Protein-Protein Interactions and Metabolic Pathways Reconstruction of *Caenorhabditis elegans*

A thesis

submitted to

the College of Graduate Studies and Research

in partial fulfillment of the requirements for the

Degree of Doctor of Philosophy (Ph.D.)

in the

Department of Chemical Engineering

University of Saskatchewan

By

Mahmood Akhavan Mahdavi

Permission to use

In presenting this thesis in partial fulfillment of the requirement for a doctorate degree of philosophy from the University of Saskatchewan, I agree that the Libraries of this university may make it freely available for inspection. I also agree that permission for extensive copying of this thesis for scholarly purposes may be granted by Dr. Yen-Han Lin who supervised my thesis, or in his absence, by head of the Chemical Engineering Department or the dean of the College of Graduate Studies. It is understood that due recognition will be given to me and to the University of Saskatchewan in any use of the material within this thesis. Any copying, publication, or use of this thesis or parts for financial gain is prohibited without my written permission.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Chemical Engineering
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5A9

Acknowledgment

I would like to express my appreciation to my supervisor Dr. Yen-Han Lin for his useful comments and critical reviews of the results at every moment of my research.

I would also like to extend my appreciation to my respected Ph.D. advisory committee members: Drs. Ding-Yu Peng, Darren Korber, Mehdi Nemati, and Aaron Phoenix for their participation in my committee and their helpful comments during the course of this program.

I would like to thank Iran Ministry of Science, Research, and Technology (MSRT) for their invaluable financial support during my studies.

I am especially grateful to Greg Oster, laboratory systems analyst at the Department of Computer Science for his priceless assistance at any time on perl programming.

Finally, I would like to appreciate my family who provided a great support for me and a tranquil environment for my studies in the past four years.

Abstract

Metabolic networks are the collections of all cellular activities taking place in a living cell and all the relationships among biological elements of the cell including genes, proteins, enzymes, metabolites, and reactions. They provide a better understanding of cellular mechanisms and phenotypic characteristics of the studied organism. In order to reconstruct a metabolic network, interactions among genes and their molecular attributes along with their functions must be known. Using this information, proteins are distributed among pathways as sub-networks of a greater metabolic network. Proteins which carry out various steps of a biological process operate in same pathway.

The metabolic network of *Caenorhabditis elegans* was reconstructed based on current genomic information obtained from the KEGG database, and commonly found in SWISS-PROT and WormBase. Assuming proteins operating in a pathway are interacting proteins, currently available protein-protein interaction map of the studied organism was assembled. This map contains all known protein-protein interactions collected from various sources up to the time. Topology of the reconstructed network was briefly studied and the role of key enzymes in the interconnectivity of the network was analysed. The analysis showed that the shortest metabolic paths represent the most probable routes taken by the organism where endogenous sources of nutrient are available to the organism. Nonetheless, there are alternate paths to allow the organism to survive under extraneous variations.

Signature content information of proteins was utilized to reveal protein interactions upon a notion that when two proteins share signature(s) in their primary structures, the two proteins are more likely to interact. The signature content of proteins was used to measure the extent of similarity between pairs of proteins based on binary similarity score. Pairs of proteins with a binary similarity score greater than a threshold corresponding to confidence level 95% were predicted as interacting proteins. The reliability of predicted pairs was statistically analyzed. The sensitivity and specificity analysis showed that the proposed approach outperformed maximum likelihood estimation (MLE) approach with a 22% increase in area under curve of receiving operator characteristic (ROC) when they were applied to the same datasets. When proteins containing one and two known signatures were removed from the protein dataset, the area under curve (AUC) increased from 0.549 to 0.584 and 0.655, respectively. Increase in the AUC indicates that proteins with one or two known signatures do not provide sufficient information to predict robust protein-protein

interactions. Moreover, it demonstrates that when proteins with more known signatures are used in signature profiling methods the overlap with experimental findings will increase resulting in higher true positive rate and eventually greater AUC.

Despite the accuracy of protein-protein interaction methods proposed here and elsewhere, they often predict true positive interactions along with numerous false positive interactions. A global algorithm was also proposed to reduce the number of false positive predicted protein interacting pairs. This algorithm relies on gene ontology (GO) annotations of proteins involved in predicted interactions. A dataset of experimentally confirmed protein pair interactions and their GO annotations was used as a training set to train keywords which were able to recover both their source interactions (training set) and predicted interactions in other datasets (test sets). These keywords along with the cellular component annotation of proteins were employed to set a pair of rules that were to be satisfied by any predicted pair of interacting proteins. When this algorithm was applied to four predicted datasets obtained using phylogenetic profiles, gene expression patterns, chance co-occurrence distribution coefficient, and maximum likelihood estimation for *S. cerevisiae* and *C. elegans*, the improvement in true positive fractions of the datasets was observed in a magnitude of 2-fold to 10-fold depending on the computational method used to create the dataset and the available information on the organism of interest.

The predicted protein-protein interactions were incorporated into the prior reconstructed metabolic network of *C. elegans*, resulting in 1024 new interactions among 94 metabolic pathways. In each of 1024 new interactions one unknown protein was interacting with a known partner found in the reconstructed metabolic network. Unknown proteins were characterized based on the involvement of their known partners. Based on the binary similarity scores, the function of an uncharacterized protein in an interacting pair was defined according to its known counterpart whose function was already specified. With the incorporation of new predicted interactions to the metabolic network, an expanded version of that network was resulted with 27% increase in the number of known proteins involved in metabolism. Connectivity of proteins in protein-protein interaction map changed from 42 to 34 due to the increase in the number of characterized proteins in the network.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADP	Adenosine DiPhospahate
ATP	Adenosine TriPhospahate
AUC	Area Under Curve
BIND	Biomolecular Interaction Network Database
BLAST	Basic Local Alignment Search Tool
BRENDA	Braunschweig Enzyme Database
CAPRI	Critical Assessment of PREdicted Interactions
CDS	Coding Sequence
COG	Clustering Orthologous Groups
CYGD	Comprehensive Yeast Genome Database
EC	Enzyme Commission
GO	Gene Ontology Database
HPRD	Human Protein Reference Database
KEGG	Kyoto Encyclopedia of Genes and Genomes
MetaCyc	Metabolic Encyclopedia
MLE	Maximum Likelihood Estimation
NADH	Nicotinamide Adenine Dinucleotide (reduced form)
ORF	Open Reading Frame
PPI	Protein-Protein Interaction
PROSITE	Protein Sites
ROC	Receiving Operator Characteristic
SWISS-PROT	Swiss Protein Database
SVM	Support Vector Machine
SMD	Stanford Microarray Database
UNIPROT	Universal Protein Resource

1

LITERATURE REVIEW AND BACKGROUND

Traditionally, it was believed that proteins were isolated entities, floating in the cytosol and, for the most part, acting independently of surrounding proteins. Proteins were thought to move freely, and reactions occurred as a result of proteins A and B randomly colliding with one another. Today we know this picture is far too simplistic to describe the complex processes that all happen in living cells. Instead, the majority of cellular phenomena are carried out by protein complexes, or aggregates of ten or more proteins. These protein-protein interactions are critical to all cellular processes, and understanding them is key to understanding any biological system.

The growing number of fully sequenced genomes and high-throughput experimental data sets has increased our knowledge on cellular components on a genome scale and the capability of far more meaningful interpretation of metabolic responses (Fell, 2001). Information about the functions of cellular components (Gerlt and Babbitt, 2000), conserved interactions among proteins in different species (Sharan *et al.*, 2005), their genetic localizations, and mutations over evolution can be represented in the different levels of genome annotation (Reed *et al.*, 2006). One-dimensional genome annotation involves identification of genes in genomes and the assignment of predicted or known functions to the products of those genes. Advances in experimental and computational techniques have resulted in complete sequencing of hundreds of organisms (Janssen *et al.*, 2003) and identifying many new genes and proteins. Bioinformatics tools that are used to derive one-dimensional annotations including protein functions are now publicly available (Pellegrini, 2001). We will review these methods in sections 1.1 and 1.2.

Two-dimensional annotation specifies the interaction among cellular components. Physical and functional interactions between cellular components lead to a network reconstruction that effectively represents two-dimensional annotation information. Metabolic network reconstruction is one aspect of two-dimensional annotation, which is basically a structured database in which one-dimensional annotation is placed into a biological context. Thus, two-dimensional annotation builds on one-dimensional

annotation by considering cellular components and their interactions. It should be noted that, in some cases two-dimensional annotation can lead to a one-dimensional genome re-annotation. Metabolic network reconstruction will be discussed in section 1.3.

Protein-protein interaction maps (interactomes) and consequently protein function assignments are two basic sets of information that can be incorporated into network reconstruction process (Hatzimanikatis *et al.,* 2004). Protein-protein interaction is the main target of proteomics (Archakov *et al.,* 2003). There are bioinformatics tools by which proteins are identified through their interactions as well (Huang *et al.,* 2005). Even computational techniques are available to design interactions between proteins (Kortemme and Baker, 2004). Organisms' interactomes can now be characterized by computational approaches (Colizza *et al.,* 2005; Needham *et al.,* 2006). These interactomes contain conserved and essential protein complexes in which proteins interact permanently or transiently to perform a biological process (Butland *et al.,* 2005). This information results from integrating genomic data under certain circumstances (Lu *et al.,* 2005) or comparison of protein interaction maps (Liang *et al.,* 2006). However, overlap among interactomes is not satisfactory. As an example, in a comparison among the interactomes of four model organisms including yeast, worm, fly, and human, of over 70000 binary interactions only 42 were found common to all four organisms (Gandhi *et al.,* 2006).

Proteins are assigned function based on their interactions. When an interaction between two proteins is predicted computationally or confirmed experimentally this is evidence that the two proteins may have a functional relationship. Similarly, when two proteins have functional links this is a strong indication that the two proteins may interact with each other (Vazquez *et al.,* 2003). Organisms' functional maps are assembled based upon their interactomes (Grant and Wilkinson, 2003). Recently, a faster and more accurate algorithm has been proposed for protein function assignment using protein interaction information (Sun *et al.,* 2006). Many computational approaches have been proposed to predict protein-protein interactions at the one-dimensional annotation level. Additionally, experimental high-throughput technologies have also been discovered to produce tremendous amount of protein-protein interaction data. Even with all these methods a large fraction of the genes in the genomes are still uncharacterized. In the

following sections we will review currently available experimental and computational methods for prediction of protein-protein interactions.

1.1 Experimental protein-protein interaction techniques

Currently, there are many experimental techniques available to generate protein-protein interaction information. Among all these techniques, yeast two-hybrid is one of the most common high-throughput methods able to generate a large amount of data in one set of experiment. Other techniques such as co-immunoprecipitation, and affinity chromatography detect protein-protein interactions one-at-a-time. Yeast two-hybrid technique has also been used to study cell death mechanism in which a few proteins are involved (Wallach *et al.*, 1998).

The principle behind yeast two-hybrid is the activation of a downstream reporter gene by the binding of a transcription factor to an upstream activating sequence (Fields and Song, 1989). As seen in Figure 1.1 it uses two protein domains that have specific functions: a DNA-binding domain (BD) that is capable of binding to DNA, and an activation domain (AD), that is capable of activating transcription of the DNA. Both of these domains are required for transcription, whereby DNA is copied in the form of mRNA and then translated into protein. For the transcription of DNA, it requires a protein called transcriptional activator (TA) that possesses both domains. This protein binds to the promoter, a region located upstream from the gene (coding region), that serves as the binding site for the transcriptional protein. Once the TA has bound to the promoter, it is able to activate transcription via its activation domain and the transcription of reporter gene occurs. If either of these domains is absent, the transcription will fail.

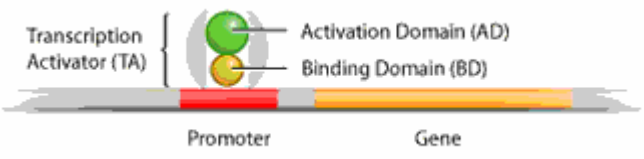

Figure 1.1. Transcription proteins in yeast two hybrid technique (Taken from *the science creative quarterly* at www.scq.ubc.ca)

The key principle in yeast two-hybrid is that the BD and the AD do not necessarily have to be on the same proteins. Basically, the two proteins whose interaction is going to be investigated are genetically engineered and their plasmids are incorporated into a strain of yeast in which the biosynthesis of certain nutrients is lacking. One plasmid contains the binding domain fragment (bait protein) and activating domain is contained in the other plasmid (prey protein). The bait protein is typically a known protein that is used to identify its new partners.

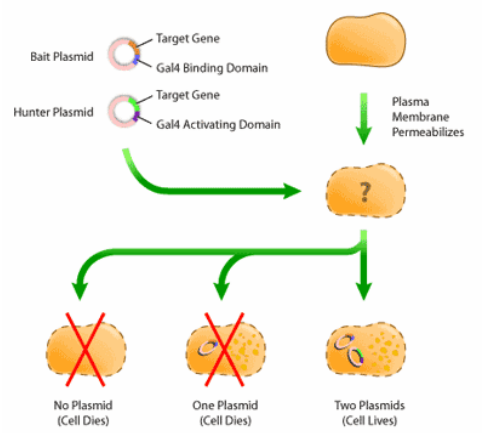

Figure 1.2. Interaction of bait and prey proteins. If bait and prey proteins interact, transcription of receptor happens and the yeast strain grows on a media that is lacking an essential nutrient. (Taken from *the science creative quarterly* at www.scq.ubc.ca)

If the bait and prey interact (i.e. bind) then the AD and BD of the transcription factor are indirectly connected and the transcription of the reporter gene takes place. As a result, the plasmids allow the mutant yeast to grow on the medium lacking nutrients because the transcription of reporter gene is followed by encoding enzymes that allow the synthesis of the nutrients that mutant strain is unable to produce. A common transcription factor for yeast two-hybrid screening is GAL4.

Currently, more than 95% of experimental data on protein interactions are obtained by the yeast two-hybrid technique. In *C. elegans*, 7081 experimental protein-protein interactions were reported by Li *et al.* (2004), approximately 6800 of them obtained through yeast two-hybrid assays. Schwikowski *et al.* (2000) identified 2358 protein-protein interactions in *Saccharomyces cerevisiae* using yeast two-hybrid technique. This technique has also been used to detect 10021 interactions in *Drosophilo melanogaster*

(Uetz and Pankratz, 2004). The human interactome is a developing resource that yeast two-hybrid plays an important role in its completion (Ramani *et al.,* 2005). In a recent study, bait and prey proteins in *E. coli* K12 were purified by electrophoresis and 2667 interactions were identified by data explorer and/or proteomics solution (Arifuzzaman *et al.,* 2006). *In silico* two-hybrid has been also used to detect physically interacting proteins (Pazos and Valencia, 2002).

1.2 Computational protein-protein interaction approaches

Many computational approaches have been proposed to predict protein-protein interactions (Franzot and Carugo, 2003). These methods utilize different information to predict interactions ranging from genomic and sequence information related to primary structures, to domains, motifs, and other functional units related to secondary structures of proteins. Computational approaches to predict protein-protein interactions have been reviewed from different prospective (Yu and Fotouhi, 2006; for example). Here we categorize these methods into six groups based on the type of the information upon which interactions are predicted.

1.2.1 Genomic information in protein-protein interaction prediction

With the availability of complete genome sequences, genomic information became the basis for genomics-based prediction techniques. Early methodologies rely on homology among primary structure of proteins that was believed was able to reveal general function of some proteins in different organisms (Bork *et al.,* 1998). For example, from sequence homology, 30% of yeast genes had known human homologs, and 40% were similar enough to other genes in other organisms (Brent and Finely, 1997). Homology is defined as similarity in DNA or protein sequences between individuals of the same species or among different species. If the similarity occurs among those proteins in different organisms, they are orthologs and if it occurs among proteins from the same organism they are paralogs (Sonnhammer and Koonin, 2002). Homology has been used to classify protein structures (Dietmann and Holm, 2001) and some proteins have been identified based on their homologous partners (Bolten *et al.*, 2001). Orthologous genes have been clustered and maintained in databases such as COG (Clustering Orthologous Groups) and these databases have been used to annotate hypothetical proteins across

more than 200 prokaryotes (Doerks *et al.,* 2004). Also some programs are available to predict protein-protein interactions based on orthologous proteins (Huang *et al.,* 2004). Species-specific proteins have been identified through detecting homology in different organisms (Huynen *et al.,* 1998). Those proteins which have no ortholog among a set of genomes may represent specific features of an organism. Comparative genomics is another way to detect specific proteins across organisms. This type homology-based genomics analysis has been used to identify eukaryotic genes responsible for specific protein interactions (Rubin *et al.,* 2000).

Homology-based computational techniques are based merely on primary structure similarity which creates some random relationships among proteins. On the other hand, part of proteins encoded by an organism can not be functionally assigned by pure homology searching methods. Hence, it is believed that combination of homology with evolution may improve prediction of relationships among proteins (Eisen, 1998). Phylogenetic trees which show the ancestral history of genes and their products are appropriate indicators of interactions among proteins (Pazos and Valencia, 2001). Thus phylogenetic analysis was introduced to genomics studies to improve gene function and protein interaction predictions (Eisen and Wu, 2002). The relationship between evolution and gene function has recently been emphasized by incorporating this information into protein-protein networks and characterizing more unknown proteins (Koonin and Wolf, 2006). Co-evolution of gene and proteins has also been a source of information to predict interacting proteins even though their phylogenetic relationship is excluded from the assessment (Sato *et al.*, 2005).

Non homology-based methods such as conventional phylogenetic profiles (Pellegrini *et al.,* 1999), protein fusion (Marcotte *et al.,* 1999), gene neighbourhood (Dandekar *et al.*, 1998), and transgenic distance (Strong *et al.,* 2003) address that part of the proteomes which can not be detected by homology-based techniques. These methods link a pair of non-homologous proteins based on fusion or speciation events that happened over evolution and eventually assign proteins with function (Marcotte, 2000). Although these computational techniques use homology searching tools to detect the presence of a whole or partial sequence in other organisms, the final linkages are not based on similarity between a pair of sequences. As most of these methods use the BLAST program as a tool

to search homologous sequences, this program will be described briefly and then the underlying hypothesis of each method will be discussed.

The BLAST program (Altschul *et al.*, 1990) is used to compare a new sequence with those contained in nucleotide and protein databases by aligning the novel sequence with previously characterized genes. The emphasis of this tool is to find regions of sequence similarity which will provide functional and evolutionary clues about the structure and function of the novel sequence. Regions of similarity detected via alignment tool can be either local or global. Global alignment is based on the whole sequence of the query and is not a suitable way to find similarity. Then local alignment was proposed which is far more effective than global alignment. This type of alignment is based on Smith-Waterman algorithm in which the program scores the best alignment of any substring of one string with any substring of the other string. Smith-Waterman algorithm implements a technique called dynamic programming which takes alignment of any length, at any location, in any sequence, and determines whether an optimal alignment can be found. Based on these calculations scores are assigned to each character-to-character comparison so that positive for exact matches and negative for insertions or deletions are given. At the end scores of all comparisons of a sequence are added together and the highest scoring alignment is reported. The running time of Smith-Waterman algorithm makes it impractical for use. Therefore, BLAST carries out a significant amount of pre-processing on the query and database. In the pre-processing phase most letters are eliminated from similarity searching. Then BLAST identifies high scoring pairs of three-letter substrings.

In BLAST an expect value (E-value) is assigned to a match between two sequences as a measure of similarity as follows:

$$E = K.m.n.e^{-\lambda S} \tag{1.1}$$

where, m and n are lengths of two sequences, K and λ are statistical parameters, and S is the similarity score. These parameters are related together as follows:

$$S' = \frac{\lambda S - \ln K}{\ln 2} \tag{1.2}$$

where S' is a bit score, normalized similarity score. Then E-value corresponding to a given bit score is:

$$E = m.n.2^{-S'} \tag{1.3}$$

E-value is a parameter that describes the number of hits one can expect to see by chance when searching a database of the same size. This means that the lower the E-value, or the close it to "0", the more "significant" the match is.

In phylogenetic profiling using the whole genome sequence of an organism, patterns of presence or absence of all proteins of the genome in a set of reference genomes are constructed. When two proteins have similar patterns a link between the two proteins is established as shown in Figure 1.3. The presence or absence of a protein in a reference genome is jugged by the similarity of the query sequence with sequences within the reference genome. Similarity is measured based on E-value. When E-value is greater than a threshold two proteins are considered similar. The choice of threshold, and the number of reference genomes depend on the size of query database and the species and will be different from case to case. Nevertheless, some suggestions such as 145 genomes and the threshold of 10^{-4} as general guidelines are provided (Shi *et al.,* 2005). Recently, the phylogenetic profiles method was used to identify genes involved in orphan metabolic activities (Chen and Vitkup, 2006). These genes could not be detected by homology-based techniques.

Protein fusion method is based on the idea that a pair of distinct proteins in one organism may be expressed as a fused protein in another organism as illustrated in Figure 1.4. Identification of fused proteins is based on local alignment of a protein against another protein or a protein against a domain. When similarity between protein A and domain 1 of protein C and also between protein B and domain 2 of protein C is significant (E-value higher than a threshold), proteins A and B are fused into protein C. Thus, a relationship between the two proteins A and B is predicted. Searching fusion events across organisms result in huge number of protein-protein interactions in an organism. However, it is clear that not all fusion events are equally valuable for inferring interactions. For this matter, a statistical measure was developed to score all fusion events and specify the significance of a link (Verjovsky Marcotte and Marcotte, 2002). Briefly a benchmark was developed for testing interaction predictions, and comparison of the significance score of the link against the benchmark shows that the significance score correlates well with the degree of relatedness of the linked proteins. Another approach to detect protein fusion is applying mathematical relations to protein sequence databases to

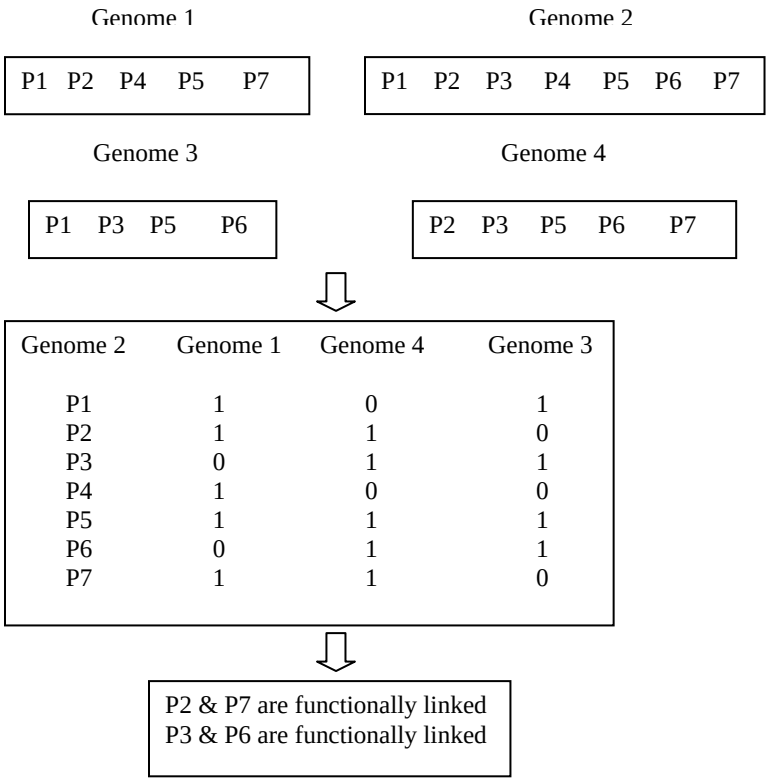

Figure 1.3. The illustration of phylogenetic profiles method. The presence or absence of each protein in each genome is demonstrated by a profile comprising '0' and '1' representing absence and presence, respectively.

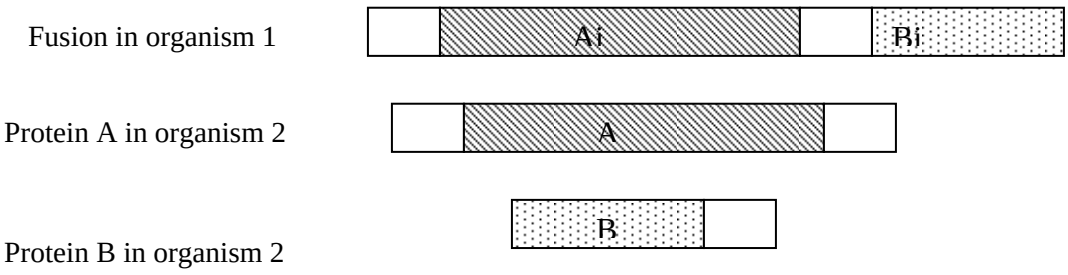

Figure 1.4. The illustration of protein fusion method. Two separate proteins in organism 2 indicated as a fused protein in organism 1. Protein A and protein B in organism 2 are found as two domains of a protein in organism 1.This fusion event is an indication that proteins A and B may have a link.

find fusion events (Truong and Ikora, 2003). With the availability of many protein and domain sequence databases, this mathematical approach seems promising.

The gene neighbour method infers a functional link between two proteins if they are neighbours on one chromosome in organism X and their orthologs in organism Y are also neighbours to each other. The assumption is that protein-protein interactions impose evolutionary constraints to keep the genes together. This functional association is independent of relative gene orientation. The main limitation of the method is that it is suitable only for bacterial genomes since the conservation of the gene neighbouring is kept well in the bacteria (Eisenberg *et al.*, 2000). This analysis also results in a number of false predictions because the constraint of proximity is not strong and some distant interactions are not identifiable.

The transgenic distance method is based on the notion that prokaryotic operon organization enables the highly controlled co-expression of multiple genes, by transcribing them together on a single transcript. Thus, the encoded proteins often have functional relationships. It is shown that the intergenic spacing between genes in a common operon is shorter than the intergenic spacing of genes encoded by separate transcription units. Therefore, imposing a transgenic distance threshold, when the distance between two genes is less than threshold the two genes are considered on the same operon. In contrast to previously mentioned methods, this method focuses on the analysis of a single genome. Examination of different prokaryotic operons detects that the genetic distance above 200 bp is less likely to result in a reliable interaction (Strong *et al.*, 2003).

Combination of genomics-based methods to predict protein-protein interactions may strengthen the robustness of predictions. It has been shown that when two or more methods agree on a link the probability of being related is higher, however, the level of correlation between different methods may vary (Hoffman and Valencia, 2003).

1.2.2 Statistical measures and protein-protein interaction prediction

Genes with identical patterns of occurrence across organisms are more likely to interact; however, the requirement that the profiles be identical restricts the number of links that can be established by such phylogenetic profiling. Thus, there are a group of

methods that rely on scoring phylogenetic patterns and match them based on those scores rather than identical profiles. Various scoring functions such as mutual information (Date and Marcotte, 2003), Jaccard coefficient (Yamada *et al.*, 2004), and chance co-occurrence probability distribution (Wu *et al.*, 2003) are used to match profiles together. These scoring functions provide more information than the simple presence or absence of genes.

As a measure of phylogenetic profile similarity, the mutual information score is calculated between pairs of phylogenetic profiles. Profile for each protein *i* is a vector with elements p_{ij} corresponding to each organism *j* in the set of reference organisms, where $p_{ij} = -1/\log E_{ij}$, and E_{ij} represents the E-value of the top-scoring sequence alignment between protein *i* and all of the proteins in organism *j*. The mutual information is calculated as follows (Huynen *et al.*, 2000):

$$MI(A,B) = H(A) + H(B) - H(A,B) \tag{1.4}$$

where, $H(A) = -\sum p(a) \ln p(a)$ and represents the entropy of the probability distribution *p(a)* of gene A occurring among the organisms in the reference database, and $H(A,B) = -\sum\sum p(a,b) \ln p(a,b)$ represents the relative entropy of the joint probability distribution *p(a,b)* of occurrence genes A and B across the set of reference genomes. Once the pairs of profiles are ranked based on mutual information scores, specifying a threshold, corresponding proteins are linked accordingly when their mutual information score is higher than the threshold.

Another measure of similarity between phylogenetic profiles is Jaccard coefficient. This coefficient is calculated between two binary profiles. These profiles represent the presence or absence status of genes in a set of reference genomes. Jaccard coefficient is calculated as follows:

$$JC(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1.5}$$

where $A \cap B$ means the number of organism that have both genes A and B, and $A \cup B$ means the number of organisms that have gene A or gene B. Jaccard coefficient of a pair of genes is usually used along with another property of the pair such as pathway

distance (Yamada *et al.,* 2004) to conclude a relationship between genes. Nonetheless, it is a strong evident that two genes are suitable candidates for interaction.

Chance co-occurrence probability distribution has been also used as a measure of similarity between two phylogenetic profiles. This measure is used to relax the restriction of identical profiles between two proteins, based on the probability that a given arbitrary degree of similarity between two profiles would occur by chance, with no biological pressure. The interaction predictions are drawn with the criterion used to reject the null hypothesis. The probability *P(z|N,x,y)* of observing by chance (i.e. no functional pressure) *z* co-occurrence of genes *X* and *Y* in a set of *N* genomes, given that *X* occurs in *x* genomes, and *Y* occurs in *y* genomes is calculated as follows:

$$P = \frac{w_z \overline{w}_z}{W} \tag{1.6}$$

where w_z is the number of ways to distribute *z* co-occurrence over the *N* genomes, $\overline{w}_z$ is the number of ways of distributing *x*-*z* and *y*-*z* genes over the remaining *N*-*z* genomes, and *W* is the number of ways of distributing *X* and *Y* over *N* genomes without restriction. The final equation is as follows:

$$P = \frac{(N-x)!(N-y)!x!y!}{(x-z)!(y-z)!(N+z-x-y)!z!N!} \tag{1.7}$$

Predictions are established upon lower probabilities of matching proteins by chance. A cut-off threshold should be specified to obtain interacting pairs.

1.2.3 Structure-based Prediction of protein-protein interactions

Early studies on protein interactions and functions showed that there was relationship between protein structure and interaction (Hegyi and Gerstein, 1999). These studies mostly relied on secondary structures such as domains to correlate protein interactions to structural properties (Elofsson and Sonnhammer, 1999). There were some models which considered protein structures as a network of amino acids and sub networks provide interfaces for protein-protein interactions (Del Sol *et al.,* 2005). Moreover, conservation of some sequence patterns consolidated this hypothesis (Espadaler *et al.,* 2005; Aytuna *et al.,* 2005). Crystallography is a common technique to detect the structure of proteins; however, interactions between crystal packing may vary according to the effect of

complex formation (Zhu *et al.,* 2006). Advances in this technique have provided the advantage of using protein tertiary and quaternary structures to inferring protein-protein interactions. These methods range from threading approach (Lu *et al.,* 2002), docking methods (Smith and Sternberg, 2002), and CAPRI experiment (Janin *et al.,* 2003) to protein interaction prediction based on surface patch comparison (Carugo and Franzot, 2004) and oligomeric protein structure networks (Brinda and Vishveshwara, 2005). Structure-based methods are dependent to the number of known structures and existing structural complexes in each organism. In threading approach one attempt to align the sequence of the protein of interest to a library of known folds and find the closest matching structure (Lu *et al.,* 2003). The goal of threading is to extend sequence-based approaches by recognizing the structures that can be analogous (i.e. the two proteins are not necessarily evolutionary related). Docking method is based on the identification of binding sites in a protein structure and subsequently determining of structure of protein complexes (Jackson and Sternberg, 1995). Although the study of protein-protein docking was boosted by the rapid increase in available protein structures, the main limitation of docking algorithms is that they can not always assess which proteins interact and which do not. Because it usually takes hours to predict the interacting sites for a pair of potentially interacting proteins. The current status of docking methods has been reviewed elsewhere (Mendez *et al.,* 2003).

The critical assessment of predicted interactions (CAPRI) experiment was designed to testing protein docking algorithm in blind predictions of the structure of protein-protein complexes. CAPRI is a protocol by which structural prediction of protein complexes that is offered by crystallographers can be assessed further and regions of interactions can be detected (Wodak and Mendez, 2004). Comparison of protein surface patches is based on three-dimensional structure of proteins. In this analysis the surface of each protein, represented by solvent accessible atoms, is divided into small patches. The geometry of each patch is described by the atom distributions along its principle axes. The geometry between two patches is estimated by comparing their atom distribution along axes. Only those patch combinations whose atom distribution values are higher than a threshold, may translate into interactions between proteins that correspond to the surface patches. Oligomeric protein structures and their comparison with monomeric protein structure

networks provide insight into new protein associations. Specifically, the interface hubs, hydrogen bonds, hydrophobic interactions and other interactions essential for protein associations are identified through this comparison. These hub interactions are the key information to identify protein complexes.

1.2.4 Machine learning in prediction of protein-protein interactions

There are computational methods to predict protein-protein interactions which employ machine learning techniques. These methods use different information to predict protein-protein interactions such as primary structures (Bock and Gough, 2001), and conserved network motifs (Albert and Albert, 2004). Interaction mining was also used to train learning systems to recognize correlated patterns within protein interaction pairs (Bock and Gough, 2003). Data mining can be applied to different data sources. Study set gene files and gene-association files associated with genes which contain description of gene function can be a source of mining (Castillo-Davis and Hartle, 2003). Published literature can be another source of mining novel interactions which are identified through independent studies (Marcotte *et al.,* 2001). This type of data mining has also been used to search functions for interacting proteins (Chen and Xu, 2004). Even protein-protein interaction maps can be explored to find hidden interactions which are evolved as a result of network behaviour (Hu, 2005). In all different data mining approaches mentioned above, a certain identifier is trained using machine learning techniques.

Support vector machines (SVM) (Noble, 2006) have been used to construct supervised classifiers in order to identify interacting proteins (Huang *et al.,* 2004). The effect of training dataset on the performance of SVM prediction has been studied (Lo *et al.,* 2005) to enhance the efficiency of predictions. SVM is a useful tool to predict interactions among proteins which are involved in a specific biological process such as binding (Han *et al.,* 2004). Nevertheless, it can be used on a genome scale to predict interactions with reasonable precision (Alashwal *et al.,* 2006). Classifiers trained by SVM learning system can be constructed based on physiochemical properties of amino acids which are extracted from composition of amino acids in a protein (Nanni and Lumini, 2006).

1.2.5 Gene expression analysis and protein-protein interactions

With the availability of gene expression map of some model organisms, such as *C. elegans* (Kim *et al.,* 2001), gene expression data has been widely used to predict protein-protein interactions (van Noort *et al.,* 2003). Also, gene expression profiling data was used to specify the function of some macro molecules such as oligo-nucleotides (Tolstrup *et al.,* 2003). These methods predict interacting proteins through integration of micro array data in different biological conditions and construction of co-expression profiles for genes (Zhou *et al.,* 2005). When two genes are co-expressed in a series of biological events in a correlated fashion, it indicates that the two genes and their translated proteins may have functional relationships. Identification of protein interactions via expression information may also assist finding more physical cooperation of proteins to accomplish a biological task. Most of this cooperation is evolutionary conserved and may be specified through other prediction techniques (Gunther and Gaasterland, 2001). In yeast, expression of genes is highly correlated among those conserved over the evolution (Mata and Bahler, 2003). In order to construct an expression profile for a particular gene, the employed clustering technique plays an important role. There exist numerous clustering techniques in the literature (D'haeseleer, 2005); however, the selection of the appropriate one depends on the objective of clustering and the accuracy of the analysis.

1.2.6 Domain-based protein-protein interaction prediction

Proteins interact through their functional subunits (Ponting and Russell, 2002). Protein domains, active sites, motifs (collectively called signatures) are sub-sequence functional and conserved patterns that are essential to the functioning of individual cells and are the interfaces in interactions at protein level (Littler and Hubbard, 2005). With the completion of full genome sequence of many organisms, genome-wide characterization of protein domains is now practical (Murvai *et al.,* 2000). Although proteins are specified by unique amino acid sequences, the domain content of a protein sequence is crucial to specify interactions in which the particular protein is involved.

Protein domain information has been used to predict protein-protein interactions. Naively, when two proteins were known to interact, their homologs in other organisms were assumed to interact based on comparative analysis (Bansal, 1999). Domain contents of interacting partners were utilized as input to predict more accurate predictions in

another organism (Wojcik and Schachter, 2001). Intermolecular or intramolecular interactions among protein families that share one or more domains were implemented to infer interactions among proteins (Park *et al.*, 2001). Domain-domain relationships were used to predict interactions at protein level. In the association method (Sprinzak and Margalit, 2001) interacting domains were learned from a dataset of experimentally determined interacting proteins, where one protein contained one domain and its interacting partner contained the other domain. The probabilistic model of maximum likelihood estimation (MLE) (Deng *et al.*, 2002) outperformed the association method through taking the experimental errors into account. Following a recursive calculation procedure, in MLE method, probabilities for domain-domain interactions were predicted based on the observation of interaction between their corresponding proteins. Then the prediction was extended to protein level assuming that two proteins interact if and only if at least one pair of domains from the two proteins interacting. Potentially Interacting Domain pairs (PID) were extracted from an experimentally confirmed pair dataset using PID matrix score (Kim *et al.*, 2002) as a measure of domain interaction probability. In another study the strengths of protein pairs were incorporated into the association method to enrich probability estimations (Hayashida *et al.*, 2004). As many domain structures are shared by different organisms, the integration of data from multiple sources may strengthen the reliability of domain associations and protein interactions (Liu *et al.*, 2005). Domain contents of *S. cerevisiae* proteins have been used to train an SVM classifier to distinguish interacting protein from non-interacting one. Protein interactions were measured based on the mean of similarity among domain contents of two query proteins (Zaki *et al.*, 2006). In all these methods, if a probability score meets a certain threshold, domains and subsequently related proteins are considered ‘interacting’. However, these methods do not distinguish between single-unit proteins and multi-unit proteins.

To overcome the limitation of conventional domain-based approaches to consider interactions of single domain pairs, domain combination based methods were proposed. Domain combination based approach predicted protein interactions based on the interactions of multi-domain pairs or the interactions of groups of domains (Han *et al.*, 2004). Recently, interactomes (Li *et al.*, 2004; Rain *et al.*, 2001; Uetz and Pankratz,

2004; Rhodes *et al*., 2005) and databases, such as DIP (Salwinski *et al.*, 2004) were used as reliable sources for mining interacting domains and may contribute to inferring uncharacterized interacting proteins (Riley *et al.*, 2005). Therefore, domain contents of proteins play a crucial role in predicting protein interactions. Domain-based PPI prediction techniques rely on statistically significant related domains. When the interaction probability score between two domains (in two different proteins) is greater than a threshold value, such a relationship is extended to the corresponding proteins and the potential interaction is inferred.

1.3 Metabolic network reconstruction

Metabolic network reconstruction allows for an in depth insight into molecular mechanism of cellular activities in a particular organism. A reconstruction breaks down metabolic pathways into respective genes, enzymes, and reactions and analyzes them in terms of their biological relationship. Briefly, a reconstruction involves collecting all the relevant metabolic information related to a specific organism from various sources and then compiling them in a way that is capable to performing various types of analyses. Metabolic reconstruction consists of a few steps that are crucial to proper relationships among different elements of the network (Francke *et al.*, 2005). The beginning step is searching information that correlates between genome and metabolism. It can be found in different databases such as KEGG (Kanehisa and Goto, 2000) in which a search can be conducted based on a protein name or enzyme commission (EC) number in order to find the associated gene. Presently, KEGG is the most comprehensive database that contains metabolic information at different levels including genes, proteins, pathways, reactions, and metabolites for many organisms. Similar to KEGG resource, MetaCyc (Caspi *et al.*, 2006) provides metabolic information retrieved from scientific experimental literature. It is an encyclopaedia of metabolic pathways containing a wealth of information on metabolic reactions derived from over 600 different organisms. Also, BRENDA (Schomburg *et al.*, 2002) is a comprehensive enzyme database that allows searching an enzyme by name or EC number. This database can be searched for an organism and all its relevant enzyme information. Moreover, when an enzyme search is carried out, BRENDA provides a list of all organisms containing the particular enzyme of interest. A

collection of different databases and their characteristic features can be found in (Baxevanis, 2003).

The next step in metabolic reconstruction is the verification of the data to ensure consistency and accuracy of the data. This provides an added level of assurance for the reconstruction that the enzyme and the reaction it catalyzes do actually occur in the organism. Any new reaction not present in the database need to be added to the reconstruction. The presence or absence of certain reactions of metabolism will affect the whole picture because products in one reaction go on to become the reactants for the next reaction i.e. products of one reaction combine with other proteins or compounds to form new compounds in the presence of different enzymes.

In order to simulate a metabolic network, information related to reactions and enzymes are incorporated into a stoichiometric matrix where rows and columns correspond to metabolites and reactions, while the elements are the stoichiometric coefficients (Vo *et al.,* 2004). Information collected in this matrix is used to build or revise metabolic pathways. Combining the stoichiometric matrix and gene-protein-reaction (GPR) structure the missing reactions and enzymes can be recognized. These reaction and enzymes can be found through mining in the literature or conducting experiments. The main advantage of metabolic reconstruction is that it reveals the knowledge gap in relationship among different biological elements of cellular system e.g. genes, proteins, and reactions. Moreover, it provides the opportunity to augment metabolic networks through the integration of relevant protein interacting information.

1.4 Conclusion

Protein-protein interaction information is the building block of metabolic network reconstruction. Proteins in a cell are not isolated entities; instead, they create associations to perform a biological task. Thus, identification of protein interactions is crucial to understand cellular activities and then incorporate them into metabolic networks which provide a large picture of all cellular activities in an organism. So far, in *C. elegans*, a small portion of proteins and their interactions have been identified due to the complicated multi-cellular structure of this organism. Thereby, the metabolic network of *C. elegans* is still incomplete and much work is yet to be done to achieve a greater picture of metabolic processes in this organism. More validated protein-protein interactions need

to be available and then incorporated into the network to expand the current metabolic network. On the other hand, experimental techniques to elucidate more protein-protein interactions are expensive and labour intensive. Consequently, computational interaction prediction approaches have been widely used to infer more protein-protein interactions in a shorter amount of time and supply adequate information to improve metabolic reconstruction studies. However, the growing numbers of computational approaches are not only insufficiently accurate but also they suffer from mass false positive predictions. These issues have been addressed in this research. The current metabolic network of *C. elegans* was reconstructed and known protein-protein interactions were specified. A new method of predicting protein interactions was introduced and a framework for reducing false positive predictions was proposed. Then newly predicted and validated interactions were incorporated into the current network and an expanded metabolic network was achieved. More details on the objectives of this research is presented in next chapter.

References

Alashwal H., Deris S., Othman R.M. (2006) One-class support vector machines for protein-protein interactions prediction. *Int. J. of Biomedical Sciences,* **1(2):**120-127.

Albert I., Albert R. (2004) Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics,* **20:**3346-3352.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. *J. Mol. Boil.* **215,** 403-410.

Archakov A.I., Govorun V.M., Dubanov A.V., Ivanov Y.D., Veselovski A.V., Lewi P., Janssen P. (2003) Protein-protein interactions as a target in proteomics. *Proteomics,* **3:**380-391.

Arifuzzaman M., Maeda M., Itoh A., Nishikata K., Takita C., Saito R., Ara T., Nakahigashi K., Huang H.-C., Hirai A., Tsuzuki K., Nakamura S., Altaf-Ul-Amin M., Oshima T., Baba T., Yamamoto N., Kawamora T., Ioka-Nakamichi T., Kitagawa M. Tomita M., Kanaya S., Wada C., Mori H. (2006) Large-scale identification of protein-protein interaction of *Escherichia coli* K-12. *Genome research,* **16:**686-691.

Aytuna A.S., Gursoy A., Keskin O. (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, **21(12):** 2850-2588.

Bansal A.K. (1999) An automated comparative analysis of 17 complete microbial genomes. *Bioinformaics*, **15,** 900-908.

Baxevanis A.D. (2003) The molecular biology database collection: 2003 update. *Nucleic Acids Research*, **31(1)**, 1-12.

Bock J.R., Gough D.A. (2001) Predicting protein-protein interactions from primary structure. *Bioinformatics,* **17,** 455-460.

Bock,J.R., Gough D.A. (2003) Whole-proteome interaction mining. *Bioinformatics,* **19:**125-135.

Bolten E., Schliep A., Schneckener S., Schomburg D., Schrader R. (2001) Clustring protein sequences-structure prediction by transitive homology. *Bioinformatics,* **17,** 935-941.

Bork P., Dandekar T., Diaz-Lazcoz Y., Eisenhaber F., Huynen M., Yuan Y. (1998) Predicting function: from genes to genome and back. *J. Mol. Biol.* **283,** 707-725.

Brent R., Finely R.L. (1997) Understanding gene and allele function with two-hybrid methods. *Annu. Rev. Genet.* **31,** 663-704.

Brinda K.V., Vishveshwara S. (2005) Oligomeric protein structure networks: insight into protein-protein interactions. *BMC Bioinformatics,* **6:**296.

Butland G., Peregrin-Alvarez J.M., Li J., Yang W., Yang X., Canadien V., Starostine A., Richards D., Beattie B., Krogan N., Davey M., Parkinson J., Greenblatt J., Emili A. (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature,* **433:**531-537.

Carugo O., Franzot G. (2004) Prediction of protein-protein interactions based on surface patch comparison. *Proteomics,* **4,** 1727-1736.

Caspi R., Foerster H., Fulcher C.A., Hopkinson R., Ingraham J., Kaipa P., Krummenacker M., Paley S., Pick J., Rhee S.Y., Tissier C., Zhang P., Karp P.D. (2006) MetaCyc: A multi-organism database of metabolic pathways and enzymes. *Nucleic Acids Research,* **34:**D511-D516.

Castillo-Davis C., Hartle D.L. (2003) GeneMerge: post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics,* **19(7),** 891-892.

Chen L., Vitkup D. (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biology,* **7:**R17.

Chen Y., Xu D. (2004) Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. *Nucleic Acids Research,* **32(21)**, 6414-6424.

Colizza V., Flammini A., Maritan A., Vespignani A. (2005) Characterization and modeling of protein-protein interaction networks. *Physica A,* **352**, 1-27.

Dandekar T., Snel B., Huynen M., Bork P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **9,** 324-328.

Date S.V., Marcotte E.M. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology,* **21,** 1055-1062.

Del Sol A., Fujihashi H., O'Meara P. (2005) Topology of small-world networks of protein-protein complex structures. *Bioinformatics,* **21(8)**, 1311-1315.

Deng M., Mehta S., Sun F., Chen T. (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Research,* **12**, 1540-1548.

D'haeseleer P. (2005) How does gene expression clustering work? *Nature Biotechnology*, **23:** 1499-1501.

Dietmann S., Holm L. (2001) Identification of homology in protein structure classification. *Nature Structural Biology,* **8(11),** 953-957.

Doerks T., von Mering C., Bork P. (2004) Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes. *Nucleic Acids Research,* **32(21):**6321-6326.

Eisen J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research,* **8,** 163-167.

Eisen J.A. and Wu M. (2002) Phylogenetic analysis and gene functional predictions: phylogenomics in action. *Theoretical Population Biology*, **61**, 481-487.

Eisenberg D., Marcotte E.M., Xenarios I., Yeates T.O. (2000) Protein function in the post genomic era. *Nature,* **405,** 823-826.

Elofsson A., Sonnhammer E.L.L. (1999) A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics,* **15(6),** 480-500.

Espadaler J., Romero-Isart O., Jackson R.M., Oliva B. (2005) Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics,* **21(16)**, 3360-3368.

Fell D.A. (2001) Beyond genomics. *TRENDS in Genetics,* **17(12),** 680-682.

Fields S., Song O. (1989) A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-246.

Francke C., Siezen R.J., Teusink **.** (2005) Reconstructing the metabolic network of a bacterium from genome. *TRENDS in Microbiology,* **13(11):**550-558.

Franzot G., Carugo O. (2003) Computational approaches to protein-protein interaction. *Journal of Structural and Functional Genomics,* **4,** 245-255.

Gandhi T.K.B., Zhang J., Mathivanan S., Karthick L., Chandrika K.N., Mohan S.S., Sharma S., Pinkert S., Nagaraju S., Periaswamy B., Mishra G., Nandakumar K., Shen B., Deshpande N., Nayak R., Sarker M., Boeke J.D., Parmigiani G., Schultz J., Bader J.S., Pandey A. (2006) Analysis of the human protein interactome and comparison with yeast, worm, and fly interaction datasets. *Nature Genetics,* **38:**285-293.

Gerlt J.A., Babbitt P. (2000) Can sequence determine function? *Genome Biology,* **I(5)**.

Grant B.D., Wilkinson H.A. (2003) Functional genomic maps in *Caenorhabditis elegans*. *Current Opinion in Cell Biology*, **15**, 206-212.

Gunther C.S., Gaasterland T. (2001) Characterizing the relationship between protein fusion and gene co-expression. *Genome Informatics,* **12,** 34-43.

Han D.-S., Kim H.-S., Jang W.-H., Lee S.-D., Suh J.K. (2004) PreSPI: design and implementation of protein-protein interaction prediction service system. *Genome Informatics,* **15,** 171-180.

Han L.Y., Cai C.Z., Lo S.L., Chung M.C.M., Chen Y.Z. (2004) Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA,* **10:**355-368.

Hatzimanikatis V., Li C., Ionita J.A., Broadbelt L.J. (2004) Metabolic networks: enzyme function and metabolite structure. *Current Opinion in Structural Biology*, **14,** 300-306.

Hayashida M., Ueda N., Akutsu T. (2004) A simple method for inferring strengths of protein-protein interactions. *Genome Informatics,* **15,** 56-68.

Hegyi H., Gerstein M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288,** 17-164.

Hoffman R., Valencia A. (2003) Protein interaction: same network, different hubs. *TRENDS in Genetics*, **19(12):**681-683.

Huang H.-D., Lee T.-Y., Wu L.-C., Lin F.-M., Juan H.-F., Horng J.-T., Tsou A.P. (2005) MultiProIdent: Identifying proteins using database search and protein-protein interactions. *Journal of Proteome Research*, **4**: 690-697.

Huang T.-W., Tien A.-C., Huang W.-S., Lee Y.-C.G., Peng C.-L., Tseng H.-H., Kao C.-Y., Huang C.-Y.F. (2004) POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, **20(17):**3273-3276.

Huang Y., Frishman D., Muchnik I. (2004) Predicting protein-protein interactions by a supervised learning classifier. *Computational Biology and Chemistry*, **28,** 291-301.

Hu X. (2005) Mining and analyzing scale-free protein-protein interaction network. *Int. J. Bioinformatics Research and Applications,* **1(1):**81-101.

Huynen M., Dandekar T., Bork P. (1998) Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Letters,* **426,** 1-5

Huynen M., Snel B., Lathe W., Bork P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inference. *Genome Research,* **10,** 1204-1210.

Jackson R.M., Sternberg M.JE. (1995) A continuum model for protein-protein interactions: application to the docking problem. *J. Mol. Biol.,* **250:**258-275.

Janin J., Henrick K., Moult J., Eyck L.T., Sternberg M.J.E., Vajda S., Vakser I., Wodak S.J. (2003) CAPRI: A critical assessment of predicted interactions. *PROTEINS: Structure, Function, and Genetics,* **52:**2-9.

Janssen P., Audit B., Cases I., Darzentas N., Golddovsky L., Kunin V., Lopez-Bigas N., Peregrin-Alvarez J.M., Pereira-Leal J.B., Tsoka S., Ouzonis C.A. (2003) Beyond 100 genomes. *Genome Biology,* **4:**402.

Kanehisa M., Goto S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research,* **28,** 27-30.

Kim S.K., Lund J., Kiraly M., Duke K., Jiang M., Stuart J.M., Eizinger A., Wylie B.N., Davidson G.S. (2001) A gene expression map for *Caenorhabditis elegans*. *Science,* **293,** 2087-2092.

Kim W.K., Park J., Suh J.K. (2002) Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. *Genome Informatics,* **13,** 42-50.

Koonin E.V., Wolf Y.I. (2006) Evolutionary systems biology: links between gene evolution and function. *Current Opinion in Biotechnology,* **17:**481-487.

Kortemme T., Baker D. (2004) Computational design of protein-protein interactions. *Current Opinion in Chemical Biology,* **8:**91-97.

Li S., Armstrong C.M., Bertin N., Ge H., Milstein S., Boxem M., Vidalain P.O., Hao T., Goldberg D.S., Li N., Martinez M., Rual J.F., Lamesch P., Xu L., Tewari M., Wong S.L., Zhang L.V., Berriz G.F., Jacotot L., Vaglio P., Reboul J., Hirozane-Kishiawa T., Li Q., Gabel H.W., Gabel H.W., Elewa A., Baumgartner B., Rose D.J., Yu H., Bosak S., Sequerra R., Fraser A., Mango S.E., Saxton W.M., Strome S., Van den Heuvel S., Piano F., Vandenhaute J., Sardet C., Gerstein M., Doucette-Stamm L., Gunsalus K.C.,

Harper J.W., Cusick M.E., Roth F.P., Hill D.E., Vidal M. (2004) A map of interactome network of the metazoan *C. elegans*. *Science,* **303,** 540-543.

Liang Z., Xu M., Teng M., Niu L. (2006) NetAlign: a web-based tool for comparison of protein interaction networks. *Bioinformatics,* **22(17):**2175-2177.

Littler S.J., Hubbard S.J. (2005) Conservation of orientation and sequence in protein domain-domain interactions. *J. Mol. Biol.*, **345**, 1265-1279.

Liu Y., Liu N., Zhao H. (2005) Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics,* **15,** 3279-3285.

Lo S.L., Cai C.Z., Chen Y.Z., Chung M.C.M. (2005) Effect of training datasets on support vector machine prediction of protein-protein interactions. *Proteomics,* **5:**876-884.

Lu L., Lu H., Skolnick J. (2002) MULTIPROSPECTOR : an algorithm for the prediction of protein-protein interactions by multimeric threading. *PROTEINS: Structure, Function, and Genetics,* **49,** 350-364.

Lu L., Arakaki A.K., Lu H., Skolnick J. (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Research,* **13,** 1146-1154.

Lu L.J., Xia Y., Paccanaro A., Yu H., Gerstein M. (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Research,* **15:**945-953.

Marcotte C.J.V. and Marcotte E.M. (2002) Predicting functional linkages from gene fusions with confidence. *Applied Bioinformatics,* **1(2),** 93-100.

Marcotte E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Current Opinion in Structural Biology*, **10**, 359-365.

Marcotte E.M., Pellegrini M., Ng H.-L., Rice D.W., Yeates T.O., Eisenberg D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science,* **285,** 751-753.

Marcotte E.M., Xenarios I., Eisenberg D. (2001) Mining literature for protein-protein interactions. *Bioinformatics*, **17(4)**, 359-363.

Mata J., Bahler J. (2003) Correlations between gene expression and gene conservation in fission yeast. *Genome Research,* **13,** 2686-2690.

Mendez R., Leplae R., De Maria L., Wodak S.J. (2003) Assessment of blind predictions of protein-protein interactions: current status of ducking methods. *PROTEINS: Structure, Function, and Genetics,* **52:**51-67.

Murvai J., Vlahovicek K., Barta E., Cataletto B., pongor S. (2000) The SBASE protein domain library, release 7.0: a collection of annotated protein sequence segments. *Nucleic Acids Research,* **28(1),** 260-262.

Nanni L., Lumini A. (2006) An ensemble of K-local hyperplanes for predicting protein-protein interactions. *Bioinformatics,* **22(10):**1207-12010.

Needham C.J., Bradford J.R., Bulpitt A., Westhead D.R. (2006) Inference in bayesian networks. *Nature Biotechnology,* **24:**51-53.

Noble W.S. (2006) What is a support vector machine? *Nature Biotechnology*, 24(12):1565-1567.

Park J., Lappe M., Teichmann S. (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.,* **307**, 929-938.

Pazos F., Valencia A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering,* **14(9):**609-614.

Pazos F., Valencia A. (2002) *In silico* two-hybrid system for the selection of physically interacting protein pairs. *PROTEINS: Structure, Function, and Genetics,* **47:**219-227.

Pellegrini M. (2001) Computational methods for protein function analysis. *Current Opinion in Chemical Biology,* **5**, 46-50.

Pellegrini M., Marcotte E.M., Thompson M.J., Eisenberg D., Yeates T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* USA **96**, 4285-4288.

Ponting C.P., Russell R.R. (2002) The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.,* **31,** 45-71.

Rain J.C., Selig L., De Reuse H., Battaglia V., Simon S., Lenzen G., Petel F., Wojcik J., Schachter V., Chemama Y., Labigne A., Legrain P. (2001) The protein-protein interaction map of *Helicobacter pylori*. *Nature,* **409,** 211-215.

Ramani A.K., Bunescu R.C., Mooney R.J., Marcotte E.M. (2005) Consolidating the set of known human protein-protein interaction for large-scale mapping of the human interactome. *Genome Biology*, **6:**R40.

Reed J.L., Famili I., Thiele I., Palsson B.O. (2006) Towards multidimensional genome annotation. *Nature Reviews Genetics*, **7,** 130-141.

Rhodes D.R., Tomlins S.A., Varambally S., Mahavisno V., Barrette T., Kalyana-Sundaram S., Ghosh D., Pandey A., Chinnaiyan A.M. (2005) Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology*, **23,** 951-959.

Riley R., Lee C., Sabatti C., Eisenberg D. (2005) Inferring protein domain interactions from database of interacting proteins. *Genome Biology,* **6:**R89.

Rubin G.M., Yandell M.D., Wortman J.R., Miklos G.L.G., Nelson C.R., Hariharan I.K., Fortini M.E., Li P.W., Apweiler R., Fleischmann W., Cherry J.M., Henikoff S., Skupski M.P., Misra S., Ashburner M., Birney E., Boguski M.S., Brody T., Brokstein P., Celniker S.E., Chervitz S.A., Coates D., Cravchik A., Gabrielian A., Galle R.F., Gelbart W.M., George R.A., Goldstein L.S.B., Gong F., Guan P., Harris N.L., Hay B.A., Hoskins R.A., Li J., Li Z., Hynes R.O., Jones J.M., Kuehl P.M., Lemaitre B., Littleton J.T., Morrison D.K., Mungall C., O'Farrell P.H., Pickeral O.K., Shue C., Vosshall L.B., Zhang J., Zhao Q., Zheng X.H., Zhong F., Zhong W., Gibbs R., Venter J.C., Adams M.D., Lewis S. (2000) Comparative genomics of the eukaryotes. *Science,* **287,** 2204-2215.

Salwinski L., Miller C.S., Smith A.J., Pettit F.K., Bowie J.U., Eisenberg D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32,** D449-D451.

Sato T., Yamanishi Y., Kanehisa M., Toh H. (2005) The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics,* **21(17):**3482-3489.

Schomburg I., Chang A., Schomburg D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Research,* **30:**47-49.

Schwikowski B., Uetz P., Fields S. (2000) A network of protein-protein interactions in yeast. *Nature Biotechnology,* **18,** 1257-1261.

Sharan R., Suthram S., Kelley R.M., Kuhn T., McCuine S., Uetz P., Sittler T., Karp R.M., Idekar T. (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci.* USA, **102**, 1974-1979.

Shi T.L., Li X.Y., Cai Y.D., Chou K.C. (2002) Computational methods for protein-protein interaction and their application. *Current Protein and Peptide Science*, **6(5):**443-449.

Smith G.R., Sternberg M.JE. (2002) Prediction of protein-protein interactions by docking methods. *Current Opinion in Structural Biology*, **12:**28-35.

Sonnhammer E.L.L., Koonin E.V. (2002) Orthology, paralogy, and prposed classification for paralog subtypes. *TRENDS in Genetics*, **18(12),** 619-620.

Sprinzak E., Margalit H. (2001) Correlated sequence-signatures as markers of protein-protein interactions. *J. Mol. Biol.*, **311**, 681-692.

Strong M., Mallick P., Pellegrini M., Thompson M.J., Eisenberg D. (2003) Inference of protein function and protein linkages in Mycobacterium tuberculosis based on prokaryotic genome organization : a combined computational approach. *Genome Biology*, **4,** R59.

Sun S., Zhao Y., Jiao Y., Yin Y., Cai L., Zhang Y., Lu H., Chen R., Bu D. (2006) Faster and more accurate global protein function assignment from protein interaction networks using the MFGO algorithm. *FEBS letters*, **580:**1891-1896.

Tolstrup N., Nielsen P.S., Kolberg J.G., Frankel A.M., Vissing H., Kauppinen S. (2003) OligoDesign: optimal design of LNA (locked nucleic acid) oligonucleotide capture probes for gene expression profiling. *Nucleic Acids Research*, **31(13),** 3758-3762.

Truong K., Ikora M. (2003) Domain fusion analysis by applying relational algebra to protein sequence and domain databases. *BMC Bioinformatics*, **4:**16.

Uetz P., Pankratz M.J. (2004) Protein interactions on the fly. *Nature Biotechnology*, **22(1),** 43-44.

van Noort V., Snel B., Huynen M.A. (2003) Predicting gene functions by conserved co-expression. *TRENDS in Genetics*, **19,** 238-242.

Vazquez A., Flammini A., Maritan A., Vespignani A. (2003) Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, **21(6)**, 697-700.

Vo T.D., Greenberg H.J., Palsson B.O. (2004) Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *The Journal of Biological Chemistry,* **279(38):**39532-39540.

Wallach D., Boldin M.P., Kovalenko A.V., Malinin N.L., Mett I.L., Camonis J.H. (1998) The yeast two-hybrid screening technique and its use in the study of protein-protein interactions in apoptosis. *Current Opinion in Immunology,* **10**:131-136.

Wodak J.S., Mendez R. (2004) Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. *Current Opinion in Structural Biology,* **14:**242-249.

Wojcik J., Schachter V. (2001) Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics,* **17 Suppl.**, S296-S305.

Wu J., Kasif S., DeLisi C. (2003) Identification of functional links between genes using phylogenetic profiles. *Bioinformatics,* **19(12)**, 1524-1530.

Yamada T., Goto S., Kanehisa M. (2004) Extraction of phylogenetic network modules from prokaryote metabolic pathways. *Genome Informatics,* **15,** 249-258.

Yu J., Fotouhi F. (2006) Computational approaches for predicting protein-protein interactions: A survey. *Journal of Medical Systems,* **30(1):**39-44.

Zaki N., Deris S., Alashwal H. (2006) Protein-protein interaction detection based on substring sensitivity measure. *International Journal of Biomedical Sciences,* **1(2):**148-154.

Zhou X.J., Cao M.C., Huang H., Wong A., Nunez-Iglesias J., Primig M., Aparicio O.M., Finch C.E., Morgan T.E., Wong W.H. (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nature Biotechnology*, **23(2)**, 238-243.

Zhu H., Domingues F.S., Sommer I., Lengauer T. (2006) NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics,* **7:**27.

2

OBJECTIVES

According to the knowledge gap explained in previous chapter the following objectives were planned to achieve in this research:

1. Reconstruction of metabolic network of *C. elegens* to evaluate the current number of known protein-protein interactions in the genome of this organism. Currently available interaction information in different databases was integrated to achieve the interaction map of all enzymes known to active in different metabolic pathways.
2. Developing a new computational protein-protein interaction prediction method to predict novel interactions and to infer previously uncharacterized proteins.
3. Evaluation of validation of predicted protein interacting pairs and quantification by means of statistical techniques.
4. Augmentation of the reconstructed metabolic network of *C. elegans* by introducing a new two-dimensional genome annotation using predicted protein-protein interaction information to achieve a larger map for *C. elegans* protein interactions.

To achieve the above mentioned objectives, bioinformatics was used as a tool to explore numerous databases, parse suitable data, and integrate different pieces of information.

As an overview, what was done in this research was generating reliable protein-protein interactions using a newly developed computational interaction prediction method. Then with the aid of a proposed algorithm the predicted interactions were filtered and the number of false positives was substantially decreased. Next, the predicted and filtered data was incorporated to the current metabolic network of *C. elegans* resulting in an expanded version of the network. Along with the expanded network, new functions were inferred for unknown proteins embedded to the expanded network.

This dissertation has been organized according to the stated objectives. In Chapter 1 previous work on protein-protein interaction prediction and reconstruction of metabolic

networks has been reviewed. Moreover, the widely used experimental high-throughput screening technique, yeast two-hybrid, has been described from molecular point of view. In Chapter 2 the general objectives of the research and the organization of the report has been briefly explained. In Chapter 3, the current metabolic network of *C. elegans* has been reconstructed and the resulting protein-protein interaction map of *C. elegans* has been inferred. A simple procedure has been employed to integrate different levels of information and place them in a network context. The current situation of *C. elegans* metabolic network has been evaluated at the end of this chapter. A new computational protein-protein interaction prediction approach has been introduced in Chapter 4. This method has been developed based on the concept that the similarity between profiles of signature content of proteins may play a role in functional or physical interactions. This method has been compared with equivalent approaches and it has been shown that this method outperforms the peer approaches. Statistical analysis of the results has also proved that inferred interactions are significant. Due to the overall high false positive results in computational approaches, a global framework has been proposed in Chapter 5 in an attempt to reduce the number of false positives in every predicted dataset. The framework is a post-prediction processing procedure to remove predicted interacting protein pairs which do not comply with ontology and annotations. After applying the proposed algorithm to different datasets and comparing them with high-confidence experimental datasets, the mass reduction of false positives has been statistically evaluated. The new protein-protein interaction dataset with reduced number of false positives has been incorporated into the current metabolic network of *C. elegans* in Chapter 6 to achieve an expanded network of this organism. When new interactions are placed into a biological context some uncharacterized enzymes, missing relationships, and consistent interactions are revealed. General discussion, overall conclusions and recommendations are presented in chapter 7. All output files, datasets, and PERL computer programs along with adequate comments are presented in Supplementary Data. The organization of Supplementary Data is described in Appendix.

3

RECONSTRUCTION OF METABOLIC NETWORK OF CAENORHABDITIS ELEGANS

Contribution of this chapter to the overall study

In order to make any contribution to expand the metabolic network of the studied organism the current situation of the network should be assessed. In this chapter a new strategy to reconstruct metabolic networks emphasizing on the use of recent genomic information available in public databases was developed. The resulted network was studied quantitatively and qualitatively.

3.1 Abstract

With the completion of sequencing of *C. elegans* in 1998, the metabolic network reconstruction of this species became possible. As of yet several global metabolic network reconstruction algorithms have been proposed, many of which are more appropriate for bacterial and prokaryotic genomes. *C. elegans*, as a multi-cellular eukaryotic model organism, needs to be studied individually to specify some specific cellular organizations, such as metabolic pathways and their relationships. Further, most of network reconstruction algorithms focus on the strings of biochemical reactions to reconstruct the network, while the role of enzymes in the interconnecting behaviour of a network and revealing hidden mechanisms to perform biological tasks has not yet been well studied. With the use of conventional reconstruction algorithms, and considering functional approach to interconnect metabolic enzymes at different pathways, the metabolic network of *C. elegans* was reconstructed. In this reconstruction, different levels of current biological information including genes, enzymes, and reactions were related together. Then a mechanism was proposed to identify biological relationships among pathways upon specifying key enzymes. These enzymes were revealed by examining the most-connected pathways, resulting in the identification of primary pathways. Key enzymes contributed to the interconnecting nature of the network, based on which different pathways were functionally linked together. Metabolic paths in the network represented linked pathways and the metabolic paths with the highest values represented the most probable routes taken by the organism where endogenous sources of nutrient are

available to the organism. A specific example, contribution of energy metabolism pathway to replicate DNA molecules, was demonstrated to perceive how functionally related pathways collaborate.

3.2 Introduction

Two-dimensional genome annotation refers to reconstruction of networks based upon one-dimensional genome annotation. In fact, metabolic reconstruction is assembling a puzzle with many different pieces (Marcotte, 2003). Metabolic networks are examples of protein networks that represent the entire network of biochemical reactions carried out by a living cell. The complete description of a metabolic network not only includes small molecules, large molecules, intermediates, and metabolic products of cellular reactions, but also the characteristics of relevant enzymes. In a metabolic network distinct sequences of reactions are grouped in pathways. Enzymes that catalyze different reactions in a pathway are encoded by protein-encoding genes.

Therefore, a metabolic network is a complete picture of the metabolisms of species based on the sequence of the genes that encode metabolic enzymes. For examples in prokaryotes, about 900 *E. coli* genes encode enzymes which are distributed into 130 different pathways. These genes account for about 21% of the genes in the *E. coli* genome. In *Penicillium chrysogenum* new metabolic activities in two novel pathways were identified as a result of metabolic network analysis (Christensen and Nielson, 2000). Metabolic network of *H. influenza* contains 448 metabolic reactions operating on 443 metabolites (Edwards and Palsson, 1999). In eukaryotes, *S. cerevisiae* has 5900 protein-encoding genes; among them 1200 genes (~20%) encode enzymes involved in metabolism (energy reservoirs). In the fruit fly (*Drosophila melanogaster)* 2400 (~17%) out of 14100 genes are involved in metabolic pathways (Horton *et al.*, 2002). The *C. elegans*, a small multi-cellular animal, is another extensively studied model organism which approximately 5400 (~25% of the proteome) of its proteins have been studied by experimental techniques (Mawuenyega *et al.,* 2003). Its genome has already been sequenced (The *C. elegans* Sequencing Consortium, 1998) and is available in web access databases (Stein, 1999). This is a model organism because many of its specialized cells and tissues are also found in larger species such as human (Hekimi *et al.,* 1998).

To reconstruct metabolic networks researchers have exploited different strategies. On early attempts a three-step procedure was implemented to reconstruct prokaryotic metabolic networks that included gathering a list of metabolic genes, assigning reactions to the genes, and adding physiological information about the organism to the record related to each gene (Covert *et al.,* 2001). The same three-step procedure was also applied to analyze metabolic pathways of parasites (Fairlamb, 2002). Another study (Forster *et al.,* 2003) focused on metabolic reconstruction of *S. cerevisiae* as the first comprehensive network for a eukaryotic organism. In this work the metabolic reactions were categorized between cytosol and mitochondria, and transport steps between these two compartments were included. Famili and Palsson (2003) proposed a systemic analysis of genome-scale biochemical conversion properties using singular value decomposition, aiming at comparing overall properties of genome-specific metabolic networks. This approach focused on the systemic aspect of metabolic reactions, but not the relationship among associate metabolites and other elements within a genome.

With the ignorance of the currency metabolites such as ATP, NADH, etc., Ma and Zeng (2003) reconstructed a global metabolic network for 80 organisms of interest, resulting in the different average path length between any pair of metabolites in three domains of life: eukaryotes, archaea, and bacteria. They reconstructed the metabolic network using a revised bioreaction information database in which reversible reactions were represented by undirected connections and directed connections corresponded to irreversible reactions. It was then clear that the choice of connectivity exerted a significant influence on the estimation of path length of a network. In an another report, Sun and Zeng (2004) used the similar network reconstruction strategy along with a modified method to prepare their data set which consists of simultaneous gene finding from genome database and gene annotation. After these two parallel processes, the network reconstruction was performed. Miyake *et al.* (2004) proposed a graph analysis method to identify the metabolic sub-networks or building blocks of metabolic networks. They used compound-reaction relations as the dataset. This dataset was searched for highly conserved sequential reactions to identify sub-networks.

The initial reconstruction of human mitochondrial metabolic network was already performed based on recently published proteomic data. The dataset in this work consisted

of 189 reactions and 230 metabolites mostly involved in energy metabolism (Vo *et al.*, 2004). These reactions were distributed among three cellular compartments including mitochondrial, cytosol, or extracellular and as a result main metabolic functions in these three locations were determined. In another study, a computational method was proposed to identify human metabolic pathways based on complete human genome (Romero *et al.*, 2004); however, the sophisticated human metabolic network is still far from complete. The metabolic network of a pathogenic strain of *Staphylococcus aureus* was also reconstructed to elucidate some properties of this resistant strain to many antibiotics (Becker and Palsson, 2005). Metabolic network reconstruction of bacteria has already been established (Francke *et al.,* 2005). Recently, a semi-automated approach was introduced to accelerate the process of genome-scale metabolic network reconstruction (Notebaart *et al.,* 2006). This approach took the advantage of availability of manually curated networks to predict gene-reaction relationships and expanded current networks. A few attempts were made to integrate different levels of information on *C. elegans* to perform biological hypothesis (Walhout *et al.,* 2002); however, these efforts were not focused on the reconstruction of metabolic networks.

As a general outline to reconstruct metabolic networks, different levels of information should be integrated as follows to obtain a detailed description of biochemical transformation. At the first level, the metabolite specificity of a gene product should be defined. Although primary metabolites are often the same for homologous enzymes across organisms, the use of coenzymes might vary. The second level of detail accounts for the stoichiometry and directionality of reactions considering thermodynamic properties of metabolites and cofactors. At the third level, the cellular compartment in which the reaction takes place has to be determined. Pathway association of some enzymes is in accordance to their cellular compartments. Although pathway boundaries are rather arbitrary, considering close pathways in same common cellular compartment is not far from the reality.

In this chapter, metabolic data concerning *C. elegans* was retrieved from biological databases such as KEGG. This data included various levels of metabolism including genes, proteins (enzymes), metabolites, and reactions. This data was examined with other databases such as SWISS-PROT as a validation step. In this step the metabolite

specificity of reactions were specified and multi-function enzymes which appear in different pathways were identified. Directions of reactions were determined in terms of reversibility or irreversibility. Then, an algorithm was developed to integrate different levels of information and assign genes, enzymes, metabolites, and reactions to pathways. The boundaries of pathways were considered as it was in KEGG. Next, based on post-genomic definition of protein interaction, protein-protein interaction map of the studied organism was assembled which represents a summary of all current interaction information on this organism.

3.3 Methods

3.3.1 Dataset preparation

In order to reconstruct the metabolic map the Kyoto Encyclopaedia of Genes and Genomes (Kanehisa and Goto, 2000; Release 32) database was used as reference. The KEGG database contains genomes, reactions, pathways, and EC tables of many sequenced species. This database is one of the most comprehensive databases in which different levels of biological information such as genomics, proteomics, transcriptomics, and metabolomics are integrated and pathways are reconstructed based upon published data (Nakao *et al.*, 1999). It is updated weekly and available for public access. The first step toward the network reconstruction is to retrieve information relating to *C. elegans* from KEGG and save to a local computer, including pathways, reactions, and genes. These three sets of information contain pathway numbers and the descriptions, all reactions carried out in the pathways, and gene entries (ORF names) along with their nucleotide sequences and the amino acids sequences of encoded proteins. Three perl scripts were developed and used to extract relevant biological information from downloaded files. The outputs from perl scripts were stored in the following three files (`celPath, celReact, celGene`) accordingly, and integrated to reconstruct metabolic network for *C. elegans* (see Supplementary Data, Chapter 3). `celPath` is a list of all 94 metabolic pathways within the *C. elegans* genome and their descriptions. `celReact` is a list of all reaction-enzyme relations, and `celGene` contains a list of

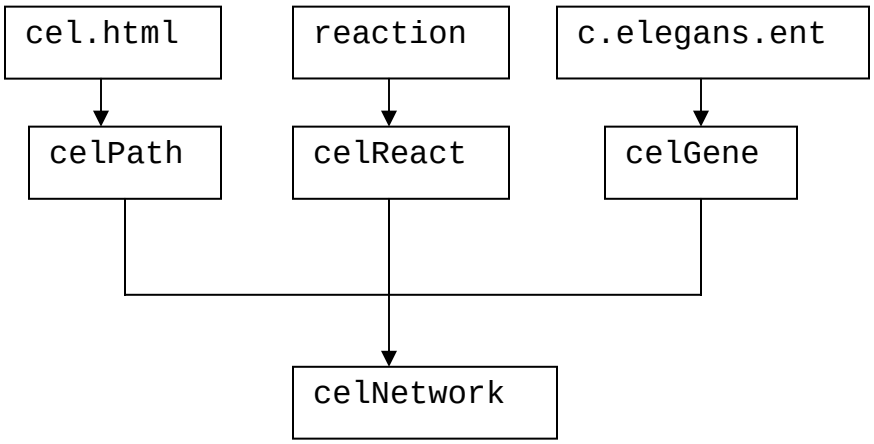

Figure 3.1. The flowchart for the reconstruction of metabolic network of *C. elegans*. In this flowchart three KEGG reference files: cel.html, reaction, and c.elegans.ent, were used as input data for three perl scripts. The outputs from these scripts were used as input data for another perl script to reconstruct the metabolic network.

22740 gene entries in which some have gene names, encoded enzymes and associated pathway(s) with the gene. These entries were checked for whether genes were missing using SWISSPROT (Bairoch and Boeckman, 1992) and WormBase (Stein *et al.*, 2001; wormpep152). The extraction process is illustrated in Figure 3.1.

3.3.2 Data integration and network reconstruction

We automatically integrated the information and categorized enzymes into each particular pathway and the resulting information was collected into '`celNetwork`' file (see Supplementary Data, Chapter 3). In this integration process pathway numbers are the central information and all other data are directed toward pathways such that each record of results is introduced with a pathway number. Based on the fact that a pathway is a collection of biochemical reactions and each reaction is catalyzed by an enzyme and each enzyme (protein) is encoded by a gene, we have created a list of genes associated with a pathway and other sequential biological processes including encoding enzymes and catalyzing reactions. For each gene, other pathways that this particular gene is participating in are found, however, only the reaction which is catalyzed in the pathway of interest is reported and other reactions which may be catalyzed by this enzyme are excluded form this record of results. A partial listing of this file (only one record of results) is presented in Figure 3.2. Each record of information in this file is related to one pathway only. In total, there are 94 information records collected in '`celNetwork`' involving 792 genes.

3.3.3 Protein-protein interaction map

The classical view of protein interaction focuses on the action of a single protein molecule. In metabolism, this action may be the catalysis of a given reaction or the binding of a small or large molecule. In the post genomic era, this local interaction may help to find molecular function of a protein; however, it does not represent the role of a protein as an element in the network of interactions. The idea is that each protein in living matter functions as part of an extended web of interacting molecules. In the expanded view of interaction, proteins that participate in a common structural complex or metabolic pathway are defined as interacting proteins (Eisenberg *et al.,* 2000). Several prokaryotic and eukaryotic protein interaction maps have been reported successfully based on this

```
cel00040   Pentose and glucuronate interconversions - Caenorhabditis
elegans
1. Y105E8B.9  *****   cel00040 cel00500 cel00531 cel00860
   [EC:3.2.1.31]
2. F35H8.6  *****   cel00040 cel00150 cel00500 cel00860
   [EC:2.4.1.17]
     R01379  UDPglucuronate + H2O <=> UDP + D-Glucuronate
3. K08E3.5a  *****   cel00040 cel00052 cel00500 cel00520
   [EC:2.7.7.9]
     R00289  UTP + D-Glucose 1-phosphate <=> Pyrophosphate + UDPglucose
4. C18C4.3  *****   cel00040 cel00150 cel00500 cel00860
   [EC:2.4.1.17]
     R01379  UDPglucuronate + H2O <=> UDP + D-Glucuronate
5. F29F11.1  *****   cel00040 cel00500 cel00520
   [EC:1.1.1.22]
     R00286  UDPglucose + H2O + 2 NAD+ <=> UDPglucuronate + 2 NADH + H+
6. T04H1.7  *****   cel00040 cel00150 cel00500 cel00860
   [EC:2.4.1.17]
     R01379  UDPglucuronate + H2O <=> UDP + D-Glucuronate
7. B0310.5  *****   cel00040 cel00150 cel00500 cel00860
   [EC:2.4.1.17]
     R01379  UDPglucuronate + H2O <=> UDP + D-Glucuronate
8. T07C5.1a  *****   cel00040 cel00150 cel00500 cel00860
   [EC:2.4.1.17]
     R01379  UDPglucuronate + H2O <=> UDP + D-Glucuronate
9. T07C5.1b  *****   cel00040 cel00150 cel00500 cel00860
   [EC:2.4.1.17]
        R01379  UDPglucuronate + H2O <=> UDP + D-Glucuronate
```

Figure 3.2. Partial listing of 'celNetwork.txt'. Each record of information in this file starts with a pathway number (cel00040 in this case) and the description of the pathway. In this typical pathway there are 9 associated genes. For each gene entry (for example T07C5.1b) the encoded enzyme (in the form of EC number) as well as the reaction catalyzed by this enzyme is shown. In addition, the other pathways that this gene (T07C5.1b) is participating in are indicated as cel00150, cel00500, and cel00860. In the cases that the gene name is unknown star signs are printed. There are some cases in which although the enzyme translated by the gene is known (Y105E8B.9), the reaction catalyzed by this enzyme in this particular pathway (cel00040) is yet to be determined.

new definition of interaction (Enright *et al.*, 1999). Furthermore, it is believed that proteins form permanent or transient complexes to provide a response to external stimuli (Szilagyi *et al.*, 2005). Proteins aggregated in these complexes work together to accomplish part of an entire biological process. Sometimes one single protein ought to work with several other proteins to transmit a signal or regulate a biochemical reaction. Most of permanent complexes are in accordance to pathways or cellular components. Therefore, in order to infer protein-protein interactions upon constructed metabolic network of *C. elegans*, proteins participating in same metabolic pathways were considered interacting.

3.4 Results and Discussion

With the current reconstructed metabolic network of *C. elegans* (see Supplementary Data, Chapter 3), known proteins in this network are connected together in a pair-wise fashion, based on the notion that proteins in same metabolic pathways interact with each other. There are 32902 interactions involving 792 proteins in 94 metabolic pathways in the current protein-protein interaction map of *C. elegans* (see Supplementary Data, Chapter 3).

3.4.1 Connectivity in the protein-protein interaction map

The average connectivity of each protein in the current map is 42 interactions. This complies with the estimation that each protein generally interacts with about 5 to 50 proteins (Huzbun and Fields, 2001). However, correlation between connectivity and other protein biochemical properties such as hydrophobicity has been suggested (Deeds *et al.*, 2006). In some protein-protein interaction maps distribution of edges among nodes follow a power law model (Hoffman and Valencia, 2003). There are proteins which catalyze the same reactions in different pathways. These proteins may contribute to the interconnectivity of the protein interaction network and serve as the hubs of the network. Hub proteins are conserved structures with the higher number of connectivity compared to other proteins. Thus, the possibility of finding new interactions for these proteins is higher than low connected proteins. In the current protein interaction map the proteins in energy metabolism pathways are the most connected ones as most metabolic reactions need energy to proceed.

The properties of protein interaction map depend on the accuracy and validity of genomic information utilized in the reconstruction of metabolic network. At the time being, numerous techniques are available enabling researchers to produce huge amount of biological information on different species; however, the reliability of this information is still in question. Currently the agreement among three major databases of genomic information including KEGG, MIPS, and GO is surprisingly poor (Bork *et al.*, 2004), even though these databases are the main sources of metabolic and genomic information that metabolic networks are reconstructed upon. Therefore, the reconstructed metabolic network presented here is built based on most recent information publicly available to research community.

3.4.2 Quantitative analysis of the reconstructed network

All biological elements related to each pathway including genes, enzymes and reactions catalyzed by each enzyme are integrated in the reconstructed metabolic network. Each gene is accompanied with the encoded protein (enzyme) and the enzyme is followed by the reaction catalyzed by that enzyme. For example, there are 38 genes currently associated in the glycolysis pathway and 29 genes are currently found in the TCA cycle. There are 22,740 ORFs in the *C. elegans* genome, including 21,357 protein genes (coding sequences or CDS), and 753 RNA genes. Of 22,740 ORFs, 1,361 have known entries which count for 6% of the entire genome. Of the entire known entries, 792 entries, involved in metabolism, have known pathways such as glycolysis, citric acid cycle and so on. The remaining 569 proteins are annotated proteins that their pathways are still unknown. The relationship between the unassigned 569 annotated proteins and the pathway association requires further investigation.

3.4.3 Qualitative analysis of the reconstructed network

The pathway-gene relation information can be represented by an undirected two-mode network. A two-mode network consists of two set of units (vertices) and relations (edges). In this representation, there are 94 vertices corresponding to 94 pathways as the first set of units, and 792 vertices corresponding to genes as the second set of units. Pathway-gene relations, connecting two sets of units together, are linked by ‘undirected

lines' known as edges. Each pathway is introduced by an index such as cel00100 and genes are shown by their ORF synonyms.

Knowing the fact that each metabolic network comprises several pathways which include common (genes shared by different pathways) and uncommon (genes associated with only one particular pathway) groups of genes, these genes are classified according to their connection degrees, which is defined as the total number of edges coming in or going out of a vertex. As a result, there were 363 genes with connection degrees 1, each of which was associated with only one pathway, encoding an enzyme which catalyzes a distinct reaction in that pathway. Another group contains 429 genes which have the connection degrees greater than 1, involving in different pathways, catalyzing the same or distinct reactions, and contributing to the interconnecting nature of the network. Distribution of this group of genes among pathways is illustrated in Figure 3.3. As shown in the figure more than 44% of genes are shared by only two pathways, whereas, less than 1% of them are shared by 15 pathways. The higher the number of associated pathways the lower the percentage of these genes will be.

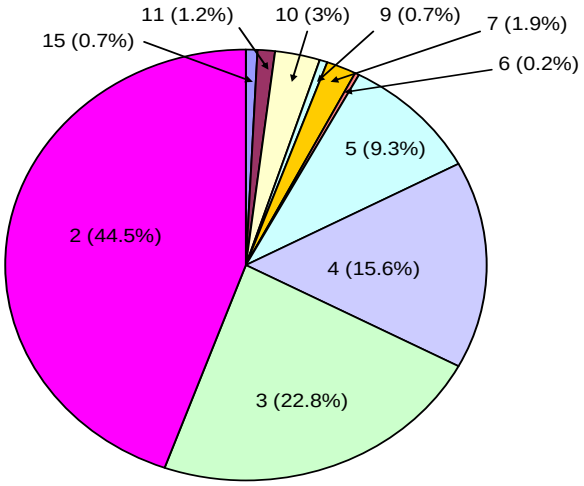

Figure 3.3. Distribution of 429 key enzymes across pathways. The numeral in each sector shows the number of pathways that each member of this group participates in, and inside the parenthesis is percentage of each group out of 429 key enzymes. For example, those key enzymes that participate in 2 pathways are 44.5% of the whole key enzymes. This figure for enzymes found in 15 pathways is 0.7%.

In the first set of units, nodes with the highest connection degrees represent pathways with the highest number of associated genes. These pathways involve energy metabolism (including phosphorylation reactions, synthesis of ATP, and breakdown or polymerization of fatty acids), regulation of purine and pyrimidine metabolism as building blocks of all nucleotides, metabolism of sucrose as major transport compound and starch as important storage for carbohydrate residues, degradation of benzoate using coenzymes, biosynthesis of different glycerolipids that are regulated in different stages of age in the nervous system, regulation of synthesis of different phosphate derivatives of myo-inositol functioning as the second messenger for different extra cellular signals and releasing Ca^{2+} from intracellular storage, and metabolism of tryptophan which is an essential amino acid to the immune system.

On the other hand, the second set of vertices contains 785 genes in the reconstructed network. Taking all genes with connection degrees greater than 1 (called key enzymes) into consideration, key enzymes can be used as indicators of functionally related pathways. In order to accomplish a specific biological task several pathways must function co-ordinately. These key enzymes play central roles in modulating such a coordinated work. For example, to replicate DNA molecules, the following six pathways act together as depicted in Figure 3.4. These pathways are: ATP synthesis (cel00193), oxidative phosphorylation (cel00190), TCA cycle (cel00020), pyruvate metabolism (cel00620), purine metabolism (cel00230), and DNA polymerase (cel03030). To anabolically synthesize DNA molecules, both DNA replication pathway (cel03030) and purine metabolism pathway (cel00230) share 11 DNA polymerization enzymes, and the required ATP is partially furnished by pyruvate metabolism pathway (cel00620) providing with 3 pyruvate kinases. To acquire more ATP, these pyruvate kinases convert phosphoenolpyruvate to pyruvate which is subsequently oxidized by dihydrolipoamide dehydrogenase, resulting in the accumulation of acetyl-CoA to fuel the TCA cycle (cel00020). In the cycle, the succinate dehydrogenase complex, consisting of 5 succinate dehydrogenases, utilizes succinate, a downstream product of acetyl-CoA, as a substrate and convert it to fumurate in which the oxidation phosphorylation (cel00190) is involved. Fumurate is further hydrated to malate. In the presence of malate dehydrogenase, malate is converted to oxaloacetate and ubiquinol (QH2), resulting in the production of NADH

to replenish the reservoir of reducing powers. In cases where ATP is over-supplied, pyruvate carboxylase and phosphoenolpyruvate carboxykinase (see Figure 3.4) shared by pyruvate metabolism pathway (cel00620) and TCA cycle (cel00020) convert pyruvate to phosphoenolpyruvate. Or, if there is a short fall of NADH, the lactate dehydrogenase is triggered (see Figure 3.4) to convert pyruvate to lactate along with the production of NADH. The phosphorylation pathway (cel00190) has 44 key enzymes (28 H^+ transporting enzymes and 16 ATPases) that are in common with the ATPase pathway (cel00193), all NADHs produced in the TCA cycle are used to generate an electrochemical gradient of protons across the inner membrane of mitochondrion in the way that, when electrons pass through 28 electron carriers (H^+ transporting enzymes), the electron flow toward final oxidizing agent, O_2, causes a flow of protons from the inner to outer membrane of mitochondrion, creating a gradient of proton concentration. The 16 ATPases catalyze the phosphorylation of ADP to ATP as the protons move back across the membrane. Therefore, additional amounts of ATP are generated to assist the DNA polymerization.

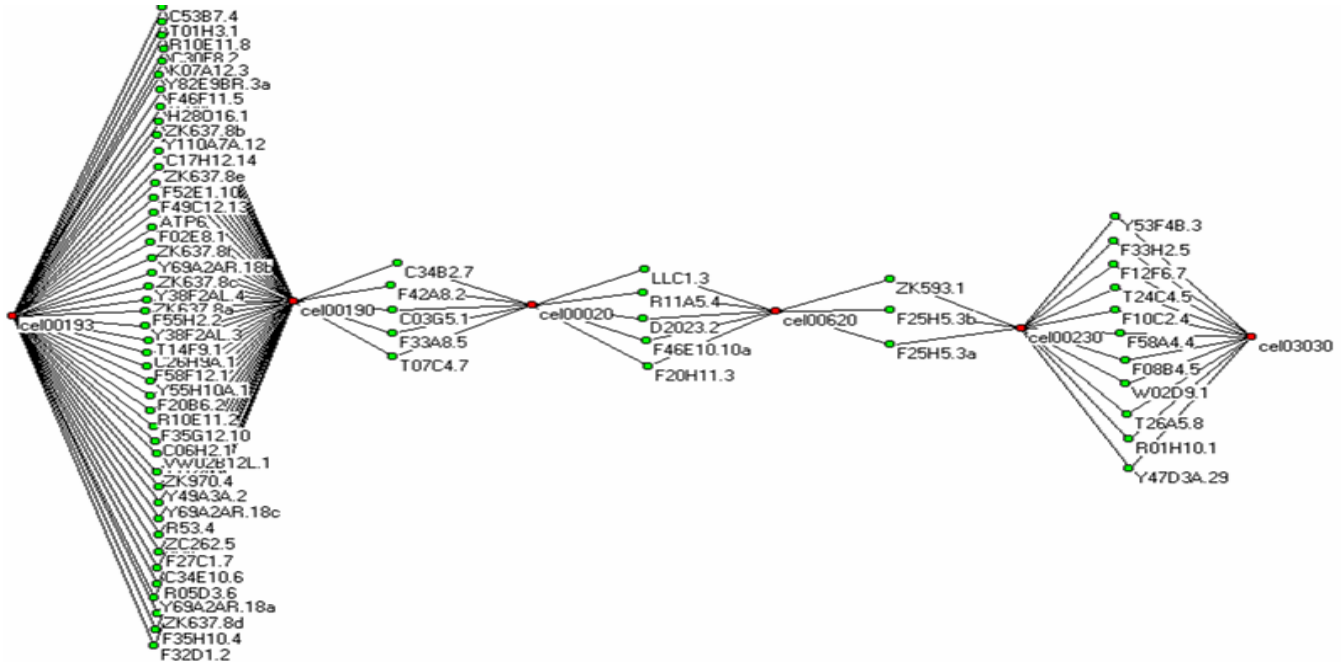

Figure 3.4. Collaboration among 6 pathways in DNA molecule replication path.

Pathway collaboration can be elucidated by locating key enzymes and the pathway connectivity can be depicted by metabolic path length. Among 429 identified key enzymes (i.e. connection degrees greater that 1) and 94 pathways, the longest shortest metabolic path length is 5. As illustrated in Figure 3.5, there were 6 pathways involved in the DNA synthesis process. Depending on extraneous conditions or the availability of starting materials, more than one route can be taken by *C. elegans* to replicate its DNA molecules. Figure 3.5 illustrates all other possible shortest paths and the number of common enzymes between each pair of pathways. The straight path was described in the previous paragraph and it is believed that this path is taken up by the organism when endogenous sources of nutrient (starch and glycogen) are available. In cases where extracellular nutrient is available, the energy metabolism path passes through glycolysis pathway (cel00010) (see Figure 3.5), because macromolecules are broken down to glucose and it is converted to pyruvate through glycolysis pathway. In the starvation cases amino acids can be used as source of energy, alanine and aspartate are the best amino acids for this purpose. Thus, the energy metabolism path passes through alanine and aspartate metabolism pathway (cel00252). In cases where pyruvate is converted to phosphoenolpyruvate to store energy, CO_2 released from conversion of pyruvate to acetylCoA can be used to synthesize oxaloacetate and then malate from phosphoenolpyruvate catalyzed by malate dehydrogenase through carbon fixation pathway (cel00710). Malate is then converted to pyruvate by pyruvate kinases. These alternate paths allow *C. elegans* to detour as one or some of key enzymes in one route being inactivated owing to extraneous variations. Such naturally built-in features greatly enhance the survival of a species. The line (edge) values shown in Figure 3.5 express the number of key enzymes required between two consecutive pathways. The whole number of key enzymes between initial and terminal pathways in a path is represented by path values, defined as the sum of line (edge) values. In the exemplified path (i.e. the straight route in Figure 3.5), there are 68 enzymes function in this path, while some of them may be turned off under normal condition and be activated only under stress. The higher the value of a metabolic path, the more chance for the organism will survive under harsh conditions. Figure 3.5 is also an indication that to proceed a biological process there is more than one combination of pathways that a regulatory system can be chosen from.

One possible approach to alter the enzyme activation or pathway coordination is through the manipulation of extraneous cultivation environment.

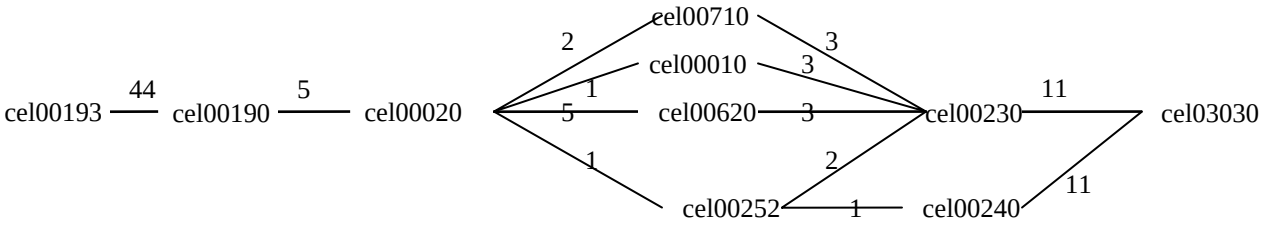

Figure 3.5. All possible shortest paths among two typical pathways. The two pathways cel00193 and cel03030 in the reconstructed *C. elegans* metabolic network are connected through intermediate pathways.

3.5 Conclusion

The reconstructed metabolic network of *C. elegans* provides an insight into the current situation of known proteins within the genome. A functionally more meaningful metabolic network was reconstructed in conjunction with those functionally-assigned genes and was represented by an undirected two-mode graph to investigate its topological property. In this network each protein was connected to 42 other proteins by average and some proteins had partners in 15 different pathways. Protein relationships outside pathway boundaries contributed to the interconnectivity of the network which elucidated some hidden routes to synthesize essential metabolites at different organism's living condition. Analysis of the network showed that how reactions and enzymes at different pathways were working together to accomplish a biological task. Currently, this reconstructed network consists of approximately 6% of all genes in *C. elegans* genome while this network covers gene that are solely involved in metabolism.

References

Bairoch A., Boeckman B. (1992) The SwissProt protein sequence data bank. *Nucleic Acids Research*, **29,** 2019-2022.

Becker S., Palsson B. (2005) Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiology*, **5:**8.

Bork P., Jensen L.J., von Mering C., Ramani A.K., Lee I., Marcotte E.M. (2004) Protein interaction networks from yeast to human. *Current Opinion in Structural Biology*, **14**, 292-299.

Christensen B., Nielson J. (2000) Metabolic network analysis of *Penicillium chrysogenum* using ^{13}C-labled glucose. *Biotechnology and Engineering*, **68(6),** 652-659.

Covert M.W., Schilling C.H., Famili I., Edwards J.S., Goryanin I.I., Selkov E., Palsson B.O. (2001) Metabolic modeling of microbial strain *in silico*. *TRENDS in Biochemical Science*, **26(3)**, 179-186.

Deeds E.J., Ashenberg O., Shakhnovich E.I. (2006) A simple physical model for scaling in protein-protein interaction networks. *PNAS*, **103(2):** 311-316.

Edwards J.S., Palsson B.O. (1999) Systems properties of the *Haemophilus infulenzae* Rd metabolic genotype. *The Journal of Biological Chemistry*, **274,** 17410-17416.

Eisenberg D., Marcotte E.M., Xenarios I., Yeates T.O. (2000) Protein function in the post genomic era. *Nature*, **405,** 823-826.

Enright A.J., Iliopoulos I., Kyrpides N.C., Ouzounis C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402,** 86-90.

Fairlamb A.H. (2002) Metabolic pathways analysis in trypanosomes and malaria parasites. *Phil. Trans. R. Soc. Lond.* **B357,** 101-107.

Famili I., Palsson B.O. (2003) Systemic metabolic reactions are obtained by singular value decomposition of genome-scale stoichiometric matrices. *Journal of Theoretical Biology*, **224,** 87-96.

Forster J., Famili I., Fu P., Palsson B.O., Nielsen J. (2003) Genome-scale reconstruction of *Saccharomyces cerevisiae* metabolic network. *Genome Research*, **13,** 244-253.

Francke C., Siezen R.J., Teusink B. (2005) Reconstructing the metabolic network of a bacterium from genome. *TRENDS in Microbiology,* **13(11):**550-558.

Hekimi S., Lakowski B., Barnes T.M., Ewbank J.J. (1998) Molecular genetics of life span in *C. elegans*: how much does it teach us? *TRENDS in Genetics,* **14(1),** 14-20.

Hoffman R., Valencia A. (2003) Protein interaction: same network, different hubs. *TRENDS in Genetics,* **19(12):**681-683.

Horton H.R., Moran L.A., Ochs R.S., Rawn D.J., Scrimgeour K.G. (2002) Principles of Biochemistry. 3rd Ed. *Prenice-Hall,* N.J., USA

Hazbun T.R., Fields S. (2001) Networking proteins in yeast. *Proc. Natl. Acad. Sci.* USA, **98,** 4277-4278.

Kanehisa M., Goto S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids research,* **28,** 27-30.

Ma H., Zeng A.P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics,* **19,** 270-277.

Marcotte E.M. (2003) Assembling a jigsaw puzzle with 20,000 parts. *Genome Biology*, **4:**323.

Mawuenyega K.G., Kaji H., Yamauchi Y., Shinkawa T., Saito H., Taoka M., Takahashi N., Isobe T. (2003) Large-scale identification of *Caenorhabditis elegans* proteins by multidimensional liquid chromatography- tandem mass spectrometry. *Journal of Proteome Research,* **2,** 23-25.

Miyake S., Takenaka Y., Matsuda H. (2004) A graph analysis method to detect metabolic sub-networks based on phylogenetic profile. *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference,* August 16-19 Stanford, CA, pp. 634-635.

Nakao M., Bono H., Kawashima S., Kamiya T., Sato K., Goto S., Kanehisa M. (1999) Genome-scale gene expression analysis and pathway reconstruction in KEGG. *Genome Informatics,* **10,** 94-103.

Notebaart R.A., van Enckevort F.H.J., Francke C., Siezen R.J., Teusink B. (2006) Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics,* **7:**296.

Romero P., Wagg J., Green M.L., Kaiser D., Krummenacker M., Karp P.D. (2004) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology,* **6:**R2.

Stein L.D. (1999) Internet access to the *C. elegans* genome. *TRENDS in Genetics,* **15(10)**, 425-427.

Stein L.D., Sternberg P., Durbin R., Thierry-Mieg J., Spieth J. (2001) Wormbase: network access to the genome and biology of *Caenorhabditic elegans*. *Nucleic Acids Research*, **29,** 82-86.

Sun J., Zeng A.P. (2004) IdentiCS- Identification of coding sequence and *in silico* reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence. *BMC Bioinformatics,* **5:**112.

Szilagyi A., Grimm V., Arakaki A.K., Skolnick J. (2005) Prediction of physical protein-protein interactions. *Phys. Biol.,* **2,** S1-S16.

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science, **282**, 2012-2018.

Vo T.D., Greenberg H.J., Palsson B.O. (2004) Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *The Journal of Biological Chemistry*, **279(38):**39532-39540.

Walhout A.J.M., Reboul J., Shtanko O., Bertin N., Vaglio P., Ge H., Lee H., Doucette-Stamm L., Gunsalus K.C., Schetter A.J., Morton D.G., Kemphues K.J., Reinke V., Kim S.K., Piano F., Vidal M. (2002) Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Current Biology*, **12,** 1952-1958.

4

PREDICTION OF PROTEIN-PROTEIN INTERACTIONS USING SIGNATURE PROFILING

A similar version of this chapter has been submitted to *Genomics, Proteomics, and bioinformatics*:

Mahmood A. Mahdavi and Yen-Han Lin: Prediction of protein-protein interactions using protein signature profiling. 2007.

Contribution of this chapter to the overall study

As protein-protein interaction information is the building block of reconstructing metabolic networks, the protein-protein interaction prediction methods are emerging. In this chapter a new method was developed to predict more comprehensive protein interactions to be incorporated into the reconstructed metabolic network in Chapter 1.

4.1 Abstract

Protein domains are conserved and functionally independent structures that play an important role in interactions among related proteins. Domain-domain interactions were recently used to predict protein-protein interactions (PPI). In general, the interaction probability of a pair of domains was scored using a trained scoring function. Satisfying a threshold, the protein pairs carrying those domains were regarded as "interacting". Based on the signature content of known proteins, a new approach to directly predict protein interactions without the requirement of training sets was developed. The signature contents of proteins were utilized to predict PPI pairs in *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Homo sapiens.* Similarity between protein signature patterns was scored and PPI predictions were drawn based on the binary similarity scoring function. Results showed that the true positive rate of prediction by means of the proposed approach was approximately 32% higher than that using the maximum likelihood estimation (MLE) method, resulting in a 22% increase in the area under the receiving operator characteristic curve. When proteins containing one and two signature contents were removed, the sensitivity of the predicted PPI pairs increased significantly.

The predicted PPI pairs were on average 11 times more likely to interact than the random selection at a confidence level of 0.95, and on average 4 times better than that in both phylogenetic profiling and gene expression profiling methods. The proposed approach enhances the knowledge of protein association and also aids in augmenting the reconstruction of metabolic networks.

4.2 Introduction

Domain-based interaction prediction techniques rely on statistically significant related domains. When the interaction probability score between two domains (in two different proteins) is greater than a threshold value, such a relationship is extended to the corresponding proteins and the potential interaction is inferred. Close assessment of the protein pairs whose domains possess high interaction probability scores shows that many of these protein pairs share at least one common domain. Sprinzak and Margalit (2001) reported 40 overrepresented domains pairs in protein interaction dataset of yeast. Nearly half of those domain pairs (22 of 40 pairs) contained similar domains and the rest of them were functionally close domains. Non-identical pairs could not pass the threshold, even though the threshold was considered very loose. Okada *et al.* (2005) studied the role of common domains in the extraction of accurate functional associations in interacting partners. It has been shown that, when two proteins share a similar domain structure their interaction confidence score is higher than that of two proteins with non-similar domains (Ng *et al.*, 2003). Common domains are conserved structures and may relate to evolutionary traits of species (Littler, and Hubbard, 2005). When two proteins share common domains, the co-evolution of these domains would provide strong evidence that they are biologically related and the probability of interaction between associated proteins is higher (Ramani and Marcotte, 2003).

Discovery of new patterns in the structure of proteins play a central role in detecting novel interactions. This discovery may happen either through mining literature and published studies (Hao *et al.*, 2005) or comparative analysis of certain group of domains with known functions (Hesselberth *et al.*, 2006). With the combination of protein interaction data form different species and gene ontology a set of high-confidence domain-domain interactions were constructed that was used to predict protein-protein interactions in other organisms (Lee *et al.*, 2006). Another attempt to detect conserved

sub-structures in proteins relies on identifying potentially missing interactions in the dataset which are found in yeast two-hybrid datasets. These missing interactions are predicted based on the relationships of complementary binding domains which are built upon a mathematical model (Morrison *et al.,* 2006).

In this chapter we propose a new genome-wide approach to predict protein-protein interactions based on the observation that proteins with common signatures are more likely to interact. The signature content of a protein is represented by a binary profile, called signature profile, and then the similarity between two profiles is scored based on a binary similarity function. Imposing a threshold, the two proteins are considered 'interacting' if they satisfy the threshold. Despite conventional signature-based methods which score the relationship between two signatures and extrapolate such a relationship to predict protein-protein interactions, our approach directly scores protein relationships based on the signature content of each individual protein and the extent of commonality in signature patterns. The more signatures in common, the higher the similarity score will be between two different profiles. This approach is applied to three organisms including *Saccharomyces cerevisiae, Caenorhabditis elegans,* and *Homo sapiens.* Predicted interactions are compared with signature-based MLE approach (Deng *et al.,* 2002) over a test dataset and two other non signature-based prediction techniques including phylogenetic profiles (Pellegrini *et al.,* 1999), and gene expression profiles (van Noort *et al.,* 2003). Although at the time being a small portion of genes in each genome has been identified with their signatures, the approach is capable of covering the entire genome as more genes with known signature contents are discovered.

4.3 Methods

4.3.1 Signature content information

The signature content of each protein sequence is obtained from PROSITE database (Hulo *et al.,* 2006). PROSITE is a database of protein families and domains, consisting of biologically significant sites, patterns, motifs, and domains. The entire PROSITE database was downloaded and three files were created for three organisms of interest. Each file contains the signatures found in one genome. Currently, PROSITE (release

19.27, May 2006) contained 884 signatures in *S. cerevisiae*, 738 signatures in *C. elegans*, and 1354 signatures in *H. sapiens*.

4.3.2 Experimental protein-protein interaction datasets

To evaluate and compare the predicted protein-protein interactions of our proposed approach, datasets containing experimentally obtained pairs were compiled to serve as a common reference. The dataset for yeast contains 3745 pairs that were obtained from three sources. von Mering *et al.* (2002) introduced yeast protein pairs with high confidence. Pairs confirmed by at least two experimental methods were picked from this source (1920 pairs). BIND database (Alfarano *et al.,* 2005) contains yeast protein pairs that are experimentally confirmed and manually curated (10618 pairs); and CYGD (Guldener *et al.,* 2005) contains yeast protein pairs, confirmed by experiment (10472 pairs). Combination of these three sources resulted in 16507 pairs, which consists of 4391 proteins. Those proteins that are not included in PROSITE were eliminated. As a result, 3745 pairs remained in the final dataset including 1438 proteins.

Worm dataset was constructed from BIND and Li *et al.* (2004). They reported 4960 and 6629 protein pairs, respectively. These pairs were obtained by means of yeast two-hybrid technique and manually curated. After removing repeated pairs the dataset consists of 7081 pairs, comprising 3390 proteins in *C. elegans*. Those proteins that are not included in PROSITE were dropped off resulting in 344 pairs remained in the worm dataset including 220 proteins.

Human dataset is a combination of BIND and HPRD (Peri *et al.*, 2003), containing 2332 and 23187 interactions, respectively. These pairs were obtained using either mass spectrometry or yeast two-hybrid techniques and were manually curated. Merging these two sources of interaction data, a dataset of 25000 interactions, consisting of 5726 proteins, was resulted. Only 13319 pairs contain 3975 proteins that are included in PROSITE. The experimental datasets are presented in Supplementary Data (Chapter 4).

4.3.3 Computational datasets

Phylogenetic profiles: The numbers of proteins studied in three organisms are: m=2242 in *S. cerevisiae*, m=1402 in *C. elegans*, m=8667 in *H. sapiens*. The proteins of each organism were considered as queries and aligned against a database comprising 90

genomes using BLAST program. The list of reference genomes is included in Supplementary Data (Chapter 4). Genomes were obtained from www.ncbi.nlm.nih.gov. Running BLAST program, using SEG filter over 75% similarity of the sequences, the output was a list of homolog proteins and their e-values within each genome that better match the query sequence. The best hit in each genome was taken as one bit in the profile and then profiles were created for each individual protein. These profiles should be converted into binary profiles in the form of 1 and 0 to represent the presence or absence of an individual protein in other genomes. To convert e-values to binary numbers we needed to know if the alignment score for each protein sequence P_i was statistically significant. Statistical significance of an alignment was described by the probability of finding a higher score when two sequences are compared based on a random selection. This probability depends on the number of comparisons that we are making. If the number of proteins encoded in query genome is m and the number of encoded proteins in 90 reference genomes is p the total number of comparisons is: $m \times p$. Therefore, the probability of finding a match for an individual protein sequence is $1/(m \times p)$. In this study p=370461 and m for each organism is given above. We considered this probability as a threshold based on which e-values can be translated to present or absent status. Once the binary profiles were established, they were compared to find interacting proteins. Matching profiles were considered 'interacting'.

Gene expression profiles: Genes with similar co-expression patterns are likely to interact. To find out which genes are co-expressed, the expression levels of the studied genes were extracted from normalized DNA microarray data files obtained from Stanford Microarray Database (SMD) (Ball *et al.*, 2005). Each file corresponds to an experiment. All expression values were collected in a gene expression matrix in which each row represents a different gene and each column corresponds to a different microarray experiment (100 experiments in *S. cerevisiae*, 575 experiments in *C. elegans*, and 400 experiments in *H. sapiens*). The matrix is supplied into EXPANDER program (Shamir *et al.*, 2005) for clustering. Choosing click algorithm to cluster genes the following results were obtained for each organism:

Table 4.1. The characteristics of EXPANDER output clusters

organism	**Number of clusters**	**Overall homogeneity**
S. cerevisiae	6	0.552
C. elegans	10	0.631
H. sapiens	93	0.562

Genes in the same cluster are co-expressed genes in different biological conditions. These genes were paired and considered 'interacting'.

Maximum likelihood estimation: In order to implement maximum likelihood estimation (MLE) method the compiled experimental data was randomly split into two parts including training set and test set. The training set, serving as observed interactions, was used for recursive calculations. The underlying hypothesis in this method is two proteins interact if and only if at least one pair of domains from the two proteins interact. Let D_1, D_2,….,D_M denote the M domains, and P_1, P_2,….P_N denote N proteins. P_{ij} denotes the protein pair of P_i and P_j, and D_{ij} denotes the domain pair of D_i and D_j. Treating protein-protein interactions, and domain-domain interactions as random variables, the probability of interacting two proteins under stated assumption is:

$$\Pr(P_{ij} = 1) = 1.0 - \prod_{D_{mn} \in P_{ij}} (1 - \lambda_{mn}) \tag{4.1}$$

where λ_{mn}= Pr(D_{mn}=1) denotes the probability that domain D_m interacts with domain D_n. False positive rate (*fp*) and false negative rate (*fn*) are defined based on observed interactions. Let O_{ij} be the variable for the observed interaction result for proteins P_i and P_j. O_{ij} =1 if the interaction is observed and O_{ij}=0 otherwise. Then,

$$\begin{aligned} fn &= \Pr(O_{ij} = 0 \mid P_{ij} = 1) = 1.0 - \frac{\Pr(O_{ij} = 1, P_{ij} = 1)}{\Pr(P_{ij} = 1)} \geq 1.0 - \frac{\Pr(O_{ij} = 1)}{\Pr(P_{ij} = 1)} \\ &= 1.0 - \frac{number of\ observed\ pairs}{number of\ real\ interacting\ pairs} \end{aligned} \tag{4.2}$$

$$\begin{aligned} fp &= \Pr(O_{ij} = 1 \mid P_{ij} = 0) = \frac{\Pr(O_{ij} = 1, P_{ij} = 0)}{\Pr(P_{ij} = 0)} \leq \frac{\Pr(O_{ij} = 1)}{\Pr(P_{ij} = 0)} \\ &= \frac{number\ of\ observed\ pairs}{total\ potential\ pairs - number\ of\ real\ interacting\ pairs} \end{aligned} \tag{4.3}$$

Thus, the probability of observing a protein-protein interaction is:

$$\Pr(O_{ij} = 1) = \Pr(P_{ij} = 1)(1 - fn) + (1 - \Pr(P_{ij} = 1))\, fp \tag{4.4}$$

The probability of the observed whole genome interaction dataset is

$$L = \prod (\Pr(O_{ij} = 1))^{O_{ij}} (1 - \Pr(O_{ij} = 1))^{1-O_{ij}} \tag{4.5}$$

where O_{ij}=1 if the interaction of P_i and P_j is observed and O_{ij}=0 otherwise. L is the likelihood and is a function of λ_{mn}, *fp*, and *fn*. In this calculation *fn* and *fp* are determined based on Equations 4.3 and 4.4 as 0.84 and 7.5E-4 for yeast, respectively. The number of observed interactions (training set) is given as 1873 pairs. It is reported that in yeast proteome each protein interacts with approximately 5 proteins (Hazbun and Fields, 2001). For 2242 yeast proteins in this study, it gives the number of real interactions of 11210 pairs. The total number of potential pairs is *m(m-1)/2* where in this study *m* is 2242 proteins for yeast. Then, we compute λ_{mn} using a recursive formula. First, initial values for λ_{mn} are chosen. Then $\Pr(P_{ij}=1)$ and $\Pr(O_{ij}=1)$ are computed by equations (4.1) and (4.4), respectively. Parameter λ_{mn} is updated using the following equation:

$$\lambda_{mn}^{(t)} = \frac{\lambda_{mn}^{(t-1)}}{N_{mn}} \sum_{i \in A_m, j \in A_n} \frac{(1 - fn)^{O_{ij}}\, fn^{1-O_{ij}}}{\Pr(O_{ij} = o_{ij} \mid \lambda^{(t-1)})} \tag{4.6}$$

and likelihood function is computed by Equation (4.5). Calculations continue until the value of likelihood function is unchanged within a certain error.

4.3.4 Signature content representation

A protein is characterized by the signatures existing in its sequence. Hence, each protein can be represented by a vector of *n* features, called a signature profile, where each feature corresponds to a signature and *n* is the number of signatures identified in the proteome of an organism (for example n = 885 in yeast). Let $P_i = [S_{i1}, S_{i2}, ..., S_{in}]$ represent the feature vector of protein P_i with n signatures. $S_{i1} = 1$ if signature S_1 exists in protein P_i and $S_{i1} = 0$ otherwise. Therefore, each genome is represented by a *m*-dimensional vector where *m* is the number of proteins with known signatures. In this study, m = 2242 in yeast, m = 1402 in worm, and m = 8667 in human. A similarity

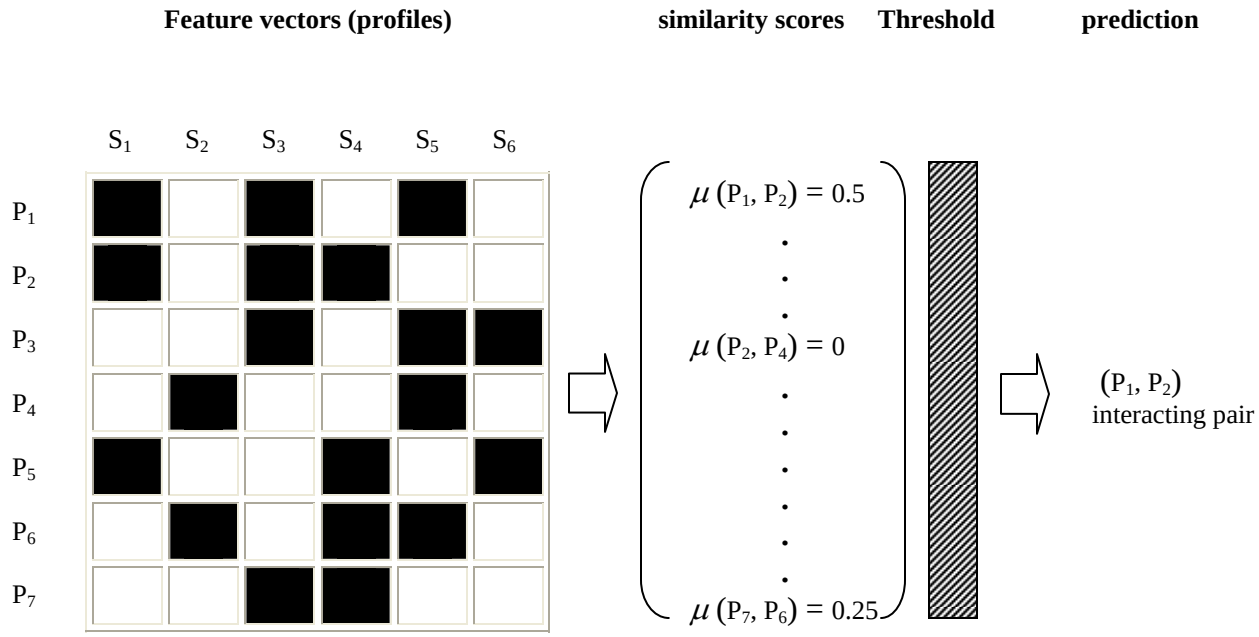

Figure 4.1. Schematic of the proposed method to predict protein interactions. Briefly, signature content of each protein is represented by a feature vector and the whole proteins containing at least one signature are represented by a m-dimensional vector. Proteins are paired in order and the similarity of feature vectors is calculated. Setting a threshold, if the similarity score is equal to or greater than the threshold, the two proteins are considered interacting.

measure was implemented to calculate the similarity between signature profiles (feature vectors). Binary Similarity Function (Rawat *et al.*, 2006) is introduced to measure the similarity between a pair of signature profiles:

$$\mu(P_i, P_j) = \frac{\sum_{l=1}^{n} (P_i \wedge P_j)_l}{\sum_{l=1}^{n} (P_i \vee P_j)_l} \tag{4.7}$$

Where, μ is the similarity score between profiles P_i and P_j. This score is calculated over n signatures contained in proteins of a genome of interest. If protein P_i contains x signatures, protein P_j contains y signatures, and both proteins contain z signatures in common, the score can then be calculated as follows:

$$\mu(P_i, P_j) = \frac{z}{x + y - z} \tag{4.8}$$

Note that $0 \leq \mu \leq 1$. The value of μ increases when there is more common signatures between the two proteins and the value of μ decreases when the number of uncommon signatures is more than common ones in P_i and P_j. If the similarity score is higher than a threshold, the two proteins are considered as an "interacting pair". The inferring procedure of protein-protein interactions is illustrated in Figure 4.1

4.4 Results

The signature profiling approach was applied to predict protein interactions for *S. cerevisiae*, *C. elegans*, and *H. sapiens*. Three different predicted PPI datasets for each organism were generated by removing proteins having none, one, and two known signatures in their sequences. The predicted protein pairs and their corresponding binary similarity values are presented in Supplementary Data (Chapter 4). To evaluate the performance of the approach, sensitivity and specificity analysis was conducted and the predicted results were compared with those obtained by MLE method in *S. cerevisiae* over a test dataset. Predicted dataset using MLE method is presented in Supplementary Data (Chapter 4). Furthermore, the fold value analysis was performed to compare the predicted results with those obtained from two non-signature-based methods including phylogenetic profiles, and gene expression profiles in *S. cerevisiae* and *C. elegans* (see

Supplementary Data, Chapter 4). In either case, the proposed approach has higher true positive rates.

4.4.1 Sensitivity and specificity analysis

The receiving operator characteristic (ROC) curve was implemented to evaluate the efficacy of the prediction of PPI pairs between our approach and the MLE method over the same dataset. The ROC curve portrays the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) for different threshold values. The true positive rate is defined as the proportion of experimentally confirmed PPI pairs (i.e., all positives) that is correctly predicted; whereas, the false positive rate is defined as the proportion of experimentally refuted PPI pairs (i.e., all negatives) that is erroneously predicted. Therefore, true positive rate and false positive rate can be formulated as follows,

$$\text{True positive rate} = \text{Sensitivity} = \frac{TP}{TP + FN} \tag{4.9}$$

$$\text{False positive rate} = 1\text{- Specificity} = \frac{FP}{FP + TN} \tag{4.10}$$

where, “TP” is the number of experimentally confirmed PPI pairs that are predicted by a method (matched), “FN” is the number of experimentally confirmed PPI pairs that are not predicted by a method, “FP” is the number of predicted PPI pairs that do not match experimentally confirmed pairs, and “TN” is the number of potential PPI pairs that are neither experimentally confirmed nor computationally predicted.

The area under the ROC curve (AUC) is a quantitative indicator for comparing the performance of PPI prediction among various PPI predicting methods. At AUC of 1, a perfect PPI prediction is obtained. As shown in Figure 4.2, the AUC of protein signature profiling approach in case of no protein removal is 0.549 and that of MLE is 0.534, indicating that more experimentally confirmed PPI pairs can be predicted by the proposed approach than the MLE method. Such an improvement based on protein signature profiling comes from the fact that the association between two proteins requires at least one signature in common. The requirement for protein signature profiling is more stringent than the interaction of two domains as implemented in the MLE method.

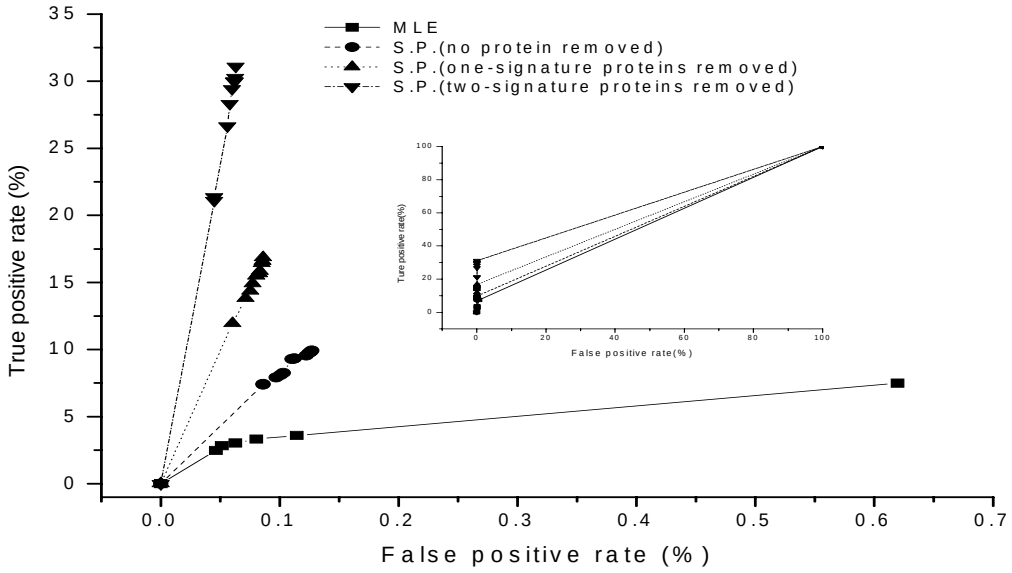

Figure 4.2. Changes of ROC curves subjected to the removal of proteins containing one- and two-signature contents.

Approximately 68% of predicted PPI pairs have the highest similarity score (i.e., 1), indicating a complete matching signature profile between two query proteins. Among this portion of predicted PPI pairs, many of these pairs contain only one or two known protein signatures. As a result, a high false positive rate was obtained as compared to that calculated by the MLE method. The cause of a high false positive rate is attributed to the low number of known signature contents in these proteins. To reduce false positive rates of predicted PPI pairs, and thus increase the accuracy of PPI prediction, proteins with one and two signature contents were removed consecutively, and the proposed approach was then applied to the remaining proteins in the dataset. As illustrated in Figure 4.2, the increase in the AUC of ROC curve was observed for both cases (see inset of Figure 4.2). The AUC increased to 0.584 when proteins with one known signature content were removed and eventually increased to 0.655 when proteins with two known signature contents were also deleted from the dataset.

It is expected that with the availability of more information on signature content of proteins, the true positive rate of the proposed approach will drastically increase along with a low false positive rate. Nevertheless, the examination of the ROC curve indicates

that protein signature profiling approach presents a competitive, or even better, result compared to other currently available domain-based methods such as MLE when applied over the same dataset.

4.4.2 Fold value analysis

The PPI pairs predicted by means of the proposed protein signature profiling approach were also compared to two other non-signature-based methods: phylogenetic profiling and gene expression profiling. Based on genomics information, phylogenetic profiling method has been reported as one of the most promising computational methods to predict PPI pairs (Marcotte *et al.,* 1999); whereas gene expression profiling method utilizes conserved co-expression patterns of genes to predict interacting protein pairs (Fraser *et al.,* 2004). To examine the efficacy of the proposed protein signature profiling approach, methods of phylogenetic profiling and gene co-expression profiling along with the proposed approach were compared against the same reference datasets.

To construct phylogenetic profiles among proteins, query proteins were blasted against reference genome database consisting of 90 species (see Methods). The co-expression patterns were constructed based on normalized DNA microarray data confirmed from Stanford Microarray Database (see Methods).

Results from above-mentioned methods applied to three model organisms were compiled in Table 4.2. As seen in the table, the signature profiling approach predicts less interacting pairs, with relatively more matched pairs with observed datasets. To quantify the statistical significance of the predicted PPI pairs among three profiling methods, a statistical parameter, called fold, was used to facilitate the comparison (Deng *et al.*, 2002). Fold is the ratio of the fraction of the predicted PPI pairs matched with experimentally confirmed dataset, to the fraction of predicted PPI pairs:

$$Fold = \frac{k_0 / K}{n / M} \tag{4.11}$$

Where k_0 is the number of matched predicted PPI pairs found in the experimentally confirmed dataset, K is the size of the experimentally confirmed dataset, n is the predicted PPI pairs satisfied a threshold value, and M is the total number of possible PPI pairs; i.e., m(m-1)/2. The m value for *S. cerevisiae*, *C. elegans*, and *H. sapiens* is 2242,

1402, and 8667, respectively. Fold is the probability of true interaction in predicted PPI pairs compared to the random prediction. The greater the fold, the higher the probability of interaction will be compared to the random pairing.

Figure 4.3 illustrates changes in fold values among protein signature profiling, phylogenetic profiling, and gene expression profiling methods applied to *S. cerevisiae*, *C. elegans*, and *H. sapiens*. Generally speaking, the proposed approach can predict more PPI pairs (at a confidence level of 0.95) than two other non-signature-based methods. As one or two protein signature contents were removed, the fold values of PPI pairs predicted by the protein signature profiling increased significantly as compared to phylogenetic profiling and gene expression profiling methods. This suggests that as proteins possessing more protein signature contents were deleted from the predicted PPI pairs, the probability of remaining predicted pairs being considered as false positive pairs would reduce noticeably. As a result, more PPI pairs with a high confidence level can be predicted.

4.5 Discussion

In this chapter, we propose that the similarity of protein signature patterns could be used to predict interaction between two proteins. Different from other domain-based approaches such as MLE method that utilizes a part of experimental PPI pairs as a learning dataset to train a scoring function in order to calculate the interaction probability, the proposed approach does not require any learning set. In fact, the entire data can be used as a query dataset. The protein signature profiling approach predicts interactions upon the extent of similarity between the signature contents of the two proteins; while domain-based methods predict interactions between protein domains and assume that two proteins will interact, if at least one pair of domains from the two proteins interact.

The significant threshold values are associated with the confidence level and the size of predicted PPI pairs. The significant threshold value in each confidence level is calculated by *(-0.1)log(P)*. P, an absolute probability, is defined as the ratio of confidence level (= 1 – significance level) over the size of predicted PPI pairs, and "0.1" is the scaling factor that scales the threshold value to its corresponding binary similarity score between 0 and 1. Figure 4.4 portrays a significant threshold value with respect to each respective confidence level for three investigated organisms. For instance, at a confidence level of 0.95 (i.e., a 1 in 20 chances of being false positive), the significant threshold

value of choosing a binary similarity score for *S. cerevisiae*, *C. elegans*, and *H. sapiens*, is 0.56, 0.53, and 0.72 respectively. At these threshold values, the predicted PPI pairs will possess a significance level of 0.05. In other words, there is a 95% probability that the predicted PPI pairs are not resulting from random events.

At a confidence level of 0.975, the corresponding significant threshold value is 0.6 for *S. cerevisiae*. From Figure 4.2, the true positive rate for the case of two-signature proteins removed, one-signature proteins removed, and no proteins removed under signature profiling approach (see legend shown in the figure) is 28.33, 14.92, and 8.25 respectively; whereas, the true positive rate for the MLE method at the same confidence level is 3.03. This indicates that the proposed approach is more sensitive than the MLE method, and the sensitivity of the approach can be manipulated by means of deleting proteins containing less signature content. As a result, more experimentally confirmed PPI pairs are predicted.

Table 4.2. Comparison of signature profiling results with/without protein removal with two other non signature-based methods against three common reference datasets.

Method	***predicted***	***observed***	***matched***	***predicted***	***observed***	***matched***	***predicted***	***observed***	***matched***
	S. cerevisiae			*C. elegans*			*H. sapiens*		
Signature profiling [a]	22176	3745	372	10147	344	27	720549	13319	3314
Signature profiling[b]	14968	1079	182	7594	79	6	602234	5441	1890
Signature profiling[c]	10916	360	112	3838	17	3	226484	2223	1069
Phylogenetic profiling	59435	3745	292	51666	344	35	3419797	13319	2921
Gene expression	575258	3745	1942	115047	344	81	606367	13319	964

a: signature profiling method with no protein removal from the dataset.

b: signature profiling method with removing proteins containing one known signature.

c: signature profiling method with removing proteins containing two or less known signatures.

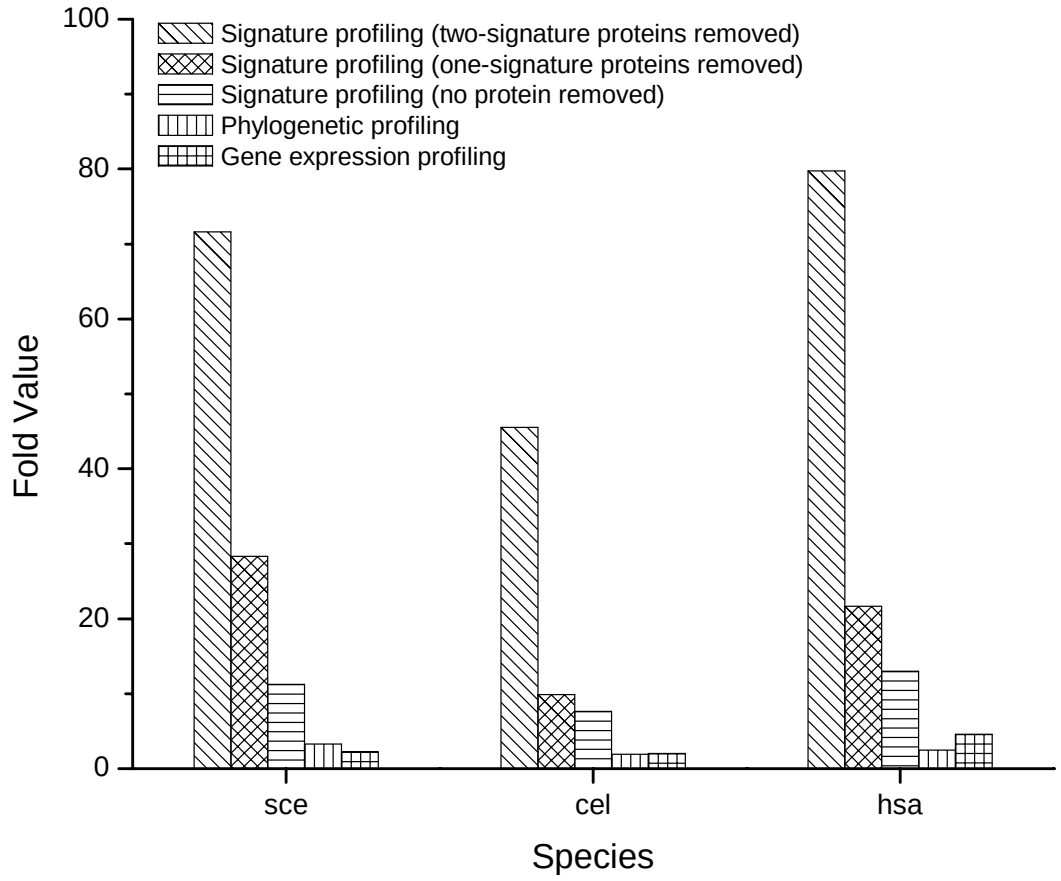

Figure 4.3. Comparison of changes of fold value among three different PPI prediction methods. Each method is applied to *Saccharomyces cerevisiae* (sce), *Caenorhabditis elegans* (cel), and *Homo sapiens* (hsa).

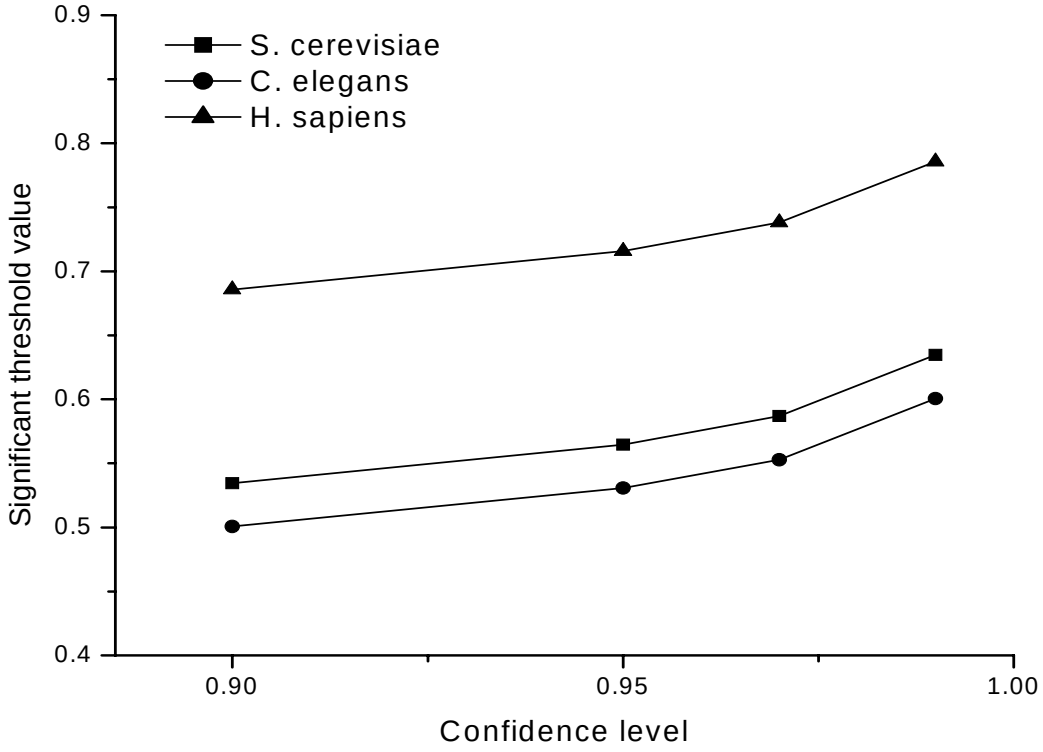

Figure 4.4. The relationship between a confidence level and the significant threshold value. Significant threshold values are correspondent to binary similarity scores.

Other than depicting the absolute relationship of fold value variations among different PPI profiling methods, Figure 4.5 presents the effect of removing proteins with different signature contents on the relative changes of fold values. As seen in the figure, by removing proteins with two signature contents from the predicted PPI pairs, the relative fold change of protein signature profiling versus phylogenetic profiling is 22.03, 23.60 and 32.41 for *S. cerevisiae*, *C. elegans*, and *H. sapiens*, respectively; whereas, the relative fold change of protein signature profiling versus gene expression profiling is 32.11, 22.66 and 17.45 for *S. cerevisiae*, *C. elegans*, and *H. sapiens*, respectively. Nevertheless at the case of no protein removal, the PPI pairs (at a confidence level of 0.95) predicted by the proposed approach is still out-performing the two non-signature profiling methods.

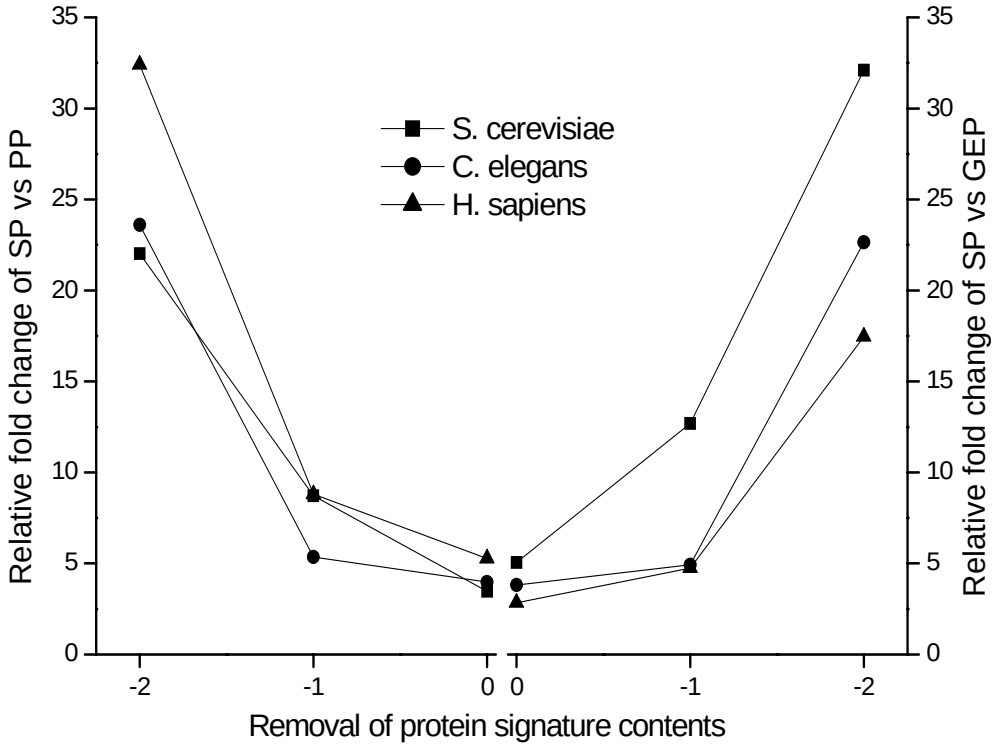

Figure 4.5. The effect of removing proteins with low number of signatures on the relative fold change. SP, protein signature profiling; PP, phylogenetic profiling; GEP, gene expression profiling; "-2", proteins containing two signature contents; "-1", proteins containing one-signature contents; "0", no removal.

New putative protein-protein interactions can be emerged from our results. In case of yeast, the experimental dataset contains 1438 proteins, while our analysis is focused on 2242 proteins whose signature contents are available. Interactions involved with other 804 (= 2242 – 1438) proteins may point out a direction for further experimental validation. For example, YBR208C and YGL062W are found interacting using our approach and these two proteins are not reported in the experimental dataset. Note that YBR208C contains seven domains six of them are shared by YGL062W. Both proteins function as carboxylases. One may postulate that a potential interaction between YBR208C and YGL062W. Such a clue may be used to guide a follow-up experiment.

Protein signature-based methods including our approach embed more intuitive biological reflection than others such as phylogenetic profiling method. Upon the notion that proteins interact through their conserved interfaces, not the whole sequence, the phylogenetic profiling method may not be able to identify true interacting partners. It relies on identifying orthologs of a query sequence in a set of genomes based on whole sequence alignment. Instead, protein signature profiling identifies interacting partners based solely on the pattern of functional interfaces, which are involved in protein interactions. The gene expression profiling method provides information on co-expression of genes in different biological events. Although this information is a strong indication that genes with similar expression profiles may have functional relationships, it provides a relatively lower degree of contribution to the prediction of physical interactions.

4.6 Conclusion

Proteins interact with each other through their functionally independent, structurally conserved, and biologically related signatures. These properties established new insight into the prediction of protein-protein interactions. Many existing domain-based prediction methods calculated the interaction probability score between two signatures. The scoring function was trained based on a learning dataset and subsequently applied to predict protein interactions. In contrast, the proposed approach did not require training information and proteins were directly paired based on their signature contents, providing that they had at least one signature in common. When proteins with a low number of known signature contents (one and two signatures) were removed from the dataset, it resulted in more predicted PPI pairs at a high confidence level. Thus, with the availability of more and more proteins with known signature contents across organisms, the coverage and accuracy of protein interacting pairs predicted by this approach is expected to increase. The predicted PPI pairs can, for instance, be incorporated into metabolic pathway reconstruction, or be used to reveal existing knowledge gaps in the association of proteins and pathways.

References

Alfarano C., Andrade C.E., Anthony K., Bahroos N., Bajec M., Bantoft K., Betel D., Bobechko B., Boutilier K., Burgess E., Buzadzija K., Cavero R., D'Abreo C., Donaldson I., Dorairajoo D., Dumontier M.J., Dumontier M.R., Earles V., Farrall R., Feldman H., Garderman E., Gong Y., Gonzaga R., Grystan V., Gryz E., Gu V., Haldorsen E., Halupa A., Haw R., Hrvojic A., Hurrell L., Isserlin R., Jack F., Juma F., Khan A., Kon T., Konopinsky S., Le V., Lee E., ling S., Magidin M., Moniakis J., Montojo J., Moore S., Muskat B., Ng I., Paraiso J.P., Parker B., Pintilie G., Pirone R., Salama J.J., Sgro S., Shan T., Shu Y., Siew J., Skinner D., Snyder K., Stasiuk R., Strumpf D., Tuekam B., Tao S., Wang Z., White M., Willis R., Wolting C., Wong S., Wrong A., Xin C., Yao R., Yates B., Zhang S., Zheng K., Pawson T., Ouellette B.F., Hogue C.W., *et al.* (2005) The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res.* **33,** D418-D424.

Ball C.A., Awad I.A.B., Demeter J., Gollub J., Hebert J.M., Hernandez-Boussard T., Jin H., Matese J.C., Nitzberg M., Wymore F., Zachariah Z.K., Brown P.O., Sherlock G. (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Research,* **33,** D580-D582.

Deng M., Mehta S., Sun F., Chen T. (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Research,* **12**, 1540-1548.

Fraser H.B., Hirsh A.E., Wall D.P., Eisen M.B. (2004) Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci.,* **101,** 9033-9038.

Guldener U., Munsterkotter M., Kastenmuller G., Strack N., van Helden J., Lemer C., Richelles J., Wodak S.J., Garcia-Martinez J., Perez-Ortin J.E., Michael H., Kaps A., Talla E., Dujon B., Andre B., Souciet J.L., De Montigny J., Bon E., Gaillardin C., Mewes H.W. (2005) CYGD: the comprehensive yeast genome database. *Nucleic Acids Res.* **33,** D364-D368.

Hao Y., Zhu X., Huang M., Li M. (2005) Discovering patterns to extract protein-protein interactions from the literature: part II. *Bioinformatics*, **21(15)**, 3294-3300.

Hazbun T.R., Fields S. (2001) Networking proteins in yeast. *Proc. Natl. Acad. Sci.* USA, **98,** 4277-4278.

Hesselberth J.R., Miller J.P., Golob A., Stajich J.E., Michaud G.A., Fields S. (2006) Comparative analysis of *Saccharomyces cerevisiae* WW domains and their interacting proteins. *Genome Biology,* **7:**R30.

Hulo N., Bairoch A., Bulliard V., Cerutti L., De Castro E., Langendijk-Genevaux P.S., Pagni M., Sigrist C.J. (2006) The PROSITE database. *Nucleic Acids Res.* **34,** D227-D230.

Lee H., Deng M., Sun F., Chen T. (2006) An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics,* **7:**269.

Li S., Armstrong C.M., Bertin N., Ge H., Milstein S., Boxem M., Vidalain P.O., Hao T., Goldberg D.S., Li N., Martinez M., Rual J.F., Lamesch P., Xu L., Tewari M., Wong S.L., Zhang L.V., Berriz G.F., Jacotot L., Vaglio P., Reboul J., Hirozane-Kishiawa T., Li Q., Gabel H.W., Gabel H.W., Elewa A., Baumgartner B., Rose D.J., Yu H., Bosak S., Sequerra R., Fraser A., Mango S.E., Saxton W.M., Strome S., Van den Heuvel S., Piano F., Vandenhaute J., Sardet C., Gerstein M., Doucette-Stamm L., Gunsalus K.C., Harper J.W., Cusick M.E., Roth F.P., Hill D.E., Vidal M. (2004) A map of interactome network of the metazoan *C. elegans*. *Science,* **303,** 540-543.

Littler S.J., Hubbard S.J. (2005) Conservation of orientation and sequence in protein domain-domain interactions. *J. Mol. Biol.*, **345**, 1265-1279.

Marcotte E.M., Pellegrini M., Thompson M.J., Yeates T.O., Eisenberg D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature,* **402,** 83-86.

Morrison J.L., Breitling R., Higham D.J., Gilbert D.R. (2006) A lock-and-key model for protein-protein interactions. *Bioinformatics,* **22(16):**2012-2019.

Ng S.-K., Zhang Z., Tan S.-H., Lin K. (2003) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res.* **31,** 215-254.

Okada K., Kanaya S., Asai K. (2005) Accurate extraction of functional associations between proteins based on common interaction partners and common domains. *Bioinformatics,* **21,** 2043-2048.

Pellegrini M., Marcotte E.M., Thompson M.J., Eisenberg D., Yeates T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* USA **96**, 4285-4288.

Peri S., Navarro J.D., Amanchy R., Kristianen T.Z., Jonnalagadda C.K., Surendranath V., Niranjan V., Muthusamy B., Gandhi T.K., Gronborg M., Ibarrola N., Deshpande N., Shanker K., Shivashankar H.N., Rashmi B.P., Ramya M.A., Zhao Z., Chandrika K.N., Padma N., Harsha H.C., Yatish A.J., Kavitha M.P., Menezes M., Choudhury D.R., Suresh S., Ghosh N., Saravana R., Chandran S., Krishna S., Joy M., Anand S.K., Madavan V., Joseph A., Wong G.W., Schiemann W.P., Constantinescu S.N., Huang L., Khosravi-far R., Steen H., Tewari M., Ghaffari S., Blobe G.C., dang C.V., Garcia J.G., Pevsner J., Jensen O.N., Roepstorff P., Deshpande K.S., Chinnaiyan A.M., Hamosh A., Chakravarti A., Pandey A. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research,* **13,** 2363-2371.

Ramani A.K., Marcotte E.M. (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.*, **327**, 273-284.

Rawat S., Gulati V.P., Pujari A.K., Vemuri V.R. (2006) Intrusion detection using text processing techniques with a binary-weighted cosine metric. *Journal of Information Assurance and Security,* **1,** 43-50.

Shamir R., Maron-Katz A., Tanay A., Linhart C., Steinfeld I., Sharan R., Shiloh Y., Elkon R. (2005) EXPANDER- an integrative program suite for microarray data analysis. *BMC Bioinformatics,* **6:**232.

Sprinzak E., Margalit H. (2001) Correlated sequence-signatures as markers of protein-protein interactions. *J. Mol. Biol.*, **311**, 681-692.

von Mering C., Krause R., Snel B., Cornell M., Oliver S.G., Fields S., Bork P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature,* **417,** 399-403.

van Noort V., Snel B., Huynen M.A. (2003) Predicting gene functions by conserved co-expression. *TRENDS in Genetics,* **19,** 238-242.

5

FALSE POSITIVE REDUCTION IN PROTEIN-PROTEIN INTERACTION PREDICTIONS USING GENE ONTOLOGY AND ANNOTATION

A similar version of this chapter has been submitted to *BMC Bioinformatics*:

Mahmood A. Mahdavi and Yen-Han Lin: False positive reduction in protein-protein interaction predictions using gene ontology annotations. 2007.

Contribution of this chapter to the overall study

Protein-protein interaction prediction techniques predict true interactions along with numerous false positives. In this chapter a global framework was proposed to reduce the number of false positives in the protein interaction dataset produced in previous chapter. Genomic information incorporated into metabolic networks should be verified to ensure the accuracy, consistency, and reliability of the data.

5.1 Abstract

Gene Ontology (GO) annotations were used to reduce false positive protein-protein interactions (PPI) pairs resulting from computational predictions. Using experimentally obtained PPI pairs as a training dataset, eight top-ranking keywords were extracted from GO molecular function annotations. The sensitivity of these keywords was 64.21% in yeast experimental dataset and 80.83% in that of worm. The specificities, a measure of recovery power, of these keywords applied to four predicted PPI datasets were 48.32% and 46.49% (by average of four datasets) in yeast and worm, respectively. Based on eight top-ranking keywords and co-localization of interacting proteins a set of two knowledge rules were deduced and applied to reduce false positive predicted protein pairs. The 'strength', a measure of improvement provided by the rules, defined based on the signal-to-noise ratio, was implemented to measure the applicability of knowledge rules applying to predicted PPI dataset. Depending on the employed PPI-predicting methods, the strength varied between two and ten-folds with respect to the randomly removing protein

pairs from datasets. Hence, GO annotations along with the deduced knowledge rules could be implemented to partially remove false predicted PPI pairs, resulting in more accurate protein interaction prediction.

5.2 Introduction

In recent years, high throughput technologies, in one hand, have provided experimental tools to identify protein interactions in large scale, generating tremendous amount of protein interaction data (Zhu *et al.,* 2003). On the other hand, computational approaches for protein interaction inference have presented inexpensive growing number of methods to predict vast number of protein pairs on genome scale (Yu and Fotouhi, 2006). However, both experimental techniques and computational approaches are affected by high false positives and false negatives (Mrowka *et al.,* 2001) that tend to poor agreement among bench mark datasets (Bork *et al.,* 2004). In the experimental front, false positive mostly stems from the technology involved. In recent years new analytical techniques have been introduced targeting more accurate screening (Campoy and Freire, 2005). Nonetheless, some techniques have been already proposed to enhance the reliability of current high-throughput screening datasets (Deane *et al.,* 2002). Searching the relationship among orthologous proteins in other organisms is one way to validate a new identified interaction (Patil and Nakamura 2005). When orthology is combined with domain content information of related proteins, the detected interacting pair of proteins is more reliable (Valencia and Pazos, 2002). In a recent work, the quality of experimental interaction datasets were improved by predicting missed protein-protein interactions using the topology of the protein interaction map observed by large-scale experiment (Yu *et al.,* 2006). In the computational front, most efforts have been focused on detecting more protein-protein interactions by means of various techniques which identify true positives along with numerous false positive and false negative predictions. Reduction of computational false positive predictions has not been adequately investigated. Verification of protein interactions based on co-expression of their orthologs is one proposal (Tirosh and Baraki, 2005).

So far, several computational approaches have been proposed to predict protein interactions (Valencia and Pazos, 2002). These approaches can be grouped into six categories based upon the ideas that are originating from as stated in Chapter 1. False

positive prediction in all computational methods is a challenge. Currently overlap among computational approaches is not statistically significant (Bard and Rhee, 2004). Furthermore, because of the lack of solid information on protein-protein interaction, the accuracy of different computational approaches remains uncertain. Nevertheless, it is a common perception that if both experimental results and computational approaches agree on a link, the confidence level of that link would be high. Therefore, one measure to evaluate the false positive content of computational predictions is the level of agreement with experimental findings. Although high-throughput screening techniques are affected by false positives, validation of computational pairs by experimental results is widely acceptable.

To enhance the overlap between computational predictions and experimental results, a common ground upon which the predicted results can be evaluated is required. Gene Ontology (GO) annotations may serve as the common ground, even though annotation is an ongoing process. Gene Ontology (GO) is the database that contains controlled vocabularies to annotate molecular attributes for different model organisms. Annotations are defined in three structured ontologies which allow the description of molecular function (F), biological process (P), and cellular component (C). Each ontology is structured in child-parent hierarchies in which a 'child' may have many 'parents' and child terms are components of parent terms. Thus, information provided by GO should be useful in further assessment of predicted PPIs and may be integrated with global filtering algorithms to reduce the number of false positives in PPI prediction techniques. Currently, several attempts have been reported to construct functional association predictors solely based on GO information. Most annotations are backed up with experimental evidence and are collected in certain databases (Reboul *et al.,* 2003). Annotation transfer was utilized to relate multi-function proteins which may operate in different locations (Hegyi and Gerstein 2001). In some studies, associations between proteins in a pair are assessed in terms of the similar GO terms (Rhodes *et al.,* 2005), while other studies evaluate functional associations based on either information content (Lord *et al.,* 2003) or GO structural hierarchy (Wu *et al.,* 2005). With the combination of GO annotations and global mRNA expression analyses a multi-stage frame-work was introduced to integrate this information, resulting in characterizing more proteins with

more detailed annotations (Jiang and Keating, 2005). In a recent study, GO annotations have been used to construct a PPI network for yeast by measuring similarity between two gene ontology terms with a relative specificity semantic relation (Wu *et al.*, 2006).

Therefore, GO can be utilized as a useful informatics resource to either predict or further analyze the predicted PPI datasets. However, ontology annotation is an incomplete process and suffers from inconsistency within and between genomes. In some cases, two confirmed interacting proteins are assigned with two different GO annotations which are not equivalent in terms of information content. One protein is assigned with a term that represents a broad type of activity, and its interacting partner is assigned with a more specific term that represents a subtype of that activity. In other cases, some proteins have not even been assigned with all three ontologies which make the interaction assessments more difficult without human intervention. Thus, molecular functions of GO annotations of related proteins should be harmonized in relation to information content and compared on a more general level. There is advantage and disadvantage associated with harmonization of GO terms. The advantage is that the relationships between proteins in a pair can be detected systematically using some keywords and it is not required to be verified manually. The disadvantage is that the integration of GO annotations and predicted PPIs might not be able to reveal the specific functions of interacting proteins. However, knowing the fact that PPI prediction techniques are merely capable to specify the general category of relationship between two proteins, this disadvantage is not a great source of concern.

In this chapter, a global framework to refine computationally predicted datasets is developed. First, two experimental PPI datasets with high confidence were prepared for two model organisms, *S. cerevisiae* and *C. elegans*. Assuming the experimentally confirmed pairs are true, the GO annotations of these interacting proteins were utilized to extract keywords which represent general category functions of the proteins. Then, a set of heuristic rules was established to be satisfied by predicted interacting proteins using extracted keywords and the fact that interacting proteins often function in the same cellular locations which assumes that two proteins acting in the same cellular components are more likely to interact than those located in different components. Next, four computational methods representing four out of six categories of prediction techniques,

mentioned earlier in this section, were selected. Using these methods, four predicted datasets were created for each organism of interest. The heuristic rules were applied to these predicted datasets. When a predicted pair of interacting proteins satisfied the rules it was considered a true positive, otherwise the pair was assumed false positive and removed from the dataset. The results show that the filtered datasets have higher true positive fractions than non-filtered datasets and the improvement is statistically significant.

5.3 Methods

5.3.1 Experimental datasets

The dataset containing experimentally obtained protein pairs was used to extract the functional keywords from GO annotations. The dataset was compiled from the following three sources: (1) von Mering *et al.* (2002) reported high confident yeast protein pairs that were confirmed by at least two experimental methods, resulting in 1920 protein pairs; (2) BIND database (Alfarano *et al.*, 2005) contains 10618 yeast protein pairs that were experimentally confirmed and manually curated; and (3) CYGD (Guldener *et al.*, 2005) contains 10472 experimentally verified yeast protein pairs. Combining three sources resulted in 16507 non-duplicated yeast protein pairs, consisting of 4391 proteins.

Worm dataset was constructed from BIND and Li *et al.* (2004). They reported 4960 and 6629 protein pairs, respectively. These pairs were obtained by means of yeast two-hybrid technique and manually curated. After removing repeated pairs the dataset consists of 7081 pairs, comprising 3390 proteins in *C. elegans*.

The two experimental datasets are presented in Supplementary Data (Chapter 5).

5.3.2 Computational protein-protein interaction methods

Four PPI predicting methods from four out of six categories discussed in chapter 1 were chosen, including phylogenetic profiles (PP), chance co-occurrence distribution coefficient (CC), gene expression profiles (GE), and maximum likelihood estimation (MLE). The criteria of choosing these methods were based on: their genome-wide applicability and competitive results in the category (Marcotte *et al.*, 1999; Butland *et al.*, 2005; Tu *et al.*, 2006; Liu *et al.*, 2005). Detail information on implementation of these methods is as follows:

1. Phylogenetic profiles (PP): The numbers of proteins studied in two organisms are, m=5863 in *S.cerevisiae* and m=12095 in *C.elegans*. The proteins of each organism were considered as queries and aligned against a database comprising 90 genomes using BLAST program. The list of reference genomes is in Supplementary Data (Chapter 5). Genomes were obtained from www.ncbi.nlm.nih.gov. Running BLAST program, using SEG filter over 75% similarity of the sequences, the output was a list of homolog proteins and their e-values within each genome that better match the query sequence. The best hit in each genome was taken as one bit in the profile and then profiles were created for each individual protein. These profiles should be converted into binary profiles in the form of 1 and 0 to represent the presence or absence of an individual protein in other genomes. To convert e-values to binary numbers it was required to know if the alignment score for each protein sequence P_i was statistically significant. Statistical significance of an alignment was described by the probability of finding a higher score when two sequences were compared based on a random selection. This probability depends on the number of comparisons made. If the number of proteins encoded in query genome is m and the number of encoded proteins in 90 reference genomes is p the total number of comparisons is: $m \times p$. Therefore, the probability of finding a match for an individual protein sequence is $1/(m \times p)$. In this study p=370461 and m for each organism is given above. We considered this probability as a threshold based on which e-values could be translated to present or absent status. Once the binary profiles were established, they were compared to find interacting proteins. Matching profiles were considered 'interacting'.

2. Gene co-expression profiles (GE): Genes with similar co-expression patterns are more likely to interact. To find out which genes are co-expressed, the expression levels of the studied genes were extracted from normalized DNA microarray data files obtained from Stanford Microarray Database (Ball *et al.,* 2005). Each file corresponds to an experiment. All expression values were collected in a gene expression matrix in which each row represents a different gene and each column corresponds to a different microarray experiment (100 experiments in *S. cerevisiae*, 575 experiments in *C. elegans*). The matrix is supplied into EXPANDER program (Shamir *et al.*, 2005) for clustering. Choosing click algorithm to cluster genes, the resulting number of clusters were 6 and 10 for yeast and worm genes, respectively. Overall homogeneity of clustering was 0.552 in

yeast and that of 0.631 in worm. Genes in the same cluster are co-expressed genes in different biological conditions. These genes were paired and considered 'interacting'.

3. Chance co-occurrence distribution (CC): Genes with identical patterns of occurrence across organisms tend to prediction of interactions; however, the requirement that the profiles be identical restricts the number of links that can be established by such pylogenetic profiling. Thus, there is a technique that relies on scoring phylogenetic patterns and matches them based on those scores rather than identical profiles. The scoring function provides more information than the simple presence or absence of genes.

Chance co-occurrence probability distribution has been used as a measure of similarity between two phylogenetic profiles. Based on the probability that a given arbitrary degree of similarity between two profiles would occur by chance, with no biological pressure, the interaction predictions are drawn with the criterion used to reject the null hypothesis. The probability *P(z|N,x,y)* of observing by chance (i.e. no functional pressure) z co-occurrence of genes *X* and *Y* in a set of *N* genomes, given that *X* occurs in *x* genomes, and *Y* occurs in *y* genomes is calculated as follows:

$$P = \frac{w_z \overline{w}_z}{W} \tag{5.1}$$

where w_z is the number of ways to distribute z co-occurrence over the N genomes, $\overline{w}_z$ is the number of ways of distributing x-z and y-z genes over the remaining N-z genomes, and W is the number of ways of distributing X and Y over N genomes without restriction. The final equation is as follows:

$$P = \frac{(N-x)!(N-y)!x!y!}{(x-z)!(y-z)!(N+z-x-y)!z!N!} \tag{5.2}$$

The general trend of –log(P) versus *z* for each protein pair (X,Y) is illustrated in Figure 5.1. Critical co-occurrence, z_c, is defined as the minimum number of co-occurrences required between two proteins to be considered as interacting proteins. Thus, as shown in this figure, protein pairs whose –log(P) is higher than a cut-off threshold (here, 8) and $z \geq z_c$ were predicted as interacting proteins.

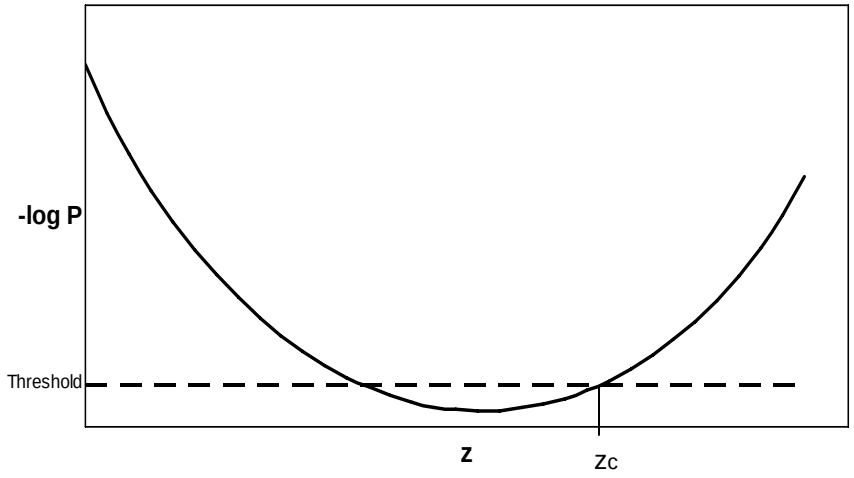

Figure 5.1. The negative logarithm of probability of z co-occurrence by chance (P) versus z. Based on the threshold value and z_c protein pairs with –log(P) located on the right-hand side portion of the curve are chosen as interacting proteins.

4. Maximum Likelihood Estimation (MLE): The underlying hypothesis in this method is two proteins interact if and only if at least one pair of domains from the two proteins interact. Let D_1, D_2,….,D_M denote the M domains, and P_1, P_2,….P_N denote N proteins. P_{ij} denotes the protein pair of P_i and P_j, and D_{ij} denotes the domain pair of D_i and D_j. Treating protein-protein interactions, and domain-domain interactions as random variables, the probability of interacting two proteins under stated assumption is:

$$\Pr(P_{ij} = 1) = 1.0 - \prod_{D_{mn} \in P_{ij}} (1 - \lambda_{mn}) \tag{5.3}$$

where $\lambda_{mn}= \Pr(D_{mn}=1)$ denotes the probability that domain D_m interacts with domain D_n. False positive rate (*fp*) and false negative rate (*fn*) are defined based on observed interactions. Let O_{ij} be the variable for the observed interaction result for proteins P_i and P_j. O_{ij} =1 if the interaction is observed and O_{ij}=0 otherwise. Then,

$$fp = \Pr(O_{ij} = 1 \mid P_{ij} = 0) \tag{5.4}$$

$$fn = \Pr(O_{ij} = 0 \mid P_{ij} = 1) \tag{5.5}$$

Thus, the probability of observing a protein-protein interaction is:

$$\Pr(O_{ij} = 1) = \Pr(P_{ij} = 1)(1 - fn) + (1 - \Pr(P_{ij} = 1))\, fp \tag{5.6}$$

The probability of the observed whole genome interaction dataset is

$$L = \prod (\Pr(O_{ij} = 1))^{O_{ij}} (1 - \Pr(O_{ij} = 1))^{1-O_{ij}} \quad (5.7)$$

where O_{ij}=1 if the interaction of P_i and P_j is observed and O_{ij}=0 otherwise. L is the likelihood and is a function of λ_{mn}, *fp*, and *fn*. In this calculation we fix *fp* and *fn* (see Chapter 4) and compute λ_{mn} using a recursive formula. First, initial values for λ_{mn} are chosen. Then $\Pr(P_{ij}=1)$ and $\Pr(O_{ij}=1)$ are computed by equations (5.3) and (5.6), respectively. Parameter λ_{mn} is updated using the following equation:

$$\lambda_{mn}^{(t)} = \frac{\lambda_{mn}^{(t-1)}}{N_{mn}} \sum_{i \in A_m, j \in A_n} \frac{(1 - fn)^{O_{ij}} \, fn^{1-O_{ij}}}{\Pr(O_{ij} = o_{ij} \mid \lambda^{(t-1)})} \quad (5.8)$$

and likelihood function is computes by Equation (7S). Calculations continue until the value of likelihood function is unchanged within a certain error.

The four prediction methods were applied to *S. cerevisiae*, and *C. elegans* genomes. The resulted eight datasets are available in Supplementary Data (Chapter 5).

5.3.3 Gene ontology annotations

The GO annotations of proteins were retrieved from the UNIPROT knowledgebase (Bairoch and Boeckman, 1992), which is collaborated with the GO database (The Gene Ontology Consortium, 2000). Annotations in both UNIPROT and GO databases are updated on a regular basis. In this study, the UNIPROT knowledgebase (Release 8, June 2006) and the GO database (Version 1.362, May 2006) were used to extract keywords for the false positive reduction on the predicted protein pairs.

5.3.4 Keywords extraction

Proteins involved in experimentally verified protein pairs were submitted to UNIPROT. Then GO and InterPro cross-reference assignments of the protein were retrieved. Through "*interpro2go*" (retrieved from Mappings to GO in GO website), all InterPro entries were mapped to GO terms and the GO terms of each protein were searched using AMIGO term search engine. The searched GO term information of each protein was collected and redundant information was removed. The remaining term definition relevant to molecular function annotation (a part of term information) was compiled and used as a training dataset. The dataset was further manually grouped into different clusters according to their general molecular activities; for instance,

GO:0003723 and GO:0000166 were placed in the same cluster because of molecule-binding activities. Refer to Supplementary Data (Chapter 5) for a complete listing of all clusters for *S. cerevisiae* and *C. elegans*.

In order to determine a representative keyword in a cluster, the number of occurrences (*n*) of a word in a cluster was counted, and the probability of finding that word in the training dataset was calculated using Poisson distribution:

$$p(n) = e^{-\lambda} \frac{\lambda^n}{n!} \tag{5.9}$$

where $\lambda = N \, x \, f$, in which N is the total number of words in a cluster, and f is the relative frequency of that word found in the whole training dataset. To avoid floating point errors and facilitate computation, *n!* was approximated by Stirling's approximation, resulting in:

$$\ln p(n) = -\lambda + n \ln \lambda - \ln(n!) \tag{5.10}$$

This calculation is valid when the total number of words in the training dataset is much greater than N or when f is small. In order to identify most comprehensive words in each cluster, grammatical terms such as proposition, and chemical formulae were purposefully eliminated. In "enzymatic function" cluster, all enzyme activities were considered as "*ase activity*" since enzymes are introduced with "*ase*" suffix in biochemistry literature. In each cluster the word with the most negative logarithmic value was selected as the representative keyword.

5.4 Results and Discussion

Using information deposited in the UNIPROT and GO databases, the experimentally obtained protein pairs for yeast and worm were processed, resulting in 1042 non-redundant GO term information (including 4391 yeast proteins) and 748 non-redundant GO term information (including 3390 worm proteins), respectively. These pieces of term information were further clustered, resulting in 35 and 25 keywords for yeast and worm, respectively (see Supplementary Data, Chapter 5).

5.4.1 Significant keywords

Low frequencies of appearance of some keywords in the training dataset prompts that all extracted keywords do not contribute equally to discriminate GO annotations. As

listed in Table 5.1, the frequency of appearance of each keyword was ranked in descending order. Eight top-ranking keywords were chosen for the following analyses, and the remaining keywords (27 in yeast and 17 in worm) were grouped and called it as "remains". In order to evaluate the significance of these top-ranking keywords, the sensitivity and specificity analysis was conducted. Sensitivity (SN) is the percentage of protein pairs that are recovered using a certain keyword or a group of keywords when they are applied back to the source (the training dataset). Specificity (SP) is the percentage of protein pairs recovered when keywords are applied to predicted datasets (the test datasets).

The sensitivity of each keyword was calculated as:

$$SN\ \ of\ \ a\ \ keyword = \frac{number\ \ of\ \ pairs\ \ represented\ \ by\ \ the\ \ keyword}{total\ \ number\ \ of\ \ pairs\ \ in\ \ training\ \ set} \times 100 = \frac{1}{x}\sum_{i=1}^{x} n_i \times 100 \qquad (5.11)$$

where x is the total number of pairs in the experimental dataset (the training dataset). If n_i =1, it indicates that two proteins in pair i are represented by a keyword; and n_i =0, otherwise. Cumulative sensitivity of all keywords was obtained as:

Table 5.1. Frequencies of extracted keywords in the yeast training set (experimental dataset).

Keywords	***frequency***
Binding	3337
ase activity	2797
Porter activity	397
Transcription activity	372
Ribosome	134
Translation activity	58
Structural activity	51
Receptor activity	23
Remaining keywords (27 keywords)	230

$$Cumulative\ \ SN = \frac{1}{x}\sum_{i=1}^{x}\sum_{j=1}^{z} n_{ij} \times 100 \qquad (5.12)$$

where z is the number of keywords. If n_{ij}=1, it shows that two proteins in pair i are represented by the common keyword j; and n_{ij}=0, otherwise. Cumulative sensitivity demonstrates the recovery power of all keywords collectively when they are applied to the source (training set). Specificity of a keyword and cumulative specificity of all keywords are similarly defined and calculated:

$$SP\ \ of\ \ a\ \ keyword = \frac{number\ \ of\ \ \ pairs\ \ represented\ \ by\ \ the\ \ keyword}{total\ \ number\ \ of\ \ \ pairs\ \ in\ \ test\ \ set} \times 100 = \frac{1}{y}\sum_{i=1}^{y} n_i \times 100 \qquad (5.13)$$

$$Cumulative\ \ SP = \frac{1}{y}\sum_{i=1}^{y}\sum_{j=1}^{z} n_{ij} \times 100 \qquad (5.14)$$

where y is the total number of pairs in the predicted dataset (the test dataset). Cumulative specificity translates into the recovery power of all keywords when they are applied to a predicted dataset (test set).

Figure 5.2 illustrates the cumulative sensitivity variations among extracted keywords in both organisms. The cumulative sensitivity of all keywords is 64.43%. While only the top 8 high-scored keywords are considered, the cumulative sensitivity is 64.21%, indicating that the remaining keywords imposed relatively insignificant contribution to the cumulative sensitivity. Similarly, in worm the same eight keywords contributed to 80.83% cumulative sensitivity and the remaining keywords increased that value to 80.88% (i.e. 0.05% increase). Thus, in trade-off between the lowest number of keywords and the highest cumulative sensitivity, it is favourable to neglect 27 keywords in yeast (17 keywords in worm) with the price of 0.22% (0.05% in worm) lower sensitivity.

In order to further examine the significance of extracted top-ranking keywords from the training dataset, the cumulative specificity of the keywords applied to four predicted protein-protein interaction datasets were calculated. These four predicted datasets were obtained using computational methods including phylogenetic profiles (PP), gene expression (GE), maximum likelihood estimation (MLE), and chance co-occurrence distribution (CC). The implementation of these methods is described in Methods. As illustrated in Figure 5.3, the cumulative specificity varies from 25% in PP dataset to 69% in MLE dataset. In all four predicted datasets specificity changes very slightly when it is

extended from top-ranking keywords to all extracted keywords. Similarly, in Figure 5.4, cumulative specificity ranges from 32% in PP dataset to 64% in MLE dataset using top-ranking eight keywords. The remaining keywords exert negligible changes to cumulative specificities in all four datasets. Therefore, these top-ranking eight keywords extracted from the experimental datasets of both organisms are capable of representing the common functions of interacting proteins either experimentally specified or computationally predicted.

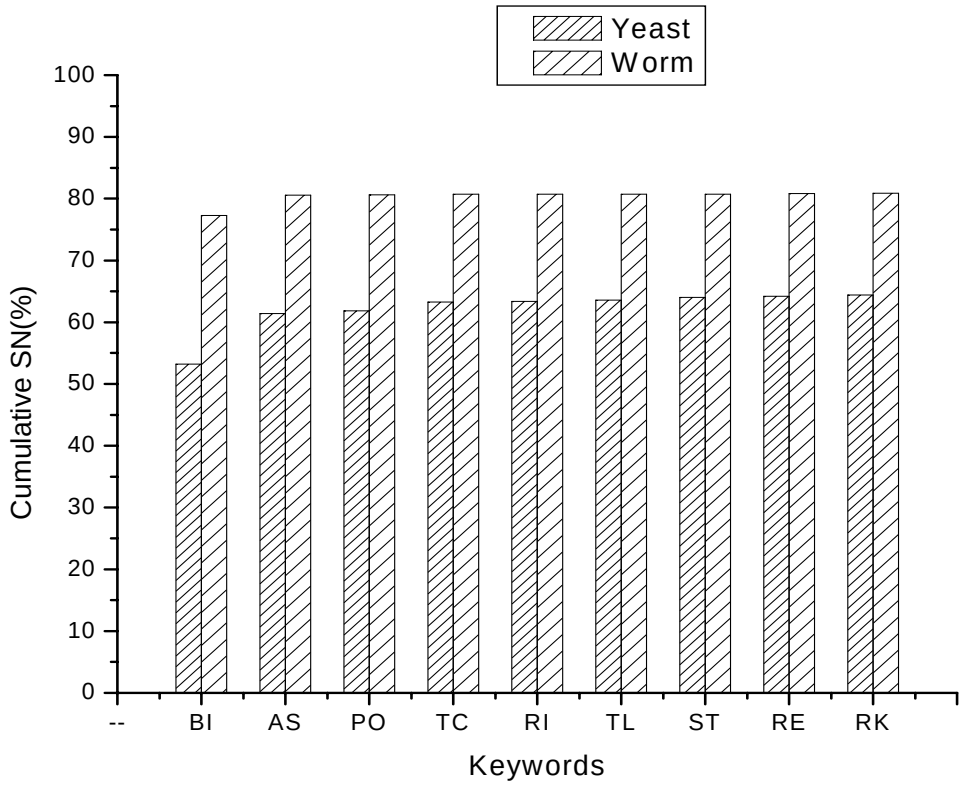

Figure 5.2. Cumulative sensitivity of keywords for yeast and worm datasets. Each column indicates sensitivity of a keyword in addition to sensitivities of previous keywords. The highest sensitivities are 64.43% and 80.88% in yeast and worm training datasets, respectively. Abbreviations for keywords are as follows: BI (binding), AS (ase activity), PO (porter activity), TC (transcription activity), RI (ribosome), TL (translation activity), ST (structural activity), RE (receptor activity), and RK (remaining keywords).

Although the eight top-ranking keywords significantly recover the experimental or predicted datasets, the cumulative sensitivity or specificity is not distributed equally as seen in Figures 5.2-5.4. Among keywords “binding (BI)” is an exception with the sensitivity of 53.22% in yeast dataset, for instance, compared to 8.20% for “ase activity (AS)”, 0.43% for “porter activity (PO)”, and so on. This drastic difference between the sensitivity or specificity of this particular keyword and that of other keywords stems from the fact that our experimental dataset is a collection of protein interactions detected mainly by two-hybrid technique. This high-throughput technique detects physical interactions among proteins in which binding of a protein to active site of another protein is a crucial step. Accordingly, most of these protein pairs are assigned with “binding” molecular function annotation in GO database. On the other hand, contribution of some keywords such as “receptor activity (RE)” in cumulative sensitivity is 0.20% which is not a remarkable contribution; however, it is significant when it is compared with 0.22% increase in cumulative sensitivity by “remaining keywords (RK)” which represents 27 keywords in case of yeast.

It should be noted that the highest obtainable cumulative sensitivity, in yeast for example, is 64.43% and 64.21% as top-ranking eight keywords were employed. Currently, it is impossible to obtain complete sensitivity (100%), as some experimental pairs do not have consistent annotations. This inconsistency comes from the fact that there are deficiencies in either annotation or experimental techniques. In case of worm the inconsistency is worse than yeast. Only 55% of worm genes are annotated and many annotations are not consistent. It is also notable that GO molecular function annotations can not be used directly as keywords. When the definition of GO molecular function was considered as a keyword, the cumulative sensitivity of the training dataset was only 45%, comparing to that of 64% the keyword extraction approach was implemented.

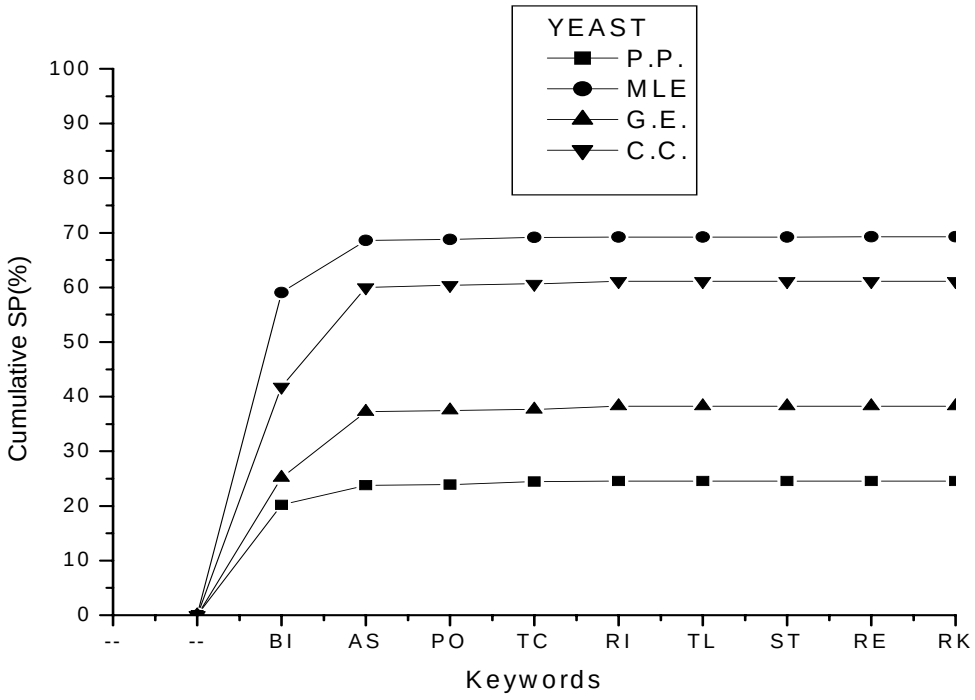

Figure 5.3. Cumulative specificity of trained keywords for yeast dataset. The keywords are applied to four predicted PPI datasets. Each data point indicates specificity of a keyword in addition to specificities of previous keywords. Abbreviations for keywords are as follows: BI (binding), AS (ase activity), PO (porter activity), TC (transcription activity), RI (ribosome), TL (translation activity), ST (structural activity), RE (receptor activity), and RK (remaining keywords). RK includes 27 keywords with negligible contribution to cumulative SP.

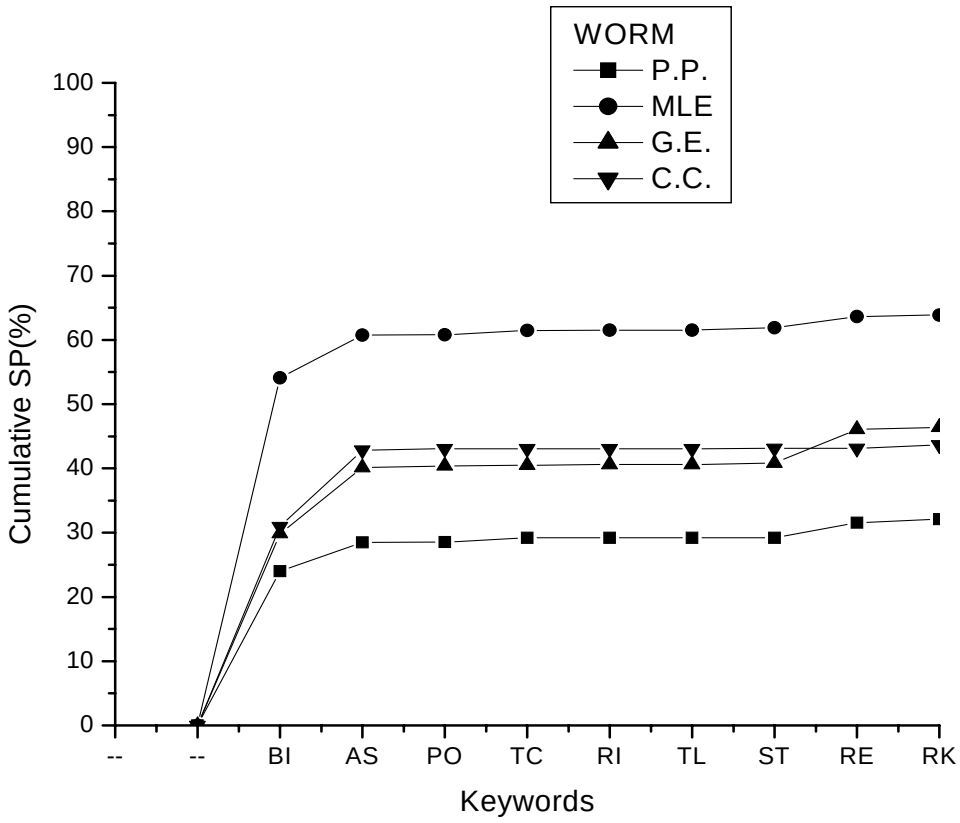

Figure 5.4. Cumulative specificity of trained keywords for worm dataset. The keywords are applied to four predicted PPI datasets. Each data point indicates specificity of a keyword in addition to specificities of previous keywords. Abbreviations for keywords are as follows: BI (binding), AS (ase activity), PO (porter activity), TC (transcription activity), RI (ribosome), TL (translation activity), ST (structural activity), RE (receptor activity), and RK (remaining keywords). RK includes 17 keywords with negligible contribution to cumulative SP.

5.4.2 Heuristic Rules

Protein interactions take place in either permanent or transient complexes formed in a cell (Cho *et al.,* 2006) suggesting that proteins are required to exist in close proximity to interact physically (Nooren and Thornton, 2003). Hence, the concept of protein-protein interactions in cellular systems is based on the following two observations: (i) interacting proteins often perform similar general functions, assuming that two proteins functioning in the same general category are more likely to interact than two proteins involved in different functions: (ii) co-localization may serve as an useful tool to predict protein interactions. Physical interactions occur when two proteins are located in the same cellular component, either a permanent cellular location or a transient complex. Motivated by the two observations, two heuristic rules were set to be satisfied by predicted interacting protein pairs. These rules are:

(I) Two predicted proteins in the pair should match one of the eight trained function keywords.

(II) Two predicted proteins in the pair should be in the same GO cellular components.

As many computational protein interaction prediction techniques suffer from mass false positive predictions, satisfying the rules filters the predicted datasets and removes the false interactions to some extent.

Based on the algorithm depicted in Figure 5.5, these two rules were applied to eight predicted PPI datasets for both yeast and worm (see Supplementary Data, Chapter 5 for source codes and output files). The algorithm reads PPI pairs predicted by PP, GE, MLE, and CE sequentially. The algorithm then examines if two proteins in the same pair possess GO annotations: molecular function and cellular component. If so, such a pair with annotations is checked with the proposed rules. Satisfying rule 1 and rule 2, this protein pair is considered as an interacting one. Finally, the filtered predicted dataset is compared with experimental dataset to assess the level of agreement with experimental results.

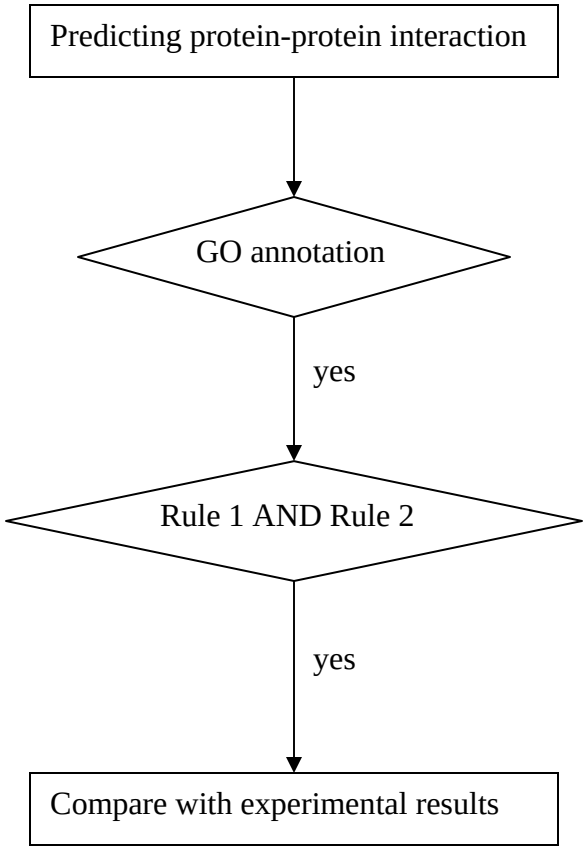

Figure 5.5. The flowchart of algorithm used to filter predicted protein interaction datasets.

In order to evaluate the improvement made by applying rules to the predicted PPI datasets, the signal-to-noise ratio (SNR) (Fujimori *et al.*, 1974) was employed. SNR is a measure of signal strength relative to background noise. In bioinformatics, SNR is translated to the ratio of capability of a computational technique in creating protein pairs to pairing proteins on a random basis. Therefore, we define SNR as the ratio of the true positive fraction of a predicted dataset to the true positive fraction of a randomly selected dataset with the same sample size. True positive fraction of a dataset is the ratio of matched protein pairs found in the experimental dataset to the total number of pairs in the same dataset:

$$SNR = \frac{\left(matched\ \ pairs \Big/ total\ \ pairs \right)_{predicted\ \ dataset}}{\left(matched\ \ pairs \Big/ total\ \ pairs \right)_{random\ \ dataset}} \tag{5.15}$$

SNR was calculated for all four predicted datasets for yeast and worm in the following two circumstances: before applying the rules to a dataset (raw dataset), and after applying the rules to a dataset (filtered dataset). The effect of the rules on the reduction of false positive predictions was measured by *strength* (*S*):

$$S = \frac{SNR_{Filtered\ \ Dataset}}{SNR_{Raw\ \ Dataset}} \tag{5.16}$$

Table 5.2 indicates values of the *strength* in four predicted datasets in each studied organism. As seen in this table when rules were applied to a predicted dataset and false positive predictions were removed, the true positive fraction of the dataset improved from approximately 2-fold to 10-fold compared to true positive fraction of the same dataset prior to applying the rules. Despite overall improvement in true positive fraction of predicted datasets, the *strength* value depends on the computational method employed to create a predicted dataset. In PP method, rules play more effective roles to eliminate false positive predictions than other three methods. The *Strength* was 9.9 in PP dataset in yeast while it was 2.32 when rules were applied to MLE dataset. The same trend was observed in worm datasets. The highest *strength* 3.94 occurred in PP. dataset and the lowest *strength* was obtained in MLE dataset. This indicates that in MLE method the rules are less effective than others due to the higher accuracy of this prediction method in the first place that relies on domain content of proteins, and protein-protein interactions are build upon domain-domain interactions. Overall *strength* values in yeast are greater than their corresponding values in worm. This is due to more availability of experimental information on yeast protein interactions than that on worm. Yeast is a well studied single cellular organism with many characterized proteins, while worm is a more complicated multi-cellular organism with numerous uncharacterized proteins.

Table 5.2. SNR and S values of predicted datasets, before (raw dataset) and after (filtered dataset) removing false positives.

	Yeast			*Worm*		
Method	**SNR*(Raw Dataset)**	**SNR* (Filtered Dataset)**	**S**	**SNR* (Raw Dataset)**	**SNR* (Filtered Dataset)**	**S**
PP	1.59	15.78	9.90	32.78	129.0	3.94
GE	1.89	8.83	4.67	27.36	66.0	2.41
CC	3.10	12.21	3.94	51.88	202.0	3.89
MLE	13.44	31.14	2.32	197.2	387.0	1.96

(*) SNR was calculated based on Equation (5.15). Random datasets were established based on the same number of protein pairs with corresponding sets and their true positive fractions were calculated based on the mean of 100 trials.

The algorithm proposed here to reduce the number of false positive interaction predictions has a global application. This algorithm can be applied to any predicted protein-protein interaction dataset and is not biased toward any computational approach. The algorithm is a post-prediction processing step so that it is applied to the resulted predicted dataset when a computational method is implemented. Thus, it can be attached to any computational strategy for further improvement of predicted results. However, it should be noted that ontology is an ongoing process and for new genomes only a few percentage of genes have been fully annotated. With more genes assigned with GO terms, the proposed filtering algorithm becomes a promising approach to reduce the number of false positive interactions and enhance the accuracy of inferring protein-protein interactions.

5.5 Conclusion

Gene ontology annotation was used as a common ground to evaluate protein pairs predicted by four different PPI-predicting methods. Molecular function annotations in Gene Ontology database were used to extract discriminating keywords, upon which heuristic rules were set. The rules were incorporated into an algorithm by which predicted datasets were filtered and false positive predictions were partially removed from the datasets. When only eight top-ranking keywords were chosen, on average 71% of molecular function annotations could be recovered as indicated by the cumulative sensitivity for both experimentally obtained and computationally predicted protein pairs.

The effectiveness of the proposed algorithm to filter false positive predicted protein pairs varies from one method to another. The proposed algorithm was unbiased and could be implemented to any existing computational method to increase the accuracy of PPI prediction. As more genes are assigned with GO annotations, the proposed filtering algorithm will become more effective accordingly.

References

Alfarano C., Andrade C.E., Anthony K., Bahroos N., Bajec M., Bantoft K., Betel D., Bobechko B., Boutilier K., Burgess E., Buzadzija K., Cavero R., D'Abreo C., Donaldson I., Dorairajoo D., Dumontier M.J., Dumontier M.R., Earles V., Farrall R., Feldman H., Garderman E., Gong Y., Gonzaga R., Grystan V., Gryz E., Gu V., Haldorsen E., Halupa A., Haw R., Hrvojic A., Hurrell L., Isserlin R., Jack F., Juma F., Khan A., Kon T., Konopinsky S., Le V., Lee E., ling S., Magidin M., Moniakis J., Montojo J., Moore S., Muskat B., Ng I., Paraiso J.P., Parker B., Pintilie G., Pirone R., Salama J.J., Sgro S., Shan T., Shu Y., Siew J., Skinner D., Snyder K., Stasiuk R., Strumpf D., Tuekam B., Tao S., Wang Z., White M., Willis R., Wolting C., Wong S., Wrong A., Xin C., Yao R., Yates B., Zhang S., Zheng K., Pawson T., Ouellette B.F., Hogue C.W., *et al.* (2005) The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res.* **33,** D418-D424.

Bairoch A., Boeckman B. (1992) The SwissProt protein sequence data bank. *Nucleic Acids Research*, **29,** 2019-2022.

Ball C.A., Awad I.A.B., Demeter J., Gollub J., Hebert J.M., Hernandez-Boussard T., Jin H., Matese J.C., Nitzberg M., Wymore F., Zachariah Z.K., Brown P.O., Sherlock G. (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Research*, **33,** D580-D582.

Bard J.B.L., Rhee S.Y. (2004) Ontologies in biology: design, applications and future challenges. *NATURE REVIEWS, GENETICS*, **5:**213-222.

Bork P., Jensen L.J., von Mering C., Ramani A.K., Lee I., Marcotte E.M. (2004) Protein interaction networks from yeast to human. *Current Opinion in Structural Biology*, **14**, 292-299.

Butland G., Peregrin-Alvarez J.M., Li J., Yang W., Yang X., Canadien V., Starostine A., Richards D., Beattie B., Krogan N., Davey M., Parkinson J., Greenblatt J., Emili A. (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature*, **433:**531-537.

Campoy A.V., Freire E. (2005) ITC in the post-genomic era…? priceless. *Biophysical Chemistry*. **115**, 115-124.

Cho K., Lee K., Lee K.H., Kim D., Lee D. (2006) Specificity of molecular interactions in transient protein-protein interactions interfaces. *PROTEINS: Structure, Function, and Bioinformatics,* **65:**593-606.

Deane C.M., Salwinski L., Xenarios I., Eisenberg D. (2002) Two methods for assessment of the reliability of high throughput observations. *Molecular & Cellular Proteomics,* **1(5),** 349-356.

Fujimori T., Miyazu T., Ishikawa K. (1974) Evaluation of analytical methods using signal-noise ratio as a statistical criterion. *Microchemical Journal,* **19(1):**74-85.

Guldener U., Munsterkotter M., Kastenmuller G., Strack N., van Helden J., Lemer C., Richelles J., Wodak S.J., Garcia-Martinez J., Perez-Ortin J.E., Michael H., Kaps A., Talla E., Dujon B., Andre B., Souciet J.L., De Montigny J., Bon E., Gaillardin C., Mewes H.W. (2005) CYGD: the comprehensive yeast genome database. *Nucleic Acids Res.* **33,** D364-D368.

Hegyi H., Gerstein M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Research,* **11,** 1632-1640.

Jiang T., Keating A.E. (2005) AVID: an integrative framework for discovering functional relationships among proteins. *BMC Bioinformatics,* **6**:136.

Li S., Armstrong C.M., Bertin N., Ge H., Milstein S., Boxem M., Vidalain P.O., Hao T., Goldberg D.S., Li N., Martinez M., Rual J.F., Lamesch P., Xu L., Tewari M., Wong S.L., Zhang L.V., Berriz G.F., Jacotot L., Vaglio P., Reboul J., Hirozane-Kishiawa T., Li Q., Gabel H.W., Gabel H.W., Elewa A., Baumgartner B., Rose D.J., Yu H., Bosak S., Sequerra R., Fraser A., Mango S.E., Saxton W.M., Strome S., Van den Heuvel S., Piano F., Vandenhaute J., Sardet C., Gerstein M., Doucette-Stamm L., Gunsalus K.C., Harper J.W., Cusick M.E., Roth F.P., Hill D.E., Vidal M. (2004) A map of interactome network of the metazoan *C. elegans*. *Science,* **303,** 540-543.

Liu Y., Liu N., Zhao H. (2005) Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics,* **15,** 3279-3285.

Lord P.W., Stevens R.D., Brass A., Goble C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics,* **19(10):**1275-1283.

Marcotte E.M., Pellegrini M., Thompson M.J., Yeates T.O., Eisenberg D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature,* **402,** 83-86.

Marcotte E.M., Xenarios I., van der Bliek A.M., Eisenberg D. (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci.,* **97(22)**, 12115-12120.

Mrowka R., Patzak A., Herzel H. (2001) Is there a bias in proteome research? *Genome Research,* **11,** 1971-1973.

Nooren I.M.A., Thornton J.M. (2003) Structural Characterization and functional significance of transient protein-protein interactions. *J. Mol. Biol.,* **325,** 991-1018.

Patil A., Nakamura H. (2005) HINT: a database of annotated protein-protein interactions and their homologs. *bp Biophysics*, **1**, 21-24.

Reboul J., Vaglio P., Rual J.F., Lamesch P., Martinez M., Armstrong C.M., Li S., Jacotot L., Bertin N., Janky R., Moore T., Hudson Jr. J.R., Hartley J.L., Brasch M.A., vandenhaute J., Boulton S., Endress G.A., Jenna S., Chevet E., Papasotiropoulos V., Tolias P.P., Ptacek J., Snyder M., Huang R., Chance M.R., Lee H., Doucette-Stamm L., Hill D.E., Vidal M. (2003) *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genetics,* **34,** 35-41.

Rhodes D.R., Tomlins S.A., Varambally S., Mahavisno V., Barrette T., Kalyana-Sundaram S., Ghosh D., Pandey A., Chinnaiyan A.M. (2005) Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology,* **23,** 951-959.

Shamir R., Maron-Katz A., Tanay A., Linhart C., Steinfeld I., Sharan R., Shiloh Y., Elkon R. (2005) EXPANDER- an integrative program suite for microarray data analysis. *BMC Bioinformatics,* **6:**232.

The Gene Ontology Consortium (2000) Gene Ontology: tools for the unification of biology. *Nature Genet.,* **25,** 25-29.

Tirosh I., Baraki N. (2005) Computational verification of protein-protein interactions by orthologous co-expression. *BMC bioinformatics,* **6**:40.

Tu K., Yu H., Li Y.-X. (2006) Combining gene expression profiles and protein-protein interaction data to infer gene functions. *Journal of Biotechnology,* **124:**475-485.

Valencia A., Pazos F. (2002) Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology,* **12:**368-373.

von Mering C., Krause R., Snel B., Cornell M., Oliver S.G., Fields S., Bork P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature,* **417,** 399-403.

Wojcik J., Boneca I.G., Legrain P. (2002) Prediction, assessment and validation of protein interaction maps in bacteria. *J. Mol. Biol.* **323:**763-770.

Wu H., Su Z., Mao F., Olman V., Xu Y. (2005) Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic Acids Research,* **33(9):**2822-2837.

Wu X., Zhu L., Guo J., Zhang D.-Y., Lin K. (2006) Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Research,* **34(7):**2137-2150.

Yu H., Paccanaro A., Trifonov V., Gerstein M. (2006) Predicting interactions in protein networks by completing defective cliques. *Bioinformatics,* **22(7):**823-829.

Yu J., Fotouhi F. (2006) Computational approaches for predicting protein-protein interactions: A survey. *Journal of Medical Systems,* **30(1):**39-44.

Zhu H., Bilgin M., Snyder M. (2003) Proteomics. *Annu. Rev. Biochem.,* **72:**783-812.

6

EXPANDING RECONSTRUCTED METABOLIC NETWORK OF *C. ELEGANS* USING NEW PREDICTED PROTEIN-PROTEIN INTERACTIONS

Contribution of this chapter to the overall study

In this chapter newly predicted protein-protein interactions were incorporated into the current metabolic network of *C. elegans* and new function for uncharacterized proteins were inferred. These new functions were outcome of the expanded version of the metabolic network resulted in this research.

6.1 Abstract

Metabolic networks are greater portrays of entire metabolic activities taking place in a living cell. This picture consists of many elements including genes, proteins (enzymes), metabolites, and reactions categorized into pathways. Growing efforts are made to identify all these elements and discover relationships among them and eventually put them in a network context. No metabolic network has been completed so far as many organisms' cellular systems especially eukaryotes are extremely complicated. Nevertheless, many attempts were made, including this chapter, to expand these sophisticated networks step by step. To expand a metabolic network more protein-protein interaction information were supplied to the current network. These pair-wise interactions were compared with the known interactions and new partners were identified. With the predicted interaction dataset provided by signature profiling method, 1024 novel interactions were introduced upon which the known metabolic proteins in *C. elegans* metabolic network increased from 17% to 22%, nearly 27% increase compared to the current network. Novel interactions were used to infer function for the unknown proteins involved in these interactions. The possible locations and the association of metabolic reactions of these unknown proteins within the network were inferred to eventually narrow down the number of experiments ought to be performed to confirm these links.

6.2 Introduction

Two-dimensional genome annotation refers to the integration of various levels of metabolic information and reconstruction of metabolic networks. Metabolic information is presented in different 'omics' including genomics, transcriptomics, proteomics, and recently metabolomics (Beecher, 2002). Metabolomics is the latest piece of this chain. It is defined as the collection of all metabolites synthesized by proteins in a living cell. Metabolite profiling of some species has been performed (Roessner *et al.,* 2002) and many metabolites have been characterized due to this global approach (Fernie, 2003). With the availability of metabolite data in biological systems, analysis of biological processes especially metabolic networks will be more accurate and comprehensive (Thomas, 2001). Computational approaches such as machine learning algorithms have been used to discover simple and robust rules in the metabolomic map of organisms (Kell, 2002). Furthermore, numerous experimental techniques are available to detect the metabolite profile of organisms. These techniques are discussed elsewhere (for example, Castrillo *et al.,* 2003) and are beyond the scope of this chapter.

With the combination of all these hierarchies now researchers are able to link these different pieces of information and discover missing links in functional associations (Hall *et al.*, 2002), novel pathways (Weckwerth and Fiehn, 2002), uncharacterized genes and their attributes (Trethwey, 2001) and eventually understand metabolic networks (Fiehn, 2001). Early studies on metabolic networks, especially dynamic mathematical models, relied on heuristic-based methods such as cybernetic framework, because of low availability of biological information (Varner and Ramkrishna, 1999). With the growing number of sequenced genomes, Jeong *et al.* (2000) proposed a large-scale mathematical model for metabolic networks. This model was applied to 43 organisms in three domains of life and despite significant variation in their individual pathways the model could demonstrate striking similarities among organization of metabolic networks. Thus, all the requirements of the mathematical model of a metabolic network were identified and the type of resources utilized in this mathematical representation has already been specified (Wiechert, 2002). Integrating information captured by multi-parallel techniques on metabolic organization of an organism is one way to develop mathematical models. In plant biology *Arabidopsis* is a pre-eminent plant model extensively studied (Fiehn *et al.*,

2001). Even comparing model organisms may provide substantial information on quantitative analysis of common reactions and their missing substrates (Krijgsveld *et al.*, 2003). Structural bases also provide valuable information on substrate specificity of metabolic reactions which is useful in flux analysis (Brinkworth *et al.*, 2003). Integration of metabolic pathways with non-metabolic pathways such as regulatory and signalling may reveal some metabolic relationships which are involved in non-metabolic activities. Mastellos *et al.* (2005) employed a text-based data mining technique, called systems literature analysis (SLA), to elucidate interactions as such. With all this information, and using powerful bioinformatics tools, gene and their products are now assigned to metabolic pathways with high precision (Popescu and Yona, 2005). Pathway assignment also specifies the phylogenetic relationship of genes as conserved property of the genomes which has been practical with the aid of gene ontology and enzyme relationships (Clemente *et al.*, 2005).

Visualizing metabolic networks is another front to understanding and interpreting the network identity. Luyf *et al.* (2002) developed a visualizing tool, ViMAc, to explore the layout of yeast metabolic network representing expression data in a metabolic context. Visualization and interpretation of genome-wide functional linkages inferred from computational methods was performed to explore the hierarchy of genes in expression data (Strong *et al.*, 2003). In this representation each linkage was displayed on a two-dimensional scatter plot, organized according to the order of gene on chromosome. These visualizing tools were not applicable to large-scale networks. Adai, *et al.* (2004) proposed an algorithm to visualize very large biological networks. This algorithm is based on a force-directed iterative layout guided by a minimal spanning tree of the network. Using the algorithm, 23 new protein families were identified. As many network algorithms produce machine-readable representation of the networks, a process diagram was proposed to further represent metabolic networks in a human-readable form which is more useable to infer biological information from the network (Kitano *et al.*, 2005).

One of the immediate outcomes of metabolic networks is function inference. Different strategies have been used to infer gene function. Accumulation of data on gene expression and gene sequencing has motivated integrating most pertinent functional data for function inference (Date and Marcotte, 2001). RNA-mediated interference targeted

elucidating function for approximately 14% of *C. elegans* unknown genes mainly on chromosome I (Fraser *et al.*, 2000). Following advances in RNA-mediated methodology, this information was integrated with other large-scale data such as microarray and protein interaction maps to enhance the speed and reliability of such function inference (Sugimoto, 2004). Intracellular concentration of metabolites were also used to establish a functional strategy for 'silent' *S. cerevisiae* genes which show no phenotype in terms of growth rate or other fluxes when they are deleted from the genome (Raamsdonk *et al.*, 2001). Probabilistic approaches were applied to metabolic networks and protein interaction maps to predict function on a genome scale (Letovsky and Kasif, 2003). Moreover, since most function inference techniques need manual curation of the information, probabilistic approaches such as gene ontology are able to assign gene functions through an iterative process that ultimately converges on the correct functions (Fraser and Marcotte, 2004). The robustness of these approaches intensively depends on the accuracy of the datasets employed that emphasizes on the importance of gold standard interaction datasets for function inference (Jansen and Gerstein, 2004). Now there are systematic genome-wide methods available to determine the function of an unknown gene and its products. These methods were reviewed by Carpenter and Sabatini (2004). Recently, structural genomics was used to predict function for un-annotated enzymes in metabolic networks (von Grotthuss *et al.*, 2006). It has been shown in another study, proteins that co-operate in these networks in form of functional modules are groups of interacting proteins that are responsible for a specific step in a biological process (Chen and Yuan, 2006).

New experimental technologies and emerging computational prediction techniques produce a huge amount of protein-protein interaction information. Thus, metabolic networks should be updated to keep up with the rapid increase in available protein-protein interaction information. In order to include new information to the current knowledge of metabolic activities, it should be verified by means of statistical evaluation techniques and experimental findings. Following this validation process, the predicted results would be accurate enough to be candidates for further experiments. Therefore, the ultimate goal of computational prediction is suggesting potential interactions for further experimental validation.

In this chapter, the new protein-protein interactions, predicted by signature profiling method and evaluated by the rules inferred from GO annotations, have been integrated into the current metabolic network of *C. elegans* resulting in a bigger picture of metabolic activities in this species. In this expanded network, new proteins are associated with pathways and new enzymatic activities can be inferred for the uncharacterized enzymes in each pathway.

6.3 Methods

Metabolic networks are reconstructed based on the catalytic activities of enzyme proteins. Basically, regulation of each metabolic reaction in a cell is the outcome of the activity of many regulatory and signalling proteins. However, in a metabolic network those regulatory and signalling proteins do not appear, since they make the network much more complicated. On the other hand, genome-wide protein-protein interaction prediction methods are not able to distinguish among different types of proteins such as metabolic, regulatory, signalling, etc. Therefore, in our working protein interaction dataset, predicted by signature profiling approach, the protein pairs in which at least one partner is involved in metabolism in the current metabolic network were selected. The selected pairs were compared against the existing protein-protein interaction map of *C. elegans* (see Section 3.3.3) and new interactions were specified. Since in each pair one protein was known, the unknown partner was assigned to the pathway in which the known partner was participating. In cases where there is more than one partner for the known protein, the interacting protein with highest binary similarity score has the highest probability to be involved in that pathway and the remaining partners were ranked for their involvement in the pathway based on their binary similarity scores. This ranking was further used for function inference. Because of the generality of function of some proteins they may appear in more that one pathway. These proteins contribute to the interconnectivity of the network. Figure 6.1 illustrates the assignment procedure of new protein interactions to existing pathways.

Once the unknown interacting proteins were assigned to pathways, gene function inference was performed based on novel protein-protein interactions in the expanded metabolic network. The function inference is upon the notion that when a known protein is involved in a reaction within a particular pathway, its unknown partner is predicted to

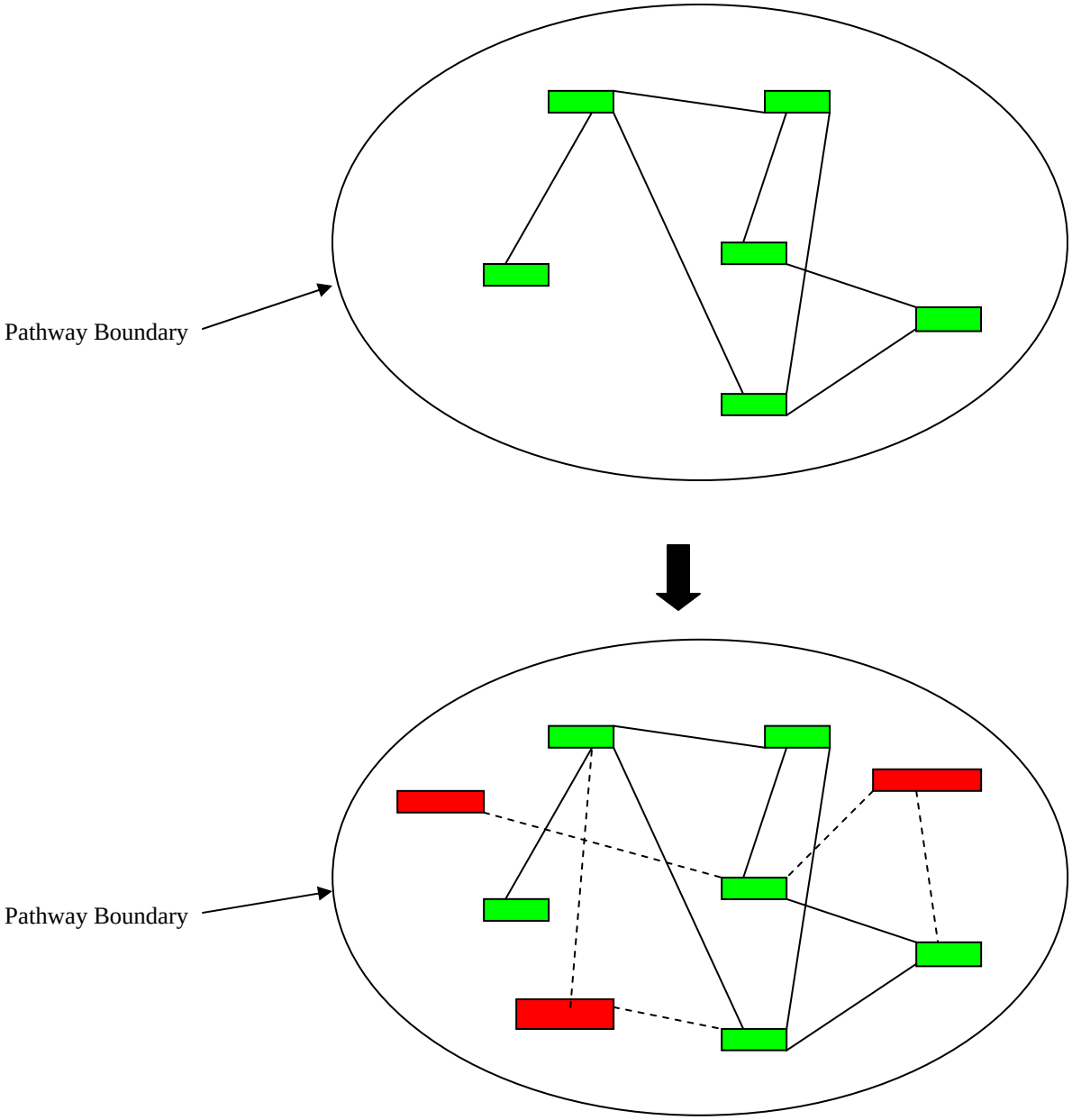

Figure 6.1. Pathway assignment procedure using new protein-protein interaction data. The upper part of the figure represents an existing web of interactions in a pathway. The lower part, demonstrates the association of newly characterized proteins with their known partners. The solid lines represent the current interactions in one particular pathway in the metabolic network. The dash lines represent new assignments to the pathway based on predicted interactions resulting from signature profiling approach.

be involved in the consecutive reaction (downstream or upstream) or in a parallel reaction that synthesize same product in an alternative path. Because the pathway structures were retrieved from KEGG, newly assigned proteins to the pathways are suitable candidates for enzymes whose reactions are known in KEGG, but their encoding genes are still unknown. Thus, it can be inferred that the unknown proteins (genes) interacting with known proteins encode those enzymes whose working reactions are given. When more than one unknown protein was involved for a particular enzymatic function, the protein with the highest binary similarity score was given the highest chance to link to that function.

6.4 Results and Discussion

6.4.1 Novel protein interactions

Signature profiling approach predicted interacting protein pairs in *C. elegans* (see chapter 4). These pairs were predicted and screened through the false positive reduction algorithm discussed in chapter 5. Selecting the pairs in which at least one partner is known to be involved in metabolism, 1235 pairs remained in our dataset. This result is complying with the observation that approximately 10-20% of proteins in organisms are involved in metabolism (van Nimwegen, 2003). The dataset of 1235 protein pairs was compared to current protein-protein interaction map of *C. elegans* (see chapter 3) to find out how many novel pairs have been predicted by the signature profiling method. Of 1235 pairs, 211 of which exist in the current map resulting in 1024 new predicted interactions (see Supplementary Data, Chapter 6). In these novel interactions one known protein is interacting with its unknown partner. The dataset of 1024 new interactions consisted of 294 proteins. Novel interactions were embedded to current protein-protein interaction map (see chapter 3) and the number of interactions increased from 32902 pairs to 33926 pairs (see Supplementary Data, Chapter 6). In this expanded protein-protein interaction map, the connectivity of each protein decreases from 42 to 34 because the number of characterized metabolic genes increases from 792 to 1009. This translates to a 27% increase in the number of characterized genes involved solely in metabolism. From the network's point of view, this is one step forward toward the completion of metabolic network of *C. elegans.*

6.4.2 Function inference and pathway association

Employing the strategy discussed in Methods (Section 6.3), all 1024 novel interactions were distributed in 94 metabolic pathways and their probable function were inferred (see Supplementary Data, Chapter 6), resulting in an expanded metabolic network of *C. elegans*. The uncharacterized proteins in novel interactions were assigned to pathways based on the pathway association of their known partners and their general functions can be unfolded. It should be noted that pathway association of new proteins are only predictions that should be further investigated by experiment. However, using binary similarity score of each pair we can rank proteins candidate for a particular metabolic function. Two examples of function inference are illustrated in Figures 6.2 and 6.3.

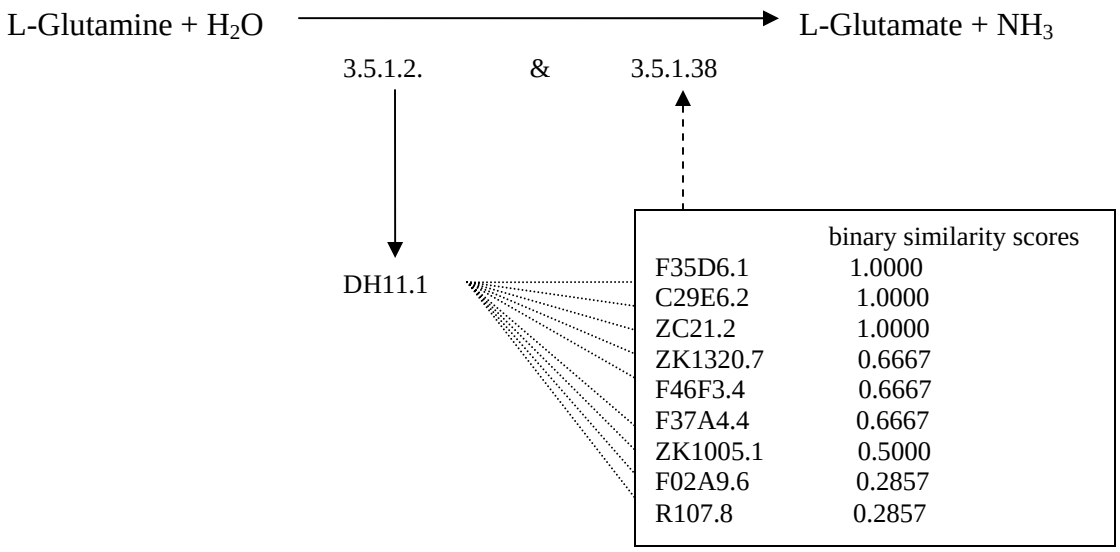

Figure 6.2. Inferring gene function. Coding genes of enzyme 3.5.1.38 may be one of the predicted interacting partners with gene DH11.1 based on their binary similarity scores.

Figure 6.2 describes an example of conclusions made based on predicted protein-protein interactions. In this example, DH11.1 is a gene that encodes for the enzyme catalyzing the hydrolysis of glutamine to glutamate. This enzyme participates in glutamate metabolism pathway and interacts with 49 other enzymes in the current protein

interaction map. In the new predicted interaction dataset, this enzyme is the known partner for 9 uncharacterized proteins. Thus, it can be concluded that the primary prediction for pathway association of these 9 unknown proteins is glutamate metabolism pathway. In order to further investigate the molecular functions of these 9 unknown proteins, and to support the primary prediction of pathway assignment, we should note that the reaction of hydrolysis of glutamine to glutamate is catalyzed by two enzymes including hypothetical protein 3.5.1.2 and hypothetical protein 3.5.1.38. The enzyme 3.5.1.2 is encoded by gene DH11.1 and the second enzyme is encoded by an unknown gene. Each of 9 predicted interacting proteins with DH11.1 could be a possible candidate for encoding enzyme 3.5.1.38 as illustrated in Figure 6.2. These candidates can be ranked based on their binary similarity scores as a criterion for experimental investigation. As seen in Figure 6.2, genes F35D6.1, C29E6.2, and ZC21.2 are more likely to be the coding gene of enzyme 3.5.1.38, since their binary similarity scores are the highest.

Some genes encode enzymes which appear in parallel reactions indicating that there may be an interaction between the genes which are encoding theses rival enzymes. For instance, as shown in Figure 6.3, ADP-ribose is converted to ribose-5-phosphate catalyzed by either ADP-ribose diphosphatese (3.6.31.13) in one step (reaction A) or ADP-sugar diphosphatase (3.6.1.21) and phosphopentomutase (5.4.2.7) in two steps (reactions B and C) in Purine metabolism pathway (cel00230). Enzyme 3.6.1.13 is encoded by gene W02G9.1, but the two other enzymes are still uncharacterized in terms of the encoding genes. W02G9.1 is found to interact with 7 other genes whose binary similarity scores are ranging from 0.5 to 1. Thus, genes with higher scores equally likely express the two uncharacterized enzymes 3.6.1.21 and 5.4.2.7 in reactions B and C, respectively. It is also predicted that these seven genes participate in Purine metabolism pathway (cel00230) along with previously known partners. There may be some other possibilities for the function assignment of each seven unknown genes in this example, however each assignment based on computational predictions need to be confirmed by experiment. The advantage of these computational assignments is that they can narrow down the number of experiments to confirm a link.

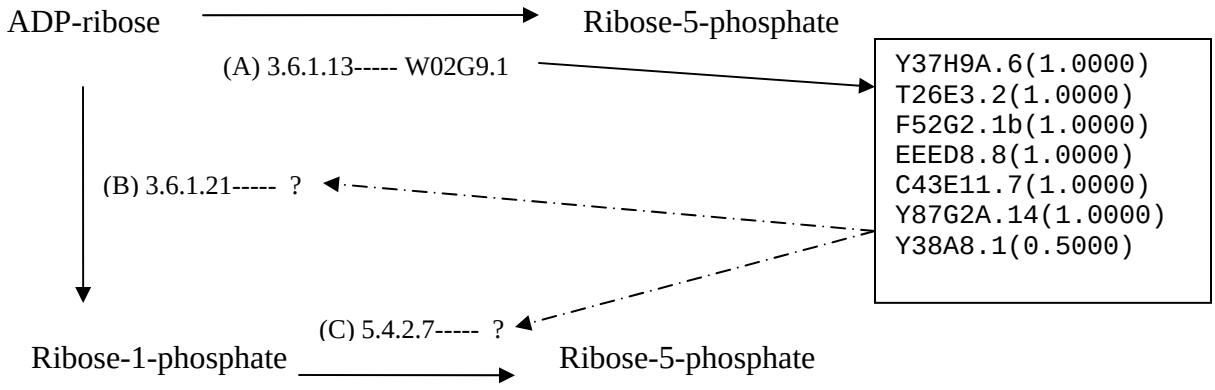

Figure 6.3. Inferring the possible enzymes encoded by *C. elegans* unknown genes. Each of the seven specified genes which are predicted to interacting with W02G9.1 may code either enzyme B or C.

6.5 Conclusion

Metabolic network of *C. elegans* has not been completely reconstructed yet. However, the efforts to understand the complex metabolic system of this multi-cellular organism are on going. Reconstruction of metabolic networks relies on the existing genomic information and organization of this information in a network context. The more the available genomic information, the more complete the network will be. Protein-protein interaction information is the key genomic data upon which the networks are built. Computational techniques are part of the tools available to predict protein interaction datasets; however, these approaches are not yet advanced enough to predict protein interactions with high reliability. In this chapter the new predictions were incorporated into the reconstructed metabolic network through pathway and possibly reaction assignment of newly characterized proteins. In the light of new assignments, the number of known metabolic proteins in this organism increased 27% and 1024 new interactions were all distributed in 94 metabolic pathways in the network of *C. elegans*. The expanded metabolic network is part of the efforts to complete the reconstruction of full metabolic network of *C. elegans* which is far to achieve. This expanded network provided guidelines to direct researches to design new experiments which focus on

determining substrate specificity of newly annotated enzymes and other detail information about metabolic reactions such as directionality and stoichiometry. Therefore, computational approaches and experimental techniques together are the two ways by which metabolic networks can be fully discovered to understand the sophisticated cellular activities inside living organisms.

References

Adai A.T., Date S.V., Wieland S., Marcotte E.M. (2004) LGL : creating of protein function with an algorithm for visualizing very large biological networks. *J. Mol. Biol.*, **340**, 179-190.

Beecher C. (2002) Metabolomics: A new "omics' technology. *American Genomics / Proteomics Technology,* **2(3),** 40-43.

Brinkworth R.I., Breinl R.A., Kobe B. (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl. Acad. Sci.* USA, **100(1)**, 74-79.

Carpenter A.E., Sabatini D.M. (2004) Systematic genome-wide screens of gene function. *NatureReviews Genetics,* **5,** 11-22.

Castrillo J.I., Hayes A., Mohammed S., Gaskell S.J., Oliver S.G. (2003) An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. *Phytochemistry,* **62,** 929-937.

Chen J., Yuan B. (2006) Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics,* **22(18):**2283-2290.

Clemente J.C., Satou K., Valiente G. (2005) Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and gene ontology. *Genome Informatics,* **16(2):**45-55.

Date S., Marcotte E.M. (2001) Exploiting big biology: integrating large-scale biological data for function inference. *Briefings in Bioinformatics*, **2(4)**, 363-374.

Fernie A.R. (2003) Metabolome characterization in plant system analysis. *Functional Plant Biology,* **30,** 111-120.

Fiehn O. (2001) Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp Funct Genom,* **2,** 155-168.

Fiehn O., Kloska S., Altmann T. (2001) Integrated studies on plant biology using multiparallel techniques. *Current Opinion in Biotechnology,* **12,** 82-86.

Fraser A.G., Kamath R.S., Zipperlen P., Martinez-Campos M., Sohrmann M., Ahringer J. (2000) Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature,* **408,** 325-330.

Fraser A., Marcotte E.M. (2004) A probabilistic view of gene function. *Nature Genetics*, **36(6)**, 559-564.

Hall R., Beale M., Fiehn O., Hardy N., Sumner L., Bino R. (2002) Plant metabolomics: The missing link in functional genomics strategies. *The plant Cell*, **14,** 1437-1440.

Jansen R., Gerstein M. (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Current Opinion in Microbiology*, **7**, 535-545.

Jeong H., Tombor B., Albert R., Oltvai Z.N., Barabasi A.-L. (2000) The large-scale organization of metabolic networks. *Nature,* **407:** 651-654.

Kell D.B. (2002) Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules. *Molecular Biology Reports,* **29,** 237-241.

Kitano H., Funahashi A., Matsuoka Y., Oda K. (2005) Using process diagrams for the graphical representation of biological networks. *Nature Biotechnology*, **23(8)**, 961-966.

Krijgsveld J., Ketting R.F., Mahmoudi T., Johansen J., Artal-Sanz M., Verrijzer C.P., Plasterk R.H.A., Heck A.J.R. (2003) Metabolic labelling of *C. elegans* and *D. melanogaster* for quantitative proteomics. *Nature Biotechnology*, **21(8),** 927-931.

Letovsky S., Kasif S. (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics,* **19 Suppl,** i197-i204.

Luyf A.C.M., de Gast J., van Kampen A.H.C. (2002) Visualizing metabolic activity on a genome-wide scale. *Bioinformatics*, **18(6),** 813-818.

Mastellos D., Andronis C., Persidis A., Lambris J.D. (2005) Novel biological networks modulated by complement. *Clinical Immunology,* **115,** 225-235.

Popescu L., Yona G. (2005) Automation of gene assignment to metabolic pathways using high-throughput expression data. *BMC Bioinformatics*, **6**:217.

Raamsdonk L.M., Teusink B., Broadhurst D., Zhang N., Hayes A., Walsh M.C., Berden J.A., Brindle K.M., Kell D.B., Rowland J.J., Westerhoff H.V., van Dam K., Oliver S.G. (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology,* **19,** 45-50.

Roessner U., Willmitzer L., Fernie A.R. (2002) Metabolic profiling and biochemical phenotyping of plant systems. *Plant Cell Rep,* **21,** 189-196.

Strong M., Graeber T.G., Beeby M., Pellegrini M., Thompson M.J., Yeates T.O., Eisenberg D. (2003) Visualization and interpretation of protein networks in *Mycobacterium tuberculosis* based on hierarchical clustering of genome-wide functional linkage maps. *Nucleic Acids Research*, **31(24)**, 7099-7109.

Sugimoto A. (2004) High-throughput RNAi in *Caenorhabditis elegans*: genome-wide screens and functional genomics. *Differentiation,* **72,** 81-91.

Thomas G.H. (2001) Metabolomics breaks the silence. *TRENDS in Microbiology,* **9(4),** 158-159.

Trethwey R.N. (2001) Gene discovery via metabolic profiling. *Current Opinion in Biotechnology*, **12,** 135-138.

van Nimwegen E. (2003) Scaling laws in the functional content of genomes. *TRENDS in Genetics,* **19,** 479-484.

Varner J., Ramkrishna D. (1999) Mathematical models of metabolic pathways. *Current Opinion in Biotechnology,* **10,** 146-150.

von Grotthuss M., Plewczynski D., Ginalski K., rychlewski L., Shakhnovich E.I. (2006) PDB-UF: database of predicted enzymatic functions for unannotated protein structures from structural genomics. *BMC Bioinformatics,* **7:**53.

Weckwerth W., Fiehn,O. (2002) Can we discover novel pathways using metabolomic analysis? *Current Opinion in Biotechnology,* **13,** 156-160.

Wiechert W. (2002) Modeling and simulation: tools for metabolic engineering. *Journal of Biotechnology,* **94,** 37-63.

7

GENERAL DISCUSSION

7.1 Discussion

Proteins in cell are not independent entities instead they create associations to perform a biological task. Thus, identification of biological associations is essential to understand cellular activities and then integrate them in a network context which provides a larger picture of all cellular activities in an organism. Due to the complicated multi cellular structure of *C. elegans* only a small fraction of proteins and their interactions in this species has been elucidated. Thereby, the metabolic network of *C. elegans* is still under investigation and more research is yet to be done to achieve a complete portrait of metabolic processes in this organism. More robust and comprehensive protein-protein interaction datasets need to be available to be accommodated into the network to expand the current metabolic network. Experimental techniques to screen protein-protein interactions are expensive and time consuming. As an alternative, computational approaches have been widely used to detect more protein-protein interactions in a less time consuming way and supply adequate information to improve metabolic reconstruction studies. However, computational approaches are not only inaccurate but also suffer from mass false positive predictions. The predicted datasets need to be refined to improve the reliability of final links. The contribution of this research to all concerns mentioned above is clarified in the remainder of this section. These concerns form the objectives of this research as outlined in Chapter 2.

The first objective of this study was reconstructing the metabolic network of the studied organism based on current genomic information obtained form public databases. As described in chapter 3 the current metabolic network of *C. elegans* was reconstructed and 792 known proteins were specified in 94 metabolic pathways. These proteins were involved in 32902 pair-wise interactions called as current protein-protein interaction map. The 792 known proteins and the relationships among their encoding genes were used to create the visualized representation of the reconstructed network. This network demonstrated the current situation of known proteins within the genome. The network

became more meaningful in conjunction with those functionally-assigned genes which contributed to the interconnectivity of the system and were represented by an undirected two-mode graph to investigate its topological property. In the resulted protein-protein interaction map each protein was connected to 42 other proteins by average and some proteins had partners in 15 different pathways. Protein relationships outside pathway boundaries contributed to the interconnectivity of the network which revealed alternative routes to synthesize essential metabolites at different environmental conditions. Analysis of the network showed that how reactions and enzymes at different pathways were working together to accomplish a biological task. This reconstructed network consisted of 792 known metabolic proteins which accounts for approximately 17% of proteins involved in metabolism and 3.5% of all genes in *C. elegans* genome.

In comparison with previously reconstructed networks, discussed in Chapter 3, the approach employed here was more relying on genomic information and protein-protein interaction data. In other reconstruction strategies metabolites were the center points. Because of the shortage of genomic information in the past, relationship among metabolites, i.e. reactions, were either directly searched through public databases or identified by means of experiments. In other words, the required reaction information to reconstruct the network was collected manually. Interactions were considered only physical contacts dealing with a pair of reactant and product. In our strategy, pathways played the central role and the knowledge of association of genes and proteins in pathways was required to reconstruct the network. In order to assign genes and proteins to pathways, protein-protein interaction information was essential. This information came from genomic data of different organisms which do not solely focus on physical contacts. One aspect of genomic data is signature content information of proteins. This information was used in this research to elucidate protein-protein interactions for the fulfilment of the second objective.

Given the importance of protein interaction information in reconstructing metabolic networks, in Chapter 4 a new method of predicting protein-protein interactions was proposed. The underlying hypothesis of this method was based on the observation that proteins interact with each other through their functionally independent, structurally conserved, and biologically related signatures when they have some signatures in

common. These properties established new insight into the prediction of protein-protein interactions. Existing domain-based prediction methods used the interaction probability score between two signatures. The scoring function was trained based on a learning dataset and subsequently applied to predict protein interactions. In contrast, the proposed approach did not require training information and proteins were directly paired based on their signature contents, providing that they had at least one signature in common. Removing proteins with a low number of known signatures (one and two signatures) from the dataset the confidence level of the prediction significantly increased. Thus, as more and more proteins with known signature contents across organisms are discovered, the coverage and accuracy of protein interacting pairs predicted by this approach is expected to rise. The proposed method was applied to three model organisms including *S. cerevisiae*, *C. elegans*, and *H. sapiens* resulting in three predicted PPI datasets that contained many novel pair interactions. Critical comparison between the proposed approach and similar approaches was performed in Chapter 4. The predicted PPI pairs along with other datasets predicted by other computational methods were used as potential building blocks of reconstructing metabolic networks.

In order to increase the reliability of protein pair interactions predicted by means of all computational methods including the proposed method, as targeted in the third objective of this research, a filtering algorithm was proposed in Chapter 5 to partially remove false positive interactions from predicted datasets. The algorithm utilized gene ontology annotation as a common ground to specify computational predictions which were confirmed by experimental results. Molecular function annotations of experimentally confirmed protein pairs were used as the training set to extract discriminating keywords which well represented the training set. Then based on the extracted keywords two heuristic rules were set. The rules were incorporated into an algorithm by which predicted datasets were filtered and false positive predictions were partially removed from the datasets. Statistical analyses showed that keywords were over-represented words in the datasets and only eight keywords were significantly able to recover molecular function annotations of experimental and computational interacting proteins. Furthermore, applying the algorithm to specified datasets improved the true positive fractions of the datasets compared to random pairing. The improvement varied

among datasets depending on the approach utilized to predict protein relations. The approach was unbiased toward different datasets. It can be embedded to all computational protein-protein interaction prediction methods. Currently, no genome is fully annotated in Gene Ontology and there are many genes yet to be annotated. However, the proposed approach could be readily applied to newly annotated genes to predict their functional or physical partners.

Eventually, in pursuing the fourth objective of this research, with the availability of current metabolic network of *C. elegans*, and the novel protein-protein interactions predicted by the proposed method which was further filtered by the proposed algorithm to partially remove false positive interactions, we integrated the novel predictions to the current metabolic network of *C. elegans* in Chapter 6. As a result, an augmented network was reconstructed. In the light of new assignments, the number of known metabolic proteins in this organism increased 27% and 1024 new interactions were all distributed into 94 metabolic pathways in the network of *C. elegans*. In the augmented network, known metabolic proteins increased to 1009 which accounted for 22% of *C. elegans* genes involved in metabolism and 4.4% of all genes in the genome. Connections in the network provided guidelines to direct researchers to design new experiments that focus on determining substrate specificity of newly annotated enzymes. Computational prediction of protein-protein interactions were able to narrow down the direction of future experiments and raised new thoughts to discover new proteins and their functions.

7.2 Conclusions and Recommendations

Overall, the metabolic network of *C. elegans* was reconstructed using current genomic information available in KEGG database. Proposing a new computational method to predict protein-protein interactions, the reconstructed metabolic network was expanded and more proteins were assigned function. As the reliability of genomic information incorporated into metabolic networks is crucial, a global framework was also established to increase the true positive fraction of predicted datasets and reduce the number of false positives. Thus, a new strategy based mostly upon genomic information was employed to reconstructing metabolic networks. In the initial step of reconstructing the metabolic network of *C. elegans* it was shown that the graph representation of the network reveals hidden mechanisms through which the organism may survive under

different environmental circumstances. With the discovery of more protein-protein interactions the network became further complete and more protein functions became known. The proposed computational method to predict protein interactions performed well, or even better, than equivalent methods in terms of prediction power. However, since all computational methods predict true positive interactions along with numerous false positives, predicted datasets were required to be filtered to increase the true positive fraction.

Therefore, the metabolic network of *C. elegans* became more complete using genomic information available through public sources. This approach can be applied to any species and in this research *C. elegans* was chosen as a model organism.

In order to pursue the pace of this research a few recommendations are made as follows:

1. With the emerging high-throughput screening techniques and more computational methods, more in-depth data mining approaches need to be utilized to collect as much genomic information as possible to obtain more complete metabolic networks.
2. Genomic databases and function-based references are updated on a regular basis. As metabolic networks are constructed based on that type of information, they need to be updated regularly and new information should be accommodated in the network. Changes and obsolete items should be abandoned in the updated version of the network.
3. Biological databases sometimes contain inconsistent data with other sources. It is recommended to use information that is more common among databases with more rigid referencing. Basically, when a piece of information is documented by at least two databases it is more reliable.
4. Functions inferred from the expanded version of the metabolic network are solely candidates for further experiments. Experimental investigations, narrowed down by these predictions, can reveal the practical interactions.

APPENDIX

Supplementary data is provided in a compact disc (included) and arranged based on chapters. Input/output data files and perl scripts discussed in each chapter are collected in the same folder. A complete listing of folders and their contents is as follows:

Folder	No.	File name	Description
Chapter 3	1	*celGene.txt*	A complete listing of *C. elegans* genes
	2	*celPath.txt*	A complete listing of 103 pathways in *C. elegans* genome including 94 metabolic pathways
	3	*celReact.txt*	A complete listing of metabolic reactions carried out by different enzymes in *C. elegans* metabolic pathways
	4	*celNetwork.txt*	Reconstructed metabolic network of *C. elegans* based on current information
	5	*celNetwork_prg.pl*	Source perl script that integrates information provided by files 1-3 to reconstruct the metabolic network.
	6	*currentPPImap.txt*	Protein-protein interactions obtained from the current metabolic network.
Chapter 4	1	*cel_experimental_data.txt*	A complete listing of 344 experimentally confirmed protein-protein interactions in *C. elegans*.
	2	*hsa_experimental_data.txt*	A complete listing of 13319 experimentally confirmed protein-protein interactions in *H. sapiens*.
	3	*sce_experimental_data.txt*	A complete listing of 3745 experimentally

			confirmed protein-protein interactions in *S. cerevisiae*.
	4	r*eference_genomes.doc*	A complete listing of 90 reference genomes utilized in BLAST program for phylogenetic profile method.
	5	*cel_signature_pairs_no_removal.txt*	Protein-protein interactions in *C. elegans* predicted by signature profiling method and their corresponding binary similarity scores.
	6	*cel_signature_pairs_1signature_removal.txt*	Protein-protein interactions in *C. elegans* predicted by signature profiling method and their corresponding binary similarity scores while proteins with ONE known signatures were deleted from genome.
	7	*cel_signature_pairs_2signature_removal.txt*	Protein-protein interactions in *C. elegans* predicted by signature profiling method and their corresponding binary similarity scores while proteins with TWO known signatures were deleted from genome.
	8	*hsa_signature_pairs_no_removal.txt*	Protein-protein interactions in H. *sapiens* predicted by signature profiling method and their corresponding binary similarity scores
	9	*hsa_signature_pairs_1signature_removal.txt*	Protein-protein interactions in *H. sapiens* predicted by

			signature profiling method and their corresponding binary similarity scores while proteins with ONE known signatures were deleted from genome.
	10	*hsa_signature_pairs_2signature_removal.txt*	Protein-protein interactions in *H. sapiens* predicted by signature profiling method and their corresponding binary similarity scores while proteins with TWO known signatures were deleted from genome.
	11	*sce_signature_pairs_no_removal.txt*	Protein-protein interactions in *S. cerevisiae* predicted by signature profiling method and their corresponding binary similarity scores
	12	*sce_signature_pairs_1signature_removal.txt*	Protein-protein interactions in *S. cerevisiae* predicted by signature profiling method and their corresponding binary similarity scores while proteins with ONE known signatures were deleted from genome.
	13	*sce_signature_pairs_2signature_removal.txt*	Protein-protein interactions in *S. cerevisiae* predicted by signature profiling method and their corresponding binary similarity scores while proteins with TWO known signatures were deleted from genome
	14	*sce_mle_pairs.txt*	Protein-protein interactions in *S.*

			cerevisiae predicted by Maximum likelihood estimation (MLE) and their interaction probabilities.
	15	*cel_phylogenetic_pairs.txt*	Protein-protein interactions in *C. elegans* predicted by phylogenetic profiles method.
	16	*hsa_phylogenetic_pairs.txt*	Protein-protein interactions in *H. sapiens* predicted by phylogenetic profiles method.
	17	*sce_phylogenetic_pairs.txt*	Protein-protein interactions in *S. cerevisiae* predicted by phylogenetic profiles method.
	18	*cel_gene_expression_pairs.txt*	Protein-protein interactions in *C. elegans* predicted by gene expression profiling method.
	19	*hsa_gene_expression_pairs.txt*	Protein-protein interactions in *H. sapiens* predicted by gene expression profiling method.
	20	*sce_gene_expression_pairs.txt*	Protein-protein interactions in *S. cerevisiae* predicted by gene expression profiling method.
	21	*signature-profiling-prosite.pl*	source perl script for signature profiling method
	22	*OtherMethods.pl*	source perl script for implementing phylogenetic profiles and gene expression methods.
	23	*MLE.pl*	source perl script for implementing maximum likelihood

			estimation (MLE) method.
Chapter 5	1	*sce_experimental_dataset.txt*	A complete listing of experimentally confirmed protein-protein interactions in *S. cerevisiae* compiled from multiple sources.
	2	*cel_experimental_dataset.txt*	A complete listing of experimentally confirmed protein-protein interactions in *C. elegans* compiled from multiple sources.
	3	*reference_genomes.doc*:	Complete listing of 90 reference genomes utilized in BLAST program for phylogenetic profile method.
	4	*sce_PP_raw_dataset.txt*	Predicted protein-protein interactions in *S. cerevisiae* using phylogenetic profiles method
	5	*cel_PP_raw_dataset.txt*	Predicted protein-protein interactions in *C. elegans* using phylogenetic profiles method.
	6	*sce_GE_raw_dataset.txt*	Predicted protein-protein interactions in *S. cerevisiae* using gene expression profiles.
	7	*cel_GE_raw_dataset.txt*	Predicted protein-protein interactions in *C. elegans* using gene expression profiles.
	8	*sce_CC_raw_dataset.txt*	Predicted protein-protein interactions in *S. cerevisiae* using chance co-occurrence distribution method.
	9	*cel_CC_raw_dataset.txt*	Predicted protein-protein interactions in *C. elegans* using

			chance co-occurrence distribution method
	10	*sce_MLE_raw_dataset.txt*	Predicted protein-protein interactions in *S. cerevisiae* using maximum likelihood estimation method.
	11	*cel_MLE_raw_dataset.txt*	Predicted protein-protein interactions in *C. elegans* using maximum likelihood estimation method
	12	*clusters_and_keywords.txt*	A complete listing of GO term clusters and their representative keywords in *S. cerevisiae* and *C. elegans* training sets.
	13	*sce_PP_filtered_dataset.txt*	A filtered protein-protein interaction dataset of *S. cerevisiae* using proposed algorithm. Interactions are predicted by phylogenetic profiles method.
	14	*cel_PP_filtered_dataset.txt*	A filtered protein-protein interaction dataset of *C. elegans* using proposed algorithm. Interactions are predicted by phylogenetic profiles method.
	15	*sce_GE_filtered_dataset.txt*	A filtered protein-protein interaction dataset of *S. cerevisiae* using proposed algorithm. Interactions are predicted by gene expression method.
	16	*cel_GE_filtered_dataset.txt*	A filtered protein-protein interaction dataset of *C. elegans* using proposed algorithm. Interactions

			are predicted by gene expression method.
	17	*sce_CC_filtered_dataset.txt*	A filtered protein-protein interaction dataset of *S. cerevisiae* using proposed algorithm. Interactions are predicted by chance co-occurrence distribution method
	18	*cel_CC_filtered_dataset.txt*	A filtered protein-protein interaction dataset of *C. elegans* using proposed algorithm. Interactions are predicted by chance co-occurrence distribution method.
	19	*sce_MLE_filtered_dataset.txt*	A filtered protein-protein interaction dataset of *S. cerevisiae* using proposed algorithm. Interactions are predicted by maximum likelihood estimation method.
	20	*cel_MLE_filtered_dataset.txt*	A filtered protein-protein interaction dataset of *C. elegans* using proposed algorithm. Interactions are predicted by maximum likelihood estimation method.
	21	*CC.pl*	Source perl script to implement chance co-occurrence method and proposed algorithm to the resulted dataset.
	22	*GE.pl*	Source perl script to implement gene expression method and proposed algorithm to the resulted dataset.
	23	*PP.pl*	Source perl script to implement

			phylogenetic profiles method and proposed algorithm to the resulted dataset.
	24	*MLE.pl*	Source perl script to implement maximum likelihood method and proposed algorithm to the resulted dataset.
Chapter 6	1	*signature_profiling_novelPPIs.txt*	A complete listing of novel *C. elegans* PPIs predicted by signature profiling methods.
	2	*expandedPPImap.txt*	A complete listing of protein-protein interactions known in *C. elegans* genome
	3	*Function_assignment.doc*	A complete listing of *C. elegans* proteins whose functions were inferred based on new protein-protein interaction information.
	4	*metabolic_reconstruction.pl*	Source perl script to integrate new PPI information into the current map resulting in expanded PPI map