8-2003

Cellular Function Prediction for Hypothetical Proteins Using High-Throughput Data

Trupti Subhash Joshi
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a thesis written by Trupti Subhash Joshi entitled "Cellular Function Prediction for Hypothetical Proteins Using High-Throughput Data." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Life Sciences.

Dong Xu, Major Professor

We have read this thesis and recommend its acceptance:

Jeff Becker, Loren Hauser

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a thesis written by Trupti Subhash Joshi entitled "Cellular function prediction for hypothetical proteins using high-throughput data." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Life Sciences.

Dong Xu

Major Professor

We have read this thesis
and recommend its acceptance:

Jeff Becker

Loren Hauser

Accepted for the Council:

Anne Mayhew

Vice Provost and
Dean of Graduate Studies

(Original signatures are on the file with official student records.)

Cellular function prediction for hypothetical proteins using high-throughput data.

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Trupti Joshi

August 2003

This thesis is dedicated to my parents, Subhash Joshi and Anjali Joshi and my husband, Abhijit.

ACKNOWLEDGEMENTS

I would like to thank my husband, Abhijit for his love and support. He is such a loving and caring person. I would like to thank my parents Mr. Subhash Joshi and Mrs. Anjali Joshi for their encouragement and being such wonderful parents. Words are inadequate to express how important their love and encouragement was to help me achieve my goals. I appreciate very much, the support and guidance of my sisters Swati and Smruti and my brothers-in-law Subhash and Yogesh. I wish to thank my in-laws Mr. Madhav Paranjpe and Mrs. Sucheta Paranjpe for their support and encouragement.

I would like to thank my advisor, Dr. Dong Xu, for giving me the opportunity to work on this interesting project and for his valuable mentoring. He has not only been a good teacher but also a good friend. It was a great learning experience while working on this project under his guidance. The training has given me confidence in my expertise and my abilities. Working with him has been a pleasure and a rewarding experience.

I would like to thank Dr. Jeff Becker and Dr. Loren Hauser for willing to serve on my committee and for their valuable guidance and discussions. Many thanks to Dr. Becker for his patient listening and guidance, every time I approached him with my difficulties. I wish to thank Dr. Victor Olman for the valuable discussions. I would like to take this opportunity to extend my gratitude and regards to all my teachers who have played an important role in molding me in my childhood and my adolescence and shown me the right path. I would like to thank all my friends. Special thanks to Yu Chen, it has been a pleasure working with him. I am very thankful to my roommate Sumithra, for her support. I wish to thank Kay Gardner and Gaynelle Russell for their help.

I would like to thank the UT-ORNL Graduate School of Genome Science and Technology for providing me the training opportunity. I thank Oak Ridge National Laboratory for providing me the working environment to carry out my thesis work. I also thank Ceres, Inc. for the collaboration, especially Dr. Nickolai Alexandrov at Ceres, Inc. for helpful discussions.

ABSTRACT

We have developed an integrated probabilistic prediction method, which combines the information from protein-protein interactions, protein complexes, microarray gene-expression profiles and functional annotations for known proteins. Our approach differs from the other approaches to use high-throughput data in a variety of ways. First, we utilize the GO biological process functional annotation in comparison to the MIPS classification followed by others. Second, we incorporate information from multiple sources of high-throughput data, including genetic interactions, to develop a better model for function prediction. By incorporating information from the multiple sources of high-throughput data, we identify the parameters important for protein function prediction. Third, we estimate the probability for the proteins to have a function of interest by designing a new statistical method for function prediction. Fourth, our approach assigns multiple functions to the hypothetical proteins and allows confidence assessment, based on the supportive evidences from the high-throughput data. Our work demonstrates the power of integrating multiple sources of high-throughput data with biological functional annotations, in the function prediction for unknown proteins. In addition to this, we have also developed a Web server for function prediction in yeast as well as other organisms. We have applied our method to the *Saccharomyces cerevisiae* proteome and are able to assign function to 1548 out of the 2472 unannotated proteins in yeast with our approach.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Determination of protein function is one of the most challenging problems of the post-genomic era. With the advent of the whole genome sequencing, the focus has shifted towards identification of genes and the prediction of their function. 108 bacterial, 16 archaeal and 9 eukaryotic genome sequences are complete, while 128 bacterial, 2 archaeal and 35 eukaryotic genome sequences are in progress at NCBI [1], as of June 2003. However, only 50-60 percent of genes have been annotated. The traditional method of wet laboratory experiments can assign function accurately, however the process is very time-consuming and costly. This leaves the field of bioinformatics with the challenging task of assigning function to the unannotated proteins and developing more efficient and accurate methods for function prediction. What is also important is the type of the functional annotation. A protein has two types of functions i.e biochemical and cellular. Biochemical function refers to the biochemical activity, e.g. cyclase or kinase, and is relatively easy to annotate, since it can be assigned based on sequence similarity. However, the difficult task here is to assign the cellular function, which refers to the biological objective, e.g. pyrimidine metabolism or signal transduction, of the unannotated proteins. Our aim is to be able to assign the cellular function to the proteins, which is similar to the GO biological process functional annotation.

There have been many approaches in the past to assign protein function. Information derived from sequence similarity, phylogenetic profiles, clustering patterns of co-regulated genes, protein-protein interactions, protein complexes and gene expression profiles, has been utilized in function prediction approaches. The classical

way to infer function based on sequence homologies is using programs such as FASTA [2] and PSI-BLAST [3]. Another method to predict function based on sequence information is the Rosetta Stone method [4]. Two proteins are inferred to share the same function, if they are both present together in another genome. Function can also be inferred based on the phylogenetic patterns of a protein in multiple genomes. It is believed that genes with similar functions are likely to have similar phylogenetic patterns [5].

Today there is an ever-increasing flow of biological data generated by the high-throughput methods such as yeast two-hybrid systems [6], protein complexes identification by mass spectrometry, microarray gene expression profiles and systematic synthetic lethal analysis. High-throughput experiments are designed to acquire information about thousands of genes at the same time and to study the relationships among them. High-throughput data is generated from technology driven studies at the whole genome or the proteome scale, as against hypothesis driven experiments in the laboratory, designed to study a particular protein or gene of interest. With the availability of entire genome sequences and high-throughput data, which can determine gene co-expression profiles, research scientists are focusing on whole proteome-wide studies, as against the study of single proteins or a small group of proteins in the past. This change in the strategy has made it necessary to design and implement reliable methods to assign protein function.

Many approaches have been designed to assign function based on gene expression profiles, mutant phenotypes data and protein-protein interactions. Cluster analysis of the gene-expression profiles is the most common approach used to predict function from high-throughput data. This approach is based on the assumption that genes with similar

functions are likely to be co-expressed [7, 8, 9]. Using protein-protein interaction data to assign function to novel proteins is another common approach. Proteins interact with one another in an interaction network to achieve a common objective. It is therefore possible to infer the function based on the functions of its interaction partners. Schwikowski et al. [10] applied neighbor-counting method in trying to predict the function. They assigned functions to the unknown proteins based on the frequencies of its neighbors having certain functions. Hishigaki et al. [11] used χ^2 statistics to infer protein function. Both these approaches give equal significance to the functions contributed by the l-neighbors of the protein. Deng et al. [12,13] developed a mathematical model based on the theory of Markov random fields, to infer protein function using protein-protein interaction data and functional annotation of the interaction partners. Instead of searching for the simple consensus among the functions of the interacting partners, they used the Bayesian approach to assign a probability for a hypothetical protein to have the annotated function.

However, high-throughput data is very noisy and may have many false positives and false negatives. The main drawback of these approaches lies in the use of only single type of data. There are limitations of the high-throughput technologies as is evident from the inability of the yeast two-hybrid assays to detect a few protein-protein interactions dependent on post-translational modifications or multi-body effects, while mass spectrometry may fail to detect some transient and weak interactions. In a microarray clustering analysis the genes with similar functions may not be clustered together due to lack of similar expression profiles. Though noisy and inconsistent, high-throughput data are a rich and valuable source of information, which when utilized in a careful manner can yield valuable information. Our aim is to combine the information from the various

types of high-throughput data, in an effort to remove the in-consistencies in the data, to validate the available information and to use all of it. It is a challenging task to combine the information together, since every type of high-throughput data has its own errors and is highly noisy.

We have developed a method for cellular functional annotation of the novel proteins in *Saccharomyces cerevisiae* using high-throughput biological data including yeast two-hybrid, protein complexes, genetic interactions and microarray gene expression. We decided to use the yeast *Saccharomyces cerevisiae,* as it is a well-studied and good model for the eukaryotic systems and rich high-throughput data are available for yeast. Our ultimate aim is to be able to extend the prediction method to assign function to the proteins in other organisms. At present, about 3866 yeast genes have been annotated, which leaves about 2400 genes yet to be assigned a function. All the data is encoded together into a graph of interaction network, where each node represents a protein and each edge represents an interaction between them. This is a novel approach to assign function to the unannotated proteins using multiple sources of high-throughput data. The use of multiple sources of high-throughput data allows cross validation between the different sources of information and increases the confidence in any single type of data.

We acquired the various types of high-throughput data and created a database for the centralized storage of this information, as is detailed in Chapter 2. Chapter 3 explains the estimation of the *a-priori* probabilities from the analysis of the various types of high-throughput data and the rationale behind our function prediction method. The results of function prediction and the validation of our method are discussed in Chapter 4. In

Chapter 5 we describe the web-based function prediction tool GeneFAS, developed to query the results of function prediction in yeast and in other organisms. We also try to extend our method for cross-species function prediction, as outlined in Chapter 6.

2. MATERIALS

The first step towards development of the function prediction method is the acquisition of the multiple sources of high-throughput data and the functional annotation data for the known proteins. As mentioned earlier, we use yeast *Saccharomyces cerevisiae* our model organism.

2.1 DATA SOURCES

The yeast (*Saccharomyces cerevisiae*) data was acquired from various sources. Primarily two types of data were downloaded and stored for yeast, namely main data and supportive data. The gene names, ORF names and SGDID were acquired from Saccharomyces Genome Database (SGD) [14]. The main data includes protein-protein interaction data, protein complexes data and cellular functional annotation from Gene Ontology (GO) [15] and MIPS [16]. The microarray gene expression data set is from the paper of Roberts et al. [17], including 56 experiments conditions. The protein-protein interaction data are of three types, i.e physical binary interactions, genetic binary interactions and protein complex interactions. The binary physical and genetic interactions data were acquired from MIPS Comprehensive Yeast Genome Database (CYGD). Cellular function annotation from MIPS along with other supportive data including mutant phenotype, protein classes, motif, EC number and sub-cellular localization information was also obtained from CYGD. The GO annotation was acquired from the Gene Ontology website. For the genes for which the sub-cellular localization information was unavailable at MIPS, their localization was predicted using the Yeast

Protein Localization Server [18]. All the data were complied together and collected using Perl [19].

2.2 CREATION OF YEAST DATABASE

We created a YEAST Database for the centralized storage of all the information. This allows for easy retrieval, processing and manipulation of the data. The YEAST Database was created in the XML [20] format. The use of XML for the database architecture allows us to define tags for the various types of information and also allows easy expansion of the database to accommodate new data in the future, without making any changes to the basic architecture. All the information for each ORF was stored in a separate file. The various attributes of the ORF were defined in the form of XML tags (Figure 1). Index files were created for each deeper classification of functional annotation, GO annotation, sub-cellular localization, protein classification, mutant phenotype and motifs. The indices were stored in the database as a reference to the original information (Figure 2).

2.2.1 PROTEIN-PROTEIN INTERACTIONS

The protein-protein interactions that we consider fall into two categories, physical interactions and genetic interactions. In physical interactions the proteins are involved in physical contact through a binary interaction or the formation of a protein complex. In genetic interactions, the change of one gene may affect the expression of another gene, or mutations of two genes at the same time can produce a novel phenotype that is not displayed by either mutation alone. The two-hybrid experiments allow the

```
ORF name                    <orfname name="  ">

Gene name                   <genename>        </genename>
SGD id                      <sgdid>   </sgdid>
GO annotation               <goannotation>
                            <biological-process>
                                    <goid>              </goid>
                                    <P-function>        </P-function>
                                    <P-evidence>        </P-evidence>
                            </biological-process>
                            <molecular-function>
                                    <goid>              </goid>
                                    <F-function>        </F-function>
                                    <F-evidence>        </F-evidence>
                                    </molecular-function>
                                    <cellular-component>
                                            <goid>              </goid>
                                            <C-function>        </C-function>
                                            <C-evidence>        </C-evidence>
                                    </cellular-component>
                            </goannotation>
Sub-cellular Localization   <subcellular-localization>
                                            <C>     </C>
                                            <N>     </N>
                                            <M>     </M>
                                            <T>     </T>
                                            <E>     </E>
                                    </subcellular-localization>
MIPS functional annotation  <functional-classification>
                                            <function>          </function>
                                    </functional-classification>
Protein Classification              <protein-classification>
                                            <protein-class>     </protein-class>
                                    </protein-classification>
Motif               <prosite-motif>
                            <motif>  </motif>
                    </prosite-motif>
EC Number           <EC-number>
                            <EC>     </EC>
                    </EC-number>
Mutant Phenotype    <mutant-phenotype>
                                    <phenotype>         </phenotype>
                            </mutant-phenotype>
Interaction         <interaction>
                            <physical-interaction>
                                            <interactor>        </interactor>
                            </physical-interaction>
                            <genetic-interaction>
                                    <interactor>        </interactor>
                            </genetic-interaction>
                            <complexes>
                                    <complex>           </complex>
                            <complex-interactor>        </complex-interactor>
                    </interaction>
Gene Expression     <microarray-expression>     </microarray-expression>
```

Figure 1. Example of yeast ORF XML file for the database.

<?xml version ="1.0"?>

<yeast>
<orfname name="YLR094C">
<genename>GIS3</genename>
<sgdid>S0004084</sgdid>
<goannotation>
<biological-process>
<goid>GO:0000004</goid>
<P-function></P-function>
<P-evidence>IEA</P-evidence>
</biological-process>
<molecular-function>
<goid>GO:0005554</goid>
<F-function></F-function>
<F-evidence>ISS</F-evidence>
</molecular-function>
<cellular-component>
<goid>GO:0008372</goid>
<C-function></C-function>
<C-evidence>ND</C-evidence>
</cellular-component>
</goannotation>
<subcellular-localization>
<C>0.081</C>
<N>0.605</N>
<M>0.226</M>
<T>0.053</T>
<E>0.036</E>
</subcellular-localization>
<functional-classification>
<function>99</function>
</functional-classification>
<protein-classification>
</protein-classification>
<prosite-motif>
</prosite-motif>
<EC-number>
</EC-number>
<mutant-phenotype>
</mutant-phenotype>
<interaction>
</interaction>
<microarray-expression>-0.97 -0.45 -0.32 -0.4 -0.18 0.07 -0.42 -0.0374265939597316 -0.14 0.23 0.1 -0.15 0.19 0.48 -0.01 0.24 0.66 0.2 -0.09 0.
43 0.45 0.42 0.24 0.29 0 -0.04 -0.17 0.12 -0.04 0.12 -0.01 0.25 0.08 0.03 -0.23 -0.64 -0.6 -0.29 -0.69 -0.3 -0.18 0.12 -0.34 -0.03 -0.06 -0.62
 -0.62 -0.64 -0.43 -0.42 -0.58 -0.36 -0.76 -0.69 -0.2 -0.266232464929859 -0.145391609836744 -0.54 -0.34 0.1 0.16 0.37 0.1 -0.14 -0.22 -0.76 0.
11 0.55 -0.12 -0.17 -0.45 -0.43 -0.58 -0.56 -0.4 -0.22 -0.04 0.06 0.8 0.08 -0.29 0.29 0.25 0.07 0.32 0.06 0.16 0.15 0.07 -0.172978436657681 -1
.091 0.224 1.05 0.25 0.28 0.4 0.15 -0.25 -0.04 0.62 0.11 -0.25 -0.04 -0.47 -0.22 -0.25 0.14 0.24 0.36 0.44 -0.01 0.1 -0.04 -0.09 0.38 0.26 0.0
8 -0.14 0.08 0.296 -0.26 -0.001 0.011 -0.003 0.144 -0.018 -0.224 -0.138 -0.367 1.25 1.45 1.01 -1.08 -1.2</microarray-expression>

Figure 2. Example of ORF YLR094C in the Yeast Database.

reconstruction of the binary interactions among a set of proteins in a proteome and is commonly used to identify the physical interactions. The synthetic lethality screen is a very powerful method for identifying genetic interactions [21, 22]. It identifies non-allelic and non-lethal mutations that are lethal in combination with a non-lethal mutation in a gene of interest. The physical interaction data acquired from MIPS, Uetz et al. [23] and Ito et al. [24] were combined and this set of physical interactions were used in the predictions. Genetic interactions data was obtained from MIPS, which included data from synthetic lethal screens, suppression and over-expression experiments. There are total 6516 physical binary interactions and 1019 genetic binary interactions for yeast.

2.2.2 PROTEIN COMPLEXES

Proteins in a protein complex are typically identified by enrichment of the complexes in a cell extract using a chromatographic technique that captures one protein and its associated proteins, followed by subsequent mass spectrometric identification of the proteins in the complex. The protein complexes data is obtained from Gavin *et al.* [25] and Ho *et al.* [26]. Gavin et al. [25] used TAP (tandem affinity purification) for protein complex identification. Ho *et al.* [26] used HMS-PCI (high-throughput mass spectrometric protein complex identification) method to identify protein complexes. Since in the protein complexes it is unclear which proteins are in physical contact and interact directly with which other protein of the complex, it is hard to construct a specific interaction network. Despite this lack of clarity in the protein physical interactions within a complex, the protein complexes data is a very rich resource. In order to utilize this information we have converted each protein complex into a set of binary interaction of proteins within

the complex. For each complex, we added one interaction edge to the interaction network between each protein in the complex. Thus in general, if there are n ORF's in a protein complex, n*(n-1)/2 edges are added to the interaction network. The protein complexes data that we use consist of 232 complexes, involving 1440 distinct proteins. These data, when converted to binary interactions, add 49,313 edges to the interaction network.

Once all the interaction edges are added to the interaction network, the final network has 6516 physical binary interactions, 1019 genetic interactions and 49,313 binary complex interactions.

2.2.3 MICROARRAY GENE EXPRESSION DATA

Analysis of microarray gene expression data is currently one of the most active research areas in the field of genomics. Computationally clustering individual gene expression measurements provides a novel approach to exploit and infer information in order to characterize biological processes. For example, based on the assumption that groups of genes that are co-expressed are likely to share similar function, cluster analysis of gene-expression profiles results in hypotheses of function. The microarray gene expression data was acquired from the published research of Roberts et al. [17] that considered 56 experimental conditions. For each experiment if there was data missing, we substituted it with the average ratio of all the ORF's under that specific experimental condition, to maintain the dimension of the observations. A correlation coefficient was calculated for each of the possible ORF pairs to quantify the correlation between the gene pairs.

2.2.4 SUB-CELLULAR LOCALIZATION

The sub-cellular distribution of proteins within a proteome is useful and important for a global understanding of the molecular mechanisms of a cell. Protein localization can serve as an indicator of protein function. Localization data can be used as supportive evidence to evaluate protein information inferred from other resources. In a physical protein-protein interaction, if two proteins are involved in direct physical contact with each other, they should have the same sub-cellular localization. If an interaction pair has the same sub-cellular localization, it acts as supportive evidence raising the confidence level for that interaction. Therefore, the study of the relationship between protein-protein interactions and the sub-cellular localizations of the interaction partners can help validate the protein-protein interaction data generated from high-throughput experiments, which otherwise may be very noisy. The sub-cellular localization information is acquired from MIPS. We consider five main sub-cellular localization compartmental categories namely, Cytoplasmic, Nuclear, Mitochondrial, Transmembrane, Endoplasmic Reticulum pathway proteins. (Table 1)

Based on the localization information obtained from CYGD, each ORF was assigned the value of 1 for the main localization compartment. Value 1 indicates a high localization quality with a high level of confidence in this assignment. For the ORF's with no localization information in CYGD, the localization was predicted using the Yeast Protein Localization Server. The results are in the form of prediction values between 0 and 1 for each of the five main localization compartments. The compartment with the highest prediction value is the most likely localization compartment for the ORF. The

Table 1. Sub-cellular localization compartments.

Compartment	Description
C	Cytoplasmic (excluding cytoskeletal)
N	Nuclear
M	Mitochondrial
T	Transmembrane (including plasma membrane proteins)
E	Endoplasmic Reticulum (ER) pathway proteins: ER, golgi, extracellular, peroxisomal, vacuolar, vesicular

results for all the predicted compartments for each ORF were stored in the database. If the localization server failed to predict the localization for the ORF, a value of 0 was assigned for the localization. Value 0 indicates a low localization quality as reflected from its predicted or unknown nature. In the MIPS database, 2358 ORF's have known sub-cellular localizations from experimental evidence, out of which 169 ORF's can be localized in more than one sub-cellular compartment [27, 28]. After predicting the localization for the remaining ORF's, we had sub-cellular localization for 6034 yeast ORF's. An index was created to refer to the deeper localization classification and the information was stored in the database with the index numbers (Table 2).

2.2.5 MIPS FUNCTIONAL ANNOTATION

The functional annotation for the known proteins in yeast was acquired from the MIPS database. The functions were divided into 17 broad functional categories. Each higher-level category has further functional sub-classes. We assigned a numerical index to each of these ORF's according to its hierarchical function classification. That also included a functional category sub-cellular localization, and classification not yet clear and unclassified proteins categories.

Example :

Cellular communication/signal transduction mechanism ……10

Intracellular signaling ……01

Enzyme mediated signal transduction ……03

G-protein mediated signal transduction ….05

Table 2. Sub-cellular localization categories and indices.

Sub-cellular Localization Main Category	Sub-cellular Localization Index and Deeper Categories
E	0 extracellular
T	1 cell wall
T	2 plasma membrane
C	3 cytoplasm
C	4 cytoskeleton
C	5 actin cytoskeleton
C	6 tubulin cytoskeleton
C	7 spindle pole body
C	8 intermediate filaments
E	9 ER
E	10 ER membrane
E	11 ER lumen
E	12 golgi
E	13 golgi membrane
E	14 transport vesicles
E	15 ER-golgi transport vesicles
E	16 golgi-ER transport vesicles
E	17 inter-golgi transport vesicles
E	18 golgi-plasma membrane transport vesicles
E	19 golgi-vacuole transport vesicles
E	20 endocytotic transport vesicles
E	21 other transport vesicles
N	22 nucleus
N	23 nuclear envelope
N	24 nuclear matrix
N	25 nucleolus
N	26 nuclear pore
N	27 chromosome structure
M	28 mitochondria
M	29 mitochondrial outer membrane
M	30 mitochondrial intermembrane space
M	31 mitochondrial inner membrane
M	32 mitochondrial matrix
E	33 peroxisome
E	34 peroxisomal membrane
E	35 peroxisomal matrix
E	36 endosome
E	37 vacuole
E	38 vacuolar membrane
E	39 vacuolar lumen
E	40 microsomes
E	41 lipid particles

2.2.6 GO FUNCTIONAL ANNOTATION

The GO functional annotation has three categories. Biological process refers to the biological objective and hence annotates cellular function of a protein, e.g. signal transduction, pyrimidine metabolism. Molecular function refers to the biochemical activities and hence annotates the biochemical functions of a protein, e.g. enzyme, adenylate cyclase. Cellular component refers to the localization of the proteins in the cell, e.g. ribosome, nuclear membrane. Each category has a hierarchical structure. We were interested in assigning cellular function to the unannotated proteins and hence we followed the biological process category. After acquiring the biological process functional annotation for the known proteins along with their GO ID, we generated a numerical GO INDEX, which represents the hierarchical structure of the classification. All the functions begin with 1, which represents a biological function, to distinguish them from the other molecular and cellular functions in the GO annotation. The highest level of INDEX is 13.

Example :

1-4 cell growth and/or maintenance GO:0008151

1-4-3 cell cycle GO:0007049

1-4-3-2 DNA replication and chromosome cycle GO:0000067

1-4-3-2-4 DNA replication GO:0006260

1-4-3-2-4-2 DNA dependent DNA replication GO:0006261, GO:0006262, GO:0006263

1-4-3-2-4-2-2 DNA ligation GO:0006266

Table 3 shows the GO biological process annotation for *Saccharomyces*

Table 3. GO biological process annotation for *Saccharomyces cerevisiae* proteins.

GO INDEX	Number of ORFs	Function Description
1-1	1	behavior GO:0007610
1-1-11	1	rhythmic behavior GO:0007622
1-3	362	cell communication GO:0007154
1-3-1	12	cell adhesion GO:0007155
1-3-5	6	host-pathogen interaction GO:0030383
1-3-7	3	response to endogenous stimulus GO:0009719
1-3-8	246	response to external stimulus GO:0009605
1-3-9	143	signal transduction GO:0007165
1-4	3775	cell growth and/or maintenance GO:0008151
1-4-1	24	autophagy GO:0006914
1-4-2	99	budding GO:0007114
1-4-3	506	cell cycle GO:0007049
1-4-4	4	cell growth GO:0016049
1-4-6	976	cell organization and biogenesis GO:0016043
1-4-7	5	cell proliferation GO:0008283
1-4-8	93	cellular morphogenesis GO:0000902
1-4-10	95	homeostasis GO:0019725
1-4-11	15	membrane fusion GO:0006944
1-4-12	2835	metabolism GO:0008152
1-4-13	208	response to stress GO:0006950
1-4-14	83	sporulation GO:0030435
1-4-16	732	transport GO:0006810
1-5	3	death GO:0016265
1-5-1	3	cell death GO:0008219
1-6	192	development GO:0007275
1-6-2	16	aging GO:0007568
1-6-11	55	growth GO:0040007
1-6-16	93	morphogenesis GO:0009653
1-6-20	91	reproduction GO:0000003
1-6-22	19	sex determination GO:0007530
1-8	95	physiological processes GO:0007582
1-8-6	91	conjugation GO:0000746
1-8-22	2	nutritional response pathway GO:0007584
1-8-30	2	respiratory gaseous exchange GO:0007585
1-9	7	viral life cycle GO:0016032
1-9-8	6	virus-host interaction GO:0019048

cerevisiae annotated proteins with the function description and the number of ORFs that belong to the observed GO INDEX.

The GO functional annotation appears to be a more systematic and robust classification as compared to the MIPS functional annotation. MIPS has a coarser functional classification scheme. We therefore decided to follow the GO biological process annotation for the functional assignment.

3. METHOD

Our function prediction method consisted of two steps. We estimated the *a-priori* probabilities for the different types of high-throughput data in the first step. In the second step we utilized these estimated *a-priori* probabilities to predict the functions of unannotated proteins.

3.1 ESTIMATION OF *A-PRIORI* PROBABILITIES

For each type of high-throughput data including protein-protein interactions, protein complexes and gene expression data, we considered every pair of genes represented and compared them with each other to examine if they shared similar functions and if so what level of function INDEX similarity they had. In order to accomplish this we assessed every such pair from the available data by comparing the GO INDEX for the annotated protein.

Eg. Consider ORF1 and ORF2 that have a physical binary interaction with each other.

ORF1 has a function represented by GO INDEX 1-4-3-3-4 and ORF2 has a function represented by GO INDEX 1-4-3-2. When compared with each other for the level of matching GO INDEX, they match with each other through 1-4-3 i.e INDEX level 1 (1-4) and INDEX level 2 (1-4-3).

A similar approach was used for the analysis of all the four types of data namely, physical binary interactions, genetic binary interactions, binary interactions generated from the protein complexes data and gene expression data. For the microarray gene expression data, we calculate the correlation coefficient for each gene expression pair.

3.1.1 PROTEIN-PROTEIN INTERACTIONS AND PROTEIN COMPLEXES

Figure 3 shows the results of the analysis of the protein-protein interactions data. The physical interaction data used for the analysis was combined from MIPS, Uetz et al. and Ito et al. It showed a decrease in the percentage of pairs sharing the same function with an increase in the INDEX level. The genetic interactions and the physical interactions data were more informative in terms of matching function, in comparison to the protein complexes data. This may be the effect of the different techniques used to derive these data. It showed the percentage of pairs sharing the same function was higher with the lower INDEX levels, which represent a less specific function class, in comparison to the higher INDEX levels. Our conclusion was that there was a clear relationship between the protein-protein interaction pairs and similarity in function, which can be utilized to make future predictions based on these data. This relationship is more evident in the genetic and physical interactions in comparison to the protein complexes. The relationship was strong with the lower INDEX levels and weakened with increasing INDEX levels.

3.1.2 MICROARRAY GENE EXPRESSION

We calculated the correlation coefficient for all the gene expression pairs, based on the assumption that groups of genes that are co-expressed are likely to share similar function [29]. The range of the correlation coefficient is from –1 to +1. The value +1 means a perfect correlation between the two expression profiles, while the values below 0, indicate no correlation. We calculated the value 1- correlation coefficient to plot the percentage of pairs with matching function for each INDEX level, to quantify the

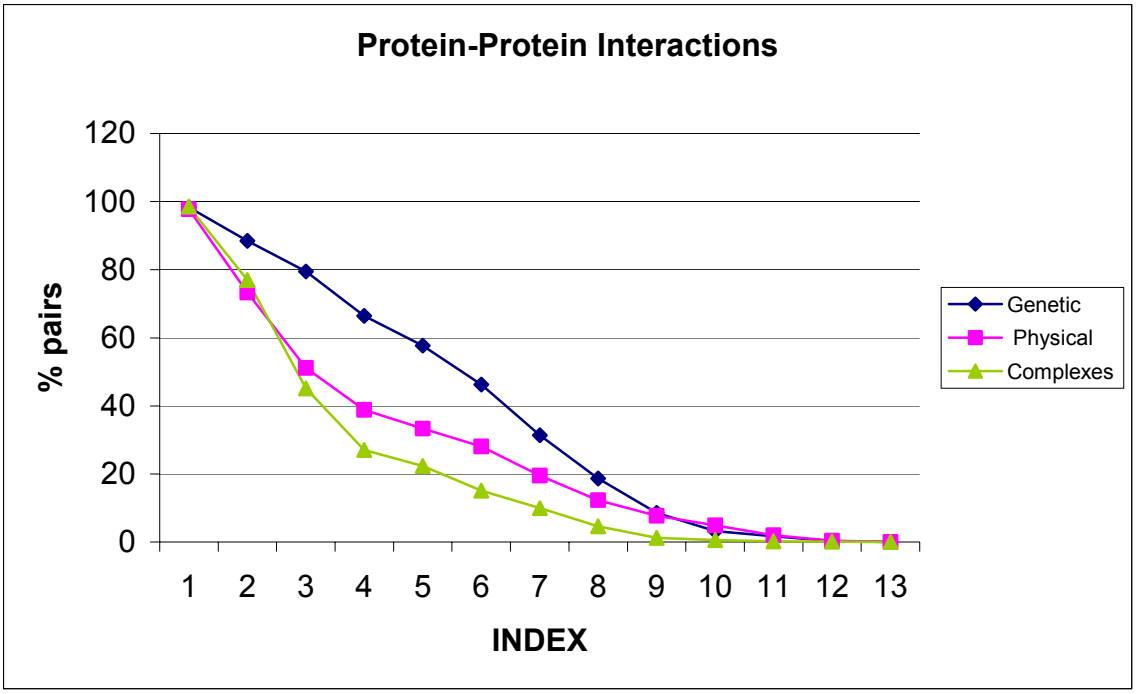

Figure 3. Percentage of pairs with matching function for INDEX levels 1 to 13 against protein-protein interactions data (physical, genetic and binary interaction from complexes).

relationship between the correlated gene expressions pairs. The range of value 1- correlation coefficient is from 0 to 2 (Figure 4). The value 0 means perfectly correlated as compared to values greater than 1, which indicate no correlation. Results of this analysis clearly showed a decrease in the probability of sharing the same function with an increase in the value (1- correlation coefficient). It also showed a decrease in the probability of sharing the same function with an increase in the INDEX level. The slope of the curves indicated a clear relationship between the microarray data value 1- correlation coefficient ≤ 0.6 and similarity in function. This relationship was lost above this value, where the curves become flat. Thus it is possible to utilize this relationship for the data with value 1- correlation coefficient ≤ 0.7 to make future predictions regarding function.

Based on these results we decided to concentrate more on the values of 1- correlation coefficient, in the range of 0 to 0.7. Figure 5 is the blowup of the region with 1- correlation coefficient values between 0 and 0.7. In this range, we recalculated the percentage of pairs with matching function for INDEX levels 1 to 13, at an interval of 0.025. The data point at 0.025 shows a smaller percent of pairs sharing same function, due to the very small data size for this data point.[♣]

3.2 PREDICTION USING *A-PRIORI* PROBABILITIES

Our method visualizes the protein-protein interaction data as a network graph in which a node represents a protein and the edges represent the interactions between the

[♣] For information regarding cross hybridization, refer to Appendix App-1.

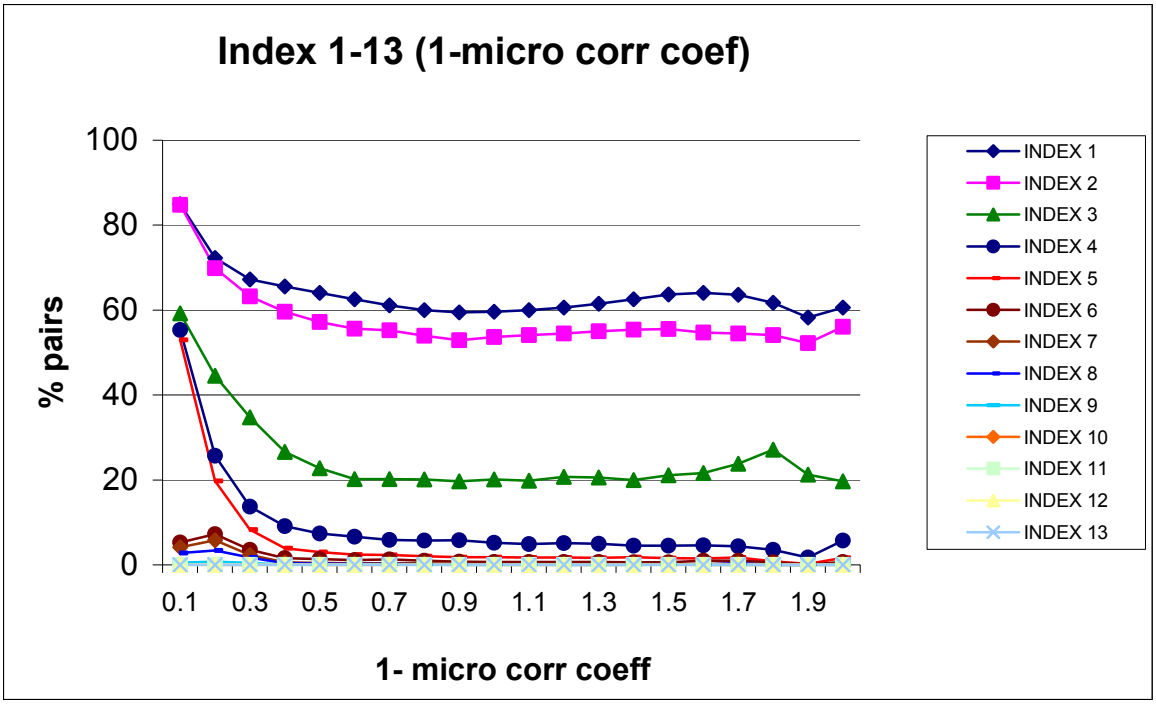

Figure 4. Percentage of pairs with matching function for INDEX levels 1 to 12 against microarray 1-correlation coefficient.

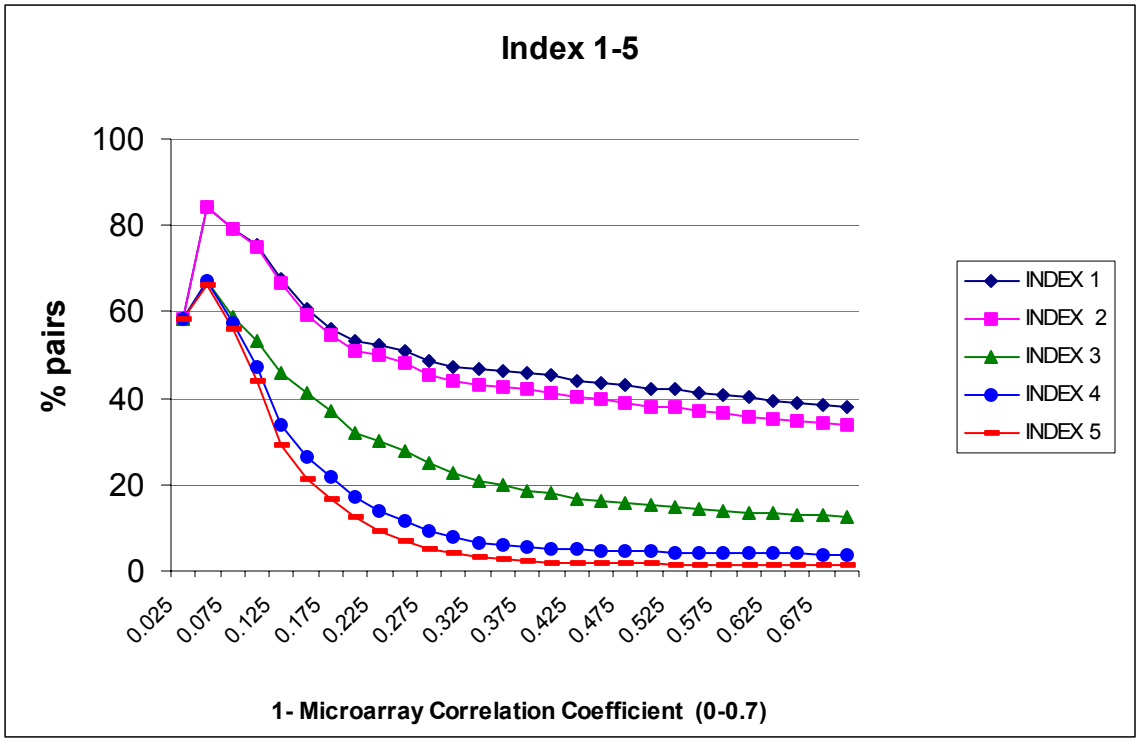

Figure 5. Percentage of pairs with matching function for INDEX levels 1 to 5 against microarray 1-correlation coefficient (0-0.7).

proteins. The function predictions were based on the assumption that the interacting proteins share at least one function in common and therefore belong to at least one common functional class. Therefore knowledge of the functional class of a few interacting proteins in the network, may lead to an accurate prediction of the function for the remaining uncharacterized interaction partners. Thus each protein can belong to one or more functional classes, depending upon the scheme of functional classification used in determining the functions. Thus the functional classification scheme used would affect the number of possible functions for each protein. Two types of classification schemes may exist. A coarse or less detailed functional classification scheme will have fewer function classes and will be less stringent. As compared to this a deeper, more detailed classification will have more functional classes and will be more stringent. Keeping these criteria in mind we decided to use the GO hierarchical functional classification in our prediction method as against the MIPS classification. Our approach was to assign functions to the uncharacterized unannotated proteins on the basis of the most common functions identified among the annotated interaction partners for this protein and the estimated *a-priori* probabilities. This approach is similar to the majority rule assignment. For each protein in addition to the probable function prediction we report its interactions partners along with their functions. We attribute a higher confidence to the predicted function, which have a higher rate of occurrence.

In this method we find all the possible protein-protein interactions that the query yeast protein may have, based on the collected high-throughput data. For one type of data at a time (physical interactions, genetic interactions, protein complex binary interactions and microarray gene expression with correlation coefficient ≥ 0.8, we identified the

possible interactors for the query protein. We compared the function for the query protein and each interactor in terms of the GO INDEX namely, series of the numbers, which represents the function.

For example, if for a query protein the interactor had a GO INDEX 1-3-4-2, the possible GO function INDICES for the query protein were 1-3, 1-3-4 and 1-3-4-2.

For each potential GO INDEX we assigned a score based on the type of high-throughput data, which identified this interactor. This score was assigned based on the *a-priori* calculated probabilities from the analysis of that type of high-throughput data for each INDEX level 1 to 13. If one or more interactors were identified, then based on their function GO INDEX, the query protein potential INDICES were predicted.

For example, if for a query protein the interactor 1 had a GO INDEX 1-3-4-2, and interactor 2 had a GO INDEX 1-3-4-3, then the possible GO function INDICES for the query protein were 1-3, 1-3-4, 1-3-4-2 and 1-3-4-3.

The final GO INDEX predictions for the query protein were sorted based on the reliability score for each predicted GO INDEX. The reliability score was a combined score calculated based on the types of the high-throughput data used for the prediction.

For each GO INDEX let P_1 = probability from genetic interactions,

P_2 = probability from physical interactions,

P_3 = probability from complex interactions,

P_4 = probability from microarray gene expression, then

$(1-P_1)$ gives the probability of a protein not sharing the same function as its physical interaction partner, and respectively for all the other types of data.

We considered all the types of data to be independent of each other and thus do not estimate the parameters in the form of weights. Instead we combined them all together and then estimated the probability for the protein to have the same function as that of the known interaction partners, by calculating the prediction score as,

Reliability score = 1- [(1-P_1) (1-P_2) (1-P_3) (1-P_4)]

But since the value is a very small number, we could loose precision due to this method of calculation. Instead we calculated the final reliability score by taking a log value of the terms,

Reliability score = 1- exp [(Log(1-P_1) + Log(1-P_2) + Log(1-P_3) + Log(1-P_4))]

3.2.1 NORMALIZATION

If more than one interactors for a query protein from a single type of data, had the same GO INDEX function, then the GO INDEX function prediction for the query protein was normalized based on the number of interactors supporting it.

For example, if for a query protein the interactor 1 had a GO INDEX 1-3-4-2, and interactor 2 had a GO INDEX 1-3-4-3, then the possible GO function INDICES for the query protein were 1-3, 1-3-4, 1-3-4-2 and 1-3-4-3. In this case the scores for INDEX 1-3 and 1-3-4 had two interactors supporting this function, so the scores for them were normalized as follows:

If $(1-P_1)$ is the probability for this INDEX from this high-throughput data type, and K is the number of interactors supporting this INDEX from this high-throughput data type, we normalized the score as,

$$\mathbf{(1-P_1) = (1-P_1')^{K}}$$

and we now used the score $\mathbf{(1-P_1')}$ instead after normalization.

3.2.2 TESTING AND TRAINING

For the purpose of testing and training, we split the 3866 annotated proteins with known GO id into two sets. The training set had 2866 proteins and the testing set had the remaining 1000. All the values were re-calculated for the testing and the training sets and the corresponding values were used. The testing set of 1000 was done in 10 sets of 100 each.

At one time only 100 proteins out of the 3866 known proteins were considered as having no known function and then with the new calculated probabilities the prediction was done. As expected the sensitivity and specificity values matched for the testing and training sets (Figure 6).

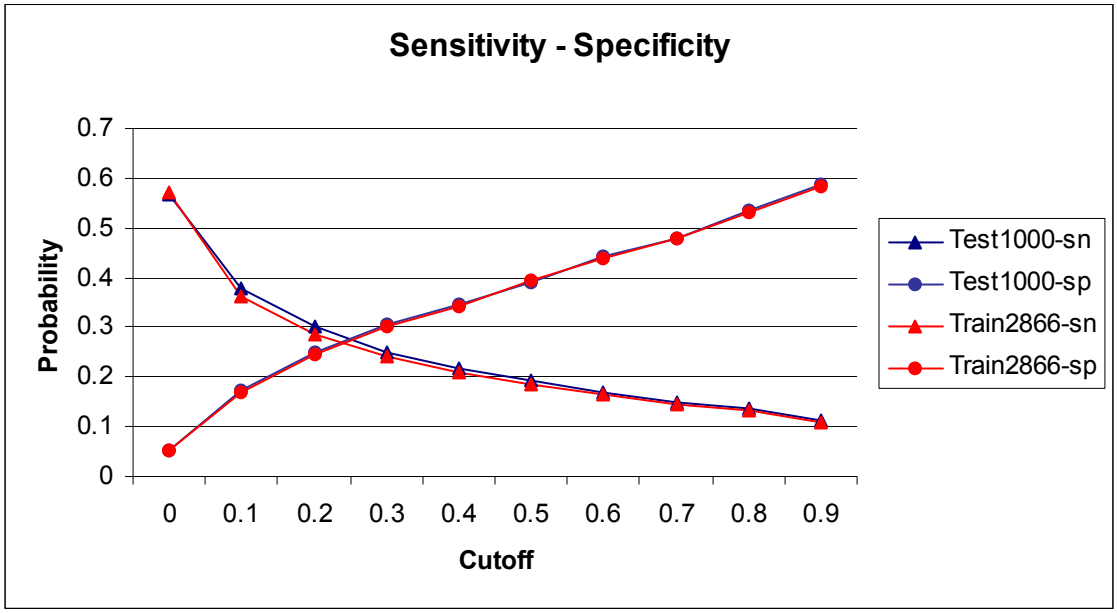

Figure 6. Sensitivity and specificity for testing-training.

4. RESULTS AND DISCUSSION

Sensitivity and specificity are two important measures to evaluate the accuracy of the prediction method. We estimated the sensitivity to determine the success rate of the method and specificity to assess the confidence in the predictions of the method. Let us take a brief look at the two measures. For a given set of proteins *K*, let n_i be the number of the known functions for protein P_i. Let m_i be the number of functions predicted for the protein P_i by the method. Let k_i be the number of predicted functions that are correct (known). Thus sensitivity (SN) and specificity (SP) are defined as,

$$SN = \frac{\sum_1^K k_i}{\sum_1^K n_i}$$

$$SP = \frac{\sum_1^K k_i}{\sum_1^K m_i}$$

SN and SP give a quantitative evaluation of the prediction accuracy and the reliability of the predictions. One can also observe the change in the SN and SP with respect to the variation in the cutoff for reliability score. Both the values are within the range 0 to 1. In practice, if a method tends to predict most of the known functions, it has high sensitivity and it tends to also predict many more functions than the known functions, thus having a low specificity. Thus for a particular method and a data set, an

increase in sensitivity typically correlates with a decrease in specificity and vice-versa. A good prediction method will yield higher sensitivity and specificity values.

Figure 7 shows the sensitivity and specificity of the method with the use of microarray gene expression data, with different cutoff values for the correlation coefficient from 0.6 to 0.8. It is evident from the plot that the sensitivity decreases slightly with the increase in cutoff values for 1-correlation coefficient, whereas the specificity improves substantially with the increase in cutoff values for 1-correlation coefficient. As a tradeoff between the sensitivity and specificity, we decided to use the microarray gene expression data with values for 1-correlation coefficient >= 0.8, to achieve the best results for the method.

Our initial approach was to use a curve–fitting method for the microarray gene expression data and use the equations for the *a-priori* probabilities estimated from the microarray data analysis. Mathematica [30] was used to fit the microarray data points to the different exponential curves for the different INDEX levels and to obtain the equations for the different curves. The general equation for the curve is

$$Y = C + \alpha e^{-\beta X}$$

where C, α and β are constants and X is the value 1-correlation coefficient for the microarray data. But, we were not satisfied with the sensitivity and specificity obtained by this approach. So we tried a different approach for estimating the *a-priori* probabilities i.e linear lookup. In this we estimated the *a-priori* probabilities for the different intervals for the value 1-correlation coefficient for the microarray data.

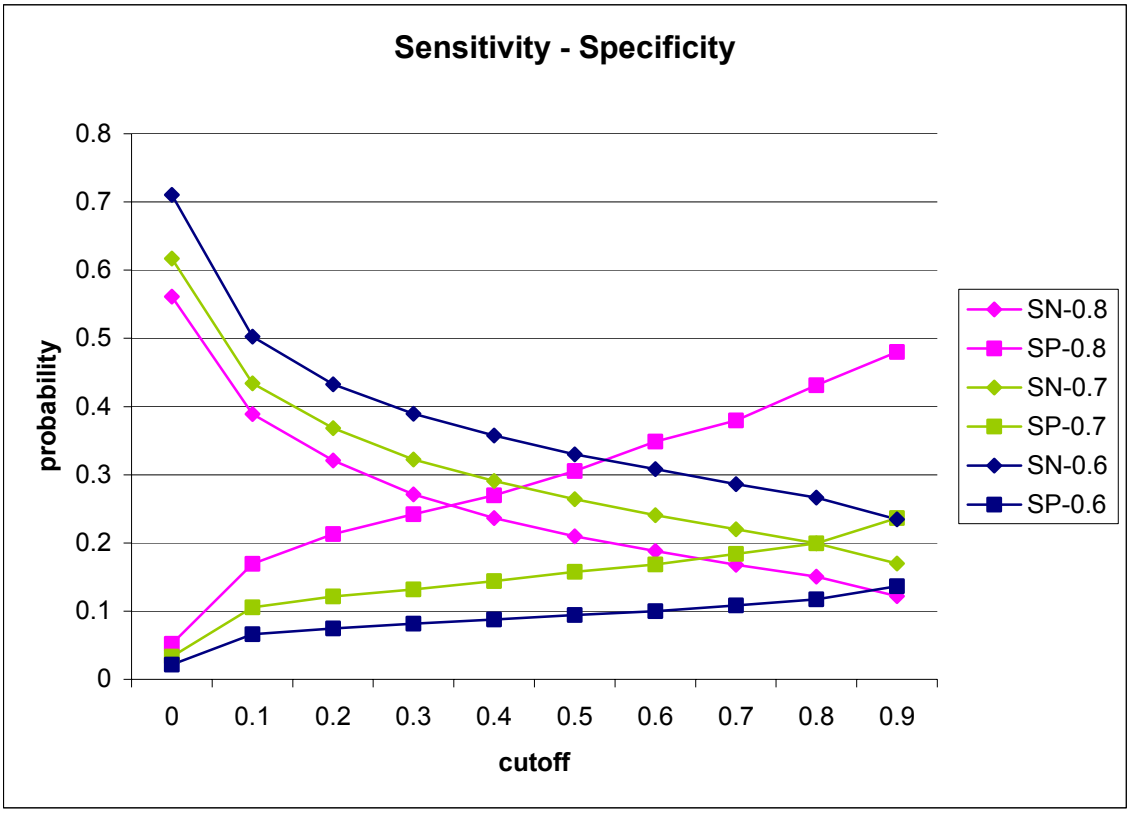

Figure 7. Sensitivity and specificity for microarray correlation coefficient 0.6-0.8.

For example, if the value 1-correlation coefficient was anywhere in the interval 0.15 to 0.25, we used the estimated *a-priori* probability for the value 0.2. Yet, unsatisfied with the results, we decided to use the linear interpolation approach. In the linear interpolation approach, we calculated the *a-priori* probability for the specific microarray 1-correlation coefficient value, based on the neighboring two data points. For example, let X be the 1-correlation coefficient value for point A for which we want to estimate Y, the *a-priori* probability for the value X. Let X_1 and X_2 be the 1-correlation coefficient value for the two neighboring data points on either side of A and Y_1 and Y_2 be the respective *a-priori* probabilities. Then the *a-priori* probability Y for the value X is given by,

$$Y = Y_1 + \left(\frac{Y_1 - Y_2}{X_1 - X_2} \right)(X - X_1)$$

Figure 8 shows the results obtained with the curve-fitting, linear lookup and linear interpolation approaches, used to utilize the microarray gene expression data. As is evident from the results (Figure 8) the performance of the method was better with the linear extrapolation method as against the curve fitting and the linear lookup approaches.

The performance of the method when judged against the cutoff value for reliability score from 0 to 0.9, has a sensitivity and specificity of around 58% (Figure 9, 10). The difference in the reliability score and the probability may be due to the fact that we consider all the types of data as independent data, whereas in reality this may not be

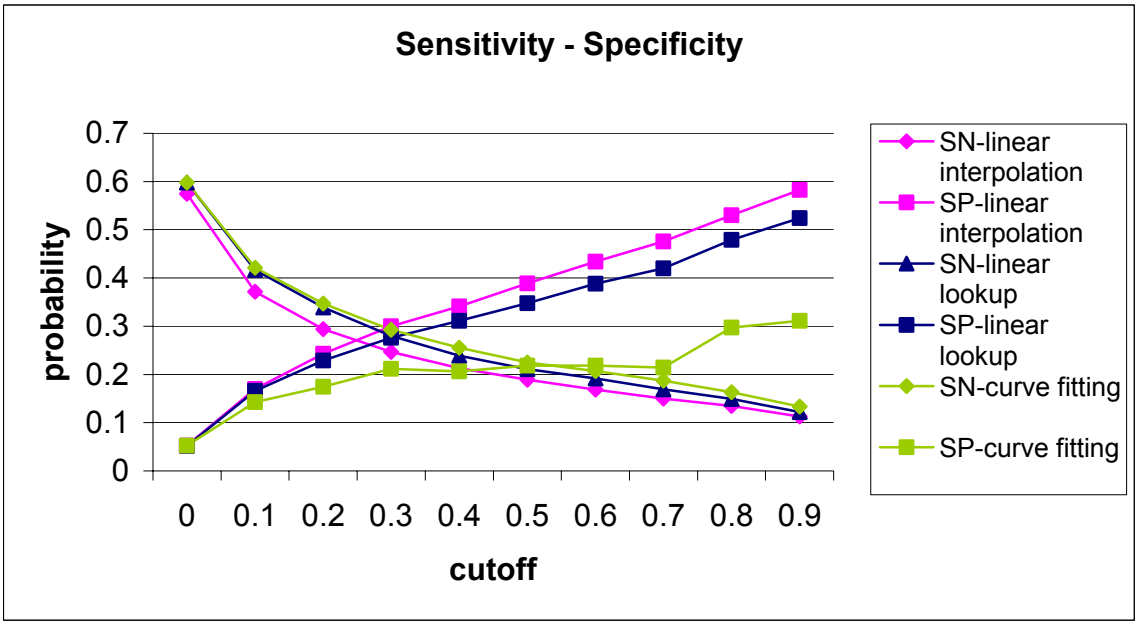

Figure 8. Sensitivity and specificity for curve fitting, linear lookup and linear interpolation methods for utilizing microarray correlation coefficient data.

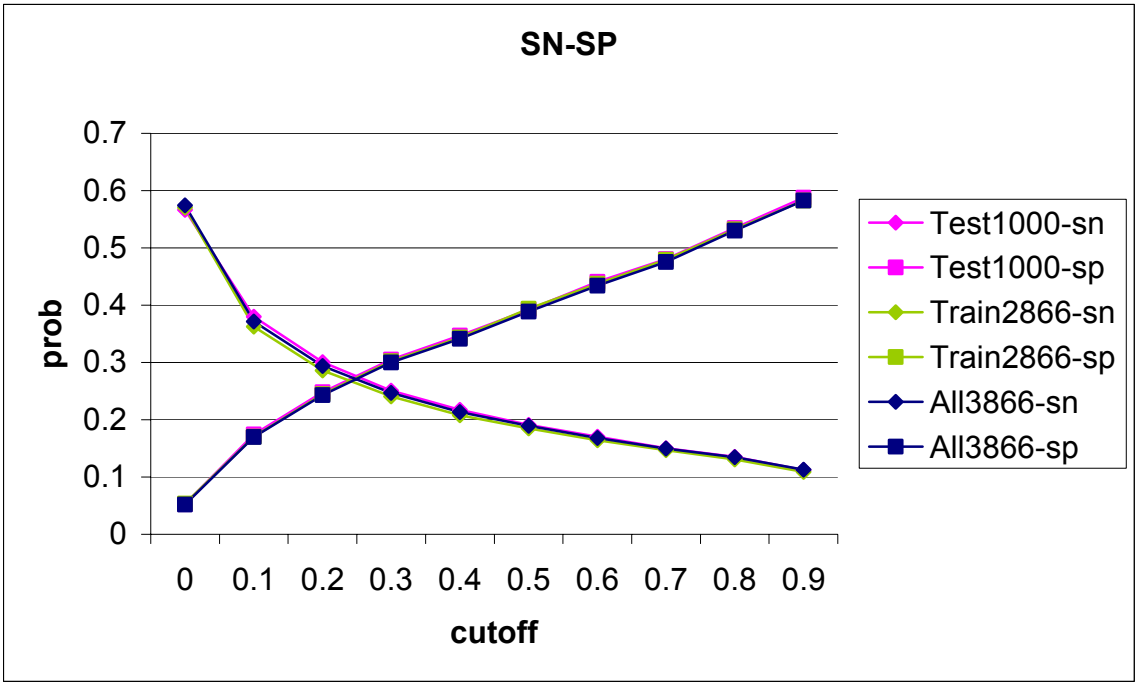

Figure 9. Sensitivity and specificity for the testing set, training set and all 3866 known protein.

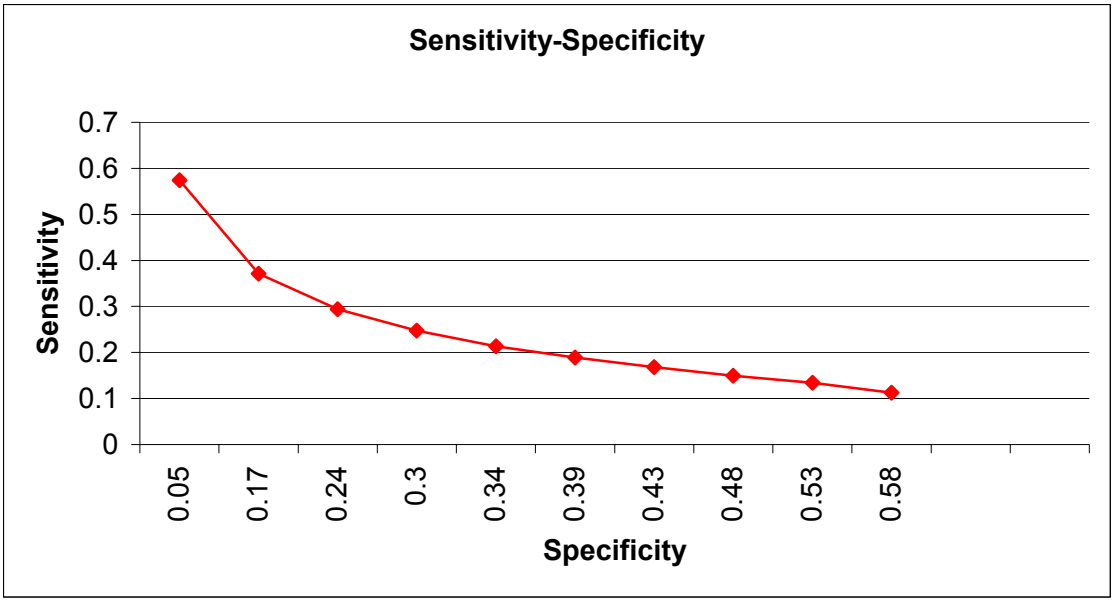

Figure 10. Sensitivity and specificity of the method.

the case. The final prediction along with the predicted GO INDEX, reliability score, function description and GO id also outputs the probability of predictions being correct.

Using our method, we have been able to assign function to 1548 out of the 2472 unannotated proteins in yeast. Figure 11 shows the distribution of the 1548 unannotated yeast proteins for which GeneFAS predicted function, against the reliability score cutoff and INDEX level. It shows the percentage of the unannotated proteins predicted for a particular reliability score cutoff as well as for particular INDEX levels. For example, consider the reliability score cutoff to be 0.8, as shown by the purple curve. For a reliability score cutoff of 0.8, 92.8 percent unannotated proteins out of 1548 have predictions for INDEX level 1, while 48.2 percent unannotated proteins out of 1548 have predictions for INDEX level 2 and so on. Table 4 shows the distribution of the 1548 unannotated yeast proteins for which GeneFAS predicted function, against the probability cutoff and INDEX level. It shows the percentage of the unannotated proteins predicted for a particular probability cutoff as well as for particular INDEX levels. For example, for a probability cutoff of 0.5, 54.6 percent unannotated proteins out of 1548 have predictions for INDEX level 1, while 25.6 percent unannotated proteins out of 1548 have predictions for INDEX level 2 and so on.

Our approach differs from the other approaches in many aspects. We follow GO functional annotation in comparison to MIPS annotation followed by others. In our approach we incorporate more data by including genetic interactions in addition to the physical interactions, protein complexes and microarray gene expression. We develop a new statistical model and provide confidence assessment for the predictions. We have also developed a web server to query the results of function prediction and allow cross-

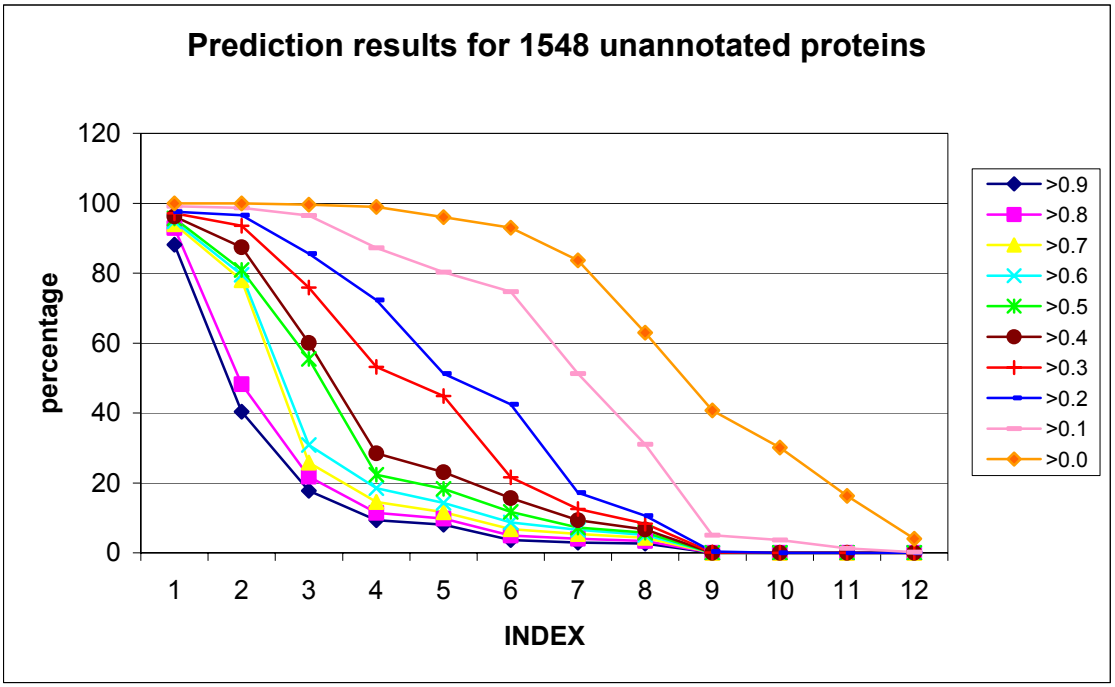

Figure 11. Percentage of 1548 unannotated yeast proteins for which GeneFAS predicted function, against the reliability score cutoff and INDEX level.

Table 4. Percentage of 1548 unannotated yeast proteins for which GeneFAS predicted function, against probability cutoff and INDEX level.

Probability \ INDEX	≥ 0.9	≥ 0.8	≥ 0.7	≥ 0.6	≥ 0.5	≥ 0.4	≥ 0.3	≥ 0.2	≥ 0.1	≥ 0.0
1	20.2	24.9	29.5	36.4	55.8	55.8	92.9	95.6	97.5	100
2	7.88	11	13.8	16.7	24.7	24.7	48.3	80.9	96.6	100
3	0.45	1.03	2.58	4.52	9.63	9.63	21.8	55.4	85.6	99.6
4	0	0	0.06	1.16	4.26	4.26	11.4	22.4	72.4	99
5	0	0	0.71	2.07	4.26	4.26	9.75	18.3	51.3	96
6	0	0	0	0	0.13	0.13	4.97	11.8	42.5	93
7	0	0	0	0	0	0	4.01	7.24	17.2	83.7
8	0	0	0	0	0	0	3.42	5.75	10.6	63
9	0	0	0	0	0	0	0	0	0.32	40.8
10	0	0	0	0	0	0	0	0	0	30.1
11	0	0	0	0	0	0	0	0	0	16.3
12	0	0	0	0	0	0	0	0	0	4.07

species application of the method. We are able to assign function to 1548 out of the 2472 unannotated proteins in yeast.[♣] Our method has an advantage over the other purely computational methods methods, as it involves both, the use of raw data as well as experimental evidences and analysis.

[♣] For detailed list of 1548 unannotated proteins in yeast for which GeneFAS can predict function, refer to Appendix App-2.

5. GeneFAS : Web-based tool

GeneFAS is the web based tool developed for functional annotation of yeast genes using multiple sources of high-throughput data, such as yeast two-hybrid data, genetic interactions data, protein complexes data derived from mass-spectrometry, microarray gene expression data and functional annotation for known yeast proteins (Figure 12). The website is http://compbio.ornl.gov/genefas/index.html (Figure 13)

5.1 Usage

The user can either enter a yeast orf /gene name or enter a protein sequence from any other organism.

1. Select the first option.

Enter yeast ORF name eg. YAL001C or yal001c or Gene name eg. ADE5 or ade5.

Select the appropriate type eg. ORF name or Gene name.

OR

2. Select the second option.

Enter a protein sequence from any organism in raw or fasta format.

Select the E-value, matrix and the number of hits to be displayed from the options.

AND

For either case select an option from the type of high-throughput data to be used.

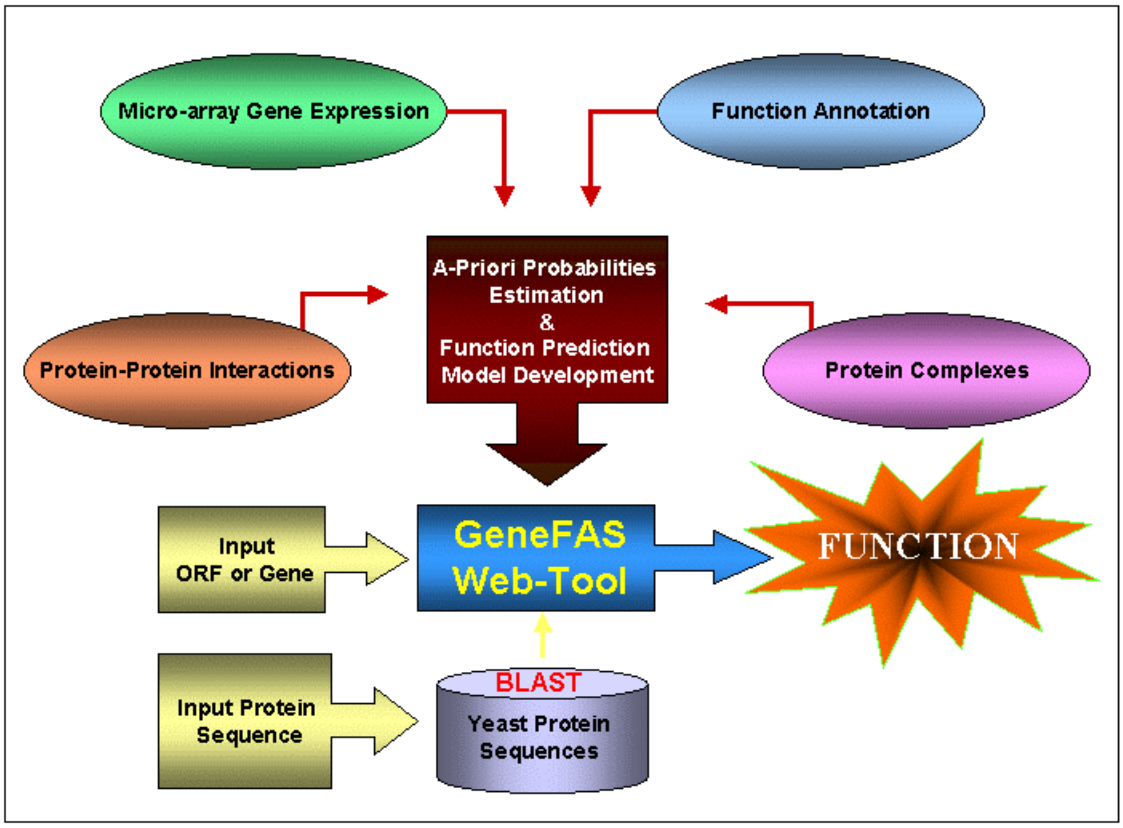

Figure 12. Architecture of GeneFAS.

Figure 13. GeneFAS input interface.

If a complete yeast ORF / gene name is entered the tool outputs the function prediction for the entered gene using the selected high-throughput data to be used for prediction (Figure 14). If a partial matching yeast gene name is entered the tool outputs a list of all possible gene names with the respective orf names and the user can then choose the gene and the high-throughput data type to output the function prediction (Figure 15).

If a protein sequence is entered (Figure 16), the tool blasts the query sequence against the database of all yeast proteins and outputs a list of the hits from the yeast as per the selected E-value, matrix and number of hits. From this list the user can select a yeast protein and the high-throughput data type to predict the function for the selected yeast protein. Each hit is linked to the BLAST alignment for the query sequence and the hit and has the bit score and E-value specified with it.

For each orf /gene the function prediction results, if the gene is already annotated the results specify the GO biological process annotation for the same followed by the function prediction for all the possible indexes based on the type of high-throughput data selected. The annotation is linked to the GO hierarchical tree for that specific GO ID.

Function prediction lists the index, reliability score, probability, function description and the GO id associated with the predicted function. The index is the GO INDEX i.e the series of numbers that represent the hierarchical structure for the GO biological process functions. The reliability score is the predicted score for the orf to have the specified function. The predicted specificity of the method is not the same as that of the expected specificity as is evident from the sensitivity and specificity plots. We state the confidence in the predictions by performing a lookup for the probability value on the basis of the sensitivity and specificity plots. Thus the probability score specifies the level

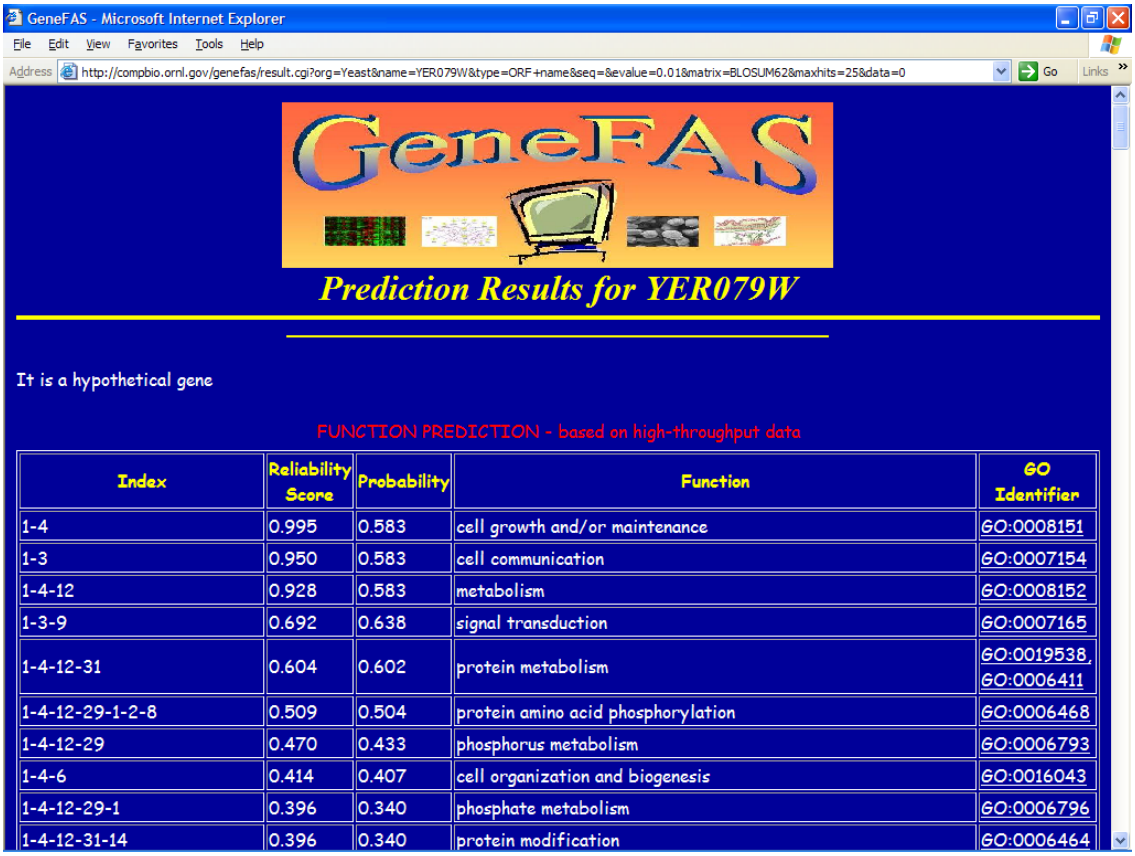

Figure 14. Function prediction results for hypothetical protein YER079W in yeast.

Figure 15. Selection options for partial matching gene names in yeast.

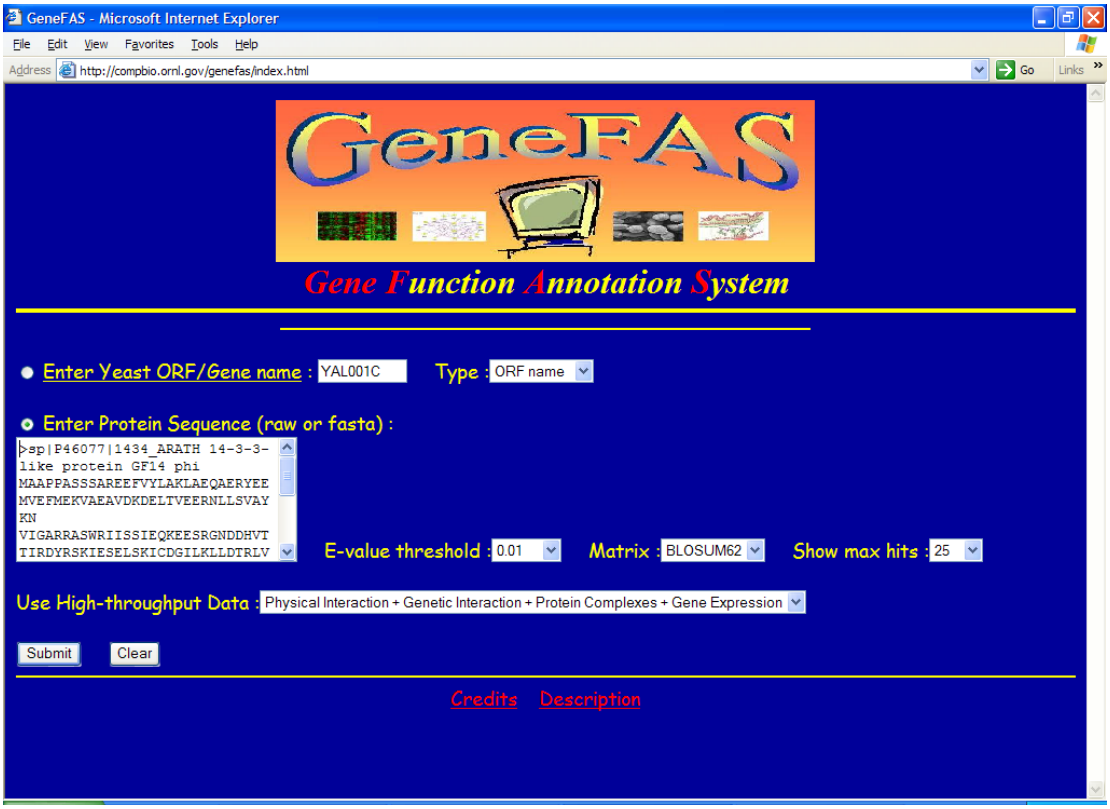

Figure 16. Arabidopsis protein sequence input for function prediction.

of confidence for this prediction. The function description and GO id describe the function represented by the index. The GO id is linked to the GO hierarchical tree for that GO id.

Following the function prediction are the evidences used to predict the function. As per the selected high-throughput data type for the function prediction the supportive evidences in the form of protein interactors from physical binary interactions, genetic interactions, protein complex interactions and microarray gene expression interactors along with their GO annotation and GO id are specified. With the microarray gene expression interactors, the value 1- correlation coefficient is also mentioned. Each of the orf evidence is also linked to the SGD.

6. CROSS-SPECIES APPLICATION

Our aim was to extend our function prediction method developed using yeast high-throughput data, to predict function for hypothetical proteins in other organisms. Toward this end we were interested in exploring the Arabidopsis data, to see how well the analysis matches with that of the Yeast. To achieve this objective we collected the data for Arabidopsis and are interested in finding the homology between Yeast and Arabidopsis.

6.1 ARABIDOPSIS DATA SOURCE

The GO id's for the annotated Arabidopsis genes was obtained from the Gene Ontology website (http://geneontology.org). The GO ids were downloaded from under the current annotations (http://geneontology.org/#annotations), from the TAIR database for the *Arabidopsis thaliana.* The Arabidopsis sequences were downloaded from the TAIR website (http://www.arabidopsis.org/). GO annotations are available for 7526 Arabidopsis genes. We created the Arabidopsis database with protein sequences for 7587 Arabidopsis proteins, downloaded from TAIR.

6.2 RESULTS

We used BLAST and FASTA to find matching hits in Arabidopsis database for every Yeast protein (Table 5). As seen in the plot, the number of pairs identified, are fewer with the FASTA as against the BLAST best hits. The hits were then grouped based

Table 5. Number of Yeast-Arabidopsis homology pairs.

Method	Number of Yeast-Arabidopsis pairs
BLAST (all hits)	89013
BLAST (best hits only)	2415
FASTA (best hits only)	2442

on the E-value ranging from 1e-1 to 1e-100, with the intervals of e-10 each. All the hits with the E-value ≤ the range E-value fall into one category. We collected the GO annotation for the known Arabidopsis genes and applied the same GO INDEX created for the yeast. The yeast Arabidopsis pair in each such category was then compared for percentage of pairs with matching function for each INDEX level. For any Yeast-Arabidopsis pair, if the E-value was ≤ the specified threshold E-value, the pairs fall in the respective category.

Figure 17 shows the probability of Yeast and Arabidopsis best hits pair identified by FASTA, to share the same function. Figure 17 also shows that there is an increase in the probability of Yeast and Arabidopsis pairs sharing the same function with the decrease in the E-value. The probability decreases with an increase in the INDEX level. However, it is important to know the baseline probability for any non-homologous Yeast and Arabidopsis pair, to share the same function. Figure 18 shows the normalized score for the Yeast and Arabidopsis FASTA best hits pair to share the same function, against the probability of non-homologous pairs sharing the same function. Clearly the probability of homologous pairs sharing the same function INDEX level, is significantly more in comparison to that of non-homologous pairs.[♣]

[♣] For detailed information regarding Yeast-Arabidopsis homology pairs identification using BLAST and FASTA, refer to Appendix App-3.

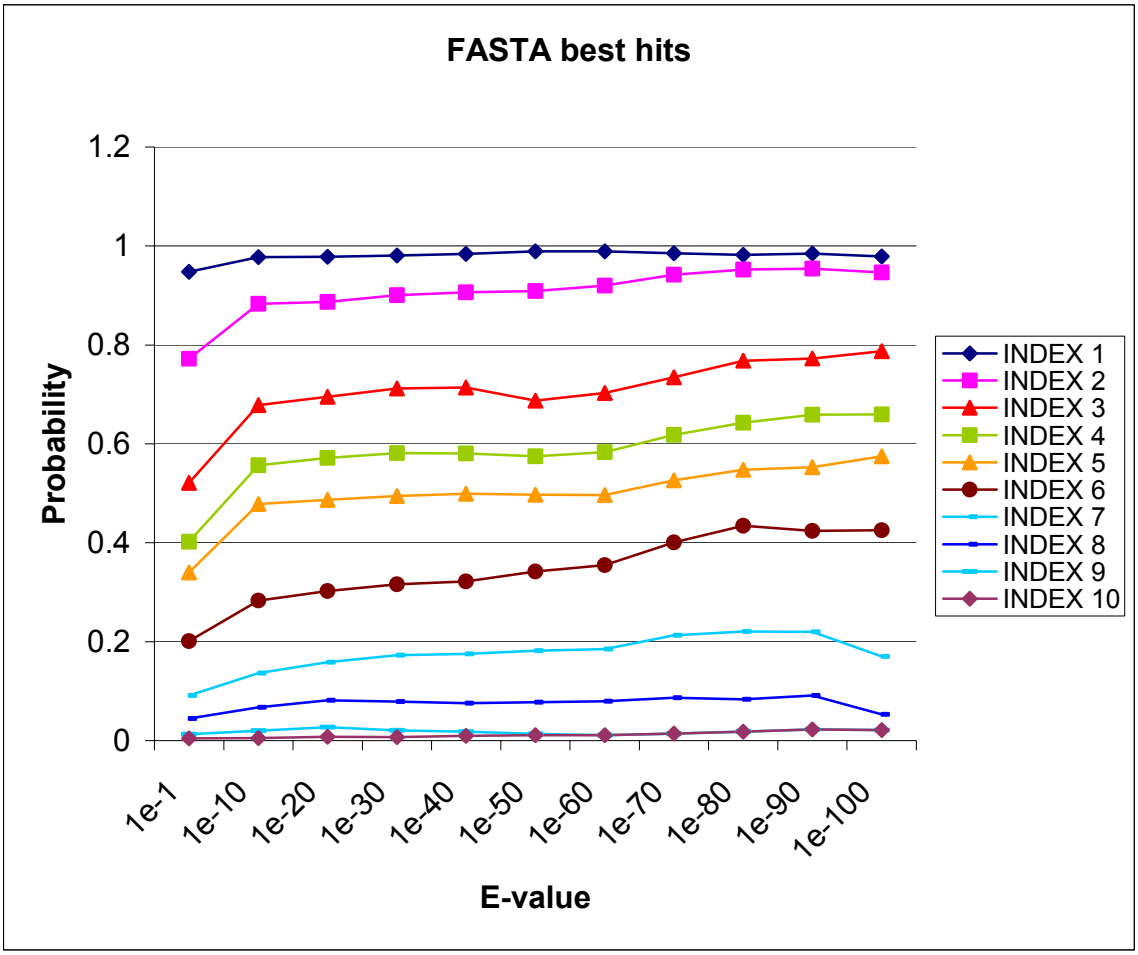

Figure 17. Probability of matching function for INDEX level 1-10 for Yeast and Arabidopsis homology pairs, for FASTA best hit pairs.

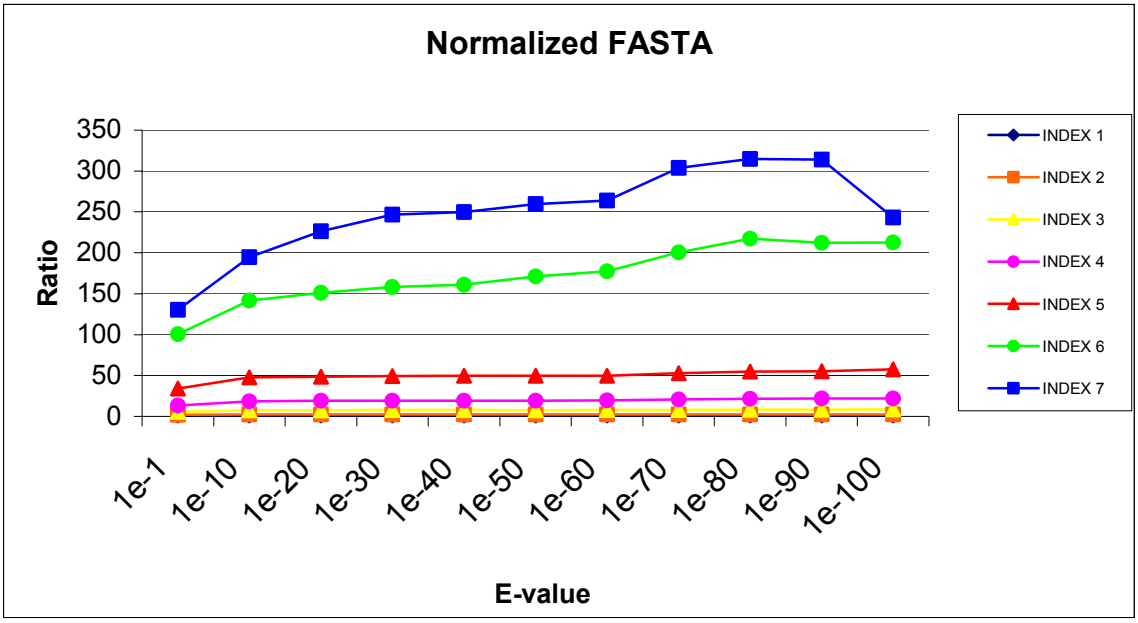

Figure 18. Normalization of FASTA Yeast and Arabidopsis best hit pairs probability of matching function for INDEX level 1-7 for against the probability of non-homologous pairs sharing the same function.

BIBLIOGRAPHY

REFERENCES

1. National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi
2. Pearson W and Lipman D, (1998), Improved Tools for Biological Sequence Comparison. P*roc. Natl Acad. Sci.USA* 85: 2444 - 2448.
3. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W and Lipman D, (1997), Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 25: 3389 - 3402.
4. Marcotte E, Pellegrini M, Ng H, Rice D, Yeates T and Eisenberg D, (1999), Detecting Protein Function and Protein-protein Interactions from Genome Sequences. *Science* 285: 751 - 753.
5. Marcotte E, Pellegrini M, Thompson M, Yeates T and Eisenberg D, (1999), A Combined Algorithm for Genome-wide Prediction of Protein Function. *Nature* 402: 83 - 86.
6. Chien C, Bartel P, Sternglanz R and Fields S, (1991), The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad. Sci. USA,* 88, 9578 - 9582.
7. Eisen M, Spellman P, Brown P and Bostein D, (1998), Cluster Analysis and Display of Genome-wide Expression Patterns. *Proc. Natl. Acad. Sci. USA* 95: 14863 - 14868.
8. Brown M, Grundy W, Lin D, Cristianini N, Sugnet C, Furey T, Ares M, and Haussler D, (2000), Knowledge-based Analysis of Microarray Gene Expression

Data by Using Support Vector Machines *Proc. Natl. Acad. Sci. USA* 97: 262 - 267.

9. Pavlidis P and Weston,J, (2001), Gene Functional Classification from Heterogeneous Data. In *Proceedings of the Fifth International Conference on Computational Molecular Biology (RECOMB2001):* 249 - 255.
10. Schwikowski B, Uetz P and Fields S, (2000), A Network of Protein-protein Interactions in Yeast. *Nature Biotechnology* 18: 1257 - 1261.
11. Hishigaki H, Nakai K, Ono T, Tanigami A and Takagi, (2001), Assessment of Prediction Accuracy of Protein Function from Protein-protein Interaction Data. *Yeast* 18: 523 - 531.
12. Deng M, Zhang K, Mehta S, Chen T, and Sun F, (2002), Prediction of protein function using protein-protein interaction data, In Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB2002). IEEE Computer Society, Los Alamitos, California. Pages 197 - 206.
13. Deng M, Chen T, Sun F, (2003), Integrated Probabilistic Model for Functional Prediction of Proteins, RECOMB2003.
14. Dwight S, Harris M, Dolinski K, Ball C, Binkley G, Christie K *et al*., (2002), *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res*., 30 : 69 - 72.
15. The Gene Ontology Consortium. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* 25 : 25 - 29.

16. Mewes H, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M *et al*., (2002), MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30 : 31 - 34.
17. Roberts C, Nelson B, Marton M, Stoughton R, Meyer M, Bennett H *et al*., (2000), Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*, 287 : 873 - 880.
18. Drawid A and Gerstein M, (2000), A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.* 301 : 1059 - 75.
19. Comprehensive **Perl** Archive Network….. www.cpan.org/
20. XML Extensible Markup Language (**XML**)…. www.w3.org/XML/
21. Tong A, Evangelista M, Parsons A, Xu H, Bader G, Page N, Robinson M, Raghibizadeh S, Hogue C, Bussey H, Andrews B, Tyers M, Boone C, (**2001**), Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science, 294 : 2364 - 2368.
22. Goehring A, Mitchell D, Tong A, Keniry M, Boone C, Sprague G, (2003), Synthetic lethal analysis implicates Ste20p, a p21-activated potein kinase, in polarisome activation. *Mol. Bio. Cell,* 4 : 1501 –1516.
23. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR *et al*., (2000), A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403, 623 - 627.
24. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y. (2001), Toward a protein-protein interaction map of the

budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci.* USA, 98, 4569-4574.

25. Gavin A, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, *et al*., (2002), Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415, 141-147.
26. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, *et al*., (2002), Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415, 180-183.
27. Chen Y, Joshi T, Xu Y, and Xu D, (2003), Towards Automated Derivation of Biological Pathways Using High-Throughput Biological Data, Proceedings of the IEEE Conference on Bioinformatics and Biotechnology, 18-25, IEEE/CS Press.
28. Chen Y and Xu D, (2002), Computation analysis of high-throughput protein-protein interaction data. *Current Peptide and Protein Science*. 4:159-181. 2003.
29. Eisen M, Spellman P, Brown P and Botstein D, (**1998**), Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA, 95, 14863-14868.
30. Stephen Wolfram, The Mathematica Book, 4th ed. (Wolfram Media/ Cambridge University Press, 1999).

WEB LINKS

1. Gene Ontology

http://www.geneontology.org

2. Saccharomyces Genome Database

http://genome-www.stanford.edu/Saccharomyces/

3. MIPS

http://mips.gsf.de/proj/yeast/CYGD/db/index.html

4. Yeast Protein Localization Server

http://bioinfo.mbb.yale.edu/genome/localize/

APPENDIX

App-1. CROSS-HYBRIDIZATION.

We compared the microarray data plots before and after the removal of pairs with ≥ 85 percent sequence similarity, to check if there is any effect of cross-hybridization on the microarray data (Figure A-1). To our observation, we encountered no significant difference in the values before and after removal of pairs with ≥ 85 percent sequence similarity, except in the first category of 1-correlation coefficient value 0.025. This difference is primarily due to the very small data size in this category. For all the other categories the values are nearly the same before and after removal of microarray pairs with ≥ 85 percent sequence similarity (Figure A-1).

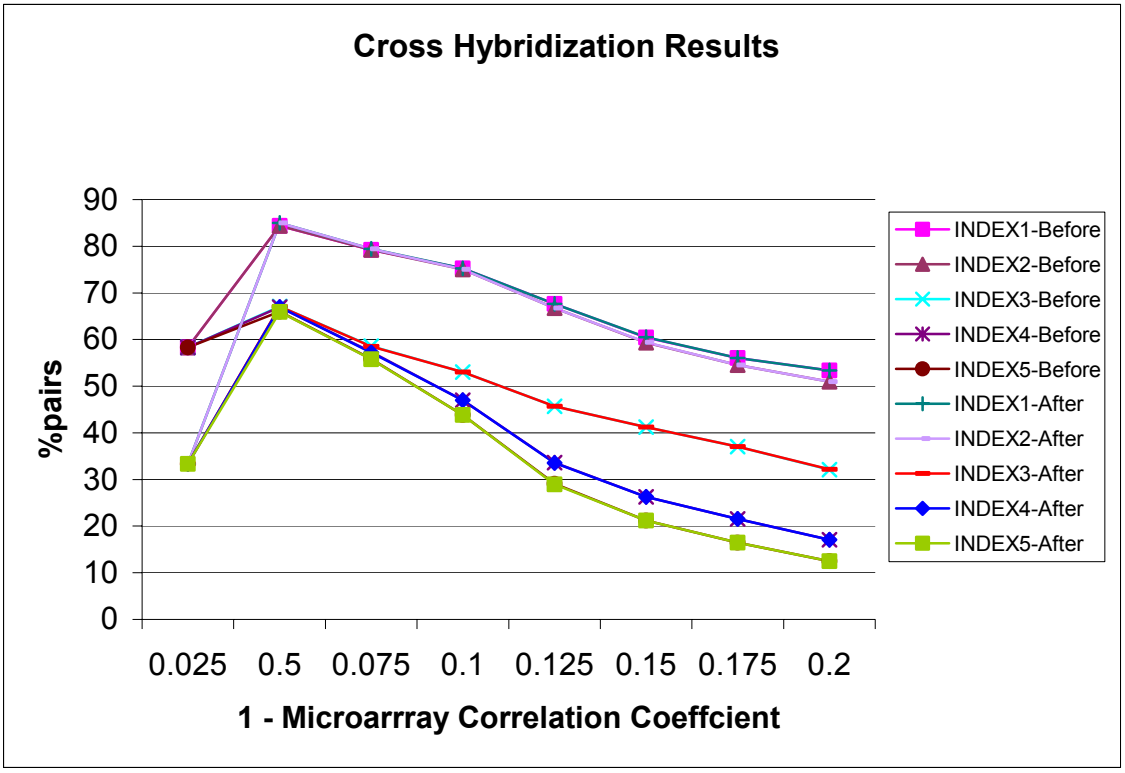

Figure A-1. Plot of microarray 1-correlation coefficient for percentage of pairs with matching function for INDEX levels 1 to 5, before and after removal of pairs with ≥ 85 percent sequence similarity.

App-2. LIST OF 1548 UNANNOTATED PROTEINS FOR WHICH GENEFAS CAN PREDICT FUNCTION.

YAL011W	YBL081W	YBR138C	YBR281C	YCR095C	YDL167C	YDR111C
YAL014C	YBL086C	YBR139W	YBR284W	YCR099C	YDL172C	YDR117C
YAL017W	YBL094C	YBR141C	YBR285W	YCR101C	YDL178W	YDR119W
YAL027W	YBL095W	YBR144C	YBR287W	YDL001W	YDL183C	YDR122W
YAL028W	YBL101W-A	YBR147W	YBR293W	YDL002C	YDL186W	YDR124W
YAL034C	YBL104C	YBR150C	YBR300C	YDL011C	YDL193W	YDR128W
YAL036C	YBL107C	YBR157C	YBR301W	YDL012C	YDL199C	YDR130C
YAL045C	YBL108W	YBR158W	YBR302C	YDL023C	YDL203C	YDR131C
YAL049C	YBL109W	YBR161W	YCL005W	YDL025C	YDL204W	YDR132C
YAL053W	YBL113C	YBR162C	YCL011C	YDL026W	YDL211C	YDR133C
YAL061W	YBR004C	YBR168W	YCL019W	YDL027C	YDL213C	YDR140W
YAL064W	YBR012W-A	YBR174C	YCL020W	YDL033C	YDL218W	YDR152W
YAL065C	YBR012W-B	YBR184W	YCL022C	YDL037C	YDL221W	YDR154C
YAL066W	YBR013C	YBR187W	YCL023C	YDL041W	YDL233W	YDR161W
YAR009C	YBR014C	YBR190W	YCL028W	YDL046W	YDL237W	YDR162C
YAR018C	YBR025C	YBR197C	YCL039W	YDL050C	YDL241W	YDR169C
YAR027W	YBR027C	YBR203W	YCL042W	YDL053C	YDL246C	YDR170W-A
YAR030C	YBR028C	YBR204C	YCL044C	YDL062W	YDR003W	YDR184C
YAR042W	YBR042C	YBR209W	YCL045C	YDL063C	YDR008C	YDR193W
YAR047C	YBR043C	YBR216C	YCL046W	YDL070W	YDR010C	YDR196C
YAR061W	YBR046C	YBR219C	YCL049C	YDL071C	YDR018C	YDR198C
YAR064W	YBR047W	YBR220C	YCL056C	YDL072C	YDR020C	YDR199W
YAR066W	YBR051W	YBR225W	YCL058C	YDL076C	YDR026C	YDR203W
YAR069C	YBR052C	YBR226C	YCL063W	YDL086W	YDR031W	YDR215C
YAR075W	YBR053C	YBR227C	YCL069W	YDL089W	YDR032C	YDR219C
YBL004W	YBR054W	YBR230C	YCL076W	YDL091C	YDR042C	YDR223W
YBL009W	YBR056W	YBR231C	YCR001W	YDL096C	YDR049W	YDR229W
YBL028C	YBR059C	YBR233W	YCR007C	YDL099W	YDR056C	YDR230W
YBL029W	YBR063C	YBR238C	YCR013C	YDL100C	YDR061W	YDR233C
YBL031W	YBR064W	YBR239C	YCR015C	YDL104C	YDR063W	YDR239C
YBL036C	YBR074W	YBR241C	YCR016W	YDL105W	YDR067C	YDR247W
YBL044W	YBR077C	YBR242W	YCR022C	YDL110C	YDR068W	YDR249C
YBL046W	YBR089W	YBR246W	YCR030C	YDL115C	YDR070C	YDR255C
YBL048W	YBR094W	YBR250W	YCR041W	YDL118W	YDR071C	YDR262W
YBL049W	YBR096W	YBR255W	YCR043C	YDL121C	YDR078C	YDR266C
YBL051C	YBR099C	YBR261C	YCR045C	YDL124W	YDR084C	YDR267C
YBL053W	YBR101C	YBR262C	YCR050C	YDL129W	YDR091C	YDR269C
YBL054W	YBR108W	YBR267W	YCR072C	YDL133W	YDR095C	YDR271C
YBL059W	YBR113W	YBR269C	YCR076C	YDL139C	YDR100W	YDR275W
YBL065W	YBR116C	YBR270C	YCR079W	YDL144C	YDR101C	YDR279W
YBL066C	YBR124W	YBR271W	YCR082W	YDL152W	YDR102C	YDR282C
YBL067C	YBR125C	YBR273C	YCR087C-A	YDL156W	YDR105C	YDR286C
YBL071C	YBR134W	YBR277C	YCR087W	YDL158C	YDR106W	YDR287W
YBL077W	YBR137W	YBR280C	YCR091W	YDL162C	YDR107C	YDR290W

YDR295C	YDR520C	YER064C	YFL054C	YGL128C	YGR058W	YGR223C
YDR304C	YDR526C	YER066C-A	YFL061W	YGL131C	YGR064W	YGR228W
YDR306C	YDR527W	YER066W	YFL067W	YGL132W	YGR066C	YGR235C
YDR307W	YDR531W	YER067W	YFL068W	YGL138C	YGR067C	YGR236C
YDR314C	YDR532C	YER071C	YFR003C	YGL139W	YGR068C	YGR237C
YDR317W	YDR533C	YER076C	YFR006W	YGL146C	YGR069W	YGR242W
YDR319C	YDR541C	YER077C	YFR008W	YGL149W	YGR071C	YGR243W
YDR326C	YDR542W	YER078C	YFR011C	YGL157W	YGR073C	YGR248W
YDR330W	YEL005C	YER079W	YFR012W	YGL161C	YGR079W	YGR250C
YDR333C	YEL008W	YER084W	YFR016C	YGL165C	YGR081C	YGR251W
YDR339C	YEL015W	YER087C-A	YFR017C	YGL168W	YGR086C	YGR259C
YDR340W	YEL016C	YER087W	YFR018C	YGL177W	YGR090W	YGR263C
YDR344C	YEL017W	YER092W	YFR024C-A	YGL179C	YGR100W	YGR265W
YDR346C	YEL018W	YER093C-A	YFR024C	YGL185C	YGR102C	YGR266W
YDR348C	YEL020C	YER101C	YFR026C	YGL188C	YGR106C	YGR268C
YDR350C	YEL023C	YER113C	YFR038W	YGL193C	YGR110W	YGR269W
YDR357C	YEL033W	YER121W	YFR039C	YGL196W	YGR111W	YGR272C
YDR359C	YEL038W	YER128W	YFR042W	YGL198W	YGR114C	YGR275W
YDR361C	YEL041W	YER138C	YFR043C	YGL204C	YGR115C	YGR277C
YDR365C	YEL043W	YER139C	YFR044C	YGL217C	YGR117C	YGR278W
YDR366C	YEL048C	YER140W	YFR046C	YGL220W	YGR122W	YGR283C
YDR367W	YEL049W	YER150W	YFR054C	YGL221C	YGR127W	YGR290W
YDR371W	YEL057C	YER156C	YFR057W	YGL226W	YGR130C	YGR291C
YDR373W	YEL059W	YER158C	YGL004C	YGL230C	YGR136W	YGR293C
YDR383C	YEL068C	YER160C	YGL010W	YGL231C	YGR139W	YGR294W
YDR412W	YEL070W	YER163C	YGL015C	YGL232W	YGR145W	YGR295C
YDR413C	YEL072W	YER170W	YGL021W	YGL239C	YGR146C	YGR296W
YDR415C	YEL074W	YER175C	YGL024W	YGL242C	YGR149W	YHL005C
YDR417C	YEL075C	YER181C	YGL036W	YGL245W	YGR150C	YHL006C
YDR425W	YEL076W-C	YER182W	YGL039W	YGL250W	YGR151C	YHL008C
YDR428C	YEL077C	YER184C	YGL046W	YGL258W	YGR153W	YHL010C
YDR430C	YER004W	YER186C	YGL047W	YGL259W	YGR154C	YHL013C
YDR433W	YER006W	YER189W	YGL052W	YGL261C	YGR160W	YHL017W
YDR438W	YER007C-A	YFL006W	YGL057C	YGL263W	YGR163W	YHL018W
YDR441C	YER010C	YFL007W	YGL059W	YGR002C	YGR168C	YHL021C
YDR444W	YER030W	YFL010C	YGL060W	YGR004W	YGR173W	YHL023C
YDR445C	YER036C	YFL012W	YGL068W	YGR016W	YGR182C	YHL026C
YDR455C	YER037W	YFL013C	YGL069C	YGR017W	YGR187C	YHL035C
YDR459C	YER038C	YFL015C	YGL074C	YGR018C	YGR189C	YHL039W
YDR474C	YER039C	YFL020C	YGL079W	YGR024C	YGR196C	YHL042W
YDR479C	YER041W	YFL030W	YGL081W	YGR031W	YGR198W	YHL044W
YDR482C	YER047C	YFL034W	YGL083W	YGR033C	YGR201C	YHL045W
YDR489W	YER048C	YFL040W	YGL091C	YGR035C	YGR203W	YHL046C
YDR496C	YER049W	YFL042C	YGL096W	YGR042W	YGR205W	YHL049C
YDR505C	YER051W	YFL044C	YGL101W	YGR043C	YGR207C	YHR003C
YDR512C	YER053C	YFL049W	YGL102C	YGR046W	YGR210C	YHR009C
YDR514C	YER054C	YFL051C	YGL110C	YGR052W	YGR213C	YHR020W
YDR516C	YER057C	YFL052W	YGL117W	YGR053C	YGR219W	YHR022C

YHR033W	YHR214W-A	YIR014W	YJL175W	YJR116W	YKL153W	YLL023C
YHR034C	YHR214W	YIR016W	YJL178C	YJR119C	YKL161C	YLL025W
YHR035W	YHR215W	YIR024C	YJL182C	YJR120W	YKL168C	YLL029W
YHR036W	YIL001W	YIR035C	YJL184W	YJR122W	YKL174C	YLL030C
YHR040W	YIL007C	YIR036C	YJL185C	YJR124C	YKL177W	YLL033W
YHR045W	YIL011W	YIR039C	YJL192C	YJR127C	YKL183W	YLL034C
YHR047C	YIL017C	YIR040C	YJL197W	YJR134C	YKL187C	YLL037W
YHR049W	YIL019W	YIR042C	YJL199C	YJR136C	YKL195W	YLL044W
YHR057C	YIL023C	YIR044C	YJL200C	YJR138W	YKL202W	YLL049W
YHR059W	YIL025C	YJL010C	YJL202C	YJR141W	YKL206C	YLL051C
YHR063C	YIL027C	YJL015C	YJL207C	YJR146W	YKL215C	YLL053C
YHR074W	YIL028W	YJL017W	YJL211C	YJR154W	YKL218C	YLL054C
YHR076W	YIL032C	YJL023C	YJL213W	YJR157W	YKL222C	YLL065W
YHR080C	YIL039W	YJL032W	YJL215C	YJR162C	YKL224C	YLR001C
YHR081W	YIL040W	YJL048C	YJL218W	YKL014C	YKR005C	YLR003C
YHR083W	YIL045W	YJL057C	YJL225C	YKL023W	YKR007W	YLR004C
YHR085W	YIL054W	YJL058C	YJR003C	YKL034W	YKR011C	YLR008C
YHR087W	YIL055C	YJL064W	YJR008W	YKL036C	YKR016W	YLR011W
YHR097C	YIL056W	YJL065C	YJR011C	YKL039W	YKR017C	YLR016C
YHR100C	YIL057C	YJL066C	YJR012C	YKL044W	YKR018C	YLR020C
YHR105W	YIL059C	YJL067W	YJR014W	YKL047W	YKR021W	YLR021W
YHR112C	YIL060W	YJL069C	YJR015W	YKL050C	YKR022C	YLR030W
YHR113W	YIL077C	YJL070C	YJR024C	YKL051W	YKR030W	YLR031W
YHR115C	YIL086C	YJL072C	YJR026W	YKL056C	YKR032W	YLR035C-A
YHR117W	YIL087C	YJL078C	YJR027W	YKL061W	YKR033C	YLR037C
YHR121W	YIL091C	YJL079C	YJR028W	YKL063C	YKR038C	YLR049C
YHR122W	YIL092W	YJL082W	YJR029W	YKL065C	YKR040C	YLR050C
YHR127W	YIL096C	YJL084C	YJR030C	YKL067W	YKR043C	YLR051C
YHR130C	YIL097W	YJL086C	YJR037W	YKL069W	YKR046C	YLR052W
YHR133C	YIL103W	YJL091C	YJR038C	YKL072W	YKR049C	YLR053C
YHR140W	YIL105C	YJL097W	YJR041C	YKL075C	YKR051W	YLR054C
YHR145C	YIL108W	YJL103C	YJR056C	YKL076C	YKR060W	YLR057W
YHR149C	YIL110W	YJL107C	YJR067C	YKL077W	YKR070W	YLR063W
YHR156C	YIL127C	YJL114W	YJR070C	YKL082C	YKR071C	YLR064W
YHR168W	YIL130W	YJL120W	YJR071W	YKL086W	YKR073C	YLR065C
YHR177W	YIL135C	YJL122W	YJR072C	YKL088W	YKR079C	YLR068W
YHR179W	YIL136W	YJL131C	YJR079W	YKL090W	YKR087C	YLR072W
YHR180W	YIL137C	YJL132W	YJR080C	YKL091C	YKR088C	YLR073C
YHR186C	YIL141W	YJL135W	YJR082C	YKL095W	YKR089C	YLR076C
YHR192W	YIL151C	YJL142C	YJR083C	YKL102C	YKR090W	YLR077W
YHR195W	YIL152W	YJL144W	YJR084W	YKL107W	YKR096W	YLR090W
YHR197W	YIL157C	YJL149W	YJR087W	YKL111C	YKR100C	YLR095C
YHR198C	YIL158W	YJL151C	YJR097W	YKL123W	YKR105C	YLR096W
YHR199C	YIL163C	YJL152W	YJR098C	YKL124W	YLL012W	YLR097C
YHR207C	YIL169C	YJL160C	YJR100C	YKL128C	YLL014W	YLR101C
YHR209W	YIL174W	YJL161W	YJR108W	YKL133C	YLL017W	YLR104W
YHR212C	YIL177C	YJL162C	YJR110W	YKL137W	YLL020C	YLR108C
YHR214C-B	YIR003W	YJL163C	YJR115W	YKL151C	YLL022C	YLR112W

YLR114C	YLR257W	YLR437C	YMR044W	YMR233W	YNL112W	YNL266W
YLR118C	YLR266C	YLR440C	YMR046C	YMR237W	YNL114C	YNL274C
YLR123C	YLR267W	YLR446W	YMR051C	YMR244W	YNL115C	YNL276C
YLR124W	YLR269C	YLR449W	YMR057C	YMR253C	YNL116W	YNL285W
YLR125W	YLR271W	YLR455W	YMR067C	YMR265C	YNL119W	YNL300W
YLR128W	YLR283W	YLR456W	YMR071C	YMR266W	YNL120C	YNL303W
YLR132C	YLR285W	YLR457C	YMR075W	YMR269W	YNL123W	YNL305C
YLR136C	YLR289W	YLR459W	YMR086W	YMR278W	YNL127W	YNL311C
YLR137W	YLR290C	YLR460C	YMR087W	YMR279C	YNL132W	YNL313C
YLR140W	YLR294C	YLR461W	YMR088C	YMR289W	YNL133C	YNL319W
YLR149C	YLR297W	YLR465C	YMR090W	YMR290C	YNL134C	YNL320W
YLR151C	YLR301W	YLR466W	YMR097C	YMR291W	YNL135C	YNL321W
YLR152C	YLR311C	YML011C	YMR102C	YMR295C	YNL144C	YNL324W
YLR154C	YLR312C	YML013W	YMR107W	YMR298W	YNL146W	YNL326C
YLR156W	YLR315W	YML014W	YMR110C	YMR304C-A	YNL152W	YNL335W
YLR164W	YLR323C	YML018C	YMR114C	YMR310C	YNL156C	YNR004W
YLR168C	YLR324W	YML020W	YMR116C	YMR315W	YNL157W	YNR005C
YLR171W	YLR326W	YML023C	YMR118C	YMR316C-B	YNL158W	YNR009W
YLR177W	YLR327C	YML027W	YMR123W	YMR317W	YNL165W	YNR024W
YLR181C	YLR331C	YML029W	YMR124W	YMR321C	YNL171C	YNR025C
YLR183C	YLR338W	YML030W	YMR130W	YMR322C	YNL174W	YNR028W
YLR187W	YLR339C	YML036W	YMR134W	YMR323W	YNL175C	YNR036C
YLR190W	YLR343W	YML040W	YMR135C	YMR325W	YNL176C	YNR040W
YLR193C	YLR345W	YML053C	YMR141C	YNL010W	YNL181W	YNR042W
YLR196W	YLR346C	YML056C	YMR144W	YNL013C	YNL182C	YNR046W
YLR198C	YLR350W	YML059C	YMR147W	YNL018C	YNL187W	YNR048W
YLR199C	YLR351C	YML072C	YMR148W	YNL022C	YNL190W	YNR051C
YLR201C	YLR352W	YML074C	YMR151W	YNL023C	YNL191W	YNR054C
YLR202C	YLR356W	YML079W	YMR155W	YNL024C	YNL194C	YNR061C
YLR211C	YLR358C	YML082W	YMR157C	YNL026W	YNL195C	YNR065C
YLR213C	YLR365W	YML089C	YMR160W	YNL028W	YNL196C	YNR066C
YLR215C	YLR366W	YML101C	YMR163C	YNL034W	YNL200C	YNR068C
YLR217W	YLR376C	YML108W	YMR171C	YNL035C	YNL201C	YNR069C
YLR218C	YLR379W	YML117W	YMR173W-A	YNL036W	YNL203C	YNR071C
YLR221C	YLR387C	YML118W	YMR178W	YNL040W	YNL205C	YNR073C
YLR224W	YLR392C	YML125C	YMR181C	YNL042W	YNL207W	YOL002C
YLR225C	YLR400W	YML131W	YMR184W	YNL043C	YNL208W	YOL003C
YLR230W	YLR409C	YML132W	YMR187C	YNL046W	YNL211C	YOL008W
YLR235C	YLR412W	YML133C	YMR192W	YNL047C	YNL212W	YOL014W
YLR236C	YLR413W	YMR002W	YMR193C-A	YNL050C	YNL215W	YOL015W
YLR238W	YLR416C	YMR003W	YMR196W	YNL056W	YNL224C	YOL022C
YLR241W	YLR419W	YMR009W	YMR204C	YNL058C	YNL226W	YOL030W
YLR243W	YLR422W	YMR018W	YMR206W	YNL063W	YNL228W	YOL032W
YLR247C	YLR424W	YMR019W	YMR209C	YNL086W	YNL235C	YOL036W
YLR251W	YLR426W	YMR027W	YMR210W	YNL087W	YNL237W	YOL042W
YLR252W	YLR427W	YMR029C	YMR215W	YNL092W	YNL253W	YOL045W
YLR254C	YLR434C	YMR030W	YMR222C	YNL105W	YNL260C	YOL046C
YLR255C	YLR435W	YMR041C	YMR226C	YNL109W	YNL265C	YOL048C

YOL050C	YOR090C	YOR286W	YPL093W	YPR031W
YOL053W	YOR091W	YOR287C	YPL095C	YPR038W
YOL063C	YOR097C	YOR289W	YPL098C	YPR044C
YOL070C	YOR102W	YOR292C	YPL102C	YPR045C
YOL073C	YOR104W	YOR298W	YPL105C	YPR048W
YOL083W	YOR105W	YOR302W	YPL107W	YPR050C
YOL084W	YOR111W	YOR309C	YPL108W	YPR053C
YOL085C	YOR112W	YOR311C	YPL109C	YPR059C
YOL087C	YOR114W	YOR314W	YPL110C	YPR061C
YOL093W	YOR121C	YOR315W	YPL113C	YPR076W
YOL098C	YOR138C	YOR318C	YPL114W	YPR077C
YOL099C	YOR141C	YOR324C	YPL136W	YPR084W
YOL101C	YOR145C	YOR325W	YPL142C	YPR085C
YOL106W	YOR146W	YOR331C	YPL146C	YPR090W
YOL107W	YOR154W	YOR333C	YPL150W	YPR093C
YOL109W	YOR155C	YOR345C	YPL158C	YPR096C
YOL111C	YOR161C	YOR350C	YPL159C	YPR098C
YOL118C	YOR164C	YOR352W	YPL165C	YPR099C
YOL124C	YOR169C	YOR353C	YPL166W	YPR106W
YOL131W	YOR172W	YOR364W	YPL170W	YPR114W
YOL132W	YOR173W	YOR367W	YPL171C	YPR115W
YOL137W	YOR175C	YOR378W	YPL181W	YPR118W
YOL146W	YOR179C	YOR379C	YPL183C	YPR126C
YOL150C	YOR186W	YOR385W	YPL186C	YPR127W
YOL152W	YOR189W	YOR387C	YPL201C	YPR130C
YOL154W	YOR205C	YOR389W	YPL205C	YPR136C
YOR004W	YOR206W	YOR390W	YPL207W	YPR143W
YOR006C	YOR214C	YOR392W	YPL208W	YPR144C
YOR007C	YOR215C	YOR393W	YPL216W	YPR148C
YOR009W	YOR220W	YPL004C	YPL222W	YPR152C
YOR019W	YOR225W	YPL005W	YPL226W	YPR157W
YOR021C	YOR227W	YPL009C	YPL229W	YPR169W
YOR034C	YOR228C	YPL014W	YPL230W	YPR171W
YOR042W	YOR243C	YPL024W	YPL236C	YPR172W
YOR050C	YOR246C	YPL025C	YPL238C	YPR174C
YOR051C	YOR248W	YPL034W	YPL245W	YPR179C
YOR053W	YOR251C	YPL044C	YPL246C	YPR188C
YOR054C	YOR252W	YPL052W	YPL247C	YPR203W
YOR059C	YOR258W	YPL054W	YPL251W	Q0032
YOR060C	YOR263C	YPL056C	YPL260W	Q0092
YOR062C	YOR264W	YPL064C	YPL263C	
YOR066W	YOR271C	YPL066W	YPL276W	
YOR068C	YOR277C	YPL068C	YPL280W	
YOR072W	YOR279C	YPL070W	YPL283C	
YOR073W	YOR282W	YPL071C	YPR003C	
YOR082C	YOR283W	YPL073C	YPR004C	
YOR086C	YOR284W	YPL074W	YPR014C	
YOR088W	YOR285W	YPL077C	YPR015C	

App-3. YEAST-ARABIDOPSIS HOMOLOGY PAIRS IDENTIFICATION.

Figure A-2 shows the results for BLAST with all hits and best hits only. Figure A-2 shows that there is an increase in the probability of Yeast and Arabidopsis pairs sharing the same function with the decrease in the E-value. The probability decreases with an increase in the INDEX level. The probability is better when only the best-hits pairs are considered. Figure A-3 shows the distribution of the total number of all Yeast and Arabidopsis pairs identified by BLAST.

We also compared the results of FASTA with BLAST for the best hits pairs. Figure A-4 shows that in addition to the above observations, the probability of sharing the same function is more with the Yeast and Arabidopsis best hits pair identified by FASTA as against that identified by BLAST. Figure A-5 shows the distribution of the total number of best hits Yeast and Arabidopsis pairs identified by FASTA and BLAST. In addition to these results, it is also important to know the baseline probability for any non-homologous Yeast and Arabidopsis pair, to share the same function (Figure A-6). It is evident from this plot that the probability of any non-homologous Yeast and Arabidopsis pair, to share the same function also decreases with the increase in the INDEX level. Clearly, the probability of homologous pairs sharing the same function INDEX level is significantly more in comparison to that of non-homologous pairs.

Yet we were not very satisfied with the results from FASTA and BLAST based on the E-value. So we instead decided to use a different approach based on the percentage of the sequence identity between the pairs identified by BLAST. Figure A-7

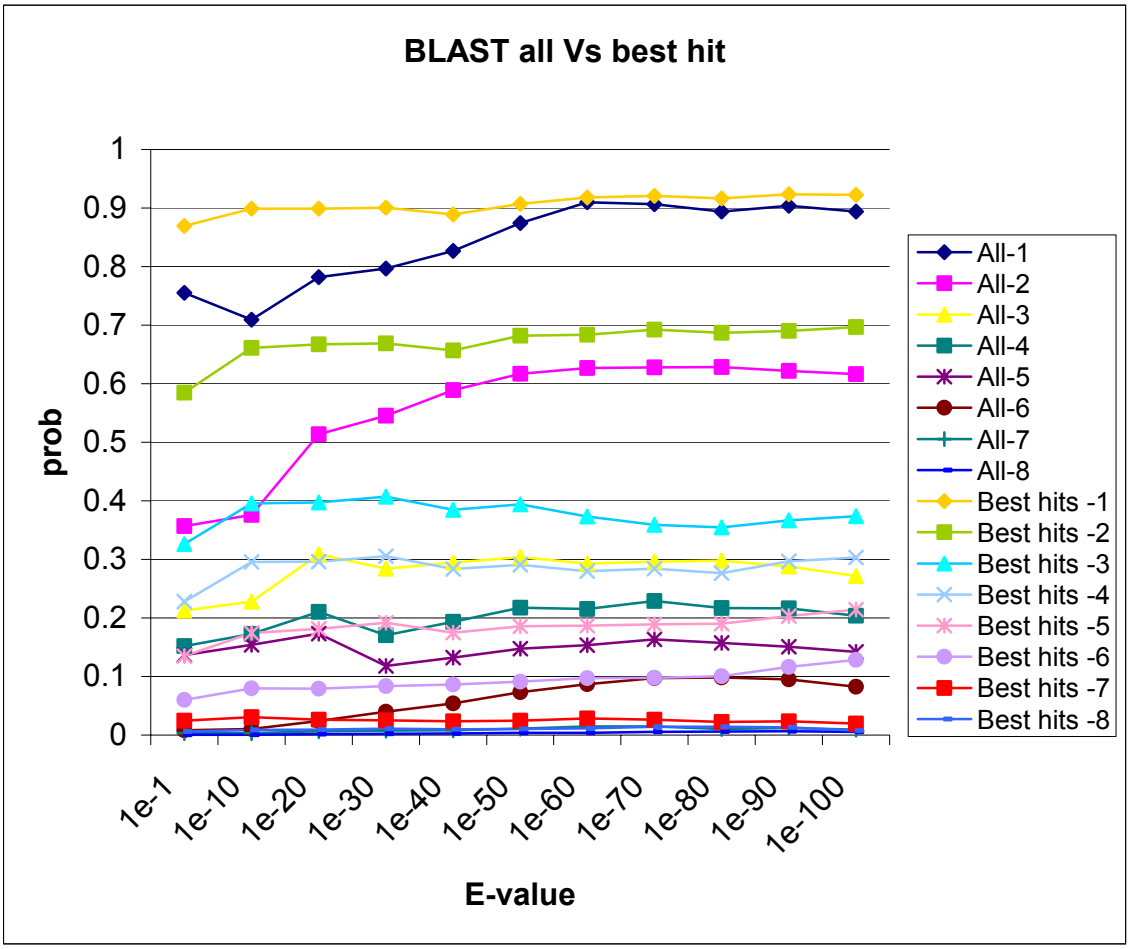

Figure A-2. Probability of matching function for INDEX level 1-8 for Yeast and Arabidopsis homology pairs, for BLAST all against best hits pairs.

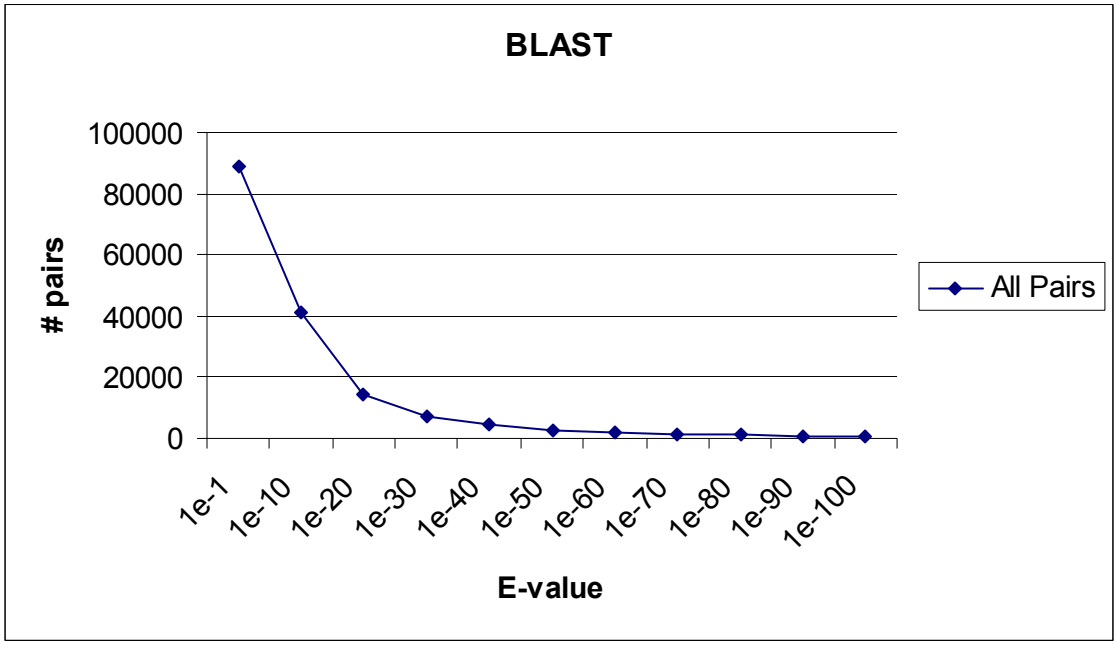

Figure A-3. Distribution of Yeast-Arabidopsis homology pairs in each E-value range for all BLAST-hit pairs.

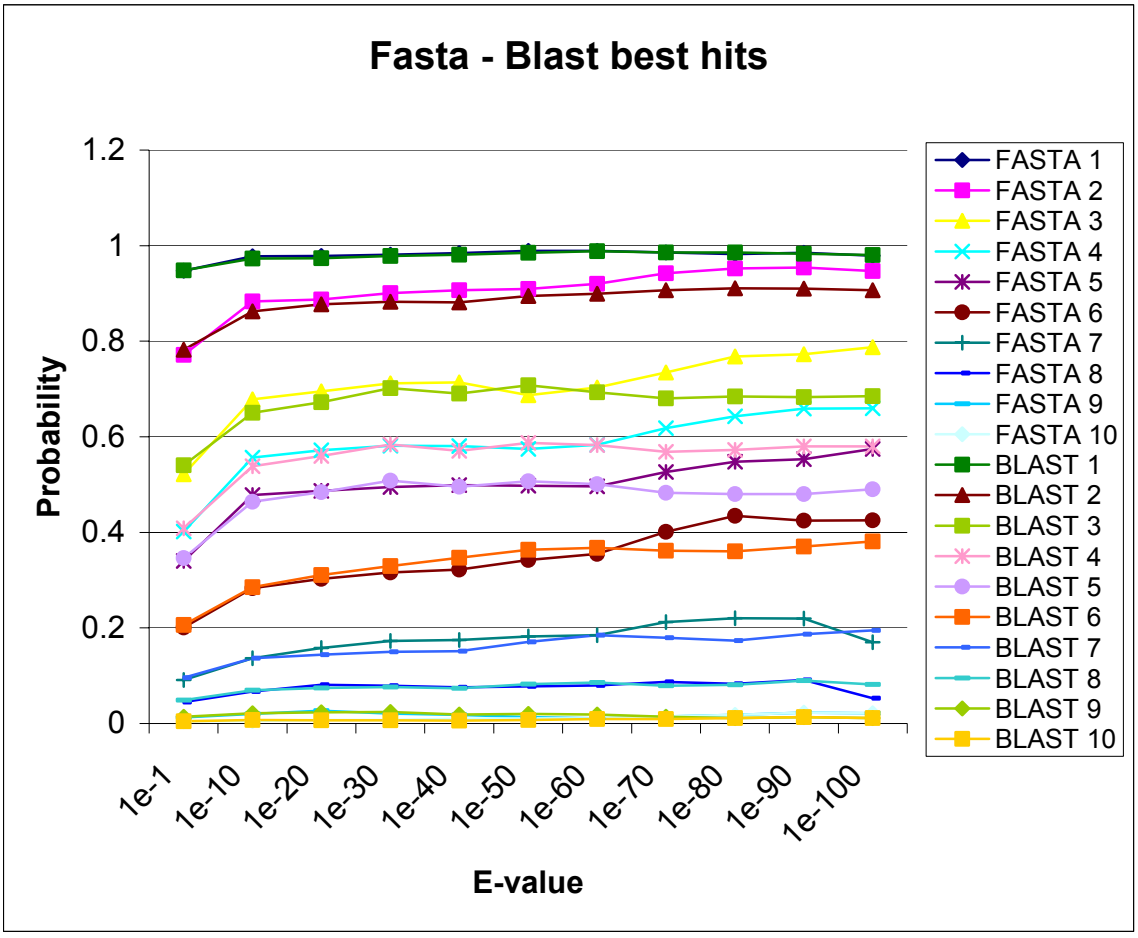

Figure A-4. Probability of matching function for INDEX level 1-10 for Yeast and Arabidopsis homology pairs, for FASTA and BLAST best hit pairs.

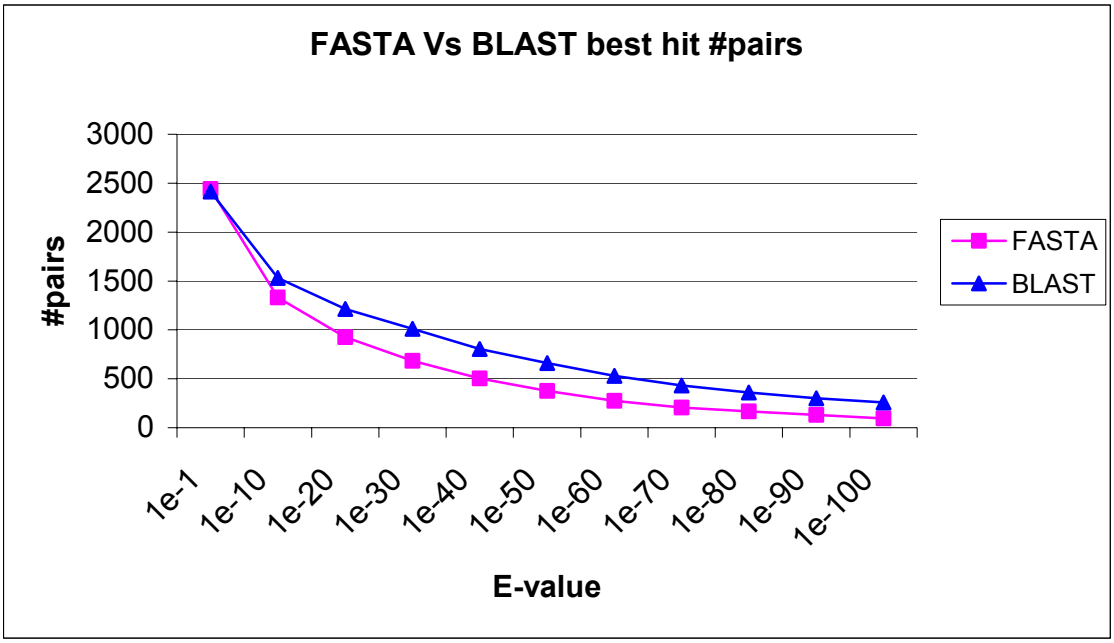

Figure A-5. Distribution of Yeast-Arabidopsis pairs in each E-value range for FASTA and BLAST best hits pairs.

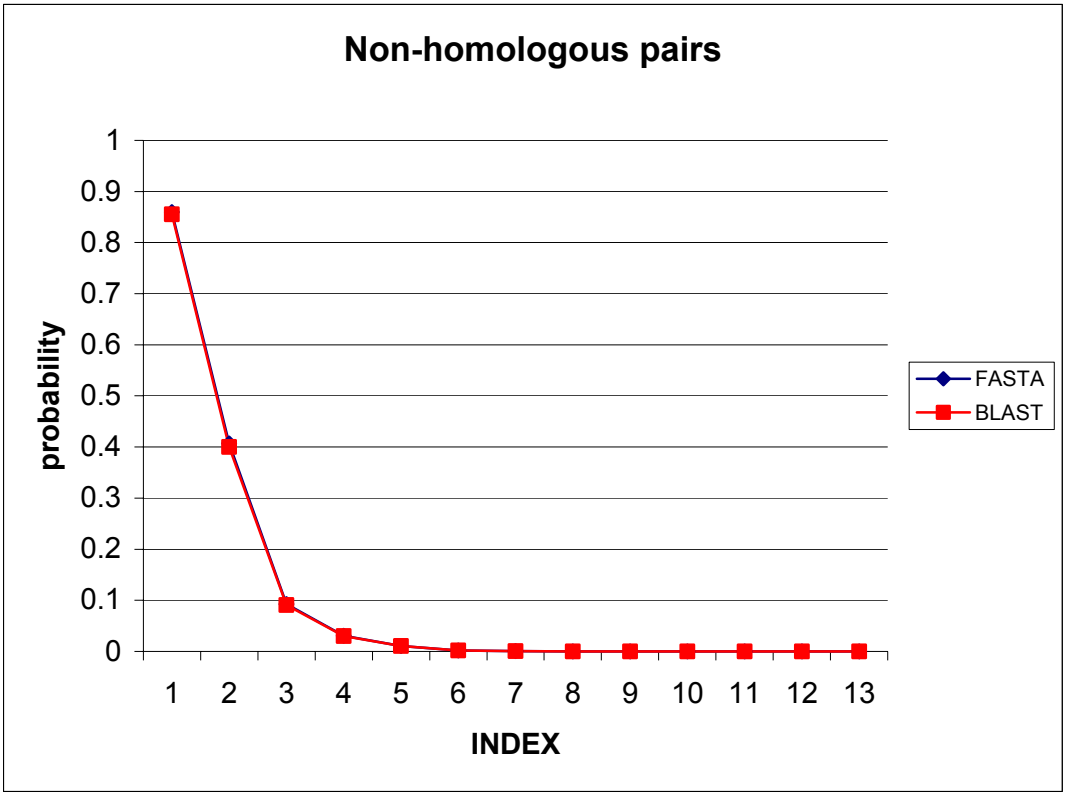

Figure A-6. Probability of matching function for INDEX level 1-13 for Yeast-Arabidopsis non-homologous pairs.

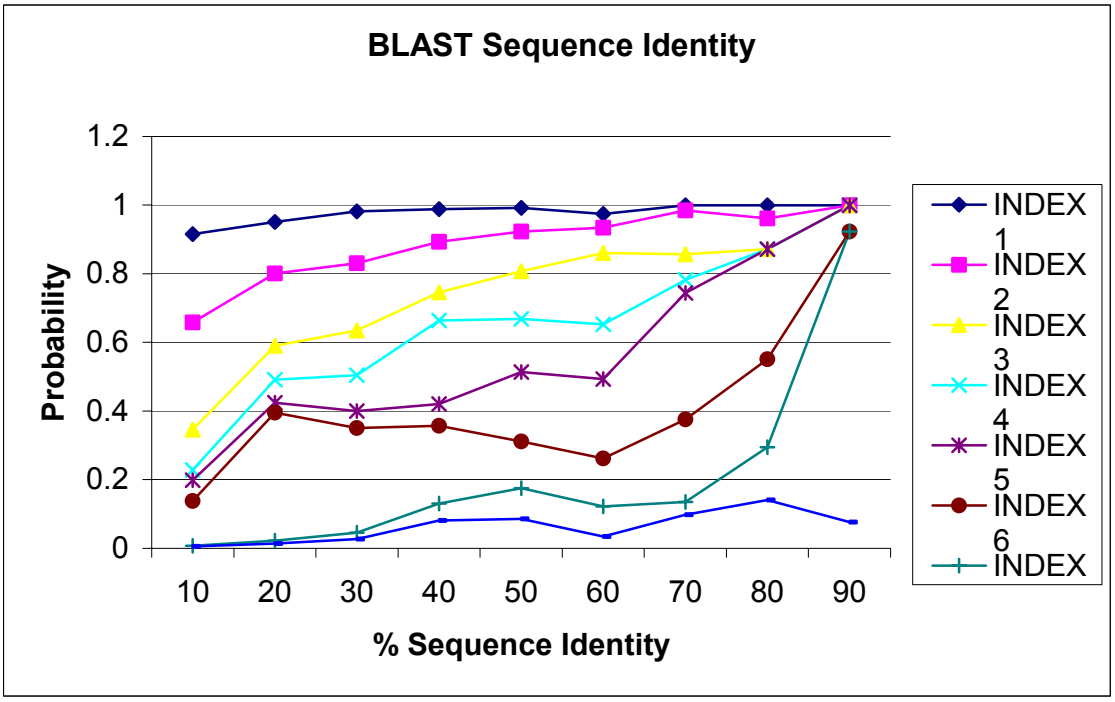

Figure A-7. Probability of matching function for INDEX level 1-8 for Yeast and Arabidopsis homology pairs, against percentage of sequence identity for BLAST.

shows the results based on the percentage of sequence identity. There is an increase in the probability of Yeast and Arabidopsis pairs sharing the same function with an increase in the percentage of sequence identity between the pairs. The probability decreases with an increase in the INDEX level. The results of sequence identity are better in comparison to those from FASTA and BLAST based on E-value. It is an indication that the function can be transferred from Yeast to its homologs in Arabidopsis.

VITA

Trupti Joshi was born on May 23, 1977, in Pune, India. She did her schooling in Karnatak High School, Pune. She was a merit rank-holder in her Higher Secondary School, which she completed in Abasaheb Garware College, Pune. She received her Bachelor of Medicine and Bachelor of Surgery degree from B.J.Medical College, Pune, with a distinction in Biochemistry. She went on to higher studies and received an Advanced Diploma in Bioinformatics at the Bioinformatics Center, University of Pune, where she ranked second with a distinction. In 2001, she came to University of Tennessee, Knoxville to pursue a Master's degree in Life Sciences. She worked with Dr. Dong Xu at Oak Ridge National Laboratory on the cellular functional annotation of yeast proteins using multiple sources of high-throughput data.