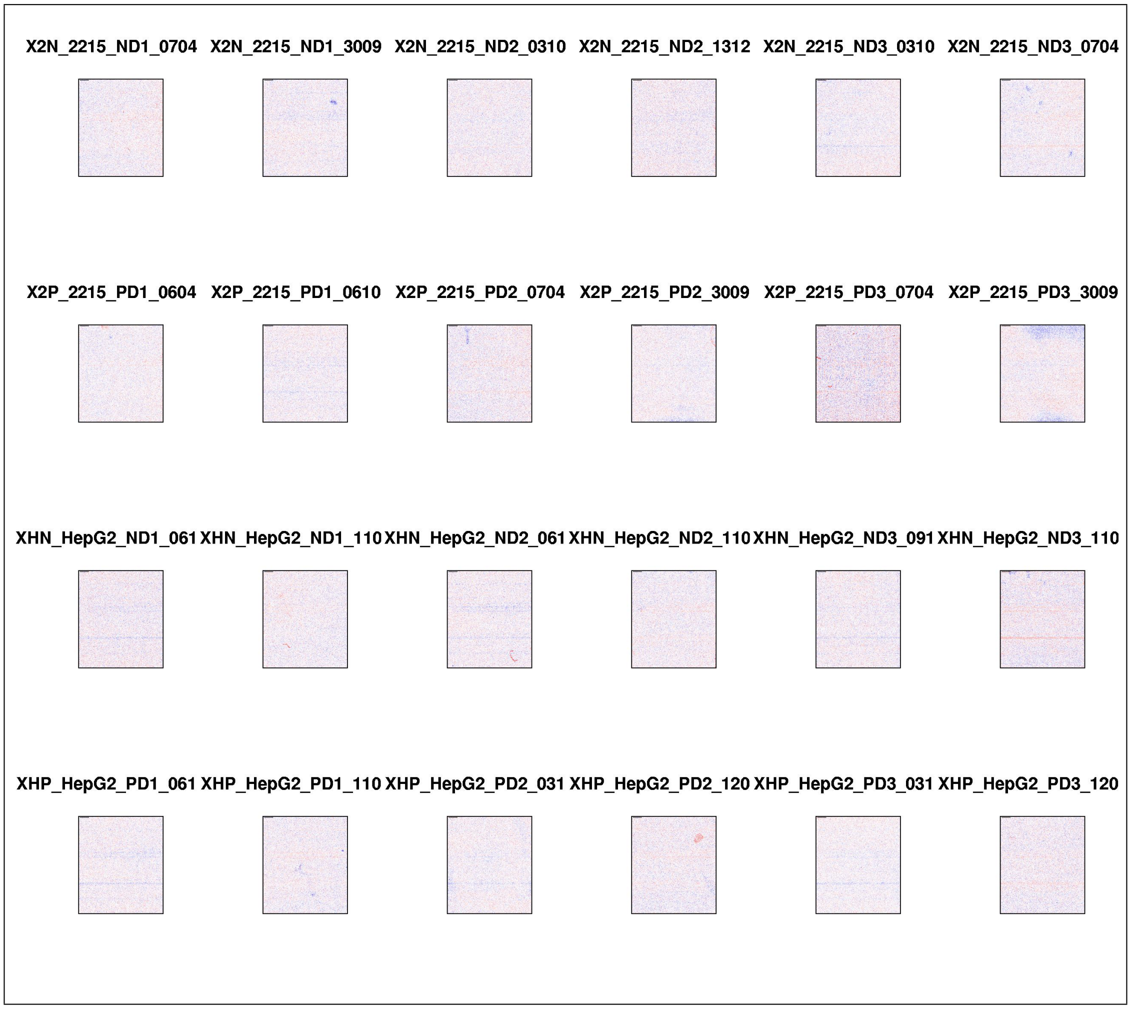


Figure 3.15: PLM residuals image. Negative residuals are colored blue and positive residuals are colored red. Intensities indicated the strength of the signal.

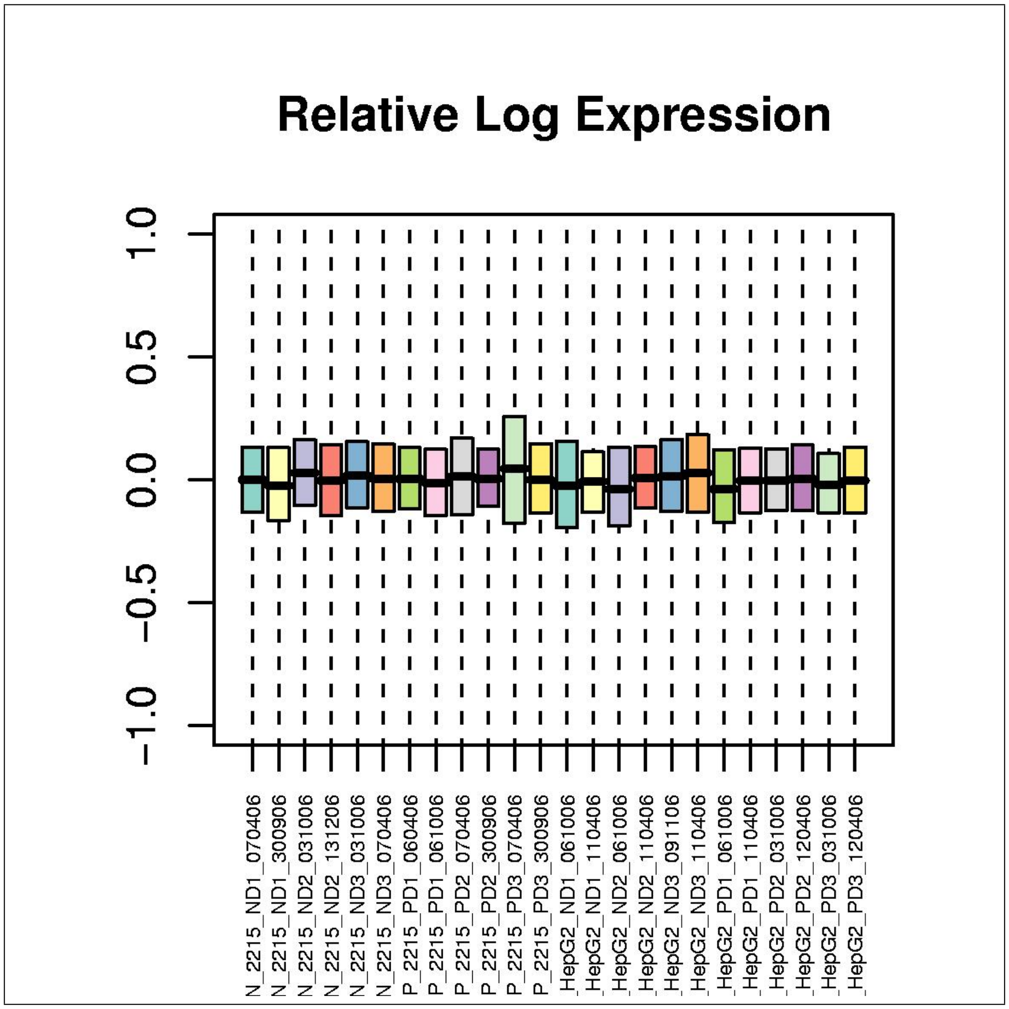


Figure 3.16: PLM RLE plot. RLEs for each probe represent deviation of the probe from the median value of that probe across arrays. This quality assessment is dependent on the assumption that measured intensities are expressed at similar levels across the arrays. The relative logs are displayed as box plots. The expectation is that the relative log expressions should be evenly distributed around zero within each array. In addition, if one or more arrays have box plots that are much larger than the other arrays, then these arrays tend to have more outliers than the other arrays.

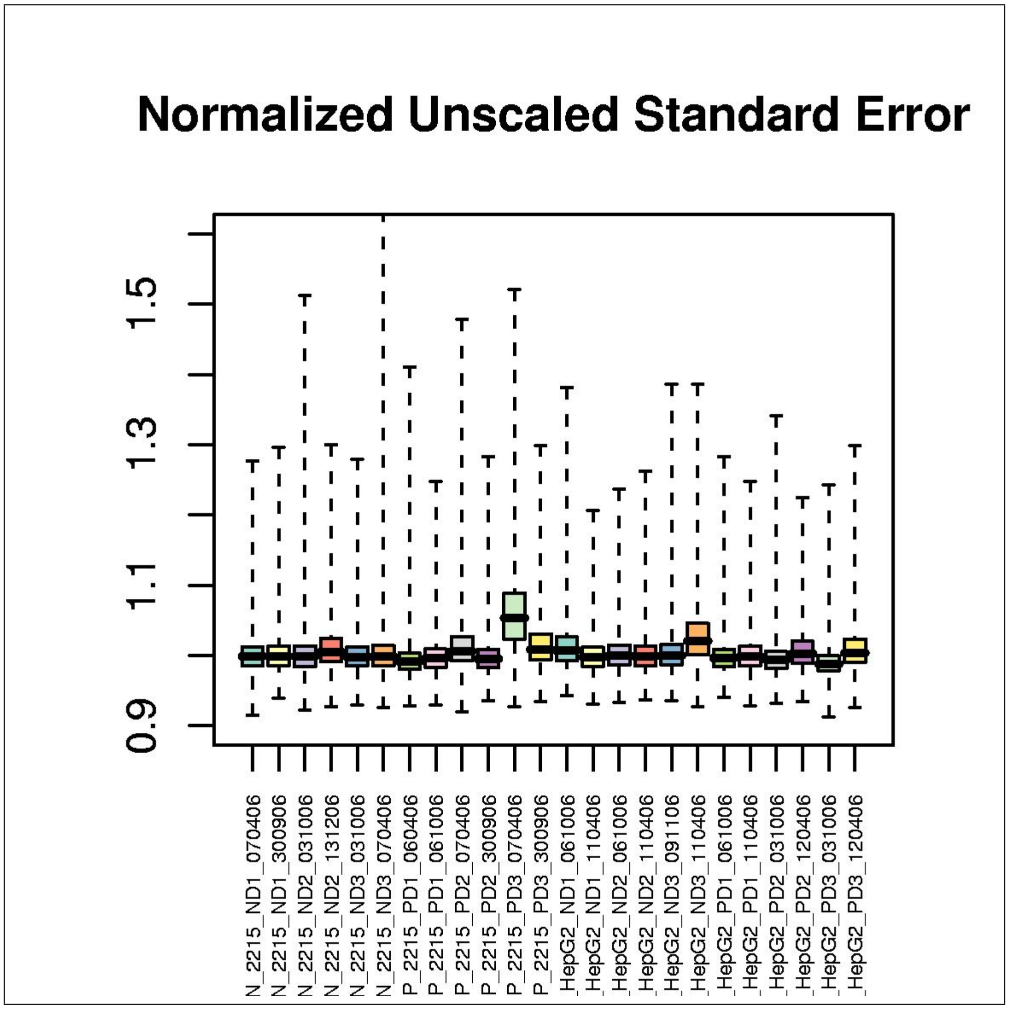


Figure 3.17: PLM NUSE plot. NUSEs represent the standard error between probe intensities within a probe set on a specific array. These errors are normalized by dividing all values of a particular probe set by the median standard error for that probe set across arrays. The expected distribution of NUSEs within an array is centered around one. A higher value indicates that the array has more variance for that probe set than the other arrays.

3.4.2.2 Normalization

Normalized probe values are downloadable for selected normalization method of either rma or gcrma in the web interface. Additionally, post-normalization boxplots as shown in Figure 3.18 are also provided.

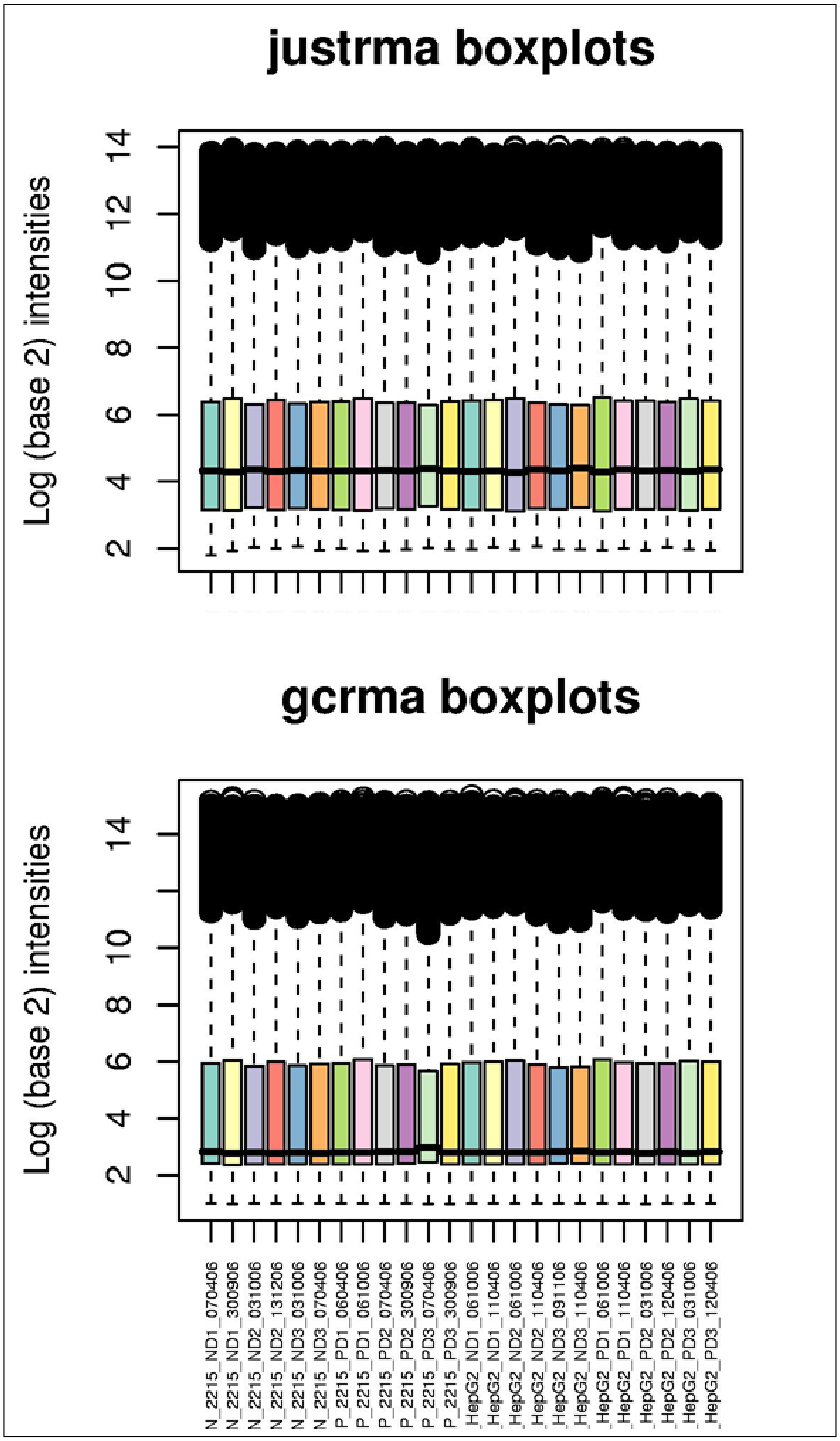


Figure 3.18: Post-normalization boxplots. gcrma gives slightly decreased normalization values; the median of rma is 4.2 and the median of gcrma is 2.8.

Table 3.4: Significant probe numbers of rma & gcrma normalization methods after t-test analysis, respectively.

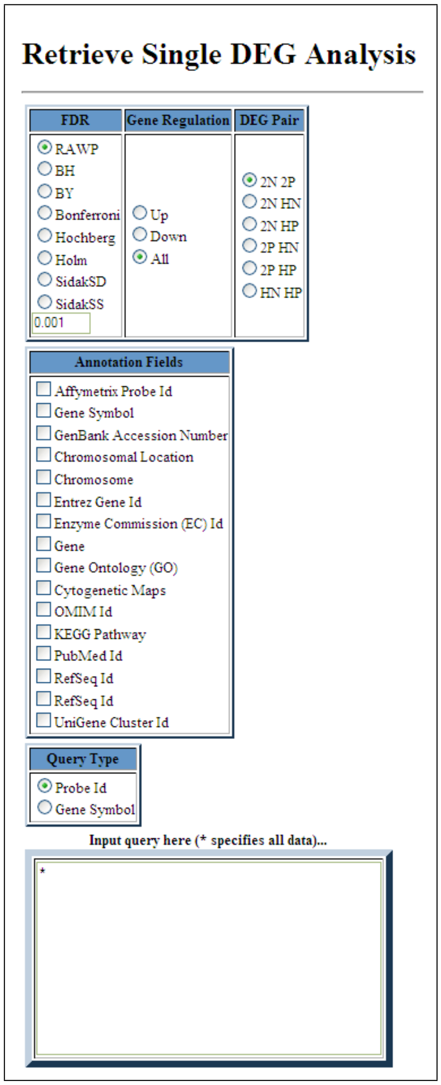
p-value	rawp	BH	BY	Bonferroni	Hochberg	Holm	SidakSD	SidakSS
0.01	3422 - 3054	665 - 575	169 - 161	65 - 61	65 - 61	65 - 61	65 - 62	65 - 62
0.02	4610 - 4179	1124 - 941	256 - 229	79 - 81	79 - 81	79 - 81	79 - 81	79 - 81
0.03	5504 - 5022	1511 - 1250	317 - 271	91 - 94	91 - 95	91 - 95	91 - 95	91 - 95
0.04	6263 - 5719	1843 - 1548	356 - 310	105 - 105	106 - 105	106 - 105	106 - 108	106 - 108
0.05	6895 - 6346	2140 - 1770	420 - 356	113 - 116	113 - 116	113 - 116	114 - 118	114 - 117
0.06	7477 - 6925	2394 - 1998	467 - 386	121 - 123	121 - 123	121 - 123	124 - 124	124 - 123
0.07	8017 - 7417	2684 - 2196	497 - 433	131 - 128	131 - 128	131 - 128	133 - 130	133 - 130
0.08	8535 - 7926	2967 - 2527	536 - 484	138 - 134	138 - 134	138 - 134	138 - 137	138 - 136
0.09	8999 - 8478	3226 - 2766	588 - 519	143 - 139	143 - 140	143 - 140	144 - 140	144 - 140
0.1	9395 - 8946	3514 - 2982	605 - 552	145 - 142	145 - 142	145 - 142	147 - 145	147 - 144
0.11	9789 - 9414	3781 - 3196	665 - 577	148 - 150	148 - 150	148 - 150	154 - 153	153 - 153
0.12	10200 - 9898	4072 - 3465	735 - 611	157 - 154	157 - 155	157 - 155	161 - 156	160 - 155
0.13	10593 - 10338	4289 - 3678	770 - 654	161 - 156	161 - 156	161 - 156	163 - 157	163 - 157
0.14	10989 - 10785	4525 - 3913	819 - 679	163 - 157	163 - 158	163 - 158	169 - 161	169 - 161
0.15	11366 - 11256	4822 - 4136	844 - 704	169 - 161	169 - 161	169 - 161	173 - 166	173 - 165
0.16	11716 - 11686	5042 - 4325	891 - 751	172 - 164	172 - 165	172 - 165	178 - 169	177 - 169
0.17	12056 - 12173	5263 - 4597	933 - 788	173 - 169	174 - 169	174 - 169	181 - 170	181 - 170
0.18	12397 - 12609	5518 - 4792	960 - 824	180 - 170	180 - 170	180 - 170	183 - 172	182 - 171
0.19	12730 - 13071	5737 - 5026	1010 - 858	181 - 171	181 - 171	181 - 171	185 - 178	185 - 178
0.2	13045 - 13521	6014 - 5274	1046 - 897	183 - 172	183 - 172	183 - 172	190 - 184	189 - 182
0.21	13338 - 13933	6252 - 5478	1099 - 922	184 - 177	185 - 178	185 - 178	195 - 186	194 - 186
0.22	13652 - 14375	6446 - 5658	1126 - 945	188 - 180	188 - 181	188 - 181	199 - 188	198 - 188
0.23	13956 - 14768	6670 - 5889	1161 - 962	192 - 185	192 - 186	192 - 186	205 - 191	204 - 191
0.24	14292 - 15175	6912 - 6134	1194 - 1008	196 - 187	196 - 187	196 - 187	208 - 195	206 - 194
0.25	14576 - 15635	7132 - 6350	1219 - 1038	199 - 188	199 - 188	199 - 188	214 - 201	214 - 200

3.4.2.3 Significant Genes Extraction

We filtered probes by setting the `Expression Value Limit` to the median values of normalized intensities from post-normalization boxplots. Afterwards, we extracted significant probes by unpaired, unequal variance, two tailed t-test method.

3.4.2.4 Data Retrieval and Merging

If significant probes lists are selected for loading in the database, users may extract the significantly regulated probes/gene symbols via `Retrieve Interface` as shown in Figure 3.19 and `Merge Interface` as shown in Figure 3.20. The number of significant probes for rawp and multiple hypothesis correction methods could also be observed from the interface and as given in Table 3.4.



Retrieve Single DEG Analysis

FDR	Gene Regulation	DEG Pair
● RAWP ○ BH ○ BY ○ Bonferroni ○ Hochberg ○ Holm ○ SidakSD ○ SidakSS 0.001	○ Up ○ Down ● All	● 2N 2P ○ 2N HN ○ 2N HP ○ 2P HN ○ 2P HP ○ HN HP

Annotation Fields

□ Affymetrix Probe Id
□ Gene Symbol
□ GenBank Accession Number
□ Chromosomal Location
□ Chromosome
□ Entrez Gene Id
□ Enzyme Commission (EC) Id
□ Gene
□ Gene Ontology (GO)
□ Cytogenetic Maps
□ OMIM Id
□ KEGG Pathway
□ PubMed Id
□ RefSeq Id
□ RefSeq Id
□ UniGene Cluster Id

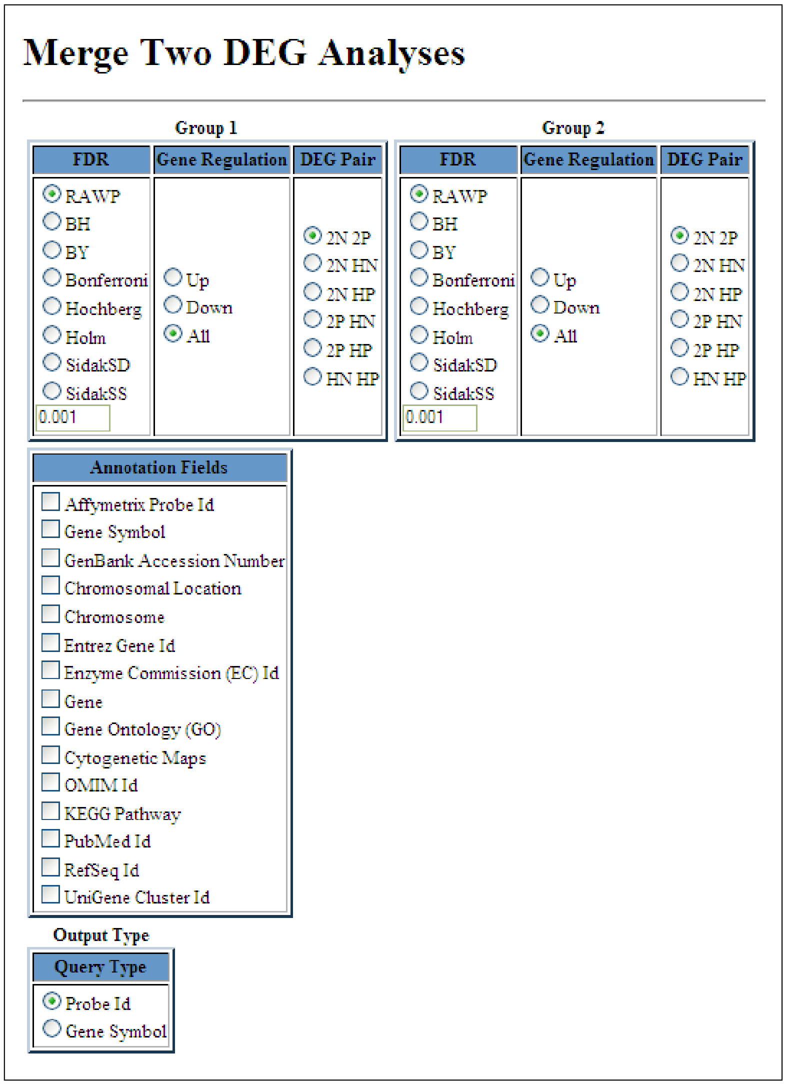
Query Type

● Probe Id
○ Gene Symbol

Input query here (* specifies all data)...

```
*
```

Figure 3.19: DEG retrieve interface for significant probes/genes upon statistical analysis.



Merge Two DEG Analyses

Group 1

FDR	Gene Regulation	DEG Pair
● RAWP ○ BH ○ BY ○ Bonferroni ○ Hochberg ○ Holm ○ SidakSD ○ SidakSS 0.001	○ Up ○ Down ● All	● 2N 2P ○ 2N HN ○ 2N HP ○ 2P HN ○ 2P HP ○ HN HP

Group 2

FDR	Gene Regulation	DEG Pair
● RAWP ○ BH ○ BY ○ Bonferroni ○ Hochberg ○ Holm ○ SidakSD ○ SidakSS 0.001	○ Up ○ Down ● All	● 2N 2P ○ 2N HN ○ 2N HP ○ 2P HN ○ 2P HP ○ HN HP

Annotation Fields
□ Affymetrix Probe Id □ Gene Symbol □ GenBank Accession Number □ Chromosomal Location □ Chromosome □ Entrez Gene Id □ Enzyme Commission (EC) Id □ Gene □ Gene Ontology (GO) □ Cytogenetic Maps □ OMIM Id □ KEGG Pathway □ PubMed Id □ RefSeq Id □ UniGene Cluster Id

Output Type

Query Type
● Probe Id ○ Gene Symbol

Figure 3.20: DEG merge interface for significant probes/genes.

Chapter 4

Conclusions and Future Work

As a result of accumulating genome and proteome data, computational analysis is irreplaceable in molecular biology today. In this study, we analyzed proteome-wide protein subcellular localization and also developed an integrated microarray gene expression data analysis, visualization, and retrieval tool. Protein subcellular localization is important for elucidating protein function and microarray gene expression data enables monitoring of the whole transcripteome simultaneously. Both are important since eukaryotic cells are divided into distinct compartments; for proper functioning of the cell, cellular components should reside in their appropriate locations together with their appropriate partners simultaneously.

This research is initially focused on representation and analysis of proteome wide subcellular localization information with a system called MEP2SL. In the MEP2SL system, using a hybrid machine learning tool called P2SL, we predicted proteome-wide subcellular localizations of nine eukaryotic model organisms including human, mouse, rat, fruit fly, zebrafish, yeast, frog, slime mold, and worm and represented them with their known experimental subcellular localizations from UniProt Knowledgebase.

The online interface of the MEP2SL system enables partial or full downloading of the predicted localization data for further computational analysis. It also provides various query options including `keyword`, `id`, `localization type` and `localization compartment`, and `sequence` queries. The resulting matches for each of these queries are represented with a table structure providing UniRef100 Id, predicted localization, and description (significance and bit score together with the pairwise alignment results in the case of `sequence` query). Each table entry is enabled to give a `details page` which presents a BLASTp utility for finding homologous sequences in the NCBI database and UniRef100 database link for further annotations together with predicted localization distribution possibilities and known experimental localizations from UniProt Knowledgebase.

To validate the prediction method used in the MEP2SL system, we analyzed our prediction in two different datasets from yeast and human and compared the accuracy of three more multi-compartmental prediction tools including PA-SUB, PSORTII, and TargetP and single-compartment tool, pTARGET with P2SL. Our accuracy criteria of `best case accuracy` and `worst case accuracy` serve as an upper and lower bound for the actual accuracy of a prediction tool. For ranking the accuracy of the tools, we use the mean of the worst and best case accuracies. In the yeast dataset from CYGD, PSORTII gave apparently the most accurate results. This result is not surprising since the training set of the PSORTII system is consisted of only yeast sequences. On the same dataset, P2SL gave the second higher odds without using any yeast sequences in its training set. In the HPRD dataset, P2SL had the most accurate predictions. TargetP followed the P2SL prediction accuracy with 3-4 percent decrease and the predictions of P2SL and TargetP systems often correlated. PA-SUB had a significantly decreased coverage compared with the other tools. pTARGET had apparently the worst results which may be due to the multi-compartmental nature of the datasets and single localization predictions of itself. As conclusion, the evaluation of the accuracy of prediction tools is not an easy work since there are many factors that may affect the results such as the prediction method of the systems, number of compartments predicted on, and the training sets used. In our evaluation criterion, category mapping used to label compartments of the datasets and prediction results of the systems may significantly change the results. However, we think the approach we used can give a rough estimate about the characteristics of the tools and P2SL has not failed this process.

We also compared the proteome wide subcellular localizations with high throughput localization experiments conducted in yeast [39], [24], [20]. In these experiments, the dominating compartments are cytosol, nucleus, and ER respectively [22]. We have these three compartments as the dominating ones in our prediction systems too. These results may be observed more clearly from Figure 3.1. From the same figure we can propose a likely conservation of subcellular localization among organisms. In addition, we confirmed that proteins are not single site acting molecules. Hence, with the affirmed performance of the MEP2SL system with experimental data and comparison with other prediction tools, and the comprehensive web interface it provides, we propose MEP2SL as a reference source for proteome wide subcellular localization prediction.

As future directions, the MEP2SL system can be extended to include further subcellular compartments such as Golgi apparatus, plasma membrane, peroxisome, and vacuole. Additionally, we can expand the system and add more prediction tools to construct a meta-database system for proteome wide subcellular localization information. This is crucial since different tools have different strengths and weaknesses for different data types. For example, we can trust prediction results of PSORTII system than the other tools for yeast proteins; and P2SL may be a more reliable source for human proteins. Furthermore, proteome wide conservation of protein subcellular localization signals should also be investigated via statistical analysis. This may further add to the exploration exploration of the protein subcellular localization phenomenon.

Second, we focused on enhancing the existing microarray gene expression data analysis tools. We constructed a web installable open source system called DEG for microarray gene expression data analysis and integrated it with a database. In DEG, the user sequentially uploads the CEL files, performs a series of quality control steps before s/he proceeds with the array normalization procedure. After selection of filtering and t-test parameters and multiple hypothesis correction procedures among `BH`, `BY`, `Bonferroni`, `Hochberg`, `Holm`, `SidakSD`, and `SidakSS` are available to choose from. Afterwards, the significantly modulated genes are extracted and they are integrated into database. The user is then fronted with an analysis id to extract further information from the system at a later time. This analysis id is also used in data merging and retrieval interfaces.

By means of DEG, we provide expression array quality control plots, normalization and significant gene extraction interfaces as well as the dynamic interfaces, developed by keeping the data together with data properties in a database. Dynamic interfaces of retrieval and merging refers to existing data in the database. The ability of the user to annotate the differentially expressed probes/gene symbols from multiple sources is an integrated feature of DEG and helps the user summarize the results of a microarray experiment. User may select among Gene Symbol, GenBank Accession Number, Chromosomal Location, Chromosome, Entrez Gene Id, EC Id, GO, Cytogenetic Maps, OMIM Id, KEGG Pathway, PubMed Id, RefSeq Id, and UniGene Cluster Id annotation fields. We aim to expand this list of selections as new identifiers and classifiers emerge. The DEG also is innovative in its structure that Gene Symbol or probe id based queries results in tables in which pre-processed expression data and statistical analysis results are combined and presented to the user for future filtering/sorting. This table can also be downloaded as a plain text file in tabular format. One of the most important features of the DEG is that the user can merge two differentially expressed gene lists originating from two different t-tests to extract the intersecting gene set. This feature may help users to refine their data further and to test the extent multiple experimental gene lists have in common.

An online yet installable tool is beneficial for the research groups who are not willing to submit their data on public analysis servers. Having a permanent data storage capability with data integration into a modular and highly scalable database presented with a comprehensive yet simple user-friendly interface, DEG provides a good starting point for generation of an expandable microarray gene expression data analysis, visualization, and retrieval suite. The integration of a database into an online installable gene expression analysis tool is a unique feature of DEG among other comparable tools in the field.

As future directions, the data merging and retrieval capabilities of DEG may be expanded to allow the processing of more than one analysis, e.g., t-tests. This may allow the interface to analyze time series data as well. Additionally, for graphical comparison of functional groups, GoTools Bioconductor package can be added to the interface.

Bibliography

[1] *CYGD Download site*. ftp://ftpmips.gsf.de/yeast/catalogues/subcellcat/.

[2] *HPRD Download Site*. http://www.hprd.org/download.

[3] *NIH Web Site*. http://www.nih.gov/science/models/.

[4] *PA-SUB Web Site*. http://pasub.cs.ualberta.ca:8080/pa/Subcellular.

[5] *pTARGET Web Site*. http://bioinformatics.albany.edu/ ptarget/.

[6] *TargetP Web Site*. http://www.cbs.dtu.dk/services/TargetP/.

[7] Schaffer AA Zhang J Zhang Z Miller W Lipman DJ. Altschul SF, Madden TL. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402, 1997.

[8] Cetin-Atalay R. Atalay V. Implicit motif distribution based hybrid computational kernel for sequence classification. *Bioinformatics*, 21, 2005.

[9] Ben Bolstad. *affyPLM: Methods for fitting probe-level models*, 2006. R package version 1.10.0.

[10] Micky Del Favero Chiara Romualdi, Nicola Vitulo and Gerolamo Lanfranchi. Midaw: a web tool for statistical analysis of microarray data. *Nucleic Acids Research*, 33 (Web Server issue):W644–W649, 2005.

[11] Chu FW. Ekins R. Microarrays: their origins and applications. *Trends in Biotechnology*, 17:217–218, 1999.

[12] Brunak S von HG. Emanuelsson O, Nielsen H. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *J Mol Biol*, 300:1005–1016, 2000.

[13] Asogawa M. Fujiwara Y. Prediction of subcellular localizations using amino acid composition and order. *Genome Inform*, 12:103–112, 2001.

[14] R. Gentleman, V. Carey, and W. Huber. *genefilter: genefilter: methods for filtering genes from microarray experiments.* R package version 1.12.0.

[15] Huber W Irizarry R Dudoit S. Gentleman R, Carey V. Bioinformatics and computational biology solutions using r and bioconductor. 2005.

[16] Kutay U. Gorlich D. Transport between the cell nucleus and the cytoplasm. *Annu Rev Cell Dev Biol*, 15:607–660, 1999.

[17] Subramaniam S. Guda C. ptarget [corrected] a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics*, 21:3963–3969, 2005.

[18] Kastenmuller G Strack N van HJ Lemer C Richelles J Wodak SJ Garcia-Martinez J Perez-Ortin JE Michael H Kaps A Talla E-Dujon B Andre B Souciet JL De MJ Bon E Gaillardin C Mewes HW. Guldener U, Munsterkotter M. Cygd: the comprehensive yeast genome database. *Nucleic Acids Res*, 33:D364–D368, 2005.

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

[20] Gerke LC Carroll AS Howson RW Weissman JS OShea EK. Huh WK, Falvo JV. Global analysis of protein localization in budding yeast. *Nature*, 425:686–691, 2003.

[21] Rafael A. Irizarry, Laurent Gautier, Benjamin Milo Bolstad, , Crispin Miller with contributions from Magnus Astrand ¡Magnus.Astrand@astrazeneca.com¿, Leslie M. Cope, Robert Gentleman, Jeff Gentry, Conrad Halling, Wolfgang Huber, James MacDonald, Benjamin I. P. Rubinstein, Christopher Workman, and John Zhang. *affy: Methods for Affymetrix Oligonucleotide Arrays*, 2006. R package version 1.12.2.

[22] Simpson JC and Pepperkok R. Localizing the proteome. *Genome Biology*, 4(12):240, 2003.

[23] Michael Acab Marc-Etienne Rousseau1 Byron Kuo1 David Goode2 Dana Aeschliman3 Jenny Bryan3 Lorne A. Babiuk4 Robert E. W. Hancock2 Karsten Hokamp, Fiona M. Roche and Fiona S. L. Brinkman. Arraypipe: a flexible processing pipeline for microarray data. *Nucleic Acids Research*, 32 (Web Server issue):W457–W459, 2004.

[24] Heyman JA Matson S-Heidtman M Piccirillo S Umansky L Drawid A Jansen R Liu Y et al. Kumar A, Agarwal S. Subcellular localization of the yeast proteome. *Genes Dev*, 16:707–219, 2002.

[25] Ting-Yuan Liu, ChenWei Lin, Seth Falcon, Jianhua Zhang, and James W. MacDonald. *hgu133plus2: Affymetrix Human Genome U133 Plus 2.0 Array Annotation Data (hgu133plus2).* R package version 1.14.0.

[26] Greiner R Lu P-Wishart DS Poulin B Anvik J Macdonell C Eisner R. Lu Z, Szafron D. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, 20:547–556, 2004.

[27] Tárraga J. Huerta-Cepas J. Burguet-J. Vaquerizas J.M. Conde L. Minguez P. Vera J. Mukherjee S. Valls J. Pujana M.A.G. Alloza E. Herrero J. Al-Shahrour F. Montaner, D. and J. Dopazo. Next station in microarray data analysis: Gepas. *Nucleic Acids Research*, 34 (Web Server issue):W486–W491, 2006.

[28] Rost B. Nair R. Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins*, 53:917–930, 2003.

[29] Horton P. Nakai K. Psort: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, 24:34–36, 1999.

[30] Erich Neuwirth. *RColorBrewer: ColorBrewer palettes*, 2005. R package version 0.2-3.

[31] Brunak S von HG. Nielsen H, Engelbrecht J. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*, 10:1–6, 1997.

[32] Emanuelsson O. Predicting protein subcellular localisation from amino acid sequence information. *Brief Bioinform*, 3:361–376, 2002.

[33] Amanchy R Kristiansen TZ-Jonnalagadda CK Surendranath V Niranjan V Muthusamy B Gandhi TK Gronborg M Ibarrola N Deshpande N Shanker K Shivashankar HN-Rashmi BP Ramya MA Zhao Z Chandrika KN Padma N Harsha HC Yatish AJ Kavitha MP Menezes M Choudhury DR Suresh S Ghosh N Saravana R Chandran S Krishna S Joy M Anand SK Madavan V Joseph A Wong GW Schiemann WP Constantinescu SN Huang L Khosravi-Far R Steen H Tewari M Ghaffari S Blobe GC Dang CV Garcia JG Pevsner J Jensen ON Roepstorff P Deshpande KS Chinnaiyan AM Hamosh A Chakravarti A Pandey A. Peri S, Navarro JD. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13:2363–2371, 2003.

[34] Katherine S. Pollard, Yongchao Ge, and Sandrine Dudoit. *multtest: Resampling-based multiple hypothesis testing.* R package version 1.12.0.

[35] Sick M Thoppae G-Harshman K Sick B. Psarros M, Heber S. Race: Remote analysis computation for gene expression data. *Nucleic Acids Research*, 33 (Web Server issue):W638–643, 2005.

[36] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[37] Stocker G Sturn A Trajanoski Z. Rainer J, Sanchez-Cabo F. Carmaweb: comprehensive r- and bioconductor-based web service for microarray data analysis. *Nucleic Acids Research*, 34 (Web Server issue):W498–503, 2006.

[38] Holger Schwender. *siggenes: SAM and Efron's empirical Bayes approaches*, 2006. R package version 1.8.0.

[39] Poustka A Pepperkok R Wiemann S. Simpson JC, Wellenreuther R. Systematic subcellular localisation of novel proteins identified by large-scale cdna sequencing. *EMBO Rep*, 1:287–292, 2000.

[40] Colin A. Smith. *annaffy: Annotation tools for Affymetrix biological metadata*, 2006. R package version 1.6.1.

[41] Attwood TK and Parry-Smith DJ. Introduction to bioinformatics. 1999.

[42] Merino-Trigo A Teasdale RD Gleeson PA. van Vliet C, Thomas EC. Intracellular sorting and transport of proteins. *Prog Biophys Mol Biol*, 83(1):1–45, 2003.

[43] Jean(ZHIJIN) Wu and Rafael Irizarry with contributions from James MacDonald Jeff Gentry. *gcrma: Background Adjustment Using Sequence Information*. R package version 2.6.0.

Appendix A

Localization Labeling for Predictor Evaluation

A.1 Localization Labels of CYGD Dataset

Table A.1: CYGD dataset protein subcellular localization labeling.

Subcellular Location	Description	Label
701	extracellular	X
710	cell wall	X
715	cell periphery	X
720	plasma membrane	A
722	integral membrane / endomembranes	A
725	cytoplasm	C
730	cytoskeleton	Y
735	endoplasmic reticulum	E
740	Golgi	G
745	transport vesicles	E
750	nucleus	N
755	mitochondria	M
760	peroxisome	P
765	endosome	A
770	vacuole	L
775	microsomes	A

A.2 Localization Labels of HPRD Dataset

Table A.2: HPRD dataset protein subcellular localization labeling.

Subcellular Location	Label	Subcellular Location	Label
Acrosome	E	Integral to membrane	A
Actin cytoskeleton	Y	Integral to plasma membrane	A
Actin filament	Y	Intermediate filament	Y
Apical membrane	A	Intracellular vesicle	S
Basolateral membrane	A	Kinetochore	C
Caveola	A	Late endosome	A
Cell junction	A	Lysosome	L
Cell surface	X	Microsome	E
Centriole	C	Microtubule	C
Centrosome	C	Mitochondrial intermembrane space	M
Chromosome	N	Mitochondrial matrix	M
Cilium	X	Mitochondrial membrane	M
Clathrin-coated vesicle	A	Mitochondrion	M
Cytoplasm	C	Nuclear matrix	N
Cytoplasmic vesicle	S	Nuclear membrane	A
Cytoskeleton	Y	Nucleolus	N
Cytosol	C	Nucleoplasm	N
Desmosome	A	Nucleus	N
Early endosome	A	Perinuclear region	C
Endoplasmic reticulum	A	Perinuclear vesicle	E
Endoplasmic reticulum lumen	E	Peroxisomal matrix	P
Endoplasmic reticulum membrane	E	Peroxisomal membrane	P
Endosome	A	Peroxisome	P
Extracellular	X	Plasma membrane	A
Extracellular matrix	X	Ribosome	C
Extracellular space	X	Sarcoplasm	E
Golgi apparatus	G	Sarcoplasmic reticulum	E
Golgi lumen	G	Secreted	X
Golgi membrane	G	Secretory vesicle	S
Golgi vesicle	G	Tubulin	C

A.3 Localization Labels of Prediction Tools

Table A.3: Protein subcellular localization labeling of prediction tools.

System	Subcellular Location	Label	Prediction Technique
PSORTII	cytoskeletal	Y	k-nearest neighborhood
	cytoplasmic	C	
	nuclear	N	
	vacuolar	L	
	peroxisomal	P	
	plasma membrane	A	
	mitochondrial	M	
	endoplasmic reticulum	E	
	vesicles of secretory system	S	
	extracellular including cell wall	X	
	Golgi	G	
P2SL	ER-targeted	E	SOM and SVM
	cytosolic	C	
	nuclear	N	
	mitochondrial	M	
PA-SUB	mitochondrion	M	Naïve Bayes classifier
	nucleus	N	
	endoplasmic reticulum	E	
	extracellular	X	
	cytoplasm	C	
	plasma membrane	A	
	golgi	G	
	lysosome	L	
	peroxisome	P	
TargetP	Mitochondrion	M	Neural networks based on N-terminal amino acid sequence
	Secretory pathway	S	
	Other	O	
pTARGET	Mitochondria	M	Score based on protein functional domains
	Nucleus	N	
	Endoplasmic Reticulum	E	
	Extracellular/Secretory	X	
	cytoplasm	C	
	Plasma Membrane	A	
	Golgi	G	
	Lysosomes	L	
	Peroxysomes	P	